Metadata and the Born-digital Imaging Workflow

Digital Preservation Unit

Department of Preservation and Conservation, University of Michigan Library

Created by
Kayla Carucci, Noa Kasman

Last Revised:
April 18th, 2017

Overview

Introduction

This report describes the use of metadata in the born-digital workflows used at the Department of Preservation and Conservation at the University of Michigan. We have explored disk imaging in BitCurator, folder packaging in Bagger, and logical transfers in Data Accessioner.

We are currently imaging external hard drives and plan to image or logically transfer content on optical discs, 3.5 inch floppy disks, and USB thumb drives.

This report will predominantly discuss manually entered and automatically generating metadata created by tools in the BitCurator environment, and the creation and use of a custom Bagger profile to store metadata in comparison to the PREMIS preservation metadata standard.

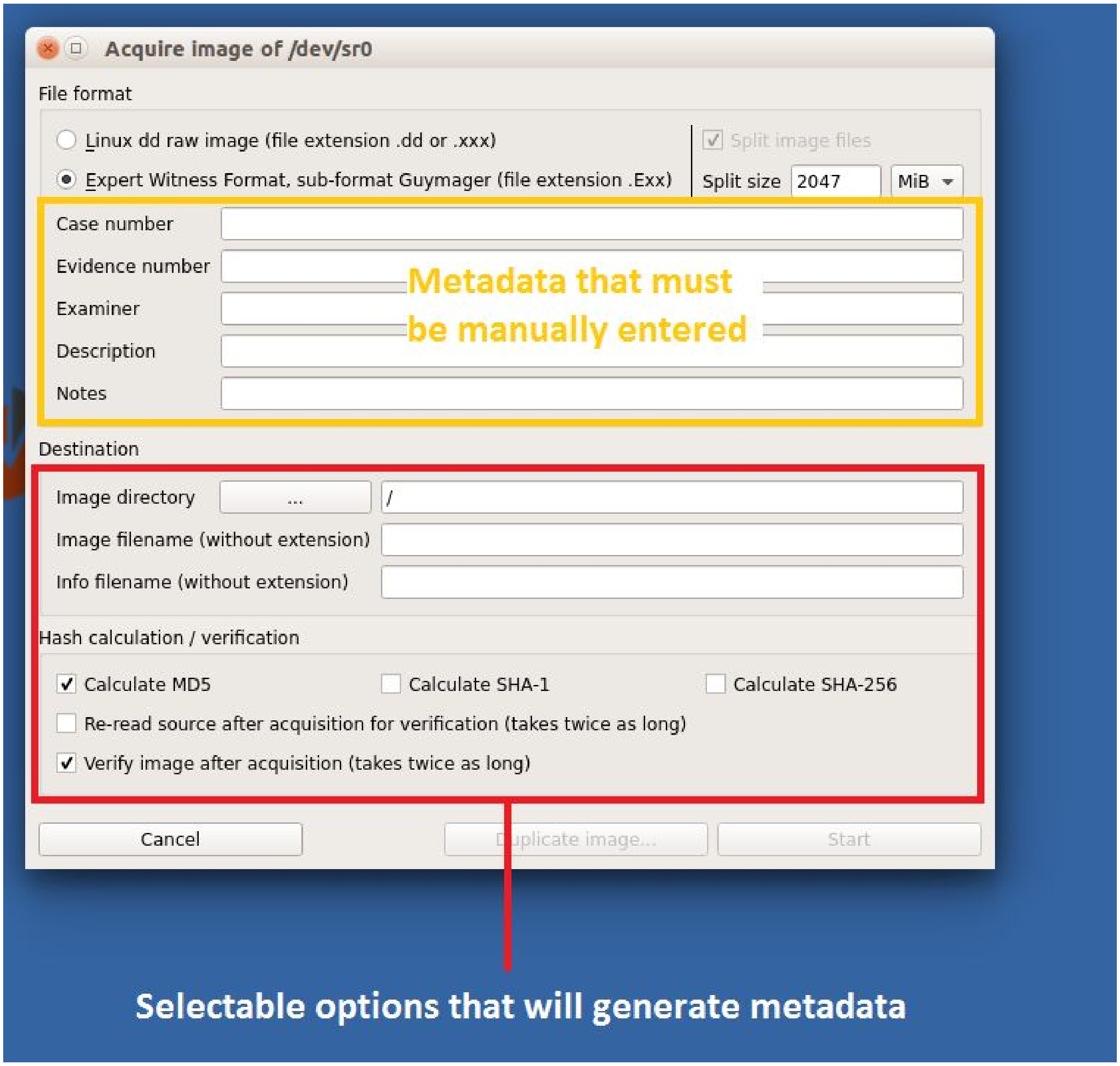
Disk Imaging Metadata

Disk Image Components

When Guymager is used to create a disk image, two files are generated:

- the first is the image file
- the second file is a .INFO file.

