Software Heritage : « une infrastructure partagée, dédiée à la recherche, à l'industrie et au patrimoine culturel »

09 décembre 2020 **RECHERCHE**

Le CNRS rejoint Software Heritage et apporte un soutien de 100 000 euros par an à cette bibliothèque universelle de codes sources de logiciels, lancée par Inria et soutenue par l'UNESCO. Son directeur, Roberto Di Cosmo en détaille les ambitions.

Professeur d'informatique[1] et chercheur chez Inria, Roberto Di Cosmo s'intéresse dès les années 1990 à ce que l'on appelle les « logiciels libres », qui, à la différence des « logiciels propriétaires », donnent accès à leur « code source » et peuvent être distribués ou modifiés. « *Le code source des logiciels est une sorte de littérature technique du XXIe siècle, écrite par des humains pour que des ordinateurs l'exécutent, et pour que d'autres humains puissent la lire, la comprendre, l'adapter et la réutiliser* », explique le chercheur.

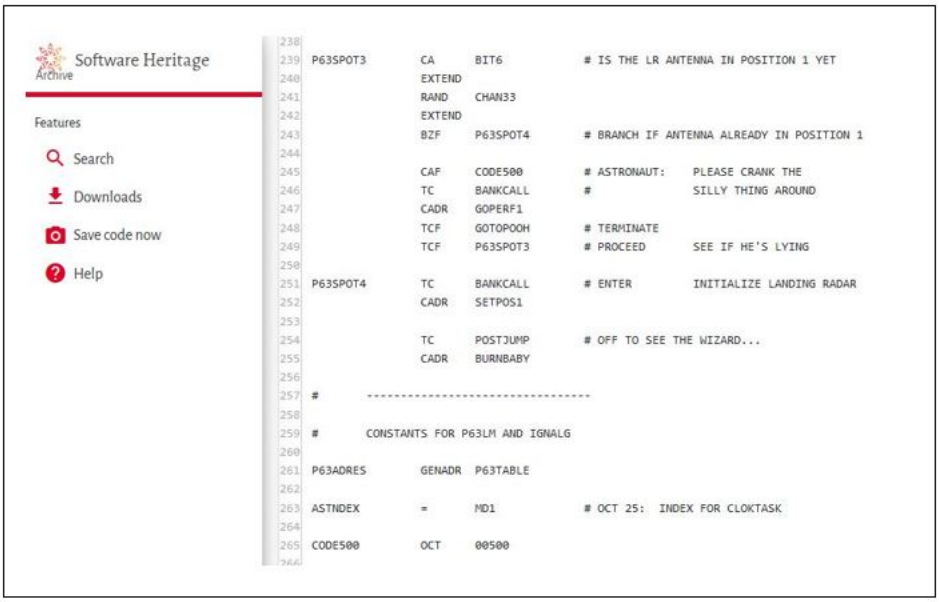
L'adoption généralisée des logiciels libres a rendu accessible ces « codes sources », mais il y avait un problème : « *Les codes sources des logiciels libres étaient éparpillés dans une variété de plateformes, et on s'est rapidement aperçus qu'aucune entité dans le monde n'avait assumé la mission de les réunir tous et de les préserver, au contraire de ce qui est fait pour le patrimoine audiovisuel ou documentaire*[2] », indique Roberto Di Cosmo. « *En effet, hors développeurs, il est difficile de prendre conscience de l'existence du code source derrière le logiciel et l'accès à ces derniers est relativement récent, allant de pair avec l'essor d'Internet et la facilitation des échanges* ». De cette observation germe l'idée de Software Heritage : construire une bibliothèque

moderne d'Alexandrie qui préserve et rend accessible les codes sources des tous les logiciels publiquement disponibles.

Un patrimoine culturel « numérique » soutenu par l'UNESCO

Le projet est présenté en 2014 à Antoine Petit, alors directeur de l'Inria, qui en comprend immédiatement l'impact sociétal. « *Software Heritage impacte la recherche, l'industrie ou encore l'Histoire, car il est bien question de sauvegarde de notre patrimoine de connaissances* », souligne Roberto Di Cosmo. L'équipe Software Heritage se met en place, commence à archiver les premiers codes sources en 2015 et annonce l'ouverture du site en juin 2016. Juste à temps ! La fermeture inattendue de grandes plateformes de développement collaboratif, comme Google Code et Gitorious en 2015, avait mis en danger plus de 1 million et demi de projets de développement (heureusement tous sauvegardés), et montré clairement le besoin d'une archive universelle à long terme, non soumise aux logiques de marché.

