NapierOne: A Modern Mixed File Data Set Alternative to Govdocs1

Simon R. Davies[a,*], Richard Macfarlane[a] and William J. Buchanan[a]

[a]*School of Computing, Edinburgh Napier University, Edinburgh, UK.*

ARTICLE INFO

ABSTRACT

It was found when reviewing the ransomware detection research literature that almost no proposal provided enough detail on how the test data set was created, or sufficient description of its actual content, to allow it to be recreated by other researchers interested in reconstructing their environment and validating the research results. A modern cybersecurity mixed file data set called *NapierOne* is presented, primarily aimed at, but not limited to, ransomware detection and forensic analysis research. *NapierOne* was designed to address this deficiency in reproducibility and improve consistency by facilitating research replication and repeatability. The methodology used in the creation of this data set is also described in detail. The data set was inspired by the Govdocs1data set and it is intended that *NapierOne* be used as a complement to this original data set.

An investigation was performed with the goal of determining the common files types currently in use. No specific research was found that explicitly provided this information, so an alternative consensus approach was employed. This involved combining the findings from multiple sources of file type usage into an overall ranked list. After which 5,000 real-world example files were gathered, and a specific data subset created, for each of the common file types identified. In some circumstances, multiple data subsets were created for a specific file type, each subset representing a specific characteristic for that file type. For example, there are multiple data subsets for the ZIP file type with each subset containing examples of a specific compression method. Ransomware execution tends to produce files that have high entropy, so examples of file types that naturally have this attribute are also present. The resulting entire data set comprises of a 100 separate data subsets divided between 44 distinct file types, resulting in almost 500,000 unique files in total. A description of the techniques used to gather the files for each file type is provided together with the actions that were performed on the files to confirm that they were of the highest quality and provided an accurate representation of their specific file type. Details are also provided on the content of the entire data set as well as instructions on how researchers can gain free and unlimited access to the final data set.

While the data set was initially created to aid research in ransomware detection, it is sufficiently broad and diverse enough to allow for its application in many other areas of research that require a varied mixture of common real-world file examples. The *NapierOne* data set is an ongoing project and researchers are strongly encouraged to leverage this data set in their own research.

1. Introduction

To facilitate research repeatability, it is necessary that well respected, realistic and easily accessible standard data sets are used. This approach is confirmed by [32, 54] who stress that test data must be representative of data likely to be encountered in real-world situations. It is a known issue within the malware research community [5, 37, 45, 50], that there is often a lack of readily available, researchable ransomware data sets with Grajeda et al. [37] finding that only 4% of malware researchers end up publishing their data sets.

The main contributions of this paper are firstly, the methodology for creating a mixed file data set that contains real-world examples of commonly used file formats, and secondly, the *NapierOne* mixed file data set itself. The data set was developed during ransomware analysis research and with a focus on the development and testing of ransomware detection systems. A common characteristic of ransomware programs is that at some point they will attempt to encrypt the user's files. Tests can be used to determine the random-ness of a file, with higher randomness, or entropy, suggesting that the contents may be encrypted. One drawback with this approach is, though, that the content of some legitimate files such as archive and image files can also appear to be random. Generally, the problem of detecting ransomware can be reduced to the problem of detecting random data being written to the file system or stored in memory [51]. File types that normally have high entropy and appear to contain random data were included in this data set for this specific reason.

1.1. Cybersecurity test data sets

For cybersecurity, test data sets can cover a wide variety of data types including malware samples, email data sets, network traffic, memory images, disk and device images. This paper will focus on data sets that contain collections of files of various types [37] and will use the term 'mixed file data set' to refer to this type of data set. Abt and Baier [4] classify this type of data set as one which contains real-world data, is publicly available and manually curated.

Previously, in an attempt to address the issue of a lack of data sets, Garfinkel et al. [30] created a publicly available Govdocs1 corpus. One part of which contains nearly one million files collected using random searches of the US *.gov* domain. The resulting files are stored in 1,000 directories, each containing 1,000 files. This corpus can be obtained from [23, 33].

✉ s.davies@napier.ac.uk (S.R. Davies)

ORCID(s):
0000-0001-9377-4539 (S.R. Davies);
0000-0002-5325-2872 (R. Macfarlane);
0000-0003-0809-3523 (W.J. Buchanan)

While the Govdocs1 corpus is a large, varied, well-known and highly cited, mixed file data set containing more than 40 different file types, recently some researchers have found it necessary to augment the corpus with additional files. A summary of these enhancements and their justifications are:

- The files in the data set are now more than 12 years old and it is unknown how accurately their type and content, reflect current usage.

- Some file types that have gained popularity since the creation of the data set are not well represented. Specifically, regarding the newer Microsoft Office document types: XLSX; DOCX; and PPTX [47].

- Closer examination of the data set reveals that a class of files known as archives are under-represented [51] within Govdocs1. Archive files are normally compressed and are often used to collect multiple files together into a single file for easier portability or to reduce storage space requirements. Archive files often store directory structures, use error detection, and sometimes use built-in encryption. There are multiple types of archives currently in use, with varying properties and characteristics, however, the only type present of any significant size in the original data set is gzip.

- Modern image file types used in web design are missing (e.g. WEBP) [51].

- No examples of encrypted files or files with little structure and very high entropy are present [48].

- Files that have high entropy, such as encrypted files, tend to have little recognisable structure [48]. No examples of such files currently appear in the Govdocs1 data set. The inclusion of these types of files would prove beneficial when testing ransomware detection systems.

- The file types contained in the Govdocs1 data set only reflect the files that were found on the .gov domain, and do not necessarily represent the general popularity of the chosen file types.

1.2. Paper contribution

In this paper, we introduce a complementary data set for the Govdocs1 corpus, known as *NapierOne* [19], which may be used to address the points made above. The Govdocs1 data set and the techniques used to create it provided an excellent template on how to create and curate the *NapierOne* data set. However, no actual data from the original Govdocs1 data set is present in the proposed new data set.

An overview of the methodology used in creating the proposed data:

1. Research was performed to identify popular file formats that could be candidates for inclusion into the mixed file data set. Usage statistics from more than 10 independent sources were gathered, aggregated and collated. This research produces a final list of file types that were included in the data set.

2. 5,000 example files of each of the identified file types were then sourced.

3. Some additional actions were then performed on some of the gathered files in order to generate additional data subsets, for example, the creation of collections of files into archives.

