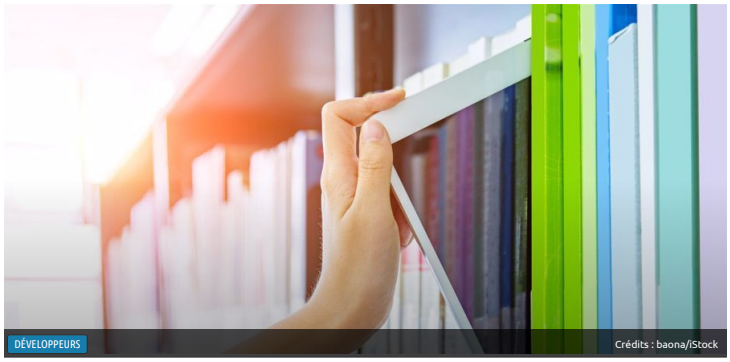
# Software Heritage, le défi d'une archive mondiale du code source

Internet n'oublie rien 21 • 0




11 min Offrir cet article

Par Guénaël Pépin
le jeudi 14 juin 2018 à 14:18

**Lancé il y a trois ans, Software Heritage ouvre officiellement ses portes, avec le soutien de grands noms du numérique. Le projet, d'abord soutenu par Inria, n'est pourtant pas au bout de ses peines pour archiver tous les logiciels de la planète. Nous en discutons avec son fondateur, Roberto Di Cosmo.**

Une bibliothèque mondiale, contenant tout le code source jamais produit. C'est l'ambition de Software Heritage, dont l'archive est ouverte au public depuis quelques jours. « *Après trois ans acharnés à construire l'infrastructure, à collecter les données, à les indexer... on est très contents d'ouvrir les portes de l'archive !* » nous lance le chercheur Roberto Di Cosmo, qui dirige une équipe d'une dizaine de personnes, dont six à plein temps.

Le projet, d'abord soutenu par Inria, gagne ici une dizaine de financeurs privés, même si près de la moitié du budget vient toujours de l'institut français. Avec un financement entre 800 000 et un million d'euros par an, le projet compte devenir une fondation, garante de ce patrimoine, comme l'Internet Archive l'est pour le web.

Software Heritage se voit surtout comme une infrastructure, autant utile à des organisations culturelles qu'aux industriels, aux chercheurs ou aux enseignants. Avec plus de 83 millions de projets et 4,5 milliards de fichiers en banque, l'archive n'en est qu'à ses balbutiements. Les difficultés techniques, juridiques et financières sont nombreuses, tout comme les surprises pour Roberto Di Cosmo.

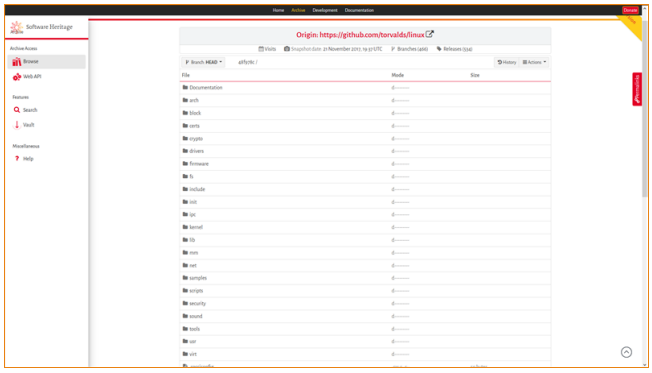
- Consulter l'archive de Software Heritage

## Une initiative nouvelle, mais une idée qui date

L'idée est née en 2014, avec de premiers travaux discrets à l'été 2015. L'annonce ne date que de juin 2016. « *On vient du monde du logiciel libre. On préfère d'abord faire les choses, puis en parler. Ce n'est pas comme une start-up, c'est une autre approche* » justifie Di Cosmo, dont le projet ne présentait que la taille de l'archive, sans accès, à l'époque.

Début 2017, un partenariat avec l'Unesco lui donne une légitimité politique. Cette première ouverture au public est donc un aboutissement, pour un projet qui se veut pérenne. Étonnamment, Software Heritage serait la première initiative de cette ampleur pour le code source.

