Document Name		
Document Title	Forensic File Carving Report	
Project	Forensic Investigation of Digital Objects (FIDO)	
Author(s)	Gareth Knight	
Date	30 June 2011	
Access	☑ Project and JISC internal	☑ General dissemination

Document History			
Version	**Date**	**Author**	**Comments**
0.1	03 May 2011	Gareth Knight	First version
1.0	30 June 2011	Gareth Knight	Added references

Catalogue Information	
Title	Forensic File Carving Report
Creator	Gareth Knight
Subject	File carving, fido, data extraction, disk image, data carving, header, footer, acquisition
Description	This report describes the different data carving methods in use, scenarios in which they are appropriate, and makes recommendations for data carving tools to be used in the FIDO project
Publisher	King's College London
Contributor	
Date	2011-06-30
Language	En-gb
Rights	Creative Commons Attribution-NonCommercial-ShareAlike 2.0 UK: England & Wales

Report purpose

The Forensic Investigation of Digital Objects (FIDO) project investigated the application of digital forensics within the working practices of a UK HE archive. The project demonstrated the value of adopting tools and techniques developed for the emerging digital forensics field, while building upon the long-standing archival theory archival and digital curation approaches. This report describes the different data carving methods in use, scenarios in which they are appropriate, and makes recommendations for data carving tools to be used in the FIDO project.

Forensic Data Carving in context

Examination refers to a process of careful inspection or review to gather information or address specific questions. For an archive handling physical records, examination is likely to incorporate activities associated with reviewing a box of papers or other items to determine the type of information they contain. For a digital forensic investigator, it will involve a detailed inspection of the digital evidence, e.g. a disk image, to identify digital information of value to the investigation. The challenge for an investigator at this stage is to identify the information that may have archival or evidential value, or serve as contextual information which may inform understanding of other information held on the disk.

The interrogation of digital media or imaged data for information is likely to include a number of activities. The simplest of these actions is to examine the 'active' data files held on the disk – files that may be located by simply navigating to the correct directory and accessing the necessary file. This may be supplemented by data recovery, using file system information to obtain data files held in unallocated space that have been marked as deleted, but have yet to be overwritten. A third, more advanced method available to the investigator is to use Data Carving[1], a technique by which raw data (undifferentiated blocks) is analysed for patterns that match the header and/or footer of a known file format and, once identified, a data file is "carved" from the data for analysis and use (Garfinkel & Metz, n.d., cited in Anon, 2010). Unlike data recovery, which is dependent upon the file system information being intact and accurate, file carving may be used to recover data that has been accidentally or deliberately deleted[2], or has been rendered inaccessible (e.g. as a result of mechanical failure or virus that has resulted in the file system allocation table being corrupted).. An archival equivalent of data carving could not be identified; however, it has some similarity to the process used in museums to identify the content or workings of Egyptian artefacts.

Data carving methods

The digital evidence given to a forensic investigator for examination are frequently bit-copies of one or more partitions of a hard disk or other storage device. The files held on primary hard disks and USB sticks are likely to change constantly, in comparison to other forms of digital media, such as floppy disks. Files may be updated, copied, moved, or deleted on a regular basis (Kloet, 2007). Although file system information may not be available to identify these files, fragments may continue to exist that may be extracted and analysed. Data carving tools use several different techniques (algorithms), alone or in combination, to identify and extract files contained within a data image. To distinguish between these methods, Garfinkel & Metz (n.d.), as cited in Anon (2010), proposed a taxonomy that may be used to distinguish between, and understand the benefits and problems, associated with each approach. This may be supplemented through recognition of other methods available.

