IJDC | *General Article*

Integrating Digital Forensics Techniques into Curatorial Tasks: A Case Study

Sam Meister
University of Montana

Alexandra Chassanoff
University of North Carolina

Abstract

In this paper, we investigate how digital forensics tools can support digital curation tasks around the acquisition, processing, management and analysis of born-digital materials. Using a real world born-digital collection as our use case, we describe how BitCurator, a digital forensics open source software environment, supports fundamental curatorial activities such as secure data transfer, assurance of authenticity and integrity, and the identification and elimination of private and/or sensitive information. We also introduce a workflow diagram that articulates the processing steps for institutions processing born-digital materials. Finally, we review possibilities for further integration, development and use of digital forensic tools.

Received 21 July 2014 | *Accepted* 9 September 2014

Correspondence should be addressed to Sam Meister, University of Montana, 32 Campus Drive, Missoula, MT 59812. Email: sam.meister@mso.umt.edu

An earlier version of this paper was presented at 9th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The acquisition of data on electronic media presents unique challenges for archivists and information professionals in cultural heritage organizations. The extraction, safe transfer, long-term preservation, and provisional access of data and metadata from physical media (floppy disks and CD-ROMs, USB drives and other fixed and removable media) to durable storage is a fundamental step in processing workflows. A growing body of open source software tools designed and developed in collaboration with the library and archives communities provide new solutions to meet this data challenge. These software tools are relatively new options for these communities; there are limited best practices or examples illustrating how such tools can be integrated into processing workflows.

Digital Curation and Digital Forensics

The increase in personal computing use throughout the 1980s and 1990s resulted in massive amounts of electronic information being generated and stored on individual desktops (and later, laptops). At the same time, the spread of computer virus attacks and worms brought focused attention to the mounting problem of "cyber-crimes." The Internet Virus of 1988, for example, resulted in federal hearings aimed at assessing the vulnerability of government telecommunication networks (Brock, 1989). Awareness that forensic evidence documenting cyber-crimes could be captured on hard drives and in other digital traces helped to propel the development of a community of practice known as digital forensics. The few texts on the history of the term date its origin to 1976, with the publication of Donn Parker's *Crime by Computer*, which provided "perhaps the first description of the use of digital information to investigate and prosecute crimes committed with the assistance of a computer" (Pollitt, 2010). More recently, digital forensics has been defined as an applied practice involving the "preservation, identification, extraction, documentation, and interpretation" of digital data as legal evidence (Kruse & Heiser, 2002).

The term digital curation was first used in 2001 as a means for describing the actions and activities related to the creation, maintenance and ongoing preservation of electronic information. A more precise definition was established by the Digital Curation Centre (DCC), whose charter states that digital curation can be defined as "maintaining and adding value to a trusted body of digital research data for current and future use; it encompasses the active management of data throughout the research lifecycle" (DCC, n.d.). Digital curation can encompass a wide umbrella of activities, ranging from documenting preservation actions undertaken on digital objects to migrating data into new formats.

Despite the acknowledged importance of digital curation practices to data stewardship in cultural heritage institutions, research projects examining possibilities and obstacles in processing born-digital materials have only recently begun to emerge. In 2009, the Maryland Institute for Technology in the Humanities (MITH) led an NEH-funded collaboration with two institutional partners to evaluate processing approaches for three institutionally significant born-digital literary collections (Kirschenbaum et al., 2009). A subset of the group went on to publish an influential Council on Library and

 doi:10.2218/ijdc.v9i2.325

Information Resources (CLIR) report, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections,* in 2010 (Kirschenbaum et al., 2010). The AIMS project, another collaborative approach among institutions, sought to summarize challenges while also providing a model framework for archivists processing born digital materials. The research project's white paper outlined institutional objectives, decision points and associated tasks in institutional workflow (AIMS Working Group, 2012).

Challenges in Workflow Integration

There is growing recognition that digital forensic techniques and methods can provide critical lifecycle information about born-digital materials for archivists and scholars. However, there are a number of technical and administrative challenges that make implementing such tools into processing workflows difficult. Institutions must make a number of policy decisions in order to effectively manage digital objects. Some of these include:

- How should such items, facing media obsolescence, be stored?
- What steps should be taken if the media itself cannot be read?
- When transferring files from media, should we create a bitstream or logical disk image?
- In providing access to materials, should we work at the level of the disk or describe individual items?

