United Nations
Educational, Scientific and
Cultural Organization

Memory of
the World

Software Source Code
Heritage of the Digital Age

D. Storti, UNESCO
Communication and Information Sector

Software Heritage

United Nations Educational, Scientific and Cultural Organization

Memory of the World

Software Heritage

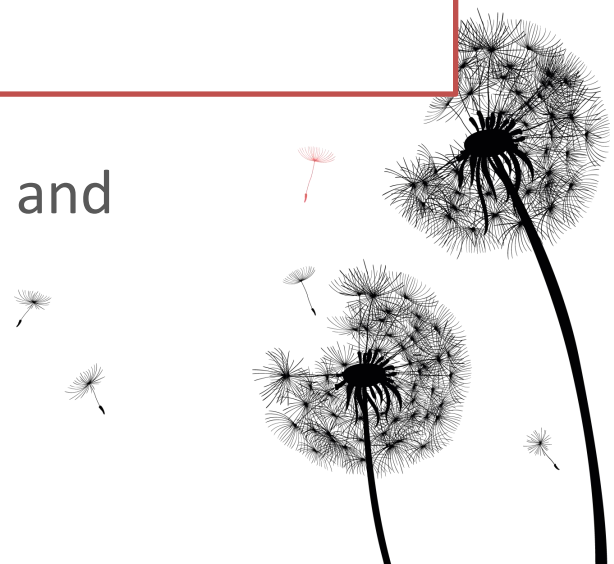
- Software is everywhere
- It embodies a collective knowledge and cultural heritage

```
4004e6: 55
4004e7: 48 89 e5
4004ea: bf 84 05 40 00
4004ef: b8 00 00 00 00
4004f4: e8 c7 fe ff ff
```

```
/* Hello World program */
#include <stdio.h>
Void mail ()
{
      printf("Hello World");
}
```

"Programs must be written for people to read, and only incidentally for machines to execute"

Harold Abelson, 1985



Margaret Hamilton, Lead software engineer for Project Apollo stands with Apollo's source code

United Nations Educational, Scientific and Cultural Organization

Memory of the World

Software Heritage

Code is increasingly complex:

