# Walk This Way:

## Detailed Steps for Transferring Born-Digital Content from Media You Can Read In-house

**By Ricky Erway**

OCLC Research

# Walk This Way:
## Detailed Steps for Transferring Born-Digital Content from Media You Can Read In-house

**Julianna Barrera-Gomez**
OCLC Diversity Fellow

**Ricky Erway**
Senior Program Officer

**OCLC Research**


**A publication of OCLC Research**

Walk This Way: Detailed Steps for Transferring Born-Digital Content from Media You Can Read In-house

Julianna Barrera-Gomez and Ricky Erway, for OCLC Research




June 2013

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 1-55653-454-X (978-1-55653-454-6)
OCLC (840612084)

Please direct correspondence to:
Ricky Erway
Senior Program Officer
erwayr@oclc.org

Suggested citation:
Barrera-Gomez, Julianna and Ricky Erway. 2013. *Walk This Way: Detailed Steps for Transferring Born-Digital Content from Media You Can Read In-house*. Dublin, Ohio: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2013/2013-02.pdf.

# Contents

# Introduction

The OCLC Research report, *You've Got to Walk Before You Can Run: First Steps for Managing Born-Digital Content Received on Physical Media,* offers eleven simple steps for getting born-digital content off of physical media that you can read using in-house resources and into a more manageable form (2012 p. 5). Project managers and archivists just getting started with born-digital materials welcomed its brevity and simplicity, but they likely had myriad questions as to how to implement them.

This document is a companion to the *First Steps* report. Like *First Steps*, the intended audience is those who are just starting to manage born-digital materials, from those wondering where to begin, to those who are actively planning their archival workflow. This paper will take the reader to the point where the digital content has been successfully transferred and preliminary processing of the files has begun. It ends with sample workflows to help readers conceptualize the process, resources that may provide additional avenues of learning and research, and ways to become engaged in the digital archives community.

The advisors who helped us with the *First Steps* report had offered a lot of ancillary recommendations not ultimately included in that report in order to keep it brief and simple. This experience-based advice forms the basis of this document, which describes the steps in more detail, and suggests software, tools, and resources for learning more about each step.

We sought to make these steps as basic as possible to accommodate the various skill sets and resources that readers (including non-archivists) may have. These steps are a guide to the transfer of born-digital material from media and the documentation of what was done to the digital files. We present these steps as discrete phases of workflow, with a variety of standalone tools for each phase of the most basic workflow, while pointing to more all-inclusive tools to meet larger-scale accessioning needs. Your final workflow may differ in the number of steps, or the order in which steps are taken, depending on your institution's goals and resources.

It is important to consider these actions in the context of existing digital preservation policies of your organization. These may include donor agreements (which can explain what information may be transferred from digital media) and policies on accessioning or

deaccessioning records or physical media. You should also consult your organization's IT policies on software use or server backups. If no policies are in place, read through the steps and begin thinking about what policies you will need to develop. The AIMS Project published a detailed report (2012) about digital collections stewardship that provides objectives for informing policy as well as a glossary for terms that non-archivists may not be familiar with. It may be helpful to consult this resource as you gain more familiarity with the processes.

All links in this report are current as of the date of publication. Many of the software resources listed are available for free download. Please note the "Supplementary Exploration" section at the end of this document, which includes "Learning Opportunities" and "Discussion Forums" to help provide avenues of engagement with colleagues and exposure to the latest trends and tools in digital preservation.

# Documenting the Project

*Before embarking on your transfer project, carefully consider the documentation needs. Now is the time to plan what information about the process will be needed in the future to understand what the project includes, the steps that were taken, and why. Developing a consistent approach may also help you incorporate this workflow into automated actions at some future point.*

**Level of Difficulty:** Easy
**Desirability:** Highly Recommended

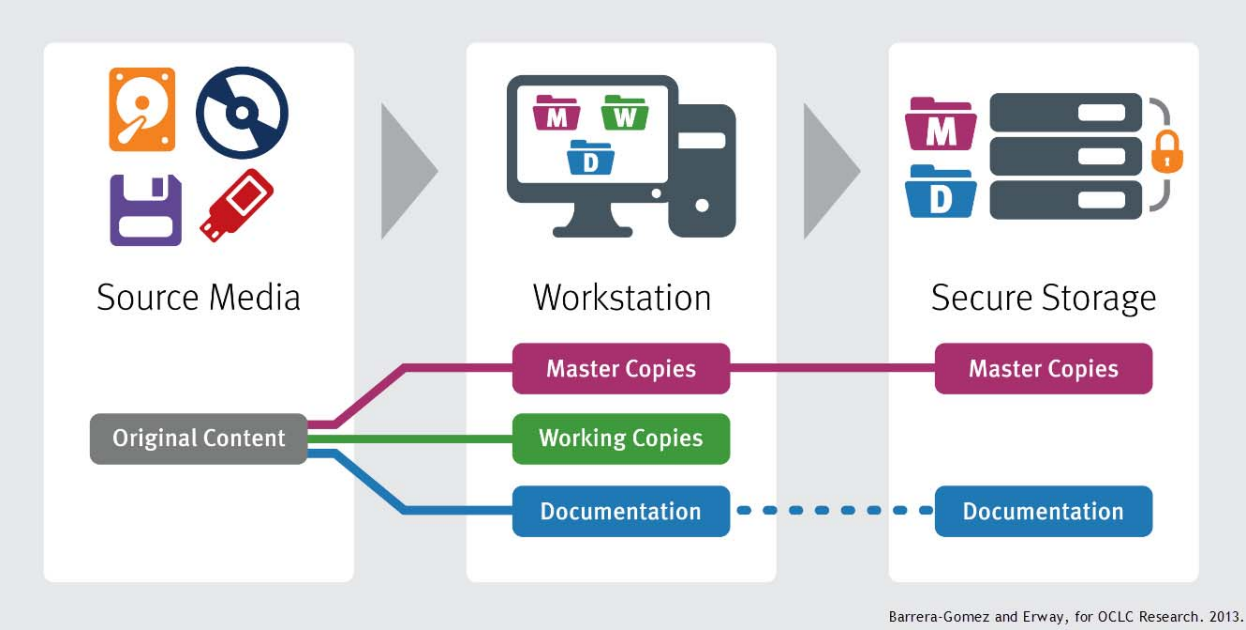
Documentation is crucial because it provides context to the entire process and forms a key part of the evidence for provenance (i.e., where the material came from); it also indicates the authenticity of the material. At the outset, ask yourself what information you or your organization will need to understand what the project included, the goals of the transfer process, what steps were taken, what was done to the content, and who was involved.