[1] Synonyms include File Carving, or simply Carving

[2] Filenames are stored in the disk file system and are, in most cases, unrecoverable. Carving tools assign new alphanumeric filenames in their absence

1. Header/Footer Carving

The image file is analysed to identify data patterns that denote the header of a known file type, followed by the first occurrence of an corresponding file type footer within a specified range (e.g. 10Mb). The data between the header and the footer is identified as a candidate file to be recovered. To illustrate, Table 1 indicates the headers and footers that are used by Scalpel to identify five common file types (Richard and Roussev 2005).

file type	*header*	*footer*
gif	nx47nx49nx46nx38nx37nx61	nx00nx3b
jpg	nxffnxd8nxffnxe0nx00nx10	nxffnxd9
htm	<html	</html>
Encrypted text	——BEGINn040PGP	
zip	PKnx03nx04	nx3cnxac

Table 1: Sample header/information used by Scapel to identify files

Header/Footer carving is most effective for analyzing data formats that possess large headers and footers (e.g. PNG images) on drives or disk images that have little or no fragmentation. In other circumstances, it may produce produce erroneous results when analyzing certain types of object. Examples include:

- *Problem Scenario 1: The disk is fragmented, locating segments of a file in different locations on a disk*
  The header of file A will be identified, however, as a result of the sequential approach taken to identification of the footer, it is matched to the footer of file B which is physically located next to the header.

- *Problem Scenario 2: The format to be carved has a minimalist header or footer*
  The likelihood that the hexadecimal values used for a header/footer is also used for other purposes will increase if the header/footer is short. A JPEG header (hexadecimal: xFFxD8) and footer (hex: xFFxD9) is relatively short, potentially resulting in the carved file lacking the final segments necessary to decode it. By comparison, a HTML header and footer is less likely to be processed incorrectly.

- *Problem Scenario 3: The format does not have a fixed header or footer*
  Some file types, e.g. unencrypted plain-text, do not have fixed a header or footer that may be used to identify the start and/or end point. As a result, it is not possible to identify these data files with any level of accuracy. If the format possesses a fixed header, but no footer, header\maximum file size carving may be performed (see below)

2. Block-Based Carving

A carving method in which the content of a raw data file are analysed on a block-by-block basis to determine the type of content. The method is particularly effective for identifying objects that contain distinct characteristics that are unique or, at least, more common to certain file types, e.g. the appearance of symbols such as '<' and '> may denote that the block contains a HTML or XML file, text characters may infer that the block contains a segment of a text document. A potential limitation is that this method assumes that each block can only be part of a standalone file or embedded object[3].

- *Problem Scenario 1: The content is part of a complex object*

[3] See Kloet, 2007, p41

Recognition problems may be encountered when processing complex file formats (e.g. MS Word, PowerPoint and Adobe Acrobat) that contain multiple types of content, such as text and images. This may result in the production of a large number of 'false positives' - carved files that were embedded within a larger object on the original disk. For many types of filestream[4], it may not be possible to determine whether the file was self-contained or embedded.

- *Problem Scenario 2: The object contain characteristics that are common to several formats*
  The statistical analysis may identify chunks of data that are ambiguous, potentially belonging to several types of content. The handling of the content, either retaining it for further analysis or discarding it, may vary between different applications.

3. Statistical Carving

A carving method that uses statistical pattern recognition of content to link data identified as belonging to the same file type (e.g. a PNG image) located at different points within the data image/disk media. Statistical carving has potential use for application to fragmented disks, where content is stored in stored on non-contiguous clusters[5]. Similar to block-based carving, the method is likely to be effective for identifying objects that contain distinct characteristics that are unique or, at least, more common to certain file types, e.g. the appearance of symbols such as '<' and '> within a HTML or XML file. However, work is still in its infancy, with developers attempting to determine the most effective method of applying it in practice.

- *Problem Scenario 1: The content is part of a complex object*
  Recognition problems may be encountered when processing complex file formats (e.g. MS Word, PowerPoint and Adobe Acrobat) that contain multiple types of content, such as text and images. This may result in the production of a large number of 'false positives' - carved files that were embedded within a larger object on the original disk. For many types of filestream[6], it may not be possible to determine whether the file was self-contained or embedded.

- *Problem Scenario 2: The object contain characteristics that are common to several formats*
  The statistical analysis may identify chunks of data that are ambiguous, potentially belonging to several types of content. The handling of the content, either retaining it for further analysis or discarding it, may vary between different applications.