Apollo modules counted

approx. 145,000 lines of code

United Nations Educational, Scientific and Cultural Organization

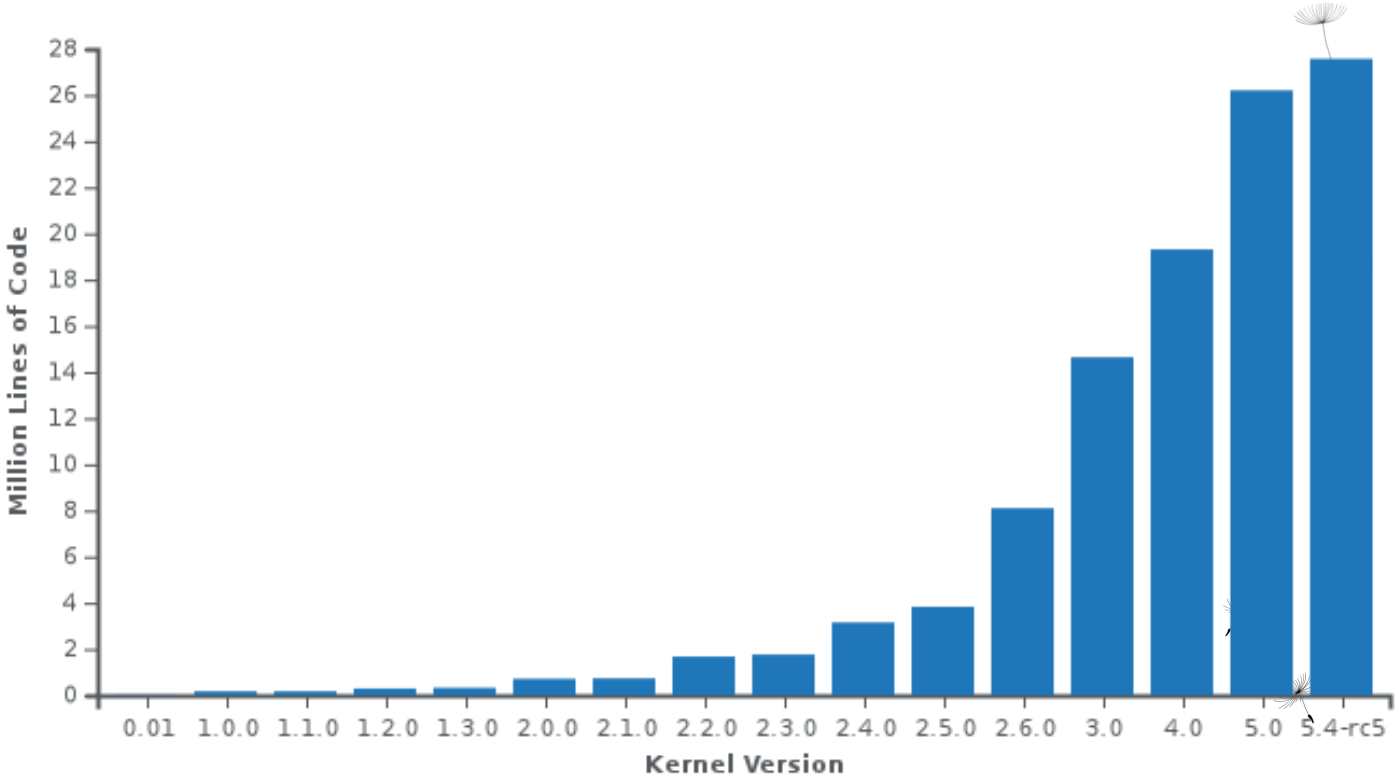
Memory of the World

Software Heritage

MacOS X Tiger = more than 86 million lines

Facebook = more than 60 million lines

Linux Kernel size in lines of code:



United Nations Educational, Scientific and Cultural Organization

Memory of the World

Software Heritage

Software is spread all around in many repositories

United Nations Educational, Scientific and Cultural Organization

Memory of the World

Software Heritage

Software is fragile and easily endangered

United Nations Educational, Scientific and Cultural Organization

Memory of the World

Software Heritage

While software **development** is skyrocketing:

- Scientific reproducibility is not guaranteed
- Research software is undervalued
- Software research is under-recognised

We are **loosing our heritage** of the digital age:

- Legacy software is abandoned or lost
- Key people are passing away
- We miss a catalog
- We miss an archive

UNESCO supports **Software Heritage**, the first and largest archive of software source code

Mission

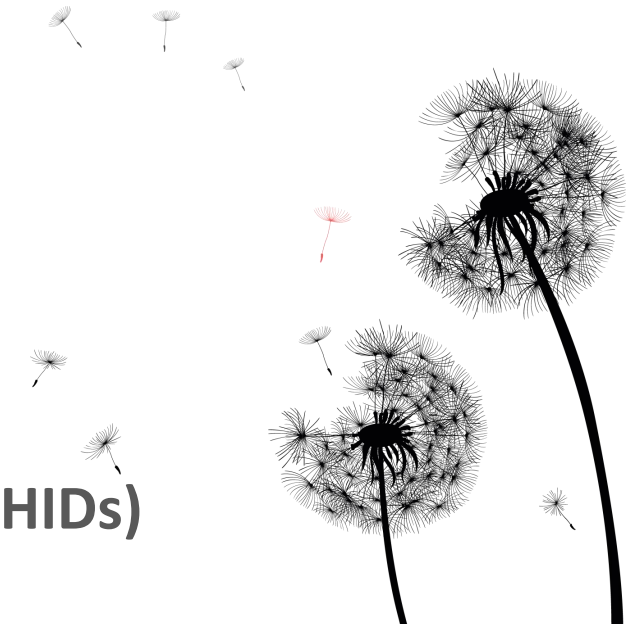
Collect, preserve and share the source code of all available software

