

Spider documentation for Tier 2

Background.....	1
Effectiveness.....	2
Installation.....	2
Running Spider.....	3
Using the graphical interface.....	3
Running from the command line.....	6
Configuration.....	7
Runtime configuration.....	8
File handling configuration.....	9
Regex (search) configuration.....	13
Logging configuration.....	16
Advanced configuration.....	19
Variables available within Spider.....	20

This document is intended as instructions on the operation of the Spider software developed by Cornell University. ITS makes no representations, warranties, or covenants whatsoever as to (i) the positive or negative effect of the software on the operation or the security of any particular network, computer system, network device, software, hardware, or any component of any of the foregoing or (ii) the accuracy, reliability, timeliness, or completeness of the documentation. ITS is providing the documentation “as is” and “as available” without representations, warranties, or covenants of any kind.

Background

Spider is an application written at Cornell University for the specific purpose of searching computers for sensitive data (see official website at: <http://www.cit.cornell.edu/computer/security/tools/>). Cornell has created versions for Windows (works under Windows 2000, XP, and Server 2003), Mac OS X and Linux, but this documentation focuses on the Windows version.

Spider searches through the content of all files looking for matches to the regular expressions it is given (see http://en.wikipedia.org/wiki/Regular_expressions for more information on regular expressions). This can be a very time, disk and CPU intensive activity (i.e. it might consume your computers resources for several hours), so configuring Spider to be most efficient is important.

With the permission of Cornell, ITS has repackaged and redistributed Spider with a custom default configuration. This configuration includes a list of file types to skip (that either would not contain sensitive data, or which format data in a way that Spider cannot interpret), some improved search parameters, and changes to the way log files are created. The ITS configuration is designed with end-users and “one-click” use in mind, but the packaged settings are also useful for IT administrators.

Effectiveness

Important: Spider is not 100% effective. There are many file types that store information in a way that Spider cannot interpret and there are limits to the search commands themselves. The ITS distribution attempts to strike a good balance between effectiveness and usability.

The ITS Spider distribution CAN detect the following items in the default settings:

- Social security numbers (SSN) in the formats “nnn-nn-nnnn” or “nnn nn nnnn” with a beginning digit of 0-7 (beginning digits of 8 and 9 are not valid SSN’s)
- Social security numbers in the format “nnnnnnnnn” issued in Colorado (beginning with 521-524 or 650-653). This format is limited to Colorado SSN’s because strings of numbers without separators are very common and will generate many false positives without restrictions. Because Coloradoans are a major group at the university in both students and employees, most collections of SSN’s will contain some Colorado issued ones.
- Credit card numbers in hyphenated form “nnnn-nnnn-nnnn-nnnn” that begin with Visa, Mastercard or Discover Card starting digits.
- Credit card numbers in hyphenated form “nnnn-nnnn-nnnn-nnn” that begin with American Express starting digits.
- Credit card numbers in the form “nnnnnnnnnnnnnnnnn” that begin with the starting digits for ACards, University US Bank travel cards, Elevations Credit Union (formerly U of C Federal Credit Union) and DiscoverCard. As with SSN’s, the non-separated form must be tightly restricted to minimize false positives, but it will also not detect all numbers, just common ones.

These items can only be detected if they are stored in ASCII format in a scanned file. The following file extensions are skipped in the ITS Spider package: WMA M4V MP3 SYS VHD M4A MDF TIF ISO AVI EXE BMP MOV GIF CAB M4P JPG WMV MPG PNG AI OGG VMDK DLL PSD WAV MSI EX_DL_MSP JAR NGR DLM AM WMF PDF

It is still possible to have sensitive data in files of these types (e.g. having an image of a social security number in a JPG file), but Spider cannot detect such pieces of data.

Note: Some file formats, like Microsoft Excel, may store information in either a form readable by Spider or not, depending on how the application is used. In most observed Excel files, the data contained in cells was searchable by Spider. While Adobe Acrobat (PDF) files contain text searchable within a PDF viewer, ITS investigation of a variety of PDF files from different sources showed that none stored the text in a format searchable by Spider.

Installation

To install Spider, you will first have Microsoft .NET 1.1 or later installed, and you must be logged in as a user with local administrator rights. Spider can run with regular user rights, but installation and configuration changes require administrator rights.

Exercise	
Module 1	Exercise 1

To install Spider, download the ITS Spider installer and run it. It will install Spider, create a Start Menu group, and set the ITS custom default settings. The installation will not require a reboot and Spider is ready to use once the installation is complete.

Exercise	
Module 1	Exercise 2

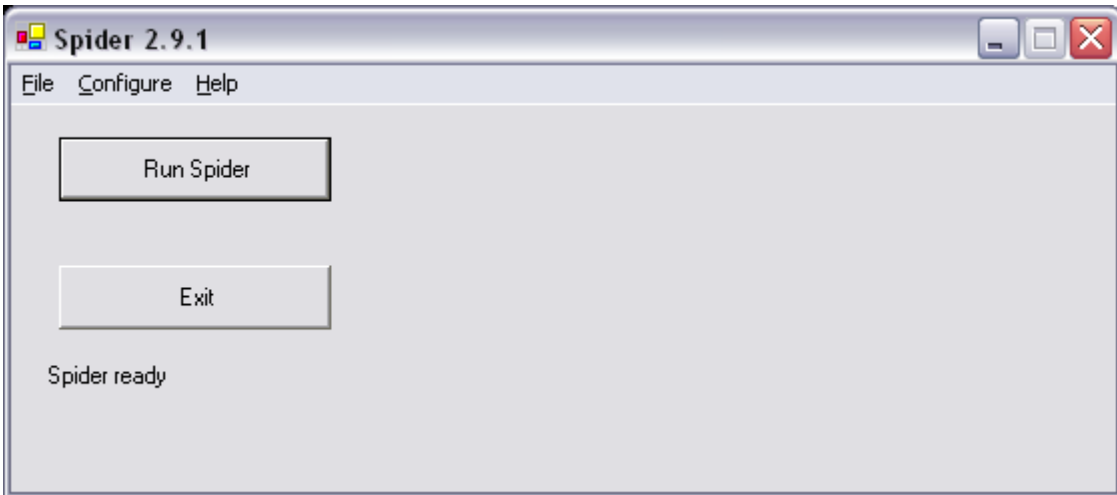
Running Spider

Spider has both a typical interactive GUI mode as well as a “silent” mode that will execute a scan with the current configuration without displaying the interface or interacting with the user. Most of this document addresses the GUI mode, but the command line mode is discussed in the last section.

Note on using Spider on a mapped drive: Spider can be used to scan any drive, including mapped drives across a network. If you choose to scan a drive across a network, keep in mind that for Spider to work it must transfer the entire contents or all scanned files across the network from the target system to the scanning system. This is not only very inefficient, it also transfers data possibly containing sensitive information using a typically unencrypted protocol. For these reasons, ITS recommends running Spider directly on the systems you wish to scan.

