Software Heritage

Does software preservation have an ethical impact on society?

Morane Gruenpeter

Software engineer and metadata specialist
Inria, Software Heritage

morane@softwareheritage.org

November 30th, 2020

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Outline

Short Bio: Morane Gruenpeter

Goal: Building the Semantic Web of Free and Open Source Software

1999-2011	Harpist
2012-2015	Licence in Computer Science CNAM
2015-2017	Master STL - M2 R&D UPMC
2017	Internship *Software Heritage* (SWH)
2018-2019	European project EU2020 *CROSSMINER* (on SWH team)
2020-2022	European project *FAIRsFAIR* (on SWH team)

Working groups for Open Science and digital preservation

- the Research Data Alliance's Software Source Code Interest Group (SSC IG),
- the FORCE11's Software Citation Implementation Working Group (SCI WG),
- the joint RDA & FORCE11 Software Identification Working Group (SCID WG)
- WikiData for Digital Preservation initiative (WikiDigi).

Goals today

Terminology

- understand terminology around software and source code

History

- a few elements to grasp recent and not so recent evolutions

Society

- software as a key to view society
- review initiatives working on software preservation

Software is all around us

Apollo 11 Guidance Computer (~60.000 lines), 1969

"When I first got into it, nobody knew what it was that we were doing. It was like the Wild West." Margaret Hamilton

The World Wide Web, 1989, at CERN on a NeXT machine

“When somebody has learned how to program a computer … You’re joining a group of people who can do incredible things. They can make the computer do anything they can imagine.”

From An Insight, An Idea with Tim Berners-Lee (2013)

What is software ?

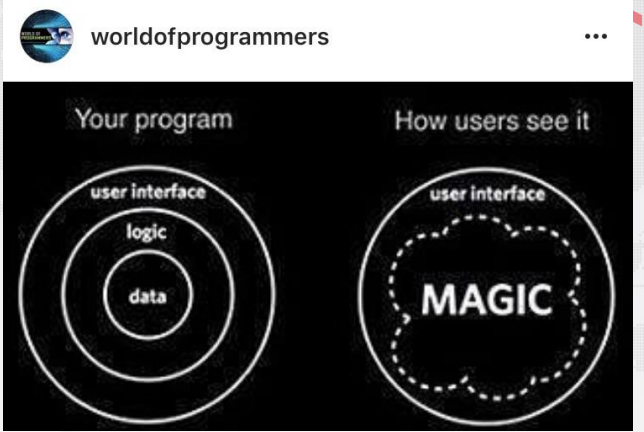


Image taken from from reddit - ProgrammerHumor

Software definition

Encyclopædia Britannica

"Software, instructions that tell a computer what to do. Software comprises the entire set of programs, procedures, and routines associated with the operation of a computer system. The term was coined to differentiate these instructions from hardware—i.e., the physical components of a computer system." link

Software as a concept

- software project / entity
- the creators and the community around it

Software artifact

- the binaries for different environments
- the software source code for each version

This is *software*?

What about *software source code* ?

Outline

The knowledge is in the source code!

"The source code for a work means the preferred form of the work for making modifications to it." GPL Licence

Hello World

Program (excerpt of binary)

```
4004e6: 55
4004e7: 48 89 e5
4004ea: bf 84 05 40 00
4004ef: b8 00 00 00 00
4004f4: e8 c7 fe ff ff
4004f9: 90
4004fa: 5d
4004fb: c3
```

Program (source code)

```
/* Hello World program */

#include<stdio.h>

void main()
{
    printf("Hello World");
}
```

Source code is *special*

Executable and *human readable* knowledge — copyright law

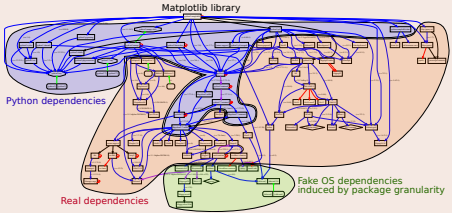
"Programs must be written for people to read, and only incidentally for machines to execute."
Harold Abelson

