Software Heritage Archive

Pelayo Reguera García

Diego Villanueva Berros

Rubén del Rey Álvarez

EN-10

1. Introduction

1.1.- Roberto Di Cosmo

Roberto Di Cosmo is an Italian computer scientist who graduated from the Scuola Normale Superiore di Pisa, became a PhD graduate in the University of Pisa and then became a professor at the École normale supériure in Paris.

In 2016 he started the Software Heritage initiative.

1.2.- What is it?

As Roberto said in the podcast, the Software Heritage is like the Library of Alexandria of source code, a place where anyone can find all public available software developed by anyone.

The aim of the Software Heritage Archive is that all the code that it stores all public code no matter where it is stored or where it is stored. This does not only include platforms like Github, GitLab, Bitbucket or GitPocket, but also package managers, or distributions that share software. The idea is to provide an infrastructure that has every piece of software developed in history.

1.3.- What does it store?

The Software Heritage Archive stores open-source code, but not only open-source code. It collects any publicly available software, even if it is not open source. This means that there is some software that is not possible to use by anyone, so its license should be checked to see if it can be used or not.

The approach taken by the Software Heritage Archive to store the code is to build adapters for the different platforms, and store the code and its history, into gigantic data structures that stores all the code and its information, which is not easy at all. This makes searching for code very easy for anyone, as it is like searching on google. Another advantage of this approach is that all code is presented the same way, no matter where it comes from.

1.4.- The cost and financing it.

One of the biggest problems of this kind of projects is financing it, since it needs an enormous infrastructure. At the start of the project, one proposal was to create a private company, where they would sell services to stakeholders. The problem of this approach is that if the company shuts down, the archive would be lost. For this reason, they decided to start a nonprofit, multi-stakeholder, international organization. Thanks to this decision, they signed an agreement with UNESCO. Currently, they have funding from 20 different organizations. These organizations include companies, academias, universities or even ministries.

2. How frequently archive new data and how to do it

The mechanism for registering these contents, consist of three different ways. The first one is to take all the contents from GitHub for example and add them to a queue where the code is being compared step by step to avoid archive repeated content. This procedure is made once each few months, so is a bit slow. If for any reason, it is required to save something immediately, it can be requested to do it on an exclusive repository from save.softwareheritage.org as long as the version-control system is supported. As the last option, a company can also sign an agreement with Sofware Heritage to have the capability of using an interface and being able to archive their software with specific metadata and quality control.

3. How to archive code SAFELY long term (and for how long) and challenges

Until now, how they earn money for at least maintain the system, how many times they archive repositories, how much data manage and how they retrieve that amount of code questions were treated. However, one of the most difficult challenges of this project is how to stay alive for a long time.

This project works by the moment with 180 million projects, which are 12 billion code files. This is one petabyte of information stored. Apart from this code files, relations among files and repositories like, different commits done, structure of directories, revisions, releases, etc, are also stored, therefore, due to these relations, a no relational database is strongly needed. It can perform reads in an efficient way and at the same time is able to archive quickly

They have entire copies of all the database in external services to keep the code save in case of failure of the system caused by a hacker, an erroneous action done by a developer or whatever. In addition, they have tools for checking the integrity of the data and in case something has been modified, it can be restored with only searching for that content in one of the other separate copies and duplicate again the node in the original graph. Checking the state of each repository is thanks to the use of cryptographic identifiers. Another threat are laws, as occurred some years ago, that Roberto's organization had to collaborate with other open-source organizations to avoid to be passed one legislation that would affect a lot to these sector.

Regarding the graph that Roberto Di Cosmo mentions many times, is a Merkle graph. This solution was chosen as it allows to create as many relations as you need among nodes, so for example, this is the main key to save space because when many repositories use the same resources, instead of making a copy for each repository, they linked them to the resources. With this method, the team was able to reduce the actual use of space three hundred times. This means that, in case they were not applying that policy, the database will occupy 300 petabytes.

4. Software Heritage ID

Software Heritage needed a way to uniquely identify each file, directory or commit uploaded so they could keep track of versions and check that there are no repeated elements. For this they debated between using an intrinsic or extrinsic attribute. The main difference is that extrinsic attributes need an authority that keeps track of which number is assigned to which file, so, if it is hacked or malfunctions, there is no way to check the validity of the IDs. In the other hand, intrinsic attributes only need agreeing upon a standard to work, when it is established, any organization can generate an ID from the file so it cannot be compromised. The latter was the choice of SWH and for computing it they used SHA-1 with a little variation. They called this ID is called "SWHID" and it is registered with IANA but now, they are working to make it an ISO standard.

5. Get involved

As Software heritage is an open-source project everyone can participate, to do so you can check their official webpage, www.softwareheritage.org to find ways to contribute. Finally, it is encouraged to browse their page as it can satisfy your curiosity about how a certain software was developed or you could find solutions to certain problems you are facing in your own work, it really is like a Library of Alexandria of code.