Exploring Archival Collections through Forensically Packaged Disk Images

Productive Scholar Series
New Media Center, 130 Lewis Library
October 23, 2014
Presenter: Jarrett M. Drake
Digital Archivist, Seeley G. Mudd Manuscript Library
Overview

Part 1  Identify core issues related to born-digital archival collections

Part 2  Identify impact of born-digital archival collections on researchers
Responsibilities

University Archives
- Describe collections
- Enhance workflows
- Reference!

University Library
- Chair Born-Digital Group
- Contribute to many others!

Society of American Archivists
- Access to Electronic Records Working Group

Digital archivist
University Archives

- Established in 1959 by the Board of Trustees
- Holds over 400 distinct archival collections documenting the history of the University
- Maintains > 100GB of 30,000 digital files in more than 20 collections
Analog Archival Collections

- Information recorded and used on human-readable media
- Description turns chaos → order
- “More Product, Less Process” (MPLP) approach helps reduce backlogs
Born-Digital Archival Collections

- Information created and used on machine-readable digital media

- New data points to consider

- If context lost, chaos → order → chaos

- “Three moves equals a fire”
Issues of Born-Digital Collections

Acquire

Traditional practice

Control

Preserve

Protect
Acquisition

- Requires earlier contact with records’ creator
- Demands archivists be equipped to handle technical oddities

Hardware
- Hard Drive
- Optical Media
- USB

Operating System
- Windows
- OS X
- Linux
Preservation

- Molded papers stink and erode slowly, but computer viruses don’t and corrupt quickly.
- All files are equal, but some files are more equal than others.

<table>
<thead>
<tr>
<th>Image</th>
<th>Tagged Image File</th>
<th>Value</th>
<th>Instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG2000</td>
<td>.tif, .tiff</td>
<td>1</td>
<td>Save uncompressed</td>
</tr>
<tr>
<td>PNG</td>
<td>.png</td>
<td>1</td>
<td>Use lossless data compression</td>
</tr>
<tr>
<td>Scalable Vector Graphics</td>
<td>.svg</td>
<td>1</td>
<td>Use for vector graphics; ensure no JavaScript binding</td>
</tr>
<tr>
<td>Drawing eXchange</td>
<td>.dxf</td>
<td>2</td>
<td>Convert to .svg where possible</td>
</tr>
<tr>
<td>GIF</td>
<td>.gif</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>BMP</td>
<td>.bmp</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Photoshop</td>
<td>.psd</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

"Privacy is at the heart of the ethical issues surrounding born-digital materials." – 2010, Council on Library and Information Resources

http://www.clir.org/pubs/reports/reports/pub149/pub149.pdf
Part 2: Impact on Researchers

- Born-digital archives have forced archivists to make substantive changes in the way we curate materials.
- These changes led archivists to look outside the archival profession for solutions.
- One approach has been to apply digital forensics procedures to cultural heritage materials.
“Digital forensics...is concerned with discovering, authenticating, and analyzing data in digital formats to the standard of admissibility in a legal setting.”*

Why Digital Forensics?

- Archivists can better demonstrate chain of custody, or **provenance**
- Archivists can ensure that their actions do not corrupt a file’s embedded metadata
- Archivists can become more transparent in the actions they take on collections

Researchers can ask new and interesting questions that otherwise would not avail themselves to research!
Forensic Disk Image

“...sector-by-sector copies of all the data that reside on a storage medium.”*

Accessing Disk Images

- Permission from the archival repository
- Operating system that allows for safe mounting of disk images

www.bitcurator.net
Using Disk Images

- Scan image
- Extract entities
- Reveal deleted files
- Build topic models
### Extracted Entities

List of features that can be scanned using Bulk Extractor

**Credit card numbers**
- ccn.txt
- ccn_histogram.txt
- domain.txt
- domain_histogram.txt
- email.txt
- email_domain_histogram.txt
- email_histogram.txt
- exif.txt
- hex.txt
- jpeg_carved.txt
- json.txt