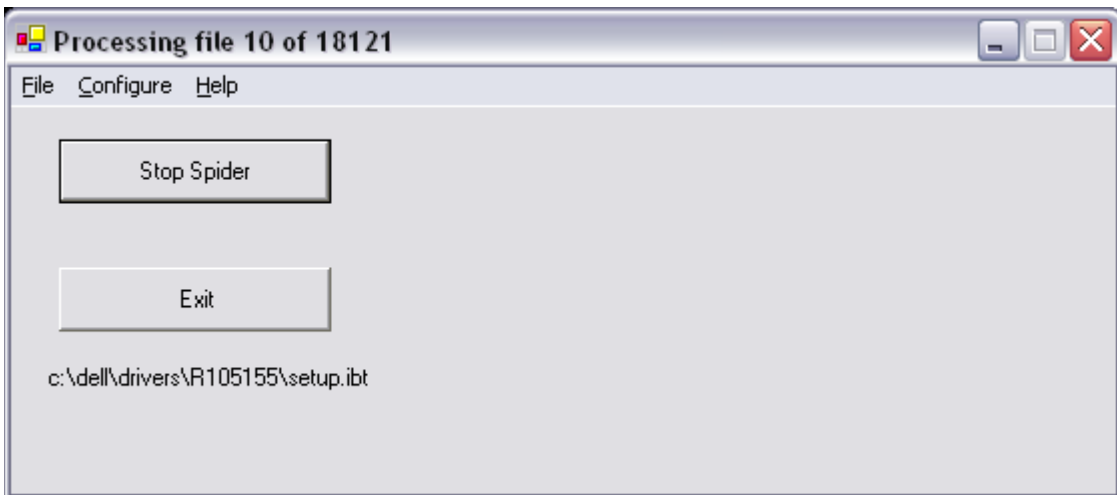
Using the graphical interface

To run Spider, select the start menu, then “Programs” or “All Programs”, then the “Cornell Spider” folder, then click on “spider”. The following window will appear:



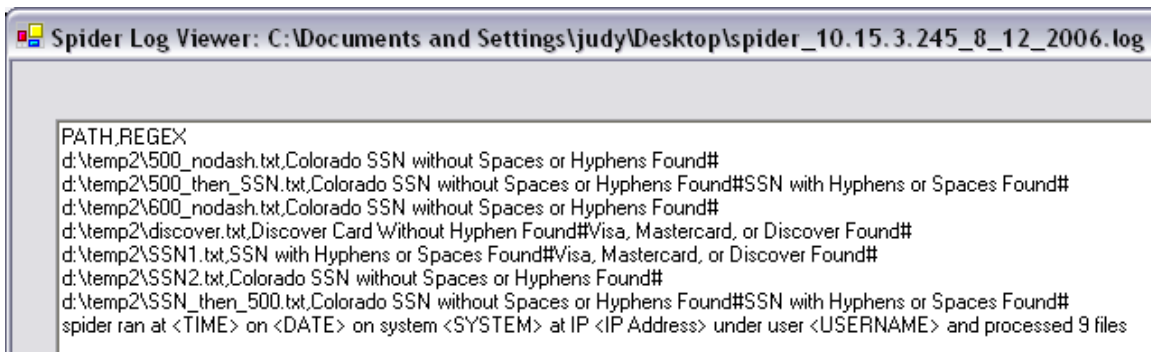
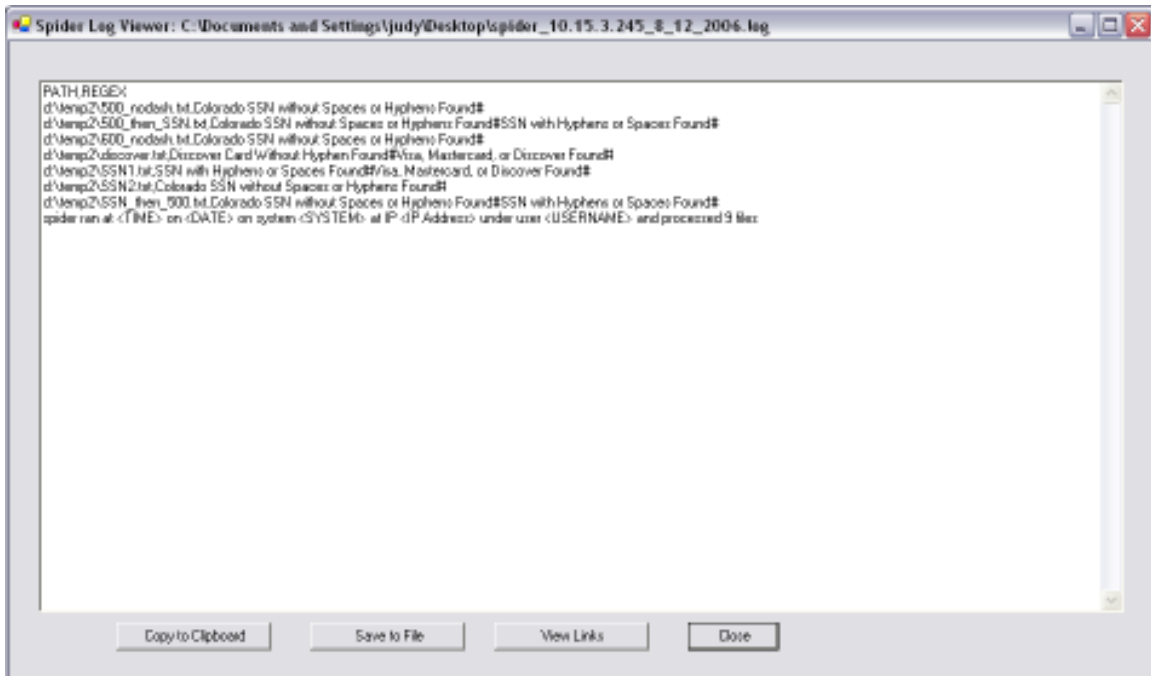
The two buttons on this window operate just as they are labeled. “Run Spider” will start a scan with the current configuration and “Exit” will close Spider.

When you start a scan by clicking “Run Spider”, the window will change to display the file currently being scanned and the “Run Spider” button becomes a “Stop Spider” button, as seen here:



Once a scan is complete, Spider will automatically open the resulting log file. In the ITS distribution, the log file is saved to the current user’s desktop and is named “spider_<IP address>_<Date>.log”. The name and location of the log file can be changed in Spider’s configuration options.