It is important to document your project and manage the associated metadata that will be either automatically generated or manually created. If your organization is using a content management system (e.g., Archivist's Toolkit or Archon), create an accession record in your system to provide a link from this project's documentation to your other holdings.

To manage the project on your workstation, create a project directory with folders. Give it a meaningful name, or use the donor or accession number from your content management system. Within the directory, create folders such as these:

- Master Folder (to hold the master copy of the file)
- Working Folder (to hold working copies of the master copy)
- Documentation Folder (to hold metadata and other information associated with the project)

During the course of the project, you will generate master copies of your source media's original content, and from this you will generate working copies that will be used to view and run software during the various processes. Using the working copies will reduce the risk of altering metadata or accidentally changing or deleting files from the master copies. These master copies, and the documentation, will be sent to secure storage where the master copies are preserved and the documentation is updated as needed.




## Transferring Born-Digital Content at a Glance

After finalizing the steps in "Securing Project Files" (p. 17), the Working Folder of copies will be useful for the "Assess Content" steps (p.22). These steps will provide an opportunity to learn more about the content of your accession and provide information that you can use to describe the contents and link it to its associated collection. When all steps have been completed, you may keep the Working Folder on your workstation (for further use and to create copies for users to access), save these as revised master copies if you have altered the files (such as redaction or arrangement), or you may delete them and make new working copies of the master copies as needed.

# Preparing the Workstation

*Prepare a dedicated workstation to connect to the source media and to use throughout the project. Start with a single type of media from a collection to aid in efficiency and in keeping track of your materials and metadata.*

## Set up a "Clean" Workstation

**Level of Difficulty:** Easy to Moderate
**Desirability:** Mandatory

Use a computer that is regularly scanned for viruses. This workstation serves the same purpose as the quarantine room many archives use for new acquisitions that have not yet been reviewed for mold, insects, etc. Keeping this computer virus-free is critical. Virus check your workstation using your institution's current virus protection system before beginning to copy materials, and plan to keep the software up to date and the workstation regularly scanned. See the "Check for Viruses" step (p. 14) for software options.

Consider keeping this computer non-networked until a network connection is needed (such as during file transfers or when updating software or virus definitions) to reduce the risk of vulnerability of your workstation. Keeping the workstation non-networked will help mitigate common internet risks, such as viruses introduced through email. For collections that may have restrictions, restricting network connectivity could also prevent unauthorized network access to your in-process project.

You will need the relevant peripherals such as cables, drives, or ports for transferring digital files from the source media. You may need particular software or drivers to access the hardware—consult with your IT department or other qualified technicians to set this up appropriately.

Do not open files on your source media; it could change the files' metadata (such as creator and date) or the content. We strongly recommend that you use a write blocker (software or hardware) to prevent your actions on the workstation from altering the files on your source media.

### *Resources*

- Olsen, Porter. 2012. "Digital Curation Workstation," *MITH (Maryland Institute for the Digital Humanities)* (blog). 26 November. http://mith.umd.edu/digital-curation-workstation.
  This is an example of a workstation created by MITH.

## Install Write Blockers

**Level of Difficulty:** Easy to Moderate
**Desirability:** Mandatory

Write blockers can be software or hardware tools. If software is used, it must be compatible with your operating system. Hardware write blockers are physical devices that are connected between the media and your workstation and are not software-dependent.

Write blockers allow information to be read from the media while preventing the computer from overwriting or altering file metadata and content. Use a write blocker to prevent changes to the original content, especially when connecting magnetic source media (diskettes, hard drives, or tape) to your workstation. 3 ½ inch floppy disks usually have switches that can be flipped open to block writing. 5 ¼ inch floppy disks usually have notches that can be covered to block writing. Write-once CDs and DVDs do not require write blockers. Writable optical disks are much less likely to be altered than their magnetic and flash-based counterparts, but using an optical drive without writing capability is nevertheless recommended.

### *Tools and Software*

**Hardware tools:**

- **Tableau**: http://www.tableau.com/index.php?pageid=products
- **ICS Drive Lock**: http://www.ics-iq.com/Super-DriveLock-Write-Blocker-Write-Protector-p/f.gr-0028-0000.htm
- **WiebeTech USB WriteBlocker:** http://www.wiebetech.com/products/USB-WriteBlocker.php
- **Digital Intelligence Computer Forensics Hardware:** http://www.digitalintelligence.com/forensichardware.php

- **FC5025 floppy controller card**: http://mith.umd.edu/vintage-computers/fc5025-operation-instructions
- **KryoFlux floppy controller:** http://www.kryoflux.com
- **DiscFerret:** http://discferret.com/

**Software tools:**

- **ForensicSoft SAFE Block:** http://www.forensicsoft.com/
- **MacForensicsLab Write Controller:** http://www.macforensicslab.com/ProductsAndServices/index.php?main_page=product_info&cPath=1&products_id=339
- **GitHub Disk Arbitrator:** https://github.com/aburgh/Disk-Arbitrator

### *Resources*

- "Write Blockers" 2012. Forensics Wiki. Last modified 23 July. http://www.forensicswiki.org/wiki/Write_Blockers
- Newton, Derek. 2010. "Write Blockers—Hardware vs Software." *Information Security Insights* (blog). http://dereknewton.com/2010/05/write-blockers-hardware-vs-software/.
  This is a detailed discussion of the pros and cons of software and hardware write blockers.

