Library Technology Conference 2014

Macalester College, St. Paul, MN

Using Freely Available Tools to Protect Your Digital Content

Carol Kussmann, University of Minnesota Libraries

Sara Ring, Minitex
What Brought You to this Session?
Taking the Temperature

What worries you the most about protecting your digital content?
Introduction

Small Steps Now = Healthy Digital Files in the Future
Checksum Tools
What is a Checksum?

How to verify that a file has not changed?
- Generate a checksum!

Checksum = Mathematical algorithm run on a file

Resulting number is usually referred to as a checksum value
- Fingerprint
- Hash value
- Cryptographic hash function
- Checksum
# Checking Your Digital Content: How, What and When to Check Fixity

[http://z.umn.edu/k6q](http://z.umn.edu/k6q)

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Level of Effort and Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRC</td>
<td>Typically checking network error</td>
<td>Low level of effort and moderate level of detail. CRC function values, which are variable but typically 32 or 64 bit, are relatively easy to implement and analyze.</td>
</tr>
<tr>
<td>MD5</td>
<td>Cryptographic hash function</td>
<td>Moderate level of effort and high level of detail. CPU and processing requirements to compute the hash values are low to moderate depending on the size of the file. The output size of the hash value is the lowest of the cryptographic hash values at 128 bits.</td>
</tr>
<tr>
<td>SHA1</td>
<td>Cryptographic hash function</td>
<td>Moderate level of effort, high level of detail and added security assurance. Due to its higher 160-bit output hash value, SHA-1 requires more relative time to compute for a given number of processing cycles CPU and processing time than MD5.</td>
</tr>
<tr>
<td>SHA256</td>
<td>More secure cryptographic hash function</td>
<td>High level of effort and very high level of detail, and added security assurance. With an output hash value of 256 bits, SHA-256 requires more relative time to compute for a given number of processing cycles CPU and processing time than SHA-1.</td>
</tr>
</tbody>
</table>
- Calculate MD5 checksums
- Save as CSV file
- Checksum on a Single File
- Checksums on Group of Files
- Checksums to put with files burned to CD/DVD
- Multiple checksum algorithms available
- Verify Checksums
- Calculate a variety of checksums
- Save values as CSV/Tab-Delimited File
Karen’s Directory Printer

- Calculate MD5, and SHA-checksums
- Save as .txt file
File Duplication Detection Tools
Find duplicate files
Delete duplicate files

<table>
<thead>
<tr>
<th>File Name</th>
<th>Ext</th>
<th>Size</th>
<th>Modified Date</th>
<th>In Folder</th>
</tr>
</thead>
<tbody>
<tr>
<td>home_tree</td>
<td></td>
<td>634.34 KB</td>
<td>12/5/2013 10:55 AM</td>
<td>C:</td>
</tr>
<tr>
<td>home_tree</td>
<td></td>
<td>634.34 KB</td>
<td>12/5/2013 10:55 AM</td>
<td>C:</td>
</tr>
<tr>
<td>Manage_DPOE_CRK</td>
<td>pptx</td>
<td>32.05 KB</td>
<td>12/2/2013 12:49 PM</td>
<td>C:</td>
</tr>
<tr>
<td>PowerPoint</td>
<td>pptx</td>
<td>32.05 KB</td>
<td>12/2/2013 12:49 PM</td>
<td>C:</td>
</tr>
</tbody>
</table>

Total Duplicates: 2 | Wasted Disk Space: 666.40 KB
Files Scan: 11 | Time Taken: 00:00:00

2 files marked for deletion, total size 666.40 KB
Hash My Files

- Find duplicate files
- Save reports as CSV/Tab-Delimited file

<table>
<thead>
<tr>
<th>Filename</th>
<th>MD5</th>
<th>Extension</th>
<th>File Size</th>
<th>File Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>authentication.docx</td>
<td>a34424592e896eb78de333c343605c7</td>
<td>docx</td>
<td>16,566</td>
<td></td>
</tr>
<tr>
<td>checksum tests.md5</td>
<td>e12783f8564468b35e7c7004c9dc0064</td>
<td>md5</td>
<td>368</td>
<td></td>
</tr>
<tr>
<td>checksums.md5</td>
<td>d8c7425ac1b253202794ea9648f7f76f</td>
<td>md5</td>
<td>532</td>
<td></td>
</tr>
<tr>
<td>DPCHandbook.pdf</td>
<td>9ba58ef567d23c6f63c5b0a268db72</td>
<td>pdf</td>
<td>1,067,906</td>
<td></td>
</tr>
<tr>
<td>home_tree</td>
<td>999b6d52d3cebbf9a153ccbd13580b32</td>
<td>png</td>
<td>649,567</td>
<td>1</td>
</tr>
<tr>
<td>home_tree.png</td>
<td>600b6d52d2eabf0b2a152ccbd13580b32</td>
<td>png</td>
<td>640,567</td>
<td>1</td>
</tr>
<tr>
<td>home_tree.png.bak</td>
<td>164f3e97e43c8104ce5f13626b41468f</td>
<td>bsk</td>
<td>649,601</td>
<td></td>
</tr>
<tr>
<td>Manage_DPOE_CRK.pptx</td>
<td>37f4c9d03aab3f2063e9275706c4</td>
<td>pptx</td>
<td>32,823</td>
<td>2</td>
</tr>
<tr>
<td>OAS Blue.docx</td>
<td>ffe8c923461f2177202b850ddf7df51</td>
<td>docx</td>
<td>2,079,281</td>
<td></td>
</tr>
<tr>
<td>PowerPoint.pptx</td>
<td>37f4c9d03aab3f2063e9275706c4</td>
<td>pptx</td>
<td>32,823</td>
<td>2</td>
</tr>
<tr>
<td>Thumbs.db</td>
<td>4d33af964538e2494f83dead6bd1bccc</td>
<td>db</td>
<td>84,480</td>
<td></td>
</tr>
</tbody>
</table>
File Format
Identification Tools
- File property capture
- Format identification
- Generates specific reports (PDF)
- Save results as CSV
Karen’s Directory Printer

- File property capture
- Save reports as .txt file
Metadata Quality Control Tools
Metadata Quality Control

- Compares **expected** technical metadata to **actual** technical metadata
- Produce CSV pass/fail report
Demonstrations -
Now Your Turn
Demonstrations

- DROID
- Karen’s Directory Printer
- Duplicate File Finder
- Hash My Files
- Exact File
- Metadata Quality Control (MDQC)
Thanks!!

Carol Kussmann; kussmann@umn.edu
Digital Preservation Analyst, University of Minnesota Libraries

Sara Ring; ring0089@umn.edu
Digitization & Metadata Training Coordinator, Minitex