Availability of datasets for digital forensics e And what is missing

Name Nam Gihoon

Place 319

2017-11-20

Introduction

- In order to produce high-quality research results, we argue that three critical features must be examined

1. Quality of the datasets. - This helps guarantee that results are accurate and generalizable. Researchers need data that is correctly labeled and similar to the real world or originates from the real world.

2. Quantity of the datasets.

- This ensures that there is sufficient data to train and validate approaches/tools which is especially important when utilizing machine learning techniques

3. Availability of data.

- This is critical as it allows the research to commence and ensures reproducible results helping in improving the state of the art.

Introduction

- We contend that is important to have easily accessible datasets

- Penrose et al

“in the scientific method it is important that results be reproducible. An independent researcher should be able to repeat the experiment and achieve the same results. Most research has been done with private or irreproducible corpora generated by random searches on the WWW.”

Introduction

- In this work we analyzed a total of 715 cybersecurity and cyber forensics research articles from the years 2010e2015 from five different conferences/journals with respect to the utilization of datasets.

1. dataset's origin generated
2. Availability
3. Kinds of datasets

Missing

- Our findings illustrate that the majority of available datasetswere experiment generated (over 1/2) and only around 1/3 originated from real world data.
- we show that researchers (re-)use available datasets frequently but when they have to create their own dataset, it is rarely shared with the community (less than 4%).

Limitations

- All of our data analysis was performed by manual inspection. We note that human error might have been introduced, but we attempted to alleviate the errors by conducting multiple runs

Related work

- In their article, the authors analyzed 106 network security papers over four years (2009e2013) and concluded with three main findings

(1) many researchers manually produced their datasets

(2) datasets are often not released after the work is completed and

(3) there is a lack of standardized datasets that are labeled that can be used in research

- These weaknesses combined, produced one of the major disadvantages facing the cybersecurity forensics community to this day, which is low reproducibility, comparability and peer validated research

Methodology

- While this work was influenced by Abt and Baier (2014), the difference between both studies is that we do not exclusively focus on network traffic but on all kinds of datasets that may be useful for cybersecurity/forensics research, e.g., malware, disk images or memory dumps.

- our study expands to a broader number of articles, results from Google searches and provides an overview of existing datasets.

Definition of a dataset

- For this work we define a dataset as a collection of related, discrete items that has different meanings depending on the scenario and was utilized for some kind of experiment or analysis.

Analyzing peer-reviewed articles

- The first phase entailed the collection and analysis of publications from digital forensics and security conference proceedings as well as journal publications3 spanning six years (from 2010 to 2015)

1. Origin of datasets(Is the dataset computer generated, experiment generated or user generated)

2. Availability of datasets (Are datasets available to the community?)

- Was the utilized dataset available prior to the research?(re-usage)

- If the dataset was created, was it released? (availability)

- If the dataset was available prior to the research, is the origin disclosed/is it freely available? (proprietary to one ‘group’)

Analyzing peer-reviewed articles

3. Kinds of datasets (What datasets exist and can be used by researchers?

- Were any third party databases, services or online tools used in the creation of datasets?

4. What is missing (What datasets or other things are currently missing? This will be addressed in Sec. What is missing.)

Results overview and origin

- Due to the significantly higher adoption of datasets in the digital forensics domain, the remaining analysis focused on conferences/journals that embodied digital forensics as a main thematic topic.

(i) International Conference on Digital Forensics & Cyber Crime (ICDF2C) had 60 out of 107 that used datasets;

(ii) Association of Digital Forensics, Security & Law (ADFSL, Conference) contained 29 out of 87 articles that utilized datasets;

(iii) Digital Investigation (Journal) contained 108 out of 190 articles that employed datasets.

Experiment generated datasets

- Over half of the datasets found in this study were experiment generated, where researchers created specific scenarios to conduct their experiments. There are several reasons for having such a heavy shift towards this kind of data.

1. in many cases, there is a lack of real world datasets available to the digital forensics community

2. Another reason is that using experiment generated data allows researchers to test and verify such data, especially when conducting experiments on new technologies as that is common within the area of cybersecurity and digital forensics

User generated datasets

- With over 36%, user generated datasets (a.k.a. real world datasets) were the second most used type of data.

- According to Baggili and Breitinger (2015), experimenting on real world data is crucial for developing reliable algorithms and tools

- “how can we learn from our past when we do not have real, accessible data to learn from?”

- One of the major reasons is clearly copyright and privacy laws which prohibit sharing with the community (Abt and Baier, 2014). If real world data was used, we found the following different origins:

1. Dataset was released
2. Collaboration with law enforcement:
3. Source of data is online:

Computer generated datasets

- The final category is computer generated datasets or synthetic data which may have several origins, e.g., an algorithm, bots, /dev/ urandom or simulators

- Our analysis revealed that almost 5% of the analyzed articles employ those datasets which is not necessarily a surprise

- often researchers in digital forensics want to solve real world problems and therefore cannot use simulated or generated data. One argument for generated data is the exact knowledge of the ground truth

Usage of third party databases, services or online tools

- In our research, we realized that about 20.4% (39/191) articles used third party databases, services or online tools to retrieve information.