## Connect the Source Media

**Level of Difficulty:** Easy
**Desirability:** Mandatory

Once your workstation has been set up, you are ready to connect the source media. Examine the media for any defects such as cracks or breaks that could damage the disk content or your workstation, if inserted. If the media contains any removable ephemera (such as sticky notes or other non-permanent adhesives that could become loose and be potentially harmful for your disk drive) consider removing these and setting them aside. If desired, take a digital photograph of the media showing any labels or ephemera to further document it. Store digital photographs in your documentation folder.

Insert the media in the appropriate drive or connect it to the appropriate port. Do not attempt to open any files at this stage.

### *Resources*


- Brown, Adrian. 2008. *Care, Handling and Storage of Removable Media*. The National Archives Digital Preservation Guidance Note 3. Washington DC: The National Archives. http://www.nationalarchives.gov.uk/documents/information-management/removable-media-care.pdf.

- Byers, Fred R. 2003. *Information Technology: Care and Handling of CDs and DVDs—A Guide for Librarians and Archivists.* National Institute of Standards and Technology Special Publication 500-252. Gaithersburg, Maryland: National Institute of Standards and Technology and Council on Library and Information Resources. http://www.itl.nist.gov/iad/894.05/docs/CDandDVDCareandHandlingGuide.pdf.

# Transferring the Data

*The following steps address how to safely copy the digital data from your source media, generate metadata about the digital files, and link the documentation of these procedures to your project. Once you have transferred the data to your workstation, you will create working copies of your master copies. The working copies will be used in other processing steps.*

## Copy the Files or Create a Disk Image

**Level of Difficulty:** Easy to Complex
**Desirability:** Mandatory

There are two approaches to copying the content from the physical media: 1) copying the files individually or in groups or 2) making a disk image. The first approach is adequate for some purposes and can be a practical way for archivists new to this work to get started. With the second approach, more information is captured and it's easier to ensure authenticity.

Disk imaging makes an exact, sector-by-sector bitstream copy of a disk's contents, retaining original metadata. The intent of the disk imaging process is to make a single file containing an authentic copy of the files and the file system structure on a disk, allowing you to store that file somewhere less vulnerable than the source media. Disk images that image everything (including deleted files and unallocated space) are called "forensic images," while those that omit deleted files and unallocated space are referred to as "logical copies." Disk images of large-capacity media may need to be split up into multiple smaller files if you have size constraints. Any decisions to capture forensic images must be made in accordance with your institution's policies and any donor agreements about what may be extracted from digital media.

## *Tools and Software*

### File Copying Tools

- **Duke Data Accessioner:** http://library.duke.edu/uarchives/about/tools/data-accessioner.html
  Created by the Duke University Archives, this tool has "a simple GUI interface to allow technical services staff an easy way of migrating data off disks and onto a file server for basic preservation, further appraisal, arrangement, & description. It also provides a way to integrate common metadata tools at the time of migration rather than after the fact. With a simplified interface and being written in Java it is intended to be easily adopted by smaller institutions with little or no IT staff support." It integrates checksum tools, JHOVE/DROID for file identification, and metadata extraction tools that generate XML reports.

- **rsync:** http://en.wikipedia.org/wiki/Rsync

- **Zip:** This is a file format for data compression and archiving. Zip folders contain multiple files bundled within either as-is or compressed to reduce file size. Copying files from the original media to a Zip file is one of the most basic ways to copy files. Operating systems often come with built-in tools for creating zipped folders:

  - Window's Compressed Folder tool
  - Mac OS X's Archive Utility

  For other tools to create zip archives using different compression algorithms, see http://en.wikipedia.org/wiki/Comparison_of_file_archivers.

### Disk Imaging Tools

- **FTK Imager (Forensic ToolKit Imager):** http://accessdata.com/support/adownloads
  FTK Imager, created by Access Data, is one of the commonly used tools for forensic disk imaging in the archives community. It can create disk or logical images: the logical images are in a commercial format, while the forensic image options include both non-proprietary and proprietary formats. It works on Windows and can read Mac, Linux and Windows file systems. FTK Imager also can automatically generate checksums during transfer and at later stages of managing the file (see the Run Checksums or Hashes section); create a text file with date and time stamps and checksums; and allow you to view deleted files and unallocated space. This commercial software can be downloaded for free.

- **BitCurator:** http://www.bitcurator.net/ and http://wiki.bitcurator.net
  Created as part of the BitCurator Project (a collaborative effort between the School of Information and Library Science at the University of North Carolina at Chapel Hill and the Maryland Institute for Technology in the Humanities), the BitCurator Environment is a fully functioning Linux system built on Ubuntu 12.04 that has been customized to meet the needs of digital archivists. It can be run either as a standalone operating system or as a virtual machine. Once installed, the BitCurator environment incorporates a number of useful digital forensics tools that can easily be integrated into digital curation workflows.

  A sampling of these tools includes:

  - **Guymager:** a tool for creating disk images in one of three commonly used disk image formats (dd, E01, and AFF)
  - **Custom Nautilus scripts:** A collection of enhancements to Ubuntu's default file browser that allow users to quickly generate checksums, identify file types, safely mount drives, and more
  - **The Sleuth Kit:** an open-source digital investigation platform
  - **fiwalk:** an open source tool for processing disk images, producing Digital Forensics XML and human-readable metadata on file system structure and contents
  - **bulk_extractor:** a program that extracts information (including Personally Identifying Information, or PII) from disk images without parsing the file system. bulk_extractor generates reports on that information in both human and machine readable formats, and includes a GUI front-end, Bulk Extractor Viewer
  - **sdhash 2.x:** A tool to evaluate file similarity using similarity digests
  - **Ghex:** an open-source hex editor that allows viewing a file in hexadecimal format

  In addition, the BitCurator team is in the process of building Python-based reporting tools that reprocess and provide visualizations based on the output of forensics tools that produce Digital Forensics XML; these tools are currently distributed separately via GitHub and will be integrated into the BitCurator environment as the project progresses.

