C12 Another Forensic Image Data Set (AFIDS)

Mark D. Guido, MS, The MITRE Corporation, 7515 Colshire Drive, Mclean, VA 22102; and Michael McCarrin, PhD*, Naval Postgraduate School, 507 Koshland Way, Santa Cruz, CA 95064*

The goal of this presentation is to raise awareness and interest in a new and available research data set that will be composed of real-world forensic images from around the world. This increased awareness of AFIDS is primarily targeted at the Digital & Multimedia Sciences Section of the American Academy of Forensic Sciences due to its diverse member population across industry, academia, and law enforcement/government/military arenas.

This presentation will impact the forensic science community by providing a real-world, reasonably safe, and unique dataset to aid researchers in tool development and testing of forensic tools and related capabilities.

In 2009, Garfinkel et al. made the case for standardized forensic corpora to foster scientific approaches and progress in digital forensics.[1] Of the datasets that resulted from this effort, the Real Data Corpus (RDC) was by far the largest and most representative of the variety and complexity of data encountered during digital forensic investigations. Its use in the development and de-bugging of widely used tools such as bulk extractor and The Sleuth Kit testify to the success of the project; however, though this dataset is still actively maintained, challenges relating to cost as well as institutional policies have prevented the addition of new drive images for several years. Consequently, new applications and operating systems are absent, and the versions of software represented are long out of date. The current RDC continues to suffer from entrenched problems related to policy, operations, and oversight.

The development of AFIDS is therefore proposed. The intent is to define a community resource to enable the advancement of digital forensics research while incorporating from the outset lessons learned from previous work on large-scale restricted datasets such as the RDC and the National Software Reference Library (NSRL). AFIDS can leverage Amazon® Web Services GovCloud infrastructure to reduce costs of long-term storage and to allow controlled access by researchers without the additional privacy risks associated with distributing copies of the data. AFIDS can provide a managed interface by which researchers can submit “queries” to the dataset (ranging from simple functions to custom forensic analysis tools) that will be assessed in terms of their risk of exposing personally identifiable information before being run across the dataset via Cloud technologies. It is expected this risk assessment will facilitate Institutional Review Board oversight and reduce researcher overhead. Query results, moreover, can be cached and made available, preventing the need to repeat expensive computations.

In conclusion, a study to identify improvements over the current RDC is already being conducted; the plan is to bring AFIDS online and make it available for researchers from industry, academia, and law enforcement/government/military. The current status of AFIDS will be provided to attendees.

Reference(s):

1. Garfinkel, Simson, Paul Farrell, Vassil Roussev, and George Dinolt. Bringing science to digital forensics with standardized forensic corpora. *Digital Investigation.* 6 (2009): S2-S11.

Real Data Corpus, Forensic Images, Research

*Presenting Author