

Lightgrep™ - Fast Search for Forensics

Lightgrep for EnCase[[1]](#footnote-2) provides you with new options for search in a familiar environment: Keywords that are compatible with Perl; filters for search terms to help you find the real information you're looking for faster; text extraction for Office 2007 and similar document formats.

Table of Contents

[Minimum Requirements 4](#_Toc318653573)

[Known Issues 4](#_Toc318653574)

[Limitations and General Information 4](#_Toc318653575)

[Search Tab 6](#_Toc318653576)

[Options 6](#_Toc318653577)

[Lightgrep Output Folder 6](#_Toc318653578)

[Create Bookmarks 7](#_Toc318653579)

[Create Excel Report 7](#_Toc318653580)

[Search only selected items 7](#_Toc318653581)

[Search for Keywords 7](#_Toc318653582)

[Search for Entities 7](#_Toc318653583)

[Ignore all items in loaded hash library (search slack) 7](#_Toc318653584)

[Ignore items with hash category ‘KNOWN’ (search slack) 7](#_Toc318653585)

[Search Status 7](#_Toc318653586)

[Search (Button) 8](#_Toc318653587)

[Keywords 9](#_Toc318653588)

[Options 10](#_Toc318653589)

[Import From EnCase 10](#_Toc318653590)

[Add Keyword List 11](#_Toc318653591)

[Import / Export XML 13](#_Toc318653592)

[Keyword Tree 13](#_Toc318653593)

[Editing Keywords 13](#_Toc318653594)

[Supported Pattern Syntax 14](#_Toc318653595)

[Individual Bytes 14](#_Toc318653596)

[Metacharacters 15](#_Toc318653597)

[Character Classes 15](#_Toc318653598)

[Character Class Metacharacters 15](#_Toc318653599)

[Named Character Classes 15](#_Toc318653600)

[Repetition 15](#_Toc318653601)

[Conjunction, Alternation, Grouping 16](#_Toc318653602)

[Entity Search 17](#_Toc318653603)

[Options 18](#_Toc318653604)

[Entities 18](#_Toc318653605)

[Entity Filter 18](#_Toc318653606)

[Validation Lists 18](#_Toc318653607)

[Import / Export Entities 18](#_Toc318653608)

[Add Sample Entities 18](#_Toc318653609)

[Advanced Tab 19](#_Toc318653610)

[Options 20](#_Toc318653611)

[File Signature List 20](#_Toc318653612)

[Import From EnCase 20](#_Toc318653613)

[Reset Defaults 20](#_Toc318653614)

[Run Lightgrep automatically 21](#_Toc318653615)

[Lightgrep executable path 21](#_Toc318653616)

[Excelutil executable path 21](#_Toc318653617)

[Number of Threads 21](#_Toc318653618)

[Log Level 21](#_Toc318653619)

[Starting a Search 22](#_Toc318653620)

[Monitoring Search Progress 23](#_Toc318653621)

[Reviewing Output 24](#_Toc318653622)

[Bookmark Output 24](#_Toc318653623)

[Excel Output 25](#_Toc318653624)

# Minimum Requirements

* EnCase Forensic or Enterprise version 6.16.2 (32-bit or 64-bit) or greater
* Windows XP SP3, Windows 7 (32-bit or 64-bit) – the application has not been tested with Windows Vista.
* 4 GB of RAM preferred minimum
* Microsoft .Net Framework 3.5 SP1 is required for Excel output

