

The Software Heritage Acquisition Process (SWHAP)

motivations and overview

Roberto Di Cosmo
SWHAP Days

Director, Software Heritage
Inria and Université de Paris Cité

September 30th 2022



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE



1 Tackling the challenge

2 The SWHAP approach

3 Conclusion

Limitations of popular approaches, revisited

A - Since the 1970's 1990's

.zip or .tar file on:

- ftp-server
- web page
- document archive (+ DOI)

B - Since the 2000's

Rely on *software forges*

- institutional or project ones
- free commercial ones: BitBucket, GitHub, GitLab, ...

Assessing pros and cons of forges


Pros




- better *user experience*
- *version control* is built-in

Cons




- no *preservation* guarantee
- can be easily *misused*

An example is worth a thousand words



 [agend](#) / [warcraft-2000-nuclear-epidemic](#) Public














 Watch 4  Fork 3  Star 16

[Code](#) [Issues](#) [Pull requests](#) [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#)

 master  1 branch  0 tags


[Go to file](#) [Add file](#) [Code](#)


 **agend** copy as-is from original CD 018cf4b on Aug 16, 2015  1 commit


 3DSurf.cpp	copy as-is from original CD	7 years ago
 3DSurf.h	copy as-is from original CD	7 years ago
 AntiBug.cpp	copy as-is from original CD	7 years ago
 AntiBug.h	copy as-is from original CD	7 years ago
 Build.cpp	copy as-is from original CD	7 years ago
 COMPO.TXT	copy as-is from original CD	7 years ago
 CWAVE.H	copy as-is from original CD	7 years ago
 Cdirsnd.cpp	copy as-is from original CD	7 years ago
 Cdirsnd.h	copy as-is from original CD	7 years ago
 Crowd.cpp	copy as-is from original CD	7 years ago
 Cwave.cpp	copy as-is from original CD	7 years ago
 DPLAY.H	copy as-is from original CD	7 years ago
 DPLOBBY.H	copy as-is from original CD	7 years ago

About

This is source code found on disk of game called Warcraft 2000 Nuclear Edition. This is attempt to create game in 1998 based on fusion of Warcraft and Starcraft. As stated in readme file it's uncompleted and developers give a way source code for free.

 16 stars

 4 watching

 3 forks

Releases

No releases published

Packages

No packages published

 [agend](#) / [warcraft-2000-nuclear-epidemic](#) Public

 Watch 4  Fork 3  Star 16

[Code](#) [Issues](#) [Pull requests](#) [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#)

Back to the drawing board

Collect

- *find* source code and related materials
- *gather* in a physical and/or logical place for later processing

Curate

- *analyze, cleanup and structure* the materials
- identify *authors* of *versions* of source code, with its *dates*
- identify *owners*, obtain *authorizations*
- add quality *metadata*, in a *standard format*

Archive

save the curated materials to appropriate *archives*

Present

make the materials accessible to a *wide audience*

a key requirement: *traceability* all along the way



1 Tackling the challenge

2 The SWHAP approach

3 Conclusion

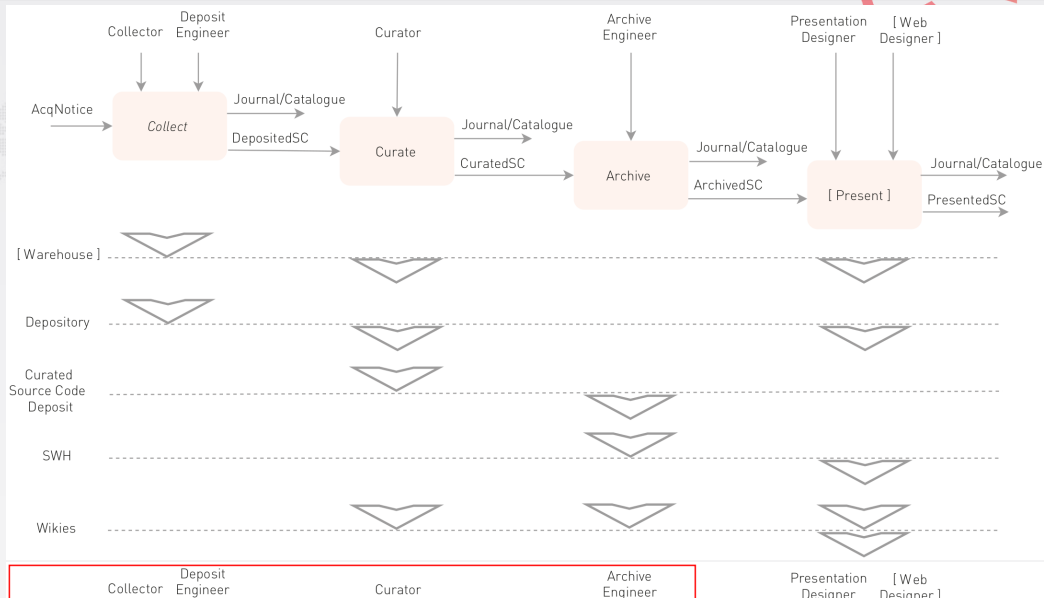
Paris Call on Software Source Code

“[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”