Availability of datasets

Creating vs. re-using datasets

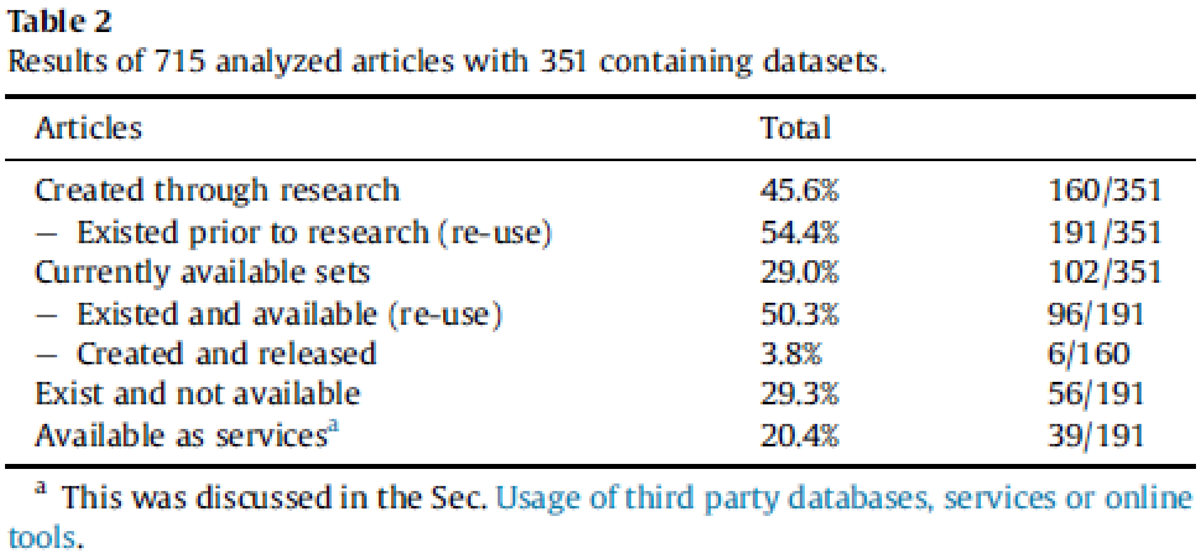


Table 2
Results of 715 analyzed articles with 351 containing datasets.

Articles	Total	
Created through research	45.6%	160/351
– Existed prior to research (re-use)	54.4%	191/351
Currently available sets	29.0%	102/351
– Existed and available (re-use)	50.3%	96/191
– Created and released	3.8%	6/160
Exist and not available	29.3%	56/191
Available as services[a]	20.4%	39/191

[a] This was discussed in the Sec. Usage of third party databases, services or online tools.

- The first row in Table 2 provides an overall summary and indicates that 45.6% of the articles analyzed produced their own datasets in their experiments while 54.4% of the articles utilized datasets that existed (re-use of an existing set).

Availability of datasets

Creating vs. re-using datasets

- This almost equalshare seems reasonable as researchers often train algorithms based on simulated/experiment data while on the other hand for evaluating performance/comparing two algorithms often real world datasets are favored, e.g
- Coming to the high usage of self-made datasets, some researches clearly stated they were required to create their own dataset since nothing was available
- This indicates that researchers re-use datasets if they are available and do not necessarily favor building their own.

Currently available datasets

- The current availability is discussed in the second row of Table 2 e only 29.0% (102) of all sets are available for research and thus allow reproducible results. The vast majority (96) of the sets already existed where on the other hand only 3.8% of the newly created ones were released. Examining the origin of currently available sets revealed, that 59.8% (61/102) employed real world datasets. Subsequently, 38.2% of available datasets were recognized as experiment generated and 2.0% as computer generated datasets.