# Known Issues

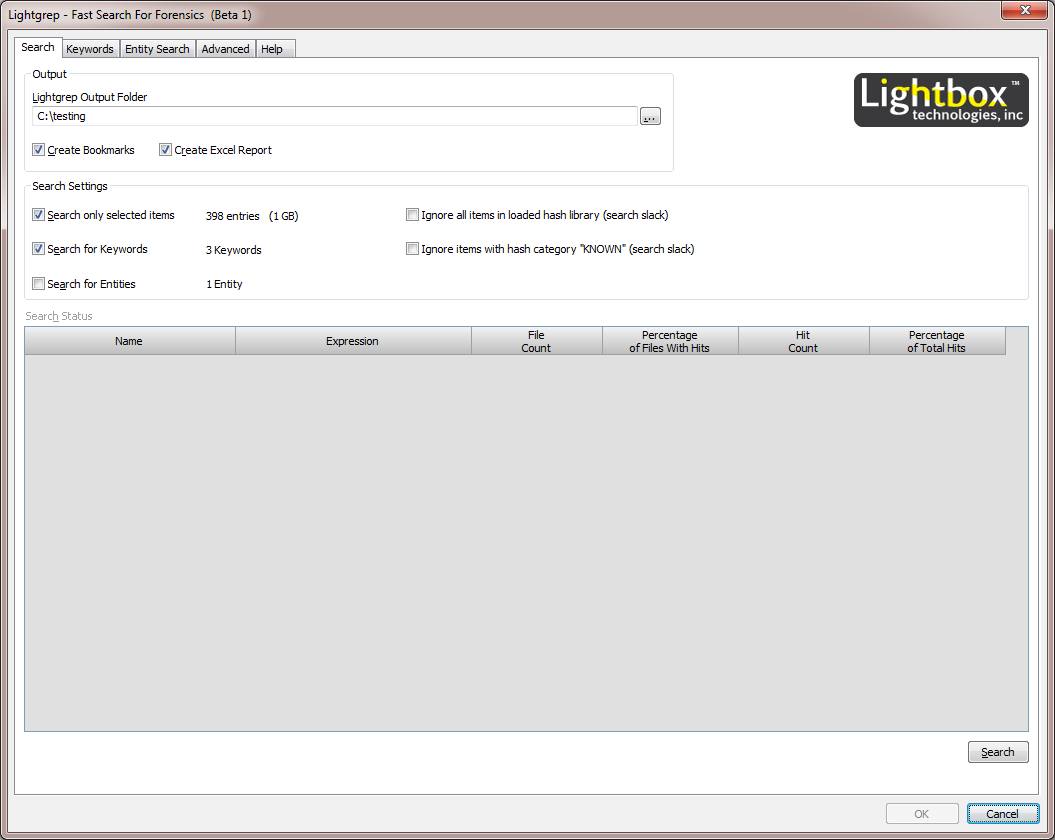
* BETA 1: The beta test is only being offered in a 64-bit version, which requires your Windows installation to be 64-bit. EnCase can be either 32-bit or 64-bit.
* BETA 1: Only one instance of Lightgrep can be running at any given time. This means that multiple searches with Lightgrep cannot be run at the same time, even if using different instances of EnCase. You will receive a warning if trying to start a second instance of Lightgrep search. Due to this limitation within the beta, you may notice a slight delay when starting the script to check for running instances. This limitation will not be present in the 1.0 release.
* BETA 1: Number of files searched as displayed in the Search Results section of the Search tab may be greater than the number of entries selected to search due to items with text extraction being counted twice (once for binary file slack, once for extracted text).
* BETA 1: Selecting the UTF-8 encoding will only search for ASCII characters. This limitation will not be present in the final release.
* BETA 1: Selecting the UTF-16 LE encoding will only search for the UTF-16 equivalent of ASCII characters. For example, for the expression “abc” Lightgrep will search for “\x00a\x00b\x00c”. This limitation will not be present in the final release.
* BETA 1: Attempting to use characters that can only be represented by UTF-8 or UTF-16 will result in a validation error. Full UTF-8 and UTF-16 support will be available in the final release.
* BETA 1: Canceling the script waits until the current entry is fully searched to complete. Thus, if searching a device with very large unallocated space, it may take a large amount time to cancel.
* BETA 1: On the Keywords tab, right-clicking and choosing “Copy Selected Items Here” will result in an incomplete copy of the keywords and an error, once the dialog closes.
* BETA 1, v1.0: Not compatible with EnCase version 7.

# Limitations and General Information

* Sweeping Bookmarks in EnCase v6 have a limit of 16,384 bytes. Although Lightgrep does not have a technical limit on the length of search hits, the limit within EnCase presents a practical limitation when the desired output format is an EnCase Bookmark.
* Entries without a starting extent cannot be bookmarked. These usually are files containing Invalid Cluster in the Description field, although some files marked as Internal, such as the streams $UsnJrnl·$J and $BadClus·$Bad, also fall into this category.
* While Lightgrep is able to search files flagged as Overwritten, bookmarks containing hits may point to the overwriting file, not to the overwritten entry that you originally selected. This is a limitation of the bookmarking function in EnCase v6.
* The “Lightgrep Output Folder” location cannot be set to a protected location in Windows Vista or Windows 7. Protected locations include, but may not be limited to, those locations represented by the following Windows environment variables: *SystemRoot*, *ProgramFiles(x86)*, *ProgramW6432* (e.g. C:\Windows , C:\Program Files, C:\Program Files (x86)). Lightgrep for EnCase will attempt to automatically identify if you are trying to save output to one of these locations and ask you to correct it.
* Lightgrep stores settings in two files: *Lightgrep Search.ini* and *Lightgrep Search.xml*. The .ini file stores simple options that have been selected in the GUI; the .xml file stores Entities and Keywords. The .xml file uses approximately 70 MB of space per 100,000 keywords.
* Lightgrep uses a constant and relatively small amount of memory while searching, but can require more memory up front while parsing the keyword list depending on the number of keywords. 4 GB of RAM with no other applications running should allow you to easily parse and search several hundred thousand fixed string keywords. Approximately 8 GB of RAM is required to parse and search 1 million fixed string keywords. When searching 1 million plus keywords, it can take several minutes before the search begins while the search engine parses the list of keywords.

