Digital corpora
#OSDFCON
2:20pm - 2:35pm
October 16, 2019
Herdon, VA

DIGITAL CORPORA – SHORT TALK

https://www.digitalcorpora.org/

In collaboration with:

Simson L. Garfinkel

simsong@acm.org

The views in this presentation are those of the author, and not those of the US Census Bureau, the Department of Commerce, the US Navy, the US Department of Defense, or the United States Government.

GEORGE MASON UNIVERSITY

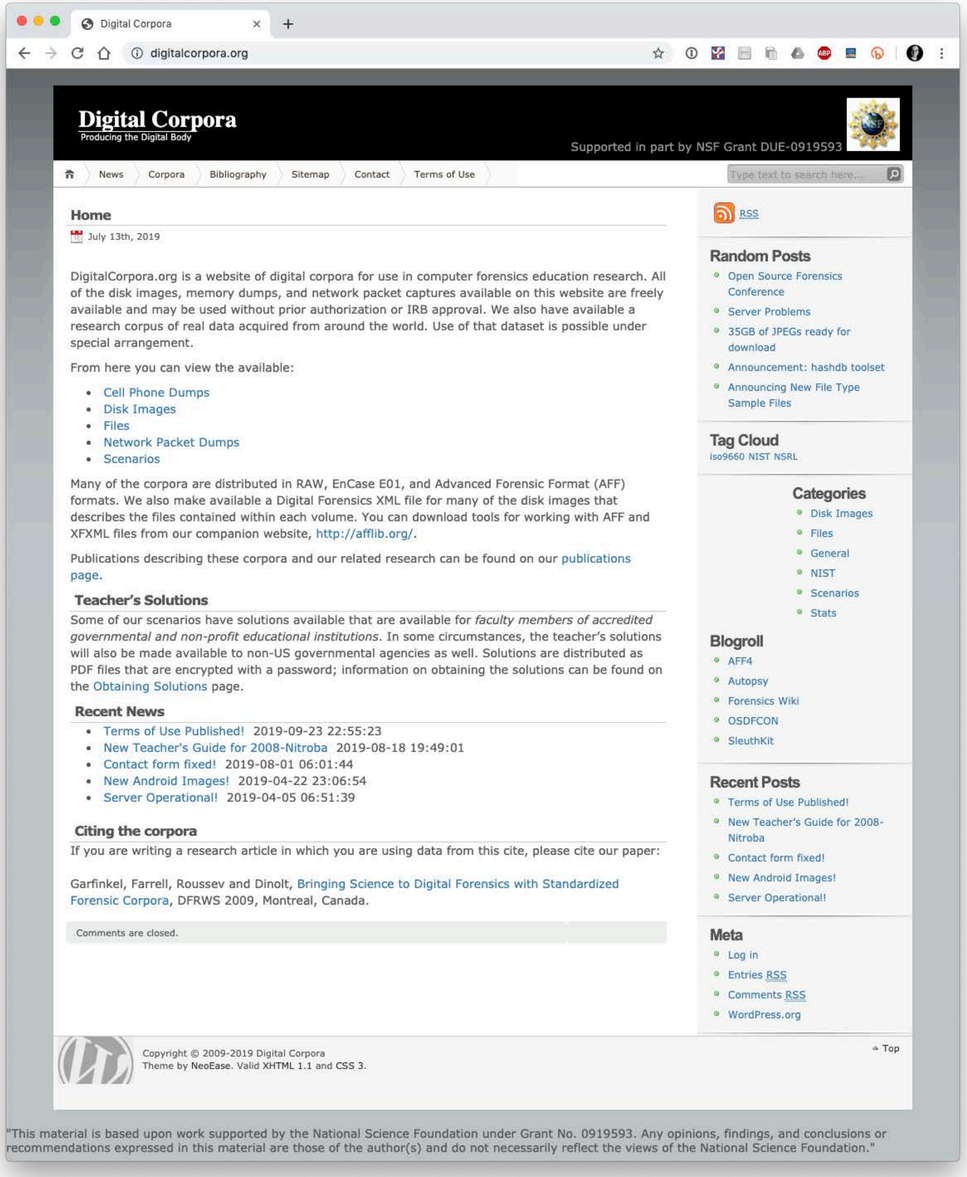
Digital corpora:
complex digital artifacts for digital forensics education and tool testing.

https://digitalcorpora.org/

Originally developed under NSF Grant No. 0919593

Originally developed at Naval Postgraduate School

Significant growth in recent years.



