The Process

The Software Heritage Acquisition Process (SWHAP) is build of the following steps:

- collect
- curate
- present

To realize the process we use the following components:

- the depository
- the workbench
- the vault

The final product of the process is the **curated source code (CuratedSC)** that will be stored into the vault, that is the Software Heritage Archive (HAL). Both the depository and the workbench are introduced to being able of tracing the origin and the evolution of the artifacts leading to the CuratedSC. In the end we must be able to answer the questions "What we have ?", "Where and when we found it ?", "In what way has been archived and transformed" ?
After being either pulled or pushed, the software is Collected from an origin, and stored in its digital copy into the depository. The origin, in the case of physical repositories, can be a physical place, we speak of warehouse. From the warehouse a digital "as is" copy is made into the depository.

Both the depository and the workbench have :

- a Catalogue, that is, similar to a library catalogue, a complete list of items, typically one in alphabetical or other systematic order;
- a Journal, that is, a registry where all the done operations are written. For the depository and the warehouse the journal contains records for acquisitions, their date, notes about the origin of the artifacts, information about where they are archived. For the workbench the journal is more detailed: is a sort of lab notebook where every activity is tracked.

The Instantiation of the process: the di.unipi case

To acquire legacy softwares of Department of Computer Science at the University of Pisa, we instantiated the model of SWHAP using GitHub as support.

In particular, we choose GitHub as material implementation of the depository and of the workbench as :

- it is a well established platform for storing open source project and for collaborate with others. It offers an extensive and reachable disk space at a convenient price - it is free for open source projects;
- at the moment, the Software Heritage has already a crawler that feed the vault from GitHub
- it offers facilities to compile the journal (in the form of commit history, where each changes is tracked) and the catalogue.

The main GitHub repository involved are

1. **DIUNIPI4SWH** This is the front-page of the project. It presents the project, the software acquisition process, collect the documentation, contains the catalogue of acquired softwares and the journal of acquisitions. It also links the template repository that has to be used to start each acquisition process. It is curated by one or more maintainers in charge of accepting records for the catalogue.
2. **DIUNIPI SWH TEMPLATE** This repository defines the skeleton of directories and files that has to be used for each software acquisition.

The process here is as follows:

1. For every new software acquisition, the DI.UNIPI SWH TEMPLATE, wich contains th deppository skeleton template, is forked into a LAB repostitory. The forked repository is named with the pattern LAB, where is in the form of .
2. The acquisition process begin filling the DEPOSITORY TMP directory with all the digital version of original materials and the traces are written into the specific Depository Journal. Once the

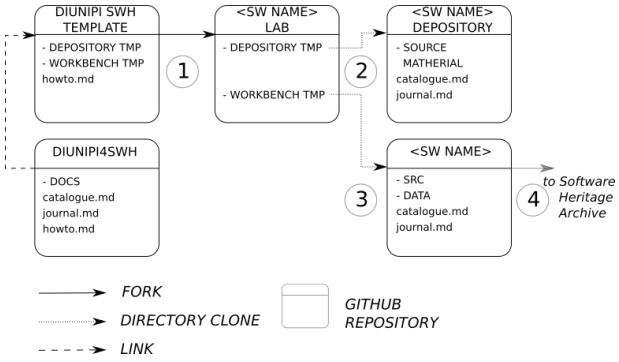

Figure 1: SH_UNIPI_PROCESS

acquisition process is terminated the specific Depository catalog is compiled the DEPOSITORY TMP directory is cloned into a DEPOSITORY repository. This repository will be set as read only (that is, writable only by the owner).

3. The curation process begin filling the WORKBENCH TMP directory from material of the DEPOSITORY TMP directory. In particular, the SRC folder will contains a synthetic git build following what done by Spinellis (Spinellis 2017). The DATA directory will contains all additional information. As for depository, once the curation process is completed, the WORKBENCH TMP directory is cloned into a repository and will be set as read only. This is the curated software, result of acquisition process.
4. Once the curated software repository is done, the curator propose a record into the catalog of DIUNIPI4SWH repository. The owner of DIUNIPI4SWH repository will accept the record and will submit the curated software to the Software Heritage Archive (HAL).

Use cases

We focused our attention to the following legacy software:

- **CMM** - A garbage collector written by Giuseppe Attardi and Tito Flagella
- Grossi Tarabella musica elettronica
- Martelli Montanari
- Macchina ridotta
- Compilatore fortran cep
- Index Dantesco di Padre Busa
- Programma primo Ping di Lenzini
- Parser di linguistica computazionale

Web Ref

Latest generated pdf

Directory clone

DIUNIPI4SWH

google doc precedente

Code

Generare pdf con bibliografia ~ pandoc –filter pandoc-citeproc –bibliography=WorkingNotepad.bib –variable classoption=twocolumn –variable papersize=a4paper -s WorkingNotepad.md -o WorkingNotepad.pdf ~

Spinellis, Diomidis. 2017. “A Repository of Unix History and Evolution.” *Empirical Software Engineering* 22 (3). Springer US: 1372–1404. doi:10.1007/s10664-016-9445-5.