The Great Library of Source code

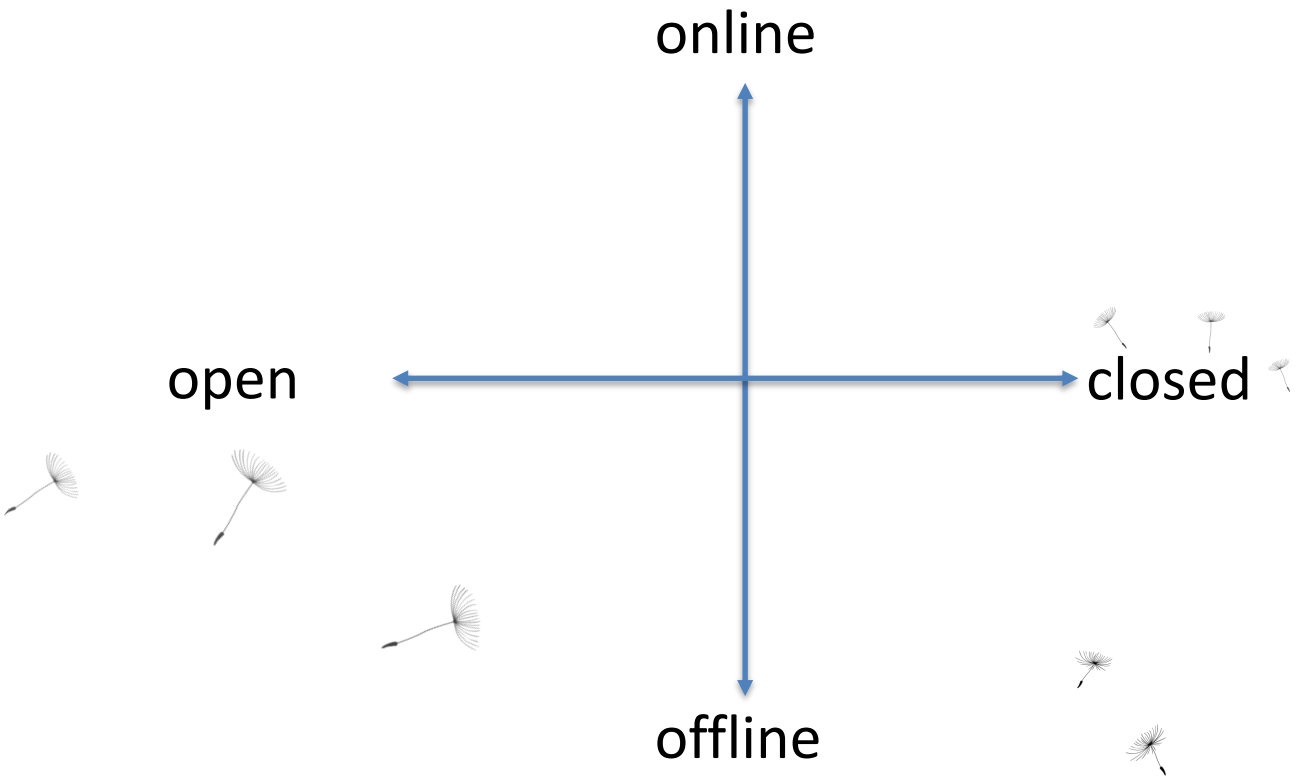
Launched in June 2018 at UNESCO

- 153+ million software projects archived
- ~10 billion source files

SoftWare Heritage persistent IDentifiers (SWHIDs)



How is it possible to collect all source code? It depends on its availability



United Nations Educational, Scientific and Cultural Organization

Memory of the World

Software Heritage

In November2018, a group of experts, from research, industry and memory institutions gathered in UNESCO and issued the Paris Call on Software Source Code as Heritage for Sustainable Development

> “[We call to] support efforts to gather and preserve the artifacts and narratives of the history of computing, while the earlier creators are still alive”

https://en.unesco.org/foss/paris-call-software-source-code

United Nations Educational, Scientific and Cultural Organization

Memory of the World

Software Heritage

In October 2019, **UNESCO**, the **University of Pisa** and **Inria** responded to the Call issuing **SWHAP**, the first detailed guidelines to curate and properly archive landmark legacy source code on Software