- **Disk Copy/Disk Utility:** Utilities bundled with Mac OS X, specifically Disk Copy in Mac OS X v10.2 and earlier, and Disk Utility in Mac OS X v10.3 and later.
- **Guymager:** http://guymager.sourceforge.net/
- More on disk imaging tools http://www.forensicswiki.org/wiki/Category:Disk_Imaging

### *Resources*


- AIMS Work Group. 2012. *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship.* Charlottesville, VA: University of Virginia Library. http://www.digitalcurationservices.org/files/2013/02/AIMS_final.pdf. See "Chapter 2: Accessioning" for more on the process of accessioning digital files and the decision points, tasks, and impacts.
- Gengenbach, Martin J. 2012. "'The Way We Do it Here' Mapping Digital Forensics Workflows in Collecting Institutions." A Master's Paper for the M.S. in L.S degree. August 2012. http://digitalcurationexchange.org/system/files/gengenbach-forensic-workflows-2012.pdf. See the Findings section (pp.27-68) for detailed tool lists and workflows of eight prominent digital archiving programs.
- John, Jeremy L. 2012. *Digital Forensics and Preservation.* DPC Technology Watch Report 12-03. Great Britain: Digital Preservation Coalition. http://www.dpconline.org/advice/technology-watch-reports.
- Kirschenbaum, Mathew.G., Richard Ovenden, and Gabriela Redwine. 2010. *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Washington, DC: Council on Library and Information Resources. http://www.clir.org/pubs/abstract/reports/pub149.
- Wilsey, Laura, Rebecca Skirvin, Peter Chan, and Glynn Edwards. 2013. "Capturing and Processing Born-Digital Files in the STOP AIDS Project Records: A Case Study." *Journal of Western Archives*, 4, no. 1. http://digitalcommons.usu.edu/westernarchives/vol4/iss1/1
- Woods, Kam, Christopher A. Lee and Simson Garfinkel. 2011. "Extending Digital Repository Architectures to Support Disk Image Preservation and Access." *Proceedings of the Eleventh ACM/IEEE Joint Conference on Digital Libraries (JCDL 2011).* Ottawa, Ontario, Canada, June 13-17, 2011, (pp. 57-66). New York, NY: ACM Press. http://ils.unc.edu/callee/p57-woods.pdf.

## Check for Viruses

**Level of Difficulty:** Easy to Moderate
**Desirability:** Highly Recommended

In addition to regularly scanning your workstation, you will need to scan every batch of copied files to check for viruses. This process may be done during different steps in the transfer process depending on whether you use disk imaging or file copying and what your goals are for preserving metadata. To avoid any alterations to the master files or metadata, run virus software only on a working copy generated from the master copy.

When you find a virus, the next steps will vary depending on the nature of the content and the virus. One approach is to note the names and locations of the infections and, if possible, obtain a clean copy from the donor. Another approach is to create a "cleaned" copy using antivirus software (be sure to document this action). Keep infected files quarantined until you determine the best course of action.

### *Tools and Software*

Your workstation is probably equipped with your institution's current virus protection system. If not, free virus checking tools are available for download:

- **ClamAV (Windows):** http://www.clamav.net/lang/en/
  This is a user interface or command line tool.
- **ClamXav (Mac):** http://www.clamxav.com/
  This is a command line only tool.
- **AVG (Windows)**: http://free.avg.com
- **Microsoft Defender:** http://www.microsoft.com/en-us/download/details.aspx?id=17
- **avast:** http://www.avast.com/en-us/free-antivirus-download

Additionally, you may want to check for spyware, especially if material is being copied from a hard drive. Spyware checking software is freely available.

- **Ad-Aware (Windows):** http://www.lavasoft.com/
- **Spybot (Windows):** http://www.safer-networking.org/
- **MalwareBytes (Windows):** http://www.malwarebytes.org/

## Record the File Directory

**Level of Difficulty:** Easy to Moderate
**Desirability:** Recommended

An important step in identifying your digital content is to make a copy of the directory tree. After creating a working copy of the files or disk image, use a directory tree "printer" to document the folders and files that are in your project, including file name, date, size, and file extension.

### *Tools and Software*

- **fiwalk**: http://www.forensicswiki.org/wiki/Fiwalk
  This command line tool in forensic analysis can be used to create a list of a disk image's file system.
- **Karen's Directory Printer:** http://www.karenware.com/powertools/ptdirprn.asp
  This easy-to-use GUI interface can be used to print metadata about digital files with the option to generate checksums.
- **FTKImager:** http://www.accessdata.com/support/product-downloads
  This tool can automatically create a file list after creating the disk image.
- **NARA File Analyzer and Metadata Harvester:**
  https://github.com/usnationalarchives/File-Analyzer
  This is a directory analyzer with the ability to generate checksums.

## Run Checksums or Hashes

**Level of Difficulty:** Easy to Complex
**Desirability:** Highly Recommended

A checksum, or hash, is a unique value based on the contents of a file and is generated by specific algorithms (e.g., MD5 or SHA-256). Comparison of checksums generated from the same file at different times identifies whether and when the file has changed. Creating checksums is not difficult and may be done during several processes described earlier (such as creating a disk image, generating a directory list, or using the Duke Data Accessioner). It is very easy to create a hash for a single file and then to compare that hash to one generated for another copy of the file. An automated technique is necessary, however, when processing a large number of files.

It is important to note that while a changed checksum can alert a repository to the fact that something in a file or folder has changed, it cannot indicate what exactly has changed, nor can it reverse the change. Regularly hashing the file or image you have copied and checking those new hashes against the hashes made at the time of the transfer should be part of your digital curation workflow. During the lifecycle of your digital collections you will need to periodically verify the checksums to ensure that files remain unchanged.

