

# 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

## 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

The screenshot shows a GitHub repository page for 'agend / warcraft-2000-nuclear-epidemic'. The repository is public and has 4 watchers, 3 forks, and 16 stars. The main content is a commit titled 'copy as-is from original CD' by user 'agend' on August 16, 2015. The commit includes a list of files, all of which are copies of files from the original CD. The files listed are: 3DSurf.cpp, 3DSurf.h, AntiBug.cpp, AntiBug.h, Build.cpp, COMPO.TXT, CWAVE.H, Cdirsnd.cpp, Cdirsnd.h, Crowd.cpp, Cwave.cpp, DPLAY.H, and DPLOBBY.H. The right sidebar contains an 'About' section with a description of the project, 'Releases' (no releases published), and '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

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

# 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 e1c:f4b 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

# 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”



## 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”



## 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

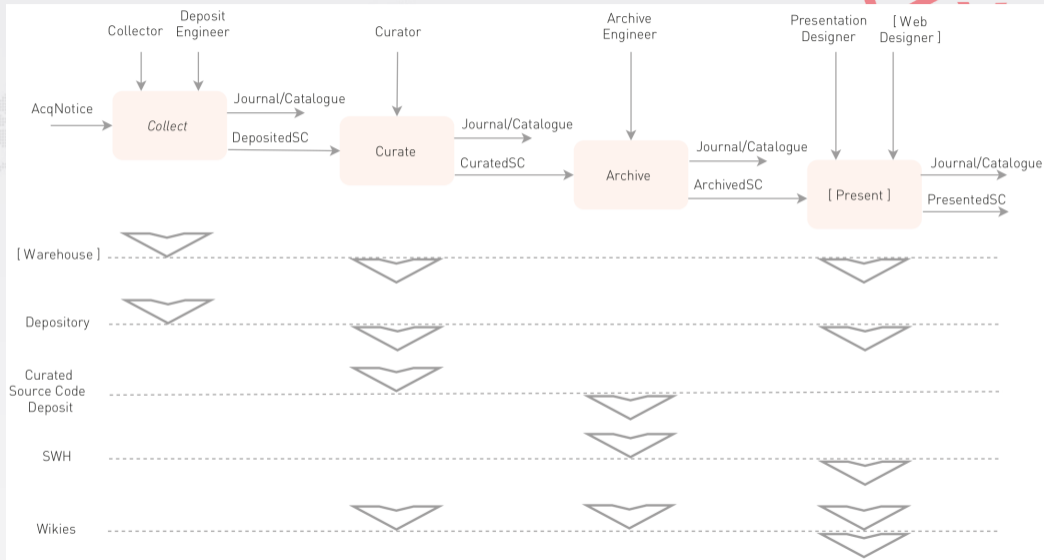
## Paris Call on Software Source Code

“[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”

