Current Activities in the

November 7, 2006

NIST United States Department of Commerce
National Institute of Standards and Technology

Project Support

The NSRL project is supported by

- **NIST, Office of Law Enforcement Standards**
- **DoJ, National Institute of Justice, FBI**
- **DoD, DCCC**
- **DHS, ICE, USSS**
- **State & Local Law Enforcement**
- **Vendors**

Other federal agencies and industry organizations provide resources.

Overview

- What is the NSRL?
- How is it used?
- Beyond MD5 hashes
- Installation and Registry
- Unverifiable Metadata
- Known Disk Blocks
- Distributing Large Data Sets
- Storage and Reprocessing

What is the NSRL?

The NSRL is conceptually three objects:

- **A physical collection of software**
- **A database of meta-information**
- **A subset of the database,**
  the Reference Data Set (RDS)

Physical Collection

The collection is treated as case evidence.

- **Software is kept in a locked room with limited access.**

If metadata is questioned in court, it can be regenerated from original media.

The collection is 8,000+ applications, in over 35 languages, for many OSes.

Based on "popular" titles – most encountered, most pirated, most applicable at the time.

Database

The database contains over 55 million file signatures.

All of the metadata is stored to uniquely identify a file in a directory on a piece of media in an application.

The hashes are only from files on original media.

Archive-type files (CAB, ZIP, UU, TAR, ISO, DMG, dd) are hashed, then extracted and contents are hashed.

Database schema is available, database contents can be made available.

Reference Data Set

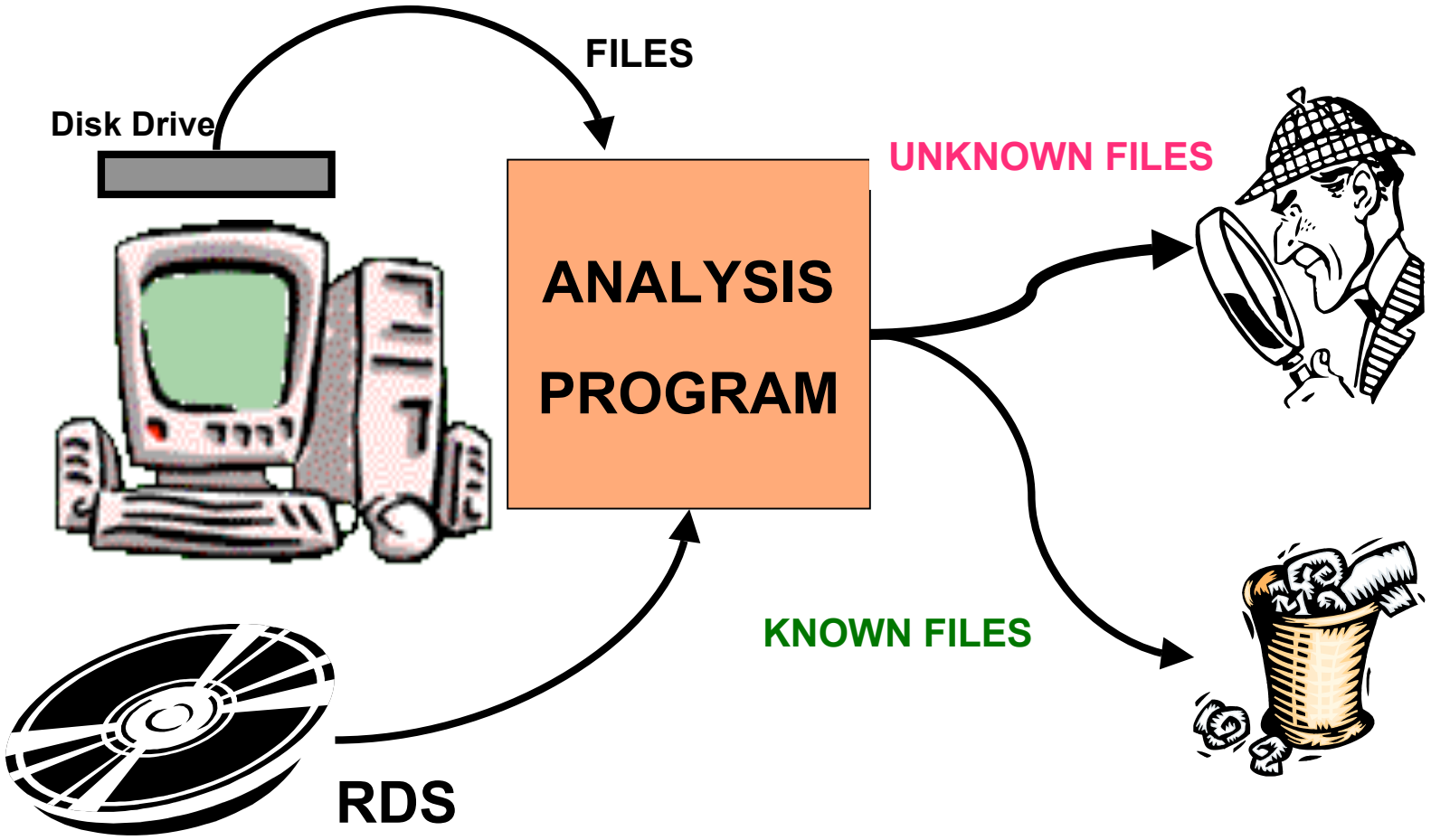
Version 2.14 was released September 2006.

