Aggregating Temporal Forensic Data Across Archival Digital Media

Walker Sampson
University Libraries
University of Colorado Boulder
Boulder, Colorado, United States
walker.sampson@colorado.edu

***Abstract*—In this paper, I introduce a method for generating an aggregated timeline of file system activity derived from the disk images of archival digital media. Using a collection of 1,059 floppy disks from the Woody and Steina Vasulka collection as a case study for this process, I evaluate the technical issues of the dataset and describe the variances in date and time use for different file systems and the operating systems that use them. I discuss the utility of the timeline as a research and archival aid, and conclude that such timelines are promising resources which can provide a wider evaluative context for collections of digital media.**

***Index Terms*—data analysis; data collection; data integration; file systems; forensics**

I. Introduction

Practitioners of digital preservation have seen an encouraging rise in tools, techniques, and debate in communities charged with the preservation of born-digital media. Digital forensics have been a focus of these developments, and for good reason: forensic tools allow archivists to preserve more data and metadata than would otherwise be available, while the methodology and principles underlying the tools suggest a coherent and repeatable workflow for managing born-digital accessions [1]. Along with the general adoption of the Open Archival Information System reference model, now codified in ISO 14721:2012 [2], born-digital preservation practitioners can more easily share terms, expectations, and processes.

The differing goals between traditional forensic practice in the context of criminal or legal investigations, and that of forensic work in the context of the archive, promise new areas of growth for the field [1][3][4]. For example, disk imaging has become an often recommended, and increasingly practiced, first step in processing digital media [5][6][7]. This process has been improved through projects such as BitCurator, which can help practitioners more easily perform disk imaging, while using best-practice documentation, such as the Digital Forensics XML and the PREMIS Data Dictionary for Preservation Metadata standard during the process.

BitCurator provides Simson Garfinkel's *fiwalk* tool to catalog the contents of a disk image file by file, recording the file system metadata associated with each file and directory. The outputs of *fiwalk* are very valuable, but a process that automatically generates the file system metadata *fiwalk* garners from disk images would be of especial use for large collections. This paper examines that particular challenge: gaining preliminary knowledge of a large collection of floppy disks by aggregating the file system metadata data across those disks. The resulting dataset can indicate the file contents of the entire collection "at a glance," while the aggregated file system level activity may help archivists and researchers identify salient areas to study by highlighting time periods—and the relevant floppy disks in those periods—featuring above average levels of writes, modifications or creation events. Alternatively, the researcher may target particular file system events of interest. More generally, this dataset can suggest how forensic information may or may not be useful to the researcher and archivist in understanding the shape of the collection.

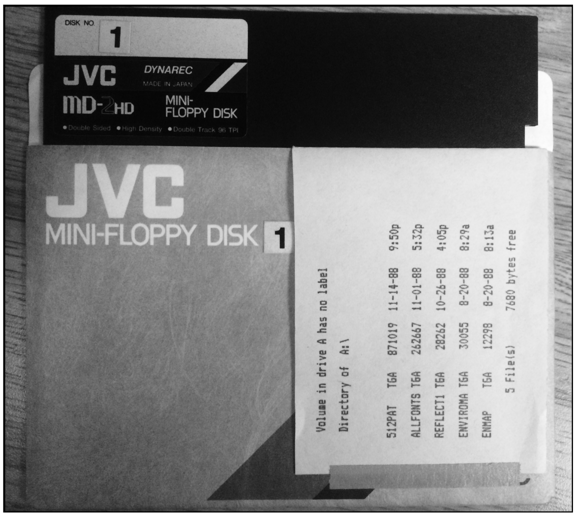
II. Background

The group of floppy disks examined here belong to the Woody and Steina Vasulka art collection. This is an expansive multimedia collection constituting much of the artists' body of work, and is presently in the process of transfer and initial documentation at the University of Colorado Boulder. Woody and Steina Vasulka are experimental media artists, and have been credited as pioneers of video art [8]. Each began their work in Europe, before meeting in Prague and later moving to New York City in 1965. They founded The Kitchen in Manhattan in 1971, a nonprofit art and exhibition space still in operation today. Their work has served as a compelling argument for and demonstration of computer and video technologies in aesthetic and theoretical expression [8]. The artists have been practicing since the 1960s to the present. Their close consideration and use of new technology throughout their career highlights the importance of the born-digital subset of the collection: it is voluminous, containing drafts of components of larger installations or videos, early software used by the artists, along with their correspondence and other writings. Moreover, by dint of the collectible file system events and observable labeling on their disks, the collection can indicate how the artists worked with their tools.

