le 5 novembre 2021 seulement

Te quedan **9** artículos gratis este mes

SUSCRÍBETE

Transformación Digital

PROGRAMACIÓN >

Software Heritage: una Biblioteca de Alejandría del código fuente con 11.000 millones de ficheros

El científico Roberto di Cosmo dirige un archivo que aspira a reunir las instrucciones de funcionamiento de todos los programas informáticos del mundo y ha conseguido que la Unesco declare al 'software' patrimonio cultural de la humanidad

MANUEL G. PASCUAL

Madrid - 05 NOV 2021-02:19 UTC

Una pantalla con código fuente.
ZOKARA (GETTY IMAGES)

Roberto di Cosmo tiene una obsesión: quiere reunir en un mismo lugar todo el código fuente del mundo. Esa especie de Biblioteca de Alejandría de la programación no puede tener ánimo de lucro y debe ser accesible para cualquiera, desde investigadores hasta empresas privadas o particulares. Que todos, o quienes quieran, conozcan la arquitectura de las aplicaciones informáticas que usamos ayudará a entenderlas y a mejorarlas. A generar más conocimiento, a la prosperidad de la sociedad.

Este año se cumplen cinco desde que el sueño de este científico italiano afincado en París empezó a convertirse en realidad. Gracias a su empeño personal, la Software Heritage Inititative vio la luz en verano de 2016 en la sede del centro de investigación INRIA, ubicado en la capital francea. Desde entonces ha recopilado más de 11.000 millones de ficheros únicos de más de 160 millones de repositorios. Todo ese código cabe en un petabyte (un millón de terabytes, que a su vez son 1.000 gigabytes), el equivalente a los datos que amasaría el Telescopio espacial Hubble

durante 455 años. La copia maestra de ese superarchivo la tiene Software Heritage, aunque hay otras dos en los servidores en la nube de Microsoft (Azure) y de Amazon (AWS).

Di Cosmo y su equipo lograron que la Unesco declarara en 2017 al *software* patrimonio cultural de la humanidad, que debe preservarse como la música o la literatura. Y a eso se dedica Software Heritage, a cuya financiación contribuyen instituciones públicas como el Ministerio de Innovación de Francia y varias universidades, pero también bancos como Société Générale o empresas como Microsoft, Google, Intel o Huawei.

La entonces directora general de la Unesco, Irina Bokova (con un pañuelo), visita en 2017 junto al también entonces presidente de Francia, François Hollande, la sede del centro de investigación INRIA tras firmar ambas instituciones un acuerdo de colaboración para preservar y compartir el 'software' como un patrimonio cultural más.
ALAIN JOCARD (AFP)

“Lo que hacemos es el equivalente a crear una especie de Google del código”, comenta Di Cosmo en su perfecto castellano con acento argentino, gentileza de su mujer. Está de visita en Madrid para participar en un congreso sobre ciencia abierta celebrado en la Universidad Politécnica. La institución que fundó y dirige tiene mucho que decir al respecto. “Hace falta construir una infraestructura que permita fácilmente almacenar, referenciar, difundir y describir de forma accesible para todos el código fuente que se usa en la investigación”, subraya. La exitosa

colaboración de la comunidad científica para desarrollar la vacuna de la covid es un argumento de peso a favor de esta histórica reivindicación.

El código fuente son unas líneas de texto escritas en algún lenguaje de programación que permiten que se ejecuten los programas informáticos. Muchas compañías y desarrolladores guardan celosamente esos códigos: viven de venderlos o de desarrollar productos a partir de ellos. Pero también hay quien publica sus creaciones para que los demás puedan aprovecharlas. La cultura del *software* libre, que tiene su origen en los años ochenta e impulsó Richard Stallman, promueve esa visión de la programación: la transparencia del código fuente de los programas, compartir con la comunidad los desarrollos propios para que otros puedan perfeccionarlos o tomarlos como punto de partida de proyectos mayores.

