January 24, 2008

Forensic String Searching Tool Requirements Specification

Public Draft 1 of Version 1.0

National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

Table of Contents

1.0 Introduction

There is a critical need in the law enforcement community to ensure the reliability of computer forensic tools. A capability is required to ensure that forensic tools consistently produce accurate, repeatable and objective test results. The goal of the Computer Forensic Tool Testing (CFTT) project at the National Institute of Standards and Technology (NIST) is to establish a methodology for testing computer forensic tools by the development of functional specifications, test procedures, test criteria, test sets, and test hardware. The results provide the information necessary for toolmakers to improve tools, for users to make informed choices about acquiring and using computer forensics tools, and for interested parties to understand the tools' capabilities. This approach for testing computer forensic tools is based on well-recognized international methodologies for conformance testing and quality testing. This project is further described at http://www.cftt.nist.gov/.

The CFTT is a joint project of the National Institute of Justice (NIJ), the research and development organization of the U.S. Department of Justice; NIST's Office of Law Enforcement Standards (OLES) and Information Technology Laboratory (ITL); and is supported by other organizations, including the Federal Bureau of Investigation, the Department of Defense Cyber Crime Center, and the Department of Homeland Security's Bureau of Immigration and Customs Enforcement and U.S. Secret Service. Since all documents are posted on the web for public review, the entire computer forensics community participates in the development of the specifications and test methods.

2.0 Purpose

This document defines requirements for digital forensic string search (DFSS) tools used in computer forensics investigations.

The requirements in this document are used to derive assertions to be tested. The assertions are described as general statements of conditions that are checked after a test is executed. Each assertion is checked in one or more test cases that specify detailed initial conditions, test scenarios, and expected test results.

These requirements were initially developed by a focus group of individuals who were expert in the use of disk acquisition tools and have performed investigations that depend on the results of these tools. As this document evolves through comments from the focus group and others, new versions will be posted at http://www.cftt.nist.gov/.

3.0 Scope

The scope of this specification is limited to software tools that search images acquired from digital storage media and also tools that directly search digital storage media.

The proper or improper use of a tool is not within the scope of this specification.

4.0 Background

A general model for operation of a DFSS tool is helpful for writing tool requirements and identifying terms useful for discussing string searching. This abstract framework gives structure and provides a general outline with which to proceed during the analysis. At an abstract level string searching involves the following:

- something to search with,
- someplace to search,
- something to search for, and
- search results.

A *search engine* implements a *search algorithm* that performs the search. A *digital forensic string search* tool provides an interface between a user and a search engine. The DFSS tool interfaces to at least one search engine, but may interface to additional search engines.

Someplace must be accessible to the DFSS tool for searching. This place is called the *search universe*. The actual search may be restricted to a subset of the search universe.

In the simplest case, the user is looking for a match from the search universe to a target *search string*. So the tool can search for a key word like *gun* or *knife*, but it might be directed by the user to search for both. In general, the user has a case specific list of search terms. In another case if the user wants the tool to find social security numbers groups of nine digits can be specified as a regular expression (i.e., a pattern) such as **[0-9]{9,9}** (a string of nine digits with no separators). In other cases the user might need to search for text that is not represented in ASCII, such as searching for the Chinese word 虎 (*hu* or tiger). There are multiple possible encodings for the character (e.g., UNICODE, GB, Big 5, SHIFT JIS, etc). It should be noted that English text may also use multiple encodings, e.g., old Univac series computers used an encoding called *field data* and some IBM systems used *extended binary coded decimal interchange code* (EBCDIC) to represent text. The encoding for *Z* in ASCII is 01011010, in EBCDIC is 11101001 and in field data (six bits) is 011111.

As a practical matter, the *something to search for* is not just a search string but is a collection of parameters.

After a search is performed the results must be presented to the user in a meaningful and useful way. The actual strings matched, in the case of a pattern search, and the location of the matched strings must be presented in the response such that the matched strings and surrounding context can be extracted for analysis and reporting.