The scope of materials studied are just over a thousand floppy disks in both 5.25" and 3.5" formats, with a handful of 8" floppy disks as well. As with any archive collection containing hundreds or thousands of items, quantity is a challenge in processing. In the case of numerous digital media — in this case floppy disks — that challenge is compounded by the opaque nature of the media. Floppy disk contents are unintelligible without the time-intensive work of generating a bit-level copy of the media and collecting the file content and

W. Sampson, "Aggregating temporal forensic data across archival digital media," *2015 Digital Heritage*, Granada, 2015, pp. 391-394. doi: 10.1109/DigitalHeritage.2015.7419531

metadata there. Even with such work completed, the archivist can be left with an unwieldy collection of either disk images, disk files, or both, with little sense of the overall patterns: file formats, content types, or user activity. What remains for the archivist then is a collection of separate disk images, each containing unique contents and events, with no immediate method of bridging that content into a comprehensive whole.



```
Volume in drive A has no label
Directory of  A:\

512PAT   TGA   871019  11-14-88   9:50p
ALLFONTS TGA   262667  11-01-88   5:32p
REFLECT1 TGA    28262  10-26-88   4:05p
ENVIROMA TGA    30055   8-20-88   8:29a
ENMAP    TGA    12298   8-20-88   8:13a
        5 File(s)      7680 bytes free
```

1. One 5.25” disk featuring a printout of the file contents and dates.

This contrasts sharply to the case of multiple physical folders, documents, photos and other analog media, where the archivist can take a high-level glance at the accession to reasonably speculate on the contents, document types, and the user’s previous handling of those materials. While floppy disk labels can suggest contents, it is problematic to accord too much definitiveness to the labeling, as disk contents could change frequently for users. Fig. 1 depicts one of the more diligently recorded disk labels maintained by the artists, with a file listing and times attached to the sleeve. Such labeling is a slim minority in the collection. Even were it not so, the user would need to be assiduous in updating a disk's labeling, and no automated method of collecting that data would be available in any case.

III. Methodology

The Vasulka floppy disks were imaged with the KryoFlux disk controller, which acquires both a sampling of the floppy disk tracks and a mountable disk image. The resulting disk images were each set in a folder, along with the imaging log and raw data tracks. A collection of command line forensic tools written by Brian Carrier and packaged in *The Sleuth Kit* forensic suite are used to capture file and file system information from these disk images. Execution of these tools is automated through a custom Python script[1] which iterates over a master directory containing the disk images. To clarify the provenance of the final dataset, I will briefly describe the script and how it deploys the tools from *The Sleuth Kit*.

The script's end output is a compiled spreadsheet of all individual disk image timelines into a single timeline that illustrates the file system activity and file contents for the collection. The *fsstat* tool first reads the disk image to discern the file system type. The *fls* tool is then also run against the disk image to produce an intermediary “body file” which contains a timeline of the file system activities. The *mactime* tool is then used against this body file to generate a final, more legible spreadsheet of the file system events.

Both the *fls* and *fsstat* programs cannot read disk images in the Hierarchical File System (HFS) format. In the case that either program is unable to discern the file system or read the disk image, the script instead uses the *unhfs* command line Java utility, packaged with Erik Larsson's *HFSExplorer* program, to extract the file contents from the disk image. The *mac-robber* tool is then run against this folder of exported disk image files to produce a body file. As with the positive reads from the *fls* tool, *mactime* is also run against this body file to produce a final, legible spreadsheet of the timeline. The script's final step concatenates these spreadsheets into a single table. This table represents the aggregated file system events and files of all the parsed disk images in the collection.