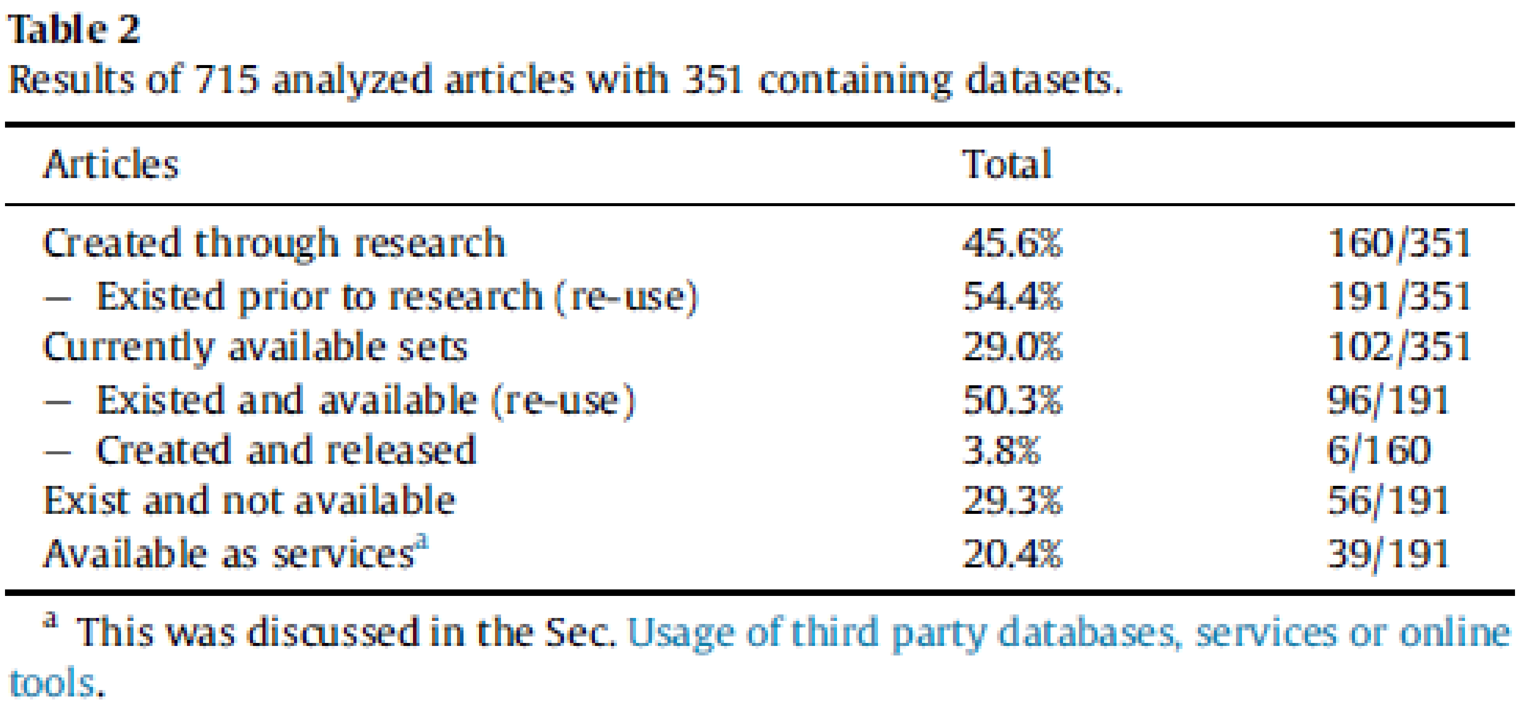


Table 2
Results of 715 analyzed articles with 351 containing datasets.

Articles	Total	
Created through research	45.6%	160/351
– Existed prior to research (re-use)	54.4%	191/351
Currently available sets	29.0%	102/351
– Existed and available (re-use)	50.3%	96/191
– Created and released	3.8%	6/160
Exist and not available	29.3%	56/191
Available as services[a]	20.4%	39/191

[a] This was discussed in the Sec. Usage of third party databases, services or online tools.

Non available datasets

- This section focuses on datasets that exist but were not available. Specifically, we discovered 29.3% (56/191) articles with datasets that we were unable to verify and classify as currently available. We organized this set of articles into three groups:

1. Source is unknown:
2. Source has privacy restrictions:
3. Source not accessible:

Non available datasets

1. Source is unknown:

- major problem because not knowing the source of the datasets may raise questions about the quality and integrity of such data.

- it completely hinders researchers from reproducing experimental results

2. Source has privacy restrictions:

- these were mostly real world datasets generated by Universities, Government agencies and law enforcement and could not be released.

3. Source not accessible: - About 1/7 of the articles had accessibility problems, such as temporarily unavailable, download link broken or not maintained anymore.

Kinds of datasets

- we found over 70 different datasets though our article analysis and organized them in 21 categories with major ones discussed in the following subsections.

- Each subsection will provide references/links to the available datasets, and provide a brief overview, e.g., origin, amount of samples, total size, etc. (when obtainable).

Malware datasets (computer and mobile)

Android. In total, three repositories were frequently used.

(1) Drebin (Arp et al., 2014) is a collection of 5560 Android samples from 179 different malware families collected between 2010 and 2012 and was used by Talha et al. (2015) to test permission based malware detection.

(2) Contagio Mobile Mini-Dump (A.4.7.1) is part of the larger computer malware repository Contagio Malware Dump. In contrast to other repositories, this website is more like a traditional blog with an upload/download functionality. Thus, users can download the repository but also extend it. According to the website, there are over 200 malware posts and each post might contain more than one malware sample, collected from 2011 to 2016. Lastly,

(3) Jang et al. (2015) possess a dataset (A.4.7.2) of 9990 malware samples which can be requested for research purposes. Part of this dataset included samples from the repository Contagio Mobile Mini-Dump and Virus Share (A.3.2.2) (exact amount not mentioned in article).

Malware datasets (computer and mobile)

- Computer malware. In total, four repositories were utilized in the analyzed articles:

(1) Contagio Malware Dump is similar to its counterparts and has around 400 posts.

(2) VX Heaven (A.3.2.3) which is a virus information website that contains over 271,000 computer malware samples. However, it is unknown how often thewebsite is updated and as thewebsite states, the last time the malware collection was scanned was by Kaspersky Anti-Virus in 2006.

(3) Virus Share which was the most comprehensive malware collection that was referenced with over 27 million samples. Although not stated, it seems that this repository is a mix of mobile and computer malware. Additionally, it is one of the most updated sites with new entries every month. Consequently, this malware site is one of the most secure in relation to the acquisition of malware since access to the site is by invitation only. If access is needed an e-mail is required to be sent to the admin stating reasons to be added. Lastly

(4), the forumKernelMode.info (A.3.2.4) was mentioned by Al-Shaheri et al. (2013). According to the post dates which range from 2010 to 2016, this forum seems still active but registration is required. Unfortunately, the amount of malware samples in this forum is unverifiable but it seems to have a mix of mobile and computer malware as well.

E-mail datasets

- Besides that, Armknecht and Dewald (2015) used about 75,724 real world e-mails from the Apache online e-mail repository which was never intended to be a dataset but provides real world examples. Lastly, we found about 12 e-mails in Digital Corpora's experiment

File sets/collections

- File sets are collections of files with various types like text, html, pdf, doc, ppt, jpg, xls, gif, zip or csv. They are frequently used for different purposes (e.g., to test/improve forensic file formats like AFF4 (Schatz, 2015)).