The iterative nature of existing institutional workflows can make streamlining processes challenging. In a 2012 survey conducted with institutions working with born-digital materials, one respondent summed up the problem thusly:

> 'We're doing things this way now, but we know that it's not the right way, and let's just wait and see what happens, because maybe they'll come up with something better' (Gengenbach, 2012).

The speed of processing tools and associated prohibitive costs can also be cumbersome. Analyzing media sources, an important function for many institutions acquiring entire collections of born-digital artifacts, can often take several days with existing tools (Knight, 2012). Institutions may not have the budgetary resources (both in terms of personnel training and equipment) necessary to implement workflows that accommodate digital forensic practices.

Integrating Tools into Workflows: The BitCurator Project

One research project that has emerged from previous research efforts mentioned above is the BitCurator project. A joint effort led by the School of Information and Library Science at the University of North Carolina, Chapel Hill (SILS) and the Maryland Institute for Technology in the Humanities (MITH), BitCurator is an open source digital forensics environment that incorporates a variety of functionality and processing for born-digital materials. BitCurator is a fully functioning Linux system built on Ubuntu 12.04 that has been customized to meet the needs of digital archivists, and it can be run

either as a stand-alone operating system or as a virtual machine. The project began its second phase of funded development in October 2012.

BitCurator acts in two critical curatorial capacities. First, it addresses traditional archival concerns within a born-digital context. Software components perform a variety of tasks related to authenticity, provenance and trustworthiness. Custom-built Nautilus scripts, for example, can generate hash values for all files on a hard drive, which can then be used as an audit trail to ensure the integrity of materials. Fiwalk and bulk extractor can quickly and efficiently parse large amounts of digital data to find specific items of interest. A built-in python module, iredact.py, can identify and redact sensitive, personally identifiable information from the disk. The use of a write-blocker can serve as intermediary protection to ensure that source media is not overwritten. While the detection of duplicate objects is helpful, fuzzy hashing can also be fruitful to find similar (but not identical) objects in a data repository. Finally, the software produces a DFXML output. DFXML is a language for describing forensic processes and associated metadata. Though still in early stages of development, DFXML can be easily packaged into existing METS containers, a commonly used encoding standard for managing digital objects in repositories.

BitCurator also captures and records essential attributes and characteristics about born-digital media for future curation activities. Increasingly, how we work as scholars will be captured in our "use" of digital tools - the digital environment in which media are created, altered, deleted and accessed. For example, consider an institution acquiring an author's archive of born-digital literary works. BitCurator contains tools that can enable the reconstruction of digital research environments, including documentation of research traces and activities on the web. One can imagine that future researchers might be interested in contextual factors that lend insight into scholars' research activities and their everyday lives. The tool bulk extractor, for example, can provide a histogram of all the URL domains found on a hard drive. Such rich metadata will undoubtedly facilitate future understanding of scholarly research environments. BitCurator also acknowledges that recording preservation activities are critical to the digital curation lifecycle model. Using the Library of Congress metadata standard known as PREMIS, BitCurator documents and records the preservation actions undertaken during the acquisition, processing, and analysis of the born-digital object. The XML output can then be packaged alongside the digital object and stored in a repository.

Case Study: Curating an Archival Collection using BitCurator

The collection we have chosen offers an exemplary use case for archives interested in applying digital forensics tools and methods. It consists of a typical variety of media and data formats, including floppy disks, zip disks, CD-ROMs and USB drives. The collection also contains private, personally identifying correspondence that will need to be restricted for a designated period of time. By providing a practical model for cultural heritage institutions, we seek to contribute to a growing set of best practices for integrating digital forensics methods and tools into digital curation workflows.

 doi:10.2218/ijdc.v9i2.325

Institutional Context

The Maureen and Mike Mansfield Library at the University of Montana in Missoula, Montana has been developing policies, workflows and infrastructure to support the processing of born-digital materials since 2011. Most of these activities have taken place within the Archives and Special Collections department of the Library. To date, the design, development, testing and implementation of these activities have been carried out primarily by the Digital Archivist. This work began with the development of an initial policy framework and workflow requirements, and proceeded with testing and implementation of hardware and software to support processing of born-digital materials acquired as part of archival collections.