A. Dataset

Out of 1,059 disk images, 819 were mounted by the script described above. Floppy disk images might not mount for a variety of reasons which can include bad sectors, unknown file system types, or poor reads; the 240 absent here will need further investigation before being included in the dataset. The 819 disks collected yielded 17,983 potentially valid file system event records. This event count excludes 12,816 file system events with a zero timestamp value, as these indicate the operating system never set the time for the event. It also excludes 7,815 access times equal to the date of the dataset generation. Such times are derived from the *mac-robber* program, which modifies the last accessed times associated with files extracted from HFS disk images through the *unhfs* program (these files are written to disk, and not accessed directly from the floppy disk image).

The majority of the floppy disks are formatted in the original File Allocation Table file system (FAT12), with only 57 featuring Apple’s HFS, and a single disk on the second extended file system (ext2) used with the Linux kernel. A typical entry in this dataset is a row depicting the following fields, from left to right: the timestamp, event type, file name, the name associated with the floppy disk, and the file system type (other fields collected through the script, such as byte size and user permissions, are here redacted).

Sample Dataset Row

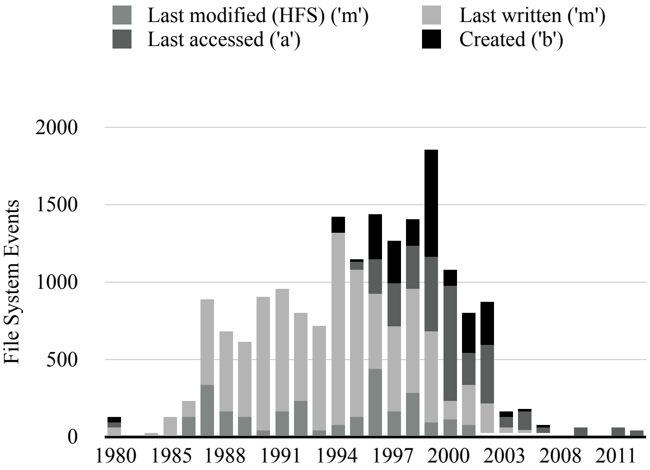
1988-04-18T22:00:00Z	m	EAGLE.TGA	narv_vasulka_truevision	fat12

The timestamp format conforms to ISO 8601 and uses a UTC time zone designation. The event type field bears further explanation. This field may contain any combination of the following notations: *m*, *a*, *c*, and *b*, where *m* indicates a file's last written or modified time, *a* indicates a file's last accessed time, *c* indicates the last time a file's status was changed, and *b* indicates a file's creation time. Nearly all event entries contained a single letter for the event in the set. A minor

[1] W.Sampson, *Disk images timeline* [Online]. Available: https://github.com/wsampson/disk-image-timeline

W. Sampson, "Aggregating temporal forensic data across archival digital media," *2015 Digital Heritage*, Granada, 2015, pp. 391-394. doi: 10.1109/DigitalHeritage.2015.7419531

amount (less than 0.2%) contained both an *m* and *b*; where this is the case the event is noted as a file creation event, as the *m* notation could indicate either "created" or a "last written" event. Variances in event meaning will be discussed further in the *Analysis* section.



2. Chart illustrating all valid file system events for 819 floppy disks.

All data points are illustrated in a stacked column chart in Fig. 2. The "Last modified (HFS)" category applies to *m* events in the HFS system (HFS does not support a last accessed time and there were no creation events for HFS disks). The "Last written" category applies to all FAT12 events marked *m*; the "Last accessed" applies to FAT12 events marked *a*. The "Created" category applies to FAT12 events containing the *b* notation. The single ext2 disk contains too few events to merit inclusion in the illustration.

B. *Analysis*

A few broad observations can be made from the chart. We see this collection spans activity from 1980 to 2012, and we see disk activity for the collection appears to have increased from the early 80s to the late 90s. We observe the majority (55%) of file system events are the FAT12 file system's last written event type, which indicates the last time the contents of a file were modified, or the time a file was created. The ambiguity of this value is indicative of a larger consideration for the dates associated with the file system events. A brief discussion of the file system in question, FAT12, will help to better understand this aspect.