4. Robust validation was then performed on all of the files in the data set, including virus scanning, submission to virustotal [63], duplication removal and file format verification.

5. Each data subset was then documented and curated, with a copy finally being placed for public use on the www.napierone.com website [19].

The presence of particular files types in the final data set was determined from research into the general popularity of the file type and not by its presence in a particular source.

While the data set was initially created to aid research into ransomware detection, it is felt that it is sufficiently broad and diverse enough to allow its use in many other areas of research such a file analysis, compression technique comparison, image compression comparison and Microsoft Office file format analysis to name just a few examples.

2. Related Work

Cybersecurity-related research data sets can be classified into various types depending on the data they contain. The following five classifications have been proposed [4, 30]; Malware Samples, Disk Images, Memory Images, Network Data and File Data. The remainder of this paper will concentrate on file type data sets. This type of data set contains a rich and varied collection of typical file types that may be found on modern computer systems and can be used to generate target file collections that are attractive to ransomware attacks. A recurring theme in many of the papers surveyed is that there exists a lack of publicly available mixed file data sets available for malware research and this is a large obstacle faced by ransomware researchers [6]. Currently, many studies build their own data sets by downloading raw samples from public repositories. However, many of these studies do not follow standard approaches for creating these data sets. Building a high quality, publicly available data set could be of great use, as it would contribute to building robust and accurate detection models.

Berrueta [9] identifies that there are no common metrics of accuracy and performance in ransomware detection and the fragmentation of scientific research on ransomware combined with a lack of coherent investigation methodology is a major challenge in this research [17]. This view is supported by Maigida [45] who state that the lack of readily available, researchable ransomware data sets is also hindering the speedy development of detection and prevention solutions. The availability of up-to-date ransomware data sets is critical in evaluating newly proposed detection methods as the advancement of ransomware techniques could quickly render old data sets irrelevant. Almost no reviewed proposal

for a ransomware detection technique provided enough detail on how the data set was created or sufficient description of its actual content to allow it to be recreated by other researchers interested in reconstructing their environment and validating the research results.

It has been reported by Grajeda et al. [37] that as many as two-thirds of the data sets used by researchers are experimentally generated with only a third being created from real-world data. They go on to say that in 96% of cases, these data sets are not released for public use or scrutiny. This is against the recommendations of some researchers in the field [4, 27, 37] who stress the importance of sharing data sets, allowing researchers to replicate results and improve the state-of-the-art [37]. Building a public, ready-to-use ransomware data set would facilitate upcoming studies [37] and encourage more researchers to further investigate ransomware and produce solutions for various issues [37]. Two separate independent surveys agree with this conclusion, finding that these data sets would contribute to building robust and accurate detection models [6, 50]. In addition to this, the development of a universal testing platform allowing each of the methods to be evaluated in isolation using the same data sets would provide transparency in the evaluation processes as each test would be performed in a consistent, reproducible manner allowing direct comparisons to be made [30, 48, 50]. Berrueta [9] goes further and states that comparability and reproducibility are neither facilitated by the problem at hand nor by the way researchers present their results.

The research performed by Grajada [37] shows that many researchers prefer not to share their data sets [4]. Offered reasons for which being, firstly, researchers may not have the capability of sharing the set due to its size and the researcher's lack of available resources. A second factor may be related to the data set content as well as law and privacy concerns. Thirdly, the importance of sharing their data had not been considered. Finally, manually compiling data sets can be time-consuming, sometimes requiring months of work, so researchers having access to data sets with limited public accessibility have a clear competitive advantage. Grajeda et al. [37] research goes on to show that less than 4% of researchers shared their data set while on the other hand almost 50% make use of existing data sets. In other words, whenever a repository or a sophisticated data set is available, researchers appreciate and utilise it.

The following defines some of the publicly available mixed data sets.

2.1. Govdocs1

The most well-known publicly available data set used in malware analysis today was developed by Garfinkel [30, 33] in 2009 and is known as Govdocs1. The data set was designed to enable reproducibility of forensic research but makes no claims regarding the popularity of the file types it contains. One part of the data set contains approximately one million files collected using random searches of the *.gov* domain. It has been used by many ransomware researchers [14, 27, 37, 44, 46, 47, 48, 50, 51, 52, 55], supporting the claim that this

Table 1. Govdocs1 File Types and File Counts

Ext	# files	Ext	# files	Ext	# files
bmp	75	jar	34	sql	632
chp	2	java	323	squeak	1
csv	18,396	jpg	109,281	swf	3,691
data	1	js	92	sys	8
dll	7	kml	995	tif	3
doc	80,648	kmz	949	tmp	196
docx	169	log	10,241	ttf	104
dwf	474	pdf	232,791	txt	84,091
eps	5,465	png	4,125	wp	17
exe	5	pps	1,629	xbm	51
exported	5	ppt	50,257	xls	66,599
gif	36,301	pptx	219	xlsx	46
gz	13,870	ps	22,129	xml	41,994
hlp	660	pst	1	zip	27
html	191,409	pub	76		

data set is a well-known and respected source of test data. In 2017 Grajeda et al. [37] reported that this was the most popular data set currently in use. The files are stored in 1,000 directories, each containing 1,000 files of various types, as well as 10 randomly assigned streams for development and testing purposes.

This data set is now more than ten years old and during this time the use of some file types has diminished and new types have emerged. The sample size of some file types within the corpus now does not accurately reflect modern use. For example, there exist only 169 examples of files with the DOCX format, whereas, in reality, this format has become prevalent. Microsoft Office document file types, such as XLSX, DOCX and PPTX, are often targeted by ransomware [6, 42, 43], so their inclusion in the data set would be beneficial in ransomware research testing. A table illustrating the file types present in this corpus together with its sample size is shown in Table 1.

Apart from the small data set size of modern Microsoft Office formats, there are also few examples of different archive types and other highly entropic file types and no examples of encrypted files. Entropy is often used in ransomware detection systems, so the presence of high entropy files in the data set, would be useful during the development and testing of such systems. To address these points, some researchers [47, 48, 51] have attempted to enrich the original data set with additional file types, before using it, making the resulting hybrid data set more realistic and relevant to their research.