- The most prominent and comprehensive dataset may be the GovDocs1 corpus from Digital Corpora which consists of ~1 million documents gathered by crawling the .gov domain. Given that massive size, a common subset is the t5-corpus which was created by Roussev (2011) and contains 4457 files of various types and is commonly used for testing approximate matching, e.g., by Breitinger and Roussev (2014). Lastly, Roussev and Quates (2013) also created the msx-13 corpus which contains 22,000 MS Office 2007 user generated random files (e.g., docx, xlsx, pptx) crawled from the Internet.

RAM dumps

- Our study found six repositories having over 90 dumps where
- all of them were experiment generated (obviously RAM cannot be fully controlled and therefore it can be considered as a mixture of user and experiment data). The first set was published by Minnaard (2014) where the authors acquired their own RAM data from different operating systems and devices. The authors state the complete RAM archive is available on request, but a sample with over 1 GB of data can be downloaded (A.4.9.1). A second set consisting of five 1 GB RAM dumps (Windows, 2000, 2003, Vista Beta 2, and XP) is provided by the CFReDS Project (A.4.9.3). According to the website, the "systems were not engaged in any malicious or even network based activity at the time of imaging." Two more dumps of WinXP 32-bit machines were released by the DFRWS' forensic challenge (A.4.9.2). Another experiment generated dataset which was used by Case and Richard (2015) originates from The Art of Memory Forensics book (Ligh et al., 2014) and can be downloaded from the corresponding website (A.4.9.4). This single dump has a size of 3.8 GB. Lastly and the most comprehensive collection of memory dumps with 88 samples and a total size of over 44 GB can be downloaded from Digital Corpora (A.4.9.6).

Images of computer drives

- Especially in digital forensics, complete disk images are valuable to create and test tools as well as procedures. Leading theway is the Real Data Corpus (RDC) from Digital Corpora which according to their website11 “is a collection of raw data extracted from datacarrying devices that were purchased on the secondary market around the world.” As of 2011, the non-U.S corpus contained 1289 hard drive images ranging in size from 500 MB to 80 GB. According to Garfinkel et al. (2009) there is also a U.S RDC which contains 1228 hard disk images, however, we could not locate it on the website nor does it say anything about it at the time of writing. A second but way smaller set is provided by the CFReDS Project (A.5.15.3) which contains three images extracted with different imaging tools (Encase, iLook, & Compressed dd). The original image was made with 5 partitions (OS Extended Journaling, OS Extended, another OS Extended, OS Standard & UNIX File System) created on a MAC OS X. According to the website, the purpose of having images extracted from 3 different tools was to test if those tools would recognize the file systems created on the Mac OS X.

Images of other devices

Cell Phones:

- In total, we found 26 images within the two repositories CFReDS (A.4.11.1) and Digital Corpora (A.4.11.2). The former one contains 14 images; 7 from a Nexus One and 7 from a Nexus S-1 while the latter one has 12 images from Black Berry Torch 9800, HTC One V, iPhone 3GS and the Nokia 6102i. Gaming systems: Although there are a variety of consoles out there which get analyzed, we only identified 2 sets with Xbox images. The first one 3.1.1 was released by Moore et al. (2014) and according to them it was released so the “forensic community may expand upon our work”. The second one 3.1.2 came through the nps-2014 XBox-1 scenario comprising of 4 disks; 2 originals and 2 modified by experiments. No other game console image was found.

Images of other devices

SIM card:

- SIM card images were not utilized in any article, nonetheless, we discovered at least 3 images in the CFReDS (A.4.14). Apple iPod & Tablet: Although not utilized in any of the articles, Digital Corpora offers a total of 10 iPod disk images (A.5.18) and 25 disk images of various tablets (A.5.19) (brands not disclosed). Flash Drives: As far as real world flash drive images go, Digital Corpora offers a total of 643 flash images (e.g., USB, Memory Stick, SD and other), with sizes from 128 MB to 4 GB with real world data. Furthermore, it offers the nps-2009-canon2 (A.5.16) and nps-2013-canon1 sets which is a collection of 7 images of 32 MB SD cards which were used by Lambertz et al. (2013) & Garfinkel et al. (2010) for testing image/picture carving tools.

Network traffic

- This section summarizes a variety of different network traffic sources which include PCAP files acquired through tools such as Wireshark or logs (i.e., port and protocol data, IP and operating systems source information and so on). The following datasets were found through our study: The first set was generated for the DFRWS 2009 forensic challenge (A.4.12.2) and thus contains experiment generated PCAP files where most of the traffic is HTTP traffic on port 80. A second shared PCAP dump (A.4.12.3) was created by Karpisek et al. (2015). The dataset was compiled by the researchers for the purpose of acquiring WhatsApp traces that they were able to decrypt. The dataset is comprised of 3 PCAP files containing WhatsApp register and call traffic. A wireless network repository named CRAWDAD was discovered in our study (A.4.13) from which datasets of mobility traces of taxi cabs in San Francisco were acquired. This website also contains hundreds of other types of wireless network traffic (e.g., TCP traces, Bluetooth, accelerometer, 802.11p packets, etc.) released since 2002.

