Title: Crossing the Bitstreams: A Call for Collaborative Application of Forensics to Digital Curation Work
Authors: Jeremy Leighton John, Matthew Kirschenbaum, Mark Matienzo, Don Mennerich, Christopher A. Lee, Porter Olsen, Kam Woods
Date: December 13, 2012

Digital forensics tools and methods can support various digital curation tasks including extraction of metadata, assurance of data integrity, capture of contextual information, documentation of chain of custody, and identification of sensitive personal information for restriction or redaction. In recent years, several institutions and projects have focused on applying digital forensics to acquisition and management of born-digital materials in collecting institutions - libraries, archives and museums (LAMs). This work has yielded helpful guidance documents, examples of workflows, and more recently, an open-source software environment (BitCurator) designed for use by LAMs. While there are a variety of efforts to coordinate activities across projects and institutions, there is still great potential for further alignment.

One promising area of alignment is in the provision of information about digital storage media. The first time that one acquires or examines an unfamiliar or obsolete type of storage medium, there can be a steep learning curve in becoming familiar with the medium. This can involve an understanding of physical connectors, power plugs, drives and enclosures, but also likely failure modes and dependencies on specific hardware, software, and firmware in order to read data off the device. There have been efforts to pool information related to media to serve as a common professional resource, including Mediapedia at the National Library of Australia and the Trustworthy Online Technical Environment Metadata (TOTEM) Registry work of the KEEP project. The Computer Product Manuals Collection at the Charles Babbage Institute can also serve as a useful resource. The work of LAM institutions could benefit substantially from further efforts to organize, manage and disseminate such information.

Once the hardware issues have been resolved, one can create a sector-by-sector copy of the data from the medium, generating what is common referred to as a disk image. There are many digital forensics tools that can then perform analysis and data extraction actions on the data in the disk image. However, most current forensics software is designed to work with a relatively small set of common filesystems, such as FAT and NTFS (Windows), ext (Unix) and HFS+ (Mac). Not only will many tools not be able to read older and less common filesystems, but the tools will not even be able to detect what filesystems they are. A great contribution would be a body of shared information about how to detect various encoding schemes and filesystems on disks. For example, what distinct pattern of signals or bits appears at the beginning of Commodore 64 disk (CBMFS filesystem) or older Macintosh disks (HFS filesystem)? By sharing this information, LAM professionals would be better able to conduct basic triage on their collections and develop strategies for what do with the media in their care.

A much wider set of considerations relate to pattern detection more generally. For example, there are a variety of algorithms and regular expressions that one can use to identify file types at the file live (e.g. using headers and extensions) or sector level (e.g. using end-of-line markers), as well as other features in the data such as credit card numbers or other personally identifying information that warrants redaction or restriction. Rather than reinventing such algorithms and expressions each time they are needed, LAM professionals could benefit from sharing and collaborative development of pattern detection methods related to common curatorial tasks.

Finally, there are various opportunities related to metadata. LAM professionals want not only to extract information from disks, but also to incorporate the information into their collection management and access environments. Several open-source digital forensics tools share a common set of metadata elements -- Digital Forensics XML (DFXML) -- and LAMs take advantage of those conventions when incorporating metadata into their systems. However, mappings from DFXML to existing LAM metadata schemes are still in early development, and there is great potential for further work on cross-walks, transformations and application profiles for given settings and situations. LAMs would be further served by a standardized DFXML schema that would provide a baseline for the development of applications able to generate metadata for less common filesystems.

We believe that it will be increasingly important and beneficial for the digital preservation community to more actively work with disk images as entities in their own right. This is a promising area for international collaboration and coordination.