- **Detailed process** for Legacy Software
 - *curation*
 - reconstruction of the development history
 - collecting metadata
 - *archival*
 - in Software Heritage
- **Traceability** of all process phases
 - using modern version control tools

SWHAP in a nutshell: four phases workflow



An example SWHAP outcome: TAUMus - curator and source code branch

The screenshot displays the Software Heritage Archive interface for the repository `https://github.com/Unipisa/TAUMus`. The interface includes a sidebar with navigation icons, a top search bar, and a main content area. The repository is dated 25 June 2021, 10:13:09 UTC. The main content area shows the repository's structure, including a table of files and a section for the `README.md` file.

Repository Information:

- URL: `https://github.com/Unipisa/TAUMus`
- Date: 25 June 2021, 10:13:09 UTC
- Branch: `HEAD` (4506b19 /)
- Commits: 7c5eabad1a5cc1b87918b399d433b2aze8 authored by CarloQMontangero on 16 March 2021, 10:05:34 UTC

Files Table:

	Mode	Size
README.md	-rW-r--r--	2.7 KB
codemeta.json	-rW-r--r--	1.4 KB







README.md Content:

TAUMus

TAUMus is the software controlling the real-time computer-music system TAU2-TAUMUS, developed in the 70's of the XX century at IEE and CNUCE in Pisa under the leadership of Maestro P. Grossi.

This repository has a branch containing a small excerpt of the development history of the source code: some samples of session scripts that

An example SWHAP outcome: TAUMus - curation history



Browse the archive

Enter a SWHID to resolve or keyword(s) to search

<https://github.com/Unipisa/TAUMus>

25 June 2021, 10:13:09 UTC


<> Code Branches (2) Releases (2) Visits






★ Branch: HEAD Save again


sort by: ☒ revision date ☐ DFS ☐ DFS post-ordering ☐ BFS

Revision	Author	Date	Message	Commit Date
3e4e117	CarloQMontangero	16 March 2021, 10:05:34 UTC	typos	16 March 2021, 10:05:34 UTC
62cd163	Guido	03 March 2021, 10:48:57 UTC	Update README.md	03 March 2021, 10:48:57 UTC
4f1f1a5	Guido	02 February 2021, 10:23:09 UTC	Update codemeta.json minor fix 2	02 February 2021, 10:23:09 UTC
988ded8	Guido	02 February 2021, 10:22:49 UTC	Update codemeta.json minor fix	02 February 2021, 10:22:49 UTC
6ce07c4	Guido	02 February 2021, 10:22:00 UTC	Update codemeta.json fixed referencePublication	02 February 2021, 10:22:00 UTC
c9430f4	CarloQMontangero	02 March 2020, 17:07:00 UTC	Typo	02 March 2020, 17:07:00 UTC
e3a4b4f	Guido	11 December 2019, 10:16:01 UTC	Update README.md added badges	11 December 2019, 10:16:01 UTC



An example SWHAP outcome: TAUMus - source code branch







 Browse the archive


Enter a SWHID to resolve or keyword(s) to search


 <https://github.com/Unipisa/TAUMus> 


 25 June 2021, 10:13:09 UTC

< > Code


 Branches (2)


 Releases (2)


 Visits


 Branch: refs/heads/SourceCode

c673de3 /






 History

 Download

 Save again

 Tip revision: be97ff85eb836773e0af90490bea376e52fce579 authored by Pietro Grossi on 16 October 1972, 08:54:00 UTC

v1.1 -

File	Mode	Size
 Taumus_sessions		
 PROGRAM_FOR_AT1.FOR	-rW-r--r--	920 bytes
 SCALA.FOR	-rW-r--r--	727 bytes
 SUBROUTINE_CALMUS.FOR	-rW-r--r--	2.0 KB
 TWO_VOICES_RANDOM_MUSIC.FOR	-rW-r--r--	1.5 KB

An example SWHAP outcome: TAUMus - source code history

Software Heritage Archive

🔍

📄

📷

🔖

?