**Email headers**
- rfc822.txt
- telephone.txt
- telephone_histogram.txt
- url.txt
- url_histogram.txt
- url_searches.txt
- url_services.txt
- windirs.txt
- wordlist.txt
- zip.txt
Enron email dataset can be found at http://www.cs.cmu.edu/~enron/
### Extracted Entities: Domains

<table>
<thead>
<tr>
<th>Domain Name</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>enron.com</td>
<td>1,265,648</td>
</tr>
<tr>
<td>aol.com</td>
<td>201,330</td>
</tr>
<tr>
<td>ENRON.com</td>
<td>149,910</td>
</tr>
<tr>
<td>hotmail.com</td>
<td>145,500</td>
</tr>
<tr>
<td>yahoo.com</td>
<td>77,250</td>
</tr>
<tr>
<td>bpa.gov</td>
<td>55,180</td>
</tr>
<tr>
<td>football299.fantasy.sportsline.com</td>
<td>52,740</td>
</tr>
<tr>
<td>nyiso.com</td>
<td>48,470</td>
</tr>
<tr>
<td><a href="http://www.carrfut.com">www.carrfut.com</a></td>
<td>45,320</td>
</tr>
<tr>
<td>haas.berkeley.edu</td>
<td>40,030</td>
</tr>
<tr>
<td><a href="http://www.nytimes.com">www.nytimes.com</a></td>
<td>38,420</td>
</tr>
<tr>
<td>pgn.com</td>
<td>37,920</td>
</tr>
<tr>
<td>football222.fantasy.sportsline.com</td>
<td>37,380</td>
</tr>
<tr>
<td>bracepatt.com</td>
<td>37,150</td>
</tr>
<tr>
<td>earthlink.net</td>
<td>33,730</td>
</tr>
<tr>
<td>ect.enron.com</td>
<td>32,320</td>
</tr>
<tr>
<td>carrfut.com</td>
<td>31,960</td>
</tr>
<tr>
<td>pira.com</td>
<td>29,170</td>
</tr>
<tr>
<td>houston.rr.com</td>
<td>28,270</td>
</tr>
<tr>
<td>williams.com</td>
<td>27,610</td>
</tr>
<tr>
<td>msn.com</td>
<td>26,500</td>
</tr>
<tr>
<td>schemas.openxmlformats.org</td>
<td>25,020</td>
</tr>
</tbody>
</table>
Extracted Entities: Emails

Histogram File: email_histogram.txt
- brudy@umich.edu (n=18)
- chonawee@umich.edu (n=14)
- enronresearch@umich.edu (n=8)
- drakejm@umich.edu (n=5)
- fwb@umich.edu (n=4)
- imceane@22bruce+20c+20rudy+22+20+3brudy@umich.edu (n=4)
- leekj@umich.edu (n=4)
- annec@umich.edu (n=3)
- haleyw@umich.edu (n=3)
- jaminkoo@umich.edu (n=3)
- jtsumi@umich.edu (n=3)
- pionke@umich.edu (n=3)
- enronresearch-errors@umich.edu (n=2)
- enronresearch-members@umich.edu (n=2)
- keppo@engin.umich.edu (n=2)
- skildau@umich.edu (n=2)
- drakejm@umich.edu (n=2)
- 4939523-POD-1196 (n=1)
- 1286418342-ZIP-56251 (n=1)
- 1286552991-ZIP-2888 (n=1)
- 2315078299-ZIP-2832 (n=1)
- 2315079676-ZIP-1874 (n=1)

Referenced Feature File: email.txt
- drakejm@umich.edu

Histogram File: email_domain_histogram.txt
- princeton
- umich
- yale
On the Horizon

- Archivists advocating for the use of *all* their collections
- Archivists soliciting formal feedback from users of born-digital collections
- BitCurator Access: more at access.bitcurator.net