Data Formats of the NSRL Reference Data Set (RDS) Distribution

Introduction

This report describes the format of data included in the distribution of the National Software Reference Library (NSRL) Reference Data Set (RDS). The NSRL is a repository of commercial-off-the-shelf (COTS) software. The RDS consists of file profiles made up of file identifying information and hash values from the individual files provided in a software package's installation media. An extract of the NSRL database is produced periodically and made available through the U.S. Department of Commerce's (DoC) National Institute of Standards and Technology (NIST) Special Database program as NIST Special Database #28. (See "Physical Distribution Media.")

As software is received in the NSRL, it is processed on a production machine that runs programs to compute a file signature, or hash value, for each file within the installation media. The file name, hash value, and other descriptive information about each file are stored in the NSRL database. When sufficient changes to the database have been made, a new distribution package of the extracted information is produced and distributed.

Typical uses for this data set include the forensic examination of computers and stored data seized by law enforcement agencies, copyright infringement investigations, and similar types of functions. The structure of the data in the RDS is described in the remainder of the report. The sections that follow cover the RDS terms and logical record structure.

RDS Data Elements

The following table describes the data elements used in the NSRL RDS distribution package. *Char* represents data of type character using ASCII encoding in 8-bit bytes, with the exception of the FileName element (see Note, below). *Integer* represents data of type integer including variations of the integer type (short, long, etc.)

Note: The FileName element is unique among the elements of type *char* in that it is derived directly from the software distribution media. The character encoding of this element may vary, and will be the same as the encoding used by the software manufacturer.

All of the data is stored in the distribution files in human-readable form. No binary data or nonstandard characters are used. *Char* fields are represented by alphabetic, numeric, and punctuation character strings surrounded by double quotes ("). *Integers* are represented by unquoted strings of decimal digits.

Table 1. RDS Data Elements

DATA ELEMENT	TYPE	MAXIMUM LENGTH (IN CHARACTERS)	DESCRIPTION
ApplicationType	Char	50	Character string that identifies a general use of the software product
CRC32	Char	8	32-bit Cyclic Redundancy Checksum (file signature) of a specific file as defined in CCITT X.25 link-level protocol and FIPS PUB 71
FileName	Char	255	Name of a specific file within a software product
FileSize	Integer	15	Size in bytes of a specific file
Language	Char	150	Character string that identifies the language(s) used in the software product
MD5	Char	32	128-bit Message Digest 5 (file signature) of a specific file as defined in IETF RFC 1321
MfgCode	Char	15	Character identifier of a specific vendor or manufacturer
MfgName	Char	150	Identifying name of the vendor or manufacturer of the software product, e.g., “Microsoft”
OpSystemName	Char	150	Identifying name of the operating system on which the software product executes, e.g., “Windows NT”
OpSystemCode	Char	15	Code identifier of a specific operating system version
OpSystemVersion	Char	15	Characters that identify individual versions of an operating system on which the software product executes, e.g., “4.0”
ProductCode	Integer	15	Identifier of a specific software product, e.g., “103”; maps to the NSRL database
ProductName	Char	150	Identifying name of the software product, e.g., “Netscape Communicator”
ProductVersion	Char	15	Characters that identify individual versions of a software product, e.g., “3.0”
RDSVersion	Char	40	Character string that identifies the date and version of the RDS distribution.
SHA-1	Char	40	160-bit Secure Hash Algorithm message digest (file signature) of a specific file as defined in FIPS PUB 180-2
SpecialCode	Char	1	A single character field that identifies special file signature entries, such as malicious code signatures or other types of special entries

Logical Record Structure

A logical record forms one item or grouping of information from the data elements defined in the above table within the NSRL RDS. There are five such logical record types: (1) file record, (2) manufacturer record, (3) product record, (4) operating system record, and (5) version record. Each is described in the following tables. Examples of each type of record are also provided. Figure 1 illustrates how these files relate to each other.