# Search Tab

The Lightgrep Search Tab as shown in presents the main program options that should be reviewed before performing a search, and also provides an area to display ongoing output about search statistics.



Figure

## Options

### Lightgrep Output Folder

The folder specified here will be used both as a temporary folder location and as the output location for Lightgrep search results. Generally speaking, less than 10 megabytes of space is used by temporary files from the application, depending on the number and size of keywords. The size of results files can vary greatly depending on the number of hits found.

### Create Bookmarks

This option will create sweeping bookmarks of all search hits in the current EnCase Case. The output is very similar to EnCase search hits. Search hits will be organized into the same folder structure as present in the Keywords tree (see Figure 2). If you have a folder named “Financial” containing 5 keywords, you will also have a bookmarks folder named “Financial” which will contain one folder for each of the 5 keywords that have hits.

### Create Excel Report

This option will create an Excel 2007 format overview report, containing statistics about the results of the search. The Microsoft .NET Framework 3.5 SP1 is required for output. It is not required to have Microsoft Excel installed in order to produce the report.

### Search only selected items

This option determines whether all Entries or only selected (blue checked) Entries will be searched. Note that Records are not searched directly, but their content can be searched by selecting the appropriate Entries. Information about the number and size of selected Entries is displayed for reference.

### Search for Keywords

When this is selected, Lightgrep will search for the keywords selected in the Keywords tab. See Figure 2.

### Search for Entities

When this option is selected, Lightgrep will search for entities using the options on the Entity Search tab. Keyword hit results will be restricted based on each entity’s filter code. See the Entity Search tab description for further information.

### Ignore all items in loaded hash library (search slack)

This option will prevent searching of any item that is present in the currently loaded hash library. To learn more about loading hash sets into the active hash library, please consult the EnCase Help file sections named “Hash Sets” and “Rebuilding a Hash Library.” Slack space will always be searched regardless of this setting. Selecting this item disables the following option.

### Ignore items with hash category ‘KNOWN’ (search slack)

This option will prevent searching of any item with a hash category of ‘KNOWN’ (case insensitive) that is present in the currently loaded hash library. To learn more about loading hash sets into the active hash library, please consult the EnCase Help file sections named “Hash Sets” and “Rebuilding a Hash Library.” Slack space will always be searched regardless of this setting. This item will be disabled if the preceding option is selected.

### Search Status

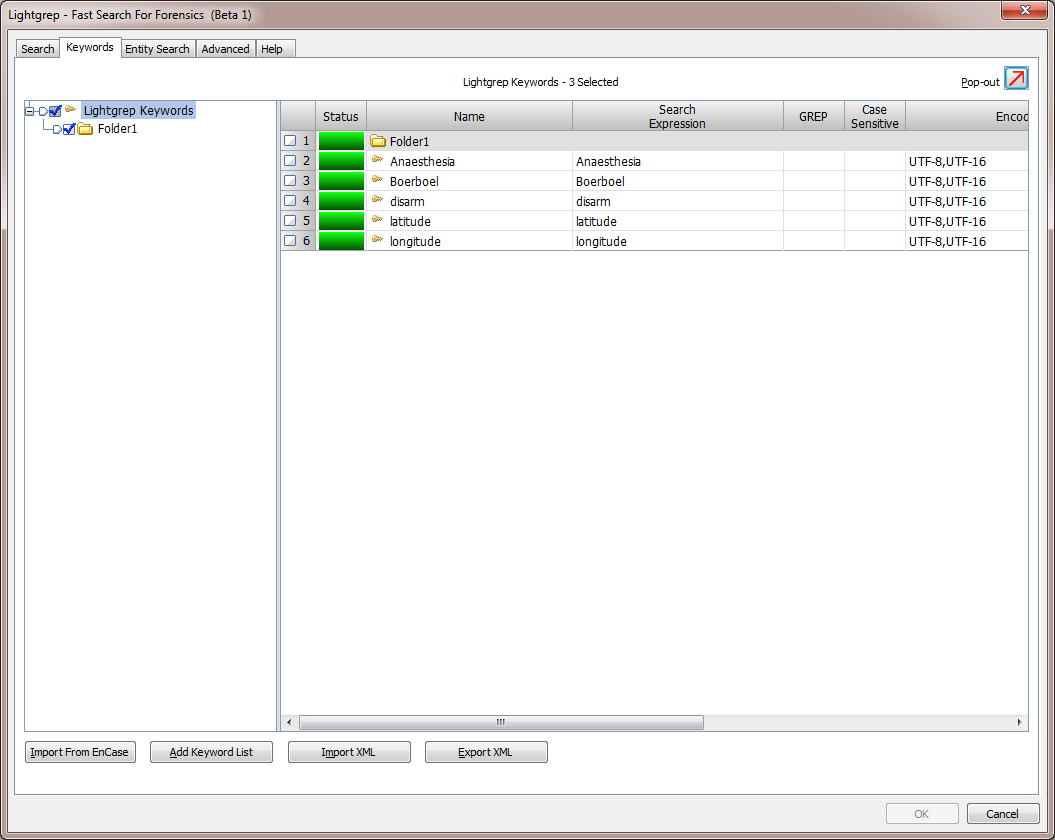
The Search Status window displays statistics about the currently running search. You will see the keyword name, search expression, a count of how many files have hits for each keyword, the percentage of files with hits, the total count of hits produced by that keyword, and the percentage of total hits related to that keyword.