2.2. t5

When evaluating ssdeep and sdhash for similarity matching, Roussev [53, 54] created the *t5* corpus based on the first four directories of the Govdocs1 corpus. Files that were smaller than 4KB and larger than 16.5MB being excluded from this data set. The resulting corpus has 4,457 files including DOC, GIF, HTML, JPG, PDF, PPT, TXT and XLS. Approximately 45.7% of the files are text-based (TXT or HTML). The files were selected from neighbouring directories in the Govdocs1 corpus in the hope that there would

be some similarity between them as they are likely to have originated on the same server. As this data set is based on the Govdocs1 data set, the same issues also exist with this data set such as that there are only a limited number of modern file types present. The t5 corpus has been used extensively in research into approximate matching [12, 46, 52, 57], which is a technique used to identify the similarity between digital artefacts (sequence of bytes) [14].

2.3. msx-13

Continuing in his research, Roussev [55] developed the *msx-13* corpus containing over 22,000 Microsoft Office 2007 documents. DOCX, XLSX and PPTX files were gathered using Google queries across 10 domains which favour English language content: .com, .net, .org, .edu, .gov, .us, .uk, .ca, .au and nz. The data set also contains MSZ archive files which are zip containers containing deflate-encoded content and embedded objects in their native format, however, they found great variations in the content of the files making it impossible to come up with a general classification scheme. It was also noted that the gathered PPTX files were much larger and had many more embedded objects than DOCX and XLSX.

2.4. Other Sources

Some researchers describe in detail how they generated their data sets. For example, Pont [51] outlines the precise technique they used to generate their data. Initially, their data set was based on the Govdocs1 corpus but then enriched with multiple image types and archive files [49]. This appears to be the most up-to-date comprehensive data set available, but due to its recent release, it is unknown if it has been leveraged by other researchers. Jung [42] also developed a data set that while it is only limited to PDF files, it does contain encrypted versions of these files as well. The methods used to generate the data set used by DeGaspari [20] are also well described, discussing the files types present and their distribution, but the actual data set has not yet been made public by the researchers and the description is not sufficiently detailed enough to be able to reproduce the data set. During the research, the following websites were also searched for prospective data set examples [23, 62].

2.5. Data Set Conclusions

Considering all of this evidence from the reviewed literature, it is clear that data sets used within the ransomware testing community are currently facing several issues. These include the fact that many researchers are manually creating their own data sets which are often not released after the work is completed [37] and there is a lack of standardised data sets that are suitable for ransomware research [4, 6]. These weaknesses combined produce one of the major disadvantages facing the research community, that of low reproducibility and comparability [37].

If a modern, diverse, representative mixed file data set could be created and made publicly available, its effect would be to improve the quality and pace of research especially in domains such as ransomware analysis and digital forensics [37]. Building such a data set would be of great use, as it would contribute to building robust and accurate detection models [6]. Experimenting with real-world data is crucial for developing reliable algorithms and tools as "how can we learn from our past when we do not have real, accessible data to learn from?" [8].

3. Methodology

Some of the forensic corpora reviewed in the previous section have proven to be an excellent resource in the field of ransomware research. The philosophy guiding the development of the *NapierOne* data set was to mimic their success, by complementing these data sets, with one that contains more examples of data formats that have become more prominent, as well as including files that have similar characteristics to encrypted files. The techniques used in determining the data set candidate files types are discussed below followed by how the actual files for each type were sourced.

3.1. File Selection

An important aspect of building a representative data set relates to file type usage and popularity. It is known that Google gathers statistics on file types while it performs its website indexing searches. However, the statistics are only gathered on a limited number of file types [35, 36]. While it was not possible to discover a definitive ranked list of files types currently in use, it was decided to adopt a consensus approach. This involved querying various sources of possible usage information and gathering approximate lists of up to their Top 40 file types. These lists were then compared and aggregated with the lists gathered from other sources, resulting in a fair representation of what are currently popular file types are in use today. The list produced is not proposed as definitive but rather a best guess consensus. A list of the sources for file type usage information is discussed below. The alphabetical identification used for each data source is then repeated in the data set description shown in Table 5.

A - VirusTotal Statistics taken from file submissions to this website [63] during that past 12 months were recorded.

B - W3Techs This website [64] provides information about the usage of various types of technologies on the web. Statistics relating to the most popular images types were gathered from here.

C - Search Engine Statistics Searches were performed for more than 100 different candidate file types. Search trends and statistics were also analysed and recorded in an attempt to identify the popularity of the file type. Although Google only index certain file types [35], their justification for doing so could also be considered a reason to include these files types in the data set.

D - File Preservation There exists research into the investigation of digital data preservation. One aspect of this research is to identify popular file formats. The findings from this research [1, 2, 3, 41, 56] were also taken into account, as this research identifies certain files as being important and

have the potential for remaining in active use for the foreseeable future.

E - MIME Type Statistics A MIME type [28, 29] is a standard label used to indicate the nature and format of a document, file, or assortment of bytes. It is mainly used in Internet-based technology so that software can know how to handle the data. It serves the same purpose on the Internet that file extensions do on Microsoft Windows. When crawling the internet, these labels may be recorded and the frequency of them can be used to identify the popularity of file formats. Statistics have thus been gained from the WayBackMachine [65] and for CommonCrawl [15], who record and publish monthly statistics on MIME types that are encountered.

F - fileinfo.com This website [26] performs research into file types and regularly publishes ranked lists of popular file types. These popularity lists are calculated from both the requests that the website receives and general web traffic analysis.

G - Ransomware Targets One of the anticipated initial applications of this data set will be in the field of ransomware research. For the data set to be useful in this research, it should contain a good selection of file types that are regularly targeted by ransomware [6, 42, 43], typical examples of which being DOCX and XLSX. This additional criterion did not impact the main goal of listing common file types as these file types appear in multiple lists.

H - High Entropy As the data set will initially be used in ransomware detection testing, it was considered a benefit for the data set to contain files that had this attribute. One technique used to identify files that have been affected by ransomware is to test the file's entropy so including files that normally have this attribute, such as archive files, would provide realistic challenges in ransomware identification. This secondary consideration does not impact the overall goal of generating the file type popularity list.

I - File-extension These are others websites [24, 25] specialising in file type research and regularly publish file type usage statistics.

J - OS Installation File type count statistics were gathered from clean windows installations (XP 2003 SP1, Vista 2006 and Windows 10 20H2) as well as four clean 64Bit Linux installations (MX Linux 19.4, Debian 5.9.1, Ubuntu 18.04.1, and Ubuntu 20.10).