4 CD set contains zipped flat text files according to public spec on website. Can be imported into popular forensic tools.

Contains 11,514,592 unique SHA-1/MD5 values.

Expect 250K-1M new unique values per quarterly release.

How the RDS is Used



Metadata Intent

The project sponsors were concerned with identification of known application files, to allow known files to be ignored, focusing investigation on user-generated data.

NIST does not assign “malicious” nor “notable” values to applications.

The NSRL does assign application categories, e.g. image manipulation, steganography, encryption. Original directory/path location is noted.

The NSRL metadata has been used to determine the “pedigree” of NARA systems. Can determine the upgrade path of a PC such as from NT3.5 to NT4 to W2K.

NSRL Impact

Referenced in 2001 seizure of bogus MS media in CA.

Referenced by Simpson Garfinkel in 2002 efforts with reclaimed disks.

Imported into EnCase, FTK, Ilook, Hashkeeper, Maresware.

Essential to FBI CART, copied for every field office.

Used by private organizations to eradicate P2P use.

Used by ISPs to track app sharing on servers.

Used by sysadmins to confirm valid OS file state.

Used by FDA in FL Botox case.

International use - UK NHTCU, EU JRC, etc.

Beyond MD5 Hashes

With respect to MD5 collision news:

- **The NSRL project does not see any fatal ramifications from the collision announcements.**
- We have not seen a "pre-image" attack; that is, the researchers did not identify a known file in the NSRL and attempt to generate a different file with a matching hash value.
- There are known MD5 collisions and weaknesses; the NSRL data provides an MD5 to SHA-1 mapping to facilitate the migration away from MD5.
- SHA-1 will be superceded in 2010 by FIPS 180-2, Secure Hash Standard (SHA-224, 256, 384,512). The NSRL will provide a SHA-1 to SHA-256 mapping.

SHA-256, Whirlpool, etc.

- Very easy for NIST to add algorithms to the NSRL hashing cluster code.
- Mappings between hash values will be maintained.
- NSRL will continue to collect “outdated” hash values, e.g. MD4.
- Willing to work with researchers to run algorithms against file corpus.

  http://www.nsrl.nist.gov/documents/yapc2004/index.html

Installation and Registry

WIndows REgistry Dataset (WIRED)

- NSRL software (OS, applications) installed on reference machines
- Capturing dynamic, contextual data
- Registry keys, values resulting from installation and persisting after “deletion”
- Available for comments - http://xsun.sdct.itl.nist.gov/~dwhite/WIRED/WIRED-060511.iso

Unverifiable Metadata

- NSRL software acquisition bottleneck
  - $2,500 per month
  - Hard to keep informed of new software
  - Priority app types may not be commercial
- Web crawling limitations
  - Does not fit “shelf escrow” model
  - Possible legal restrictions