Exercise	
Module 2	Exercise 1
Module 3	Exercises 1,2



The log file shows the file location and name for each match, along with the name of the match criteria. There may be more than one type of match in a single file and two match names will be shown (as seen in multiple lines above). The possible match criteria names and meanings in the ITS package are:

- Colorado SSN without Spaces or Hyphens Found – a Colorado issued social security number in the format nnnnnnnnn
- SSN with Hyphens or Spaces Found – a social security number in the format nnn-nn-nnnn or nnn nn nnnn
- Visa, Mastercard, or Discover Found – a Visa, Mastercard or Discover Card in the format nnnn-nnnn-nnnn-nnnn
- American Express Card Found – an American Express card in the format nnnn-nnnn-nnnn-nnn
- Discover Card Without Hyphen Found – a Discover Card in the format nnnnnnnnnnnnnnnnn

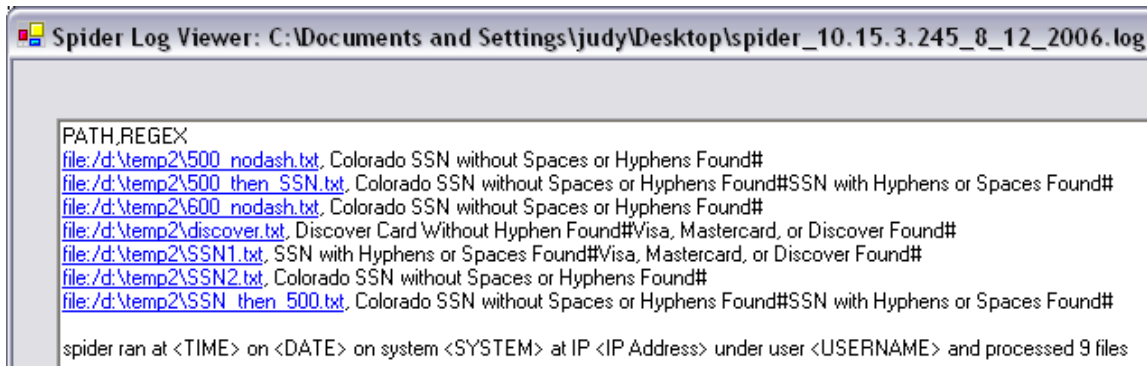
- Elevations Card Found – an Elevations Credit Union card in the format nnnnnnnnnnnnnnnnnn
- CU Acard Found – a university Acard in the format nnnnnnnnnnnnnnnnnn
- CU USBank Card Found – a US Bank card (issued for university travel expenses) in the format nnnnnnnnnnnnnnnnnn

The last line of the log gives a summary of important information including:

- Date and time of the scan
- Computer name and IP address
- Username that the scan ran under
- Total number of files scanned

Both the information logged for the matches and the information contained in the last line can be changed in the Spider configuration options.

The “View Links” button will convert the file path into a clickable link to open the file:



This can be very useful in verifying hits, but only works on the computer that Spider ran on (i.e. if you open a log file on another computer, the links will not point to the scanned files).

Running from the command line

Spider has several command line switches that allow it to be run without user intervention and even as a scheduled task. Note that encrypting log files is not compatible with an automated scan.

The Spider command syntax is as follows:

spider.exe /options

and the options are:

/R: [path or file]

causes Spider to assume the supplied path is its start directory and recursively scan it. If provided a file instead, Spider assumes it contains a list of paths, one to a line, that it should sequentially process.

/D: [path or file]

causes Spider to process the supplied path non-recursively, i.e., without descending into subdirectories. If supplied a file, Spider will scan that file and quit.

/run

Spider will start and run unattended. This option will use the current settings as configured in the GUI, except for the starting path and log path options also given in the command line. Make sure Spider is properly configured before executing an unattended scan.

/L: [path]

write the log file to the specified path. This allows Spider to take advantage of Windows environment variables to create machine-specific log paths, for example: “c:\path\to\spider\spider.exe /L: z:\%COMPUTERNAME%.log” will create Spider logs based on the system's host name.

Spider will accept conventional Windows paths (c:\foo) or UNC paths (\\PC1\C\$). Running from the shell also accepts Windows environment variables like %COMPUTERNAME%.

For example, the following command:

```
Spider.exe /R: c:\test /L: c:\%COMPUTERNAME%.log /run
```

Will start an unattended, recursive scan of c:\test and write a log file to c:\<computer name>.log.

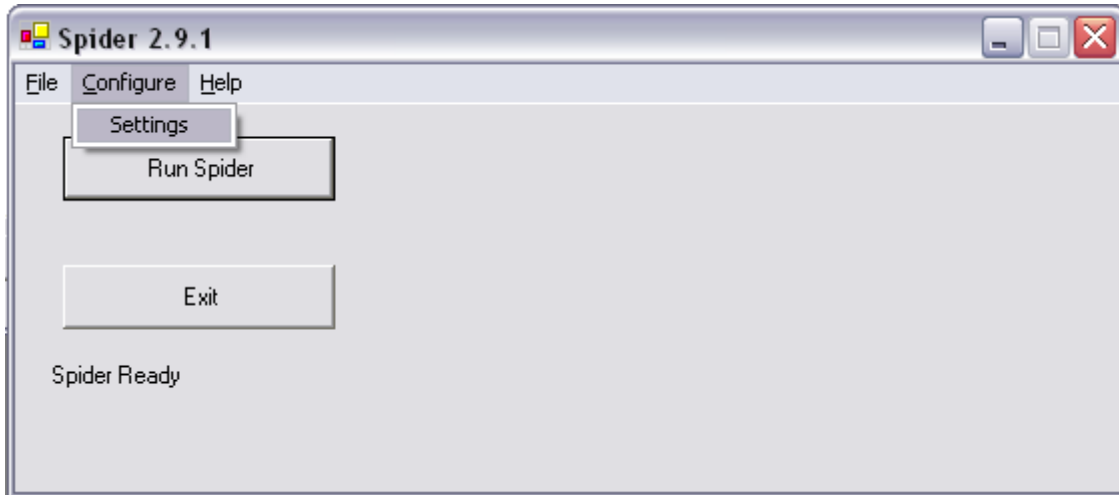
Configuration

Spider has numerous configuration options and the online documentation covers most options well (<http://www.cit.cornell.edu/computer/security/tools/spider-windows.html>).

Changing Spider's configuration requires local administrator level access because of the registry location used to store the settings. This also prevents users from changing the configuration in managed environments.

