

Preserving software's legacy

In-Depth Latest News Published: December 5th, 2018 - Christina Cardoza

Twitter

Email

+ More

All throughout our lives we are reminded of events from the past. History teaches us about what happened before us to help us understand how society came to be as it is today. But today we live in a digital age, and while leaders, laws, wars and other parts of our history will always be important to know; what about software? Technology is everywhere and it is rapidly changing every day. Should we care about where it all started?



RELATED CONTENT: [The ubiquity of shared code](#)

The Software Heritage was launched with a mission to collect, preserve and share all software source code that is publicly available. It is currently working towards building the largest global source code archive ever. The Software Heritage was founded by the French Institute for Research in Computer Science and Automation Inria, and it is backed by partners and supporters such as Crossminer, Qwant, Microsoft, Intel, Google and GitHub.

According the heritage, software is an essential part of our society and lifestyle. It has become crucial for businesses and industries to succeed. It's enabled the emergence of social and political organizations. It has provided us the ability to communicate, pay bills, purchase goods, access entertainment, find information, and more. It would be a shame to our future generations if we were to lose access to that.

"Cultural heritage is the legacy of physical artifacts and intangible attributes of a group or society that are inherited from past generations, maintained in the present and bestowed for the benefit of future generations," the organization wrote on its [website](#). "Software in source code form is produced by humans and is understandable by them; as such it is an important part of our heritage that we should not lose. Software is furthermore a key enabler for preserving other parts of our cultural heritage that we would de facto lose if we lose the software needed to access them. Preserving software is essential for preserving our cultural heritage."

According to Robert Di Cosmo, founder and CEO of the Software Heritage, the organization specifically looks at source code because it is written and understood by humans, and contains high-level programming languages that explain what they want machines to do.

There are three main properties of the source code collected: availability, traceability and uniformity. The way the heritage collects the software is like a search engine, explained Stefano Zacchiroli, founder and CTO of the Software Heritage. It crawls specific places where software lives and where developers go to develop, collect or distribute software. Some of the current places include Debian, GitHub, GitLab, Gitorious, Google code, GNU, HAL, Inria, and Python. It will constantly continue to go back to these places to look for new software and updates.

Currently, the Software Heritage has archived more than 80 million software projects, according to Di Cosmo. Zacchiroli explained it is not only bits of code itself that the heritage is collecting. It is collecting the development history from the first version that was stored to the commit data and all the releases, information and artifacts attached to a project along the way.

“It is very important to archive software of programs because even when the machines on which the programs were supposed to be running will no longer be available, source code will still be a valuable part of knowledge,” Di Cosmo said in a [video](#). “Today, the amount of the software that has been developed around the world is actually doubling every two to three years, so we need a common platform where we collect all the software and can study it to find the efforts, to improve the quality, to understand what is going on and to prepare a better future in software development.”

Aside from preservation, Zacchiroli hopes the archive will be used for scientific and industrial applications. For scientific use, Zacchiroli says scientist can mirror the archive and run experiments using the dataset. “Software preservation is a pillar of reproducibility, because software is used in essential ways during all phases of research in all fields of science. To be able to reproduce an experiment, knowing the exact version of the software used is essential,” the Software Heritage wrote on its website. “Software Heritage will ensure availability and traceability of software, the missing vertex in the triangle of scientific preservation.”

For industrial applications, Zacchiroli explained that every IT product nowadays contains some [open-source](#) software. The problem, however, is open-source software usually changes or uses different environments, and developers are some times only tracking a specific version. According to Zacchiroli, since the Software Heritage tracks the origin, history and evolution of projects, it will be able to spot and fix vulnerabilities easier. “Software Heritage will provide a universal, vendor-neutral, persistent reference software archive. On this foundation, a wealth of new applications will emerge, improving all aspects of the software process, and leading to better software for everybody,” the heritage wrote.


The Software Heritage is not the only organization looking to preserve software. The Software Preservation Network is an organization active in the United States that is also working to ensure long-term preservation and access to software. “We connect and engage the legal, public policy, social science, natural science, information & communication technology and cultural heritage preservation communities that create and use software. Our work currently involves: legal licensing and information policy research; an international registry of software collections; and software development contributions to technical infrastructures that facilitate long-term access to software,” the organization wrote on its [website](#). According to Software Heritage’s Di Cosmo, Software Heritage and the Software Preservation Network’s roles are complementary and the two actively collaborate on areas such as software project metadata and archiving legal issues.

Going forward, the Software Heritage hopes to increase coverage, provenance information, and improve its source code indexing and search capabilities.

“We are so focused on developing new things that we forget about storing, preserving what we have developed up to now,” Di Cosmo said in the video.

ARTICLE TAGS

[open source](#), [software](#), [source code](#)

 **Subscribe to SDTimes**

 [Twitter](#)

 [Email](#)

 [More](#)



About Christina Cardoza

Christina Cardoza is the News Editor of SD Times. She is responsible for the oversight of the daily news published to the website as well as the company's weekly newsletter, News on Monday. She covers agile, DevOps, AI, machine learning, mixed reality and software security. She is an undeniable nerd who loves Marvel comics and Star Wars. On Follow her on Twitter at [@chriscatdoza](#)

[View all posts by Christina Cardoza](#)