United Nations Educational, Scientific and Cultural Organization — Memory of the World

Software Heritage

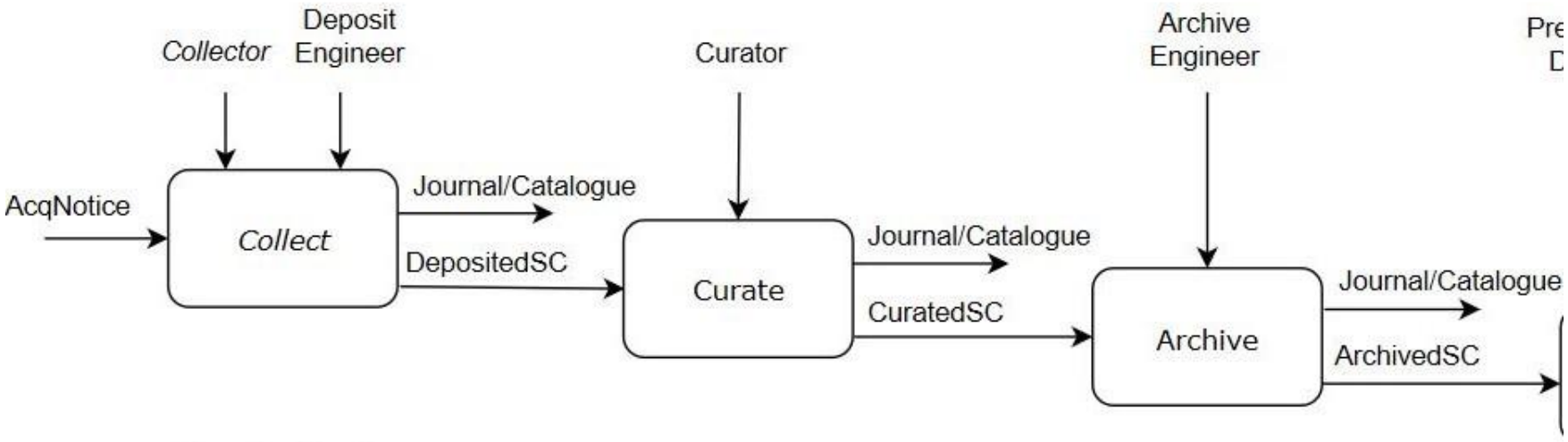
SWHAP

Software Archeology

The process is inspired by a naïf view of the way archeologists work:

- First, on site, they collect and identify the finds.
- Then, in the museum, they safely store, curate, and exhibit them.
- Often, they come back on site for getting more information

Obviously, reality is much more complex and intertwined



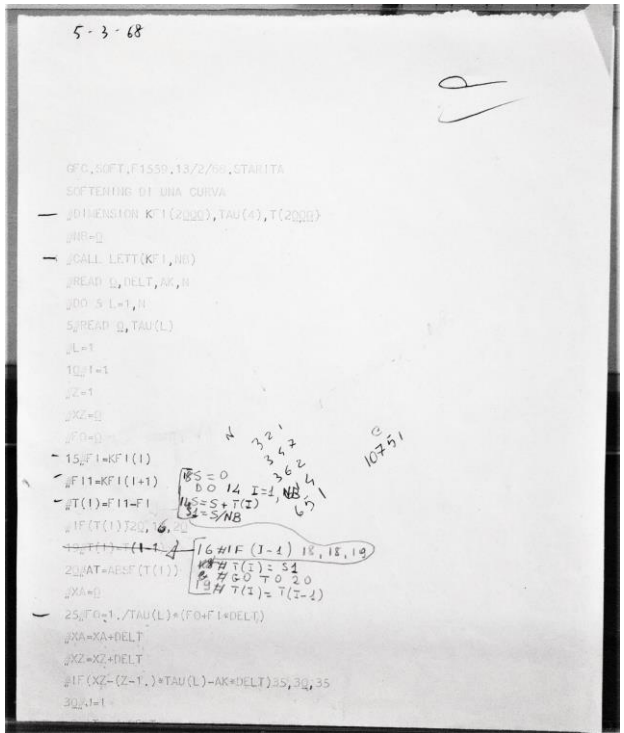
United Nations Educational, Scientific and Cultural Organization