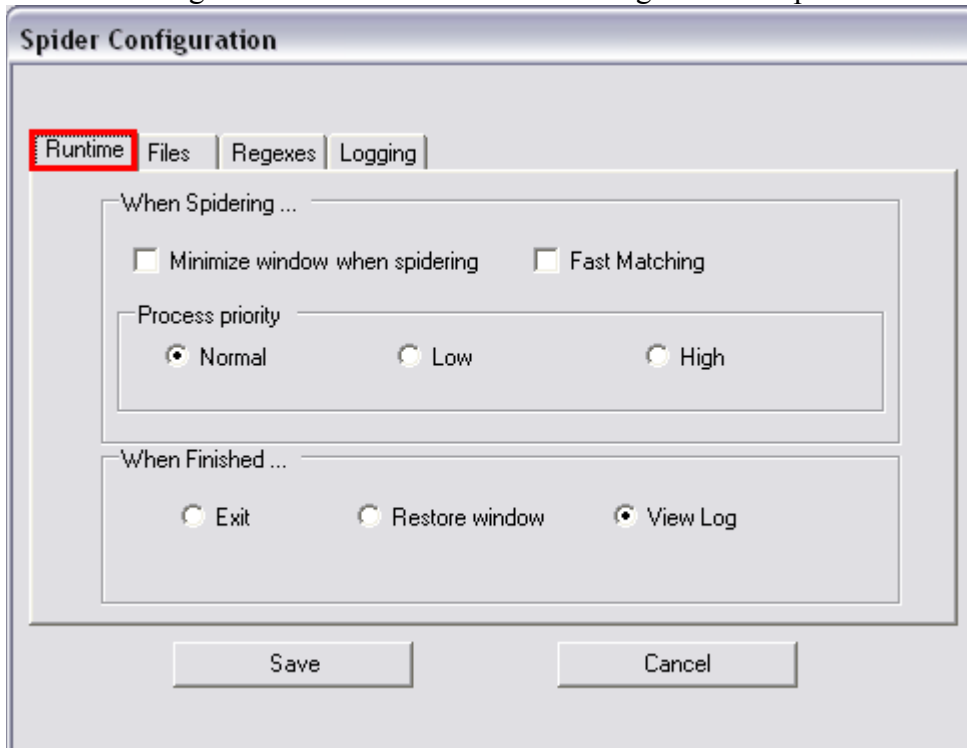
Spider configuration information is stored in the Windows registry at “HKEY_LOCAL_MACHINE\SOFTWARE\Cornell University\Spider” so configuration of Spider can be done by exporting this key on a configured computer and then importing it onto other computers.

Spider has four major configuration areas: runtime, files, regex and logging, and each is described below. To access all of the configuration options, select Configure -> Settings from the Spider menu



Runtime configuration

The first configuration tab is “runtime” and configures how Spider runs.



Minimize window when spidering –

When enabled, causes the Spider window to minimize while Spider is working. **Be aware that if you do other work on the same computer while Spider is running, Spider may not be able to scan files that are in use by other applications.**

Fast Matching –

When enabled, Spider will stop searching a particular file when the first match is found. This speeds up scanning, but will not alert you to the presence of other potential hits (like both an SSN and a credit card number in a file).

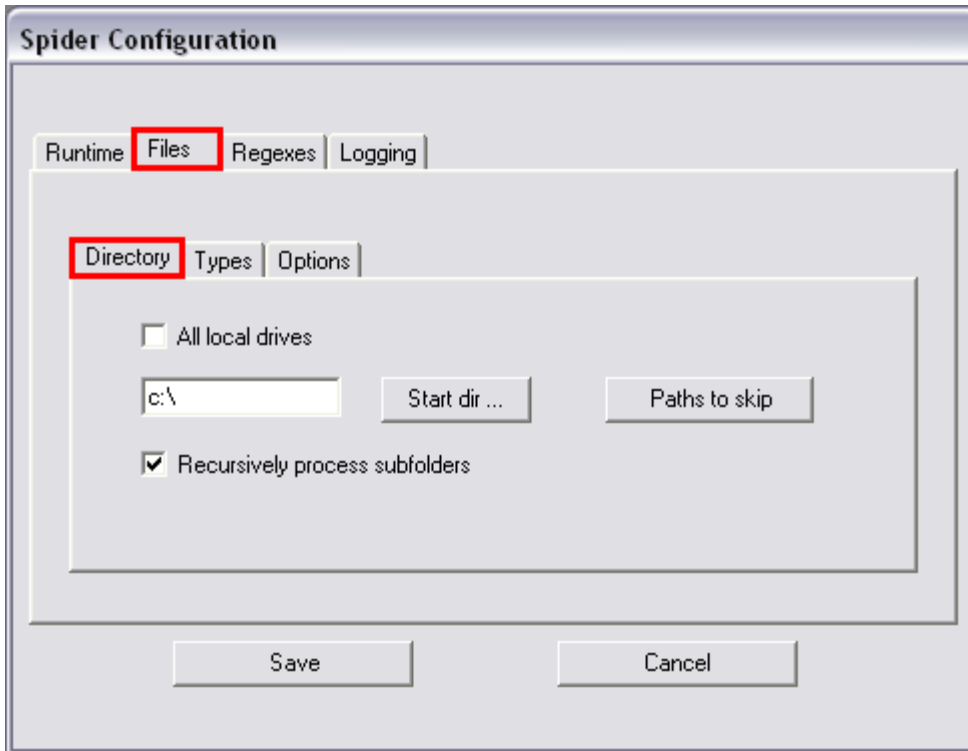
Process priority

- Normal: Spider allows Windows to assign a priority to its file processing task.
- Low: Spider will ask Windows for a lower process priority while scanning files. This preserves resources and is less disruptive while the system is in use.
- High: Spider will ask Windows for a higher priority while scanning files. This will cause Spider to complete its task sooner, but with noticeable impact on system responsiveness

When Finished

- Exit: Spider will exit immediately after it is done scanning files.
- Restore window: Spider will restore itself from a minimized state when it is done scanning files.
- View Log: Spider will spawn the log viewer and display its scan results when it is done scanning files.

File handling configuration



Files tab → Directory tab

All local drives:

When enabled, tells Spider to systematically process all local drives, starting with the root directory. This option might cause Spider to error and stop scanning when an empty floppy drive is present. Empty CD drives do not appear to be a problem.

or - (if you click the All local drives box, the option below will gray out)

Start dir:

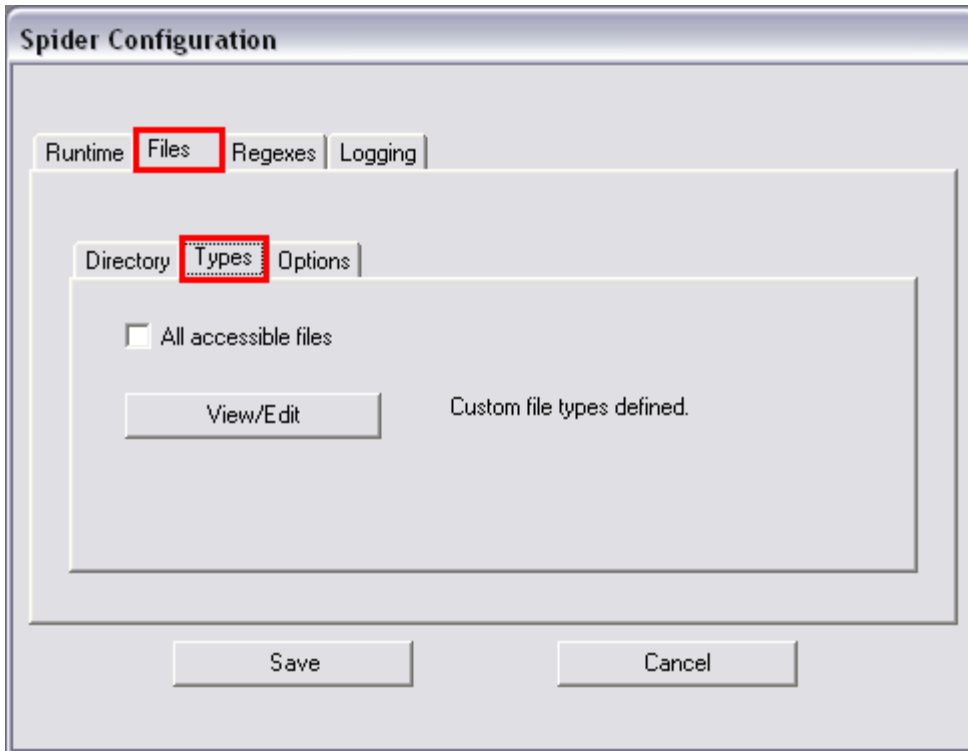
tells Spider to start processing with the selected directory, initially the C drive. To specify the drive or directory of your choice, you can either type in the white box, or click the Start dir... button, browse, and then click OK.

Paths to skip:

Lists folders that you do not need Spider to scan. This button opens a new window, where you click Add New for each path you want Spider to skip.

Recursively process subfolders:

When enabled, tells Spider to descend into directories it finds and process the files there. This is Spider's standard behavior and is recommended in most circumstances.



Files tab → Types tab

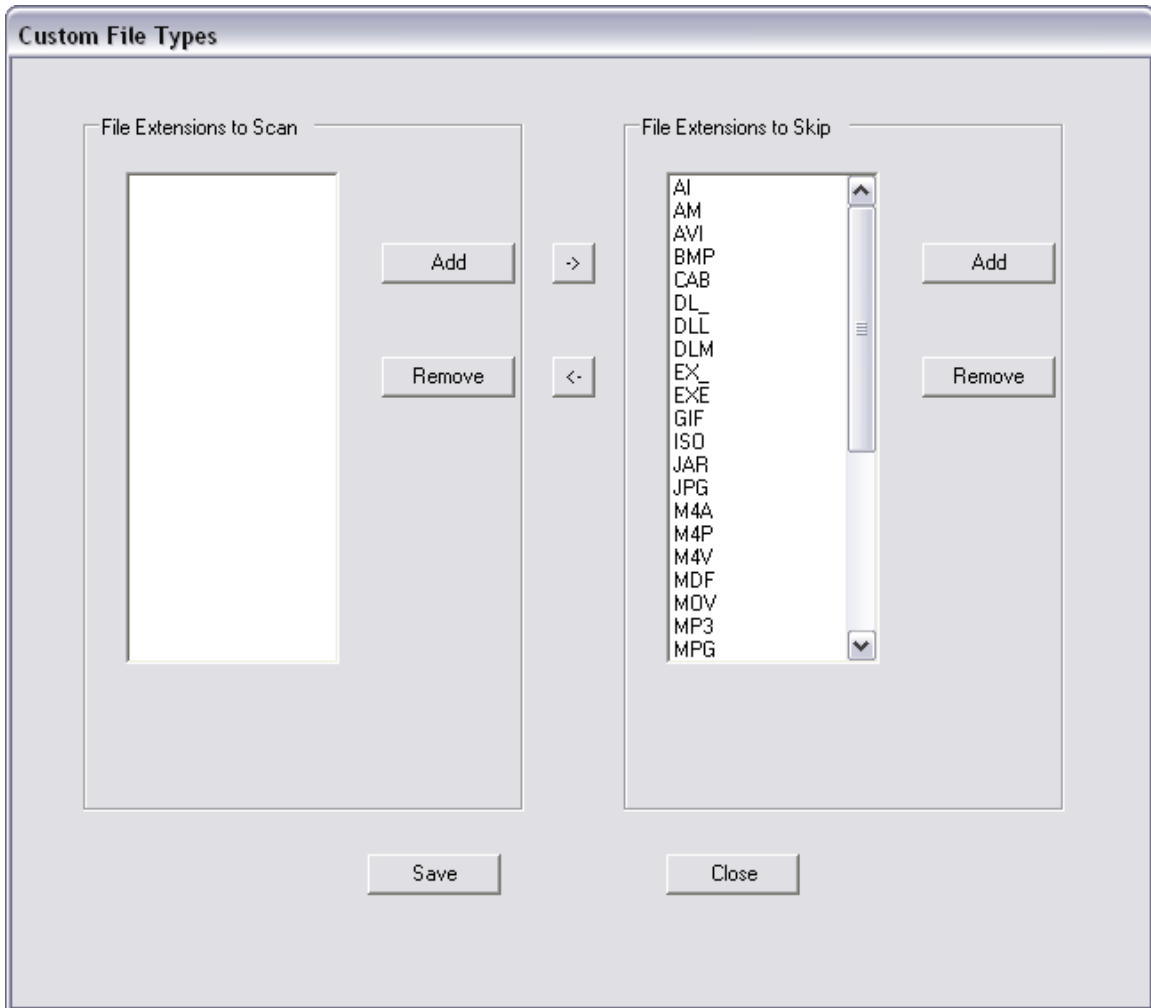
All accessible files:

When enabled, tells Spider to process any file it can successfully read, ignoring the custom list of file types.

- or -

View/Edit: allows you to specify file extensions that should explicitly be scanned or ignored.

