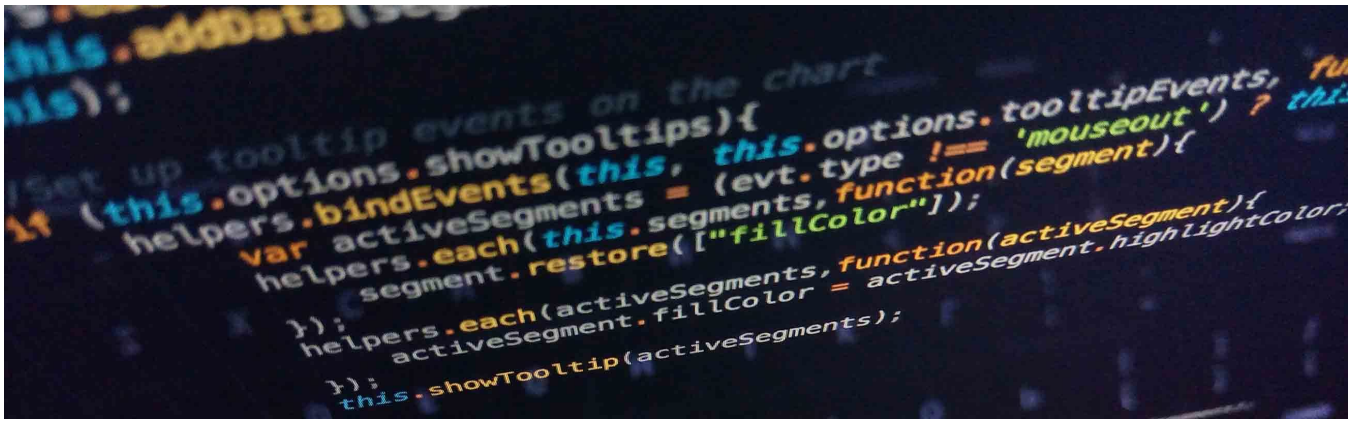
UNDARK



VARIABLES / *News & Features*

A Wayback Machine for Source Code

Modern digital life relies on layers of shared and dependent code that is, over time, vulnerable to deletions. Will an archive help?

05.02.2018 / BY Dalmeet Singh Chawla

2 COMMENTS

IN MARCH 2016, software developer Azer Koçulu famously broke the internet by taking 11 lines of open source computer code he had written offline. The problem: millions of software packages written in the programming language JavaScript had been built on top of Koçulu's code, or they were built on top of other packages that, in turn, were built on top of the code Koçulu wrote. "I think I have the right of deleting all my stuff," Koçulu wrote bluntly in an email at the time.

Whether that's true or not, it demonstrates the fragility of our modern digital existence, with layer upon layer of dependent code vulnerable to even minor deletions. It also, activists argue, highlights the need for better backup and preservation, particularly given that precious little modern software is built from scratch, and instead relies on pre-existing bits of code from what can seem like a limitless number of sources. And that's at least one rationale behind the "Software Heritage" project, a sort of Wayback Machine for software. The project plans to

create an archive of computer code source files as they appear on the web — an undertaking that has implications not just for history, but for science and research, too.

Since 2015, archivists at the Software Heritage project, which is hosted by the French Institute for Research in Computer Science and Automation, have been collecting open source code available at various online repositories and websites. To date, the archive contains more than 4 billion source files from more than 80 million projects, says Roberto Di Cosmo, a computer scientist who is directing the project in Paris. In cases where open source code disappears, or the server it is stored on is hacked, destroyed or lost, the platform aims to become the go-to place for a backup version.

In the coming weeks, Di Cosmo and colleagues plan to release the archive for anyone to access for the first time. Adding code to the platform, however, will continue in the same fashion, Di Cosmo says. He speculates that the archive currently contains only around a quarter of the world's open source software, noting that code is often published in hard-to-access places on the internet.

The initiative has attracted interest from some big names, with U.S. technology firms Microsoft and Intel and French bank Société Générale each investing more than $120,000 per year since 2016. Several other sponsors have also backed the project with smaller investments. Last year, the project signed a partnership deal with The United Nations Educational, Scientific and Cultural Organization (UNESCO). An Intel spokesperson told Undark: "Intel's sponsorship demonstrates our investment in corporate social responsibility through cultural and historical preservation," noting that the move also helps attract code experts who may not have considered partnering or collaborating with the firm.

Until now, Di Cosmo says, the project has focused on backing up software from well-known databases such as GitHub — another project sponsor. In future, the process will be simpler since code from various places will be automatically copied onto the platform, he notes, and every version of each source file will be stored.