≡ Browse the archive

Enter a SWHID to resolve or keyword(s) to search

📄 <https://github.com/Unipisa/TAUMus>

🕒 25 June 2021, 10:13:09 UTC

<> Code

🔗 Branches (2)

📦 Releases (2)

📅 Visits

🔗 Branch: `refs/heads/SourceCode`

📷 Save again

sort by: ☒ revision date ☐ DFS ☐ DFS post-ordering ☐ BFS

↻ Revision	Author	Date	Message	Commit Date
↻ be97ff8	Pietro Grossi	16 October 1972, 08:54:00 UTC	v1.1 - Contributors: Leonello Tarabella	08 October 2019, 08:51:13 UTC
↻ a99524e	Pietro Grossi	16 September 1972, 07:54:00 UTC	v1.0 - Contributors: Leonello Tarabella	08 October 2019, 08:51:11 UTC

Newer

Older

An example SWHAP outcome: TAUMus - separate author and curator

Software Heritage Archive

🔍

⬇️

📷

🔖

?

☰

Browse the archive

Enter a SWHID to resolve or keyword(s) to search

🔗 <https://github.com/Unipisa/TAUMus>

🕒 25 June 2021, 10:13:09 UTC

<> Code

🌿 Branches (2)

📦 Releases (2)

📅 Visits

🔗 Revision [be97ff85eb836773e0af90490bea376e52fce579](#) authored by [Pietro Grossi](#) on [16 October 1972, 08:54:00 UTC](#), committed by [TAUMus Curation Team](#) on [08 October 2019, 08:51:13 UTC](#)

V1.1 -

Contributors: Leonello Tarabella

1 parent 🔗 a99524e

Files

Changes

🌿 Branch: [refs/heads/SourceCode](#)

c673de3 /

🕒 History

⬇️ Download

📷 Save again

🔗 Tip revision: [be97ff85eb836773e0af90490bea376e52fce579](#) authored by [Pietro Grossi](#) on [16 October 1972, 08:54:00 UTC](#)

V1.1 -

File	Mode	Size
📁 Taumus_sessions		
📄 PROGRAM_FOR_AT1.FOR	-r--r--r--	920 bytes
📄 SCALA.FOR	-r--r--r--	727 bytes
📄 SUBROUTINE_CALMUS.FOR	-r--r--r--	2.0 KB

R. Di Cosmo roberto@dicosmo.org (CC-BY 4.0)

SWHAP motivations and overview

September 30th 2022

10 / 12

An example SWHAP outcome: TAUMus - view source code evolution

Software
Heritage
Archive



Revision **be97ff85eb836773e0af90490bea376e52fce579** authored by Pietro Grossi on 16 October 1972, 08:54:00 UTC, committed by TAUMus Curation Team on 08 October 2019, 08:51:13 UTC

V1.1 -

Contributors: Leonello Tarabella

1 parent -> a99524e

Files

Changes

Showing 1 changed file with 2 additions and 3 deletions (1 / 1 diffs computed)

Compute all diffs

modified: SUBROUTINE_CALMUS.FOR

SUBROUTINE_CALMUS.FOR

Unified

Side-by-side

View file

@@ -4,8 +4,8 @@

4 DIMENSION NNN(1700), I(10), NN(10), FFRE(20), TT(20), I:

5 1 FR (5000), T(5000) NPLLOD

6 REAL KFT(8)

7 - READ(5,10)N1,N2

8 -10 FORMAT(2I4)

9 N=4

10 LN=1

11 KK=0

@@ -34,7 +34,6 @@

34 1 FORMAT(1X,20I6)

35 LL=0

36 33 DO 35 M=1,K

37 - DO 35 M=1,K

38 I(M)=I(M)+1

39 IF(I(M).LE.N).GO TO 30

@@ -4,8 +4,8 @@

4 DIMENSION NNN(1700), I(10), NN(10), FFRE(20), TT(20), I:

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

@@ -34,7 +34,6 @@

34 1 FORMAT(1X,20I6)

35 LL=0

36 33 DO 35 M=1,K

37 I(M)=I(M)+1

38 IF(I(M).LE.N).GO TO 30



1 Tackling the challenge

2 The SWHAP approach

3 Conclusion

Summary of the SWHAP process

Leverage modern forges and version control tools

- clear separation of software *authors* from *curators*
- *traceability* of all the *curation process*
- reconstruction of the *evolution of software*

Leverage Software Heritage

- archives the full version control system
- keeps previous snapshots of a version control system

Combined result: enables an *iterative process* supporting

- addition of new raw material (e.g. intermediate versions)
- fixing mistakes in the curation process

let's see this in action!