

# Creating signatures for ClamAV

## 1 Introduction

CVD (ClamAV Virus Database) is a digitally signed container that includes signature databases in various text formats. The header of the container is a 512 bytes long string with colon separated fields:

```
ClamAV-VDB:build time:version:number of signatures:functionality
level required:MD5 checksum:digital signature:builder name:build
time (sec)
```

`sigtool --info` displays detailed information about a given CVD file:

```
zolw@localhost:/usr/local/share/clamav$ sigtool -i main.cvd
File: main.cvd
Build time: 09 Dec 2007 15:50 +0000
Version: 45
Signatures: 169676
Functionality level: 21
Builder: sven
MD5: b35429d8d5d60368eea9630062f7c75a
Digital signature: dxsus0/HWP3/GAA7VuZpxYwVsE9b+tCk+tPN6OyjVF/U8
JVh4vYmW8mZ62ZHYMlM903TMZFg5hZIXc jQB3SX0TapdF1SFNzoWjsyH53eXvMDY
eaPVNe2ccXLfEegoda4xU2TezbGfbSEGoU1qolyQYLX674sNA2Ni6l6/CEKYYh
Verification OK.
```

The ClamAV project distributes a number of CVD files, including *main.cvd* and *daily.cvd*.

## 2 Signature formats

### 2.1 MD5

The easiest way to create signatures for ClamAV is to use MD5 checksums, however this method can be only used against static malware. To create a signature

for test.exe use the --md5 option of sigtool:

```
zolw@localhost:/tmp/test$ sigtool --md5 test.exe > test.hdb
zolw@localhost:/tmp/test$ cat test.hdb
48c4533230e1ae1c118c741c0db19dfb:17387:test.exe
```

That's it! The signature is ready for use:

```
zolw@localhost:/tmp/test$ clamscan -d test.hdb test.exe
test.exe: test.exe FOUND
```

```
----- SCAN SUMMARY -----
Known viruses: 1
Scanned directories: 0
Engine version: 0.92.1
Scanned files: 1
Infected files: 1
Data scanned: 0.02 MB
Time: 0.024 sec (0 m 0 s)
```

You can change the name (by default sigtool uses the name of the file) and place it inside a \*.hdb file. A single database file can include any number of signatures. To get them automatically loaded each time clamscan/clamd starts just copy the database file(s) into the local virus database directory (eg. /usr/local/share/clamav).

## 2.2 MD5, PE section based

You can create a MD5 signature for a specific section in a PE file. Such signatures shall be stored inside .mdb files in the following format:

```
PESectionSize:MD5:MalwareName
```

The easiest way to generate MD5 based section signatures is to extract target PE sections into separate files and then run sigtool with the option --mdb

## 2.3 Body-based signatures

ClamAV stores all body-based signatures in a hexadecimal format. In this section by a hex-signature we mean a fragment of malware's body converted into a hexadecimal string which can be additionally extended using various wildcards.

### 2.3.1 Hexadecimal format

You can use `sigtool --hex-dump` to convert any data into a hex-string:

```
zolw@localhost:/tmp/test$ sigtool --hex-dump  
How do I look in hex?  
486f7720646f2049206c6f6f6b20696e206865783f0a
```

### 2.3.2 Wildcards

ClamAV supports the following extensions for hex-signatures:

- `??`  
Match any byte.
- `a?`  
Match a high nibble (the four high bits).  
**IMPORTANT NOTE:** The nibble matching is only available in libclamav with the functionality level 17 and higher therefore please only use it with `.ndb` signatures followed by `":17"` (`MinEngineFunctionalityLevel`, see 2.3.5).
- `?a`  
Match a low nibble (the four low bits).
- `*`  
Match any number of bytes.
- `{n}`  
Match  $n$  bytes.
- `{-n}`  
Match  $n$  or less bytes.
- `{n-}`  
Match  $n$  or more bytes.
- `{n-m}`  
Match between  $n$  and  $m$  bytes ( $m > n$ ).
- `(aa|bb|cc|...)`  
Match aa or bb or cc..

- `!(aa|bb|cc|..)`  
Match any byte except aa and bb and cc.. (ClamAV $\geq$ 0.96)
- `HEXSIG[x-y]aa` or `aa[x-y]HEXSIG`  
Match aa anchored to a hex-signature, see [https://www.clamav.net/bugzilla/show\\_bug.cgi?id=776](https://www.clamav.net/bugzilla/show_bug.cgi?id=776) for discussion and examples.
- `(B)`  
Match word boundary (including file boundaries).
- `(L)`  
Match CR, CRLF or file boundaries.

