Software Code's "Wayback Machine" Gets a Boost

CONOR REYNOLDS
18TH FEBRUARY 2019

➕ INCREASE / DECREASE TEXT SIZE ➖

📷 Add to favorites

"The lack of Software Intelligence around open source versioning and licensing puts many companies in danger of losing valuable IP"

Call it the Wayback Machine of code: a searchable open archive of software source code across iterations; from buggy beta versions, to sophisticated contemporary release.

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 spans 5.6 billion source files from more than 88 million projects.

Software Heritage is itself built on open-source code. It gathers source files by trawling through repositories that developers use to create and share code, such as GitHub, Gitlab, GoogleCode, Debian, GNU and the Python Package Index, with users able to trace detailed revision history of all the codebase versions that it stores.

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 occurrences of any given source file.

For the curious, that will allow "unprecedented insight" into software evolution, the two believe. CAST, however, is also keen to emphasise the opportunity afforded by plugging Software Heritage into the company's paid for application portfolio analysis and software composition analysis tool CAST Highlight.

The tool aims to make it easy to find the "first edition" of software code

Software Composition Analysis is Getting Important

CAST added software composition analysis to this tool in 2018, when it bought Antelink, a "knowledge base" of open source components founded by Inria, the public science and technology institution dedicated to computer science learning.

Plugging the two tools 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 copyrights for all known software in use, CAST describes its searchability as crucial where a Bill of Materials is required.

This might include occasions like when an enterprise is outsourcing software development, buying software assets or during 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).

"The lack of Software Intelligence around open source versioning and licensing puts many companies in danger of losing valuable IP, as most executives are unaware of their risk exposure," said Vincent Delaroche, Founder and CEO at CAST.

"Business leaders should be aware when 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's project developers meanwhile 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 our culture," they state on their website. Those curious can drag and drop source code files (.c, .java, .py, ...) or enter their SHA1 into to search bar here.