En 2017, l'UNESCO se joint à l'aventure, affirmant l'importance du code source comme « *un véritable patrimoine culturel qu'il est important de préserver* », et devient un partenaire privilégié de l'initiative, qui a depuis réuni de nombreux soutiens, allant des acteurs de la recherche à l'industrie. « *Notre objectif est de construire une infrastructure partagée, dédiée à la recherche, à l'industrie et au patrimoine culturel et servant les besoins de toute la planète, le tout dirigé à terme par une institution internationale.* »



Software Heritage protège déjà 9 milliards de fichiers de code source issus de plus de 140 millions de projets logiciels. Parmi les plus célèbres, on retrouve le code source du système de navigation d'Apollo 11, qui permit le premier pas sur la Lune en 1969, ou celui du navigateur NCSA Mosaic, qui popularisa l'utilisation du web des 1993. © Software Heritage

Le code source des premiers pas sur la Lune

Aujourd'hui, l'équipe de Software Heritage regroupe une douzaine de personnes allant du chercheur, à l'ingénieur ou doctorant et postdoctorant, qui développent l'infrastructure autour d'un « moissonneur » qui va chercher directement les codes sources accessibles en ligne, même si les utilisateurs peuvent également intégrer leurs codes sources au dispositif. L'archive de Software Heritage, qui pèse plusieurs centaines de téraoctets, protège déjà 9 milliards de fichiers de code source issus de plus de 140 millions de projets logiciels. Parmi les plus célèbres, on retrouve le code source du système de navigation d'Apollo 11, qui permit le premier pas sur la Lune en 1969, ou celui du navigateur NCSA Mosaic, qui popularisa l'utilisation du web des 1993, mais également de jeux mythiques des années 1990 tels que Doom ou Quake.
« *Comme dans une bibliothèque, on trouve des passages passionnants et d'autres sans intérêt ; certains sont faciles à lire, d'autres sont des merveilles de créativité qui nécessitent une étude approfondie pour les comprendre pleinement… Ces derniers sont évidemment mes préférés !* »

L'équipe de Software Heritage regroupe des chercheurs, ingénieurs ou doctorants et postdoctorants. © Software Heritage

Construire le pilier logiciel de la Science ouverte

Dans une démarche de Science ouverte, Software Heritage met aujourd'hui à disposition des chercheurs une infrastructure unique qui met le code source des logiciels sur un pied d'égalité avec les articles et les données. Il est ainsi possible d'archiver très facilement tous les codes sources pertinents et de le référencer de façon précise avec un identifiant intrinsèque, unique et pérenne,

appelé SWHID , qui est disponible pour tous les contenus, indépendamment de leur origine[3].

« *C'est la combinaison d'une archive universelle et des identifiants SWHID qui est vraiment révolutionnaire : elle garantit de toujours retrouver exactement l'objet désigné par le chercheur, jusqu'à la ligne de code, en pointant sur l'archive.* » Cela devrait simplifier nettement le travail pour tous ceux qui utilisent des logiciels dans leur recherche, à la fois pour qui rédige les articles, et pour qui les lit : plusieurs revues ont commencé à adopter cette approche, et d'autres devraient suivre.

Software Heritage est aussi interconnecté avec l'archive ouverte HAL, ce qui permet déjà d'effectuer déjà des dépôts de codes sources en les rattachant aux laboratoires des contributeurs, et de nouvelles fonctionnalités en cours de développement rendront ce processus encore plus facile.

En parallèle du travail de long terme pour construire le pilier logiciel de la Science ouverte, l'équipe de Software Heritage continue de gérer les urgences : il y a juste quelques mois, ils ont sauvé de justesse 250 000 projets qui ont été supprimés de la plateforme collaborative Bitbucket. Un sauvetage à l'image de la mission que s'est donné Software Heritage.

Notes

1. CNRS/Université de Paris
2. En France, ce sont les missions respectivement de l'INA et de la BNF.
3. Voir les articles « Archiving and Referencing Source Code in Software Heritage », ICMS, volume 12097 of Lecture Notes in Computer Science, pages 362–373, 2020 et « Announcing biblatex-software » ACM SIGSOFT Software Engineering Notes, 45(4):22--23, 2020.