

# National Software Reference Library (NSRL) Project

Douglas White

Information Technology Laboratory

March 2003

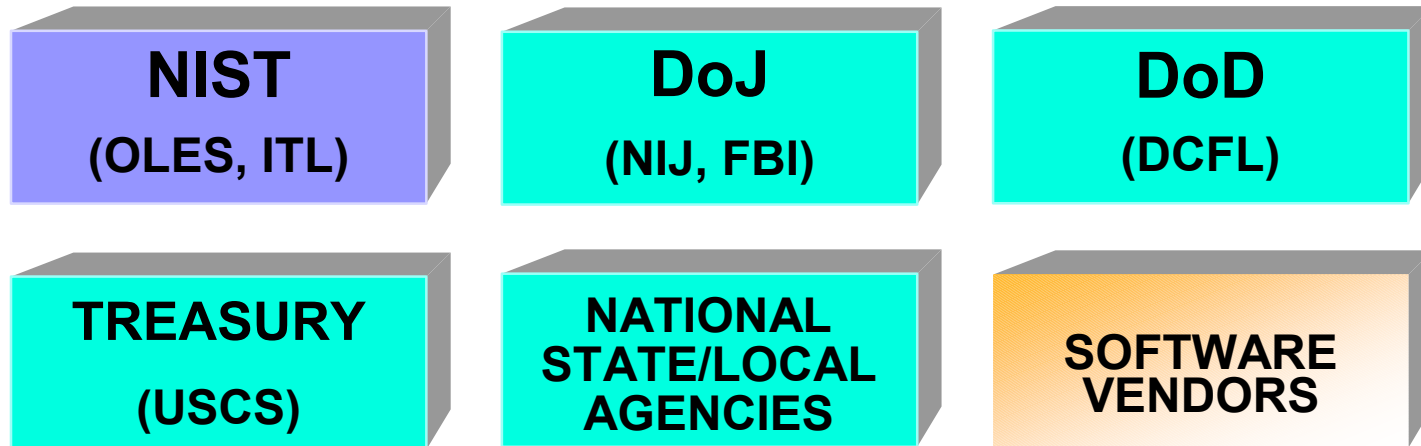
**NIST** United States Department of Commerce  
National Institute of Standards and Technology

# Introduction

The National Software Reference Library is:

- A physical collection of over 3,800 software packages on secured shelves
- A database of file “fingerprints” 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 and researchers

# Law Enforcement Origins



- Mission: assist Federal, state and local agencies
- NIST is a neutral organization in relation to vendors
- NIST provides an open rigorous process assuring data quality
- NIST NSRL will become an international software repository

# Addressing Industry Needs

- No unbiased organizations were involved in implementing investigative tools
- Law enforcement had no control over quality of data provided by available tools – data was market-driven
- Traceability - No repositories of original software available for reproducing data
- Each tool provided a limited set of capabilities

# Computer Forensics in ITL

- Located in Software Diagnostics and Conformance Testing (SDCT) Division
  - Includes development of specifications and conformance tests for use by agencies and industry
  - Work is funded by Federal agencies and NIST internal funds
- Homeland Security support of agencies investigating terrorist activities

# Computer Forensics in ITL

- Goals of Computer Forensics Projects
  - Introduce new automated processes into the computer forensics investigative process
  - Provide stable foundation built on scientific rigor to support the introduction of evidence and expert testimony in court

# NSRL Software Collection

- Media in format as available to the public
- Consumer products available in stores
- Developer products available as vendor services
- Malicious software
- “Cracked” software



# NSRL Software Collection

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

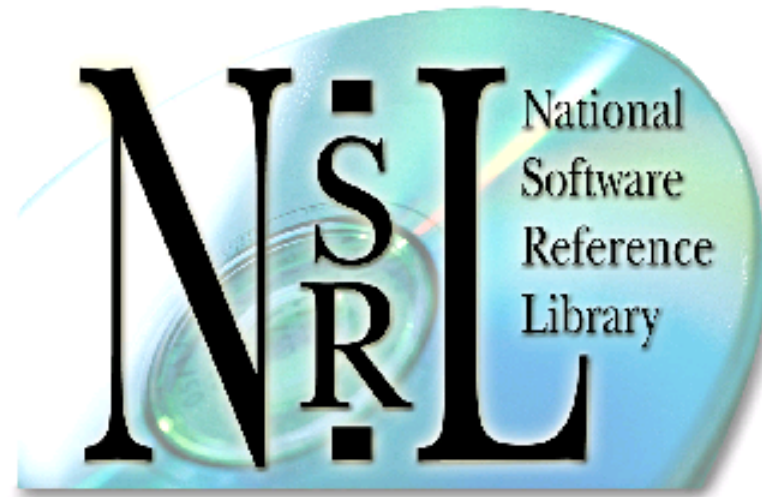
[www.nsrl.nist.gov](http://www.nsrl.nist.gov)



