Up Close and Personal: Individual Digital Traces as Cultural Heritage and Discovery through Forensics Tools

Christopher A. Lee
School of Information and Library Science
University of North Carolina
216 Lenoir Drive, CB #3360
1-(919)-962-7024
callee@ils.unc.edu

ABSTRACT
The documentary traces of individuals (personal traces) have long been recognized and preserved as fundamental components of cultural heritage. They serve as the most personalized possible cultural heritage, reflecting individual behaviors, documenting shared experiences and shaping individual and collective senses of identity. The nature of personal documentary traces has undergone dramatic evolution in recent years, including various aspects of one's "digital footprint." Many cultural institutions have begun applying digital forensics to create authentic copies of data on disks; reflect the original order of materials; establish more trustworthy chains of custody; discover and expose associated contextual information; and identify sensitive information that should be filtered, redacted or masked in appropriate ways. Many of the same approaches can be adapted and applied by individuals and families who are managing their own collections of personal traces. This demonstration will illustrate features of the open-source BitCurator environment that allow individuals to discover and navigate the personal traces of themselves and others.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—collection, dissemination, systems issues.

General Terms
Digital Forensics, Personally Identifying Information, Cultural Heritage

Keywords
Personal traces, digital heritage, BitCurator

1. PERSONAL TRACES AS THE MOST PERSONALIZED CULTURAL HERITAGE
The documentary traces of individuals (what I will call personal traces) have long been recognized and preserved as fundamental components of cultural heritage. The traces have traditionally taken the form of personal mementos such as postcards, photographs and other artifacts related to specific life events, as well as products of more sustained personal activities such as diaries, journals, and series of correspondence.

Historically, individuals and families have accumulated and managed collections of personal traces. Many of the materials have been lost due to the "hazards of time" [10], but others have been maintained across multiple generations, based on the curatorial efforts of private individuals. These personal traces serve as the most personalized possible cultural heritage: primary sources created and maintained by oneself and one's loved ones. They reflect individual behaviors, document shared experiences and shape individual and collective senses of identity. Personal traces serve as raw materials for the stories that people tell about themselves, and curatorial control over the traces can significantly impact one's ability to influence the stores that others can tell.

Collections of personal traces form a few prominent individuals and families have made the transition into collecting institutions (libraries, archives and museums). These collections tend to be given labels such as "manuscripts," "personal papers," or "personal archives" of those individuals or families. Many cultural institutions were initially seeded by personal collections of influential people.

Much of the public's experience of cultural heritage depends on the personal traces of others. In some cases, individuals encounter them directly as artifacts in museum exhibits. In other cases, the traces are quietly in the background as source or reference material for exhibits, historical reenactments, news reports, popular books and films (historical fiction and documentaries). In still other cases, members of the public make direct use of individual traces by consulting them as researchers in archival repositories.

The archival literature uses the term *fonds* to refer to "the entire body of records of an organization, family, or individual that have been created and accumulated as the result of an organic process reflecting the functions of the creator" [8]. In the case of personal archives, "the *fonds* of an individual is a site where personality and the events of life interact in documentary form" [6]. The *fonds* is an "intellectual construct" rather than a "physical entity" [2]. The materials that constitute an individual's *fonds* are often not co-located and are often distributed across various systems.

Thinking of the aggregate of all one's personal traces – even if they never reside in the same place all at once – is a powerful way to bridge the practices and needs of individuals caring for their

PATCH, February 24, 2014, Haifa, Israel

own materials with the practices and needs of professionals working in cultural institutions. In both cases, it can be important to understand the scope of what traces are there, carry associated contextual information across time in order to support meaningful use of the traces, and identify potentially sensitive information that should not be disclosed or shared with others except in limited cases.

2. DIGITAL TRACES AND INCREASING PERSONALIZATION

The nature of personal documentary traces has undergone dramatic evolution in recent years. Some changes have involved adoption of digital technologies that closely reflect previous documentary forms, e.g. digital photographs, email correspondence. Other changes involve documentary forms that have little precedent, e.g. server logs, tags on photos, relational data in social media.

In a digital environment, one could think of *fonds* as an individual's "digital footprint" broadly conceived. This can include various traces left behind by an individual both on specific computer devices (e.g., desktop computers, laptop computers, tablets, phones, external storage media) and in networked resource spaces (e.g., blogs, email accounts, Twitter feeds).

In order for individuals to have persistent access to the personal traces that they have left across various digital environments, they must make a concerted effort to capture and maintain the traces themselves. They cannot rely solely on the providers of online services who do not have the incentive – and often do not have the architectural capability – keep static representations of previous system state (e.g. old user profiles, expired profile information, previous "friend" relationships). Capturing information from any given digital environment is usually feasible, but an effort to collect all of one's traces can involve a dizzying array of protocols, application programming interfaces (APIs), screen scraping tools, and data format.