### Search (Button)

The Search button starts the search using all previously selected criteria.

# Keywords

Lightgrep keywords are a little different from EnCase keywords. For one, Lightgrep keywords are PCRE compatible. Lightgrep supports more operators, longer repetition, and longer search hits. Unfortunately this means we can't use EnCase keywords, so you won't see your EnCase keywords here automatically. Supported keyword syntax is discussed in the section Supported Pattern Syntax.



Figure

## Options

### Import From EnCase

Clicking this button will open the *Import Keywords from EnCase* dialog as shown in Figure 3. The leftmost tree pane shows folders in both Global Keywords and Case Keywords from EnCase. Use this pane and the pane on the right to select (blue-check) the keywords you would like to import.

The right tree pane shows Lightgrep Keywords folders. The folder selected (highlighted) here will receive the imported keywords.



Figure

At the bottom of the dialog, a checkbox labeled “Retain folder structure when importing” allows you to keep the imported keywords organized as they were in EnCase. The folders named “\*\*GLOBAL KEYWORDS\*\*” and “\*\*CASE KEYWORDS\*\*” are not retained, and are displayed only for informational purposes. In the example from Figure 3, the root of Lightgrep Keywords is selected on the right; this will produce folders as shown below in Figure 4.



Figure

When done, click the OK button and the EnCase keywords will be imported and converted into Lightgrep Keywords. A summary of the import will be presented as in Figure 5.



Figure

### Add Keyword List

Clicking this button will open the *Add Keywords from List or File* dialog as shown in Figure 6. By default, the “List” option is selected. Copy and Paste a list of keywords, one per line to import. Then, select the proper options for the keywords. The right tree pane shows Lightgrep Keywords folders. The folder selected (highlighted) here will receive the imported keywords.



Figure

If the user selects the “File” option, they will be presented with a path selection box. Choose the path to a text file containing keywords, one per line. Then, select the proper options for the keywords. The right tree pane shows Lightgrep Keywords folders. The folder selected (highlighted) here will receive the imported keywords.



Figure

### Import / Export XML

These options allow the user to export selected keyword folders to an XML file for later import or for import on another installation of Lightgrep for EnCase. Once exported, the user can click the “Import XML” button and import the exported keywords to the folder of their choosing.

### Keyword Tree

The Keyword Tree as shown in gives a familiar interface to view, organize, edit, and delete keywords. Keywords can be organized into folders on the left (tree) side of the interface just as any other dialog within EnCase. The heading / title directly above the keyword tree shows how many keywords are currently selected. **Only keywords which are blue checked will be searched by Lightgrep.**

### Editing Keywords

The user can create a new keyword by right-clicking anywhere in the tree and selecting “New” from the context menu or edit an existing keyword either by right-clicking on the keyword and selecting “Edit” or by double clicking on the keyword. Upon creating a new keyword or opening one to edit, the user will be presented with the *Edit Keyword* dialog as shown below in . The fields in the keyword dialog are:

* Search Expression – the fixed string or GREP pattern to search for
* Name – familiar representation of the keyword
* GREP – search using regular expression pattern matching (see Supported Pattern Syntax)
* Case Sensitive – only search for the exact case shown in the expression (i.e. “test” will not hit on “Test” or “teSt” if this is selected)
* Encodings – blue check the desired encodings to search for
* Supported Patterns – displays supported syntax for GREP patterns
* Sample Keyword Hits – displays sample strings that the expression could hit on, or an error message if invalid; these are not necessarily representative of data that may be present in the currently open Case



Figure

## Supported Pattern Syntax

### Individual Bytes

***b*** where ***b*** is any literal byte except **0x00** and metacharacters

**\a** **0x07**, bell

**\e** **0x1B**, escape

**\f** **0x0C**, form feed

**\n** **0x0A**, new line

**\r** **0x0D**, carriage return

**\t** **0x09**, horizontal tab

**\*o*** **0*o***, where *o* is an octal digit

**\*oo*** **0*oo***, where each *o* is an octal digit

**\*ooo*** **0*ooo***, where each *o* is an octal digit, less than or equal to 0377

**\x*hh*** **0x*hh***, where each h is a hexadecial digit

**\c** ***c***, where c is any character not listed above

### Metacharacters

Outside of character class specifications, the following characters have special meanings: \, |, (, ), ?, +, \*, ., [. To specify a literal backslash, pipe, left parenthesis, right parenthesis, question mark, plus, asterisk, period, or left square bracket, escape it with a backslash. Additionally, literal left curly braces ({) must be escaped outside of character class specifications if they could otherwise be interpreted as the start of ranged quantifiers, but need not be escaped otherwise.

