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

Outline

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

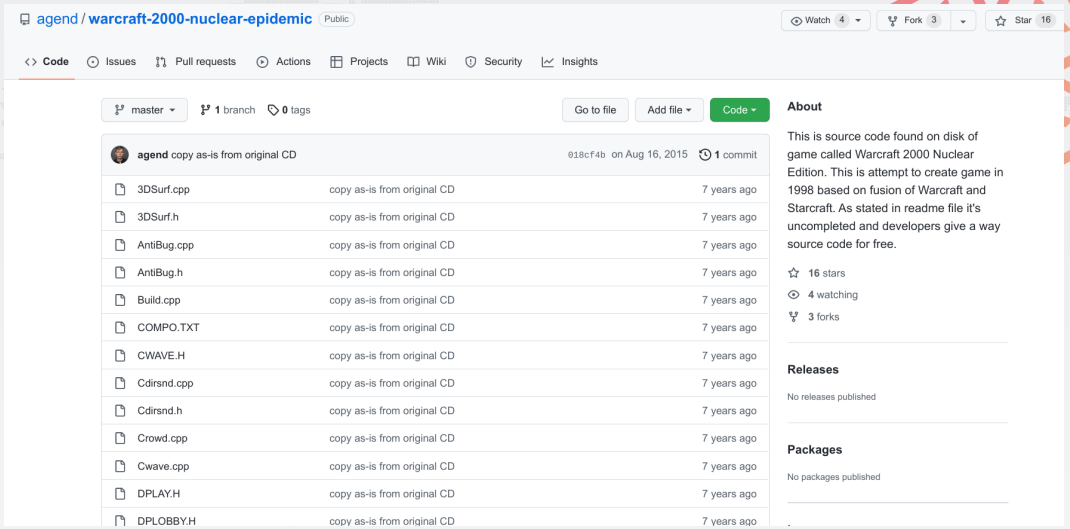
Pros

- 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

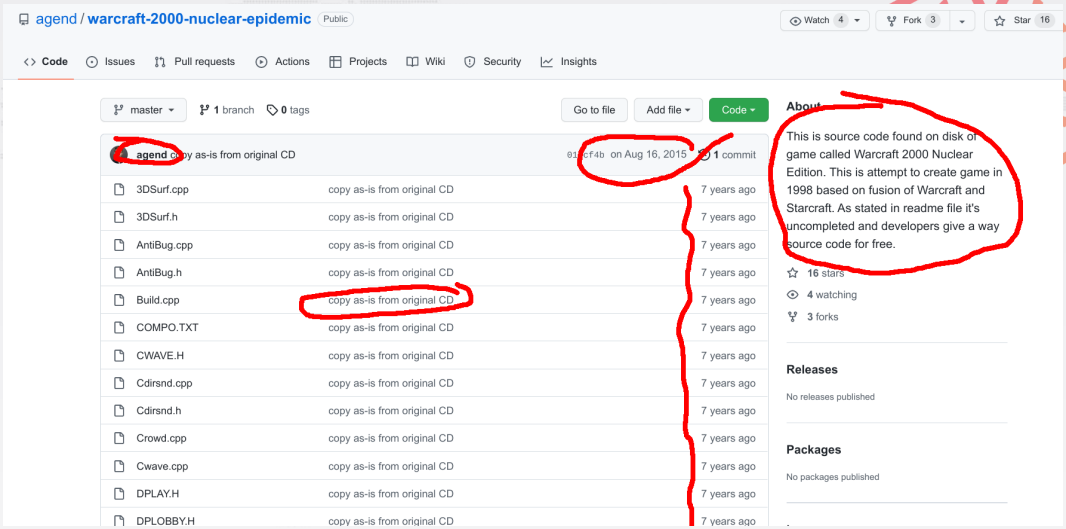
Releases

No releases published

Packages

No packages published

An example is worth a thousand words



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

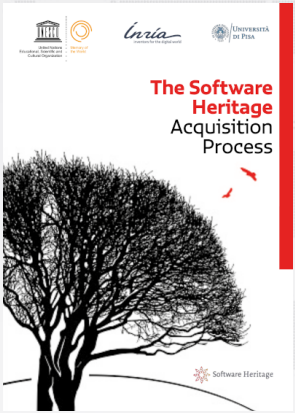
Outline

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"