### *Tools and Software*

**Disk imaging or Disk copying tools that incorporate checksums**:

- **BitCurator:** http://www.bitcurator.net/
- **FTK Imager (Forensic ToolKit Imager):** http://accessdata.com/support/adownloads
- **Duke Data Accessioner:** http://library.duke.edu/uarchives/about/tools/data-accessioner.html

See "Copy the Files or Create a Disk Image" section (p. 10) for more details on these tools.

**File directory printing tools that incorporate checksums**:

- **Karen's Directory Printer:** http://www.karenware.com/powertools/ptdirprn.asp
- **Beyond Compare:** http://www.scootersoftware.com/
- **NARA File Analyzer and Metadata Harvester:** https://github.com/usnationalarchives/File-Analyzer

See "Record the File Directory" section (p. 17) for more details on these tools.

**Collection management tools that incorporate checksums:**

- **Archivematica:** https://www.archivematica.org/wiki/Main_Page
  Archivematica is "a free and open-source digital preservation system that is designed to maintain standards-based, long-term access to collections of digital objects. Archivematica uses a micro-services design pattern to provide an integrated suite of software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model." Developed under a collaboration led by Artefactual Systems and the City of Vancouver Archives.
- **Curator's Workbench:** http://www.lib.unc.edu/blogs/cdr/index.php/about-the-curators-workbench

Developed at the University of North Carolina, Chapel Hill, "the Workbench helps archivists manage files before they are stored in an institutional repository or dark archive. As the files are selected, arranged, and described, a METS file is generated by the software that documents these processes. In addition, checksums and UUIDs are generated for each object and MODS descriptive metadata elements can be mapped to individual objects and folders."

**Standalone checksum tools:**

- **Jacksum:** http://www.jonelo.de/java/jacksum/
- **Md5summer (Windows):** http://www.md5summer.org
- **Md5deep:** http://md5deep.sourceforge.net
  This command line tool can also be used as a directory printer.

### *Resources*

- Prom, Chris. 2012. "Checksum Verification Tools: Guest Post by Carol Kussmann" *Practical E-Records* (blog). Accessed April 2013.
  http://e-records.chrisprom.com/checksum-verification-tools/
  This is a review of five checksum generating and verification tools.

# Securing Project Files

*These steps address how to take the first step in preservation of your digital files by moving master copies and documentation to a secure location, as well as how to handle the source media once the digital transfer is complete.*

## Consolidate Documentation

**Level of Difficulty:** Easy
**Desirability:** Highly Recommended

Throughout the project, note which steps were taken and any issues that have arisen, such as viruses discovered or software that did not perform as expected. Formalize these notes as a record of your process, and store them with your project materials or in another associated location. Consider creating a document, such as a text file or spreadsheet (named "readme" or labeled with the accession or project number) that describes the project, the files, the pertinent information from the inventory, and all subsequent actions and issues along with the name of the person doing each step. If you are using a collection management system, you may add this information to your accession record directly.

This record should be tied to the documentation actions performed in the "Documenting the Project" step (p. 5) and stored in the project folder on your workstation. Documenting as you work will make it possible for others to see what was done with the files and the various preservation events during the life of the digital content.

## Prepare for Storage

**Level of Difficulty:** Moderate
**Desirability:** Highly Recommended

After completing all previous steps, you can begin to prepare your digital files for storage. You will be storing the Project Folder that you created on your workstation and two of the subfolders: the master files (file copies or disk images in the Master Folder) and all the associated metadata you have generated (such as the file directory listing, checksums, etc.)

along with the accession record. If your organization does not have a digital archive, arrange for space on a backed-up network server that is secure.

At this point, you will relate this project to other collection components. The following are all-in-one tools developed to aid archivists in managing and maintaining ties between digital records and metadata, as well as in creating links to the collection itself.

### *Tools and Software*

See the "Run Checksums or Hashes" section on p. 15 for more information on these tools:

- **Archivematica:** https://www.archivematica.org/wiki/Main_Page
- **Curator's Workbench:** http://www.lib.unc.edu/blogs/cdr/index.php/about-the-curators-workbench

### *Resources*

- Gengenbach, Martin J. 2012. "'The Way We Do it Here' Mapping Digital Forensics Workflows in Collecting Institutions." A Master's Paper for the M.S. in L.S degree. August, 2012. http://digitalcurationexchange.org/system/files/gengenbach-forensic-workflows-2012.pdf
  See "Findings" (pp.27-68) for detailed tool lists and workflows of eight prominent digital archiving programs.

## Transfer to a Secure Location

**Level of Difficulty:** Easy to Moderate
**Desirability:** Highly Recommended

At this point, the master files you have generated will be stored on a secure server that is backed up regularly. Ideally, an additional copy will be stored in a different physical location to prevent loss in the event of disaster. Be sure that your organization's IT disaster preparedness plan reflects your digital preservation needs.

Transfer the Master Folder and the Documentation Folder to your digital archive or secure server space designated for preservation of your born-digital files. Run checksums and verify that the transfer was successful.

The files in this server location and in its backups are the preservation master copies and must be kept safe from unintentional alteration. As emphasized throughout this report, any

time you wish to work with a file, or provide access to it, use a working copy instead of working directly with the master files.

### *Tools and Software*

- **Archivematica**: https://www.archivematica.org/wiki/Main_Page
- **BagIt Transfer Utilities**: http://en.wikipedia.org/wiki/BagIt
  BagIt is a tool that was developed by the Library of Congress and its partners in the National Digital Information Infrastructure and Preservation Program (NDIIPP) for the purpose of validation and transfer of data that conforms to the BagIt specification. It generates a file directory with checksums to validate the data was successfully transferred.

## Store or Deaccession the Source Media

**Level of Difficulty:** Easy to Moderate
**Desirability:** Optional

At this point, consider what to do with the physical media from which you have successfully transferred digital content. You may return the source media to physical storage or destroy it. If the decision is to destroy the source media, make sure that a secure method of destruction is used. Make this decision in conjunction with relevant policies and the donor agreement.