# 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 9 GB for 3,800 applications with 13,400,000 files

NIST Special Database #28



**Reference Data Set**

**Version 1.5 03/03/2003**

**NIST**

# 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

# Use of the RDS

- 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
- Provide rigorously verified data for forensic investigations

# RDS Field Use Example

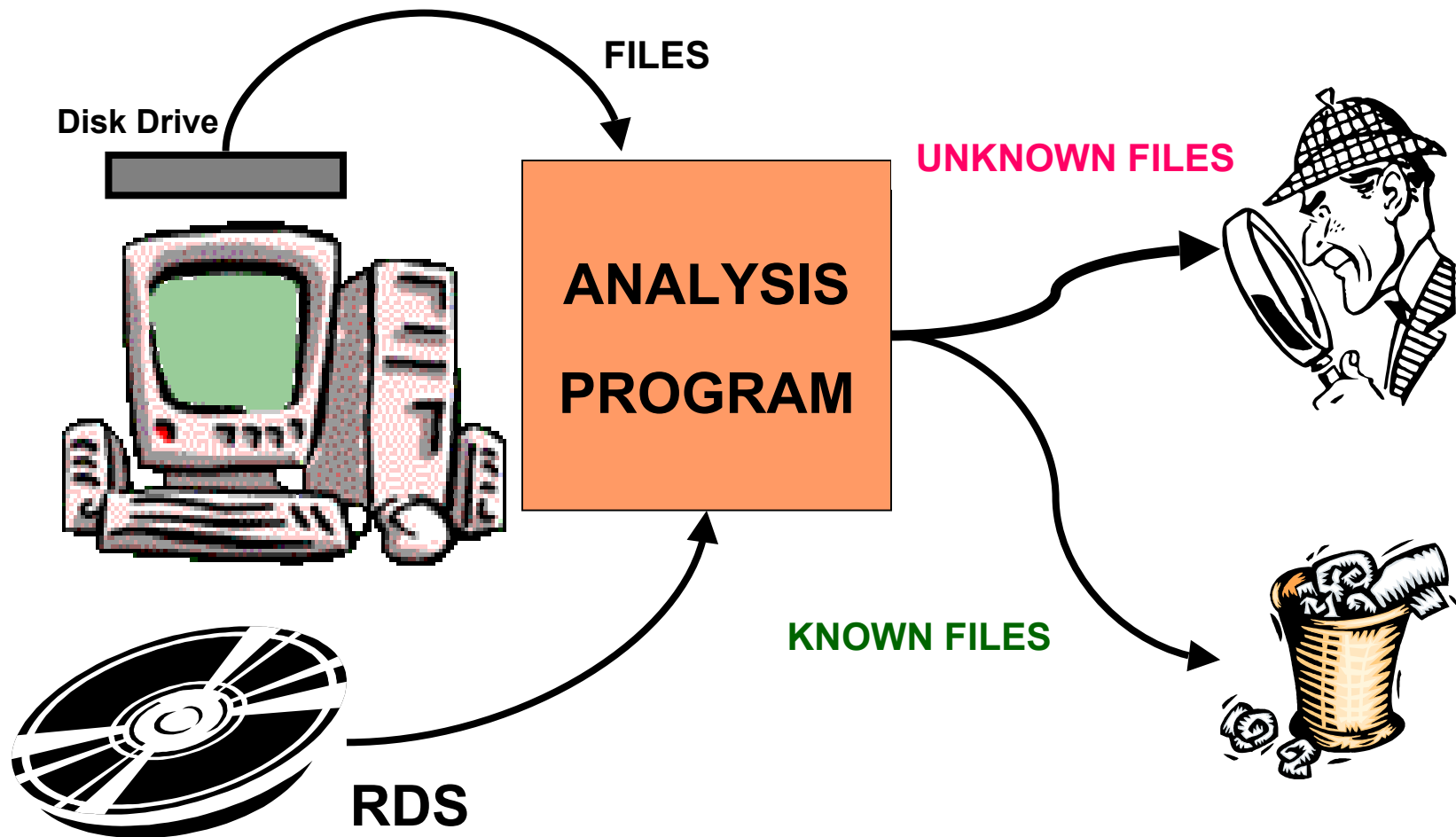
You are looking for facility maps on a computer which is running Windows 2000.