La vision n'est pourtant pas nouvelle. Roberto Di Cosmo relève surtout un article de 2006 de Leonard Shustek, le directeur du Musée de l'histoire de l'ordinateur à Mountain View. « *Il y avait toutes les bonnes idées ! Il expliquait le besoin de préserver le code source. Mais ils ne l'ont pas fait* » regrette le chercheur. Il estime tout de même le moment propice, y compris d'un point de vue technologique.




**Le noyau Linux sur Software Heritage**

## Un travail de longue haleine, étape par étape

Pour l'instant, Software Heritage s'est concentré sur le code open source, facilement accessible. Comprendre celui disponible sur de grandes plateformes ouvertes comme GitHub, la distribution Debian ou Sourceforge. Ces services sont indexés régulièrement via un crawling automatique.

« *Ce sont paradoxalement les codes les plus en danger, comme l'a montré la fermeture de Google Code. En agissant vite, on peut faire quelque chose, sinon on perd tout* » explique Roberto Di Cosmo. Disparu en 2016, le service de Google a été rapidement contacté par Software Heritage, qui a obtenu une copie de la base. Grâce à l'intervention de Vinton Cerf, l'un des fondateurs d'Internet, employés par le groupe depuis une décennie. Gitorious, mort en 2015, a aussi mobilisé l'équipe, qui a obtenu des copies par des tiers. « *Ça a pris un temps infini.* ».

Pour le chercheur, la difficulté à collecter du code repose sur deux critères : techniques et juridiques. Les combinaisons forment un quadrant. Certains seront simples à obtenir des deux points de vue, quand d'autres poseront des problèmes techniques ou légaux, voire les deux.

« *Il y a des logiciels pour lesquels nous devons passer par un avocat, des logiciels propriétaires, avec des licences qui ne permettent pas de les réutiliser. D'autres sont perdus. Techniquement, ce n'est pas évident. Des chercheurs ont pu garder des copies ici ou là, sur des bandes magnétiques ou allez savoir où. Il faut donc mettre en place une collecte distribuée (crowdsourcing) du matériel* » détaille notre interlocuteur. Les pires cas sont donc les logiciels propriétaires anciens.

Pour autant, « *même si c'est ambitieux, sur un demi-siècle* [de développement informatique], *c'est peut-être jouable* », espère Di Cosmo. D'autant que le code informatique est bien plus nombreux et ouvert aujourd'hui, facilitant le travail sur la grande majorité du corpus.

## 200 To de données en base

L'archive actuelle pèse 200 To, en plus de données historiques sur les projets. Elle est hébergée via trois copies, deux internes au projet et une autre sur le cloud Azure de Microsoft, l'un des principaux sponsors de Software Heritage. Elle discute également d'un partenariat avec l'Internet Archive, mais les négociations patineraient.

La bibliothèque est aidée par les techniques actuelles, comme **l'arbre de Merkle**, qui permet de dédupliquer les fichiers identiques d'un projet à l'autre. Un avantage par rapport à l'Internet Archive et son historique web lancé en 1996, qui conserve une copie de chaque page, image ou vidéo pour chaque version.

« *L'idée était de ne pas avoir une seule copie, ni de l'héberger sur une seule technologie* », lance le responsable du projet. « *Mais ça ne suffit pas. Imaginez que je deviens fou et que je brûle ces archives. À terme, on compte le déployer sur un réseau de miroirs mondial indépendant.* »

Software Heritage ne se pose d'ailleurs pas de limite sur la récupération. Le code source de malwares, par exemple, sera traité comme tout le reste. Aucun tri n'est effectué. « *On peut le faire, la taille ne va pas exploser* » pense Di Cosmo, qui rappelle qu'un projet peu connu aujourd'hui pourrait devenir essentiel demain, justifiant son archivage. « *En 1995, les concepteurs de PHP ne se doutaient pas qu'il deviendrait un langage majeur.* »

