Software Heritage, the “Library of Alexandria of software,” launches today

arstechnica.co.uk/business/2016/06/software-heritage-the-library-of-alexandria-of-software-launches-today/

By Glyn Moody

One of the images used on the Software Heritage website.

The Software Heritage project aims to "collect, organise, preserve and make easily accessible the source code of all software that is publicly available." Calling itself the "Software Wikipedia," the "Internet Archive for source code," and even the "Library of Alexandria of software," it believes "a single and universal archive making software source code readily available will facilitate access to the knowledge contained therein, support programming education, and create a reference catalogue with all knowledge about this software."

At launch, the project says that it had already "ingested in the Software Heritage archive a significant amount of source code, possibly assembling the largest source code archive in the world." That currently amounts to 2 billion source files, 600 million commits, and 22.7 million projects.

Major holdings include public, non-fork repositories from GitHub, source packages from the Debian distribution (as of August 2015, via the snapshot service), and tarball releases from the GNU project (as of August 2015). As that list shows, the emphasis is on free software, although the project's website repeatedly refers to "publicly available" code as its target. Ars has asked for clarification on what this means, but has not yet received any response.

The project has been initiated by the French research institute INRIA, which also provides most of the people working on it. Founder and CEO of the Software Heritage project is Roberto Di Cosmo. A blog post on the launch explains that the team has been working on the site for "over a year."

Although INRIA is leading the project, the aim is to "encourage the emergence of an open network of peers and mirrors that will share with us the responsibility of maintaining available several copies of all the software we collect."

Rather ironically, the main industry partner of the project seems to be Microsoft. The company's open technologies chief, Jean Paoli, is quoted as saying: "We applaud the Software Heritage as an open project that will help curate and conserve human knowledge in the form of code for future generations as well as help today’s generations of developers find and re-use code worldwide. We are proud to be one of the first industry partners for this initiative and to provide the Azure infrastructure to ensure the data is highly available."

There are a number of testimonials supporting the project from well-known names in the computing, academic

and business worlds, and the project is seeking "contributions from institutions, foundations, corporations, and individuals that wish to substantially support our mission."

As well as creating a huge archive of source code, the Software Heritage project aims to "index, organise, make referenceable and accessible" all its holdings. "We will provide unique identifiers, intrinsically bound to the software components. This will ensure that a resilient web of knowledge can be built on top of the Software Heritage archive. Software Heritage will foster the emergence of a variety of services, ranging from documentation to classification, from search to distribution."

One of the driving forces behind the project is the increasing importance of software for science, hitherto rather neglected, and the need to preserve it along with other resources: "Science relies more and more on software. To guarantee scientific reproducibility we need to preserve it. Amassing source code at this scale will be challenging, but will also enable the next generation of software studies."

The Software Heritage project also believes that industry will benefit from its work: "Software is present in all industrial processes and products. The universal source code archive we are building will help industry with provenance tracking, long-term archival, and software bill of materials."