Datasets Library

This page was moved to https://confluence.educopia.org/display/BC/Resources
Click in the link above if you are not automatically redirected in 10 seconds.

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://www.forensicswiki.org/wiki/Forensi_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/national-gallery-dc-2012-attack
- 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
- Elena Kagan's email (approximately 19,000 messages) from her tenure in the White House, along with an organizational chart and staff list
- the Enron email collection (approximately 600,000 messages), along with organizational charts
- a set of documents 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 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://dftt.sourceforge.net/

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

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

Forensic Focus Test Images and Forensics Challenges:

- http://www.forensicfocus.com/images-and-challenges

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/node/1303796/cfreds-current-data-sets

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/experiments/webarchive-datasets/?loclr=twndi