5.0 Definitions

Term	**Definition**
Forensic search tool:	Interfaces to one or more *string search engines*.
String search engine:	Implements a string search algorithm that takes a *query* and a *search universe* and returns a *response*.
Query:	A set of *search parameters* that specify a *match set*.
Search parameters:	• Search pattern: a string specifying in some *pattern matching language* substrings of the search universe, subject to meeting criteria established by other search parameters, that satisfy a query. • Search arena: a subset of the search universe to be searched by the query. • Search engine: the search engine to be used for the query. • Character representation: the interpretation of the bit patterns of the search arena by the query. • Ignore case: the upper case and lower case variations of a character match. • Text direction: the direction words (left-to-right or right-to-left) are written in the text. • Search direction: the direction (beginning-to-end or end-to-beginning) that the search arena is traversed by the search. • Hit count: the number of search hits to return in response to the query. A hit count parameter can be further refined as a *number of files* or *hits within a file* count. • Search type: One of *pattern match*, *word*, *stem, synonym, fuzzy, phonetic*, or *index*. These are the most common, but a search engine may define other search types. • Disambiguation rule: A rule for selecting a match from multiple candidates. • Other: These are the most common search parameters, but a search engine may define others.
Pattern matching language:	A language, such as the regular expression language of UNIX used by the **grep** tool, for specifying strings that satisfy a query.
Totally ordered search arena:	If there exists a unique ordering of all bytes within a search arena then the search arena is totally ordered.
Ordered within objects search arena:	A search arena that is an unordered set of ordered objects (e.g., files).
Location Description:	A location description is composed of four items: 1. Matching string 2. Object identification (e.g., file name and path) 3. Offset within the object (e.g., sector address) 4. Length of matching string The offset and length are sometimes omitted.
Search hit:	The string matching the query and a *location description*.
Match set:	The set of substrings from the search universe specified by a query.

Query response:	The set of search hits returned by a query.
Search universe:	The search universe may be either the content of some type of digital media or an image file taken from some type of digital media. The media may be either unformatted or formatted with one or more file systems (e.g., FAT, NTFS, HFS, etc.).

6.0 Requirements

This section lists requirements for forensic string search tools. The requirements in section 6.1 must be met by all tools. Requirements for optional tool features are presented in section 6.2.

6.1 Basic Requirements

SS-BR-01. The response returned by a query is equal to the match set for the query.
SS-BR-02. The tool shall search using one or more specified character representations.

6.2 Optional Requirements

SS-OR-01 If the tool allows specification of a search arena that is a subset of the search universe then all search hits in the response are from the search arena.
SS-OR-02 If the tool searches an ordered search arena and a hit count of *n* is specified then the response is the first *k=min(n,m)* matches located in the direction specified, where *m* is the number of strings in the match set.
SS-OR-03 If the tool searches an unordered search arena and a hit count of *n* is specified then the response is any *k=min(n,m)* matches, where *m* is the number of strings in the match set.
SS-OR-04 The tool shall display text of a query and the query response in the specified text direction.
SS-OR-05 If the tool performs a *stemming* search, the matches shall be at least as good as from the Porter stemming algorithm (M.F. Porter, 1980, An algorithm for suffix stripping, *Program*, **14**(3) pp 130−137).
SS-OR-06 If the tool performs a *synonym* search, synonyms of the query string shall match.
SS-OR-07 If the tool performs a *fuzzy* search, close misspellings of the query string shall match.
SS-OR-08 If the tool performs a *phonetic* search, words that are pronounced the same as the query string shall match.
SS-OR-09 If the tool provides predefined queries then the response returned is equal to the match set for the query.
SS-OR-10 The tool may combine queries with logical operations such as *and*, *or*, or *not*.

The following requirements are for specifying a search pattern. The requirements are a generalization of a subset of the regular expression specification of IEEE Std 1003.1 (POSIX).

A search pattern is character string composed from designated literal and meta characters of the pattern matching language implemented by the search tool.

There may be some contextual exceptions to each of the following requirements.

SS-OR-11 A *literal* character matches itself.
SS-OR-12 An *any meta character* matches any single character.
SS-OR-13 An *anchor meta character* matches some location in an object (e.g., beginning of file, end of line, etc).
SS-OR-14 A *class meta character* matches a single character from some character class (e.g., uppercase). Note that specification of the class may require an identifying substring.
SS-OR-15 An *escape meta character* removes any special meaning of a meta character and allows the specification of the literal value of the meta character that shall match the literal value of the character.
SS-OR-16 A *group meta character* specifies grouping of substrings within the pattern string.
SS-OR-17 A *repeat meta character* specifies the repetition of a substring *n* times, where *n* may be specified as one of the following: $n = 0$, $n \geq 0$, $n \geq 1$, or $k \leq n \leq j$.