The decision to pursue installation of open source software as part of these activities was made in large part due to the relatively low resources needed to initially test and analyze these tools against workflow requirements. The motivation to install, test and initially implement the BitCurator software environment was driven by this same low-resource commitment factor, along with a desire to contribute to the continued development of the tool through testing and analysis in the context of processing an existing collection of born-digital materials. Additionally, the relatively minimal testing and implementation requirements of open source tools, such as BitCurator, has provided an opportunity to move forward with the implementation of a workflow to process born-digital materials acquired within archival collections, while in parallel a much larger digital preservation program is being developed to support the long-term preservation of and access to the Library's wide array of digital content.

Collection Description

The case study collection consists of the born-digital materials created by Patricia Goedicke, a poet and professor of English at the University of Montana. The born-digital materials constitute a portion of a much larger collection, the Patricia Goedicke and Leonard Wallace Robinson Papers, that primarily consists of correspondence, notebooks, drafts of poetry and fiction, lecture notes for classes, personal, and business records. Types of content include emails, word processing documents, spreadsheets and image files. The born-digital materials were originally acquired on various forms of electronic storage media, including 5.25 inch and 3.5 inch floppy disks, Zip disks, CDs, DVDs, and a USB flash drive. The analog materials within the collection have been arranged and described, with the born-digital materials organized into a separate archival series. During the processing of the paper materials, a decision was made to restrict some correspondence materials in order protect the privacy of living individuals. This is a common access restriction strategy within archival collections, and the implications of implementing this strategy in the context of working born-digital materials will be discussed below.

Workflow Description

To illustrate the application of digital forensics methods and tools to a real world collection, the series of workflow steps followed to process the collection will be described in sequential order. These workflow steps are visually represented within the associated workflow diagram in Figure 1.

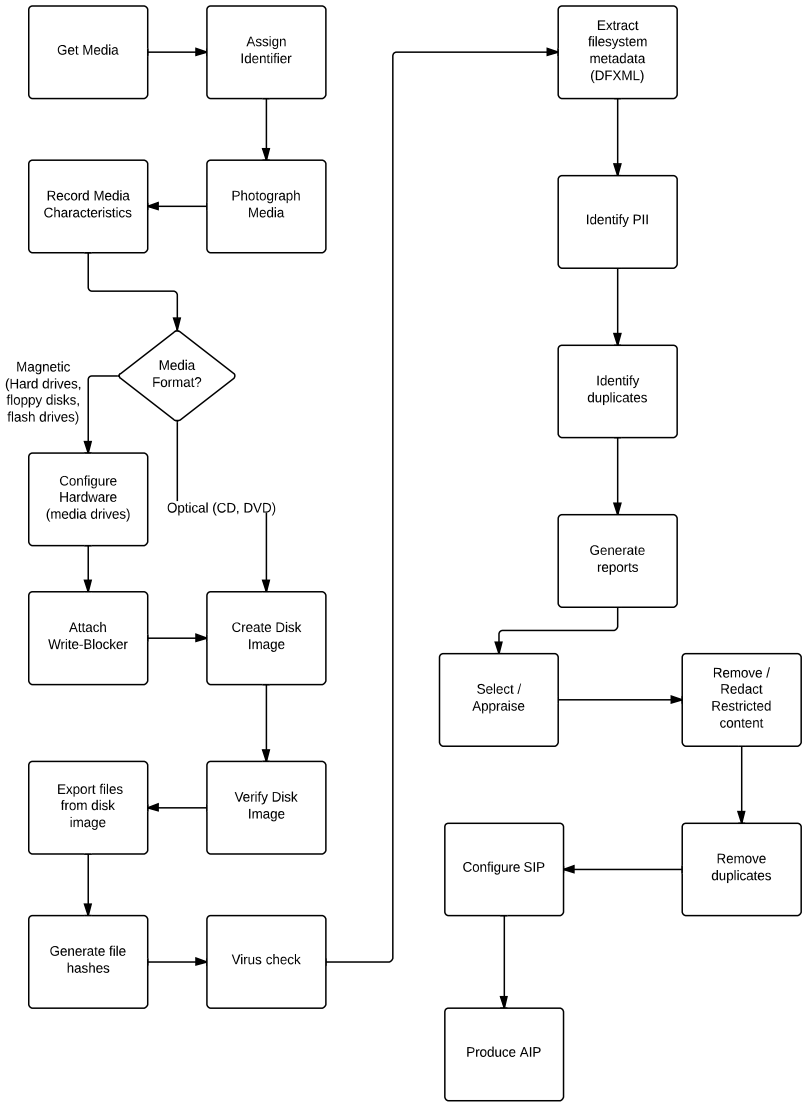

Figure 1. Case study workflow diagram.

