Software Code’s “Wayback Machine” Gets a Boost

CONOR REYNOLDS 30TH FEBRUARY 2009

The lack of software intelligence surrounding source code history and licensing puts many companies in danger of losing valuable IP.

Software Heritage is a non-profit initiative developed and hosted by the French Institute for Research in Computer Science and Automation.

Officially created in 2015, the project has been growing over the years. It now stores 5.6 billion source files from more than 70 million projects.

Software Heritage is itself built on open-source code. It gathers source files by crawling through repositories that developers use to create and share code, such as GitHub, GitLab, Google Code, Biele, CNX and the Python Package Index, with users able to trace detailed versions of all the codebase versions that exist.

Software Heritage: Adds “Provenance Index”

Now the non-profit, which is backed by industry leaders including Intel, Google and Microsoft, has announced a new partnership with Paris-based software company CAST that will see the two create a “provenance index” of code.

The provenance index enables users on the Software Heritage platform to search for the original sources of any given source file.

Software Composition Analysis is Getting Important

CAST added software composition analysis to this tool in 2019, when it brought together a “knowledge base” of open-source components found by its, the public science and technology institution dedicated to computer science learning.

Plug-and-play technology together will provide rapid identification of third-party source code across more than five billion known source code files, enabling better detection of external code, license risks and vulnerabilities, CAST said.

With Software Heritage containing information about known application security vulnerabilities in addition to copyright for all known software in use, CAST identifies its needability and usefulness as crucial where a bill of materials is required.

This might include decisions like when an enterprise is outsourcing software development, buying software assets or shutting a merger or acquisition.

A software bill of materials is a list of components in a piece of software, vendors often create products that bring together open-source and commercial software components.

“Today it's hard to understand what software is doing behind a company’s firewall,” said Vincent Delahaye, founder and CEO at CAST.

“A software bill of materials reports to software developers what open source and other external components in code expose their organization to non-compliance, legal action and possible loss of proprietary IP,” he added.

Software Heritage is project developers a means to have given themselves just one clearly defined mission: “We are committed to collect, index, preserve and make easily accessible the source code of the software that lies at the heart of culture,” they wrote on their website. Those curious can download the source code files (.c, .java, .py) or enter their SHA1 IDs to see what it has here.