### *Resources*

- Adelstein, Peter Z. 2009. *IPI Media Storage Quick Reference, 2nd Ed*. (Rochester NY: Image Permanence Institute, Rochester Institute of Technology). https://www.imagepermanenceinstitute.org/webfm_send/301.
  This is a reference about storing conditions for magnetic and optical media.
- Brown, Adrian. 2008. *Care, Handling and Storage of Removable Media*. The National Archives Digital Preservation Guidance Note 3. Washington DC: The National Archives. http://www.nationalarchives.gov.uk/documents/information-management/removable-media-care.pdf.
- Byers, Fred R. 2003. *Information Technology: Care and Handling of CDs and DVDs—A Guide for Librarians and Archivists.* National Institute of Standards and Technology Special Publication 500-252. Gaithersburg, Maryland: National Institute of Standards

and Technology and Council on Library and Information Resources. http://www.itl.nist.gov/iad/894.05/docs/CDandDVDCareandHandlingGuide.pdf.

- Kirschenbaum, Matthew, Erika L. Farr, Kari M. Kraus, Naomi Nelson, Catherine Stollar Peters, Gabriela Redwine, and Doug Reside. 2009. "Digital Materiality: Preserving Access To Computers As Complete Environments." In: *iPress 2009: The Sixth International Conference on Preservation of Digital Objects Proceedings, 5-6 October 2009, San Francisco, California.* http://escholarship.org/uc/item/7d3465vg. This reference provides a discussion on preserving media and digital environments for users.

## Validate File Types

**Level of Difficulty:** Moderate to Complex
**Desirability:** Optional

Many of the tools for disk imaging, file copying, file directory printing, and collection management incorporate features for validating and identifying file types. This information will help you determine whether you can open and read the contents of digital files. Using working copies, you can try to open files with current software to see if they will render in a way that is useful for your purposes.

If you have legacy files with no file extension or files you cannot render, hex editors can be useful for discovering more about these files by looking at the file properties. Hex editors allow you to examine files (including disk images) as individual byte representations, typically expressed in both hexadecimal notation and, where possible, the corresponding ASCII value of the byte. Examining information in the file header may help you make a determination as to file type. Additionally, hex editors can be used to recover strings of text even if the file cannot be properly rendered. Numerous standalone hex editors are available freely, and disk imaging tools such as FTK Imager and BitCurator incorporate these.

After identifying and characterizing your files, you may want to consider the sustainability of the file formats. As software changes over time, or is no longer supported, software-dependent files can be at risk of becoming unusable. The Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP) has resources on the sustainability of digital formats, including detailed format descriptions and other information that can help you decide if files should be converted to lower-risk formats. It is worth noting that the more formats your organization chooses to maintain, the more effort will be needed over time to maintain them.

## *Tools and Software*

Further research may be needed to identify and characterize files and formats. Popular applications include:

- **JSTOR/Harvard Object Validation Environment (JHOVE/JHOVE2):** https://bitbucket.org/jhove2/main/wiki/Home
  This tool is used to identify formats, and validate and characterize files.
- **Digital Record Object Identification (DROID):** http://sourceforge.net/projects/droid/
  DROID is bundled within JHOVE2, and is used to identify formats, characterize files, and generate checksums.
- **File Information Toolset (FITS):** http://code.google.com/p/fits/
  FITS bundles JHOVE & DROID which is used to Identify formats, and validate and extract metadata.
- **File Identification for Digital Objects (FIDO):** http://www.openplanetsfoundation.org/software/fido
- **HxD (Hex editor for Windows):** http://mh-nexus.de/en/hxd
- **0xED (Hex editor for OS X):** http://www.suavetech.com/0xed

## *Resources*


- Prom, Chris. 2012. "Characterizing Files" *Practical E-Records: Software and Tools for Archivists* (blog). Posted 29 April 2011. http://e-records.chrisprom.com/resources/software/accessioningingest/identifying-and-characterizing-files/.
- Kirschenbaum, Mathew G., Richard Ovenden, and Gabriela Redwine. 2010. *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Washington, DC: Council on Library and Information Resources. http://www.clir.org/pubs/abstract/reports/pub149
  For additional information on data recovery and use of hex editors, see section 2.5, "Data Recovery" (pp. 39-46).
- Library of Congress. 2013. "Sustainability of Digital Formats: Planning for Library of Congress Collections." NDIIPP (National Digital Information Infrastructure and Preservation Program). Last updated 20 March. http://www.digitalpreservation.gov/formats.

- The National Archives. 2013. "The Technical Registry PRONOM." Digital Preservation. Accessed 23 April. http://www.nationalarchives.gov.uk/PRONOM/Default.aspx. This is an online registry of technical information about file formats, maintained by the National Archives of the United Kingdom.

- Nelson, Naomi L., Seth Shaw, Nancy Deromedi, Michael Shallcross, Cynthia Ghering, Lisa Schmidt, Michelle Belden, Jackie R Esposito, Ben Goldman and Tim Pyatt. 2012. *SPEC Kit 329: Managing Born-Digital Special Collections and Archival Material.* Washington, DC: Association of Research Libraries. August. http://publications.arl.org/Managing-Born-Digital-Special-Collections-and-Archival-Materials-SPEC-Kit-329 .
  See "Format Policies" for more on sustainability of file formats.

# Assess Content

**Level of Difficulty:** Easy
**Desirability:** Optional

By now you have successfully created a working copy of the digital content onto your workstation and have created metadata to document what it contains. You have documented basic information about your files such as file names, dates, format types, and sizes, along with notes about your process and any issues you encountered. Keep this information together and use it to update your accession record and make preliminary appraisal decisions.

Ideally, your organization has already assessed the overall value of the collection of which your files are a part, but you may not be familiar with the content of the individual files, or you may not intend to keep all of them (such as duplicates or out-of-scope material). Now that your digital data has been successfully transferred and master copies are secure, use the working copies to begin exploring the content. The following steps will help you learn more about the nature of your digital data and depending on what you discover, you may choose to utilize one of more of them. If your primary goal was to extract the material from the physical media, decisions about what to keep and how to arrange your new digital collection may come later.

