**Date : 17/11/2016**
Heure : 18:12:18
Journaliste : Louis Adam

www.zdnet.fr
Pays : France
Dynamisme : 0



Visualiser l'article

# Software Heritage veut devenir l'internet Archive du code open source -

Open Source : L'Inria a profité du salon Paris Open Source Summit pour faire un point d'étape sur son projet Software Heritage. Un projet mené par 4 chercheurs, mais qui se fixe une ambition de taille : collecter tous les codes sources accessibles sur le Net.

Nous avions raté la première présentation publique de Software Heritage, mais le projet porté par l'Inria et des chercheurs de l'IRILL (Initiative de Recherche et Innovation sur le Logiciel Libre) mérite pourtant le détour. Ne serait-ce que par l'ampleur de ce projet : Software Heritage propose en effet de collecter et de préserver l'ensemble des codes sources ouverts produits par l'humanité. Pour son principal promoteur, Roberto di Cosmo, le projet vise à la conservation de la connaissance présente dans les codes sources produits par les développeurs au fil des années.

Visualiser l'article


*Roberto Di Cosmo présentait le projet à l'occasion du Paris Open Source Summit*

« Nous réalisons qu'il est très facile de perdre des informations et de perdre cet accès à la connaissance. Il suffit d'un bug, d'un crash de disque dur, d'une attaque ou même une décision business qui décide de fermer

un service comme pour Google Code, et ce sont autant de codes sources qui disparaissent » expliquait ainsi Roberto Di Cosmo lors de sa présentation sur la scène du Paris Open Source Summit.

Pas un github de plus

Software Heritage n'est-il donc qu'une plateforme de développement de plus, à l'instar de Github ? Non, précise Stefano Zacchiroli, CTO du projet, l'ambition du projet est bien différente « Ce n'est pas une plateforme de développement collaboratif, mais un effort de stockage et d'archivage de l'ensemble des codes sources. » Le but n'est pas de fournir un outil, mais bien de devenir l'Internet Archive du code source. Ainsi, Software Heritage ne se contente pas de stocker une seule version du code source, mais entend compiler l'ensemble des versions afin de pouvoir étudier leur évolution au fil du temps.

Le projet se tourne essentiellement vers les chercheurs qui souhaitent disposer de ce répertoire pour leurs études. « On peut ainsi imaginer des études portant sur la sécurité et la qualité du code via des analyses empiriques. Par exemple, on pourrait faire de la recherche de motif de bug : on sait que lorsque l'on constate certains comportements de développement, cela favorise certains types de bugs. Cela permet d'envisager des approches machine learning ou big data appliquées au logiciel libre » explique Stefano Zacchiroli. Une approche qui pourrait également permettre le développement d'outils de correction automatique du code pour les problèmes les plus courants.

Le projet a été présenté pour la première fois au public en juin 2016, mais ses initiateurs travaillent dessus depuis bientôt un an et demi. Un départ discret qui a permis aux chercheurs de mettre en place une première version fonctionnelle du projet : « Aujourd'hui, la base de données représente environ 200 To de données. On compte un peu plus de 45 millions de projets consultables sur l'archive, ce qui représente environ 3 milliards de fichiers différents » explique Stefano Zacchiroli. Le projet reprend ainsi l'intégralité des codes sources partagés via Github, mais l'ensemble du projet GNU, les dépôts Debian, ainsi que les archives de feu gitorious. Les projets récupérés chez Google Code ne devraient pas tarder à être mis en ligne.

Software Heritage vise l'exhaustivité : chaque fichier, chaque modification apportée au code est associée à un identifiant (un hash type checksum) qui permet de le retrouver parmi l'ensemble des codes sources collectés par le projet et de s'assurer qu'il n'a pas été modifié depuis sa récupération. Plus anecdotique : cet étiquetage de chaque fichier permet de retracer le « destin » d'une ligne de code à travers sa réutilisation au sein de différents projets et au fil des années.

Stocker sur la durée, un défi à relever

Car le projet fait face à un vrai défi d'archivage. « La mise en place de ces hash nous permet notamment de lutter contre le phénomène de bistrot et de nous assurer que le code stocké sur l'archive ne se perd pas avec le temps. » explique Stefano Zacciroli. « Nous avons plusieurs solutions pour faire face aux différents risques. Nous avons ainsi deux copies en interne de la base de données, ainsi qu'une troisième en externe hébergée sur le cloud de Microsoft. L'objectif, c'est évidemment de démultiplier les copies grâce à de nouveaux partenaires afin de s'assurer de la redondance des données stockées. »

Oui vous avez bien lu : le projet est mené et maintenu par des libristes convaincus et Microsoft en est le premier sponsor. Si l'arrivée surprise de Microsoft à la tête de la Linux Foundation ne vous avait pas suffi pour comprendre, voici donc une nouvelle démonstration de l'amour que porte aujourd'hui l'éditeur au monde du logiciel libre après l'avoir pendant longtemps publiquement conspué. On peut trouver la stratégie grossière, mais business is business et aujourd'hui, l'open source est roi dans le cloud. Rien de vraiment très étonnant donc.

Software Heritage souhaite de toute façon multiplier les partenaires. « L'objectif à long terme, c'est de devenir une fondation complètement indépendante et de multiplier les partenariats » précise Stefano Zacchiroli. Parmi les plus récents, Software Heritage a ainsi annoncé un partenariat avec les prestigieux Bell Labs détenus par Nokia ou encore avec l'Unesco. Et invite les entreprises et organisations intéressées à leur proposer une collaboration via leur site web.