### Character Classes

[*ccspec*] matches any byte in *ccspec*  
[^*ccspec*] matches any byte not in *ccspec*

Where *ccspec* is one or more of the following:

* a byte,
* a byte escape,
* a range formed by bytes or byte escapes, or
* a named character class (excerpt .).

### Character Class Metacharacters

* The right square bracket ] must be escaped inside character class specifications unless it is the first member.
* The hyphen - forms ranges inside character class specifications using the characters immediately preceding and following it, except when - is the first or last character in the class specification or if the previous character is already part of a range, in which case - stands for itself.
* A literal \ must always be escaped.
* All other characters stand for themselves.

### Named Character Classes

**.** equivalent to **[0x00-0xFF]**  
**\d** equivalent to **[0-9]**  
**\D** equivalent to **[^0-9]**  
**\s** equivalent to **[\t\n\f\r]**  
**\S** equivalent to **[^\t\n\f\r]**  
**\w** equivalent to **[0-9A-Za-z\_]**  
**\W** equivalent to **[^0-9A-Za-z\_]**

### Repetition

***S*\*** equivalent to***S*{0,}*****S*+** equivalent to***S*{1,}*****S*?** equivalent to ***S*{0,1}*****S*{*n*,}** equivalent to***S*{*n*,*M*} *S*{*n*,*m*}** matches at least ***n*** and at most ***m*** repetitions of ***S***, greedily***S*\*?** equivalent to***S*{0,}?*****S*+?** equivalent to***S*{1,}?*****S*??** equivalent to***S*{0,1}? *S*{*n*,}?** equilvalent to***S*{*n*,*M*}?*****S*{*n*,*m*}?** matches at least *n* and at most *m* repetitions of *S*, reluctantly

Where ***S*** is an atomic pattern, 0 ≤ ***n*** ≤ ***m***, and ***M*** is a largish integer which presently depends on the size of the function stack on the machine where Lightgrep is run.

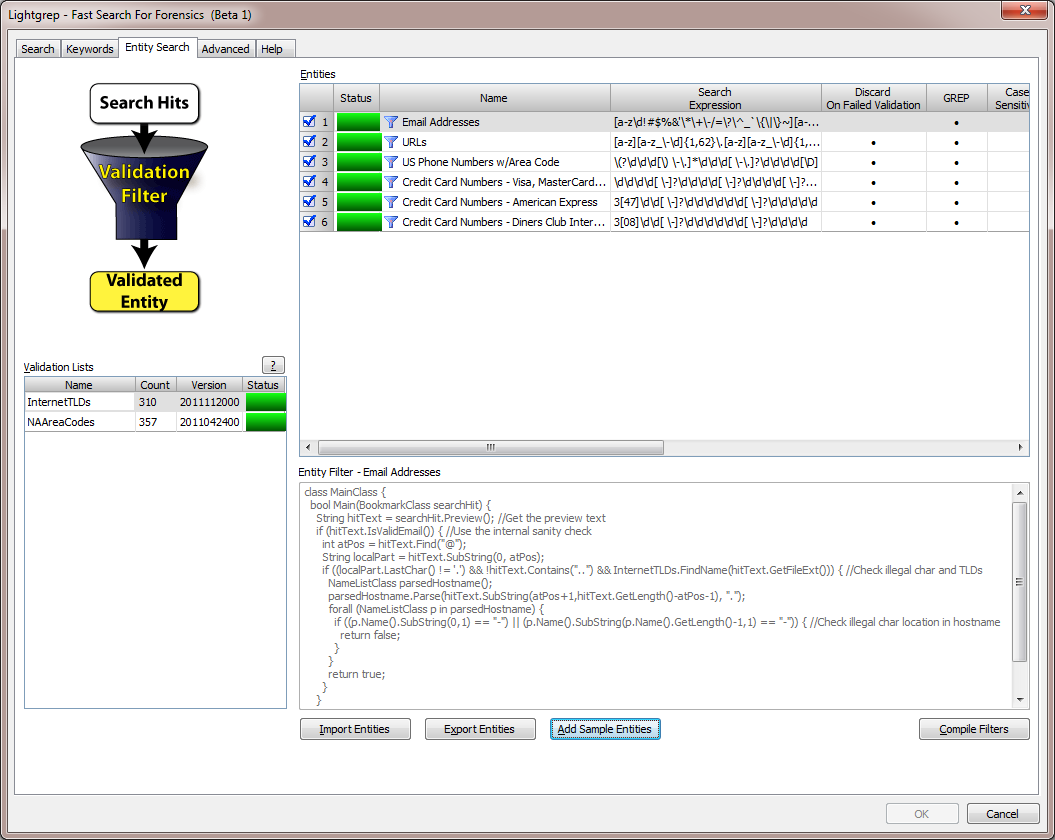
### Conjunction, Alternation, Grouping

***ST*** matches something matching ***S*** followed by something matching ***T******S*|*T*** matches ***S*** or ***T***, preferring matches for ***S***  
**(*S*)** equivalent to ***S***, but is atomic

Individual bytes, character classes, and groups are atomic.