K - Govdocs1 Inclusion of the file types in the original Govdocs1 data set [23] were also taken into consideration when deciding if they should appear in this data set. However, the actual files being sourced from elsewhere.

The results for all the sources were then sorted by popularity and aggregated into a list of approximately 40 file types. To encourage consensus, the file types needed to appear in at least two lists before they were considered in the final results.

The contents of the new data set were generated in the following ways:

- Examples of files using the most commonly used document types such as DOC, DOCX, PDF, PPT, PPTX, XLS, and XLS were be gathered from the UK government domain, *gov.uk*. Files collected from this domain are covered by the 'Open Government License agreement for Public Sector Information' [59]. This agreement allows for the free distribution of the documents as long as they are accompanied by the correct attribution.

- Image files in the TIF format were taken from the RAISE data set [16]. This data set permits redistribution of its contents as long as it is accompanied by the correct attribution. Subsequent image file type data subsets, for example, BMP, DWG, EPS, GIF, JPG, PNG, SVG and WEBP were generated from this original TIF data set. In addition, examples of image files for each of these image types were taken from the CommonCrawl [15] dataset, thus providing real-world examples of these file types.

- Archive data subsets were generated using files from the documents subset.

- Audio files were gathered from the Free Music Archive [22]. Its license permits free and unrestricted usage and distribution.

- Video files, in MP4, format, were gathered from the Kinetics-700 data set [11, 21, 34]. Its license also allows free usage and distribution when using the correct attribution. Other video data sets in different formats were generated by transcoding this MP4 data set. Real-world examples of video files were also sourced from the CommonCrawl [15] dataset..

- Examples of files in the EPUB format were taken from Project Gutenberg [39], which is an online library of free eBooks.

3.2. Ethics

Current research institutional guidelines concerning the handling of Personally Identifiable Information (PII) present within data sets are, that if the data has been publicly sourced, as the files in the data set have been, there is no requirement for further anonymising or obfuscation of the data. The justification for this being that the information is already in the public domain. This is considered to be a positive advantage as this data set would then be useful for researchers that rely on the file metadata being present. As well as also preserving the data in its original form.

However, there remains a possibility in the future, depending on changing guidelines or regulations which may require the data to be cleaned. Possible future actions that could be performed on the data could be to have any versioning, comments and metadata removed. An attempt to anonymise any PII from the content of the files could also be undertaken if required. To date, these actions have not yet been performed on the data set.

```
<sequence-number>-<original ext>-<extra>.<extension>

Where:
sequence-number is a number between 0 and 5,000
original ext is the files original file extension string
extra contains extra information describing the file's contents
extension is the file extension value.
```

Fig. 1. File naming convention

```
0001-docx-password.docx

Where:
0001 first example in the subset.
original ext Originally a docx file.
extra File is password protected.
extension file is of type docx.
```

Fig. 2. Example file name.

In accordance with organisation guidelines and best practice recommendations [37, 58] a Data Management Plan (DMP) has been developed for this data set and is available on request. This report contains precise information relating to the data set's management, curation, retention, security, access and responsibilities. Attribution information is provided with each data subset and these should be followed when the subset is used.

4. Data Set Creation

Over 40 distinct file types were identified from the research described in Section 3.1. Some file type data sets were further divided into multiple data subsets. These subsets were used to contain examples of files that demonstrate different characteristics. For example, multiple data subsets were created for the ZIP file type. Each subset containing examples of files compressed using different compression algorithms. In most case, there exist 5,000 example files for each data set and subset. The resulting *NapierOne* [19] data set contains nearly 500,000 unique files distributed between 100 separate data sets and subsets. A detailed overview of the structure of the data set is shown in Figure 5.

4.1. Data Set Structure

As with the Govdocs1 data set, once the data had been downloaded it needed to be analysed, validated, characterised and curated [30]. To provide consistency across the entire data set, a standardised naming convention was used. Reasons for renaming the files to follow this convention were: firstly the file names reflect to some extent the content and structure of the file itself and so are self-documenting. The renaming processes provides a small degree of anonymisation and finally, files containing related content are linked via their sequence numbers. A description of the file naming convention used is shown in Figure 1. The inclusion of the sequence number facilitates the possibility of being able to cross-reference certain related files across data subsets, such as in tracing image files or archives that use the same underlying content. The file's extension value appears as part of the file name, as well as the file's actual extension. It is in both places because in some cases, such as when a file is encrypted by ransomware, the actual extension is modified. In these circumstances, the file's original extension can still be seen from the file name.

For example, for the first file in the DOCX data subset that has been password-protected, the name would be the value shown in Figure 2. Examples of how the sequence number is used to link related files are shown in Tables 2 and 3. The authors retain the ability to determine each file's original file name if required. When renaming the files, a record is kept of the mapping between the original and new file names to allow the original name to be determined if required.

The files within the data set are organised by file type into a directory hierarchy. On the top level are the main file types, such as DOC, PPTX, ZIP, and so on. Below these main type classification directories are one or more subdirectories, one for each type of file format variation. For example, a BZIP directory containing files created using the bzip compression method of the ZIP program, would be found under the main zip directory. The directory structure of the data set can be deduced from Table 5. Contained within each directory that holds data, is a description file in both HTML and PDF formats, which describes the data held within that data subset. The description file includes amongst other things: a description of the data files; where the data was sourced; the licence details; and contact information.

4.2. File Sources

The *NapierOne* [19] data set is a collection of multiple data subsets each associated with one specific file type or a specific, configuration of a file type. However, on a conceptual level, the individual data subsets can be grouped as follows.

Document files. Various techniques were employed to gather file examples from websites covered by the *gov.uk* domain. These included using the API functionality provided by the following gateways [10, 13, 18], direct data set download and files available for download on sub-domain websites. Using these techniques it was possible to gather examples of files with the following file formats: CSV, DOC, DOCX, JPG, PDF, PPT, PPTX, TXT, XLS, XLSX and WEBP. Two additional data subsets were also created for each of these main file types. One where the magic number for the file has been removed and a second where the original file is protected with a password. The same password being used for all files, the password used was *napierone* [19]. Data sets for the OpenOffice file formats, ODS and OXPS, were derived from the Microsoft Office files. Some other web-related files such as CSS, HTML, Javascript, JSON and XML were also sourced from these websites. ICS calendar data was generated using the following program [7].