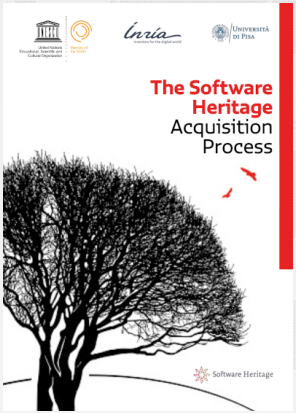
Paris Call on Software Source Code

"[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

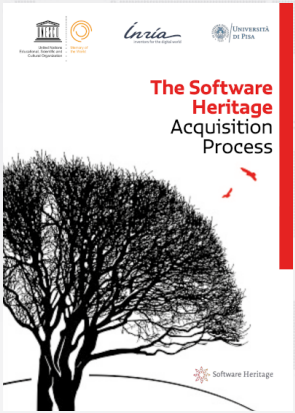
"[We call to] support efforts to gather and preserve the artifacts and narratives of the history of computing, while the earlier creators are still alive"



- Detailed process for Legacy Software
  - *curation*
    - reconstruction of the development history
    - collecting metadata
  - *archival*
    - in Software Heritage

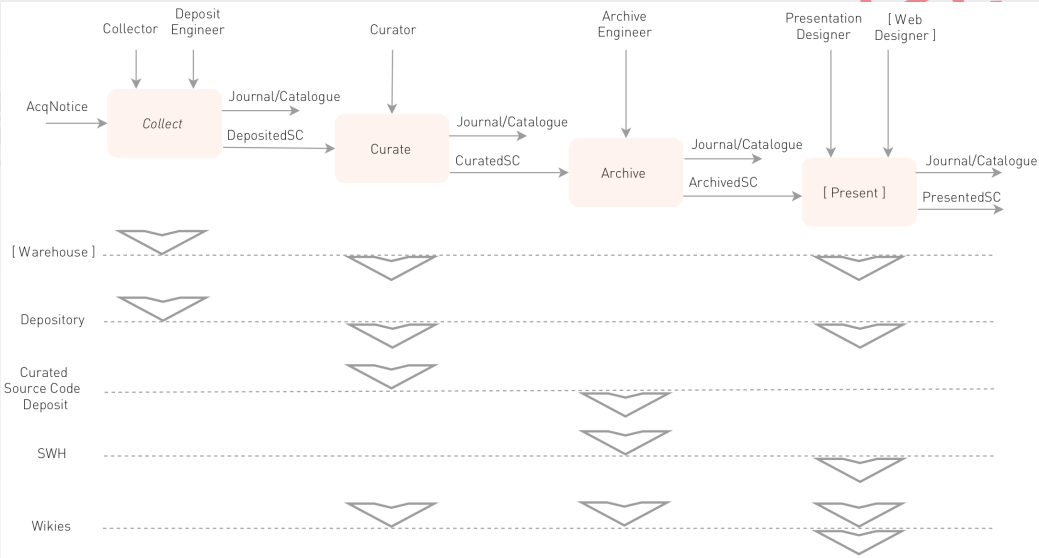
Paris Call on Software Source Code

"[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"