The range signatures `*` and `{ }` virtually separate a hex-signature into two parts, eg. `aabbcc*bbaacc` is treated as two sub-signatures `aabbcc` and `bbaacc` with any number of bytes between them. It's a requirement that each sub-signature includes a block of two static characters somewhere in its body.

### 2.3.3 Basic signature format

The simplest (and now deprecated) signature format is:

```
MalwareName=HexSignature
```

ClamAV will scan the entire file looking for HexSignature. All signatures of this type must be placed inside `*.db` files.

### 2.3.4 Extended signature format

The extended signature format allows for specification of additional information such as a target file type, virus offset or engine version, making the detection more reliable. The format is:

```
MalwareName:TargetType:Offset:HexSignature[:MinFL:[MaxFL]]
```

where `TargetType` is one of the following numbers specifying the type of the target file:

- 0 = any file
- 1 = Portable Executable, both 32- and 64-bit.

- 2 = file inside OLE2 container (e.g. image, embedded executable, VBA script). The OLE2 format is primarily used by MS Office and MSI installation files.
- 3 = HTML (normalized: whitespace transformed to spaces, tags/tag attributes normalized, all lowercase), Javascript is normalized too: all strings are normalized (hex encoding is decoded), numbers are parsed and normalized, local variables/function names are normalized to 'n001' format, argument to eval() is parsed as JS again, unescape() is handled, some simple JS packers are handled, output is whitespace normalized.
- 4 = Mail file
- 5 = Graphics
- 6 = ELF
- 7 = ASCII text file (normalized)
- 8 = Unused
- 9 = Mach-O files

And `Offset` is an asterisk or a decimal number `n` possibly combined with a special modifier:

- `*` = any
- `n` = absolute offset
- `EOF-n` = end of file minus `n` bytes

Signatures for PE, ELF and Mach-O files additionally support:

- `EP+n` = entry point plus `n` bytes (`EP+0` for `EP`)
- `EP-n` = entry point minus `n` bytes
- `Sx+n` = start of section `x`'s (counted from 0) data plus `n` bytes
- `Sx-n` = start of section `x`'s data minus `n` bytes
- `SL+n` = start of last section plus `n` bytes
- `SL-n` = start of last section minus `n` bytes

All the above offsets except \* can be turned into **floating offsets** and represented as `Offset,MaxShift` where `MaxShift` is an unsigned integer. A floating offset will match every offset between `Offset` and `Offset+MaxShift`, eg. `10,5` will match all offsets from 10 to 15 and `EP+n,y` will match all offsets from `EP+n` to `EP+n+y`. Versions of ClamAV older than 0.91 will silently ignore the `MaxShift` extension and only use `Offset`.

Optional `MinFL` and `MaxFL` parameters can restrict the signature to specific engine releases. All signatures in the extended format must be placed inside \*.ndb files.

### 2.3.5 Logical signatures

Logical signatures allow combining of multiple signatures in extended format using logical operators. They can provide both more detailed and flexible pattern matching. The logical sigs are stored inside \*.ldb files in the following format:

```
SignatureName;TargetDescriptionBlock;LogicalExpression;Subsig0;  
Subsig1;Subsig2;...
```

where:

- `TargetDescriptionBlock` provides information about the engine and target file with comma separated `Arg:Val` pairs, currently (as of 0.95.1) only `Target:X` and `Engine:X-Y` are supported.
- `LogicalExpression` specifies the logical expression describing the relationship between `Subsig0...SubsigN`.  
**Basis clause:** 0,1,...,N decimal indexes are SUB-EXPRESSIONS representing `Subsig0, Subsig1, ..., SubsigN` respectively.  
**Inductive clause:** if A and B are SUB-EXPRESSIONS and X, Y are decimal numbers then  $(A\&B)$ ,  $(A|B)$ ,  $A>X$ ,  $A=X, Y$ ,  $A>X, A>X, Y$ ,  $A<X$  and  $A<X, Y$  are SUB-EXPRESSIONS
- `SubsigN` is n-th subsignature in extended format possibly preceded with an offset. There can be specified up to 64 subsigs.