Media files. Two distinct methods were used to gather files of these types. Firstly real-world examples were re-
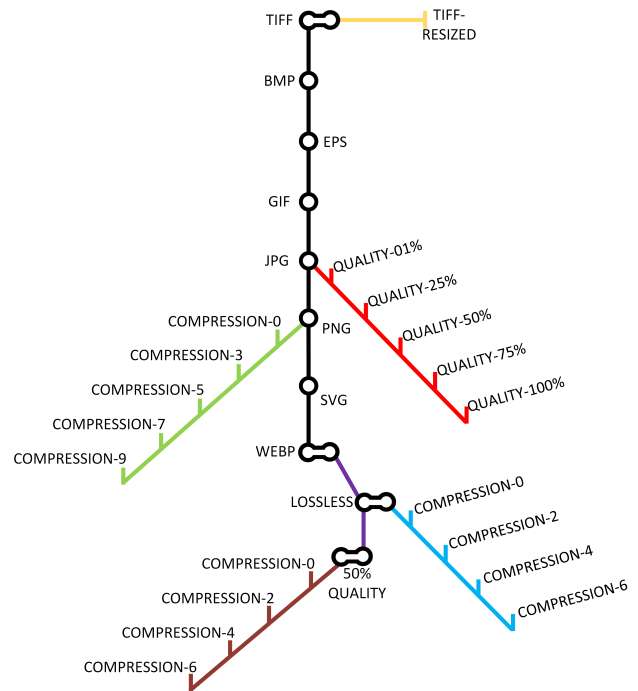
Table 2. Image File Relationship to Original TIF Image.

Filename	Description
0001-tif.TIF	Original image taken from RAISE in TIF format
0001-tif-resized.TIF	Reduced size version in TIF format
0001-webp-lossless-c0.webp	Lossless compression level 0 version in WEBP format
0001-webp-lossless-c2.webp	Lossless compression level 2 version in WEBP format
0001-webp-lossless-c4.webp	Lossless compression level 4 version in WEBP format
0001-webp-lossless-c6.webp	Lossless compression level 6 version in WEBP format
0001-webp-q50-c0.webp	50% quality compression level 0 version in WEBP format
0001-webp-q50-c2.webp	50% quality compression level 2 version in WEBP format
0001-webp-q50-c4.webp	50% quality compression level 4 version in WEBP format
0001-webp-q50-c6.webp	50% quality compression level 6 version in WEBP format
0001-svg.svg	Version of original image in SVG format
0001-png-c0.png	Compression level 0 version in PNG format
0001-png-c3.png	Compression level 3 version in PNG format
0001-png-c5.png	Compression level 5 version in PNG format
0001-png-c7.png	Compression level 7 version in PNG format
0001-png-c9.png	Compression level 9 version in PNG format
0001-jpg-q001.jpg	1% quality compression version in JPG format
0001-jpg-q025.jpg	25% quality compression version in JPG format
0001-jpg-q050.jpg	50% quality compression version in JPG format
0001-jpg-q075.jpg	75% quality compression version in JPG format
0001-jpg-q100.jpg	100% quality compression version in JPG format
0001-gif.gif	Version of original image in GIF format
0001-dwg.dwg	Version of original image in DWG format
0001-eps.eps	Version of original image in EPS format
0001-bmp.bmp	Version of original image in BMP format

Fig. 3. Image File Relationships.

trieved from the the CommonCrawl [15] dataset. Secondly, random image files in an uncompressed format were gathered from the RAISE [16] public archive to create an initial image data subset. A second data subset was then generated in the same format, based on the original data set but with a reduced size. Images in the original TIF data subset were then converted into different image formats using the Advanced Batch Image Converter program [40]. Image data subsets derived from the original TIF image subset were BMP, DWG, EPS, GIF, JPG, PNG, SVG and WEBP. Some image subsets have multiple examples of a particular file type, for example, JPG and WEBP, with differing encoding qualities and compression levels. A graphical representation of this relationship is shown in Figure 3. One side effect of this file relationship being that image files that have the same sequence number will contain a rendering of the same image. Using the naming convention outlined in Figure 1, the relationship that the original TIF image *0001-tif.TIF* has with all the other image formats is shown in Table 2. If viewed, then all these files will display the same image. The JPG and WEBP data subsets were also augmented with images gathered from the gov.uk domains.

Examples of other media data subsets present are MP3 audio files taken from the Free Music Archive [22] and MP4 video files taken from the Kinetics-700 archive [11, 21, 34] as well as sourcing real-world examples from the Common-Crawl [15] dataset. Examples of files in the e-book format EPUB were gathered from the project Gutenburg [39] which is a library of over 60,000 free eBooks.

System files. Data subsets containing files relating to system installation are also present. Files relating to Linux installations (ELF, BIN) were gathered from clean Linux installations of Debian 5.9.1, MX Linux 19.4, Ubuntu 18.04.1 and Ubuntu 20.10. DLL and EXE file examples were taken from clean Windows XP and Vista installations. A similar action was performed, on all the program files gathered from operating system distributions, as Garfinkel [30, 31] applied

to their executable files. That was to alter a portion of a file so that the original file could not be identified and could not be executed. The alteration was made in such a way as to minimise the effect on the entropy of the file. This alteration allows analysis of executable files but prevents them from being used without a license, which is believed to be sufficient redaction for the purpose of distributing the files under the *'fair use'* provisions of the US Copyright Act. Some other configuration file types such as CSS and JSON were also sourced in this way. Android installation files, APK, were gathered from the AndroZoo public data set [61].

Other files. Other data subsets present are a pseudo-random data subset named PSEUDO. This is a synthetic data subset of files created using the Python pseudo-random number generator function 'os.urandom'. Files were created using this technique with lengths varying between 512 and 2,048 bytes. A second random number data subset called PURE is present which has files that contain true random numbers generated using unpredictable physical means, for example, atmospheric noise [38]. Files were created using this technique with lengths also varying between 512 and 2,048 bytes. Data sets of these types prove useful as a benchmark when comparing against other highly entropic files such as archives.