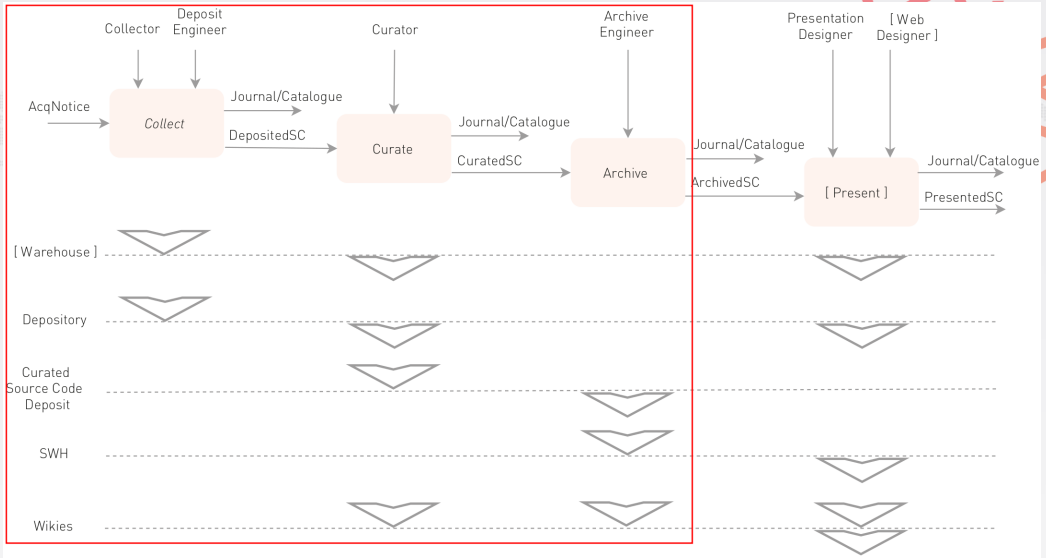


- Detailed process for Legacy Software
  - *curation*
    - reconstruction of the development history
    - collecting metadata
  - *archival*
    - in Software Heritage
- Traceability of all process phases
  - using modern version control tools

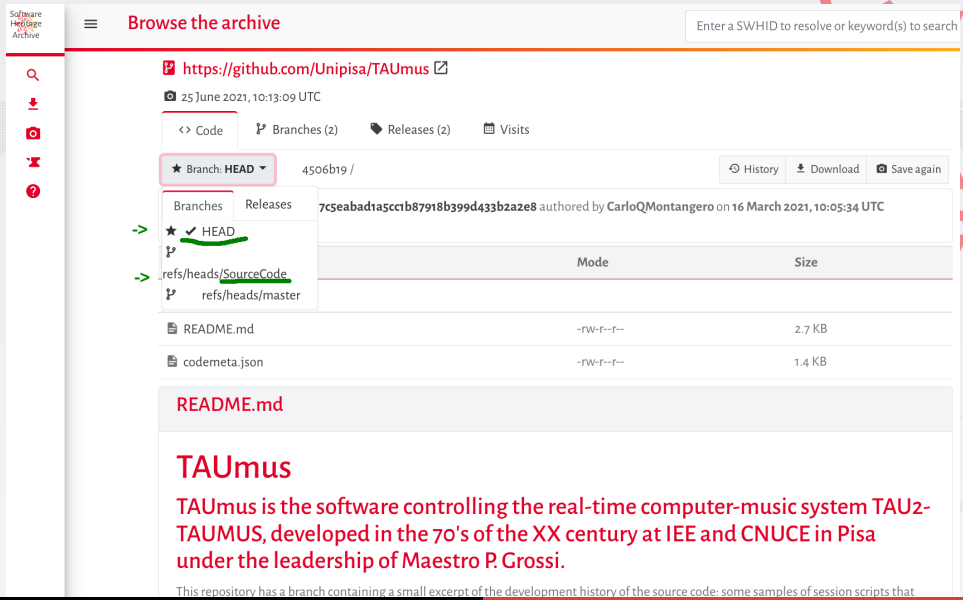
SWHAP in a nutshell: four phases workflow



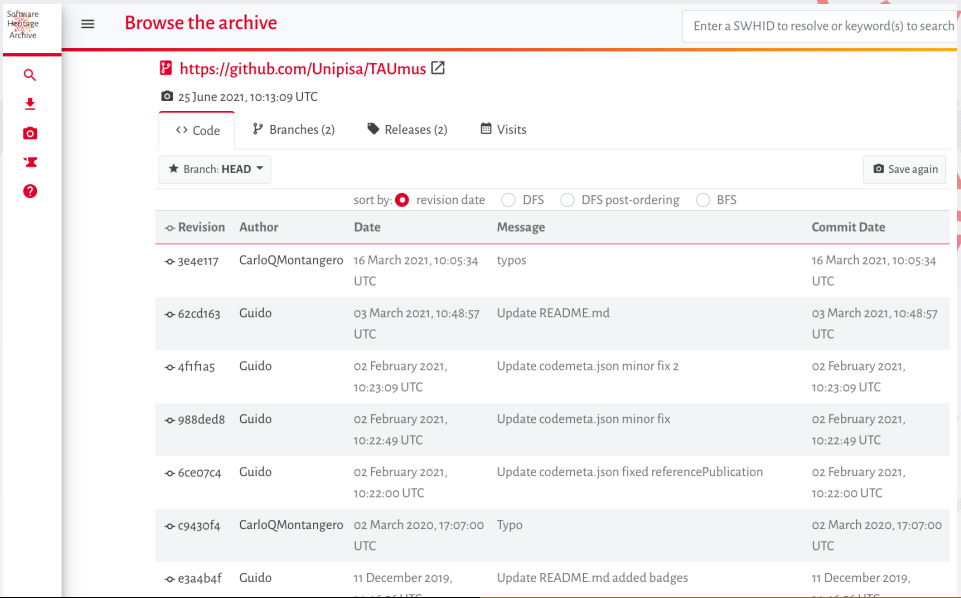
SWHAP in a nutshell: four phases workflow



