# National Software Reference Library

Douglas White
Information Technology Laboratory
July 2004

# Introduction

The National Software Reference Library is:

- A physical collection of over 5,000 software packages on secured shelves
- A database of file “fingerprints” (or “hashes”) and additional information to uniquely identify each file on the shelves
- A Reference Data Set (RDS) extracted from the database onto CD, used by law enforcement, investigators, researchers, others

# Use of the NSRL

- Eliminate as many known files as possible from the examination process using automated means
- Discover expected file name with unknown contents
- Identify origins of files
- Look for malicious files, e.g., hacker tools
- Identify duplicate files
- Provide rigorously verified data for forensic investigations

# How Did the NSRL Start?

Law Enforcement needed software hashes that could be used in investigations and in court.

- Source must be unbiased - NIST is a neutral organization
- Data produced must be of the highest quality
- Data must be traceable and repeatable
- There must be a repository of original software
- NIST provides an open rigorous process

# NSRL Software Collection

- Balance of most popular (encountered often) and most desired (pirated often)
  - Currently 32 languages, used internationally
- Software is purchased commercially
- Software is donated under non-use policy
- List of contents available on website

  www.nsrl.nist.gov

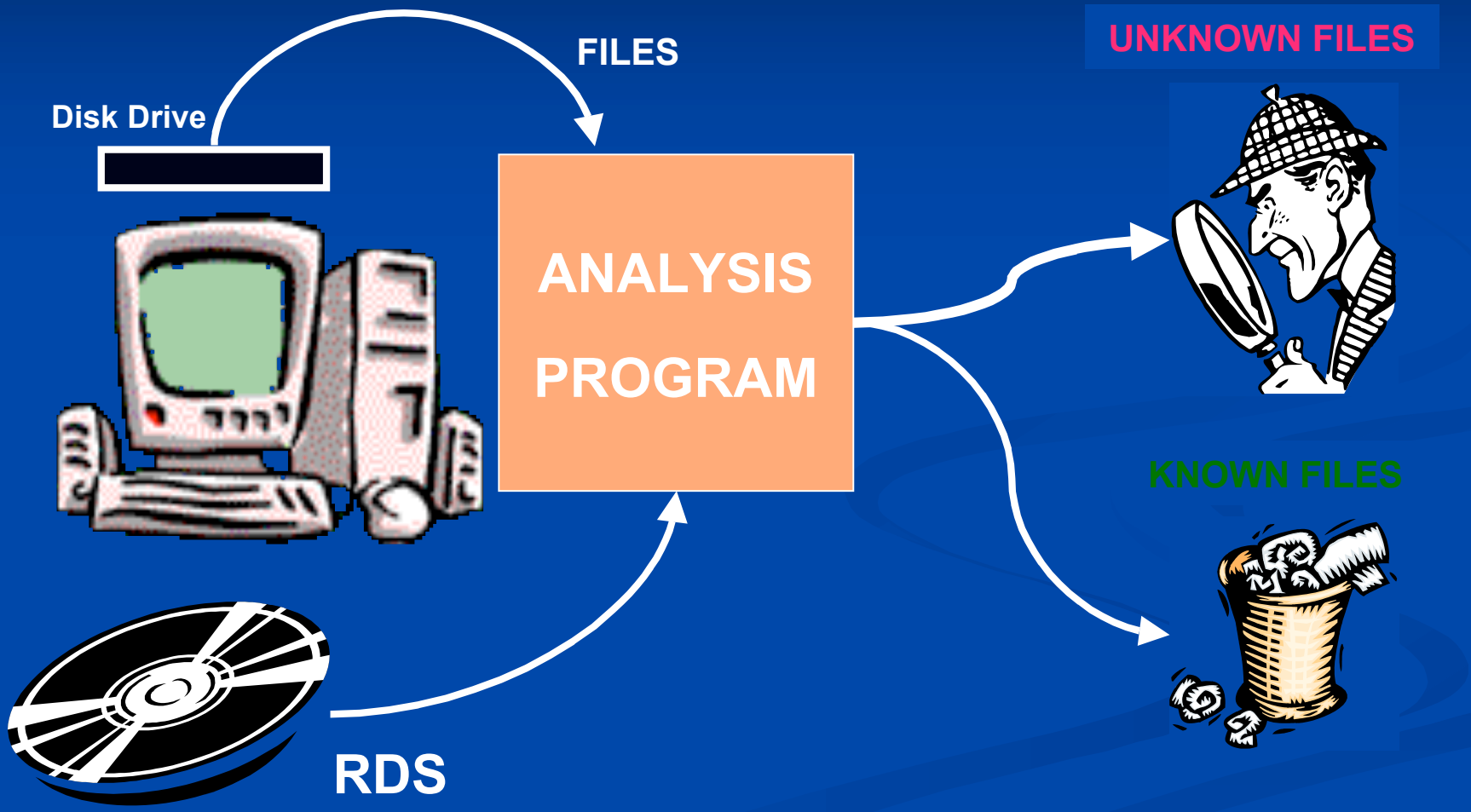
# NSRL Software Database

- Information to uniquely identify every file on every piece of media in every application
- Database schema is available on website
- 4,200 Bytes per application
- 750 Bytes per file
- Total database size now 20 GB for 5,000 applications with 31,900,000 files

# NSRL Reference Data Set

- The Reference Data Set (RDS) is a selection of information from the NSRL database
- Allows positive identification of manufacturer, product, operating system, version, file name from file "signature"
- Data format available for forensic tool developers
- Published quarterly, free redistribution
- Possible to publish critical data out of regular schedule; in February 2004 NSRL supplied 500,000 Arabic file signatures to FBI & DoD

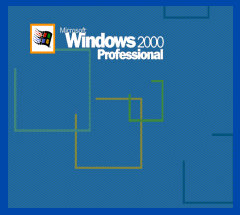
# RDS Field Use Concept



# RDS Field Use Example

You are looking for sensitive facility maps on a computer which is running Windows 2000.

Windows 2000 operating system software contains 5933 images which are known gifs, icons, jpeg files

e.g.,  