These files capture metadata at the time of disk image creation. Some of this metadata must be entered manually, while other data is captured automatically.


Acquire image of /dev/sr0

File format

○ Linux dd raw image (file extension .dd or .xxx)

● Expert Witness Format, sub-format Guymager (file extension .Exx)

☑ Split image files

Split size 2047 MiB

Case number

Evidence number

Examiner

Description

Notes

Destination

Image directory ... /

Image filename (without extension)

Info filename (without extension)

Hash calculation / verification

☑ Calculate MD5 □ Calculate SHA-1 □ Calculate SHA-256

□ Re-read source after acquisition for verification (takes twice as long)

☑ Verify image after acquisition (takes twice as long)

Metadata that must be manually entered

Selectable options that will generate metadata

Selectable options that will generate metadata

- FIle format - Linux dd raw image format (.dd or .xxx) or Expert Witness format (.Exx)
- Split size - The size limit for discrete disk image packages. If you would like to create single disk image packages for data that exceeds the default split size 2047 MiB you will need to change the split size.
- Image Directory - where the media to be imaged is located
- Image filename without extension and Info filename without extension - these file names are the same, and based on the evidence number
- Hash calculation / verification - Multiple options may be selected for this field. We use:
  - Calculate MD5
  - Calculate SHA-256
  - Verify image after acquisition (takes twice as long)

We do not use the following selections:

- ➢ Calculate SHA-1
- ➢ Re-read source after acquisition for verification (takes twice as long).

Metadata that must be manually entered

Field	**Data**
Case number	collection ID (e.g. “Robert Altman”)
Evidence number	Name assigned to source media/drive (e.g. PHC musA) This information is captured in the Data Accessioner workflow.
Examiner	First name and last name of the person who is creating the disk image
Description	Media format/type, this information should be drawn from the controlled list in the Bagger profile’s Medium/Carrier list.
Notes	Free text field. This field should capture important information on the media carrier including information on the carrier, carrier’s label, and carrier’s housing. Scan any accompanying information if necessary. When you transcribe this information, differentiate between where the information is coming from. e.g. Notes: Transcribed information from the HD box: “ARCHIVES UNBOUND”, Transcribed information from the label on the HD: "Archives Unbound The American Indian Movement and Native American Radicalism", "256479A", "GALE CENGAGE Learning". Note: The "A" in "256479A" is crossed out or filed in.) The information captured in this field should be copied into the “ItemInformationNotes” field in Bagger.

Image filename (without extension)	barcode. This field automatically populates the “Info filename” field below it.
Info filename	barcode (automatically populated)

Metadata that is automatically generated

- Device size in kbs
- Number of bad sectors, if identified
- State of disk image
    - Finished successfully
    - Finished successfully (with x bad sectors)
    - Verification status
        - Source verification FAILED. The device didn't deliver the same data during acquisition and verification. Check if the defect sector list was the same during acquisition and verification (see above). Maybe you try to acquire the device again.
        - Image verification OK. The image contains exactly the data that was written.
    - Device disconnected, acquisition paused
- Acquisition start time
- Verification start time
- Ended
- Acquisition speed
- Verification speed

.info and PREMIS files

Similarly to PREMIS, the .info file provides relevant information for the long term preservation of digital resources. However, the .info file only covers what PREMIS would consider to be one event entity, the creation of a disk image. In “Understanding PREMIS”, Caplan states:

> Each repository system must make its own decisions about which events to record as a permanent part of an object's history. PREMIS recommends that actions that change an object should always be recorded, and the Data Dictionary entry for eventType provides a "starter list" of important event types to encourage repositories to record these events consistently.[1]

Since the disk image is a creation point, after which many changes occur, the .info file is not the appropriate location to record these changes. Therefore, we have create a custom Bagger profile based on the Digital-Records-Accession-Generic-profile in Bagger 2.7.0.

Bagger Metadata and Custom Profile

The custom profile titled ‘UMICH-Disk-Imaging-profile’ was created specifically for the needs of our born-digital imaging workflow. After the .info file is created, metadata pertaining to changes to the born-digital object can be recorded here. This information will be packaged as a SIP for the department that owns the original medium.