File type with the DAT extension describe files containing some sort of text or binary data. Formatting of this data depends on the program which created or utilises the given file and can vary in content and structure. As it was not possible to gather a representative example of this file type, it was not included in the final data set. The TAR file format was not that popular, but its presence was required as an in-

Table 3. Archive File Relationships

Filename	Description
0001-7z-bzip2.7z	7zip compression using bzip method
0001-7z-encrypted.7z	7zip compression & password protected
0001-7z-hightcompress.7z	7zip using high compression
0001-7z-lzma.7z	7zip compression using lzma method
0001-7z-lzma2.7z	7zip compression using lzma2 method
0001-7z-ppmd.7z	7zip compression using ppmd method
0001-gz.gz	gzip compression
0001-rar.rar	RAR compression
0001-tar.tar	TAR compression
0001-zip-bzip2.zip	Zip compression using bzip method
0001-zip-deflate.zip	Zip compression using DEFLATE
0001-zip-encrypted.zip	Zip compression & password protected
0001-zip-highcompress.zip	Zip using high compression
0001-zip-lzma.zip	Zip compression using lzma method
0001-zip-ppmd.zip	Zip compression using ppmd method
0001-zlib.zlib	Zlib compression

Table 4. Tools Used.

Program names	Version	Comments
Advanced Batch Image Converter	1.2.2	[40]
Bulk Rename Utility	3.3.2.0	www.bulkrenameutility.co.uk
Antimalware Client	4.18.2101.9	Windows 10
Engine	1.1.17800.5	Defender
Antivirus	1.331.1067.0	
Anti-spyware	1.331.1067.0	
PDF Reader	11.0.00	Adobe
Microsoft Office	15.0.4430.2027	Office Pro 2013
WinRAR	3.42	www.rarlab.com
gzip	3.8.2	Python module
7Zip	9.20	www.7-zip.org

termediate data subset when creating other data subsets such as ZLIB. It is included in this data set for this reason.

4.3. File Structure

Once the main data types had been collected, two additional actions were performed. Firstly a directory structure was generated where 5,000 sub-directories were created named from 0001 to 5000. Each of these sub-directories was then populated with examples from the document and image data subsets, resulting in each directory containing between 5 and 15 files of various types. The purpose of this file structure was to form a hierarchy that could be used when creating data subsets of archive file types. The following archive data subsets were then created by performing compression on this file structure: 7Zip, GZIP, RAR, TAR, Zip and ZLib. As these file formats tend to have high entropy, they will be useful in ransomware detection testing. For some file types, multiple data subsets were created allowing the option to represent different compression levels or compression techniques. Again the same naming convention was leveraged, so using the sequence number, individual-related archives can be linked together. For example, Table 3 shows the different archives that were generated from the source directory 0001. Unless otherwise stated, the compression tool's default values were used in the creation of the archives.

4.4. File Collection

A single, file collection, directory was created which contained examples of file types that are typically targeted by ransomware [6, 42, 43]. Examples of these file types being DOC, DOCX, PDF, PPT, PPTX, XLS and XLSX. This directory would act as a ransomware target directory and was placed on a previously prepared, isolated test machine, and a ransomware sample was executed in an ethical manner. Valid ransomware samples being sourced from Virus-Total [63]. Once the ransomware execution had completed, the target files were examined to determine if they had been affected by the ransomware's execution. If they had, then, the affected files were checked for viruses and after confirming that they were safe, were placed in their own data subset within the *NapierOne* data set. Currently, data sub-

sets of ransomware encrypted files have been created for the following ransomware strains: NotPetya; Sodinokibi; Maze; Phobos; Netwalker; Dharma; and Ryuk.

Again leveraging the sequence number, the target file and ransomware encrypted file can be connected. A graphical representation of the relationships between the documents data subsets and the file collection and file structure hierarchies is shown in Figure 4 and an overall view of the provenance of the individual data subsets is shown in Figure 5.

4.5. File Validation

Multiple actions were taken to ensure that the data contained within the data set was of the highest quality. File validation actions that were performed on all data files were:

- All the files in the data set have been scanned for viruses using Windows Defender, details of which are given in Table 4.

- The files were also checked to confirm that they contained data and that they have the correct magic numbers.

- All the files were verified to confirm that they could be opened successfully using the appropriate application.

- Confirming that no duplicates existed in the data set. This was achieved by checking the files SHA1 hash value [30].

- Over 350,000 files that had been sourced externally as well as ransomware encrypted files were submitted to virustotal [63] to confirm that they did not contain any malicious code.

- Image files were visually checked in an attempt to ensure that no illicit content was present.

The tools used in the generation for the data set, its management and validation are described in Table 4.

