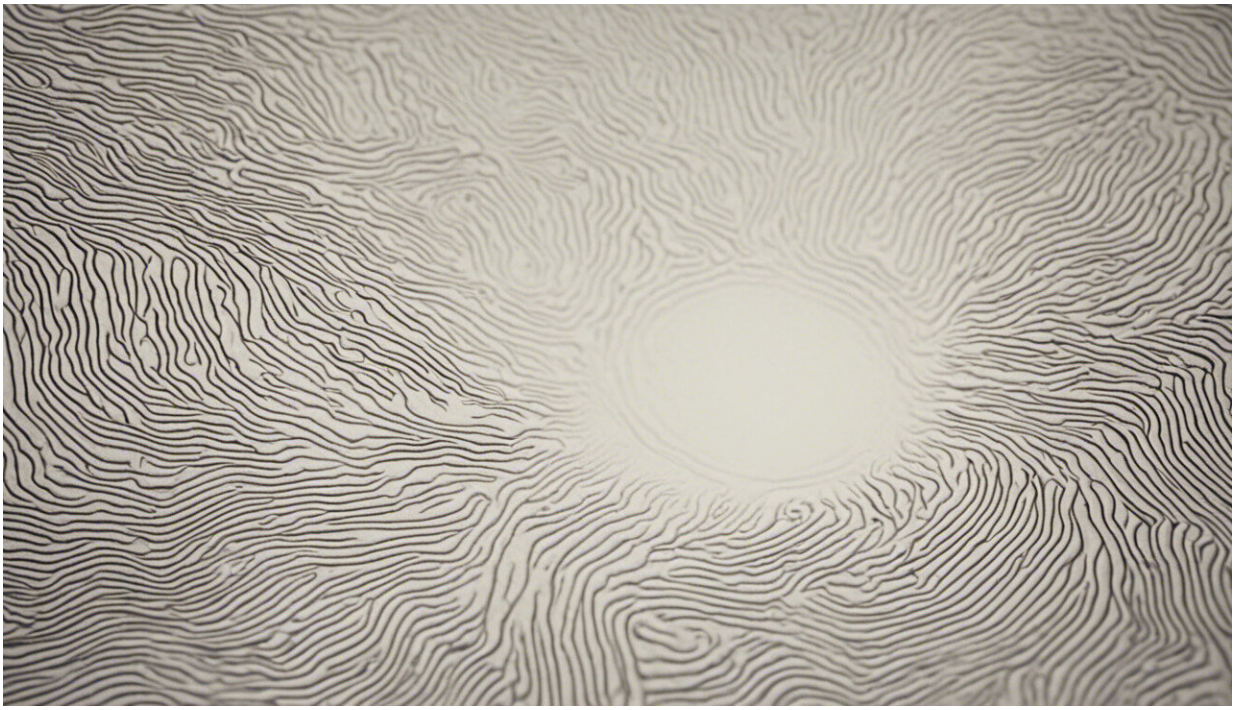
Database of software 'fingerprints' expands to include mobile apps

December 16 2016



Credit: AI-generated image (disclaimer)

A group of computer scientists at the National Institute of Standards and Technology (NIST) has been working for more than 15 years on an impossible task: to maintain an up-to-date archive of the world's software. Because the amount of software in circulation keeps growing, they will never enjoy the satisfaction of completing their assignment.

But they have succeeded in creating the largest publicly known collection of its kind in the world.

Called the National Software Reference Library (NSRL), the collection is about to get a whole lot larger. On December 15, 2016, the NSRL will expand to include its first batch of 23,000 mobile apps for Android and iOS. Another 200,000 are expected to be added during the coming year.

The NSRL is a critical tool used in law enforcement and national security investigations. Every file in the NSRL is run through a computational procedure that generates a unique digital fingerprint for that file, expressed as a string of 40 letters and numbers. NIST publishes those fingerprints in a Reference Data Set (RDS) that is updated quarterly and freely available to the public.

Software applications often include hundreds or even thousands of files—think of all the graphics files and templates that are placed on your computer when you install a word processing application. The RDS contains fingerprints for all the files in the NSRL, more than 50 million so far.

The NSRL is jointly funded by NIST and the Department of Homeland Security Science and Technology Directorate's Cyber Forensics Project , and the addition of mobile apps was done in collaboration with their Mobile App Security R&D project .

An Essential Tool for Forensic Investigators

"When we seize a computer or hard drive as part of an investigation, we need to eliminate files that are irrelevant to our investigation and focus on those that might contain evidence," said Sam Brothers, a digital protection specialist with U.S. Customs and Border Protection. In homeland security investigations, time is of the essence, so Brothers and

his colleagues use the RDS to filter out known files. "It allows us to separate the wheat from the chaff very quickly."

Occasionally, investigators use the RDS not to exclude known files, but to find them, even if the filenames have been altered. For example, after Malaysia Airlines flight MH370 disappeared somewhere over the Pacific in March 2014, the FBI called NIST. "They wanted every hash of every file associated with every flight simulator we had," recounted Doug White, the NIST computer scientist who runs the NSRL. "All the maps. All the routes. They wanted every flight path the pilot might have practiced on, so they could figure out where he might have gone."

To assist the FBI in their investigation, White and his colleagues updated the NSRL with more than 120,000 flight map-related files.

Much of the software in the NSRL is donated by the companies that publish it. Another segment of the collection is composed of free software. After that, White decides which titles to purchase with limited funds.

"Our goal is to help investigators, so we prioritize the software they are most likely to encounter in the field," White said. "We also focus on what we consider dual-use software—things that can be used for good or bad," including keystroke loggers and network monitoring tools.

A Unique National Resource

The NSRL is unique not only because of its size, but also because the original files are kept under evidence-locker conditions. For software that was distributed physically, the original discs are kept under lock and key. For software that is distributed electronically, the original distribution files are archived on secure servers. That means that the original software can be retrieved, if necessary, as evidence in court or to

verify the provenance of a title.

Although NIST generally does not distribute the software in the NSRL—only the software fingerprints—researchers can come to NIST and use the NSRL to develop and test forensic and security tools. Researchers also use the NSRL to study how software evolves over time and to trace newly discovered security vulnerabilities back to their first appearance.

Beyond the world of law enforcement and national security, the NSRL also functions as a cultural repository that is used by historians and other scholars.

How large the NSRL will ultimately grow is anyone's guess. Perhaps centuries from now, NIST scientists will still be feeding it software. On the other hand, archaeologists might dig up its discs and set them spinning to unlock a record of an earlier civilization. But for now, the NSRL is a potent tool for forensic investigators, and with the addition of mobile apps, it just got a whole lot more powerful.

Provided by National Institute of Standards and Technology