The FAT12 file system can support three times associated with a file: last written, last accessed and created [9]. While quite useful, the timestamps associated with these events are not strictly trustable. This is caused by a few variables. The first is that different operating systems using FAT12, such as MS-DOS, Windows 3.1, Windows XP or NT, etc., can and do update these times under different policies [10]. In fact, only the last written date and time is required across all FAT systems; MS-DOS and Windows 3.1 therefore use the last written times, while Windows XP and NT will use all three times. Along with this, an event as seemingly straightforward as "created" can be interpreted differently by the operating system than as an observer may expect. An example put forward in [10] is a Windows system wherein a copied file results in a created time equal to the time of copy, rather than the creation time of the original file being copied — this can result in a created date older than the last written date. In short, FAT12 cannot strictly specify how the metadata fields it provides are used.

In addition, the file system cannot check the validity of a date and time [10]. A simple example of this is an out-of-sync system clock which then yields erroneous times for the file system. In this dataset for example, four file system events are placed in or after the year 2018. Finally, we should note the differing granularity for each time value: the created timestamp is accurate to a tenth of a second, the accessed timestamp to the day, and written timestamp to 2 seconds. As noted above, all times are derived from the local system clock of the computer performing the file activity.

While the details of how an operating system uses the time values provided by a file system are critical for legal investigations, this is frequently less so in the archival context. It is important to understand the ambiguity of their values, but here the timestamps are collected to provide an overview of activity and to indicate areas of interest. To return to the case beginning this discussion, we can note that a FAT12 "last written" event indicates either an initial write of a file (i.e., when it was created), or the last time the file's content was modified. The peak year of 1999, which also contains the most "created" events, yields an abundance of user-created files: over 750 documents (with over 140 deleted which can indicate previous drafts), and 393 image files, many depicting early computer generated imagery and art. The dataset allows a user to move directly to the disks containing these files, and to observe the surrounding activity, files and floppy disks.

The qualification of "user-created" is salient: a useful distinction to make in any given year depicted in Fig. 2 is what events are user generated and which are generated by other agents, e.g., software in the case of files routinely written to the disk during an installation or other process, or the initial writes of a disk vendor or manufacturer. Such dates are not without value: timestamps derived from the initial write by a vendor date the disk, and files which are written, modified or accessed by software are often initiated by the user (this is especially the case for file system events on floppy disks), and indicate user activity by proxy.

Regardless, events and files directly linked to a user are generally of first interest to both researchers and archivists. Removing events which share a timestamp is a simple way to attempt retraction of events caused by software processes or the vendor, since it is likely that events which share identical timestamps are generated by software. With this retraction, the dataset is narrowed to 9,623 file system events. While the files in this set contain non-user generated files, such as device drivers, executables, and configuration files, most of the files appear to be created by the artists or other persons: Truevision TGA computer graphics files, JPEGs and TIFFs, text documents, 3D model files, and so on.

The event types broadly mirror the distribution in complete dataset, but do contain a higher percentage of created events, from 13% to 20%, with significantly less 'last accessed' events. These differences are less enlightening when we take into account the varying timestamp granularities in FAT12: "last accessed" accurate to the day, and "created" times to the tenth

W. Sampson, "Aggregating temporal forensic data across archival digital media," *2015 Digital Heritage*, Granada, 2015, pp. 391-394. doi: 10.1109/DigitalHeritage.2015.7419531

of a second. Only 371 created events are removed in the retracted set; the majority of these events are associated with temporary or deleted files. Within the FAT12 system then, redaction of equivalent timestamps is not effective; similar results could be achieved by filtering the created events.

C. *Deleted files*

The deleted files collected in the dataset indicate another path to pertinent material to the researcher and archivist. The *fls* command deployed by the script and used with the FAT12 file system can detect deleted files; when we observe the subset of events associated with these deleted files which still have a nonzero byte count (indicating that at least some if not all of the data is still present), we obtain a relatively narrow selection of 2,350. The *mac-robber* tool used by the Python script to capture file event metadata from HFS disks does not capture deleted files—deleted files from those disks will not be present in this set if such files exist. With this caveat in mind, many of the files associated with these events are very likely user generated files (this observation is based off both the file name and format), while others are generated through software (such as files with .TMP extension). Events associated with temporary files may indicate drafts or snapshots of documents as well. Even in the case of deleted files which are clearly transitory data used by software, these events strongly signal user intention and are of some value to the researcher.