Scenarios/cases for analysis

- We identified three scenarios or cases for analysis. The first one is the nps-2009-domexusers on Digital Corpora which is a disk image of two users (domexuser1 and domexuser2) who communicate with a third user (domexuser3) via IM and e-mail. The disk image is of a Windows XP SP3 system (NTFS format and used twice in our study). The second comprehensive scenario is the 2009- m57-patents created by Woods et al. (2011) for digital forensics and security educational purposes. According to the website, the “scenario tracks the first four weeks of corporate history of the M57 Patents company”. It consists of redacted drive images, USB drive images, RAM Images, network traffic and documentation. While this scenario was originally designed for education purposes, it was also utilized by Garfinkel and McCarrin (2015)'s experiment where it served as sample input to test hash carving techniques. The last scenario consists of three network log traces plus a USB device image from the CFReDS Rhino Hunt scenario. Additionally, this source comes with a answers.pdf which allows to fully understand the scenario.

Mixed and others

- Pictures: Besides finding a great amount of real pictures, we also found computer generated graphics and forged images tainted with steganography. Some of these datasets come from websites such as ‘Break our Steganography System’ (BOSS, A.3.3.1), which hosts a challenge that contains a testing database of 1000 512 512 pgm greyscale images and a training database of 9074 cover images.
- Language corpus (text): Language corpora are often used for Statistical Machine Translation. A common collection is the European Parliament Proceedings Parallel Corpus 1996e2011 (A.3.5.6) which contains about 21 European language versions and 60 million words per language.
- Chat logs: The dataset (A.5.20) is comprised of 1100 chat logs from 11,143 chat sessions from a single computer and recorded between 2010 and 2012 using Messenger Plus!.
- Password lists: These sets are commonly used for probabilistic
- password research such as work by Ma et al. (2014). Some comprehensive dictionaries are listed on a security wiki page (A.5.21) and have millions of leaked passwords from websites such as RockYou, Myspace, and Hotmail. According to this website, these datasets are useful “to generate or test password lists”. Note, any type of private information such as name or email is redacted.

Datasets found through Google research

- Security Repo: secrepo.com is a comprehensive list of samples of security related data. As stated on the website, “this is my attempt to keep a somewhat curated list of Security related data I've found, created, or was pointed to”. This source contains about 100 links to datasets or third party references. This includes samples of networking scanning/recon, shell traffic, security incidents, system logs, ssl certs, malware, and more. Note, the following three repositories were only found through this website. Our Google search did not lead us to either of them which shows how cumbersome finding repositories can be.

Datasets found through Google research

- Mid-Atlantic Collegiate Cyber Defense Competition (MACCDC):

netresec.com has PCAP files of three MACCDC competitions from 2010 to 2012 which comes to a total of 59 PCAP files where the 2010 competition was analyzed and summarized by Carlin et al. (2010). Additionally, this website includes links to other websites hosting cyber challenges, malware datasets, networking traffic, etc. The Cyber Systems and Technology Group of MIT Lincoln Laboratory13: According to the website, this is “the first standard corpora for evaluation of computer network intrusion detection systems” which was collected by MIT Lincoln Laboratory. The three datasets (from 1998 to 2000) are composed of file system dumps, pcap files, NT event log audit data, outside TCP dump Data, as well as “the first formal, repeatable, and statistically significant evaluations of intrusion detection systems”. The 1999 evaluation dataset was also analyzed by Mahoney and Chan (2003).

Datasets found through Google research

- The Black Market Archives14:

As its name implies, this data was acquired from Dark Net Markets (DNM) usually hosted in Tor hidden networks. The DNMs operate on selling and buying drugs, guns, and any other type of illegal or government regulated goods. The author of the site claims he collected 1.6 TB of data comprising 89 DNMs from 2013 to 2015; we found 15 papers that have cited the website/dataset. Malware samples15: This personal website lists about 12 links directed at other malware repositories/services like malshare. com or thezoo.morirt.com. The former one is an open source malware repository that permits users to download 1000 samples per day with a requested public API Key (if more samples are necessary, it requires to contact the admin). The second website is a malware repository which aims at collecting all versions of malware available for download directly from the site with no restrictions.

Datasets found through Google research

- PeekaTorrent: peekatorrent.org contains about 3.2 billion hash

values from 2.65 million torrent files totaling 66 GB of compressed data (84 GB raw) and was collected by Neuner et al. (2016). Impact Cyber Trust: Sponsored by the U.S. Department of Homeland Security (DHS) and other technology and cybersecurity organizations, this website hosts a central database of ground truth and synthetic data available for research. The data provided was donated by at least 10 organizations and ranges from 2009 to 2016, some of them include, Georgia Tech, Packet Clearing House, etc. Note, most of the datasets relate to network traffic (e.g., IDS/Firewall, DNS, IP, BGP routing data, etc.).

What is missing?

Our study shows that many researchers prefer not to share their datasets which could be for several reasons.

- First, researchers may not have the capability of sharing the set (e.g., the dataset is too comprehensive and one does not have the online resources available) which could be solved by a centralized, community based repository (see Sec. Centralized repository).

What is missing?

- A second factor may be related to privacy concerns as discussed in Sec. Data de-identification research.

- Thirdly, researchers might simply not have thought of the importance of sharing their data.

‘I probably wouldn't want to share them (at least not in a publicly accessible manner) because when I picked the content off the Internet, I didn't take into consideration that there might be some privacy or copyright issues that may come up’

Additional shortcomings

Variety

- While we found a good amount of sets online, this study also revealed on what is missing in regards to actual datasets.
- group of devices we could not find data for were Smart-TVs. Coming to a world where everything is connected (IoT), there are many more devices we should try to acquire data from, e.g., Unmanned Aerial Vehicle (UAV), streaming devices, such as Roku or Apple TV.