Scenario-based digital corpora

Complex, deep datasets.

- **Scripted scenario.**
- **Multiple characters with clearly defined motivations**
- **Specific challenges for the investigator to uncover.**
- **Multiple problems requiring different levels of skill and analysis to solve.**
- **Created over weeks or months**

Multi-modality:

- **Disk images**
- **Cell phone images**
- **Memory dumps**
- **Log files from servers**
- **Packet dumps (wiretaps)**

Day-by-day captures:

- **Useful for forensics research and tool development**
- **Not present for all scenarios**

There are many advantages to scenario-based artifacts.

No privacy-sensitive data! No PII!
- **Computer users are not real people, they are personas.**

No pornography!
- **We know that there's no pornography in the data.**
- **Especially an issue with students under 18 years old!**

No illegal content!

There are solutions!
- **Solutions are distributed on the website as encrypted PDFs.**
- **Decrypt keys available on a case-by-case basis to faculty at accredited institutions, law enforcement, and partners.**

Scenarios are distributed from the download server https://downloads.digitalcorpora.org/corpora/

```
92M  2008-nitroba/
412G 2009-m57-patents/
35G  2011-nps-1weapondeletion/
20G  2011-nps-2weapons/
19G  2011-nps-4drugtraffic/
21G  2011-nps-5control/
112G 2012-ngdc/
80G  2018-lonewolf/
128G 2019-narcos/
223G 2019-owl/
```

Packets: PCAP files

It's really hard to get full-content PCAP files

```
92M  2008-nitroba/
3.9G 2009-m57-patents/
51G  2012-defcon/
51M  2013-httpxfer/
795M  5gb-tcp-connection.pcap.gz
```

We are looking for more!

- **Especially TLS connections:**
  —for which you have the private keys;
  —where perfect-forward-secrecy is disabled;
  —where you have escrowed the master secret.

Especially Internet-of-Things!

Phones

We have a (very) small number of phone and tablet images.

- **2011-android — NexusOne (2 images)**
- **android_7**
- **android_8**
- **android_9**
- **apk.zip — 2132 pieces of circa 2012 android “blackdroid” apps**
- **Nokia_6230**
- **SE_P800**
- **SE_T630**
- **SE_T68i**

This is only 20G of data (compressed)

Please consider making a donation!

Drives — disk images for forensic tool testing

```
9.9G   m57-patents/
3.0G   nps-2008-ipod0/
2.9G   nps-2008-m57-jean/
59M    nps-2008-nano0/
95G    nps-2008-seed1/
203M   nps-2009-canon2/
295M   nps-2009-casper-rw/
8.2G   nps-2009-domexusers/
33M    nps-2009-edu-corrupt1/
21M    nps-2009-hfsjtest1/
27G    nps-2009-ipod1/
27G    nps-2009-ipod160/
424G   nps-2009-m57-patents-redacted/
55M    nps-2009-ntfs1/
471G   nps-2009-patents/
10G    nps-2009-ubnist1/
1.1M   nps-2010-emails/
395G   nps-2011-2tb/
12M    nps-2013-canon1/
2.0G   nps-2014-usb-non-deterministic/
13G    nps-2014-xbox1/
```

Files — 1M document corpus

Developed at NPS, this corpus includes 1 million files downloaded from US Government webservers.

- **US Government websites to avoid copyright issue.**

Has been used in several research efforts.

Includes:

- **JPEGs**
- **PDFs**
- **Microsoft Office**
- **Text files**
- **Log files**
- **SQL dumps**
- **Lots of random stuff**

Note: No longer a million files; we had some takedown requests.

Current Status

Currently developing a macOS/iOS deep image with multiple persona and international travel

We are taking donations!

Building a corpus is a great student project!

- **It's a lot more work than it seems!**

Thanks to George Mason University for hosting!

- **Currently 3.4T**