Another new aspect of one's *fonds* is the evidence of what personalized slice of the wider information universe was experienced by a given individual. If someone wants to recreate a past session of visiting a web site, she often cannot do this by using someone else's computer. Various aspects of the sites' layout, rendering and content can depend on her operating system, browser, browser settings, internet protocol address (IP), browsing history, cookies stored on her computer and user account information.

3. DISCOVERY AND NAVIGATING PERSONAL TRACES ON STORAGE MEDIA

Many past and current personal traces reside on digital storage media, including floppy disks, Zip disks, optical media, hard drives, solid state drives, among others. Over the past several decades, the field of digital forensics has developed principles, methods and tools for the extraction, management and analysis of data from storage media. The data include not just the immediately visible contents of files but also embedded metadata, configuration settings, system logs, deleted files, and a variety of other "hidden" traces.

A substantial portion of an individual's *fonds* can reside not only on storage media within the care of that individual but also within a diverse array of online spaces. As discussed above, there are often ways to pull information out of those spaces, and it is important for individuals and cultural institutions to pursue those opportunities to collect traces that have long-term value. However, these processes can be complicated, brittle, error-prone and incomplete.

An important way to find and preserve personal traces of online activity is to apply digital forensics methods to the data that reside on digital storage media [5]. For example, one can generate lists of email addresses and URLs appearing on a drive, as well as various artifacts associated with specific online spaces, e.g. login names, social network profiles. See Figure 1 for an illustration of using a tool called bulk_extractor (developed by Simson Garfinkel) within the BitCurator software environment to display and navigate email addresses that appeared on a disk within their original context.

4. APPLICATION OF FORENSICS TO CULTURAL HERITAGE

In recent years, many cultural institutions have begun applying digital forensics tools and methods to the materials within their care [7]. This allows them to create authentic copies of data on disks; reflect the original order of materials; establish more trustworthy chains of custody; discover and (when appropriate) expose associated contextual information; and identify sensitive information that should be filtered, redacted or masked in appropriate ways.

Many of the same approaches being taken by cultural institutions can be adapted and applied by individuals and families who are managing their own collections of personal traces. Many of the media in their care (e.g. floppy disks) are at risk of becoming unreadable and their content is not serving the interests of personal cultural heritage if they are not or cannot be accessed. Forensics methods can allow them to make copies of the data from the media, extract associated metadata such as timestamps that reflect when files were last used, as well as searching and navigating the content to discover aspects of their own personal traces.

5. BITCURATOR ENVIRONMENT

The BitCurator Project, a collaborative effort led by the School of Information and Library Science at the University of North Carolina at Chapel Hill and Maryland Institute for Technology in the Humanities at the University of Maryland, is are developing and disseminating a suite of open source tools that allow individuals to apply digital forensics methods to cultural heritage materials.[1] These tools are currently being developed and tested in a Linux environment; the software on which they depend can readily be compiled for Windows environments (and in most cases are currently distributed as both source code and Windows binaries). We are freely disseminating the software under an open source (GPL, Version 3) license. BitCurator provides users with two primary paths to integrate digital forensics tools and techniques into archival and library workflows.

[1] The software, documentation and associated guidance resources are all available at http://wiki.bitcurator.net.

1) The BitCurator software can be run as a ready-to-run Linux environment that can be used either as a virtual machine (VM) or installed as a host operating system. This environment is customized to provide users with graphic user interface (GUI)-based scripts that provide simplified access to common functions associated with handling media, including facilities to prevent inadvertent write-enabled mounting (software write-blocking).

2) The BitCurator software can be run as a set of individual software tools, packages, support scripts, and documentation to reproduce full or partial functionality of the ready-to-run BitCurator environment. These include a software metapackage file that replicates the software dependency tree; a set of software sources and supporting environmental scripts developed by the BitCurator team and made publicly available at via our GitHub repository; and all other third-party open source digital forensics software included in the BitCurator environment.

In our experience, the most approachable option for those just getting started with these tools is the virtual machine, which present the user with a pre-packaged and self-contained computer environment that can be downloaded and run for free on all major operating systems by using VirtualBox (detailed instructions are available at http:wiki.bitcturator.net).