Additional shortcoming

Updates and upgrades

- Having a closer look revealed that there are massive differences in the number of items per dataset, e.g., while there are 27 million malware samples, we only found 26 smartphone images

- A second aspect is the age of the datasets. While some sets like files are timeless (to a certain extend), other require frequent updates and need to be maintained, e.g

Additional shortcoming

Centralized repository

often these repositories are not maintained and become outdated.

- For instance, the Digital Corporawas updated the last time in 2014

Additional shortcoming

Data de-identification research

- One of the main problems impeding datasets from being released is privacy and proprietary concerns.

- If we find ways to un-personalize data by removing, changing or manipulating names, phone numbers, addresses, and other personalized data, datasets could be shared and utilized for research.

Additional shortcoming

Strategies to share complex data

- As we are moving more and more into the cloud (Platform as a Service, Software as a Service), we need strategies on how to share this kind of data among researchers

- specifically focused on the forensics aspect and offered options on how to acquire and share datasets. Nonetheless, none of the articles mentioned offered any datasets acquired through their investigations.

Additional shortcoming

Publisher support

- sharing secondary information (i.e., datasets) is mostly not well supported by publishers.
- A step into the right direction would be to enable sharing data or even force researchers to submit secondary information.

Discussion

- - Our results showthat less than 4% shared their dataset while on the other hand almost 50% make use of existing datasets.

- the lack of sharing datasets, maintenance and availability are major issues.

- Centralized repository, we believe that this could be solved through a centralized and community based repository, e.g., a github for datasets where everyone can share datasets.

- Another challenge is the availability of real world data which is of importance for researchers to produce high quality resultseonly about 1/3 of the datasets originated from real users.

- In order to allow reproducibility, improvements and faster research progress, we believe the mindset of researchers need to change and data should be released.

Conclusion & future work

- While this study comes with a comprehensive list of available datasets and repositories which can be leveraged by researchers, we also show that there is a lack of sharing data which we believe is key to improve the quality and pace of research especially in domains like digital forensics.

- section we highlight six points that we believe are needed in order to solve those current challenges: variety of datasets, updates & upgrades of repositories/datasets, a centralized repository, more research in de-identification, strategies to share complex data such as ‘cloud services’ and publisher support.

Conclusion & future wor

- For our next steps we plan on contacting some of the repositories to understand why they stopped maintaining the sites.

- Additionally, we will try to raise the awareness of our webportal with the hope that researchers contribute and keep our list up to date.

Appendix A. Overview of the datasets

- First, we identified several datasets by reviewing articles

- second we identified several sets by running Google searches and a third we identified

- third party services that we found in our articles‘ analysis. All of the findings are presented on our website http://datasets.fbreitinger.de/ which allows to contribute to the collection. In addition, we attached Tables A.3eA.5 which contain the available dataset repositories.