Table 2. FILE Record Type

RECORD FORMAT		EXAMPLE	COMMENTS
FILE RECORD			
	SHA-1	"AC91EF00F33F12DD491CC91EF00F33F12DD491CA"	
	MD5	"DC2311FFDC0015FCCC12130FF145DE78"	
	CRC32	"14CCE9061FFDC001"	
	FileName	"WORD.EXE"	
	FileSize	1217654	In bytes
	ProductCode	103	The Product record will contain more information about this product code.
	OpSystemCode	"NT4WKS"	The Operating System record will contain more information about this operating system code.
	SpecialCode	""	Blank (no value) – normal file "M" – malicious file "S" – special file

Table 3. MANUFACTURER Record Type

RECORD FORMAT		EXAMPLE	COMMENTS
MANUFACTURER RECORD			
	MfgCode	"Microsoft"	MfgCode is referenced in the Operating System and Product records. MfgCode is unique within this record set.
	MfgName	"Microsoft Corporation"	

Table 4. OPERATING SYSTEM Record Type

RECORD FORMAT		EXAMPLE	COMMENTS
OPERATING SYSTEM RECORD			
	OpSystemCode	"NT4WKS"	OpSystemCode is referenced in the File record and is unique within the Operating System record set.
	OpSystemName	"Windows NT"	
	OpSystemVersion	"4.0"	
	MfgCode	"Microsoft"	MfgCode references an entry in the Manufacturer record.

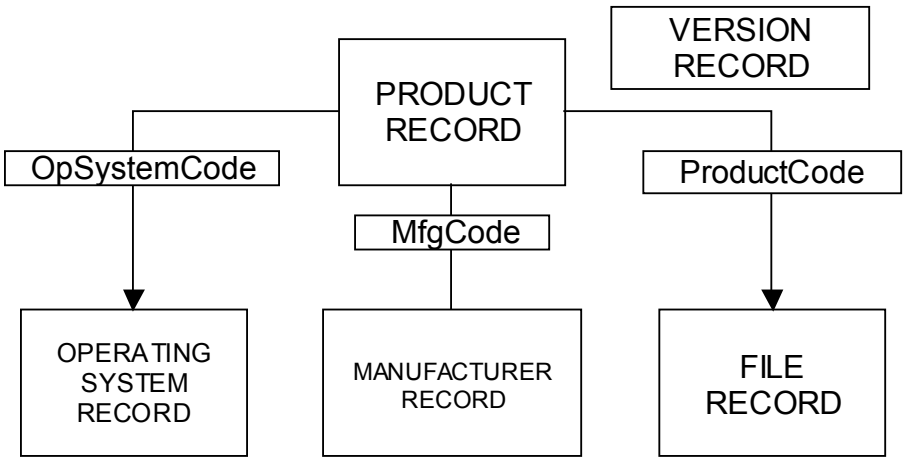
Table 5. PRODUCT Record Type

RECORD FORMAT		EXAMPLE	COMMENTS
PRODUCT RECORD			
	ProductCode	103	ProductCode is referenced in the File record and is unique within the Product record set.
	ProductName	"Microsoft Word"	
	ProductVersion	"2000"	
	OpSystemCode	"Win98"	OpSystemCode is referenced in the Operating System record
	MfgCode	"Microsoft"	MfgCode references an entry in the Manufacturer record.
	Language	"English"	If multiple languages are present, they will be comma separated within this field
	ApplicationType	"Operating System"	

Table 5. RDS VERSION Record Type

RECORD FORMAT		EXAMPLE	COMMENTS
SHA-1 RECORD			
	SHA-1	"AC91EF00F33F12DD491CC91E F00F33F12DD491CA"	This value of SHA-1 is computed from the SHA-1 values of the four other files.
	RDSVersion	"2001/03/08 0.2"	Assigned to each quarterly release of the RDS.

Figure 1. NSRL RDS Logical Record Relationships



Physical Record Structure

The RDS consists of five physical data files that correspond to the five logical record types, one file per logical record type. The character format is UTF-8 (8-bit ASCII), one logical record per physical line terminated with ASCII characters 13 and 10 (hexadecimal 0D0A). Individual fields are separated by comma (,) within each line. Character field values are surrounded by double quotation marks (“”). The first record of each file contains the field names instead of data values. Examples of the contents of each file are presented in figures 2 through 6. The first record in each figure represents the first or header record found in each file. The second record in each figure represents all subsequent or detail records in each file.