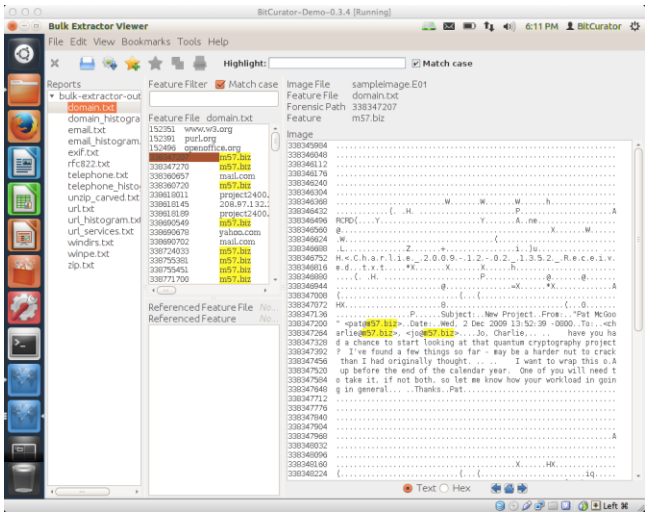

Figure 1 - Using bulk_extractor to view email addresses appearing on a disk

6. DEMONSTRATION AND FEATURED TOOLS

Tools that BitCurator is incorporating include (among others) Guymager, a program for capturing disk images; bulk extractor, for extracting features of interest from disk images (including private and individually identifying information); fiwalk, for generating Digital Forensics XML (DFXML) output describing filesystem hierarchies contained on disk images; The Sleuth Kit (TSK), for viewing, identifying and extraction information from disk images; exiftool, for viewing and extracting EXIF metadata from within digital photographs; Nautilus scripts to automate the actions of command-line forensics utilities through the Ubuntu desktop browser; DuplicatesDeletion, for quickly identifying duplicate files within a directory; and sdhash, a fuzzing hashing application that can find partial matches between similar files. For further information about several of these tools, see [1,3,4,9].

This demonstration will illustrate the specific BitCurator features that can allow individuals to discover and navigate the personal traces of themselves and others. Particular emphasis will be placed on the ability to view embedded metadata in files and traces of online activities, including email addresses, URLs, and social network service information.

7. CONCLUSIONS

One of the most powerful ways to personalize cultural heritage is to expose individuals to the personal traces of themselves and their loved ones. These traces can matter deeply for perpetuating their own stories, finding their own place within a cultural context, and deepening their appreciation for cultural heritage more generally. This demonstration will illustrate specific, tangible ways to interact with these most personal forms of cultural heritage, ranging from artifacts of the latest online interactions to long-forgotten documents stored on disks.

8. ACKNOWLEDGMENTS

BitCurator development has been supported by the Andrew W. Mellon Foundation. Members of the BitCurator team are Alexandra Chassanoff, Matthew Kirschenbaum, Christopher (Cal) Lee, Sunitha Misra, Porter Olsen, and Kam Woods. Members of two advisory boards have made valuable contributions: the Development Advisory Group (DAG) and Professional Experts Panel (PEP).

9. REFERENCES

[1] Cohen, M., Garfinkel, S., and Schatz, B. 2009. Extending the Advanced Forensic Format to Accommodate Multiple Data Sources, Logical Evidence, Arbitrary Information and Forensic Workflow. Digital Investigation 6 (2009), S57-S68.

[2] Cook, T. 1993. The Concept of Archival Fonds and the Post-Custodial Era: Theory, Problems and Solutions, *Archivaria* 35, 33.

[3] Garfinkel, S. 2012. Digital Forensics XML and the DFXML Toolset. Digital Investigation 8, 161-174.

[4] Garfinkel, S.L. Providing Cryptographic Security and Evidentiary Chain-of-Custody with the Advanced Forensic Format, Library, and Tools. International Journal of Digital Crime and Forensics 1, 1 (2009), 1-28;

[5] Garfinkel, S. and Cox, D. 2009. Finding and archiving the internet footprint. In First Digital Lives Research Conference: Personal Digital Archives for the 21st Century, London, UK, February 9-11.

[6] Hobbs, C. 2001. The Character of Personal Archives: Reflections on the Value of Records of Individuals, *Archivaria* 52, 127.

[7] Lee, C.A., Woods, K., Kirschenbaum, M. and Chassanoff, A. 2013. From Bitstreams to Heritage: Putting Digital Forensics into Practice in Collecting Institutions. http://www.bitcurator.net/docs/bitstreams-to-heritage.pdf

[8] Pearce-Moses, R. 2005. *Glossary of Archival and Records Terminology*, Society of American Archivists, Chicago, IL, 173.

[9] Roussev, V. 2011. An Evaluation of Forensic Similarity Hashes. Digital Investigation 8, S34-S41.

[10] Ward, W. P. 1982. Family Papers and the New Social History. *Archivaria,* 14, 63.