Bagger profiles are json files, where the filename ends in -profile (ie. NewProfileforBagger**-profile.json**). Therefore, the current profile can be edited in a text editor, or a new one can be created for a specific purpose. Adding a new profile is as easy as copying the file to the bagger folder, where the application was originally installed.

Metadata in Bagger

Field	**Data**	**Rationale**
Profile Name	Auto-generated, pulls the name of selected metadata profile e.g. UMICH-Disk-Imaging-profile	
BarcodeNumber/Identifier	barcode	Serves as a unique identifier that

[1] Priscilla Caplan, “Understanding PREMIS,” *The Library of Congress*, February 1, 2009, https://www.loc.gov/standards/premis/understanding-premis.pdf, 10.

		helps to differentiate one digital object from another; this number is used throughout multiple directories and reports to reference one born-digital object.
AccessionNumber/Collection	Collection ID = “Robert Altman”	This should contain the same information as the “Case Number” field in Guymager.
OriginatingUnit/Department	“Special Collections”	
MediaRetentionPriority	Drop down list. Establish the importance of retaining the media carrier.	MediaRetentiionPriority will aid in determining whether the original medium should be kept after imaging. For our purposes, original media carriers will be kept. Thus, this field should

		list the MediaRetentionPriority as “HIGH.”
recordsTitle	collection, e.g. “Robert Altman Archive”	
Medium/Carrier	Drop down list. Describes how the data was stored prior to imaging, for example on a 3.5 inch floppy disk. This field should be the same as the Guymager field “Description”.	
Medium/CarrierConditionNotes	Free text field. Information specific to the condition of the carrier, such as damage or missing features (e.g. Metal slider missing, Arrived wrapped in aluminum foil (no reason apparent))	
Medium/CarrierDetails	Free text field. Information about the medium’s proprietary format e.g. Double Density Floppy, Macintosh formatted (3M). For example: Transcribed information from the floppy disk: “SONY FORMATTED for IBM PS/2 & compatibles 1.44MB HIGH DENSITY MFD-2HD”.	
Medium/Carrier #2	SEE Medium/Carrier	
Medium/CarrierConditionNotes #2	SEE Medium/CarrierConditionNotes	
Medium/CarrierDetails #2	SEE Medium/CarrierDetails	
Hardware	Drop down list. Hardware utilized in the disk imaging or logical transfer processes.	
Software	Drop down list. Computer program used to create the disk image or to perform the logical transfer.	
WriteBlockerHardware	Drop down list. Physical method taken to ensure the original data remained unchanged.	Two write blocking methods may seem

WriteBlockerSoftware	Drop down list. Write blocking software that might have been active on the computer during the disk imaging or logical transfer processes.	redundant, but in case of failure it is better to have a backup method. This record helps to promote trust from researchers and other users.
VirusCheckRun?	Drop down list. Indicates whether a virus scan was run on the disk image using the virus scanning software listed in the VirusCheckSoftware? field.	While ingesting content, digital files can carry viruses or malware into an archive. VirusCheckRun? provides critical information for archivists and other users by letting them know whether this content was virus free upon creation.
VirusCheckSoftware?	Drop down list. Software used to conduct the scan. Add name of a different virus checking software if the ClamTk was not used	

VirusCheckResults	Drop down list. The outcome of the virus scan.	
VirusCheckResultsNotes	Free text field. Information about the virus scan results. Error messages should be included if there are any.	If a file is detected as containing a virus it can be stored separately from the archive in the quarantine, thereby creating no risk of infection; this allows the computer to re-scan for viruses, and attempt to repair the file. Quarantines also allow for virus definitions to catch up to any newer viruses that might be lurking in the file. Talk to Lance about how to handle data that does not pass a virus scan.

TestMountPerformed?	Drop down list. Describes whether the creator of the original image attempted to mount the disk image.	This information can be important, because a user would not want to mount a disk image that contained malware. Therefore, this becomes an additional reminder.
TestMountResults	Drop down list. Describes whether the test mount was successful.	A disk image that does not mount may indicate an error during the imaging process.
BagContent	Drop down list. Describes born-digital data migration process and format.	BagContent and OAISDigitalCurationLifecycle will remain static throughout the born-digital workflow. OAISDigitalCurationLifecycle should remain SIP (submission information package)
OAISDigitalCurationLifecycle	Drop down list. OAIS package type.	

		because the data will be ingested back into a collection with the original medium/carrier. However, if the medium/carrier was destroyed, the disk image may become an AIP (archival information package) or DIP (dissemination information package).
Notes	Free text field. Anything unusual throughout the imaging process, important information the image creator might want to pass on. The most common note we have used so far is: "It is impossible to run all BitCurator reports due to the formatting/file system of the floppy disk."	
ItemInformationNotes	Free text field. This field should capture important information on the media carrier including information on the carrier, carrier's label, and carrier's housing. Scan any accompanying information if necessary.	

