Software Heritage: Collecting, Preserving, and Sharing All Our Source Code (Keynote)

Roberto Di Cosmo
Inria / University Paris Diderot
France
roberto@dicosmo.org

ABSTRACT

Software Heritage is a non profit initiative whose ambitious goal is to collect, preserve and share the source code of all software ever written, with its full development history, building a universal source code software knowledge base. Software Heritage addresses a variety of needs: preserving our scientific and technological knowledge, enabling better software development and reuse for society and industry, fostering better science, and building an essential infrastructure for large scale, reproducible software studies. We have already collected over 4 billions unique source files from over 80 millions repositories, and organised them into a giant Merkle graph, with full deduplication across all repositories. This allows us to cope with the growth of collaborative software development, and provides a unique vantage point for observing its evolution. In this talk, we will highlight the new challenges and opportunities that Software Heritage brings up.

CCS CONCEPTS

• **Software and its engineering**;

KEYWORDS

Software archive

ACM Reference Format:
Roberto Di Cosmo. 2018. Software Heritage: Collecting, Preserving, and Sharing All Our Source Code (Keynote). In *Proceedings of the 2018 33rd ACM/IEEE International Conference on Automated Software Engineering (ASE '18), September 3–7, 2018, Montpellier, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3238147.3241985

OVERVIEW

Software powers our industries, fuels innovation, mediates access to digital information, is a pillar of modern scientific research: in a word, it is at the heart of our digital society.

The *source code* of this software is a unique form of knowledge: it is designed to be read by humans, the developers, and it is at the same time 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].

Hence, software source code is precious: it embodies a growing part of our scientific, technical and organisational knowledge, and is now a part of our own cultural heritage.

Software Heritage is a non profit initiative whose ambitious goal is to make sure this precious body of knowledge will not be lost. Its mission is to collect, preserve and share the source code of all software ever written, with its full development history, building a *universal source code software knowledge base.*

Software Heritage addresses a variety of needs: preserving our scientific and technological knowledge, enabling better software development and reuse for society and industry, fostering better science, and building an essential infrastructure for large scale, reproducible software studies.

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

It strives to enable 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.

It contributes to fostering 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 have started this initiative 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 [1].

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.

We have already collected over 4 billions unique source files from over 80 millions repositories, and organised them into a giant Merkle graph [2], with full deduplication across all repositories. This allows us to cope with the growth of collaborative software

ASE '18, September 3–7, 2018, Montpellier, France

ACM ISBN 978-1-4503-5937-5/18/09.
https://doi.org/10.1145/3238147.3241985

development, and provides a unique vantage point for observing its evolution.

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.

In this talk, we will highlight some of the challenges and opportunities that Software Heritage brings up.

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

BIOGRAPHY

After obtaining a PhD in Computer Science at the University of Pisa, Roberto Di Cosmo was associate professor for almost a decade at Ecole Normale Supérieure in Paris, and became a Computer Science full professor at University Paris Diderot in 1999. He is currently on leave at Inria. He has been actively involved in research in theoretical computing, specifically in functional programming, parallel and distributed programming, the semantics of programming languages, type systems, rewriting and linear logic. His main focus is now on the new scientific problems posed by the general adoption of Free Software, with a particular focus on static analysis of large software collections, that were at the core of the european reseach project Mancoosi. Following the evolution of our society under the impact of IT with great interest, he is a long term Free Software advocate, contributing to its adoption since 1998 with the best-seller Hijacking the world, seminars, articles and software. He created the Free Software thematic group of Systematic in October 2007, and since 2010 he is director of IRILL, a research structure dedicated to Free and Open Source Software quality. In 2016, he co-founded and directs Software Heritage, an initiative to build the universal archive of all the source code publicly available.

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] Ralph C Merkle. 1987. A digital signature based on a conventional encryption function. In *Conference on the Theory and Application of Cryptographic Techniques.* Springer, 369–378.

[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