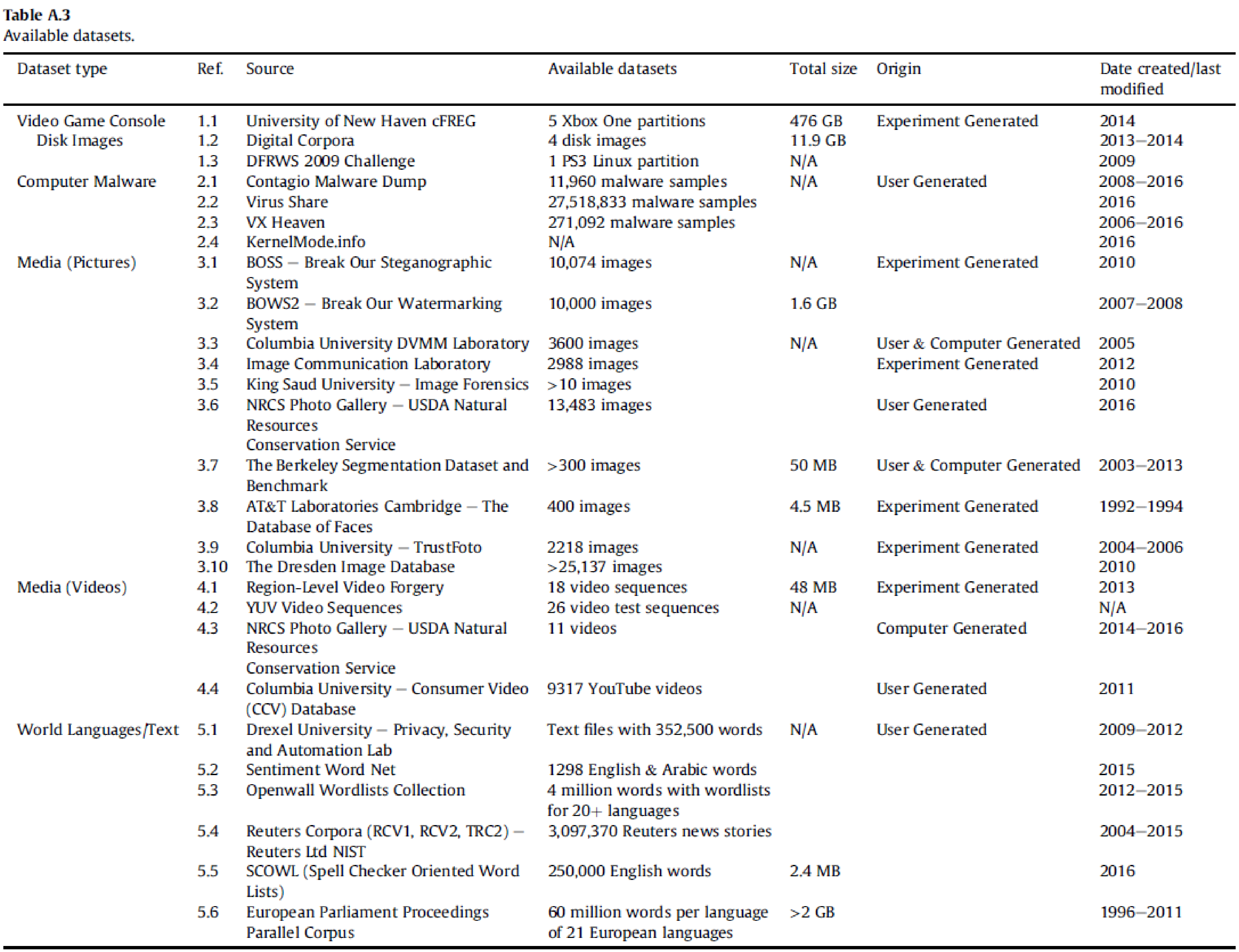
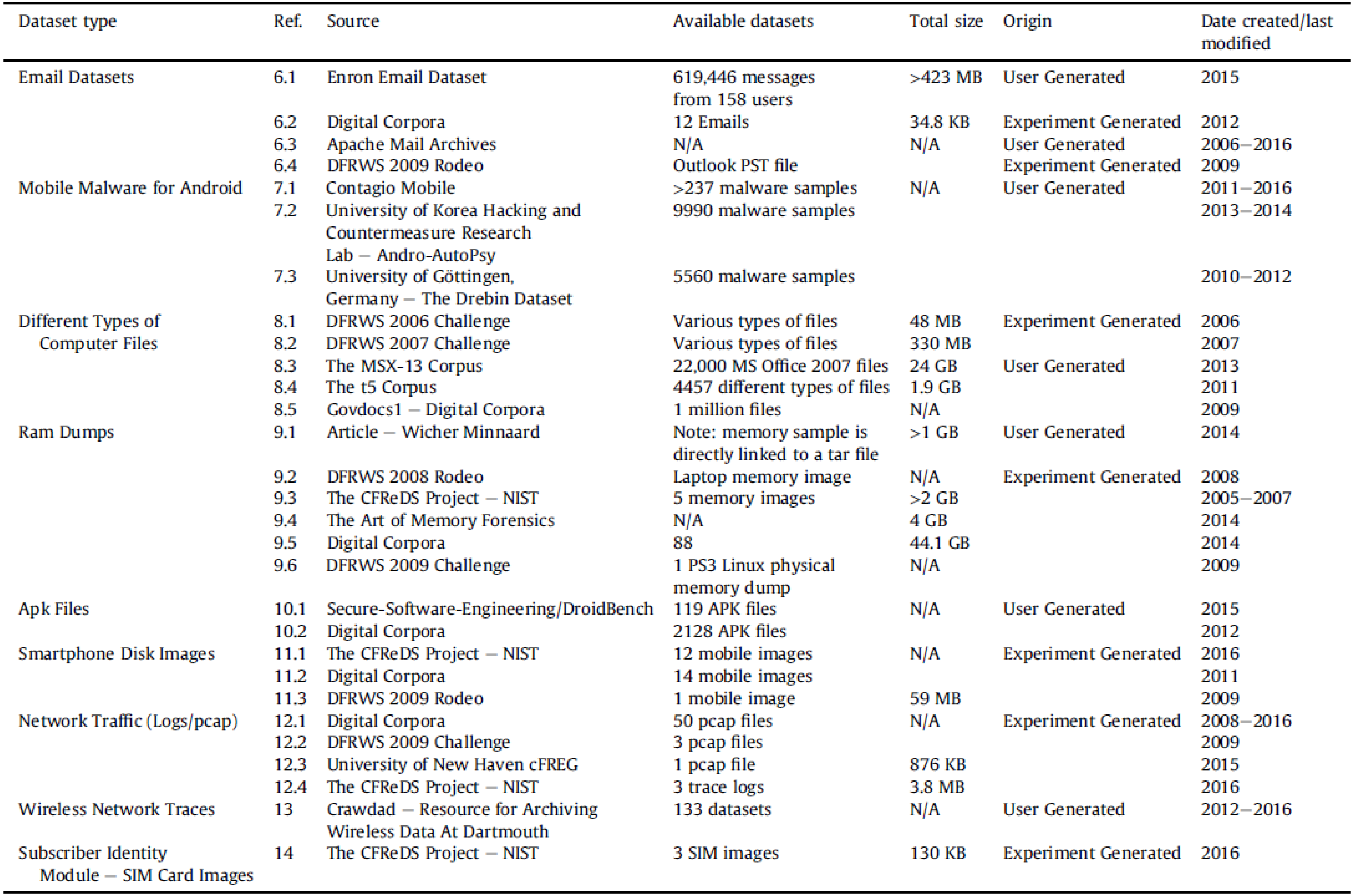


Table A.3
Available datasets.