Memory of the World

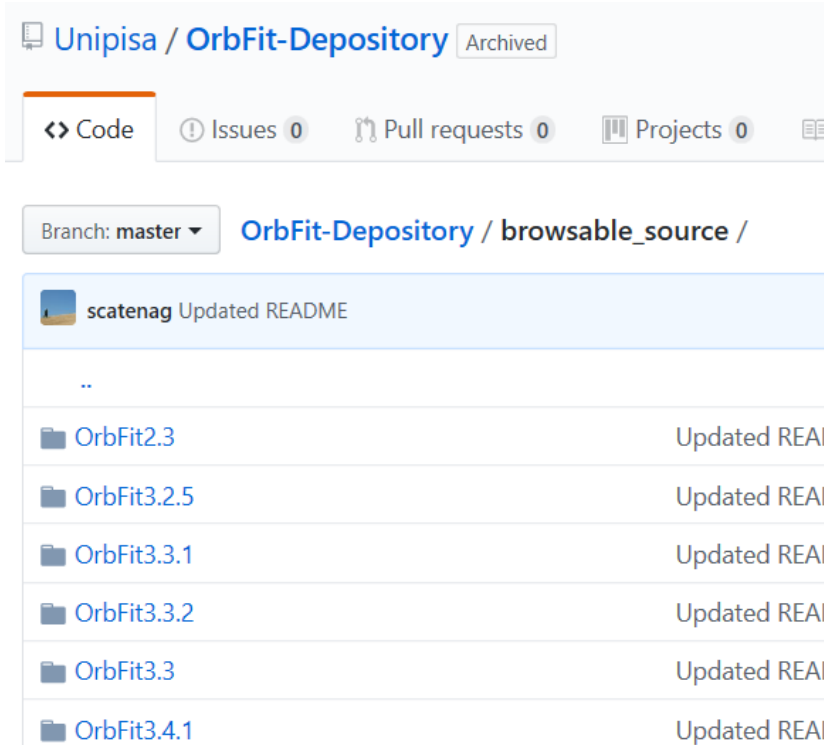
Software Heritage

The Collect phase

- to find the source code and related materials and stories
- gather it as they are in a physical and/or logical place where it can be properly archived for later processing
- diverse situations

