

Resources

Links to BitCurator blog, forums and user groups, scripts library, and user and technical documentation.

- [BitCurator-Related Resources](#)
- [Forums and User Groups](#)
- [User and Technical Documentation](#)
- [Datasets Library](#)
 - [General Resources about Digital Forensics Sample Data](#)
 - [Digital Corpora](#)
 - [CD-ROM and Floppy Disk Library – Indiana University](#)
 - [Preservation and Access Virtual Education Laboratory \(PAVEL\)](#)
 - [Corpora Designed for Targeted System Testing](#)
 - [Open Preservation Foundation](#)
 - [National Institute of Standards and Technology](#)
 - [Library of Congress Labs Web Archive Data Sets](#)
- [Scripts Library](#)
- [OSSArcFlow Resources](#)

BitCurator-Related Resources

- News, updates, and BitCurator blog posts: <https://bitcurator.net/>
- [BitCurator User Group](#): Get support and discuss issues with the community.
- [Screencasts and Video Tutorials](#): Useful screencasts on our YouTube channel.

Forums and User Groups

- Ubuntu Forums: <https://ubuntuforums.org/forumdisplay.php?f=48>
- Ask Ubuntu Help Center: <http://askubuntu.com/help>
- Oracle Virtual Box Forum: <https://forums.virtualbox.org/>
- Bulk Extractor Google Group: https://groups.google.com/forum/?hl=en#!forum/bulk_extractor-users
- The Sleuth Kit Forum: <http://forum.sleuthkit.org/>
- The Sleuth Kit User Discussion List: <https://lists.sourceforge.net/lists/listinfo/sleuthkit-users>
- FITS Discussion Group: <https://groups.google.com/forum/#!forum/fits-users>

User and Technical Documentation

- Official Ubuntu Documentation: <https://help.ubuntu.com/>
- Oracle Virtual Box Documentation: <https://www.virtualbox.org/wiki/Documentation>
- ClamAV Manual: <https://www.clamav.net/documents/clam-antivirus-user-manual>
- Bulk Extractor Forensic Wiki: https://forensicswiki.xyz/wiki/index.php?title=Bulk_extractor
- Bulk Extractor User Manual: http://digitalcorpora.org/downloads/bulk_extractor/BEUsersManual.pdf
- FSLint User Manual: <https://booki.flossmanuals.net/fslint/>
- Guymager Wiki: <https://sourceforge.net/p/guymager/wiki/Home/>
- The Sleuth Kit Wiki: http://wiki.sleuthkit.org/index.php?title=The_Sleuth_Kit
- FITS Documentation: <http://projects.iq.harvard.edu/fits/documentation>

Datasets Library

The [BitCuratorEdu](#) project team is building a library of datasets for educational, testing, and research purposes.

This document summarizes known available sample data. If you know of other examples that are already publicly available or how your own examples that you would be willing to share, please let the [BitCuratorEdu project team](#) know.

The team will continue to build this content as we gather it over the 2019-2021 project period.

General Resources about Digital Forensics Sample Data

A comprehensive review of available datasets for cyber forensics research was presented at the 2017 Digital Forensics Research Workshop.

- See original article at <https://www.sciencedirect.com/science/article/pii/S1742287617301913>
- 81 datasets listed at: <https://datasets.fbreitinger.de/datasets/>

Forensics Wiki list of forensic corpora:

- https://forensicswiki.xyz/wiki/index.php?title=Forensic_corpora

Digital Corpora

- Nitroba University Harassment Scenario: <https://digitalcorpora.org/corpora/scenarios/nitroba-university-harassment-scenario>
- M57-Patents Scenario: <https://digitalcorpora.org/corpora/scenarios/m57-patents-scenario>
- M57-Jean Scenario: <https://digitalcorpora.org/corpora/scenarios/m57-jean>
- National Gallery DC Scenario: <https://digitalcorpora.org/corpora/scenarios/national-gallery-dc-2012-attack>

- Lone Wolf Scenario: <https://digitalcorpora.org/corpora/scenarios/2018-lone-wolf-scenario>
- Govdocs1 set: <https://digitalcorpora.org/corpora/files>
- NPS-2010 emails: <https://digitalcorpora.org/corpora/disk-images/nps-2010-emails>
- Real Data Corpus (restricted access): <https://digitalcorpora.org/corpora/disk-images/real-data-corpus>

This is arguably the most directly applicable and widely used source for sample data in digital forensics education. Simson Garfinkel and his collaborators have developed several realistic corpora for digital forensics education and research, available at <http://digitalcorpora.org>.