1. Uncheck the "All accessible files" box to make the "View/Edit" button clickable.
2. Click the View/Edit button.



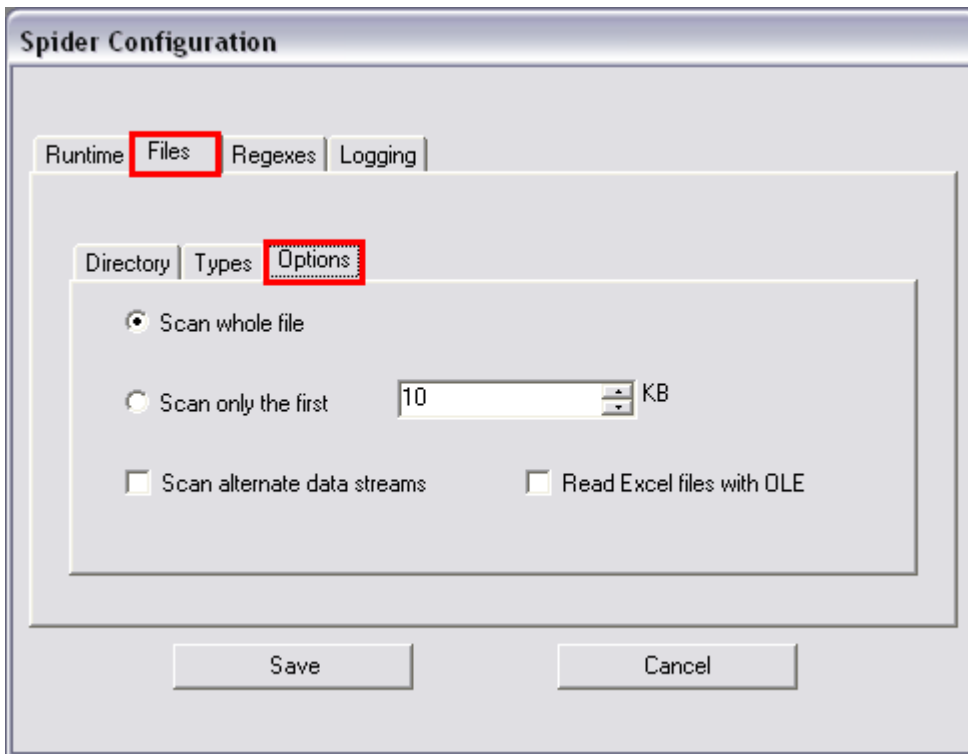
3. You will see two lists, "File Extensions to Scan" and "File Extensions to Skip." Click the Add button next to either list.
4. A popup window directs you to enter a file extension with no punctuation; for example, TIFF with no period. Do so and click OK.
5. Repeat for all the filetypes you wish to specify.
6. Click Save when done.

The following file extensions are skipped in the ITS Spider package: WMA M4V MP3 SYS VHD M4A MDF TIF ISO AVI EXE BMP MOV GIF CAB M4P JPG WMV MPG PNG AI OGG VMDK DLL PSD WAV MSI EX_ DL_ MSP JAR NGR DLM AM WMF PDF

It is still possible to have sensitive data in files of these types (e.g. having an image of a social security number in a JPG file), but Spider cannot detect such pieces of data.

Note: Some file formats, like Microsoft Excel, may store information in either a form readable by Spider or not, depending on how the application is used. In most observed Excel files, the data contained in cells was searchable by Spider. While Adobe Acrobat

(PDF) files contain text searchable within a PDF viewer, ITS investigation of a variety of PDF files from different sources showed that none stored the text in a format searchable by Spider.



Files tab → Options tab

Scan whole file:

When enabled, tells Spider to process each file from beginning to end.

Scan only the first XX KB:

When enabled, tells Spider to move on to the next file if no matching patterns are found in the first XX kilobytes of the current file. This increases the speed of the scan, but also increases the likelihood of a false negative.

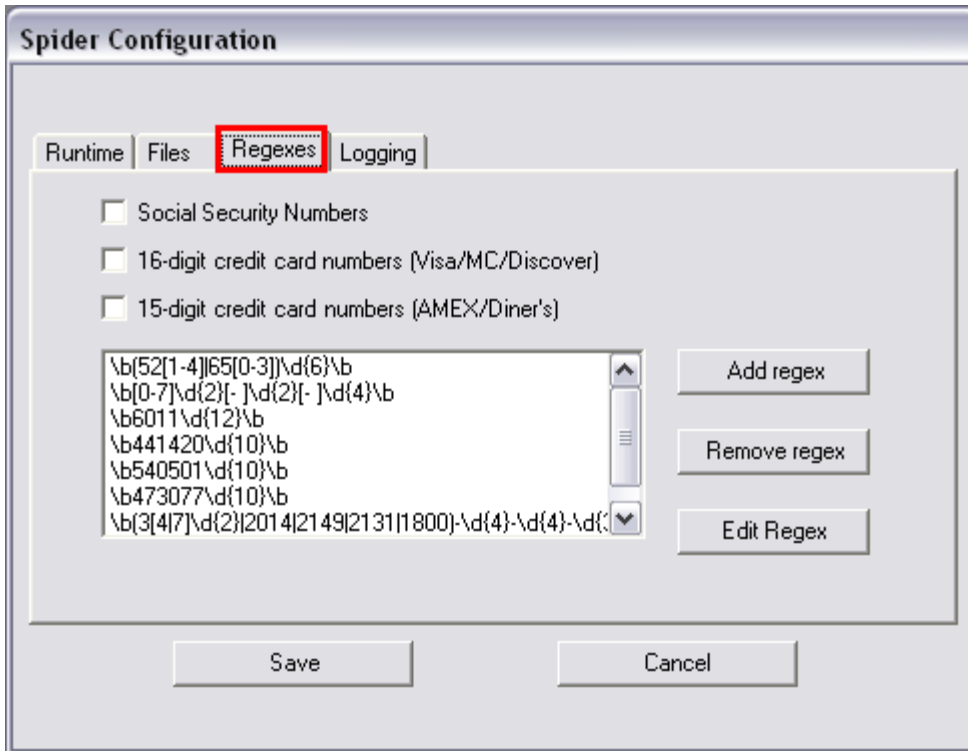
Scan alternate data streams:

When enabled, tells Spider to search each file for NTFS alternate data streams and scan those. Alternate data streams can be attached to normal files but hidden, which is a security risk.

Read Excel files with OLE:

When enabled, tells Spider to try to scan Excel files that contain content embedded from another program with Object Linking and Embedding (OLE). This does not affect most Excel files.

Regex (search) configuration



The default regex searches in the ITS distribution are:

- Visa, Mastercard, or Discover Found – $(6011|5[1-5]\d{2}|4\d{3}|3\d{3})-\d{4}-\d{4}-\d{4}$ - a Visa, Mastercard or Discover Card in the format nnnn-nnnn-nnnn-nnnn
- American Express Card Found – $(3[47]\d{2}|2014|2149|2131|1800)-\d{4}-\d{4}-\d{3}$ - an American Express card in the format nnnn-nnnn-nnnn-nnn
- Colorado SSN without Spaces or Hyphens Found – $\b(52[1-4]|65[0-3])\d{6}\b$ - a Colorado issued (beginning with 521-524 or 650-653) social security number in the format nnnnnnnn. This format is limited to Colorado SSN's because strings of numbers without separators are very common and will generate many false positives without restrictions. Because Coloradoans are a major group at the university in both students and employees, most collections of SSN's will contain some Colorado issued ones.
- SSN with Hyphens or Spaces Found – $\b[0-7]\d{2}[-]\d{2}[-]\d{4}\b$ - a social security number in the format nnn-nn-nnnn or nnn nn nnnn with a beginning digit of 0-7 (beginning digits of 8 and 9 are not valid SSN's)
- Discover Card Without Hyphen Found – $\b6011\d{12}\b$ - a Discover card in the format nnnnnnnnnnnnnnnn
- Elevations Card Found – $\b441420\d{10}\b$ an Elevations credit union card in the format nnnnnnnnnnnnnnnn
- CU Acard Found – $\b540501\d{10}\b$ - a University Acard in the format nnnnnnnnnnnnnnnn

- CU USBank Card Found– \b473077\d{10}\b - a US Bank card (issued for university travel expenses) in the format nnnnnnnnnnnnnnnn

To add your own regular expressions, click the *Add regex* button.