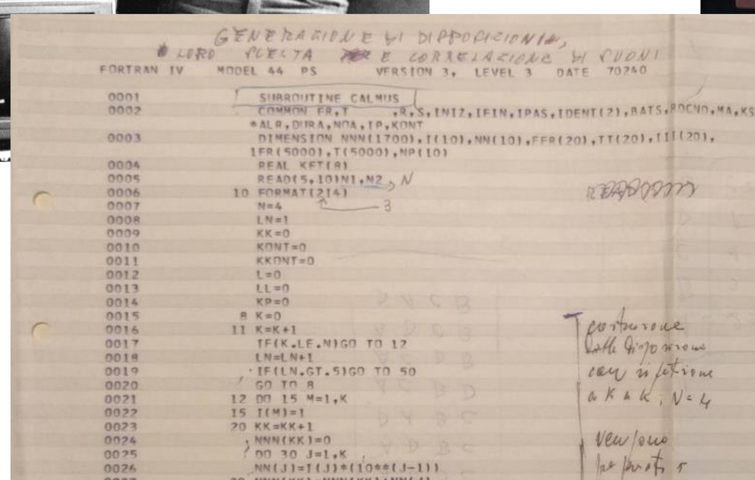


vs



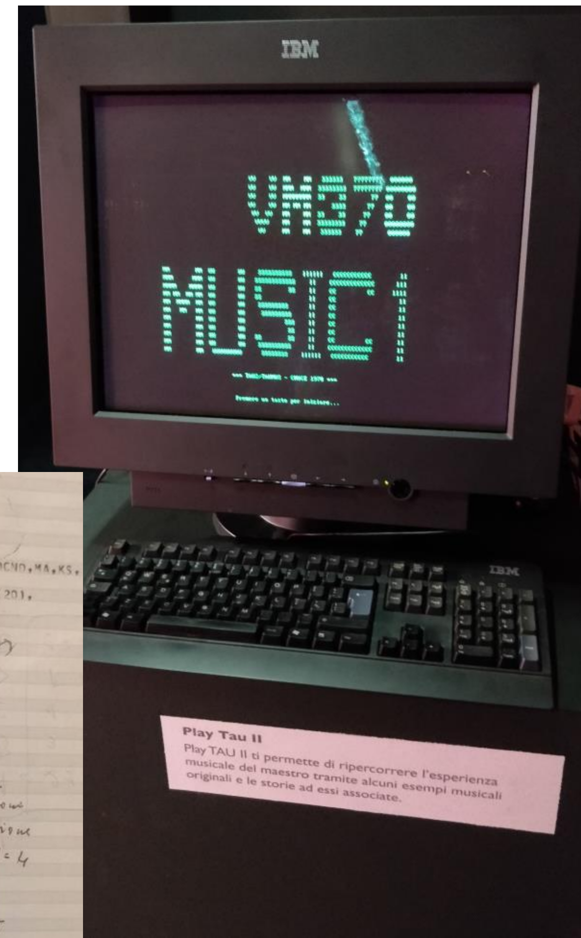
An example

TAUmus one of the first music synthetizers ('70s)



```
FORTRAN IV     MODEL 44  PS          VERSION 3,  LEVEL 3   DATE  70240

            SUBROUTINE CALMUS
            COMMON FR,T      ,R,S,INIZ,IFIN,IPAS,IDENT(2),BATS,ROCNO,MA,KS,
           *ALR,DURA,NOA,IP,KONT
            DIMENSION NNN(1700),I(10),NN(10),FFR(20),TT(20),III(20),
           1FR(5000),T(5000),NP(10)
            REAL KFT(8)
            READ(5,10)N1,N2
         10 FORMAT(2I4)
            N=4
            LN=1
            KK=0
            KONT=0
            KKONT=0
            L=0
            LL=0
            KP=0
          8 K=0
         11 K=K+1
            IF(K.LE.N)GO TO 12
            LN=LN+1
            IF(LN.GT.5)GO TO 50
            GO TO 8
         12 DO 15 M=1,K
         15 I(M)=1
         20 KK=KK+1
            NNN(KK)=0
            DO 30 J=1,K
            NN(J)=I(J)*(10**(J-1))
```