4. Header/Maximum (file) size Carving

A method for carving files out of raw data by identifying the header of a known file type and sequentially extracting all data that follows until a maximum file size has been reached. The approach may prove effective when attempting to carve file types that possess a short footer or no footer. Several authors suggest it may be used with success for file formats, such as MP3 and JPEGs containing a valid audio/image stream that can be processed, even if additional superfluous data is appended to the file.

- *Problem Scenario 1: Carving of files with little or no footer*
  The carving process will produce a number of un-necessarily large files which will likely require further processing in order to be useful

[4] Filestream is used in this context to refer to data files that may be extracted from a datastream and used without additional information, as per the PREMIS definition.
[5] See Veenman, 2007 for further information
[6] Filestream is used in this context to refer to data files that may be extracted from a datastream and used without additional information, as per the PREMIS definition.

- *Problem Scenario 2: Carving of files with little or no footer*
  The additional, superfluous data in the file may potentially contain code that can cause a buffer overflow on the processing machine.

5. Header/Embedded Length Carving

A carving method that may be performed on file formats, such as BMP, PDF, and AVI that store their total size (length) in the first few bytes of the header. Similar to other carving methods, the analysis tool examines the raw data file for headers of known file formats. If the header contains information on the length, it is used as a basis for indicating the amount of data to be extracted.

- *Beneficial scenario 1: Carving of files with a header, but little or no footer*
  The method is particularly useful for file formats that have a documented header, but do not possess a footer.

- *Problem Scenario 1: The disk is fragmented, locating segments of a file in different locations on a disk*
  The header of file A will be identified, however, as a result of the approach taken to identification of the footer, it is matched to the footer of file B which is physically located at the position indicated in the header.

6. File structure based Carving

A method for carving files from raw data using knowledge of the internal structure of file types. Research remains in its infancy, but the carving technique is considered likely to provide extremely accurate results. However, similar to statistical carving, identification may prove problematic for ambiguous content types.

7. Semantic Carving

A carving method achieved by performing a linguistic analysis of a raw data file in order to differentiate between data files. For example, a file written in English that is located next to data clusters written in French may indicate that the French text is a fragment of a previously deleted file. Semantic carving is reliant upon developments in linguistic analysis and, as a result, is still in its infancy.

- *Beneficial scenario 1: Distinguishing between content contained in different files*
  Semantic carving has the potential to differentiate between textual content associated with different files based upon the writing style.

- *Problem scenario 1: Application to non-textual resources*
  The technique has only been considered in the context of textual resources. However, it could, in principle, be applied to still and moving images by performing an analysis of colour techniques.

8. Carving with Validation

The extraction of files from raw data using one of the carving methods outlined, accompanied by some form of verification to determine if the carved file represents a valid file type, e.g. through use of JHOVE, etc. Carving with validation minimises the risk of ‘false positives’ produced as a result of non-valid data being extracted;

9. Fragment Recovery/ Split Carving

Any carving method in which two or more fragments are combined to form the original file or object.

10. Repackaging Carving

A supplementary activity in which data extracted by a carving process is modified, through the addition of new headers, footers, or other information to enable it to be accessed using common software tools. For example, an incomplete GIF image may be modified through the addition of a footer to enable the user to view the remaining content, or a ZIP carver tool may repackage the constituent parts of a ZIP file and repackage them in a manner that allows it to be unpackaged (Cohen, 2007).

11. In-place Carving

A carving technique in which pointers to the start and end of a data file are identified and recorded, rather than extracting the data itself. In-place carving was conceived, based upon the recognition that carved files are likely to require a considerable amount of disk space .It is not intended to improve the accuracy (Golden et al, 2007)

When evaluating the use of carving tools for use on real-world data, consideration must be given to the carving method(s) supported and the suitability of the approach when applied to the specific scenario. Research into different carving methods is ongoing, with effort being focused upon improving the accuracy of each approach and combining the functionality offered by several carving methods. It is evident that no single approach is truly effective for every scenario. Instead, a combined approach must be taken, merging two or more methods to improve the accuracy of the process. For example, since the publication of the taxonomy, 'Smart Carving'[7] has been proposed, combining structure-based validation along with validation of each file's unique content.