Document media

The first step in the accession workflow is to document the physical electronic media item to gain an initial level of administrative control over the item. A record for each piece of electronic media is added to a local media log database. The purpose of the media log is to document the physical characteristics of the media item and the various tasks that are performed on that media item during the accession process. This record includes an identifier that is assigned to each media item. This identifier functions as an extension of the higher level accession identifier previously assigned to the larger collection of both analog and digital materials. Additional characteristics documenting the physical media item are then added to the record, including the media format and type, manufacturer, model, age, condition and label text.

Create disk image

The next step in the accession workflow is to create a disk image of the media item. To prepare for disk imaging, hardware and software needed to acquire data from the particular media item are configured. For the case study collection, this included an array of media drives and software.

 doi:10.2218/ijdc.v9i2.325

Table 1. Disk image hardware and software.

Media	Drive (hardware)	Software
5.25 inch floppy disk	External 5.25 inch floppy drive with FC5205 USB controller	FC5205 Disk Image and Browse
3.5 inch floppy disk	External 3.5 inch floppy USB drive	FTK Imager
CD/DVD	Internal CD/DVD drive	FTK Imager
USB Flash drive	Internal USB drive	FTK Imager/Guymager

An additional element in this configuration task includes enabling write-blocking for particular media items to ensure data is not inadvertently modified during the disk imaging process. Write-blocking functionality was achieved in a variety of different ways for the case study collection.

Table 2. Write-block mechanism.

Media	Write-Block Mechanism
5.25 inch floppy disk	FC5205 USB controller designed to be read-only
3.5 inch floppy disk	Physical switch of tab on disk to read-only
CD/DVD	Automatic
USB flash drive	Tableau T8-R2 USB Forensic Bridge

Once hardware configuration is complete, disk imaging can proceed. An initial decision has been made to create forensic disk images in the Advanced Forensics Format (AFF) due to its ability to store disk images and associated metadata in an open and extensible format. For the case study collection, AFF disk images were created with the FTK Imager software for the 3.5 inch floppy disks and using the Guymager software within the BitCurator environment for the USB flash drive. The reason behind the use of two kinds of software was based on timing and sequence of events, rather than functionality or preference. Disk imaging of the floppy disk media in the collection began before initial installation, testing and implementation of the BitCurator software environment. After basic testing of BitCurator was completed it was used to create disk images for the remaining USB flash drive media. An initial attempt to use FTK Imager to create AFF disk images of the 5.25 inch floppy disks proved unsuccessful, as FTK Imager was unable to recognize the filesystems on these disks. As a result, the FC5205 Disk Image and Browse software was used to create IMG disk images for the 5.25 inch floppy disks. For the small amount of optical media, ISO disk images were created.

Export files

After a disk image is created for a piece of physical media, the files are then exported from the disk image. The motivation behind this step in the accession workflow is to provide flexibility in future preservation actions and decision-making. An interim decision has been made to include both disk images and exported files as elements of an Archival Information Package that will be created in further processing steps. The inclusion of both versions of this data allows for a potential future access scenario in which users are given direct access to the preserved disk images. The act of

exporting files supports flexibility in additional processing steps, including arrangement, description and selection decision-making. In our case study collection there are numerous examples of files not produced by the creator, such as system files and software support files. A selection decision may be made to not provide access to these types of files, resulting in a smaller set of materials that will be arranged and described for access. The step of exporting files may be re-evaluated in the future as details of the access strategy are defined.

Initial analysis

It is during the Initial Analysis accession workflow step that the functionality of the BitCurator software environment was utilized to meet workflow requirements. The main goals of this workflow step are to produce outputs that both prepare the data and metadata for long-term preservation and access, and support additional selection, arrangement and description decision-making.