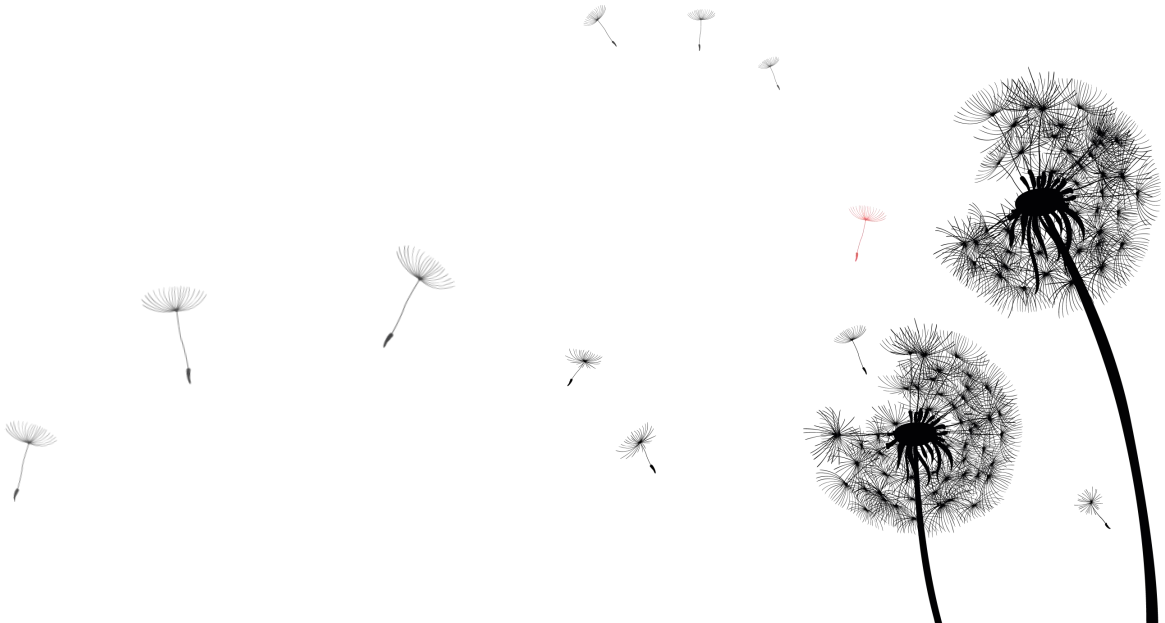
United Nations Educational, Scientific and Cultural Organization

Memory of the World

Software Heritage

Archiving Software Source Code is not just about code

It is also about preserving the history of todays' digital world: the stories of the women and men which made much of the most impactful innovations in human history possible.



United Nations Educational, Scientific and Cultural Organization

Memory of the World

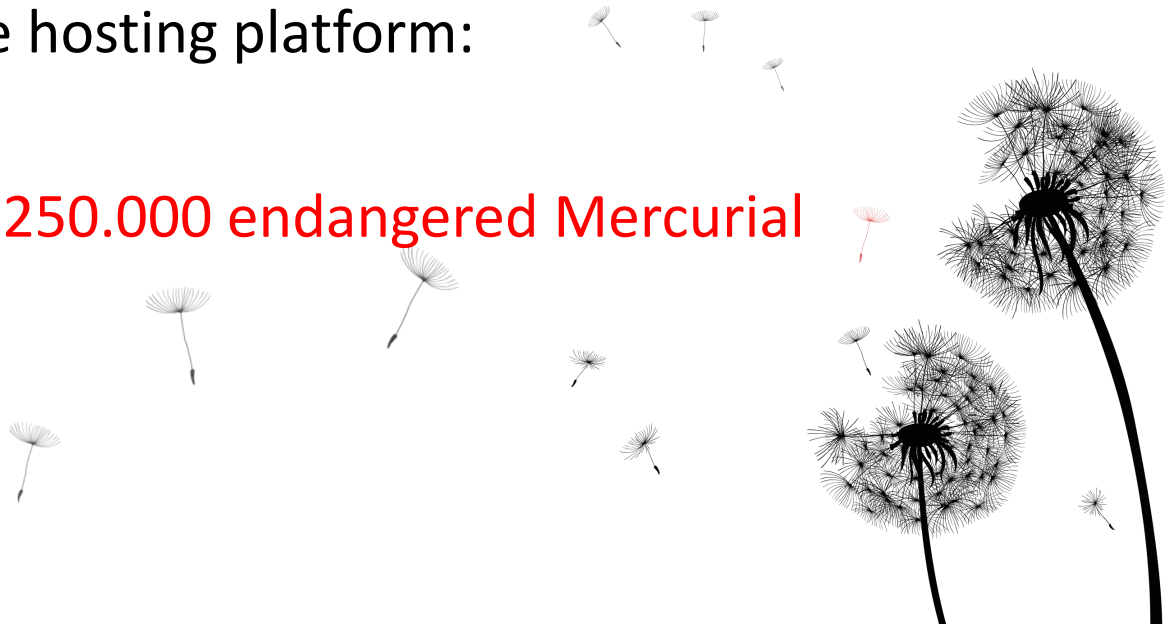
Software Heritage

Example: Mercurial

https://www.softwareheritage.org/2020/09/17/250-000-endangered-mercurial-repositories-rescued/

Following the discontinued support for the Mercurial version control system by Bitbucket code hosting platform:

Software Heritage saved 250.000 endangered Mercurial repositories!



