Viewpoint: Building the Universal Archive of Source Code

Le logiciel est un pilier de la plupart des activités de recherche scientifique dans tous les domaines de notre vie et un médiateur pour accéder à toute information numérique. C'est un héritage humain qui doit être préservé des éléments supprimés, endommagés ou égarés. Spéciale ces dernières années, nous avons assisté à la fermeture de codes clés, exposant des centaines de milliers de projets de programmes publics accessibles au public. Nous devons créer un archivage universel du code source des logiciels, afin d'attirer l'attention sur la sécurité, la sécurité, la fiabilité et la traçabilité des logiciels.

Software Héritage est une initiative lancée par Inria, l'institut français de recherche en informatique et en automatisme, dans le but de remplir cette mission avec trois objectifs principaux: collecter, conserver et partager le code source de tous les logiciels jamais écrits.

Collection

L'objectif de Software Héritage est de collecter tout le code source sans filtrage a priori.

- Pour le code source libre actuel, Inria a besoin d'un processus automatisé pour gérer tous les projets de logiciels, avec tout l'historique de développement disponible. Le défi technique consiste à créer des robots pour chaque plate-forme d'hébergement de code et à développer des adaptateurs pour tous les systèmes de contrôle de version et les formats de package.
- Inria a besoin d'une plate-forme participative pour responsabiliser les volontaires prêts à contribuer à la récupération du logiciel.
- Les logiciels fermés sont plus difficiles à récupérer. La recherche peut réussir à récupérer et à libérer son code source en fournissant un moyen de maintenir en toute sécurité les logiciels sources fermés sous embargo.

Préservation

Software Héritage utilise et développe exclusivement des outils logiciels libres et ouverts pour la construction de ses archives. Ils souhaitent également créer un réseau géographiquement distribué, mis en œuvre à l'aide de diverses technologies de stockage, dans différents domaines administratifs, contrôlé par plusieurs institutions et préservant l'historique de développement du code source.

L'approche unique de Software Héritage consiste à stocker tout le code source disponible et ses révisions dans un seul Merkle DAG (Directed Acyclic Graph), partagé entre tous les projets logiciels. Cette structure de données facilite la distribution et permet une déduplication complète, une vérification de l'intégrité et un suivi de la réutilisation dans tous les projets logiciels au niveau du fichier.

Partage

La matière première collectée par Software Héritage doit être correctement organisée pour en faciliter la réalisation. Ils ont besoin de métadonnées décrivant le logiciel et de moyens pour classer les millions de projets récoltés, extraire et réconcilier les informations existantes provenant de nombreuses sources différentes, encodées dans l'une des nombreuses ontologies logicielles, et les compléter à l'aide d'outils automatiques ou de fournisseurs multiples.

Statut actuel

Software Héritage est un projet actif, il contient plus de quatre milliards de fichiers de code source uniques et un milliard d'engagements individuels. Trois copies sont actuellement conservées, dont une sur un cloud public. Sous forme de graphique, le DAG Merkle qui sous-tend l'archive comprend 10 milliards de nœuds et 100 milliards d'arêtes. En termes de ressources, l'archive compressée et entièrement dédupliquée nécessite environ 200 To d'espace de stockage. Ces chiffres grandissent constamment.

Prochaines étapes

- Pour la phase de collecte, Inria a besoin d'aide pour récupérer les logiciels importants du passé et créer des adaptateurs pour les nombreuses plates-formes d'hébergement et formats de distribution de code source.
- Pour la phase de préservation, Inria a besoin de ressources pour héberger des miroirs, ainsi que de contributeurs disposés à essayer différentes technologies pour stocker et mettre en miroir les archives.
- Pour la phase de partage, une aide est nécessaire pour organiser le contenu, créer des mécanismes efficaces d'indexation et d'interrogation et développer des applications pour des domaines spécifiques.