Keywords used in `TargetDescriptionBlock`:

- `Target:X`: Target file type
- `Engine:X-Y`: Required engine functionality (range; 0.96)
- `FileSize:X-Y`: Required file size (range in bytes; 0.96)

- `EntryPoint`: Entry point offset (range in bytes; 0.96)
- `NumberOfSections`: Required number of sections in executable (range; 0.96)
- `Container:CL_TYPE_*`: File type of the container which stores the scanned file

Modifiers for subexpressions:

- `A=X`: If the SUB-EXPRESSION A refers to a single signature then this signature must get matched exactly X times; if it refers to a (logical) block of signatures then this block must generate exactly X matches (with any of its sigs).
- `A=0` specifies negation (signature or block of signatures cannot be matched)
- `A=X,Y`: If the SUB-EXPRESSION A refers to a single signature then this signature must be matched exactly X times; if it refers to a (logical) block of signatures then this block must generate X matches and at least Y different signatures must get matched.
- `A>X`: If the SUB-EXPRESSION A refers to a single signature then this signature must get matched more than X times; if it refers to a (logical) block of signatures then this block must generate more than X matches (with any of its sigs).
- `A>X,Y`: If the SUB-EXPRESSION A refers to a single signature then this signature must get matched more than X times; if it refers to a (logical) block of signatures then this block must generate more than X matches and at least Y different signatures must be matched.
- `A<X` and `A<X,Y` as above with the change of "more" to "less".

Examples:

```
Sig1;Target:0;(0&1&2&3)&(4|1);6b6f74656b;616c61;7a6f6c77;73746566616e;deadbeef
```

```
Sig2;Target:0;((0|1|2)>5,2)&(3|1);6b6f74656b;616c61;7a6f6c77;73746566616e
```

```
Sig3;Target:0;((0|1|2|3)=2)&(4|1);6b6f74656b;616c61;7a6f6c77;73746566616e;deadbeef
```

```
Sig4;Target:1,Engine:18-20;((0|1)&(2|3))&4;EP+123:33c06834f04100
f2aef7d14951684cf04100e8110a00;S2+78:22??232c2d252229{-15}6e6573
(63|64)61706528;S+50:68efa311c3b9963cb1ee8e586d32aeb9043e;f9c58d
cf43987e4f519d629b103375;SL+550:6300680065005c0046006900
```

ClamAV 0.96 introduced support for special macro subsignatures in the following format:  `${min-max}MACROID$`, where `MACROID` points to a group of signatures and `{min-max}` specifies the offset range at which one of the group signatures should match. The range is calculated against the match offset of the previous subsignature. The macro subsignature makes its preceding subsignature considered a match only if both of them get matched. For more information and examples please see [https://www.clamav.net/bugzilla/show\\_bug.cgi?id=164](https://www.clamav.net/bugzilla/show_bug.cgi?id=164).

## 2.4 Icon signatures for PE files

ClamAV 0.96 includes an approximate/fuzzy icon matcher to help detecting malicious executables disguising themselves as innocent looking image files, office documents and the like.

Icon matching is only triggered via `.ldb` signatures using the special attribute tokens `IconGroup1` or `IconGroup2`. These identify two (optional) groups of icons defined in a `.ldb` database file. The format of the `.ldb` file is:

```
ICONNAME:GROUP1:GROUP2:ICON_HASH
```

where:

- `ICON_NAME` is a unique string identifier for a specific icon,
- `GROUP1` is a string identifier for the first group of icons (`IconGroup1`)
- `GROUP2` is a string identifier for the second group of icons (`IconGroup2`),
- `ICON_HASH` is a fuzzy hash of the icon image

The `ICON_HASH` field can be obtained from the debug output of `libclamav`. For example:

```
LibClamAV debug: ICO SIGNATURE:
ICON_NAME:GROUP1:GROUP2:18e2e0304ce60a0cc3a09053a30000414100057e
000afe0000e 80006e510078b0a08910d11ad04105e0811510f084e01040c080
a1d0b0021000a39002a41
```



## 2.5 Signatures for Version Information metadata in PE files

Starting with ClamAV 0.96 it is possible to easily match certain information built into PE files (executables and dynamic link libraries). Whenever you lookup the properties of a PE executable file in windows, you are presented with a bunch of details about the file itself.

These info are stored in a special area of the file resources which goes under the name of VS\_VERSION\_INFORMATION (or versioninfo for short). It is divided into 2 parts. The first part (which is rather uninteresting) is really a bunch of numbers and flags indicating the product and file version. It was originally intended for use with installers which, after parsing it, should be able to determine whether a certain executable or library are to be upgraded/overwritten or are already up to date. Suffice to say, this approach never really worked and is generally never used.

The second block is much more interesting: it is a simple list of key/value strings, intended for user information and completely ignored by the OS. For example, if you look at ping.exe you can see the company being "Microsoft Corporation", the description "TCP/IP Ping command", the internal name "ping.exe" and so on... Depending on the OS version, some keys may be given peculiar visibility in the file properties dialog, however they are internally all the same.