# Entity Search

Entity search looks for GREP keywords and then uses filters to verify the results as shown in Figure 9. The default filters compare Email addresses and URLs to the list of approved top-level domains to insure validity; check US phone numbers for valid area codes; and compare credit card numbers using Luhn's algorithm. If "Discard On Failed Validation" is selected, the search hit will be ignored completely if the filter returns false, otherwise it will be noted in the bookmark comment. Be aware that filtering search hits for entities can take some time, if the filter code is involved and there are lots of hits. The productivity gain from reducing false positives, however, is well worth the slightly extended filtering time.



Figure

## Options

### Entities

Entities use the same expression rules as regular Lightgrep keywords. The user can add new entities, edit and delete existing entities in the same manner as keywords using right click controls.

### Entity Filter

This read-only text box displays the filter code for the currently selected Entity. The user must edit the Entity in order to change the filter code.

### Validation Lists

Validation Lists are NameListClass objects (see the EnScript Help within EnCase for an example of working with NameListClass types) that are pre-populated with useful lists of information. For example, the 'InternetTLDs' list is a NameListClass containing every IANA approved top level domain, such as com, edu, uk, etc. The list count and version (a string typically representing the date of the most recent update) are displayed. Double-click on a Validation List name to view its contents.

Validation Lists can be used as NameListClass objects directly in the Main() function of entity filters without initialization:

NameListClass comDomain = InternetTLDs.FindName("com");

Validation Lists are local to the Main() function; they cannot be used outside of Main() without being passed into other functions.

Current Sources:

InternetTLDs  
List of valid Internet top level domains, maintained by IANA (com, edu, uk, etc)  
http://data.iana.org/TLD/tlds-alpha-by-domain.txt

NAAreaCodes  
List of active North American area codes, maintained by the North American Numbering Plan Administration  
http://www.nationalnanpa.com/nas/public/npasInServiceByNumberReport.do?method=displayNpasInServiceByNumberReport"

### Import / Export Entities

These options allow the user to export entities to an XML file for later import or for import on another installation of Lightgrep for EnCase. Once exported, the user can click the “Import Entities” button and import the exported.

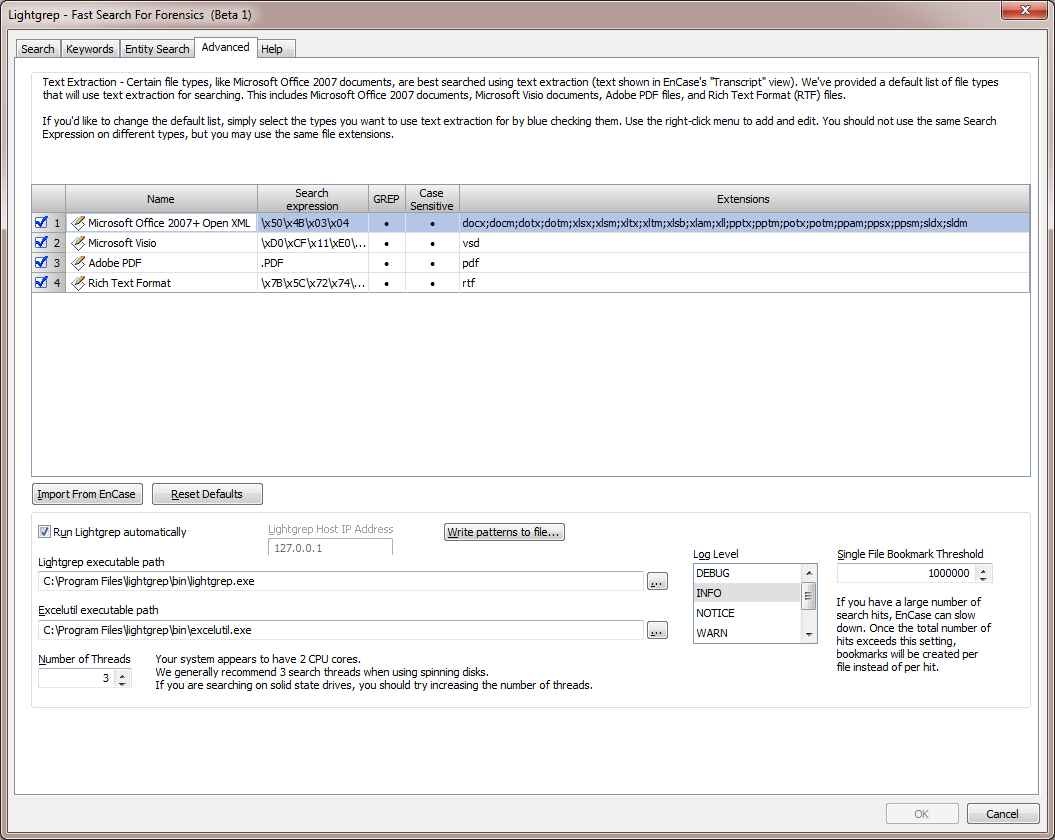
### Add Sample Entities

Clicking this button will allow the user to add the sample entities shown in Figure 9. Sample entities are provided as a starting pointing and demonstration of the power of entity filtering.