These include “scenarios,” which represent fictional but realistic events. For example, UNC SILS frequently uses the M57-Patents Scenario for classes and a variety of continuing education offerings, including conference workshops and the Digital Archives Specialist (DAS) digital forensics courses offered through the Society of American Archivists. The full hard drive images are of a manageable size for longer assignments and exercises that require a drive with a full operating system; the USB flash drive images are smaller and well-suited for short workshops, class exercises and demonstrations.

CD-ROM and Floppy Disk Library – Indiana University

Online collection of “nearly 5,000 CD-ROMs, DVDs, and floppy disks distributed by the GPO under the Federal Depository Library Program (FDLP). These tangible products have been received through the FDLP since the 1980s and consist of millions of individual files containing fundamental data on economics, the environment, population, and life and physical sciences.”

https://webapp1.dlib.indiana.edu/virtual_disk_library/

Each disk image has an associated catalog record. For use in classes, one approach is to share a subset of the images directly with students, so they don’t immediately encounter all of the descriptive metadata from the site.

Preservation and Access Virtual Education Laboratory (PAVEL)

The Preservation and Access Virtual Education Lab (PAVEL) project at the University of Michigan School of Information (2010-2012), funded by the National Endowment for the Humanities, developed a “virtual education laboratory featuring digital access and preservation tools.” This includes four data sets:

- a small set of six [files for testing preservation tools](#) (ZIP file, 6.1 MB)
- [Elena Kagan's email](#) (approximately 19,000 messages) (ZIP file, 47.9 MB) from her tenure in the White House, along with an [organizational chart](#) (PDF file, 161 KB) and [staff list](#) (PDF file, 119 KB)
- the [Enron email collection](#) (approximately 600,000 messages) (ZIP file, 112.4 MB), along with [organizational charts](#) (PDF file, 214 KB)
- a [set of documents](#) (ZIP file, 182.9 MB) from the University of Michigan College of Literature, Science, and the Arts IT department (approximately 450 megabytes of Microsoft Office files) and document providing [institutional context](#) (PDF file, 15 KB) information

Many thanks to David Wallace and Beth Yakel from the PAVEL team for allowing us to redistribute the datasets.

PAVEL data sets have the advantage of being selected for use in archival education, and they contain important supplementary information (e.g. organizational charts) for understanding their contexts of creation. However, they do not allow for the replication of many forensic tasks. For example, all of the email messages are stored as separate ASCII text files, rather than being embedded in a disk image or within a wrapper format such as PST, a scenario likely to be encountered during acquisition of a new collection.

Corpora Designed for Targeted System Testing

Brian Carrier’s valuable, but highly targeted Digital Forensics Tool Testing Images:

- <http://dfdt.sourceforge.net/>

Computer Forensic Reference Data Sets (CFReDS) from the National Institute for Standards and Technology (NIST):

- <http://www.cfreds.nist.gov/>

Open Preservation Foundation

The Open Preservation Foundation maintains a site that includes “Datasets, preservation and curation Issues with those Datasets, and Solutions to those Issues.” The experiences of solving specific Issues are written up on Solution pages, which then link to pages in the OPF Tool Registry. In many cases, this leads to “actual code that can be downloaded and re-used.”

