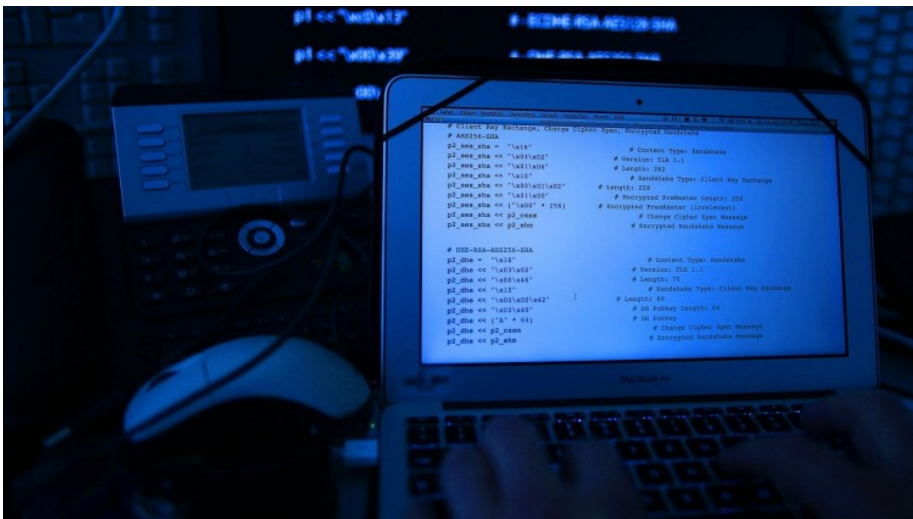


Archiv-Software Heritage

Der Quellcode des Planeten

"Die Software von heute müssen wir erhalten, damit die Software von morgen noch besser wird." Das sagt Roberto Di Cosmo vom Französischen Nationalen Institut für Informatik. Dort archivieren Wissenschaftler Source Code für die Nachwelt.

Roberto di Cosmo im Gespräch mit Manfred Kloiber



Computerbildschirm mit einem Quellcode: 23 Millionen Software-Projekte wurden Frankreich bereits archiviert. (picture-alliance / dpa / Oliver Berg)

Manfred Kloiber: Die Web-Protokolle sind alle keine geheime Kommandosache, sondern ganz öffentlich diskutierte Standards, deren Schwächen jetzt sicherlich von der Internet Engineering Task Force noch einmal überprüft werden. Das ist ja die Stärke offener Software-Produkte – jeder kann in ihnen lesen und sie auf Mängel untersuchen.

Jeder kann sie aber auch lesen, um etwas zu verstehen. Zum Beispiel, wie man Probleme löst oder sie noch besser lösen könnte. Und so verraten offene Programme noch viel mehr. Zum Beispiel, mit welchen Problemen sich Programmierer zu welcher Zeit überhaupt herumgeschlagen haben.

Forscher des französischen Nationalen Informatik-Forschungszentrum

MEHR ZUM THEMA

Geistiges Eigentum in der Videospiele-Kultur

[\[http://www.deutschlandradiokultur.de/geistiges-eigentum-in-der-videospiele-kultur-diebstahl.976.de.html\]](http://www.deutschlandradiokultur.de/geistiges-eigentum-in-der-videospiele-kultur-diebstahl.976.de.html)

dram:article_id=357158] Diebstahl, Inspiration und das juristische Bauchgefühl

Coden wie Harry Potter

[\[http://www.deutschlandfunk.de/coden-wie-harry-potter-der-einzigartige-stil-eines-software.684.de.html\]](http://www.deutschlandfunk.de/coden-wie-harry-potter-der-einzigartige-stil-eines-software.684.de.html)

dram:article_id=354136] Der einzigartige Stil eines Software-Entwicklers

Open-Source-Software

[\http://www.deutschlandfunk.de/open-source-software-wie-sinnvoll-sind-

Inria sehen deshalb in Programmen auch so etwas wie zeitgeschichtliche Dokumente. Deshalb haben sie das Projekt Software Heritage gegründet. Projektleiter Roberto Di Cosmo hat es mir erklärt.