# Advanced Tab

Text Extraction - Certain file types, like Microsoft Office 2007 documents, are best searched using text extraction (text shown in EnCase's "Transcript" view). We've provided a default list of file types that will be used text extraction for searching. This includes Microsoft Office 2007 documents, Microsoft Visio documents, Adobe PDF files, and Rich Text Format (RTF) files.

If you'd like to change the default list, simply select the types you want to use text extraction for by blue checking them. You should not use the same Search Expression on different types, but you may use the same file extensions.



Figure

## Options

### File Signature List

The File Signature List, as shown in the top portion of Figure 10, contains a list of file types that will be used to identify files for text extraction. This list contains standard EnCase file signatures. You can add or edit file signatures just as you would in EnCase.

During a search, any file types selected here will have text extracted using EnCase’s Oracle Outside-In technology. Lightgrep will search the extracted text for keywords instead of the binary content of the file. It will also search binary file slack of any files identified for text extraction, so you don’t have to worry about missing anything.

If you deselect any types, you will receive a warning: “You have deselected one or more file types. Any file type that is not selected will be searched using standard native content instead of extracted text.”

### Import From EnCase

Clicking this button will open the *Import File Signatures from EnCase* dialog as shown in Figure 11. The dialog shows the standard EnCase File Signatures list. Use this window to select (blue-check) the file signatures you would like to import. When done, click the OK button and they will be imported into Lightgrep.



Figure

### Reset Defaults

Clicking on the Reset Defaults button will remove any custom file signatures that have been created or imported from EnCase and replace them with the default values as shown in Figure 10.

### Run Lightgrep automatically

This setting should always be checked when performing searches. The user should only change this option at the instruction of Lightbox Technologies support for the purposes of debugging.

### Lightgrep executable path

The path to lightgrep.exe on the examiner system. The Lightgrep search executable is run using this path. Usually *C:\Program Files\lightgrep\bin\lightgrep.exe*.

### Excelutil executable path

The path to excelutil.exe on the examiner system. The excelutil executable is run using this path to output the Excel report. Usually *C:\Program Files\lightgrep\bin\excelutil.exe*.

### Number of Threads

The script will attempt to auto-detect how many cores you have on your box. Regardless of the number of cores, we generally recommend 3 search threads when using spinning / magnetic hard disks. If you are searching on solid state drives, and have more than 2 cores, you can try increasing the number of threads.

### Log Level

This represents the level of detail provided in the EnCase Console log. Usually this should be set to WARN. Setting this to DEBUG will cause an extremely large number of log events to be written to the Console.

### Single File Bookmark Threshold

If you have a large number of search hits, EnCase can slow down significantly. Once the total number of hits exceeds this setting, bookmarks will be created per file instead of per hit. The default value is *1,000,000*.

# Starting a Search

Once the user clicks the “Search” button on the main Search tab, the search process will begin. The first thing the user will see in the bottom status bar is “Initial filtering” as shown in Figure 12.



Figure

The “Initial filtering” process consists of several phases:

* Writing out the selected keyword and entity expressions to the file *keywords.txt*
* Filtering out entries with a physical size of 0 and also files which are completely sparse
* Sorting the entries selected for search so as to maintain consistently high usage of the tool’s multithreading capabilities
* Checking each individual entry to determine if it should be searched using extracted text (transcript)
* Writing the list of entries to be searched to the file *EntryList.txt*