Figure 2. FILE Example Data

”SHA-1”,”MD5”,”CRC32”,“FileName”,”FileSize”,”ProductCode”,”OpSystemCode”,
”SpecialCode” <13><10>
“AC91EF00F33F12DD491CC91EF00F33F12DD491CA”,“DC2311FFDC0015FCCC12130FF145DE78”,
“14CCE9061FFDC001”, “WORD.EXE”,1217654,103, ”T4WKS”,”” <13><10>

Figure 3. MANUFACTURER Example Data

“MfgCode”,”MfgName” <13><10>
“Microsoft”,“Microsoft Corporation” <13><10>

Figure 4. OPERATING SYSTEM Example Data

“OpSystemCode”,“OpSystemName”,“OpSystemVersion”,“MfgCode” <13><10>
“NT4WKS”,“Windows NT”,“4.0”,“Microsoft” <13><10>

Figure 5. PRODUCT Example Data

"ProductCode","ProductName","ProductVersion","OpSystemCode","MfgCode",
"Language","ApplicationType" <13><10>
"103","Microsoft Office","2000","Win98","Microsoft","English","Word Processor" <13><10>

Figure 6. RDS VERSION Example Data

"SHA-1","RDSVersion" <13><10>
"DD161AEFCC271124533FFFA1445764BDE12515AE","2001/03/08 0.2" <13><10>

Physical Distribution Media

The distribution of the RDS will take the form of quarterly compact discs (CD) from NIST's Standard Reference Data (SRD) Office (http://www.nist.gov/srdata) as Special Database #28. Each CD will include a full version of the RDS, i.e., each issue is cumulative and can replace previous versions. Each file will be variable in size with a full complement of files from one or more packages. Further, the files will be named NSRLFILE.TXT, NSRLOS.TXT, NSRLMFG.TXT, NSRLPROD.TXT, and VERSION.TXT.

It will be up to an individual user to determine how these files are used and provided to applications running in an investigative environment. They can be used separately as is, combined into a single database of information, or in various other combinations depending on the requirements of the particular environment in which they are used.

Additional software for demonstration purposes may be added to the distribution from time to time.

If space permits, other formats or subsets of data may be added to the distribution as requested by users. Those sets of data will be documented within the directories associated with those data sets.

A typical distribution of the RDS will contain over a million NSRLFILE records, several hundred NSRLMFG and NSRLPROD records, several dozen NSRLOS records, and only one VERSION record. The size will grow as new and updated applications are added to the database.

Appendix 1. Change Record

1. 2001/01/25 – Editorial changes, corrections to record layouts, modification of distribution naming and versioning.
2. 2001/01/29 – Addition of Virus data element to File record, modification of data records to use ASCII 10 character instead of ASCII 13 to end a record, addition of information on physical distribution to indicate additional records or replacement database.
3. 2001/03/08 – Added section on fifth file, SHA-1.txt.

4. 2001/07/06 – Modified file structures, file names, distribution information.
5. 2001/09/05 – Editorial changes and corrections.
6. 2001/10/11 – Modified manufacturer file structure, added SpecialCode field to NSRLFILE.TXT structure, and added clarifications to physical file descriptions.
7. 2001/10/23 – Modified NSRLFILE structure to remove PathHash.
8. 2001/10/30 – Modified NSRLFILE structure to remove FileExt.
9. 2001/11/26 – Corrected NSRLFILE structure to show hexadecimal 0D0A as line terminator.
10. 2002/12/30 – **MAJOR** modification of NSRLFILE structure by placement of SHA-1, MD5 and CRC32 fixed length fields at beginning of record and elimination of MD4 field. Modified NSRLPROD structure to add Language field and ApplicationType field. Corrected data elements table to reflect updated database definitions.
11. 2004/10/25 – Modified RDSVersion field in VERSION.TXT from 20 characters to 40 characters.
12. 2007/02/07 – Clarification of character encoding used in the RDS Data Elements, with special reference to that used for the FileName field.
13. 2009-07-07 – Corrected definition of table headers and ordering of example content in Figure 5, “PRODUCT Example Data”.
14. 2009-07-08 – Captioned Table 1 and renumbered subsequent tables for consistency.