El triunfo del 'software' libre

"De algún modo, el *software* libre ha ganado. Se calcula que en 2017 entre el 80% y el 90% del código de las nuevas aplicaciones fue reutilizado de otro que ya existía", apunta. "Las grandes empresas como Microsoft, que hace unos años ni siquiera usaban la palabra, ahora recurren masivamente el código abierto". Este cambio de tercio se debe a que el *software* se ha vuelto tan complejo que nadie, ninguna empresa ni ningún país, es capaz de escribirlo todo por sí mismo de cero: lo más eficiente es cortar y pegar partes de código que ya se sabe que funcionan y centrar los esfuerzos en las nuevas funcionalidades.

Que sea libre no quiere decir que no contribuya a mover la economía. Según estimaciones de la Comisión Europea, las empresas europeas invirtieron en 2018 unos 1.000 millones de euros en *software* de código abierto, lo que tuvo un impacto sobre el PNB europeo de entre 65.000 y 95.000 millones.

Roberto di Cosmo en una foto de archivo.
INRIA

Pese a su auge, su existencia no debe darse por sentada. “En 2015 cerró Google Code, el repositorio de código auspiciado por la multinacional estadounidense, poniendo en peligro 700.000 proyectos. Gitorius, otro de los sitios más populares en el mundillo, fue comprado por GitLab, que optó por cerrarlo, lo que afectó a 120.000 proyectos. Hace unos meses, Bitbucket decidió modificar un aspecto técnico y borró 250.000 proyectos. Salvar todo eso es complicado”, explica el informático.

Software Heritage recopila el material de su gran biblioteca virtual por tres vías. “Nosotros vamos a buscar todo el código fuente en todas las plataformas que conocemos, con la dificultad de que cada una habla una lengua diferente técnicamente. Así conseguimos la gran mayoría de datos”, detalla. “Pero también abrimos otras dos puertas: la posibilidad de que cualquiera indique una web con código fuente, de modo que nosotros la recuperemos automáticamente, y la colaboración con asociaciones científicas”.

Copias espejo en cada país

Di Cosmo y sus colegas optaron desde el principio por mantener varias copias de su archivo universal de código fuente. Además de la suya y de las que tienen en la nube, la Software Heritage Initiative está desarrollando un sistema de copias espejo (*disk mirroring*). Se trata de copias del archivo, pero bajo control administrativo y técnico de otras entidades. La primera estará en Italia, en la Agencia Nacional por las Nuevas Tecnologías y la Energía (ENEA). "Ellos tendrán nuestros datos, pero nosotros no podremos escribir en su archivo. Por tanto, si viene un hacker y nos borra todo, no podrá hacer lo mismo con esa copia: tendrá que hackearla también", explica.

Lo normal, opina el científico, es que los Gobiernos de los países no tarden en darse cuenta de que les interesa apoyar la iniciativa y tener su propia copia espejo. "Hoy en día el *software* es fundamental para que todo siga funcionando. Nosotros creamos una copia de todo el que podemos recoger, y los países que quieran van a tener su copia espejo. De ese modo, no pierdes los datos y además te aseguras de que, pase lo que pase, nadie te va a poder cortar el acceso al programa que usas. Así que, paradójicamente, esta iniciativa de colaboración global también responde a una necesidad de autonomía estratégica de cada país", concluye Di Cosmo.

El refugio del código fuente que dirige cuenta con un equipo mínimo. "Necesitamos entre 30 y 50 personas a tiempo completo y un presupuesto anual de entre cinco y 10 millones. "Si lo comparas con el costo de un telescopio, un barco oceanográfico o un acelerador de partículas, eso no es nada. Pero es cierto que al ser virtual nuestro trabajo es menos tangible que otros, y a la Administración se le da mejor en general financiar máquinas que gente".

Puedes seguir a EL PAÍS TECNOLOGÍA en Facebook y Twitter o apuntarte aquí para recibir nuestra newsletter semanal.

SOBRE LA FIRMA

Manuel G. Pascual

Es redactor de la sección de Tecnología. Sigue la actualidad de las grandes tecnológicas y las repercusiones de la era digital en la privacidad de los ciudadanos. Antes de incorporarse a EL PAÍS trabajó en Cinco Días y Retina.

Comentarios Normas ›

Más información