Dataset type	Ref.	Source	Available datasets	Total size	Origin	Date created/last modified
Video Game Console Disk Images	1.1	University of New Haven cFREG	5 Xbox One partitions	476 GB	Experiment Generated	2014
	1.2	Digital Corpora	4 disk images	11.9 GB		2013–2014
	1.3	DFRWS 2009 Challenge	1 PS3 Linux partition	N/A		2009
Computer Malware	2.1	Contagio Malware Dump	11,960 malware samples	N/A	User Generated	2008–2016
	2.2	Virus Share	27,518,833 malware samples			2016
	2.3	VX Heaven	271,092 malware samples			2006–2016
	2.4	KernelMode.info	N/A			2016
Media (Pictures)	3.1	BOSS – Break Our Steganographic System	10,074 images	N/A	Experiment Generated	2010
	3.2	BOWS2 – Break Our Watermarking System	10,000 images	1.6 GB		2007–2008
	3.3	Columbia University DVMM Laboratory	3600 images	N/A	User & Computer Generated	2005
	3.4	Image Communication Laboratory	2988 images		Experiment Generated	2012
	3.5	King Saud University – Image Forensics	>10 images			2010
	3.6	NRCS Photo Gallery – USDA Natural Resources Conservation Service	13,483 images		User Generated	2016
	3.7	The Berkeley Segmentation Dataset and Benchmark	>300 images	50 MB	User & Computer Generated	2003–2013
	3.8	AT&T Laboratories Cambridge – The Database of Faces	400 images	4.5 MB	Experiment Generated	1992–1994
	3.9	Columbia University – TrustFoto	2218 images	N/A	Experiment Generated	2004–2006
	3.10	The Dresden Image Database	>25,137 images			2010
Media (Videos)	4.1	Region-Level Video Forgery	18 video sequences	48 MB	Experiment Generated	2013
	4.2	YUV Video Sequences	26 video test sequences	N/A		N/A
	4.3	NRCS Photo Gallery – USDA Natural Resources Conservation Service	11 videos		Computer Generated	2014–2016
	4.4	Columbia University – Consumer Video (CCV) Database	9317 YouTube videos		User Generated	2011
World Languages/Text	5.1	Drexel University – Privacy, Security and Automation Lab	Text files with 352,500 words	N/A	User Generated	2009–2012
	5.2	Sentiment Word Net	1298 English & Arabic words			2015
	5.3	Openwall Wordlists Collection	4 million words with wordlists for 20+ languages			2012–2015
	5.4	Reuters Corpora (RCV1, RCV2, TRC2) – Reuters Ltd NIST	3,097,370 Reuters news stories			2004–2015
	5.5	SCOWL (Spell Checker Oriented Word Lists)	250,000 English words	2.4 MB		2016
	5.6	European Parliament Proceedings Parallel Corpus	60 million words per language of 21 European languages	>2 GB		1996–2011

Table A.4
Available datasets.



Dataset type	Ref.	Source	Available datasets	Total size	Origin	Date created/last modified
Email Datasets	6.1	Enron Email Dataset	619,446 messages from 158 users	>423 MB	User Generated	2015
	6.2	Digital Corpora	12 Emails	34.8 KB	Experiment Generated	2012
	6.3	Apache Mail Archives	N/A	N/A	User Generated	2006–2016
	6.4	DFRWS 2009 Rodeo	Outlook PST file		Experiment Generated	2009
Mobile Malware for Android	7.1	Contagio Mobile	>237 malware samples	N/A	User Generated	2011–2016
	7.2	University of Korea Hacking and Countermeasure Research Lab – Andro-AutoPsy	9990 malware samples			2013–2014
	7.3	University of Göttingen, Germany – The Drebin Dataset	5560 malware samples			2010–2012
Different Types of Computer Files	8.1	DFRWS 2006 Challenge	Various types of files	48 MB	Experiment Generated	2006
	8.2	DFRWS 2007 Challenge	Various types of files	330 MB		2007
	8.3	The MSX-13 Corpus	22,000 MS Office 2007 files	24 GB	User Generated	2013
	8.4	The t5 Corpus	4457 different types of files	1.9 GB		2011
	8.5	Govdocs1 – Digital Corpora	1 million files	N/A		2009
Ram Dumps	9.1	Article – Wicher Minnaard	Note: memory sample is directly linked to a tar file	>1 GB	User Generated	2014
	9.2	DFRWS 2008 Rodeo	Laptop memory image	N/A	Experiment Generated	2008
	9.3	The CFReDS Project – NIST	5 memory images	>2 GB		2005–2007
	9.4	The Art of Memory Forensics	N/A	4 GB		2014
	9.5	Digital Corpora	88	44.1 GB		2014
	9.6	DFRWS 2009 Challenge	1 PS3 Linux physical memory dump	N/A		2009
Apk Files	10.1	Secure-Software-Engineering/DroidBench	119 APK files	N/A	User Generated	2015
	10.2	Digital Corpora	2128 APK files			2012
Smartphone Disk Images	11.1	The CFReDS Project – NIST	12 mobile images	N/A	Experiment Generated	2016
	11.2	Digital Corpora	14 mobile images			2011
	11.3	DFRWS 2009 Rodeo	1 mobile image	59 MB		2009
Network Traffic (Logs/pcap)	12.1	Digital Corpora	50 pcap files	N/A	Experiment Generated	2008–2016
	12.2	DFRWS 2009 Challenge	3 pcap files			2009
	12.3	University of New Haven cFREG	1 pcap file	876 KB		2015
	12.4	The CFReDS Project – NIST	3 trace logs	3.8 MB		2016
Wireless Network Traces	13	Crawdad – Resource for Archiving Wireless Data At Dartmouth	133 datasets	N/A	User Generated	2012–2016
Subscriber Identity Module – SIM Card Images	14	The CFReDS Project – NIST	3 SIM images	130 KB	Experiment Generated	2016

Thank you