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*



Software Source Code human readable and executable knowledge

Full width | Home | Development | Documentation | Donate | login

Software Heritage Archive

Features

- Search
- Downloads
- Save code now
- Help

Permalinks

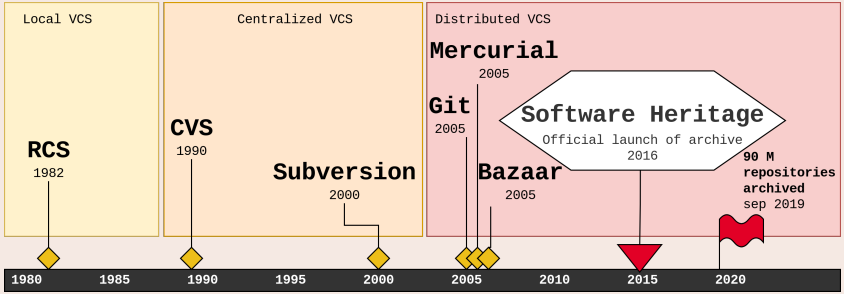
```
# THE MASTER IGNITION ROUTINE IS DESIGNED FOR USE BY THE FOLLOWING LEM PROGRAMS:  P12, P40, P42, P61, P63.
# IT PERFORMS ALL FUNCTIONS IMMEDIATELY ASSOCIATED WITH APS OR DPS IGNITION:  IN PARTICULAR, EVERYTHING LYING
# BETWEEN THE PRE-IGNITION TIME CHECK -- ARE WE WITHIN 45 SECONDS OF TIG? -- AND TIG + 26 SECONDS, WHEN DPS
# PROGRAMS THROTTLE UP.
#
# VARIATIONS AMONG PROGRAMS ARE ACCOMODATED BY MEANS OF TABLES CONTAINING CONSTANTS (FOR AVEGEXIT, FOR
# WAITLIST, FOR PINBALL) AND TCF INSTRUCTIONS.  USERS PLACE THE ADRES OF THE HEAD OF THE APPROPRIATE TABLE
# (OF P61TABLE FOR P61LM, FOR EXAMPLE) IN ERASABLE REGISTER `WHICH' (E4).  THE IGNITION ROUTINE THEN INDEXES BY
# WHICH TO OBTAIN OR EXECUTE THE PROPER TABLE ENTRY.  THE IGNITION ROUTINE IS INITIATED BY A TCF BURNBABY,
# THROUGH BANKJUMP IF NECESSARY.  THERE IS NO RETURN.
#
# THE MASTER IGNITION ROUTINE WAS CONCEIVED AND EXECUTED, AND (NOTA BENE) IS MAINTAINED BY ADLER AND EYLES.
#
#                  HONI SOIT QUI MAL Y PENSE
#
#       ***********************************************
#              TABLES FOR THE IGNITION ROUTINE
#       ***********************************************
#
#                       NOLI SE TANGERE

P12TABLE        VN      0674            # (0)
                TCF     ULLGNOT         # (1)
                TCF     COMFAIL3        # (2)
                TCF     GOCUTOFF        # (3)
                TCF     TASKOVER        # (4)
                TCF     P12SPOT         # (5)
                DEC     0               # (6)   NO ULLAGE
                EBANK=  WHICH
                2CADR   SERVEXIT        # (7)

                TCF     DISPCHNG        # (11)
                TCF     WAITABIT        # (12)
                TCF     P12IGN          # (13)

P40TABLE        VN      0640            # (0)
```

Version Control System timeline

Version control system (VCS)

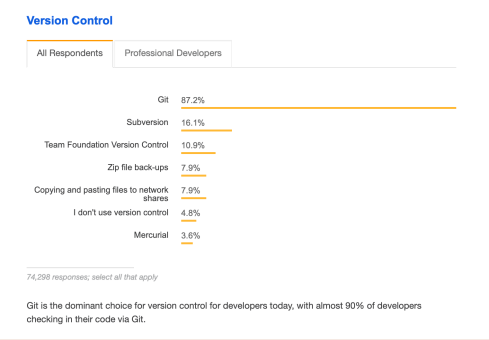
- records changes made to a (set of) *source code file* (s)
- allows to operate on versions: diff/merge/fork/recover etc.
- essential tool for software development

Three decades of evolution



Adoption

Stack Overflow [Survey 2018]



In numbers

GitHub [Octoverse 2017] [Blog 2018]

- 100.000.000+ repositories
- 40.000.000+ developers worldwide

Bitbucket [Blog 2019]

- 28.000.000+ repositories
- 10.000.000+ developers worldwide

GitLab [Blog 2019]

- 1.000.000 MRs March 19'

Let's use it!