Roberto di Cosmo: Software Heritage ist eine Initiative, die wir vor einem Jahr gestartet haben, um den ganzen Quellcode zu sammeln, der auf dem Planeten verfügbar ist. Es gibt so gut wie gar nichts mehr, was wir ohne Computer machen. Und Computer laufen eben mit Programmen. Sie enthalten sehr wichtige Informationen, technischer Art, wissenschaftlicher Art oder kultureller Art. Und bislang hat niemand entschieden, dies für die Nachwelt zu bewahren, zu strukturieren und leicht verfügbar zu machen.

Kloiber: Kann denn jeder einfach so Software zu Ihrem Projekt beisteuern?

Di Cosmo: Im Moment sind wir dabei, selbst Open-Source Software einzusammeln. Für diejenigen, die uns eine wichtige Software zur Verfügung stellen wollen gibt es auf unserer Webseite die Möglichkeit, dies anzumelden und wir sammeln dann auch diese Software ein, sobald das uns möglich ist.

Kloiber: Und in welchem Zustand muss sich die Software befinden? Muss das alles fein säuberlich dokumentiert sein oder nehmen sie auch so etwas wie Spaghetti-Code an?

Di Cosmo: Gute Frage. Eigentlich ist das ja ein Problem der Auswahl. Es gibt ja so viel Software. Sollen wir die Zeit für die Prüfung aufwenden und dann entscheiden, diese Software ist gut dokumentiert, die behalten wir und die andere, schlecht geschrieben und schlecht dokumentiert, die nehmen wir nicht?

Für Software, besonders für moderne Programme, die sehr schnell und interaktiv entwickelt wurden, ist es äußerst schwierig zu entscheiden, ob diese Software in zehn Jahren vielleicht wichtig sein wird oder nicht. Glücklicherweise ist die Datenmenge der verfügbaren Software weltweit viel kleiner als die der ganzen Videos und Bilder, die zum Beispiel für Youtube oder Flickr produziert werden. Also können wir alles speichern und jeden Code annehmen, nicht nur gut dokumentierten, sondern alles.

Kloiber: Sie nehmen ja nur Source Code. Viele Firmen veröffentlichen aber gar keinen Source Code. Führt das nicht zu einer selektiven Auswahl in Ihrer Sammlung?

Di Cosmo: In der Tat, das ist unsere Bedingung, dass wir eine Kopie des Source Codes bekommen. Es geht ja darum, dass man in dem Programm

[ausschlussklauseln.684.de.html?](http://www.deutschlandfunk.de/ausschlussklauseln.684.de.html?dram:article_id=350813)

[dram:article_id=350813](http://www.deutschlandfunk.de/ausschlussklauseln.684.de.html?dram:article_id=350813)] Wie sinnvoll sind Ausschlussklauseln?

[Internetkommunikation](http://www.deutschlandfunk.de/internetkommunikation.676.de.html?dram:article_id=348559)

[\[http://www.deutschlandfunk.de/internetkommunikation.676.de.html?dram:article_id=348559\]](http://www.deutschlandfunk.de/internetkommunikation.676.de.html?dram:article_id=348559) E-Mail, aber sicher und verschlüsselt

[dram:article_id=348559](http://www.deutschlandfunk.de/internetkommunikation.676.de.html?dram:article_id=348559)] E-Mail, aber sicher und verschlüsselt

[Proprietäre Software](http://www.deutschlandfunk.de/proprietare-software-die-gleichgueltigkeit.684.de.html?dram:article_id=334226)

[\[http://www.deutschlandfunk.de/proprietare-software-die-gleichgueltigkeit.684.de.html?dram:article_id=334226\]](http://www.deutschlandfunk.de/proprietare-software-die-gleichgueltigkeit.684.de.html?dram:article_id=334226) Die Gleichgültigkeit

[CodePhage](http://www.deutschlandfunk.de/codephage-viraler-code-repariert-fehlerhafte-software.676.de.html?dram:article_id=326701)