By using the RDS and an analysis program the investigator would not have to look at these files to complete his investigation.

# Hashes

- Like a person's fingerprint
- Uniquely identifies the file based on contents
- You can't create the file from the hash
- Primary hash value used is Secure Hash Algorithm (SHA-1) specified in FIPS 180-1, a 160-bit hashing algorithm
  - $10^{45}$ combinations of 160-bit values
- "Computationally infeasible" to find two different files less than $2^{64}$ bits in size producing the same SHA-1
  - $2^{64}$ bits is one million terabytes

# Hash Examples

| Filename | Bytes | SHA-1 |
|---|---|---|
| NT4\ALPHA\notepad.exe | 68368 | F1F284D5D757039DEC1C44A05AC148B9D204E467 |
| NT4\I386\notepad.exe | 45328 | 3C4E15A29014358C61548A981A4AC8573167BE37 |
| NT4\MIPS\notepad.exe | 66832 | 33309956E4DBBA665E86962308FE5E1378998E69 |
| NT4\PPC\notepad.exe | 68880 | 47BB7AF0E4DD565ED75DEB492D8C17B1BFD3FB23 |
| WINNT31.WKS\I386\notepad.exe | 57252 | 2E0849CF327709FC46B705EEAB5E57380F5B1F67 |
| WINNT31.SRV\I386\notepad.exe | 57252 | 2E0849CF327709FC46B705EEAB5E57380F5B1F67 |

# NSRL & National Archives and Records Administration

- Use hashing process on non-classified Presidential materials
- Identify application files
- Identify duplicate files
- Access to older installed software

# NSRL & Voting Systems Needs

- Determine that software used during elections is the *expected* software
  - Tested, certified version is definitively identifiable
  - Same during distribution, installation, setup, or use
  - "Chain of custody"
- Transparency
  - The NSRL methodology is in the public domain, available for inspection
  - Jurisdictions can share knowledge with each other

# EAC & NSRL

- Can verify that operating system file contents have not been modified
- Can verify that application file contents have not been modified
- Can verify that known static sections of files have not been modified
- At 866MHz, SHA-1 of 50MB takes ~5 sec. , MD5 of 50MB takes ~4 sec.

# Voting Research Issues

- Working with software companies to get access to software
- Distribution vs. installation hashes
- If there is any setup after the hashes are made, how do you know what changes are valid?
- Possible/practical to have on-location, time-of-certification hashing?
- Verification within time/ space/ security constraints

# Discussion

- Questions about the NSRL
- Discussion of the NSRL and Voting Systems

# Contact

Douglas White
Software Diagnostics and Conformance Testing
Information Technology Laboratory
Telephone: 301-975-4761
Email: nsrl@nist.gov
Web: www.nsrl.nist.gov