Outline

Software Heritage in a nutshell

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share *all* software source code

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

Reference catalog

find and reference all software source code

Universal archive

preserve all software source code

Research infrastructure

enable analysis of all software source code

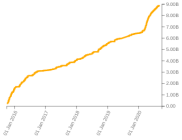
Cultural Heritage | **Industry** | **Research** | **Education**

Software Heritage

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

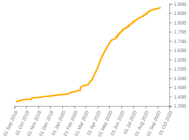
Source files

8,846,381,610



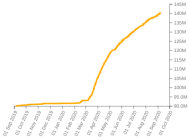
Commits

1,880,663,008



Projects

140,348,311



Directories

7,506,954,410

Authors

38,603,337

Releases

15,051,940

Growing Support

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

Sharing the vision

Sponsoring our work

Platinum sponsors

Gold sponsors

Silver sponsors

We are at a turning point

Looking at the past

- a lot of old software misplaced, lost, or behind barriers, but…
- most founding fathers are still here, and willing to share
- urgent to collect their knowledge

Only a few years left.

Looking at the future

- software development and use skyrockets: more programmers, and more code!
- essential to provide a universal platform for all the future software source code

Every year that goes by makes the problem worse.

it is urgent to take action!

Outline

The Paris call: Software Source Code is our Heritage

November, 2018 at the UNESCO headquarters experts signed *the engagement*

- Recognise software source code as a precious asset of humankind
- Support the development of shared infrastructures
- Foster international collaboration to build a common framework

see full text

The Paris call: Software Source Code is our Heritage

Quotes

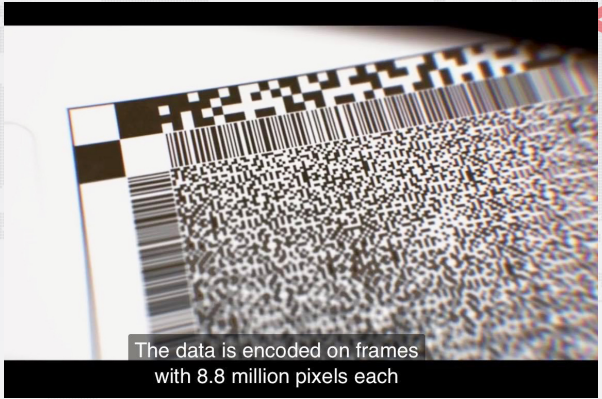
- "Preserving Software Source Code is crucial and captures human civilization"
- "It includes the need to raising awareness of the importance of SSC among decision makers, recognizing Software creators and the contributions of women and minorities to digital innovations".
- "Considering that documents produced and preserved overtime, in all their analog and digital forms through time and space, constitute the primary means of knowledge creation and expression, having an impact on all areas of humanity's civilization and its further progress".
- "Considering at the same time that the preservation of, and long term accessibility to documentary heritage underpins fundamental freedoms of opinion, expression and information as human rights".

Outline

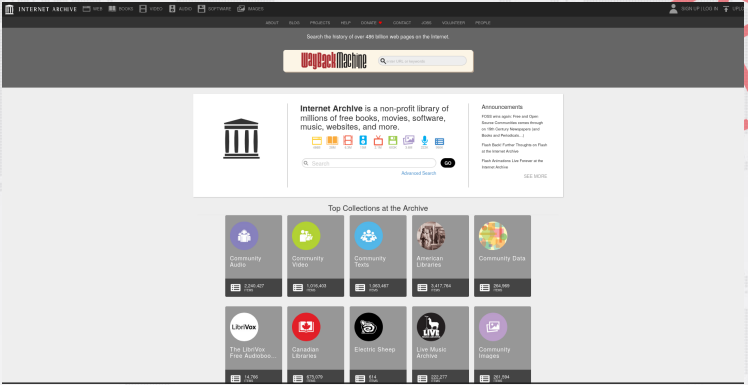
Joining forces in the urgent effort to preserve humankind's source code.

