Sipping from a Bag: Ingesting Disk Images with BagIt

Rebecca Sutton Koeser, Emory University
Elizabeth Roke, Emory University
Dorothy Waugh, Emory University

Conference Name: Open Repositories 2016
Publication Date: 2016-06-14
Type of Work: Poster
Permanent URL: https://pid.emory.edu/ark:/25593/rpff8

Final published version: http://rlskoeser.github.io/2016/06/08/diskimage-ingest/

Copyright information:
© 2016 Rebecca Sutton Koeser
This is an Open Access work distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Accessed January 14, 2018 9:27 PM EST
WHY CREATE DISK IMAGES?

Born-digital material is archival material, not a surrogate for archival material (in the way that a digitized copy of a handwritten letter would be, for example). As such, the archival principles of provenance, authenticity, and context apply to born-digital materials.

Borrowing from techniques used in digital forensics, many archivists choose to capture forensic disk images of born-digital material as a way of adhering to these three important principles. This creates an exact replica of the digital artifact at the bitstream level, thereby ensuring no inadvertent change or loss of data during transfer. The creation of checksums can then be used to verify that the capture of a disk image has been successful, which helps to authenticate the data. Furthermore, disk imaging preserves filesystem metadata and, in cases where software is stored on the disk in question, associated representation information that can help identify and render files. Taken together, these benefits help archivists establish and document the provenance, authenticity, and context of born-digital archival material.

WHY USE BAGIT?

1. Acts as a protective wrapper during transfer.
2. Validation of the entire package.
3. Maintains provenance of the SIP (object + metadata that tracks chain of custody)
4. Checksum manifest, which supports multiple algorithms (e.g. both MDS and SHA-1).

THE LANDSCAPE AT EMORY

436 disk images, 1.2 TB total
(includes migrated content)
Average 2.8 GB, largest 298GB
E01, ISO, IMG, DD, TAR*, AFF**, AD1**

5% of all Rose Library manuscript collections contain some form of born digital content.

Acquired Pre-2000 1%
Acquired 2000-2010 10%
Acquired 2010-2016 13%

*WHY CREATE TAR FILES?
The capture of forensic disk images at Emory is not always possible. In conversations with donors, Rose Library archivists aim to be transparent about transfer methods and the fact that forensic disk imaging captures a complete replica of data can cause some anxiety for donors. In these instances, we capture specific files and directories using the TAR utility.

OTHER BENEFITS OF DISK IMAGES

• Packages born-digital data, so that it can be safely moved without the risk of alteration.
• Removes the need to rely on aging and often obsolete media.
• Provides the means to create a complete copy of the data, for preservation purposes, without requiring that I interpret the data—this can then be done at my leisure as and when I have the resources and tools.
• Provides greater ease of emulation as an access point.

Statistical analysis and plot by Rob O’Reilly, corell@emory.edu

http://goo.gl/visMhD