IV. Conclusions

I have attempted to illustrate the key complications and promises of aggregated temporal forensic data in a floppy disk collection. The particular values of any given timestamp are subject to some variance in meaning, accuracy and simple presence, particularly within the FAT12 file system dominant here. More modern file systems, such as the New Technology File System and the third extended file system, may provide clearer meanings for the events dates and times. Differences in date and time support between file systems also complicates the uniformity of the dataset (e.g., HFS does not support last accessed timestamps, therefore a dearth of those timestamp values on HFS disks should not be considered a conspicuous absence). Lack of immediate tool support for less common or proprietary file systems can exacerbate this problem.

However, the aggregate timeline is useful in providing a wider context to the store of data and metadata carried in a large collection of floppy disks. As archivists encounter more collections with less defined or discernible preexisting arrangements to guide patrons [11], the timeline can offer a coherent entry point. Beyond this aim, such a timeline provides basic information on the collection, such as the date range for user activity, indications of high activity dates, and simple tallies of deleted files. Collections of other removable media, such as optical discs, flash drives, Secure Digital (SD) cards, external drives, or any combination thereof could benefit from an aggregated timeline. There is promise that such timelines will allow researchers and archivists to approach the collection as a unique body of user activity and content, particularly if a timeline were included in a collection description. Moreover, the primary component of an aggregated timeline — disk images — are already produced as part of many born-digital workflows. As such, collected timelines provide additional value to the time-intensive disk imaging process.

Future steps may include tools to discern the operating system used with removable media, which could give a finer indication of the meaning of the various timestamp values, along with further informing a researcher of the user's context. The script used here can be developed to include more sophisticated logging of the component tool outputs to accompany the final timeline, and the timeline itself could be reworked to emphasize filenames, byte sizes, or user permissions which are collected through *The Sleuth Kit* utilities. In all cases, research approaches using timelines are not intended to supersede close study of archive items, but to complement more directed evaluation of the material.

Acknowledgment

The author thanks Prof. B. Montgomery for guiding the ongoing accession of the Vasulkas' material, and to W. Sutherland-Keller, without whose assiduous work in disk imaging this work would not have been possible.

References

[1] M. Kirschenbaum, R. Ovenden, and G. Redwine, "Digital forensics and born-digital content in cultural heritage collections," Council on Library and Information Resources, Washington, DC, pp. 60-61, December, 2010.

[2] *Space data and information transfer systems—Open archival information system (OAIS)— Reference model*, ISO 14721:2012, 2012.

[3] C. Lee, K. Woods, M. Kirschenbaum, and A. Chassanoff, "From bitstreams to heritage: Putting digital forensics into practice in collecting institutions," School of Information and Library Science at the Univ. North Carolina, Chapel Hill, p. 24, November, 2013.

[4] C. Lee and K. Woods, "Automated redaction of private and personal data in collections," *Proc. Memory of the World in the Digital Age: Digitization and Preservation International Conf.,* Vancouver, British Columbia, pp. 298-313, p. 304, September, 2012.

[5] R. Erway, "You've Got to Walk Before You Can Run: First Steps for Managing Born-Digital Content Received on Physical Media," OCLC Research, Dublin, Ohio, p. 5, August, 2012.

[6] R. Fox, "Forensics of digital librarianship," *OCLC Systems & Services: International digital library perspectives*, vol. 27, no. 4, pp. 264 - 271, p. 270, 2011.

[7] J. Durno and J. Trofimchuk, "Digital forensics on a shoestring: a case study from the University of Victoria," *Code4Lib*, no. 27, January, 2015.

[8] V. Bonin. (2003). *Steina and Woody Vasulka fonds* [Online] Available: http://www.fondation-langlois.org/html/e/page.php?NumPage=422

[9] Microsoft, "FAT: General overview of on-disk format," *Microsoft Extensible Firmware Initiative FAT32 File System Specification*, p. 24, December, 2000.

[10] B. Carrier, "FAT concepts and analysis," in *File System Forensic Analysis*, Upper Saddle River, NJ: Addison-Wesley, 2005, pp. 228-236.

[11] J. Bailey, "Disrespect des fonds: Rethinking arrangement and description in born-digital archives," *Archive Journal*, no. 3, 2013.

W. Sampson, "Aggregating temporal forensic data across archival digital media," *2015 Digital Heritage*, Granada, 2015, pp. 391-394. doi: 10.1109/DigitalHeritage.2015.7419531