- A testament to the importance of software source code preservation
- A multi-partners strategy for archiving code
- A range of storage solutions, from real-time to long-term storage

Github Arctic Vault

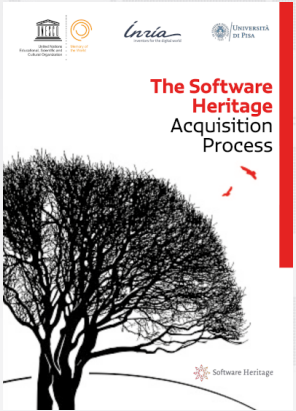


Internet archive



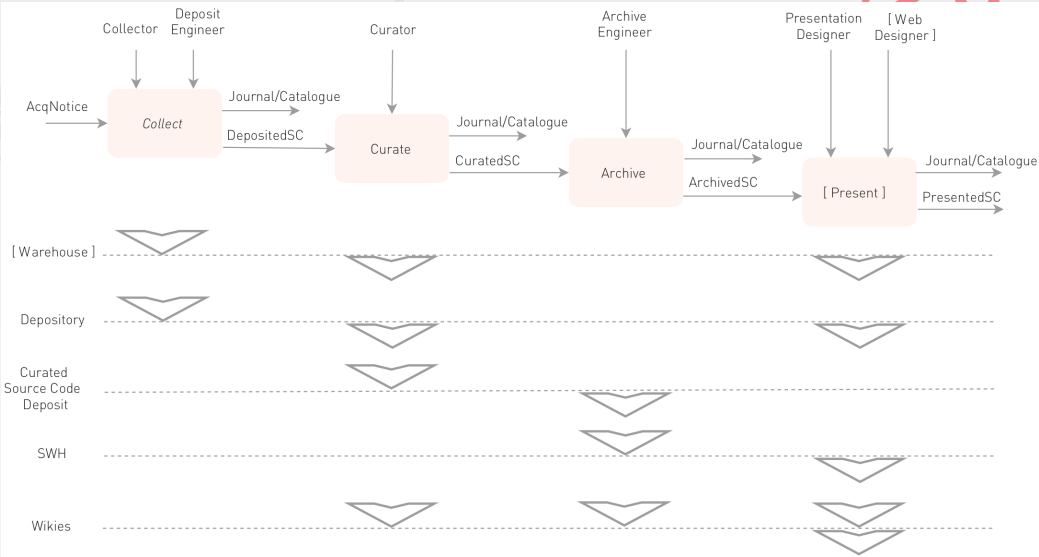
Paris Call on Software Source Code

"[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"

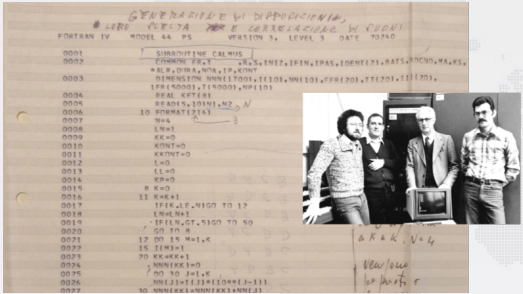


- Rescue Legacy Software from different media
  - physical
  - digital
    - legacy / unsupported
    - recent / supported
- Curate the code
  - reconstructing the development history
  - collecting metadata
- And illustrate with dedicated presentations

SWHAP: Four phases workflow to streamline the effort



SWH@PISA finds : TAUmus ('70s)



