D29 Application of Forensic Techniques to Data Preservation

John Tebbutt, and Douglas White, MS, National Institute of Standards and Technology, 100 Bureau Drive Stop 8970, Gaithersburg, MD 20899-8970*

The goal of this presentation is to learn about an automated procedure for processing large numbers of files to extract forensic metadata and how this procedure is employed for two disparate purposes: to provide court admissible computer file identification information to law enforcement agencies for use in forensic examinations of computer systems; and to aid in data reduction, management and cataloging of a large document corpus.

This presentation will impact the forensic science community by describing a set of procedures used both for the production of data for use by computer forensics investigators and the automation of data reduction and management in a large data collection.

Attendees will learn how the National Software Reference Library processes larges numbers of files to extract forensic metadata and how this same process can be applied to aid in data reduction, management and cataloging of any large corpus of files. Application to the collection managed by the National Archives and Records Administration is given as a specific example.

This talk will demonstrate that procedures originally developed for the production of forensic metadata can be successfully applied in the automation of data reduction and management in a large data collection.

The National Software Reference Library (NSRL) of the National Institute of Standards and Technology (NIST) is studying the application of techniques initially developed to aid computer forensics examiners in the investigation of mass storage devices to assist the National Archives and Records Administration (NARA) in its mission to preserve the records of government until the end of the Union.

With the increasing use of information technology, successive administrations are producing correspondingly more data to be archived. For example, the archives has estimated that the amount of data from the George W. Bush administration to [date] is on the order of [X] TB. The automation of data preservation, management and cataloging becomes ever more vital to the NARA's mission.

The NSRL routinely collects and generates metadata associated with large numbers of files. The process by which file metadata are derived is largely application agnostic and can be applied to any large corpus of digital files. As with any sufficiently comprehensive dataset, they can be mined for purposes other than that for which they were originally intended. By applying the NSRL process to the NARA dataset, we can assist with data reduction, management and cataloging.

At the file level, the metadata collected include "hard" data, such as file name, file size, and routine cryptographic hashes of the file content - useful for uniquely identifying individual files - and "fuzzy" hash values, which are based on analysis of the structure of file contents and are useful in the detection of files which are similar but not identical. At the sub-file level, the metadata consist of cryptographic hashes of a file's constituent blocks.

Data reduction is achieved by comparing the "hard" metadata of every file in the corpus and using the information to detect and discard duplicate files, replacing them with references to a single (or a very small number of) master copies. This process is totally automated and does not require input from or supervision by an archivist.

Once exact duplicates have been discarded, the remaining unique files are compared using the "fuzzy" metadata and block-level metadata and compiled into clusters of similar files. The cluster data is available for inspection by archivists and to aid in data management and cataloging.

Metadata, Data Preservation, Data Reduction

** Presenting Author*