Extract filesystem metadata

The first task in the Initial Analysis step is to extract filesystem metadata from the disk image to capture the details about the file structure for the specific media item. Extracting the filesystem details and original structure of file objects as preservation metadata is a key step in documenting contextual details about the technical environment in which file objects were created and/or managed. BitCurator uses the fiwalk tool to process disk images and export filesystem metadata as DFXML. From the user perspective, this task is easily accomplished via a simple user interface that involves selecting the disk image to process and the output storage location for the resulting XML file. The output of the DFXML file is a fundamental component in performing additional analysis tasks within BitCurator.

Identify private, personal or sensitive information

The next accessioning workflow task is to attempt to identify the existence of private, personal or sensitive information within the contents of disk images and/or files. This is an important step in generating data that may inform future processing decisions, such as determining the need for removal, redaction, or restriction of specific data or files in a collection. Understanding the nature of how private and sensitive information is created and stored is key to developing workflow requirements to identify this information. As Lee & Woods (2013) explain:

> ‘Modern computing devices often contain a significant amount of private and sensitive information. The information may appear embedded in document content and document metadata, within system files obscured by operating system arcana, or in traces of content held within local storage systems after interaction with services over a network. It may even inadvertently be retained on disk or in static memory after a device has been retired or formatted without being overwritten.’

Performing an initial investigation to identify private or personal information is critical to ensuring that such information is not inappropriately discovered and accessed.

In the case study, the bulk_extractor tool within BitCurator was used to investigate and identify the existence of private and personally identifying information. bulk_extractor operates by identifying “features of interest” within a disk image, bitstream, or location within a live file system. Typical features identified include

 doi:10.2218/ijdc.v9i2.325

'textual patterns with a specific semantic interpretation, e.g. Social Security numbers, email headers, image and geolocation metadata, and user activity history such as URLs corresponding to search results' (Lee, Woods, Kirschenbuam, & Chassanoff, 2013). To configure bulk_extractor to identify "features" on a disk image, the user selects specific types of scanners, such as the "accts" scanner that searches for credit card numbers and telephone numbers. If any data matching the designated feature criteria are found, these will be included in the output feature files in the form of .txt files. For the case study collection, all default scanners were selected to initially search disk images for any private or personally identifying information. This resulted in a number of feature report files identifying the existence of telephone numbers, email addresses, etc. within the collection. The details of these feature report files can be further explored using bulk_extractor, allowing the user to determine the location of features within specific files in a disk image. Running the default scanners within bulk_extractor produced valuable data that will inform further processing activities.

In addition to this first pass at identifying private and personally identifying information, a deeper level of investigation was required for the case study collection. The access restrictions determined for correspondence materials of specific individuals within the analog materials necessitated an attempt to determine if there were any materials related to those same individuals within the born-digital materials. An initial attempt to locate materials related to these individuals was carried out by searching for these names using the bulk_extractor tool. This task proved to be quite time-consuming, as there was not an obvious mechanism within the user interface to search for specific text across multiple disk images in the collection. A second attempt to search for specific names was performed by using the custom Nautilus file manager scripts within the BitCurator environment. These scripts include an option to search files by name or content. Searching for the names of individuals across all the files that had been previously exported from their original disk images resulted in the successful retrieval of a smaller set of files that could be explored further. Additional investigation would involve opening files with required software or viewing with a hex editor to confirm any relation to the specific individuals.

Generate human-readable reports

The final task in the accession workflow process is to generate a set of reports summarizing and presenting the data produced in previous tasks in a human-readable form. Within BitCurator, these reports are easily created via the user interface, compiling the outputs of the fiwalk and bulk_extractor tools into series of PDF documents. For the case study collection, PDF reports were generated for each corresponding disk image. The primary function of these PDF reports is to be able to share important information about the born-digital materials with other processing archivists. Generating reports that include visual elements, such as tables and graphs, provides a way to share this data outside of the BitCurator environment. As with the case study collection, future born-digital acquisitions are likely to continue to be a part of hybrid analog and digital collections. Data generated through the BitCurator software environment will assist in informing selection, arrangement, description and access restrictions, along with information collected through analog processing tasks.