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

0001          SUBROUTINE CALMUS
0002          COMMON FRCT    ,KS,INIZ,IFIN,IPAS,IDEN(2),RATS,DCOND,MA,KS,
             *KLB,DIRA,NMA,IP,KCNT
0003          DIMENSION NNN(17001),II(10),NN(10),FFR(20),TT(20),II(20),
             IFR(40000),T(40000,NP(10)
0004           REAL KETIB1
0005           READ(5,10)N1,N2
0006          10 FORMAT(2I4)
0007           N=4
0008           LN=1
0009           KK=0
0010           KCNT=0
0011           KKCNT=0
0012           L=0
0013           LL=0
0014           KP=0
0015         8 K=0
0016          11 K=K+1
0017           IF(K.LE.N)GO TO 17
0018           LN=LN+1
0019           IF(LN.GT.5)GO TO 50
0020           GO TO 8
0021          12 DO 15 M=1,K
0022          15 II(M)=1
0023          20 KK=KK+1
0024           NNN(KK)=0
0025           DO 30 I=1,K
0026           NN(I)=II(I)*(10**(I-1))
0027          30 NNN(KK)=NNN(KK)+NN(I)
```

- The control code for the music synthesizer TAU2
- FORTRAN II and TAUmus command language
- Istituto di Elaborazione dell'Informazione CNR
- Group led by the late M° P. Grossi
  - Le Sacre du Printemps (ABSTRACT)

SWHAP@PISA: Capturing metadata in branch master

Search or jump to... / Pull requests Issues Marketplace Explore

Unipisa / TAUmus

Sponsor | Unwatch 3 | Star 0 | Fork 0

Code | Issues 0 | Pull requests 0 | Actions | Projects 0 | Wiki | Security | Insights

Branch: master — TAUmus / codemeta.json — Find file | Copy path

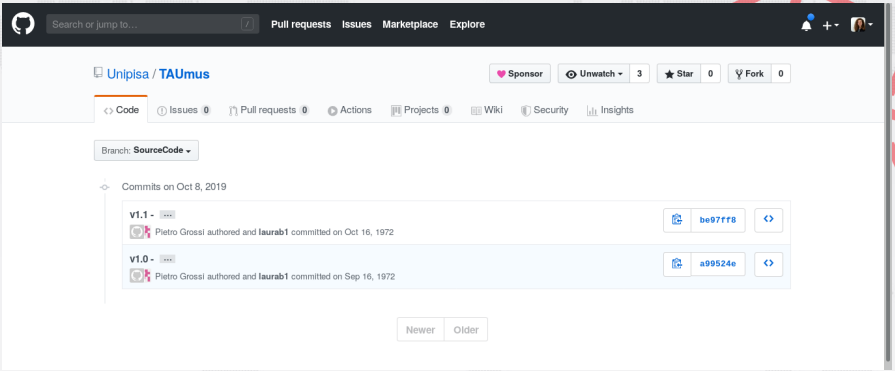
laurab1 cleaned directory — 6dce80b on Oct 8

1 contributor

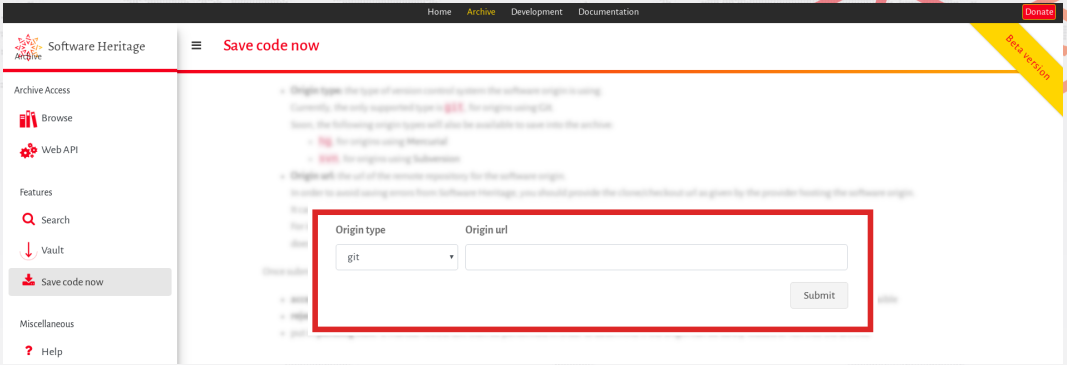
43 lines (43 sloc) 1.33 KB — Raw | Blame | History