To match a versioninfo key/value pair, the special file offset anchor VI was introduced. This is similar to the other anchors (like EP and SL) except that, instead of matching the hex pattern against a single offset, it checks it against each and every key/value pair in the file. The VI token doesn't need nor accept a +/- offset like e.g. EP+1. As for the hex signature itself, it's just the utf16 dump of the key and value. Only the ?? and (aa|bb) wildcards are allowed in the signature. Usually, you don't need to bother figuring it out: each key/value pair together with the corresponding VI-based signature is printed by clamscan when the --debug option is given.

For example clamscan --debug freecell.exe produces:

```
[...]  
Recognized MS-EXE/DLL file  
in cli_peheader  
versioninfo_cb: type: 10, name: 1, lang: 410, rva: 9608  
cli_peheader: parsing version info @ rva 9608 (1/1)  
VersionInfo (d2de): 'CompanyName'='Microsoft Corporation' -  
VI:43006f006d00700061006e0079004e0061006d006500000000004d006900  
630072006f0073006f0066007400200043006f00720070006f0072006100740  
069006f006e000000  
VersionInfo (d32a): 'FileDescription'='Entertainment Pack  
FreeCell Game' - VI:460069006c006500440065007300630072006900700
```

```

0740069006f006e000000000045006e007400650072007400610069006e006d
0065006e00740020005000610063006b0020004600720065006500430065006
c006c002000470061006d0065000000
VersionInfo (d396): 'FileVersion'='5.1.2600.0 (xpclient.010817
-1148)' - VI:460069006c006500560065007200730069006f006e00000000
0035002e0031002e0032003600300030002e003000200028007800700063006
c00690065006e0074002e003000310030003800310037002d00310031003400
380029000000
VersionInfo (d3fa): 'InternalName'='freecell' - VI:49006e007400
650072006e0061006c004e0061006d006500000066007200650065006300650
06c006c0000000
VersionInfo (d4ba): 'OriginalFilename'='freecell' - VI:4f007200
6900670069006e0061006c00460069006c0065006e0061006d0065000000660
0720065006500630065006c006c000000
VersionInfo (d4f6): 'ProductName'='Sistema operativo Microsoft
Windows' - VI:500072006f0064007500630074004e0061006d006500000000
000530069007300740065006d00610020006f00700065007200610074006900
76006f0020004d006900630072006f0073006f0066007400ae0020005700690
06e0064006f0077007300ae000000
VersionInfo (d562): 'ProductVersion'='5.1.2600.0' - VI:50007200
6f006400750063007400560065007200730069006f006e000000035002e00310
02e0032003600300030002e0030000000
[...]

```

Although VI-based signatures are intended for use in logical signatures you can test them using ordinary .ndb files. For example:

```
my_test_vi_sig:1:VI:paste_your_hex_sig_here
```

Final note. If you want to decode a VI-based signature into a human readable form you can use:

```
echo hex_string | xxd -r -p | strings -el
```

For example:

```
$ echo 460069006c0065004400650073006300720069007000740069006f006e
000000000045006e007400650072007400610069006e006d0065006e007400200
05000610063006b0020004600720065006500430065006c006c00200047006100
6d0065000000 | xxd -r -p | strings -el
```

FileDescription  
Entertainment Pack FreeCell Game

## 2.6 Signatures based on container metadata

ClamAV 0.96 allows creating generic signatures matching files stored inside different container types which meet specific conditions. The signature format is

```
VirusName:ContainerType:ContainerSize:FileNameREGEX:  
FileSizeInContainer:FileSizeReal:IsEncrypted:FilePos:  
Res1:Res2[:MinFL[:MaxFL]]
```

where the corresponding fields are:

- **VirusName:** Virus name to be displayed when signature matches
- **ContainerType:** one of CL\_TYPE\_ZIP, CL\_TYPE\_RAR, CL\_TYPE\_ARJ, CL\_TYPE\_CAB, CL\_TYPE\_7Z, CL\_TYPE\_MAIL, CL\_TYPE\_(POSIX|OLD)\_TAR, CL\_TYPE\_CPIO\_(OLD|ODC|NEWC|CRC) or \* to match any of the container types listed here
- **ContainerSize:** size of the container file itself (eg. size of the zip archive) specified in bytes as absolute value or range x-y
- **FileNameREGEX:** regular expression describing name of the target file
- **FileSizeInContainer:** usually compressed size; for MAIL, TAR and CPIO == FileSizeReal; specified in bytes as absolute value or range
- **FileSizeReal:** usually uncompressed size; for MAIL, TAR and CPIO == FileSizeInContainer; absolute value or range
- **IsEncrypted:** 1 if the target file is encrypted, 0 if it's not and \* to ignore
- **FilePos:** file position in container (counting from 1); absolute value or range
- **Res1:** when ContainerType is CL\_TYPE\_ZIP or CL\_TYPE\_RAR this field is treated as a CRC sum of the target file specified in hexadecimal format; for other container types it's ignored
- **Res2:** not used as of ClamAV 0.96