[\[http://www.deutschlandfunk.de/codephage-viraler-code-repariert-fehlerhafte-software.676.de.html?dram:article_id=326701\]](http://www.deutschlandfunk.de/codephage-viraler-code-repariert-fehlerhafte-software.676.de.html?dram:article_id=326701) Viraler Code repariert fehlerhafte Software

[dram:article_id=326701](http://www.deutschlandfunk.de/codephage-viraler-code-repariert-fehlerhafte-software.676.de.html?dram:article_id=326701)] Viraler Code repariert fehlerhafte Software

[Mobiles Betriebssystem](http://www.deutschlandfunk.de/mobiles-betriebssystem-google-und-das-android-monopol.684.de.html?dram:article_id=326431)

[\[http://www.deutschlandfunk.de/mobiles-betriebssystem-google-und-das-android-monopol.684.de.html?dram:article_id=326431\]](http://www.deutschlandfunk.de/mobiles-betriebssystem-google-und-das-android-monopol.684.de.html?dram:article_id=326431) Google und das Android-Monopol

[dram:article_id=326431](http://www.deutschlandfunk.de/mobiles-betriebssystem-google-und-das-android-monopol.684.de.html?dram:article_id=326431)] Google und das Android-Monopol

lesen kann. Das geht ja mit der ausführbaren Version nicht. es gibt eine große Zahl von Unternehmen, die Ihren Code nicht freigeben. Aber die Dinge ändern sich. Es gibt eine unglaubliche Menge an offener und freier Software, das ist der größte Teil in der Branche. Aber man sollte auch nicht vergessen: Irgendwann erlischt das Copyright. Wir hoffen, dass wir irgendwann auch von Firmen den Code bekommen, den sie jetzt noch nicht veröffentlichen wollen.

Software mit Methoden der künstlichen Intelligenz optimieren

Kloiber: Bei vielen Programme ist es ja oft so, dass sie stark auf vorgefertigte Software-Module, die sogenannten DLLs zurückgreifen, Bibliotheken. Müssten sie nicht zusätzlich zum eigenen Programm auch die DLLs und die Entwicklungsumgebung archivieren, um vollständig zu sein?

Di Cosmo: Absolut, wenn sie wirklich in der Lage sein wollen, Programme wieder lauffähig herstellen zu wollen, dann müssen sie auf alle Bibliotheken und auf die Entwicklungsumgebung zugreifen können und auch auf die Hardware, die dazu gehört. Das ist natürlich das große Fernziel. Das ist aber die Vergangenheit.

Viel wichtiger sind die Gegenwart und die Zukunft. Die Software von heute müssen wir erhalten, damit die Software von morgen noch besser wird. Durch die Sammlung aller verfügbaren Programme machen wir es möglich, die Software zu analysieren, die Fehler herauszufinden. Wir können dann auch Trends herausfinden, zum Beispiel welche Bibliotheken besonders oft genutzt werden, welche beliebter werden, welche verschwinden. Und mit den Methoden der künstlichen Intelligenz können wir Software von heute optimieren und für bessere Software von morgen sorgen.

Kloiber: Letzte Frage: Wie groß ist denn ihre Sammlung schon?

Di Cosmo: Also, wir haben im August letzten Jahres begonnen und wir haben mittlerweile 23 Millionen Software-Projekte aufgenommen. Das addiert sich auf ungefähr drei Milliarden Dateien mit Source Code. Und zwar ohne Doppler. Wir glauben, das ist die größte Sammlung von Source Code überhaupt auf diesem Planeten. Die große Herausforderung für uns ist jetzt, alle Plattformen zu finden, auf denen Source Code veröffentlicht wird. Ich hab natürlich keine Glaskugel, aber ich schätze wir haben zur Zeit etwa 15 bis 20 Prozent der verfügbaren Programme abgedeckt. Das kostet einfach Zeit, diese Sammlung vollständig zu machen.

Kloiber: Das war Roberto Di Cosmo vom Französischen Nationalen Institut für Informatik INRIA über die Software Heritage.