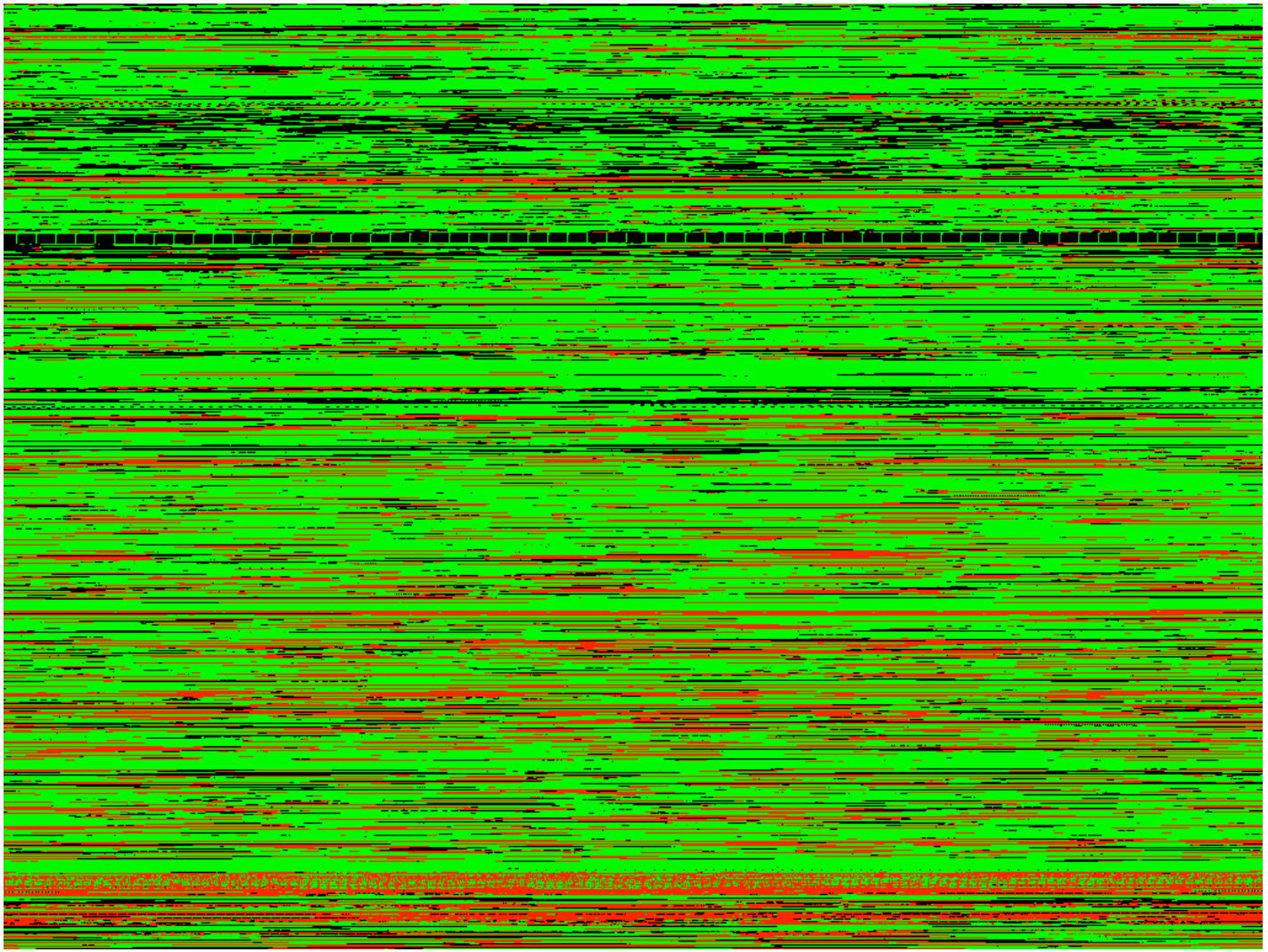
NSRL may collect and distribute this data as “of interest to the community”

Known File Blocks

- Hashes of 512 byte or 4096 byte blocks
- Cryptographic strength merged with statistical probability of identification
- Identify known data at a level independent of a file system
- Possibly used during evidence capture to reduce post-processing

Known - Unknown - Zero

2nd 512 MB in W2K NTFS VM



Distributing Large Data Sets

- In current format, 11 million unique hash values need a 4GB file
- The current 2TB of data on shelves could yield 1 billion unique block hash values

Bloom filters may provide a mechanism

- No false negatives, false positive rate can be adjusted to suit
- 100 million SHA-1s represented on a CD
- 1 billion SHA-1s represented on a DVD

Bloom Filter Distribution

- Stored as a bit stream
- Fast response to queries
- Can be used in conjunction with full RDS metadata
- Simple bitwise-or to “upgrade”
- Other researchers working this area

  http://xsun.sdct.itl.nist.gov/~dwhite/RDS/rds_2.13/bloom/

Storage and Reprocessing

- Currently, SHA-256 collection requires handling all shelved media
- NSRL will image the media to SAN
- Hashing cluster will access stored media images to apply algorithms to entire corpus

Storage and Reprocessing

Research into storage

- Currently dd images, with image metadata in database
- Advanced Forensic Format - Garfinkel
- Digital evidence bags - Turner
- Grid storage - UMCP/UCSD/NARA

Hash Processing Capability

NSRL runs on dedicated, isolated 100Mbit network.

Have 1Gb hubs, NICs in critical locations.

Windows shares limit us to 12 drives for reading media.

Current setup can process 15GB per hour, media to hashset.

Will use fiberchannel in new rack-based system.

Move to Linux/OS X Samba shares allows more read drives.

New hashing nodes will be 64-bit dual CPU blades, should quadruple throughput out of the box to 60GB/hr.

Easy to add input drives, hashing blades for growth.

Contacts

Doug White
www.nsrl.nist.gov
nsrl@nist.gov

Barbara Guttman
barbara.guttman@nist.gov

Sue Ballou, Office of Law Enforcement Standards
Rep. For State/Local Law Enforcement
susan.ballou@nist.gov

Current Activities in the National Software Reference Library

Doug White

nsrl@nist.gov www.nsrl.nist.gov

NIST United States Department of Commerce
National Institute of Standards and Technology