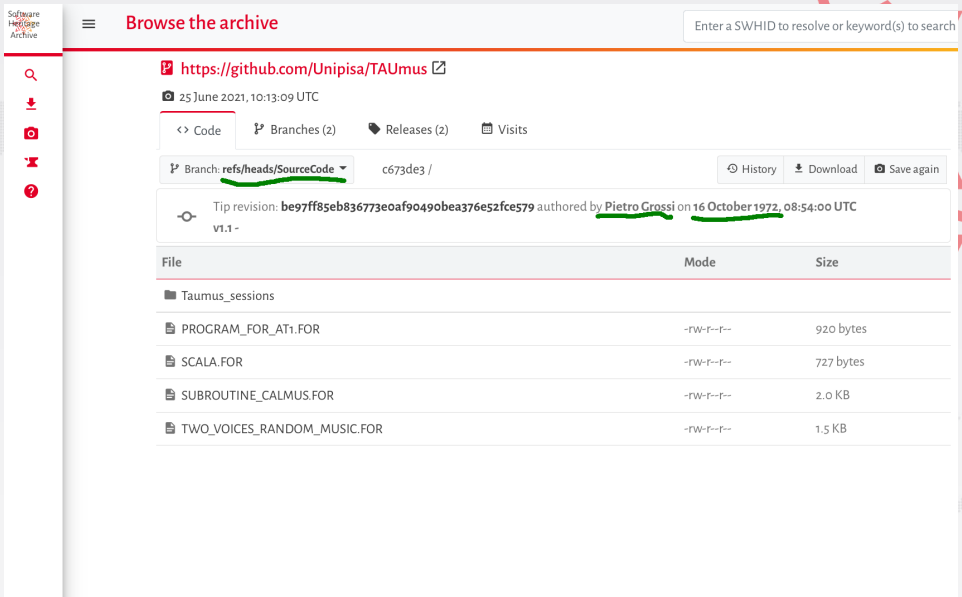
An example SWHAP outcome: TAUMus - curator and source code branch



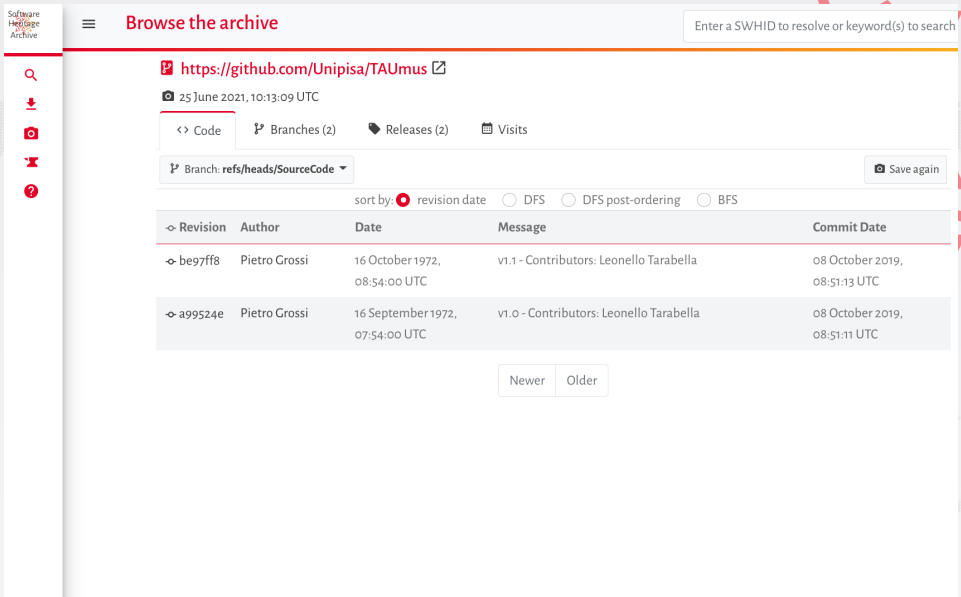
An example SWHAP outcome: TAUMus - curation history