United Nations Educational, Scientific and Cultural Organization

Memory of the World

Software Heritage

Open Science

The Paris Call and the SWHAP process need to be sustained

- Source code is a pillar of Open Science
  - Open Access; Open Data sets; Open Source
- Process vs Support (abstract vs concrete)
  - Warehouse, Depository, Curated source code deposit and Workbench
- Implementation requirements
  - Long term availability
  - Historical accuracy
  - Traceability
  - Openness
  - Interoperability

United Nations Educational, Scientific and Cultural Organization — Memory of the World

Software Heritage

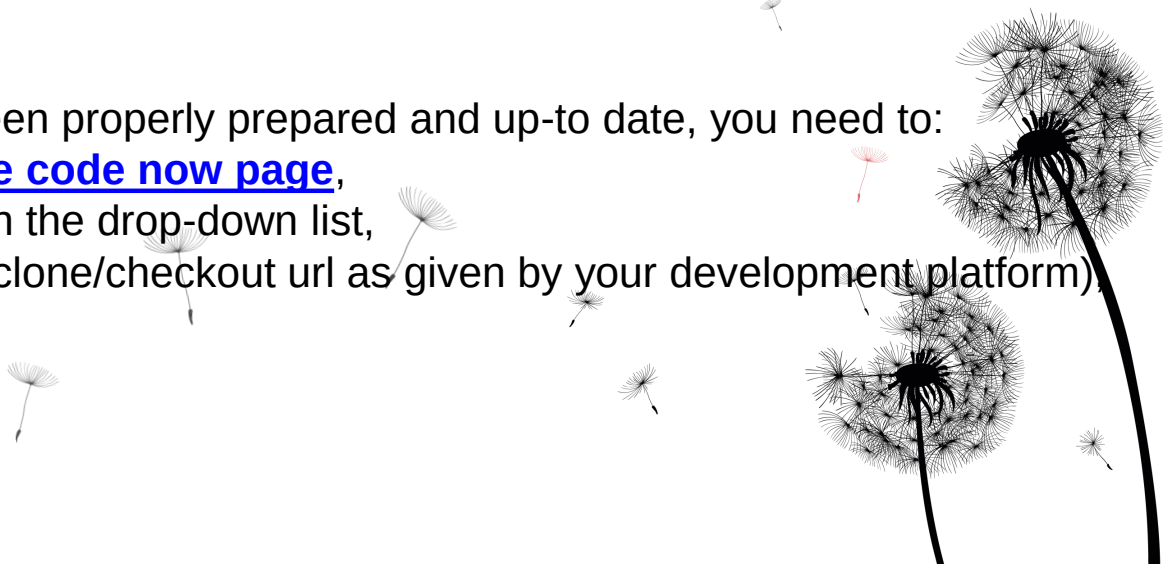
Open Science and software source code

Software Heritage allows to *archive seamlessly* your research software artifacts, and add to your research articles *precise references* to specific versions of the source code, down to *fragments of individual source files*. This allows to enhance significantly the experience of the reviewers of your work.

Step 1: prepare your public repository
Make sure your source code is hosted on a repository publicly accessible (Github, Bitbucket, a GitLab instance, an institutional software forge, etc.) using one of the version control systems supported by Software Heritage, currently Subversion, Mercurial and Git.

Step 2: save your code
Once your code repository has been properly prepared and up-to date, you need to:
go to **the Software Heritage save code now page**,
pick your version control system in the drop-down list,
enter the code repository url (the clone/checkout url as given by your development platform),
click on the Submit button.

That's it, it's all done!



United Nations Educational, Scientific and Cultural Organization

Memory of the World

Software Heritage

Thank you

https://www.softwareheritage.org/

https://www.softwareheritage.org/swhap/

https://en.unesco.org/foss/paris-call-software-source-code