Conclusion and Future Work

The results of the integration of BitCurator to process a real world collection of born-digital materials demonstrates the significant value this software environment has to offer to institutions seeking to integrate digital forensics methods into curation workflows. For the case study collection, BitCurator met all accession workflow requirements to prepare born digital materials for long-term preservation and access, and generate information to support additional selection, arrangement and description activities. Workflow tasks were easily executed through BitCurator's simple graphical user interface, an important feature in the overall design of the software tool. While other digital forensics software is available, the majority of these have been designed for criminal investigations seeking to locate specific pieces of evidence. In contrast, BitCurator offers archivists, librarians, data curators and researchers an easy-to-use interface and functionality that can be quickly integrated into existing processing workflows to meet digital curation needs.

The processing of the case study collection also revealed some potential areas for additional development. The need to search for and identify the existence of specific private and/or restricted information within a collection of born-digital materials is a common scenario and a key workflow step to be carried out before providing access. For archival collections, the restriction of information related to specific names or terms may be determined through negotiation with creators and/or donors and included in deposit agreements. Through custom file manager scripts, BitCurator currently provides an easy way to search quickly across all relevant files in a collection. An additional improvement to this search functionality could include a feature to quickly locate and view the specific search term *within* particular files.

Another potential development feature could be the incorporation of data visualization tools. The current default reports in BitCurator offers a bar graph representation of file formats within a disk image. Additional data visualization features that allow users to visually analyze data contents (dates, names, etc) could provide valuable information to assist archivists and curators in determining selection, arrangement, description and access decisions.

Acknowledgements

The BitCurator project is funded through the Andrew W. Mellon Foundation. The principle investigator is Christopher Lee and the co-principle investigator is Matthew Kirschenbaum.

References

AIMS Working Group. (2012). *AIMS born-digital collections: An inter-institutional model for stewardship*. Retrieved from http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final.pdf

Brock, J.L. & United States. (1989). *November 1988 Internet computer virus and the vulnerability of national telecommunications networks to computer viruses: Statement of Jack L. Brock, Director, Government Information ... before the Subcommittee on Telecommunications and Finance, Committee on Energy and Commerce, House of Representatives*. Washington, D.C.: United States General Accounting Office.

Digital Curation Centre. (n.d.). DCC charter and statement of principles. Retrieved from http://www.dcc.ac.uk/about-us/dcc-charter

Gengenbach, M. (2012). *"The way we do it here": Mapping digital forensics workflows in collecting institutions."* Unpublished master's thesis, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Kirschenbaum, M.G., Farr, E., Kraus, K.M., Nelson, N.L., Stollar Peters, C., Redwine, G., & Reside, D. (2009). *Approaches to managing and collecting born-digital literary materials for scholarly use.* College Park, MD: University of Maryland. Retrieved from http://hdl.handle.net/1903/9787

Kirschenbaum, M G., Ovenden, R. & Redwine, G. (2010). *Digital forensics and born-digital content in cultural heritage collections.* Washington, D.C.: Council on Library and Information Resources. Retrieved from http://www.clir.org/pubs/abstract/reports/pub149

Knight, G. (2012). The forensic curator: Digital forensics as a solution to addressing the curatorial challenges posed by personal digital archives. *International Journal of Digital Curation 7*(2), 40-63.

Kruse, W. G., & Heiser, J. G. (2002). *Computer forensics: Incident response essentials*. Boston, Mass: Addison-Wesley.

Lee, C.A., & Woods, K. (2013). Automated redaction of private and personal data in collections: Toward responsible stewardship of digital heritage. In L. Duranti and E. Shaffer (Eds.), *Proceedings of Memory of the World in the Digital Age: Digitization and Preservation: An International Conference on Permanent Access to Digital Documentary Heritage, 26-28 September 2012, Vancouver, British Columbia, Canada.* United Nations Educational, Scientific and Cultural Organization.

Lee, C., Woods, K., Kirschenbaum, M.G., & Chassanoff, A. (2013). *From bitstreams to heritage: Putting digital forensics into practice in collecting institutions.* Retrieved from http://www.bitcurator.net/docs/bitstreams-to-heritage.pdf

Pollitt, M. (2010). A history of digital forensics. In K.P. Chow & S. Shenoi (Eds.), *Advances in digital forensics VI: Sixth IFIP WG 11.9 International Conference on Digital Forensics, Hong Kong, China, January 4-6, 2010, revised selected papers*. Berlin: Springer.