## Les métadonnées, un gros défi

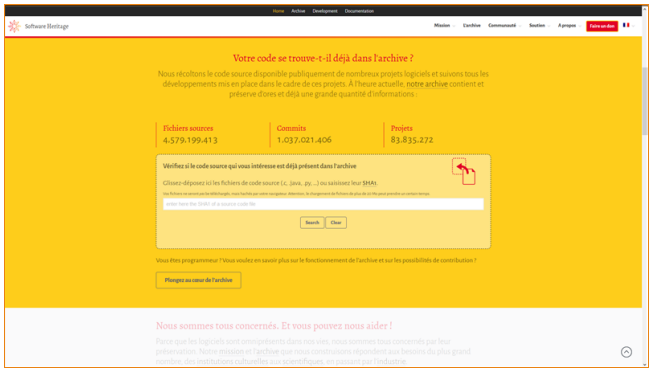
Archiver tout ce code est important, mais encore faut-il lui donner du sens. Le travail sur les métadonnées est encore difficile pour Software Heritage, qui en fait un de ses objectifs de long terme. Noms des projets, auteurs, historiques des contributions, relations entre projets, langage... Ce ne sont que quelques aspects que l'équipe doit traiter. Et elle n'est pas aidée par l'existant.

« *Ce suivi est intrinsèque au projet. On veut vraiment une traçabilité absolue de tout ce qu'on contient. Monter d'un niveau* [celui des projets] *est complexe. On peut se demander si tel projet sur Gitorious correspond bien à tel autre sur Google Code, donc quel est le vrai, qui est derrière, etc. C'est évidemment dans notre radar, mais on ne sait pas le faire* » assure encore Di Cosmo.

Problème : l'équipe a répertorié 45 ontologies différentes, c'est-à-dire des manières d'organiser logiquement le code et les projets. « *C'est n'importe quoi ! Je ne peux pas appliquer 45 ontologies sur des milliards de fichiers !* » s'agace le chercheur. Software Heritage travaille donc avec des communautés tierces pour trouver des solutions intermédiaires, sans réinventer la roue.

Autre difficulté, l'identification des langages de programmation. Selon l'équipe, il n'existe aucun outil qui puisse reconnaître automatiquement les 8 700 langages répertoriés. « *Certains en détectent quelques centaines, les plus populaires. Il faut se débrouiller pour le reste.* »

Software Heritage compte aussi lier sa base à Wikipédia, Wikidata, d'autres communautés... Mais rien n'a dépassé le stade des premiers contacts pour le moment.




**Pour le moment, la plateforme propose de vérifier si un fichier est déjà en base**

## Des soutiens privés nombreux, dont certains étonnants

L'ouverture de l'archive, il y a quelques jours, correspond aussi à l'arrivée d'une dizaine de sponsors, dont les trois principaux sont Intel, Microsoft et la Société Générale. Le ticket d'entrée, fixé à 100 000 euros, permet donc au projet de compléter le financement d'Inria. L'institut a d'ailleurs eu un rôle essentiel dans la recherche de ces soutiens, assure Di Cosmo, qui espère que Google ira au-delà du financement minimum.

« *Par rapport à la tâche, c'est peu, mais c'est déjà un signe très fort qu'ils financent un projet de long terme, qui n'est pas immédiatement un produit ou une promesse de milliards* » positive le chercheur. Peu de groupes français ont mis la main à la poche pour le moment. « *Orange, Cap Gemini, Atos ou EDF, par exemple, ne se sont pas encore engagés. Ce n'est pas une surprise. En France, on ne veut pas prendre de risque.* »

## Un revirement pour Microsoft

L'arrivée de Microsoft a de quoi étonner. En 1998, Roberto Di Cosmo avait publié *Le hold-up planétaire : la face cachée de Microsoft*, tirant à boulets rouges sur le groupe de Redmond. Il y défendait le besoin sociétal du logiciel libre. En 2001, Linux était qualifié de cancer par l'entreprise, hégémonique. Autant dire que la collaboration n'avait rien d'évident.