## Reviewing Files

The metadata you generated while copying digital files will give a preliminary indication of the types of files you have. You may recognize modern file extensions and have software that can open these files. You may have to install obsolete software to open legacy formats. The following tools may help you to open files and explore their content. Remember, attempt to open only your working copies to prevent accidental alteration or deletion of master copies and associated metadata.

### *Tools and Software*

- **Quick View Plus:** http://www.avantstar.com/metro/visit
  This can be used to open and view files in many formats to explore content.

- **TreeSize Professional:** http://www.jam-software.com/treesize/
  This can be used to view directory structure.
- **WinDirStat (Windows):** http://windirstat.info/
  This can be used to view directory structure.
- **Disk Inventory X (Mac OS X 10.3 and later):** http://www.derlien.com/
  This can be used to view directory structure.
- **IrfanView:** http://www.irfanview.com/
  **This can be used** to view raster image files.
- **Inkscape:** http://inkscape.org/
  This can be used to view vector images.
- **VLC Media Player:** http://www.videolan.org/vlc/index.html
  This can be used to open audio and video files.

### *Resources*

- AIMS Work Group. 2012. *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship*. Charlottesville, VA: University of Virginia Library. http://www.digitalcurationservices.org/files/2013/02/AIMS_final.pdf.
  See "Chapter 3: Arrangement and Description" and "Chapter 4: Discovery and Access" for more on the decision points and impacts to your archival process.
- Daines III, J. Gordon. 2013 "Module 2: Processing Digital Records and Manuscripts" In: *Trends in Archives Practice: Archival Arrangement and Description*. Edited by Christopher J. Prom and Thomas J. Frusciano. Chicago, Illinois: Society of American Archivists. http://saa.archivists.org/store/archival-arrangement-and-description-print/3033/.
  See "Preparing to Process" and "Developing Policies and Procedures" (pp. 113-115) for more on the decision points and policies for processing digital collections.

## Finding Duplicate Files

While exploring and assessing the digital records, you may want to identify duplicate files and delete redundant information. Tools are available to help you find multiple files with the same name or to identify those with the same checksum. As with paper-based material, examine and contextualize all files before removing potential duplicates to make sure that they truly are redundant.

If you have unwanted duplicate files, you may choose to simply note these and mark them for removal from future copies you may generate for patron use, or you can decide to remove them from the accession. If you choose to permanently delete duplicate files from your holdings, you will need to delete them from the Master Folder you have already moved to secure storage. Document any action you take by updating the accession record stored with it. Repeat the "Securing Project Files**"** steps (p. 17) if you decide to separate any files from your master copies.

### *Tools and Software*

- **Beyond Compare:** http://www.scootersoftware.com/
- **DROID:** http://www.nationalarchives.gov.uk/information-management/our-services/dc-file-profiling-tool.htm
- **fslint:** http://www.pixelbeat.org/fslint/
- **CloneSpy:** http://www.clonespy.com/
- **Araxis Find Duplicate Files (Mac OS X):** http://www.araxis.com/find-duplicate-files/
- **sdhash:** http://roussev.net/sdhash/sdhash.html

### *Resources*


- "Verification of Duplicate Electronic Records." 2012. *Google Groups Digital Curation Forum*. Last modified 18 December. https://groups.google.com/forum/?fromgroups=#!topic/digital-curation/apUjQYOJr4E.


## Dealing with Personally Identifying or Sensitive Information

Consider this step if you have files that may contain sensitive or confidential information such as social security numbers, financial information, research data with names of participants, or other confidential information that must be restricted or redacted. Ideally, the existence of such information was determined at the time the media was acquired. Be aware that sensitive information must be kept restricted and secure on your workstations, file servers, and any backup or transfer copies.

If you do have such information, you may need to redact or anonymize it before making it available to users.

### *Tools and Software*

- **Identity Finder:** http://www.identityfinder.com
- **The Forensic Toolkit (FTK) by Access Data:** http://www.accessdata.com
  This is expensive commercial software that is related to the free software FTK Imager.
- **BulkExtractor:** http://www.forensicswiki.org/wiki/Bulk_extractor
- **EnCase by Guidance Software:** http://www.guidancesoftware.com
- **Firefly:** http://www.cites.illinois.edu/ssnprogram/firefly/index.html
  This is a free tool created by the University of Illinois Urbana-Champaign to find social security numbers.

### *Resources*


- Cook, Timothy. 2012. "A Regular Expression Search Primer for Forensic Analysts." *SANS InfoSec Reading Room—Forensics*. http://www.sans.org/reading_room/whitepapers/forensics/regular-expression-search-primer-forensic-analysts_33929.
- Lee, Christopher A. and Kam Woods. 2012. "Automated Redaction of Private and Personal data in Collections: Toward Responsible Stewardship of Digital Heritage." *Memory of the World in the Digital Age: Digitization and Preservation Conference Proceedings September 26-28, 2012.* Edited by Luciana Duranti and Elizabeth Shaffer. Vancouver, British Columbia: Canada. http://www.ils.unc.edu/callee/p298-lee.pdf.
- University of Essex. 2013. "Anonymisation / Overview." *UK Data Archive: Consent and Ethics.* http://www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation.

## Update Associated Collection Information

**Level of Difficulty:** Easy
**Desirability:** Highly Recommended

After exploring your data and making final decisions about which files to retain, you are ready to link this project with the rest of your holdings. Create or update an associated finding aid, collection-level description, or accession record with information about the steps that were taken. Using the metadata you generated, describe the extent (number of files and total size in gigabytes), nature, and location of the digital files in storage.

### *Tools and Software*

See the "Transfer to a Secure Location" section (p. 19) for more information on these tools:

- **BitCurator:** http://www.bitcurator.net/
  This tool includes several customized scripts that generate summaries of disk contents, which can be exported as plain text, XML, Microsoft Excel, or PDF.