5. Evaluation

A detailed description of the *NapierOne* data set is presented in Table 5. The actual data set file system hierarchy levels can be deduced from the *File Type*, *Sub Classifications* and *SubSub Classifications* columns. The data subset sizes are
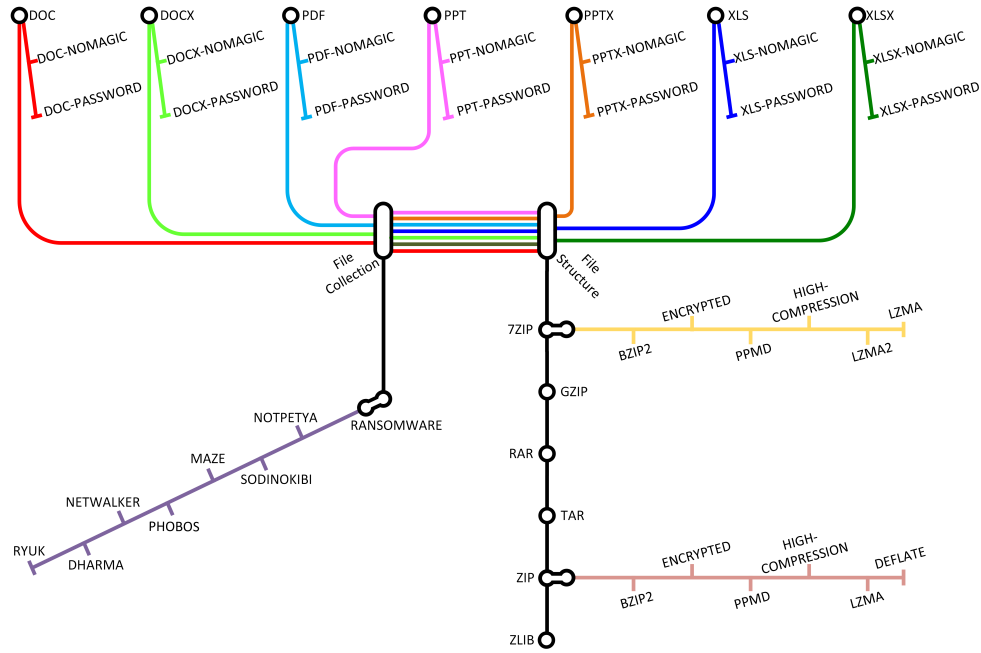
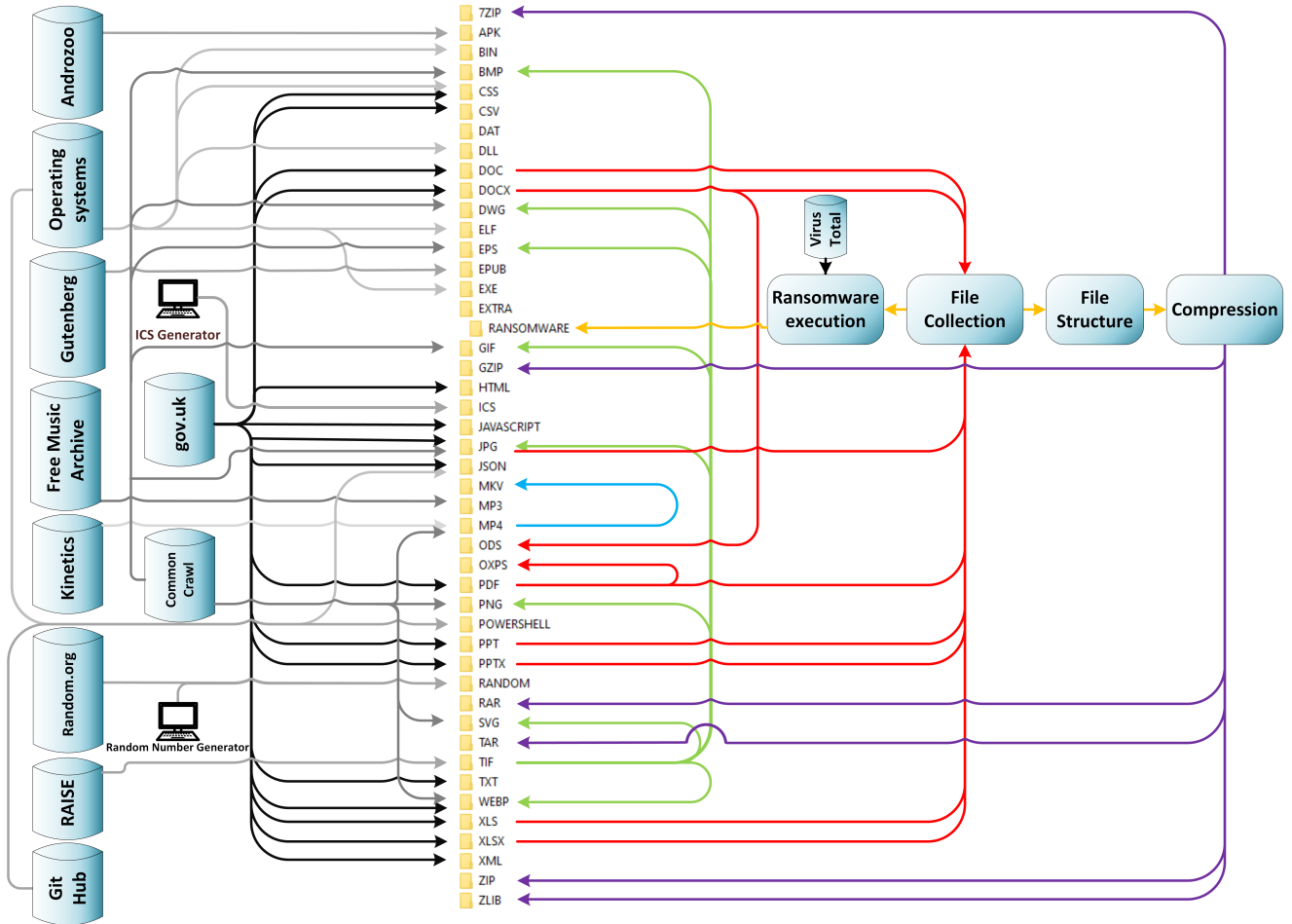**Fig. 4.** gov.uk File Collection Relationships.
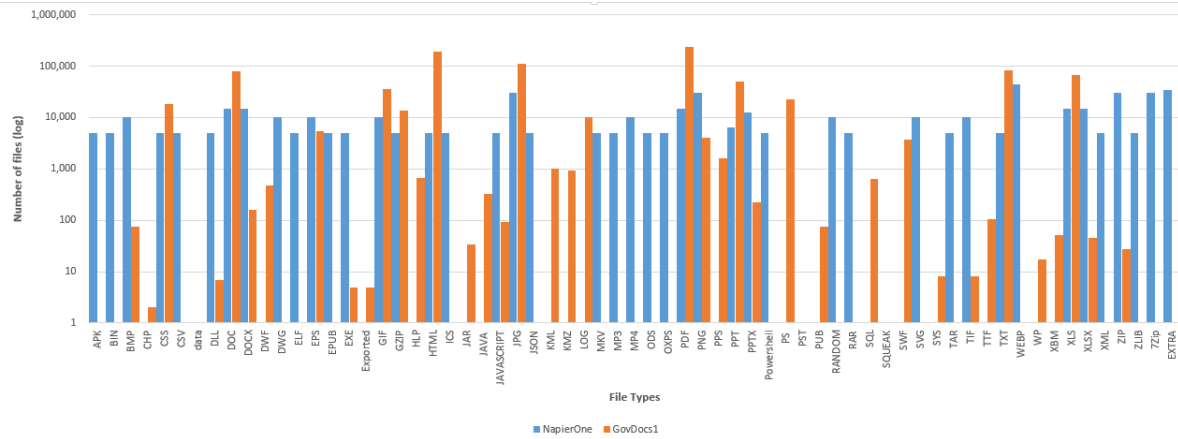

Fig. 5. Data Set Sources.

Fig. 6. Dataset Comparison.

the size of the uncompressed data set. The justification for the file type being included in the data set is represented by the columns on the right-hand side of the table. The ranked lists where the specific file type appears is illustrated using the notation described in Section 3.1.