Regular Expression Editor

Regular Expression

Regex name

Test Data

Validators

Luhn checksum

SSN area/max group

None

Enter the regular expression in the first box (see http://en.wikipedia.org/wiki/Regular_expressions for more information on regular expressions) and a name for the search in the “Regex name” box (this name will appear in the log file when a match is found).

The “Test” button and “Test Data” field can be used to test your regex search against sample data. To do so, enter a sample piece of data that you expect to match your search into the “Test Data” field and click the “Test” button. A message will appear in the window telling you if the data was a match for the regex search, like this:

Regular Expression Editor

Regular Expression

Regex name

MATCH + Validator

Test Data

Validators

Luhn checksum

SSN area/max group

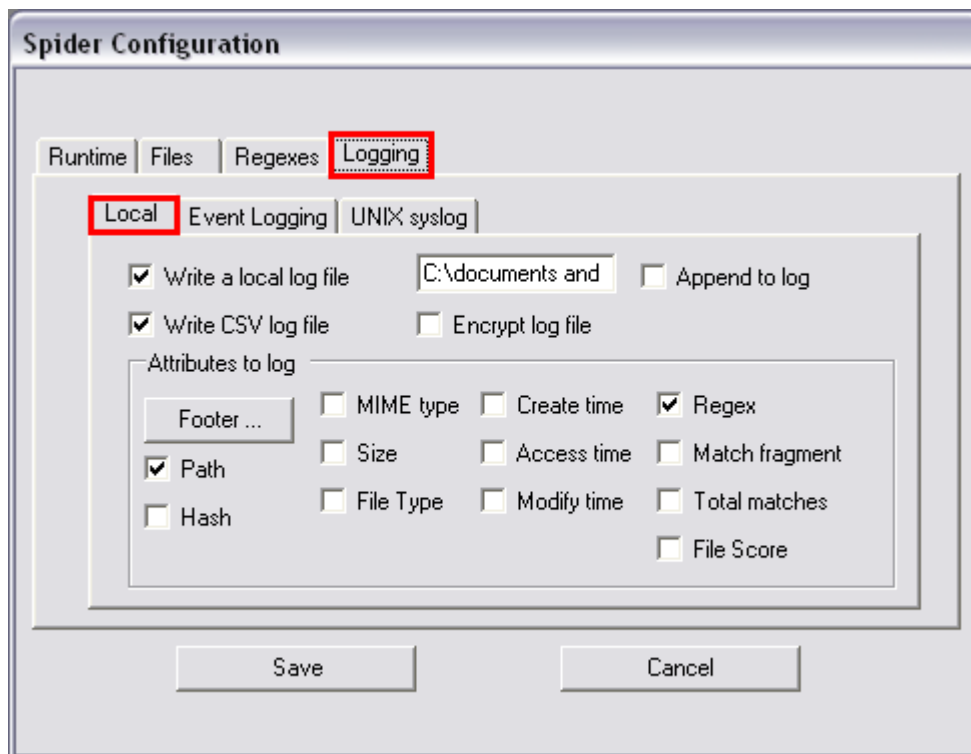
None

Note on validators: The validators are only useful when using the “file score” function in the log (see log configuration section) and will result in a lower file score if the regex matches, but the validator does not match. Hits where the regex matches and validator does not will still be logged, so this function does not serve to directly filter false positives.

The Luhn checksum validator is used to verify credit card numbers as legitimate.

The SSN area/max group validator is used to check the SSN group component (middle two digits) against a list of the valid group components for that particular area component (first three digits).

Logging configuration

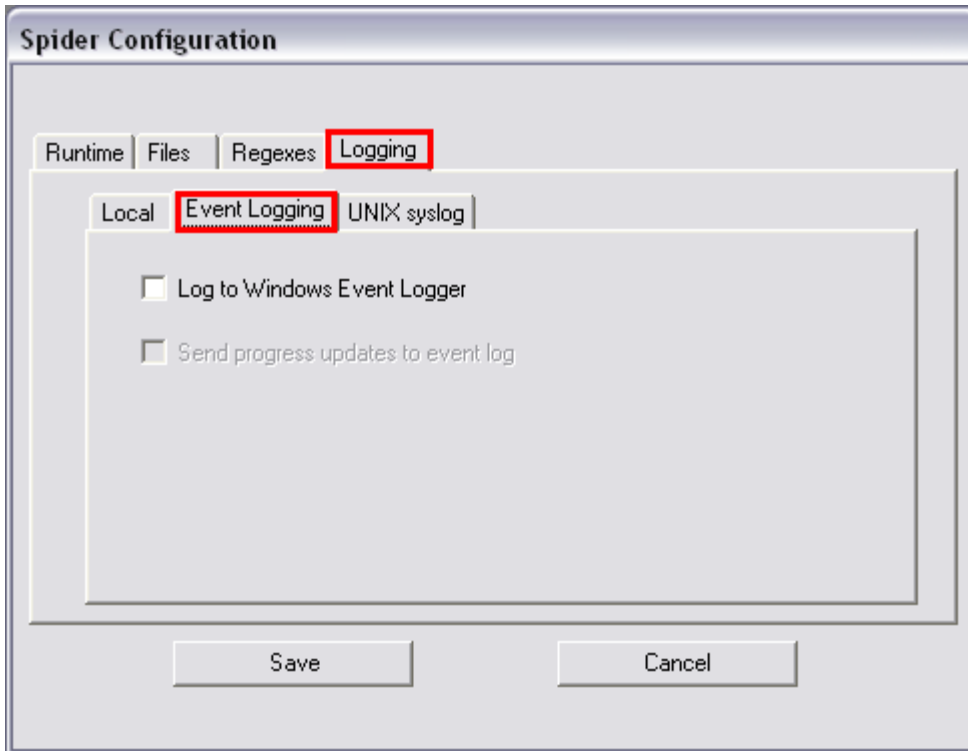


Logging tab —> Local tab

- Write a local log file (default in ITS config) - create a log file on this machine, at the location specified. This path can be a mapped drive to a server or a UNC path to another system (e.g. \\server.colorado.edu/share/log_folder logfile.log) as long as the user executing the scan has write access to the destination. This allows for central storage of the log files. The path statement also includes the name of the log file. The ITS distribution uses a log file location of the desktop of the user who ran the scan and a file name that incorporates the date and IP address. Using