- **Karen's Directory Printer:** http://www.karenware.com/powertools/ptdirprn.asp
  This is a list of file contents that can be used to summarize files.

### *Resources*

- AIMS Work Group. 2012. *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship.* Charlottesville, VA: University of Virginia Library. http://www.digitalcurationservices.org/files/2013/02/AIMS_final.pdf.
  See "Chapter 3: Arrangement and Description" and "Chapter 4: Discovery and Access" for further steps in this process.

- SAA (Society of American Archivists). 2013. *Trends in Archive Practice: Archival Arrangement and Description*. Edited by Christopher J. Prom and Thomas J. Frusciano. Chicago, Illinois: Society of American Archivists. http://www2.archivists.org/node/17121.

# Sample Workflows

It can be helpful to look at the workflows adopted by other institutions to help plan your own. Here are some pointers:

- AIMS Work Group. 2012. *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship.* Charlottesville, VA: University of Virginia Library. http://www.digitalcurationservices.org/files/2013/02/AIMS_final.pdf. See "Appendix E: Sample Processing Plans."
- Gengenbach, Martin J. 2012. "'The Way we do it Here': Mapping Digital Forensics Workflows in Collecting Institutions." Master's paper, School of Information and Library Science, University of North Carolina at Chapel Hill, August, 2012. http://digitalcurationexchange.org/system/files/gengenbach-forensic-workflows-2012.pdf.
  See "Findings" (p. 27).
- LeFurgy, Bill. 2012. "Steps in a Digital Preservation Workflow." YouTube video, 58:43, from a webinar presented at ALCTS 7 March 2012, posted 22 August. www.youtube.com/watch?v=0A6MVp8GijQ.
- Nelson, Naomi L., Seth Shaw, Nancy Deromedi, Michael Shallcross, Cynthis Ghering, Lisa Schmidt, Michelle Belden, Jackie R. Exposito, Ben Goldman and Tim Pyatt. 2012. *SPEC Kit 329: Managing Born-Digital Special Collections and Archival Materials*. Edited by Lee Anne George. Washington, DC: Association of Research Libraries. http://publications.arl.org/Managing-Born-Digital-Special-Collections-and-Archival-Materials-SPEC-Kit-329.
  See "Workflows" (pp. 154-196).
- Prom, Chris. 2012. "MAC Conference 'Streams in E-Record Workflow'" *Practical E-Records* (blog). April. http://e-records.chrisprom.com/mac-conference-streams-in-e-record-workflow/.
- Williams, Laura, Rebecca McNulty Skirvin, Glynn Edwards, and Peter Chan. 2013. "Born-Digital Archives Program: Forensics Workflow Documentation" Stanford University Libraries and Academic Information Services. Accessed 23 April 2013. https://sites.google.com/site/workflowdocumentation.

# Next Steps

You are now ready to repeat the above steps to transfer additional files and continue building your born-digital holdings. Preservation issues go beyond copying files, however. Here are a few of the bigger issues you will need to consider in the near future:

- **Infrastructure:** How will you scale your processes, and what sort of infrastructure will you need to do so?
- **Format Migration:** How will you keep your files safe from format obsolescence?
- **Preservation Lifecycle:** How will you ensure the continued life of your data?

Most importantly, you are now ready to prepare to make the content discoverable and accessible to users.

# Supplementary Exploration

## Discussion Forums

- **Digital Curation Google Group:** https://groups.google.com/forum/?fromgroups#!forum/digital-curation
- **SAA's Electronic Records Section Listserv:** http://www.archivists.org/saagroups/ers/listserv.asp
- **The AIMS Blog:** http://born-digital-archives.blogspot.com
- **Chris Prom's Practical E-Records Blog:** http://e-records.chrisprom.com
- **Digital Curation Exchange:** http://digitalcurationexchange.org
- **FutureArch Blog:** http://futurearchives.blogspot.com
- **ACRL Digital Curation Interest Group:** http://www.ala.org/acrl/aboutacrl/directoryofleadership/interestgroups/acr-igdc

## Learning Opportunities

- **CURATEcamp:** http://curatecamp.org/
- **DigCCurr Professional Institute:** http://ils.unc.edu/digccurr/institute.html
- **Rare Book School: Born-Digital Materials** (Theory & Practice)**:** http://www.rarebookschool.org/courses/libraries/l95/
- **SAA's Digital Archives Specialist Curriculum:** http://www2.archivists.org/prof-education/das

## Additional Resources

- **SAA: A Glossary of Archival and Records Terminology:** http://www2.archivists.org/glossary
- **Digital Curation Bibliography:** http://digital-scholarship.org/dcbw/dcb.htm
- **Digital Curation Center:** http://www.dcc.ac.uk/
- **Paradigm Workbook:** http://www.paradigm.ac.uk/workbook/
- **Open Planets Foundation**: http://www.openplanetsfoundation.org
- **Plato:** www.ifs.tuwien.ac.at/dp/plato/intro.html
- **Closing the Digital Curation Gap (CDCG):** http://www.dcc.ac.uk/projects/closing-digital-curation-gap
- Preserving (Digital) Objects with Restricted Resources (POWRR): Digital Preservation Tool Grid: http://digitalpowrr.niu.edu/tool-grid/

# References Cited


- AIMS Work Group. 2012. *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship.* Charlottesville, VA: University of Virginia Library. http://www.digitalcurationservices.org/files/2013/02/AIMS_final.pdf.
- Erway, Ricky. 2012. *You've Got to Walk Before You Can Run: First Steps for Managing Born-Digital Content Received on Physical Media.* Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2012/2012-06.pdf.

For more information about our work regarding born digital materials, please visit:
**http://www.oclc.org/research/publications/library/born-digital-reports.html**

6565 Kilgour Place
Dublin, Ohio 43017-3395

T: 1-800-848-5878
T: +1-614-764-6000
F: +1-614-764-6096
www.oclc.org/research

ISBN: 1-55653-454-X
978-1-55653-454-6
1403/215339, OCLC