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



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

# RDS Field Use Concept



## Haunted By Ghosts Of Hard Drives Past

CAMBRIDGE, Mass., Jan. 16, 2003

---



Simson Garfinkel, a graduate student at the MIT's Laboratory for Computer Science, holds a used hard drive he bought containing personal information. (AP)

**(AP)** So, you think you cleaned all your personal files from that old computer you got rid of?

Two MIT graduate students suggest you think again.

Over two years, Simson Garfinkel and Abhi Shelat bought 158 used hard drives at secondhand computer stores and on eBay. Of the 129 drives that functioned, 69 still had recoverable files on them and 49 contained "significant personal information" - medical correspondence, love letters, pornography and 5,000 credit card numbers. One even had a year's worth of transactions with account numbers from a cash machine in Illinois.

<http://www.cbsnews.com/stories/2003/01/16/tech/main536774.shtml>

# 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



# Hashes

- SHA-1 values can be cross-referenced by other products that depend on different hash values
- Other standard hash values computed for each file include Message Digest 5 (MD5), and a 32-bit Cyclical Redundancy Checksum (CRC32), which are useful in CF tools and to users outside LE

# 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

# Related History

- CRC concept dates from 1960's
- MD5 algorithm published in 1991
- Tripwire open source tool 1992
- Unix command "md5sum" available
- FIPS 180-1 (SHA-1) published in 1995
- Unix command "sha1sum" available
- Known File Filter project 1998
- FIPS 180-2 (SHA-512) published in 2002

# Hashes in P2P

datSourceList: 1 or more concatenated variable length Source subsequences, each with this layout:

1	srcRemoteFileName	ANSIZ	"Tiken Jah Fakoli - Y'en A Marre.mp3", "\0"	Variable length name (no path) of the file at the remote source node. Can be empty ("\0" only).
2	srcDownloadURL	ANSIZ	"\0"	Variable length filepath. Always empty in Kazaa 2.0 ("\0" only).
3	srcFileID	DWORD	3731	Small unique number, typically identical for a given file contents at all Sources. Used as virtual subdirectory name in HTTP GET requests like "http://<KazaaHost>:<KazaaPort>/3731/Tiken%20Jah%20Fakoli....mp3". Such URL's were used for file exchanges with peer nodes in Kazaa 1.x (only ?). Kazaa 2.0.2 still returns a list of these URL's in a HTML page table produced by a local <a href="http://localhost:&lt;KazaaPort&gt;/">http://localhost:&lt;KazaaPort&gt;/</a> GET root request.
4	srcContentHash	BYTE[20]	16-byte crypto std. MD5 one-way digest + 4-byte FastTrack smallhash DWORD	Fingerprint for file content. Files with identical content have an identical Message Digest value. No different content anywhere should have this same cryptographic secure, one-way MD5-value. FastTrack smallhash adds more content identifying bits. Not all file data is included in the hash calculation.
5	srcFileSize	DWORD	3751893	Size of the remote file at this Source. Counted in bytes.
6	srcNodeInetNum	DWORD	0x04,0x03,0x02,0x01 = 0x01020304 = "1.2.3.4"	32-bit IP address number of this FastTrack peer Source node.
7	srcNodeTcpPort	DWORD	0xF1,0x0A,0x00,0x00 = 0x0af1 = 2801	TCP file service port of FastTrack peer file Source. Default KaZaA file service IP-port is 1214. Can be NULL for a file service behind a proxy server and/or firewall that does not allow the remote source peer to accept incoming TCP connections. In these cases the destination node sends a PUSH Request to his supernode. This PUSH request is mediated through the (network of) supernode(s) to the source node. The source node then initiates a new TCP connection to the destination peer. Over this connection, the source sends a GIVE nnnn command to tell the destination that it can start the actual HTTP GET file download session on

KaZaA Peer-to-Peer (P2P) FastTrack File Formats

<http://kzfti.cjb.net/>

# SHA-1 Mathematics

- Bit sequence is padded to a multiple of 512
- Messages of 16 32-bit words,  $n \cdot 512$ ,  $n > 0$
- 80 logic functions are defined that accept 3 32-bit words and produce 1 32-bit word
- 80 constants defined, 5 32-bit buffers initialized
- 80 step loop:
  - Manipulate message into 80 32-bit words
  - Use shifts, functions, addition on buffers
- 160-bit SHA is string in the 5 32-bit buffers