Depuis, Microsoft affiche son amour du logiciel libre, intégrant notamment un sous-système Linux à Windows 10 et reprenant GitHub à grands frais, avec le soutien de la fondation Linux. Pour Di Cosmo, c'est un vrai revirement. « *On travaille avec eux depuis presque deux ans. J'en suis ravi* » lance-t-il.

« *Il y a 20 ans, Microsoft était mon ennemi numéro un. C'était le diable. Au début, j'étais méfiant ! Je suis allé les voir à Redmond, et je suis étonné de leur changement interne radical. Sur 60 000 ingénieurs, 10 000 ne font que du logiciel libre. Ils ont aussi interdiction de réécrire un bout de logiciel déjà disponible en libre...* » conte-t-il.

L'explication reste rationnelle, pour lui : « *À l'époque, ils étaient les maitres du monde. Il était très facile d'être dédaigneux, de voir comme des communistes révolutionnaires tous ceux qui sortaient de leur modèle économique. Depuis, le monde a changé. Même si Microsoft reste une énorme entreprise, elle s'interroge sur son futur après Windows ou Office. Le cloud est devenu essentiel. Ils se sont vite aperçus que limiter le cloud Azure à Windows ne marcherait pas* ».

## Qwant chargé de fournir un moteur de recherche

La société française Qwant, qui a monté un laboratoire avec Inria (voir notre analyse), fournira un moteur de recherche dédié à cette énorme base de code. Selon Tristan Nitot, responsable de l'open source chez Qwant, ce projet s'intègre bien au laboratoire en question. Ce qui veut donc dire qu'il durera au mieux trois à quatre ans.

La société ne signe pas de chèque, mais mobilise une partie de son équipe sur ce moteur (sans plus de détails). « *Software Heritage est un travail pour les générations futures. Il faut le faire maintenant* » estime Tristan Nitot, pour qui le logiciel libre est l'une des principales aides aux jeunes pousses, fournissant de nombreuses briques clés en main.

Il se réfère d'ailleurs au hype cycle de Gartner, où l'attention vers une tendance nouvelle atteint un pic, avant de plonger... et de remonter pour atteindre une utilisation quotidienne, qui ne fait plus parler. « *Le libre est moins hype qu'avant, mais de plus en plus utilisé. Nous n'en sommes pas encore à une vraie compréhension de l'importance des logiciels libres pour une société libre. Les développeurs ont un pouvoir grandissant, il faut leur demander des comptes* » pense le fondateur de Mozilla Europe. Alors que Qwant peine encore à libérer son propre code.

## Une fondation et des soutiens divers

À long terme, Software Heritage compte devenir une fondation, pour pérenniser la structure et le financement. « *Un projet de recherche habituel n'ira pas. Il durera à peine trois, quatre ans. Une start-up ? Pourquoi pas, on récupère facilement des dizaines de millions en faisant croire qu'on fera du machine learning sur le code... Mais que se passe-t-il dans quatre ans ? On n'en sait rien* » résume Roberto Di Cosmo.

L'équipe compte donc multiplier les partenaires et les cibles de son infrastructure, pour diluer le risque. La direction de la fondation (indépendante) doit aussi éviter la centralisation.

Pour la suite, Software Heritage mise donc sur des fonctionnalités à valeur ajoutée, comme le moteur de recherche pour justifier son intérêt. Censé être capable de travailler sur un très grand graphe et sur l'historique, il ne faudrait pas forcément l'attendre l'an prochain.

Les cartons contiennent aussi des connexions avec des articles scientifiques parlant des logiciels, des débouchés spécifiques aux industriels... « *pour transformer* [l'archive] *en une sorte de norme pour l'utilisation de code source ouvert* ».

« *C'est sûrement le moment le plus fascinant et excitant de tout ce que j'ai fait. C'est une mission majeure. C'est la première fois qu'on me dit systématiquement que c'est super, sans jamais demander à quoi cela sert* » s'enthousiasme Roberto Di Cosmo. Il rappelle d'ailleurs que le projet est ouvert aux contributions : code, soutien financier, dons... « *Toutes les portes sont ouvertes !* »