Neil Chue Hong, founding director of the U.K.'s Software Sustainability Institute at the University of Edinburgh, thinks preserving code is important. "Increasingly, knowledge is captured in the code and the decisions made when developing it," says Hong, who is not part of Software Heritage — though he doesn't think that Software Heritage will ever contain all the world's open source code. Even the nonprofit Internet Archive, which has stored hundreds of billions of websites in various iterations of design and evolution since it began in 1996, falls short of capturing every change and iteration across the entire web. Software changes even more frequently than webpages, Hong says — though he adds that he believes archiving even a fraction of the total would still be valuable.

As it stands, most of the world's open source software is currently stored by GitHub, which has "no long-term commitment to making that source code available," according to Arfon Smith, a former GitHub employee and stockholder who edits The Journal of Open Source Software. "If GitHub went bust tomorrow, there would definitely be a loss of some of this software and so it makes sense to me to keep an archive," he says. In 2015, Google also announced the shutdown of its open source project hosting site Google Code, advising developers to shift their code elsewhere.

For Di Cosmo, however, the archive isn't just about backing up software. It also presents an opportunity to make it more discoverable. Ultimately, he says, the platform will classify and categorize software that will make it incomparably better than using a search engine to hunt for code. (The version due to be released in the coming weeks will not yet be fully categorized).

THE ARCHIVE might well have impacts on science writ large. After all, modern computer code is also written by scientists for research purposes. And while science's reproducibility problem has generated calls for better sharing of information about data, protocols, and lab materials, Di Cosmo argues that the computer code underpinning much published science is often overlooked. (The oversight prompted the journal Nature to announce in March that its editors will begin seeking review of bespoke computer code used in papers submitted for publication.) Software Heritage, he suggests, could help to ensure that this code is not just backed up, but available for future scrutiny.

"Most published research uses software to either create or process data," Hong says. "If you can't get access to the software used to do this, it becomes impossible to verify the results."

Stephan Druskat, a research software engineer at the Humboldt University of Berlin, Germany, suggests that projects like Software Heritage can help to reframe our understanding of computer code as a valuable and important research product unto itself. Many critics have long felt that coding is an underrecognized and underappreciated task in academia. "One could also imagine," Druskat says, "that the deposit of research software in the Software Heritage archive could, for example, be made a requirement by publishers in the process of submitting a paper."

Smith, who now heads the Data Science Mission Office at the Space Telescope Science Institute in Baltimore, Maryland, suggests that this sort of archival approach to software used in science could significantly enhance the long-term reproducibility of research. "In theory, scientific research should become more reproducible in the long-term," he says. "At some point in the future, if I want to

repeat a result from a few years back and the only place I can find the source code associated with a result is on the Software Heritage site, then they have done their job as an archive of this software."

Information scientist James Howison from the University of Texas at Austin calls the project an "audacious effort" but says it has high potential. "It goes well beyond what other people have conceived," he says. He notes that Software Heritage could help solve the problem of reference rot, where links to software in scientific papers lead to broken webpages.

Researchers who write and rely on code have long found it difficult to receive credit for their work, particularly as bits of code become embedded in other projects that fold in still more bits of code from numerous sources. A computer program called Depsy, launched in 2015, aimed to fix that problem by measuring the extent to which code is reused by other software packages — a process known as calculating dependencies. Depsy does this by scouring public code repositories and tracking mentions of software in academic papers. Di Cosmo believes that Software Heritage could help to scale up the approach of Depsy and other initiatives by providing the largest reference archive and a common platform for analyzing the development history of all software source code.

HONG SAYS Depsy could also trawl Software Heritage's archive to work out its dependencies, though as of March, the site is no longer actively maintained. He adds that if Software Heritage makes available enriched relationship information between different software files — such as how they interconnect in terms of authors, dates, and times — that would make it easier for Depsy to calculate the impact of software. Software Heritage could also pave a way to standardize how software is cited and reported in literature, says Hong. "Current practice relies on the researcher taking the time to archive the code themselves," he notes. "If they just need to point to the Software Heritage archive, this becomes much easier."

Di Cosmo says the project's next stages involve establishing a foundation, which, he hopes, would help attract long-term endowments.

“Our plan is not to actually make money,” he says. “What we want to achieve is sustainability.”

Dalmeet Singh Chawla is a freelance science journalist based in. London.

Top visual: Irvan Smith/Unsplash

....

code archives, Dalmeet Singh Chawla, French Institute for Research in Computer Science and Automation, open source, Software Heritage Project, software preservation, source code