# Application of RDS

<b>OS/Apps</b>	<b>Files installed</b>	<b>Percent identified</b>	<b>Files unknown</b>	<b>Files on distribution CD(s)</b>
Virgin Win 98	4,266	93%	297	18,662
Virgin NT4 WS	1,659	86%	239	17,904
Virgin Win 2Kpro	5,963	86%	839	16,539
Virgin Win ME	5,169	93%	383	11,512
Win 98+Office 2K	23,464	98%	596	43,327
Win ME+Office 2K	24,112	98%	526	32,758
NIST PC #1 W2K	18,048	35%	11,839	N/A
NIST PC #2 W2K	59,135	20%	47,124	N/A
NIST PC #3 WNT	14,186	54%	6,618	N/A
NIST PC #4 W98	16,397	55%	7,404	N/A
NIST PC #5 W98	34,220	75%	8,667	N/A

# NIST Research

- Hash collisions
- Software distribution metrics
- Operating/File system effects
- Physical/Virtual machine effects
- “Mining” dynamic files
- Offsite hashing

# Software Installation Issues

- Dynamic files are “missed” by RDS
- Installed on virtual machines which can be saved in the NSRL on media
- Delineation of static sections of files for probability of identification
- Independent of installation location



# NARA Research

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

# NARA Statistics

- 93 computer systems
  - Pre-filtered to contain only software
- 51,146 individual files
- 7,610 file names
- 11,118 distinct files (SHA-1)
- 8,077 files originating in specific application(s)
- 4,326 of 8,077 exactly match application file names

# Further NARA Research

- Building profile of a “master” image
- Statistical weights for application identification
- Cross-system relationships
- Installation locations
- Old compression technologies

# NSRL Environment

- Isolated network with domain controller, DHCP
- Database server, File server, Web server
- Batch processing stations use web browser interface
- Hashing constellation
- Virtual machines for installations
- CVS source code repository


NSRL Package Information - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites History Print Copy Paste

Address http://192.168.58.4/login/script.asp Go Links

---



# Package Information

## NSRL Builder

---

*Application Name:	<input type="text" value="Your Eyes Only"/>	
*Version:	<input type="text" value="win95 1.0"/>	
Bar Code:	<input type="text" value="037646120487"/>	
*Language:	English <input type="button" value="v"/>	Please Specify: <input type="text"/>
Manufacturer:	Symantec <input type="button" value="v"/>	
*Application Type:	Utility <input type="button" value="v"/>	Please Specify: <input type="text"/>
Packaged Within:	<input type="text"/>	
Comments:	<input type="text"/>	
*Location:	G2 <input type="text"/>	

---

Contact Us:

National Institute of Standards and Technology  
 ATTN: NSRL Project  
 100 Bureau Drive, Stop 8970  
 Gaithersburg, MD 20899-8970 USA

Done Internet

# Hashing Operations

- Spring 2003 – accepting software
- Hashing constellation runs 24/7
- Processed over 13.4M files, 9M SHAs
- Byte signature file type verification
- CAB, ZIP, TAR, SFX, UU, compress

# Data Verification

- Multiple and independent techniques from different perspectives
  - We use test files with known signatures
  - Parallel database system: Match results with other system
  - Human verification
  - Database rules and constraints
  - Periodic database queries: Predefined procedures to search for and report anomalies in the database
  - User feedback: Error reports and RDS updates

# Future Operation Tasks

- More hardware platforms
- More archive tools
- Redundant hashing in constellation
- Scheduled rebatching
- Additional algorithms – AES
- Open source LAMP distribution



# NSRL/CFTT Team



# Contact

NSRL Project, Douglas White

Telephone: 301-975-8425

Email: [nsrl@nist.gov](mailto:nsrl@nist.gov)

Web: [www.nsrl.nist.gov](http://www.nsrl.nist.gov)

# Contacts

Jim Lyle

[www.cftt.nist.gov](http://www.cftt.nist.gov)

[cftt@nist.gov](mailto:cftt@nist.gov)

Doug White

[www.nsrl.nist.gov](http://www.nsrl.nist.gov)

[nsrl@nist.gov](mailto:nsrl@nist.gov)

Mark Skall

Chief, Software Diagnostics & Conformance Testing Div.

[www.itl.nist.gov/div897](http://www.itl.nist.gov/div897)

[skall@nist.gov](mailto:skall@nist.gov)

Sue Ballou, Office of Law Enforcement Standards

Steering Committee Rep. For State/Local Law Enforcement

[susan.ballou@nist.gov](mailto:susan.ballou@nist.gov)