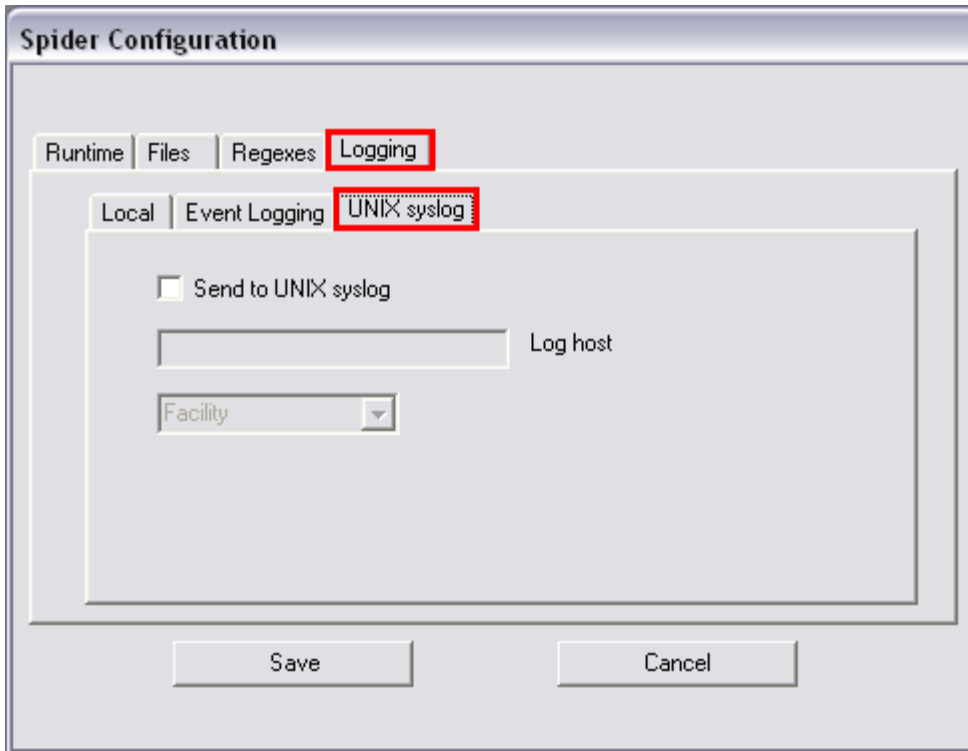
a file name with variables is important when saving logs from multiple systems to a single location.

- Append to log - add the new log to the end of an existing log
- Write CSV log file (default in ITS config) - create a comma-delimited log file containing the specified information about each file.
- Footer... - text to append to the log file. The ITS distribution includes a footer with several pieces of information that identify the date, time, username, IP address, computer name and number of files scanned. This is useful when dealing with logs from multiple systems.
- Path (default in ITS config) - full drive and path to the file containing a 'hit'.
- Hash - MD5 hash of the file, useful for identifying identical files stored with different names or paths
- MIME type - MIME type of the file (application/ms-excel, etc.)
- Size - size of the file, in bytes
- File Type - file type: Archive, Normal, Compressed, Hidden, System, etc.
- Create time - creation time of the file
- Access time - access time of the file; more often than not, this is the time Spider opened the file
- Modify time - last modification time of the file
- Regex (default in ITS config) - the regular expression matched: SSN for Social Security number, VMCD or AMEX for credit card number
- Match fragment - the section of the file that matched one of the regular expressions sought. Match fragments can only be written in an encrypted CSV file and enabling this option will automatically enable the "Encrypt log file" option.
- Total matches - total number of matches found; this must be selected if you want a File Score
- File Score- a number between 1 and 0 indicating the probability that the file contains a valid match
- Encrypt log file - Spider will prompt for a password (at least 16 characters in length) and use that to encrypt the local log file. Log file encryption is incompatible with unattended operation (using the /run flag as described in the next section).
 - **Note:** Using this setting impacts how log files are opened using Spider. When this setting is enabled, Spider will attempt to decrypt any log file you manually open in the program, including unencrypted ones (unencrypted log files can always be opened in any text editor). This means that you cannot open unencrypted log files with this setting enabled. Conversely, Spider will not attempt to decrypt a file when this option is not selected, so you cannot open an encrypted log file without this option selected.



Logging -> Event Logging

Spider is capable of logging results to the local Windows event log. To enable logging to the Windows event log, check the “Log to Windows Event Logger” box on this page. With Windows event logging enabled, Spider can also send periodic status update to the Windows event log as well (this can be helpful if you are running Spider in silent mode and wish to check on the progress).



Logging -> UNIX Syslog

Spider is also capable of logging to a remote syslog server. To enable this function, check the “Send to UNIX syslog” box and enter the hostname of the syslog server into the “Log host” text field.

Exercise	
Module 4	Exercise 1

Advanced configuration

When Spider scans a system, the access time noted on each file will be updated. For some situations (security investigation, intrusion detection systems keyed to access times, etc), one may wish to not alter the access times. Spider contains a generally undocumented option that will reset the file access times to their original values. To enable this option, create the following registry key:

HKLM\Software\Cornell University\Spider\Runtime\NoAtime

Type: DWORD

Value: 0 means default Spider behavior, 1 means reset atimes during scan

For instances where preservation of state is critical, it is best to mount a file system with read-only access before using any tools rather than having the tool reset access times.

Variables available within Spider

There are a number of variables available within Spider for customizing configuration options, particularly the logging section. Several of these are already in use in the ITS distribution for the log file name and log file footer. The following list includes the variable name (note case sensitivity), a description and an example in parentheses.

%i – IP address of computer that Spider is running on (128.138.1.1)
%d – day of the month in numeric format (27)
%D – day of the week in abbreviated English (Mon)
%P – day of the year in numeric format (2006)
%m – month in numeric format (11)
%M – month in abbreviated English (Nov)
%y – year in nnnn format (2006)
%T – time in hhmmss format (071534)
%n – system Netbios name (SYSTEM)
%N – system DNS hostname name – does not include domain, just hostname (system)
%u – username of user executing Spider (ralphie)
%C – total number of files scanned (13271)
%H – total number of hits
%R – runtime in seconds (685)
%F – CR/LF
%r – carriage return
%l – line feed
%% - percent sign

Examples:

The log file footer in the ITS distribution is:

```
spider ran at %T on %d %M %y on system %N at IP %i under user %u and processed %C files
```

which results in a log file footer like:

```
spider ran at 112601 on 13 Nov 2006 on system computer at IP 128.138.nnn.nnn under user ralphie and processed 24637 files
```

The log file path in the ITS distribution is:

```
C:\documents and settings\%u\Desktop\spider_%i_%d_%m_%y.log
```

Which results in a log file written to the desktop of the user executing Spider with a filename like:

spider_128.138.1.1_13_11_2006.log