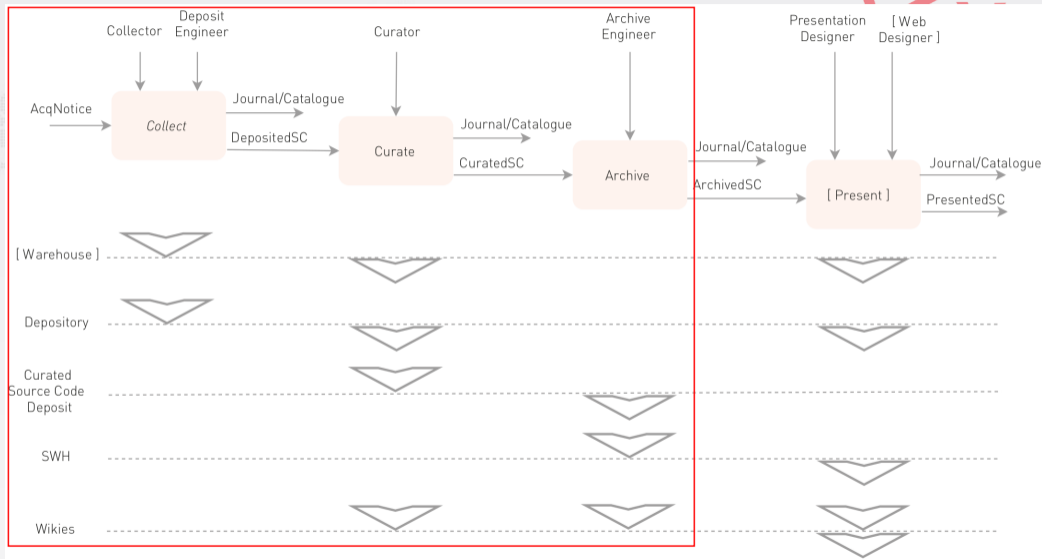


- **Detailed process** for Legacy Software
  - *curation*
    - reconstruction of the development history
    - collecting metadata
  - *archival*
    - in Software Heritage
- **Traceability** of all process phases
  - using modern version control tools

# SWHAP in a nutshell: four phases workflow



# SWHAP in a nutshell: four phases workflow



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

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: HEAD 4506b19 / History Download Save again

Branches Releases 7c5eabad1a5cc1b87918b399d433b2aze8 authored by CarloQMontangero on 16 March 2021, 10:05:34 UTC

-> ★ ✓ HEAD

-> refs/heads/SourceCode

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

## README.md

## TAUMus

TAUMus is the software controlling the real-time computer-music system 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

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: HEAD ▾

Save again

sort by:  revision date  DFS  DFS post-ordering  BFS

Revision	Author	Date	Message	Commit Date
<a href="#">3e4e117</a>	CarloQMontangero	16 March 2021, 10:05:34 UTC	typos	16 March 2021, 10:05:34 UTC
<a href="#">62cd163</a>	Guido	03 March 2021, 10:48:57 UTC	Update README.md	03 March 2021, 10:48:57 UTC
<a href="#">4f1fa5</a>	Guido	02 February 2021, 10:23:09 UTC	Update codemeta.json minor fix 2	02 February 2021, 10:23:09 UTC
<a href="#">988ded8</a>	Guido	02 February 2021, 10:22:49 UTC	Update codemeta.json minor fix	02 February 2021, 10:22:49 UTC
<a href="#">6ce07c4</a>	Guido	02 February 2021, 10:22:00 UTC	Update codemeta.json fixed referencePublication	02 February 2021, 10:22:00 UTC
<a href="#">c9430f4</a>	CarloQMontangero	02 March 2020, 17:07:00 UTC	Typo	02 March 2020, 17:07:00 UTC
<a href="#">e3a4b4f</a>	Guido	11 December 2019, 10:06:00 UTC	Update README.md added badges	11 December 2019, 10:06:00 UTC

# An example SWHAP outcome: TAUMus - source code branch

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

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
<a href="#">be97ff8</a>	Pietro Grossi	16 October 1972, 08:54:00 UTC	v1.1 - Contributors: Leonello Tarabella	08 October 2019, 08:51:13 UTC
<a href="#">a99524e</a>	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	-rw-r--r--	920 bytes
SCALA.FOR	-rw-r--r--	727 bytes
SUBROUTINE_CALMUS.FOR	-rw-r--r--	2.0 KB

# 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

<code>@@ -4,8 +4,8 @@</code>	<code>@@ -4,8 +4,8 @@</code>
4 DIMENSION NNN(1700), I(10), NN(10), FFRE(20), TT(20), I:	4 DIMENSION NNN(1700), I(10), NN(10), FFRE(20), TT(20), I:
5 1 FR (5000), T(5000) NPLLOD	5 1 FR (5000), T(5000) NPLLOD
6 REAL KFT(8)	6 REAL KFT(8)
7 - READ(5,10)N1, N2	7 + READ(5,10)N, N1, N2
8 -10 FORMAT(2I4)	8 +10 FORMAT(3I4)
9 N=4	9 N=4
10 LN=1	10 LN=1
11 KK=0	11 KK=0
<code>@@ -34,7 +34,6 @@</code>	<code>@@ -34,7 +34,6 @@</code>
34 1 FORMAT(1X, 20I6)	34 1 FORMAT(1X, 20I6)
35 LL=0	35 LL=0
36 33 DO 35 M=1, K	36 33 DO 35 M=1, K
37 - DO 35 M=1, K	
38 I(M)=I(M)+1	37 I(M)=I(M)+1
39 IF(I(M).LE.N).GO TO 20	38 IF(I(M).LE.N).GO TO 20



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*

# 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

# 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

# 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!