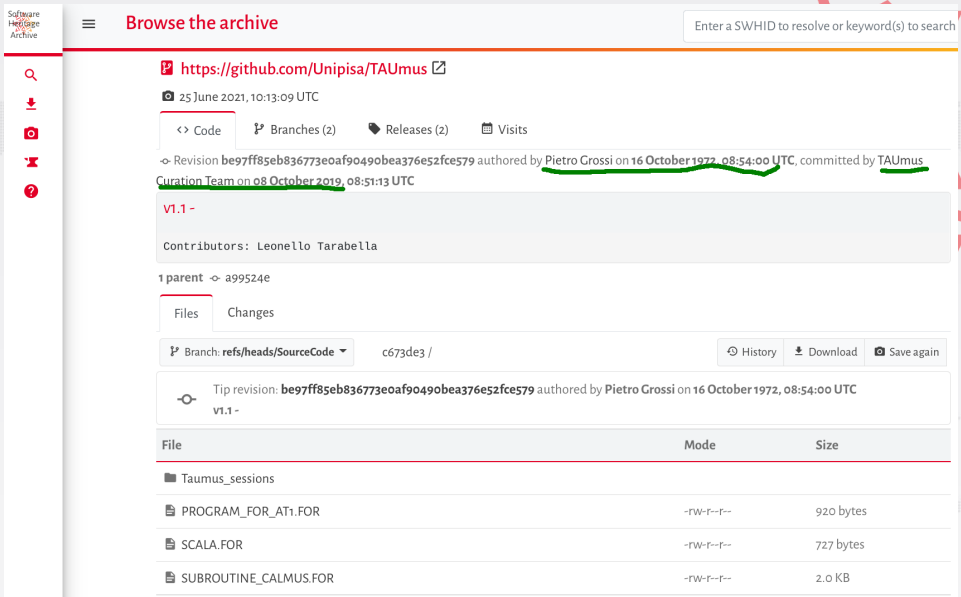
An example SWHAP outcome: TAUMus - source code branch



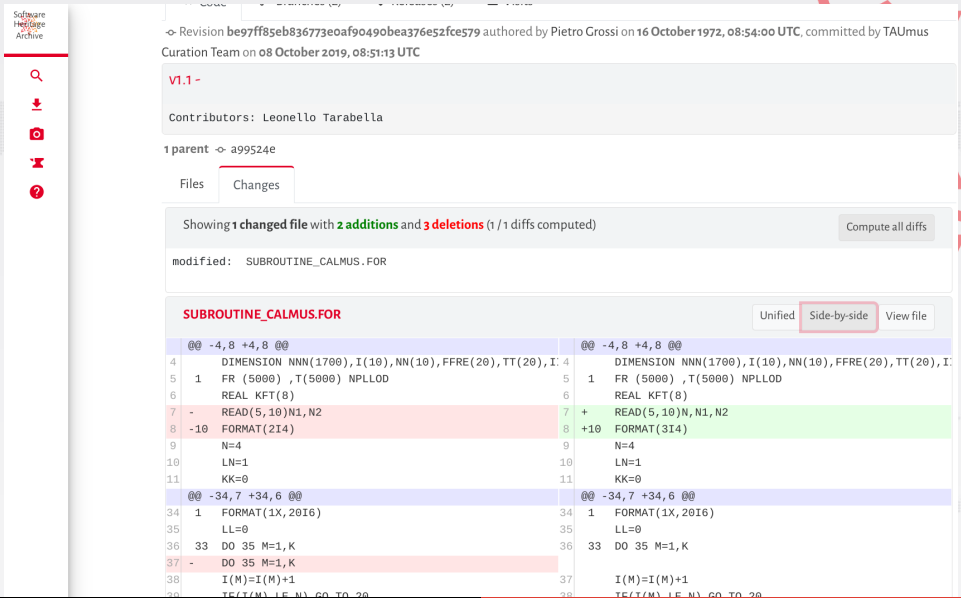
An example SWHAP outcome: TAUMus - source code history



An example SWHAP outcome: TAUMus - separate author and curator



An example SWHAP outcome: TAUMus - view source code evolution



Outline

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!