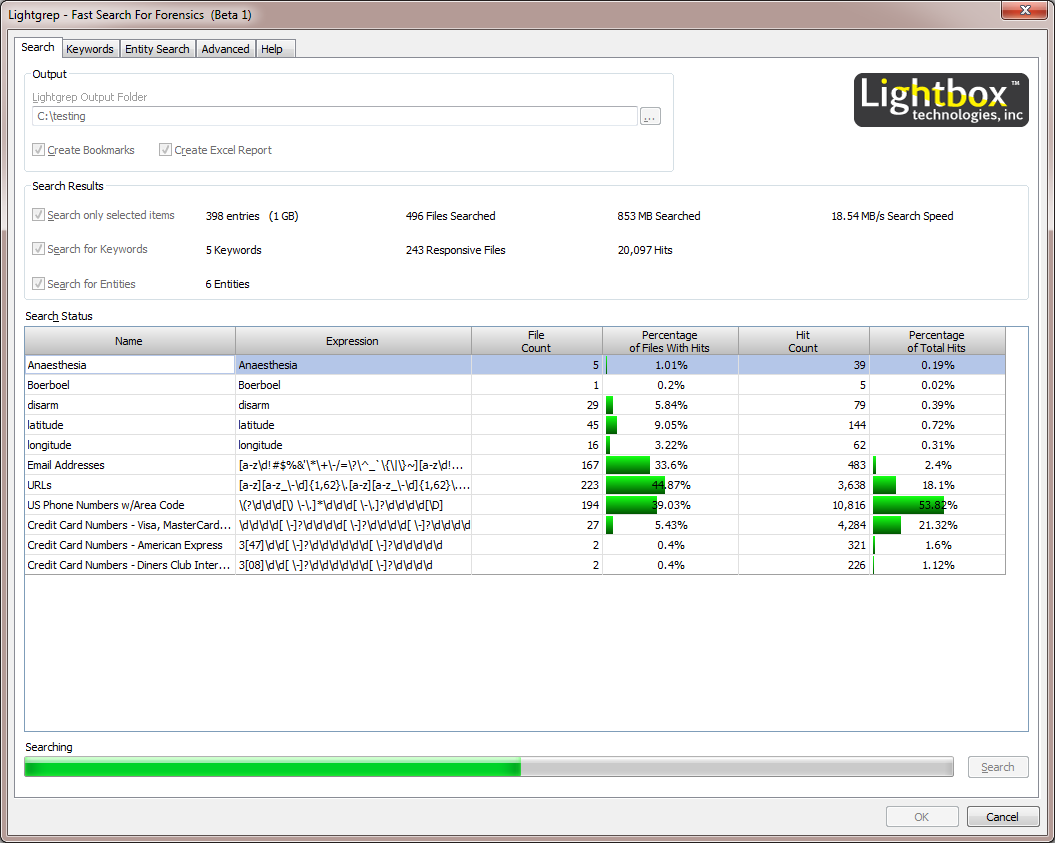
This process can take anywhere from a few seconds to a few minutes, depending on how many entries are selected, and which options are chosen by the user.

Clicking the “Cancel” button at any time during the search will cause the entire search to be aborted.

# Monitoring Search Progress

Once the search begins, the user can monitor the progress of the search through the main Search tab as shown in Figure 13. The following fields will be displayed during a search:

* Keyword or Entity name
* Search Expression
* File Count – the number of files which are responsive to each keyword
* Percentage of Files With Hits – of the total files searched, what percentage contained hits for each keyword
* Hit Count – the number of search hits produced by each keyword
* Percentage of Total Hits – of the total search hits found for all keywords, what percentage of hits were produced by each keyword



Figure

# Reviewing Output

When the search has completed, the “OK” button will be enabled. Clicking “OK” will save sweeping bookmarks containing the search hit results. If “Create Excel Report” was selected, the Excel overview will also be created. Clicking “Cancel” will end the search without saving bookmarks or the Excel report.



Figure

## Bookmark Output

Once the bookmarks have saved, you can review them just as you would other search hits. Figure 15 shows example bookmark output from the URLs entity search.



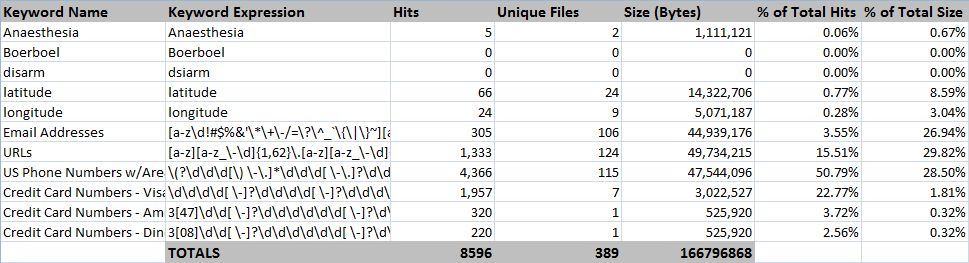
Figure

## Excel Output

The Excel report provides an overview of results by displaying charts of the top ten keywords by number of hits and file size.

|  |  |
| --- | --- |
|  |  |

A keyword frequency report shows all keywords that were searched, and the resulting hit statistics. These results are identical to those in the Search Status box which are displayed during the search.



The DataCategories tab shows all file extensions for search hits, with the number of files and size of files represented by each file extension. Files with no extension are denoted by the asterisk (\*).



The Excel report worksheets are protected so the user doesn’t accidentally alter the data. If you wish to edit the report, simply go to the Review tab in the Office Ribbon as shown in Figure 16, and click “Unprotect Sheet.” No password is needed.



Figure

1. EnCase® is a registered trademark of Guidance Software, Inc. [↑](#footnote-ref-2)