[7] For further information, see http://www.forensicswiki.org/wiki/File_Carving:SmartCarving

Carving Tools

A large number of software tools exist, for a range of operating system, which may be used by a forensic investigator to analyse digital media and extract digital information

Name	Description	Useful functionality	Method	Requirements	User Interface[8]	Licence
Disk Digger	A file recovery and carving tool		Header/footer	Windows	GUI	Commercial[9]
Foremost	A powerful file carving tool		Header/footer, internal data structure, header/maximum size		CLI	GNU GPL
Magic Rescue			Header – 3rd party tools used to id footer[10]	Linux	CLI	GNU GPL
RecoverJPG	A file carving tool for JPEGs and Quicktime files		Header/footer		CLI	GNU GPL
Scalpel	A file carver that reads a database of header & footer definitions and extracts matching files or data fragments from a set of image files or raw device	Claims file carving speeds 2-5x faster than Foremost Claims to be file system independent	In-place;	Linux, Windows, Mac OS X	CLI	GNU GPL
SFDumper	A Bash script that combines the functionality of Sleuthkit, Foremost to	[1] Single command to extract active, deleted and carved files of a common file type and de-duplicate them	Same as Foremost	Linux	CLI	Unknown - free
SleuthKit DLS	A Unix tool contained in SleuthKit that may be used for data carving		Header/footer	Linux	CLI (GUI through Autopsy)	GNU GPL
WinHex	Hex viewer that provides forensic functionality		Header/footer	Windows	GUI	Commercial

[8] Command Line Interface, Graphical User Interface
[9] Personal use licence available for $14.99, commercial licence for 49.99. http://diskdigger.org/
[10] The developer suggests various tools that may be used in conjunction with MagicRescue, e.g. dls from the Sleuth Kit. See http://www.itu.dk/people/jobr/magicrescue/ for further information.

References

Anon. 2009. “File Carving: Smart Carving”. ForensicWiki. 12 November 2009. Accessed June 29, 2011: http://www.forensicswiki.org/wiki/File_Carving:SmartCarving

Anon. 2010. “File Carving”. Forensic Wiki. 7 September 2010. Accessed June 29, 2011: http://www.forensicswiki.org/wiki/File_Carving

Cohen, M.I. 2007. “Advanced carving techniques”. Digital Investigation, Volume 4, issue 3-4 (September - December, 2007), p. 119-128. Elsevier Science Publishers B. V. Amsterdam, The Netherlands, The Netherlands. ISSN: 1742-2876 DOI: 10.1016/j.diin.2007.10.001.

Cor J. Veenman, C.J. 2007. “Statistical Disk Cluster Classification for File Carving”. IAS '07 Proceedings of the Third International Symposium on Information Assurance and Security Pages 393-398. IEEE Computer Society Washington, DC, USA. DOI: 10.1.1.132.4039

Kloett, S.J.J. 2007. “Measuring and Improving the Quality of File Carving Methods: Master’s Thesis”. Eindhoven University of Technology. October 29, 2007. Accessed June 29, 2011: http://alexandria.tue.nl/extra2/afstversl/wsk-i/kloet2007.pdf

Richard, G.G III and Roussev, V. 2005. “Scalpel: A Frugal, High Performance File Carver”. Digital Forensics Research Workshop, 2005. Accessed June 29, 2011: http://roussev.net/pdf/2005-DFRWS--scalpel.pdf

Richard III, G.G. Roussev, V. and Marziale, L. 2007. “In-Place File Carving”. Third Annual IFIP WG 11.9 International Conference on Digital Forensics. National Center for Forensic Science. Orlando, Florida, USA. January 28 - 31, 2007. Accessed June 29, 2011: http://digitalforensicssolutions.com/papers/ifip2007-final.pdf