Software Heritage: why and how we collect, preserve and share all the software source code

Roberto Di Cosmo
Software Heritage, Inria Paris
Paris, France
roberto@dicosmo.org

ABSTRACT

Software is at the heart of our digital society. It powers our industries, fuels innovation, mediates access to all digital information, is a pillar of modern scientific research, and has enabled the emergence of new forms of social and political organizations—"code is law", as Lessig said [2].

The source code of this software is a unique form of knowledge: it is designed to be read by humans (the developers), and yet it is ready to be translated into an executable form for a machine. As Len Shustek puts it, "Source code provides a view into the mind of the designer" [3].

Software source code is precious, and embodies a growing part of our scientific, technical and organisational knowledge.

Software Heritage is an open, non-profit initiative whose mission is to ensure that this precious body of knowledge will be preserved over time and made available to all. We do this for multiple reasons.

To preserve the scientific and technological knowledge embedded in software source code, that is a precious part of our heritage.

To allow better software development and reuse for society and industry, by building the largest and open software knowledge database, enabling the development of a broad range of value added applications.

To foster better science, by assembling the largest curated archive for software research, and building the infrastructure for preserving and sharing research software, a necessary complement to Open Access, and a stepping stone for reproducibility.

We do this now, because we are at a turning point: the founding fathers are still around, and willing to contribute their knowledge, but only for a limited time. And we face the risk of massive lossage of source code developed by the Free and Open Source community, with code hosting sites that shut down when their popularity decreases.

We do this in with an open approach, based on principles specifically designed to maximise the chance success in the long run, as the mission is a long term one .

Software Heritage has been started by Inria, who supports the initial effort necessary to get it up to speed, and is now transitioning to an independent structure that will welcome partners from all areas of interest, much like what was done over two decades ago for the World Wide Web Consortium.

Software Heritage archives already more than 4 billion unique source code files, spanning more than 70 million projects, with their full development history.

Building such a unique knowledge base brings about new challenges. Some are legal and organisational, others are financial. Many are research questions, ranging from classification of the software projects to compact representation of the history of development, from distributed storage to efficient query languages.

Now, we call on computer scientists and computer technologists to contribute to this grand challenge for the benefit of all.

ACM Reference Format:
Roberto Di Cosmo. 2018. Software Heritage: why and how we collect, preserve and share all the software source code. In *Proceedings of International Conference on Software Engineering (ICSE'18).* ACM, New York, NY, USA, Article 4, 1 page. https://doi.org/10.475/123_4

REFERENCES

[1] Roberto Di Cosmo and Stefano Zacchiroli. 2017. Software Heritage: Why and How to Preserve Software Source Code. In *Proceedings of the 14th International Conference on Digital Preservation, iPRES 2017, Kyoto, Japan.* https://hal.archives-ouvertes.fr/hal-01590958 Available from https://hal.archives-ouvertes.fr/hal-01590958.
[2] Lawrence Lessig. 1999. *Code and other laws of cyberspace.* Basic books.
[3] Leonard J. Shustek. 2006. What Should We Collect to Preserve the History of Software? *IEEE Annals of the History of Computing* 28, 4 (2006), 110–112. https://doi.org/10.1109/MAHC.2006.78

ICSE'18, 27 May - 3 June, Goteborg, Sweden

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4