```
1  {
2      "@context": "https://doi.org/10.5063/schema/codemeta-2.0",
3      "@type": "SoftwareSourceCode",
4      "identifier": "TAUmus",
5      "description": "Code controlling TAU2, a music synthesizer of the '70s.",
6      "name": "TAUmus",
7      "codeRepository": "https://github.com/Unipisa/TAUmus",
8      "applicationCategory" : "Computer Music",
9      "license": "Copyright (C) 1972 Pietro Grossi",
10     "version": "1.1",
```

SWHAP@PISA: Recreating development history in branch SourceCode



SWHAP@PISA: Archiving source code with Save Code Now



SWHAP@PISA: Referencing source code on SWH

Software Heritage Archive

Browse archived content for origin https://github.com/Unipisa/TAUmus

Beta version

Visits | Snapshot date: 08 October 2019, 17:49 UTC | Branches (3) | Releases (0)

Branch: refs/heads/SourceCode | 8c34070 / SUBROUTINE_CALMUS.FOR | Raw File | Select Language | Actions

SUBROUTINE_CALMUS.FOR

```
1      SUBROUTINE CALMUS
2      COMMON FR,T  ,R,S,INIZ,IFI
3      *ALR,DURA,NOA,IP,KONT
4      DIMENSION NNN(1700),I(10),NN(
5  1   FR (5000) ,T(5000) NPLLOD
6      REAL KFT(8)
7      READ(5,10)N,N1,N2
8  10  FORMAT(3I4)
9      N=4
10     LN=1
11     KK=0
12     KONT=0
13     KKONT=0
14     L=0
15     LL=0
16     KP=0
17 8   K=0
18 11  K=K+1
19     IF(K.LE.N)GO TO 12
20     LN=LN+1
21     IF(LN.GT.5)GO TO 50
22     GO TO 8
```

Permalinks

To reference or cite the objects present in the Software Heritage archive, permalinks based on **persistent identifiers** must be used instead of copying and pasting the url from the address bar of the browser (as there is no guarantee the current URI scheme will remain the same over time).

Select below a type of object currently browsed in order to display its associated persistent identifier and permalink.

content | directory | revision | snapshot

archived repository | archived swh:1:cnt:137968d1b65eabd6390647f95119f2eb24704962

swh:1:cnt:137968d1b65eabd6390647f95119f2eb24704962;origin=https://github.com/Unipisa/TAUmus

Add origin info | Add selected lines info | Copy identifier | Copy permalink

Outline

Software Heritage ethical charter

Ethical Charter for using the archive data — https://bit.ly/2KOyfeR

- Avoid harm
- Protect Personal Data
- Avoid useless copies
- Care about derived data

Ethical Charter for Mirrors — https://bit.ly/36q2nWx

- Avoid harm
- Protect Personal Data
- Maintain coherent terms of use
- Ensure fair and non discriminatory access
- Foster Collaboration

Software preservation and ethics

Discussion topics

1. Transparency and conflict of interest
2. Digital divide
3. Ownership of the physical archive
4. Responsibility for the incorrect use of source codes
5. The positive potential of this archive
6. Possible role of philosophers in interdisciplinary teams with scientists

Software Heritage

Thank you! Any questions?

contact: morane@softwareheritage.org

Jean-François Abramatic, Roberto Di Cosmo, Stefano Zacchiroli
Building the Universal Archive of Source Code, Communications of the ACM, October 2018

Roberto Di Cosmo, Morane Gruenpeter, Stefano Zacchiroli
Identifiers for Digital Objects: the Case of Software Source Code Preservation, iPRES 2018: Intl. Conf. on Digital Preservation

Acknowledgements

- Roberto Di Cosmo, Founder and Director of Software Heritage
- Leah Gruenpeter-Gold, PhD Philosophy Dept., Tel Aviv University