- <http://wiki.opf-labs.org/display/REQ/Digital+Preservation+and+Data+Curation+Requirements+and+Solutions>

National Institute of Standards and Technology

The National Institute of Standards and Technology creates and maintains data sets including disk images for testing forensic tools. Some of these are deprecated over time; the current list can be found at:

- <https://www.nist.gov/itl/ssd/software-quality-group/computer-forensics-tool-testing-program-cftt/cfreds/cfreds-0>

Library of Congress Labs Web Archive Data Sets

The 1000 .gov PDF Datasets provided by LC are a particularly useful dataset for exploring BitCurator functionality and can be applied to educational contexts.

- <https://labs.loc.gov/work/experiments/webarchive-datasets/>

Scripts Library

This is a list of scripts created by BitCurator users that can assist with digital forensics tasks and activities. If you have a script you would like added to the library, please [contact us](#) or post in our [Google Group](#).

- [Guymager Log Parser](#), Euan Cochrane:
Script to parse and process Guymager logs in bulk
- [KryoFlux Disk Format ID](#), Euan Cochrane
This program is intended to help to identify floppy disk formats. It takes a folder of KryoFlux stream files as input
Relevant Resource: [Floppy Disk Format Identifier Tool](#) (blog post)
- [hfs2dfxml](#), Dianne Dietrich
Utility to parse `hfsutils` output and produce DFXML for HFS-formatted disk images
Relevant Resource: [Advanced Topics - Scripting in BitCurator](#) (webinar, slides)
- [populate_did.sh](#) (shell script, 844 bytes), Jarrett Drake
What are the key metadata points to extract from complex, multi-level born-digital records and later represent in EAD? This shell script extracts information from directories of digital content for use in the Descriptive Identification element wrapper.
Relevant Resource: [Advanced Topics - Scripting in BitCurator](#) (webinar, slides)
- [floppy-utils](#), John Durno
Sample scripts for acquiring and processing floppy disk images
Relevant Resource: [Digital Archaeology and/or Forensics: Working with Floppy Disks from the 1980s](#) (paper)
- [rl-bitcurator-scripts](#), Matthew Farrell, Kam Woods
Scripts (bash, Python) developed to automate processes
- [disk-image-timeline](#), Walker Sampson
Generate file system event timelines from a collection of disk images automatically, and collects the individual timelines into a master CSV file
Relevant Resource: [Aggregating Temporal Forensic Data Across Archival Digital Media](#) (paper)

OSSArcFlow Resources

The [OSSArcFlow](#) project (2017-2020) was supported by a grant from the Institute of Museum and Library Services. OSSArcFlow project personnel worked with partner institutions to document and analyze born-digital workflows, and created an implementation guide and videos to support workflow documentation and analysis.

- [Digital Dossiers](#): Project partners created digital dossiers outlining the form, function, and future of digital curation at their home institutions.
- [As-Is Workflows](#): In the fall of 2017, the project team worked with partners at each institution to mockup a visual representation of their current workflow activities. Representing a "snapshot in time," these documents show how a diverse group of institutions are using OSS tools in their workflows to curate born-digital content. They also provide an essential starting point for synthesizing and comparing both the gaps and overlaps that currently exist between common OSS tools and environments.
- [Guide to Documenting Born-Digital Workflows](#): The purpose of the *Guide to Documenting Born-Digital Archival Workflows* is to encourage and assist collecting institutions of all shapes, sizes, and types to begin documenting their born-digital workflows. The *Guide* includes four main sections:
 1. **Introduction** – provides a brief background of the project, the research questions that have driven our inquiry, and how to use this *Guide* in your own work as an archivist and curator of born-digital collections.
 2. **Common Steps in OSS Born-Digital Archival Workflows** – provides brief descriptions of each of the main steps in born-digital archiving (13 in total) and what tools are commonly used to accomplish each of these steps today.
 3. **Documenting Born-Digital Workflows** – provides detailed guidance to help you use the OSSArcFlow survey, interview questions, and visualization model to document and depict your own workflow.
 4. **Using Workflows** – provides guidance and examples of how an institution can use its existing workflows to identify growth/maturity goals, to advocate for resources, and to transform roles/relationships as needed to improve its born-digital archiving practices.

If you would like to provide feedback for this page, please follow this [link to the BitCurator Wiki Google Form](#) for the Resources section.