	When you transcribe this information, differentiate between where the information is coming from. e.g. Notes: Transcribed information from the HD box: “ARCHIVES UNBOUND”, Transcribed information from the label on the HD: "Archives Unbound The American Indian Movement and Native American Radicalism", "256479A", "GALE CENGAGE Learning". Note: The "A" in "256479A" is crossed out or filed in.)	
BagCreator	First name and last name of the person who created the bag	This information should be recorded in case the bag creator is different from the disk image creator. If there are any problems, questions, or missing information, a future user will want to know who to direct their inquiry to.

Custom Metadata Profile and PREMIS

Certain data points in our profile were modelled after PREMIS event entities. “The Event entity aggregates information about actions that affect objects in the repository. An accurate and trustworthy record of events is critical for maintaining the digital provenance of an object, which in turn is important in demonstrating the authenticity of the object."[2] In PREMIS, this metadata is typically encoded, and can be viewed as XML. Unlike PREMIS, our profile uses PascalCase for its semantic units, whereas the PREMIS XML schema uses camelCase. Additionally, PREMIS utilizes XML schema, which means that each eventType has multiple components. For example, in PREMIS virus check, there are subcomponents nested within the event for the results of the scan, as well as what software was used to complete this service. Due to the nature of our bagger profile, nesting these components was not an option. Therefore, we attempt to keep our 'event' names similar, with its components listed one after another.

Although not required, the "information that can be recorded about events includes: a unique identifier for the event (type and value), the type of event (creation, ingestion, migration, etc.), the date and time the event occurred, a detailed description of the event, a coded outcome of the event, a more detailed description of the outcome, agents involved in the event and their roles, objects involved in the event and their roles."[3] Our three VirusCheck units bare the strongest resemblance to PREMIS eventTypes, specifically in this case the virus check. However, the Fixity check eventType in PREMIS is similar to our two units ImageHash and HashSignatureType; the Quarantine eventType bares resemblance to TestMountPerformed, where a user can select that a test mount was not performed and the digital object was placed in a quarantine due to a failed virus check; other PREMIS units we have imitated are the identifierValue (AccessionNumber/Barcode), and the Ingestion eventType.

Custom Metadata Profile and .info Redundancy

It is important to note that the custom profile is at times redundant with the .info file. The following profile units are also represented in the .info file: AccessionNumber/Barcode, ImageHash, HashSignatureType, Medium/Carrier. We are aware of this redundancy, but feel that it is necessary to provide all of the relevant metadata in one centralized location.

[2] Ibid.
[3] Ibid.

Metadata in Data Accessioner

Field	Data
Your Name	First name and last name of the person migrating the content
Accession Number	barcode
Collection Title	Collection ID = “Robert Altman”

Metadata Crosswalk

Bagger- Guymager - DataAccessioner- Inventory Spreadsheet

Bagger field	Guymager field	DataAccessioner field	Inventory Spreadsheet	Metadata source/value
BarcodeNumber/Identifier	Image filename/ Info filename	Accession Number	barcode	barcode
AccessionNumber/Collection	Case number	Collection Title	NA	Collection ID = “Robert Altman”
BagCreator	Examiner	Your Name	NA	First name and last name of the person who created the disk image/bag/filling in the metadata
Medium/Carrier / Medium/Carrier #2	Description	NA	Media Type	Describes the media carrier that data was stored prior to imaging.
ItemInformationNotes	Notes	NA	Title, Notes (including information on boxes/hous	This field should capture important information on the media carrier including information

			ing) columns	on the carrier, carrier’s label, and carrier’s housing. Scan any accompanying information if necessary. When you transcribe this information, differentiate between where the information is coming from. o e.g. Ntes: Transcribed information from the HD box: “ARCHIVES UNBOUND”, Transcribed information from the label on the HD: "Archives Unbound The American Indian Movement and Native American Radicalism", "256479A", "GALE CENGAGE Learning". Note: The "A" in "256479A" is crossed out or filed in.)