The signatures for container files are stored inside .cdb files.

## 2.7 Signatures based on ZIP/RAR metadata (obsolete)

The (now obsolete) archive metadata signatures can be only applied to ZIP and RAR files and have the following format:

```
virname:encrypted:filename:normal size:csize:crc32:cmethod:  
fileno:max depth
```

where the corresponding fields are:

- Virus name
- Encryption flag (1 – encrypted, 0 – not encrypted)
- File name (this is a regular expression - \* to ignore)
- Normal (uncompressed) size (\* to ignore)
- Compressed size (\* to ignore)
- CRC32 (\* to ignore)
- Compression method (\* to ignore)
- File position in archive (\* to ignore)
- Maximum number of nested archives (\* to ignore)

The database file should have the extension of .zmd or .rmd for zip or rar metadata respectively.

## 2.8 Whitelist databases

To whitelist a specific file use the MD5 signature format and place it inside a database file with the extension of .fp.

To whitelist a specific signature from the database you just add its name into a local file called local.ign2 stored inside the database directory. You can additionally follow the signature name with the MD5 of the entire database entry for this signature, eg:

```
Eicar-Test-Signature:bc356bae4c42f19a3de16e333ba3569c
```

In such a case, the signature will no longer be whitelisted when its entry in the database gets modified (eg. the signature gets updated to avoid false alerts).

## 2.9 Signature names

ClamAV uses the following prefixes for signature names:

- *Worm* for Internet worms
- *Trojan* for backdoor programs
- *Adware* for adware
- *Flooder* for flooders
- *HTML* for HTML files
- *Email* for email messages
- *IRC* for IRC trojans
- *JS* for Java Script malware
- *PHP* for PHP malware
- *ASP* for ASP malware
- *VBS* for VBS malware
- *BAT* for BAT malware
- *W97M*, *W2000M* for Word macro viruses
- *X97M*, *X2000M* for Excel macro viruses
- *O97M*, *O2000M* for generic Office macro viruses
- *DoS* for Denial of Service attack software
- *DOS* for old DOS malware
- *Exploit* for popular exploits
- *VirTool* for virus construction kits
- *Dialer* for dialers
- *Joke* for hoaxes

Important rules of the naming convention:

- always use a *-zippwd* suffix in the malware name for signatures of type *zmd*,

- always use a -rarpwd suffix in the malware name for signatures of type rmd,
- only use alphanumeric characters, dash (-), dot (.), underscores (\_) in malware names, never use space, apostrophe or quote mark.

## 3 Special files

### 3.1 HTML

ClamAV contains a special HTML normalisation code which helps to detect HTML exploits. Running `sigtool --html-normalise` on a HTML file should generate the following files:

- `nocomment.html` - the file is normalized, lower-case, with all comments and superflous white space removed
- `notags.html` - as above but with all HTML tags removed

The code automatically decodes JScript.encode parts and char ref's (e.g. `&#102;`). You need to create a signature against one of the created files. To eliminate potential false positive alerts the target type should be set to 3.

### 3.2 Text files

Similarly to HTML all ASCII text files get normalized (converted to lower-case, all superflous white space and control characters removed, etc.) before scanning. Use `clamscan --leave-temps` to obtain a normalized file then create a signature with the target type 7.

### 3.3 Compressed Portable Executable files

If the file is compressed with UPX, FSG, Petite or other PE packer supported by libclamav, run `clamscan` with `--debug --leave-temps`. Example output for a FSG compressed file:

```
LibClamAV debug: UPX/FSG/MEW: empty section found - assuming compression
LibClamAV debug: FSG: found old EP @119e0
LibClamAV debug: FSG: Unpacked and rebuilt executable saved in
/tmp/clamav-f592b20f9329ac1c91f0e12137bcce6c
```

Next create a type 1 signature for `/tmp/clamav-f592b20f9329ac1c91f0e12137bcce6c`