When creating the data subsets, the goal was for each collection to contain 5,000 example files. This was achieved in the majority of cases, with the only exception being file types in the Microsoft PowerPoint format (PPT, PPTX). The reduced size of this data set was due to the lack of example files that could be sourced from the *gov.uk* domain. Instead of adding example files from other sources and mixing their provenance, it was decided to remain with the single source and accept that the data set population size was be reduced as a consequence.

5.1. Data Set Comparison

When comparing the NapierOne and Govdocs1 data sets, several differences become apparent. While both data sets contain approximately 40 different distinct file types, the majority of the files in the Govdocs1 dataset represent only a few of these file types. For example, half of the dataset consists of only three file types (HTML, JPG and PDF). Whereas it can be seen from Figure 6 that the NapierOne data set file distribution is considerably more even across all the represented files. For nine of the file types present in the Govdocs1 data set, there exists less than 10 example files, whereas in the NapierOne dataset most file types have at least 5,000 example files. The content of the Govdocs1 data set was determined by the files found using random searches of the *.gov* domain and may not be representative of other systems or file type popularity in general. The content of the NapierOne data set was determined based on the research performed in to file usage. This real-world file usage information being gathered from more than 11 separate sources. The collected usage statistics were then collated, aggregated and ranked according to frequency and then used to determine the constituents of the NapierOne data set.

There also exist many more examples of modern file formats such as DOCS, PPTX and XLSX in the NapierOne

dataset than are present in the Govdocs1 dataset. Finally, there are many more examples of files with high entropy such as archive, compressed, random data and encrypted files in the NapierOne dataset. These types of files are useful when researching, for example, ransomware detection systems. Actual specific data set information can be found in Tables 1 and 5.

5.2. Access

As long as the attribution and license stipulations are followed, access to this *NapierOne* [19] data set for research purposes is free and unrestricted. For access to the data set, visit **www.napierone.com** or alternatively contact the authors of this paper.

It can be seen from Table 5 that some data subsets are extremely large, with the total uncompressed size of the entire data set approaching 2TB. Due to its size, the data set has been broken down into smaller parts allowing the possibility to access only portions of the data set. The final data set is broken down as follows.

- Each data subset is available in three separate archives.

  - **Tiny** This archive contains 100 examples of the specific file type. The actual files contained within this archive are the files of this type that have sequence numbers between 0001 and 0100.

  - **Small** This archive contains 1,000 examples of the specific file type. The actual files contained within this archive are the files of this type that have sequence numbers between 0001 and 1,000.

  - **Total** This archive contains all the examples of the specific file type present in the *NapierOne* data set.

- These individual archives have also been combined into three larger archives as described below:

  - **NapierOne-Tiny** This archive contains a copy of all the *Tiny* archives mentioned in the bullet point above.

Table 5. NapierOne Mixed File Data Set

File Type	Size (KB)	Files #	Min Size(KB)	Max Size(KB)	Average Size(KB)	σ	Sub classifications	SubSub Classifications	A	B	C	D	E	F	G	H	I	J	K
APK	9,310	5,000	14	3,000	2,001	778	APK	-	■					■			■		
BIN	3,346	5,000	10	164,411	669	5,512	BIN	-							■	■	■		
BMP	208,000	5,000	18,099	48,306	44,845	5,711	BMP	-		■		■		■					■
	1,750	5,000	1	1,049	375	301	BMP-FROM-WEB	-		■		■		■					■
CSS	355	5,000	1	4,698	74	234	CSS	-				■	■					■	
CSV	3,503	5,000	1	192,472	701	7,637	CSV	-				■	■	■					■
DLL	3,130	5,000	1	155,445	673	4,219	DLL	-	■								■	■	■
DOC	1,320	5,000	20	13,822	290	763	DOC	-	■		■	■	■	■		■	■		■
	1,320	5,000	20	13,822	282	746	DOC-NOMAGIC		■		■	■	■	■		■	■		■
	1,170	5,000	23	13,411	251	649	DOC-PASSWORD	-	■		■	■	■	■		■	■		■
DOCX	1,420	5,000	11	14,580	303	866	DOCX	-	■		■	■	■	■		■	■		■
	1,420	5,000	11	14,580	303	866	DOCX-NOMAGIC	-	■		■	■	■	■		■	■		■
	1,590	5,000	18	14,707	311	872	DOCX-PASSWORD	-	■		■	■	■	■		■	■		■
DWG	9,720	5,000	1,023	6,447	2,088	714	DWG	-				■					■		
	1,630	5,000	5	1,049	351	354	DWG-FROM-WEB	-				■					■		
ELF	1,560	5,000	1	130,513	312	3,064	ELF	-	■								■	■	
EPS	26,400	5,000	2,293	6,121	5,683	724	EPS	-						■			■		■
	2,290	5,000	1	1,049	493	388	EPS-FROM-WEB	-						■			■		■
EPUB	1,170	5,000	5	11,741	251	318	EPUB	-				■	■	■			■		
EXE	8,950	5,000	1	243,770	1,924	10,685	EXE	-	■					■			■	■	
GIF	3,030	5,000	108	1,172	650	141	GIF	-		■		■	■	■		■	■		■
	913	5,000	1	1,049	192	312	GIF-FROM-WEB	-		■		■	■	■		■	■		■
GZIP	18,200	5,000	6	59,749	3,909	4,858	GZIP	-	■			■		■	■		■		
HTML	303	5,000	1	6,804	64	174	HTML	-			■	■					■		
ICS	16	5,000	1	6,785	17	3	ICS	-					■	■			■		
JAVASCRIPT	86	5,000	1	6,785	17	130	JAVASCRIPT	-	■			■	■				■		■
JPG	345	5,000	1	7,445	72	318	JPG-FROM-WEB	-	■	■		■	■	■	■	■	■		■
	666	5,000	47	423	138	36	QUALITY-01-PERCENT	-	■	■		■	■	■	■	■	■		■
	3,220	5,000	79	2,012	692	308	QUALITY-25-PERCENT	-	■	■		■	■	■	■	■	■		■
	5,270	5,000	129	2,999	1,131	448	QUALITY-50-PERCENT	-	■	■		■	■	■	■	■	■		■
	8,380	5,000	285	4,479	1,799	594	QUALITY-75-PERCENT	-	■	■		■	■	■	■	■	■		■
	49,100	5,000	1,797	22,418	10,564	2,533	QUALITY-100-PERCENT	-	■	■		■	■	■	■	■	■		■
JSON	202	5,000	1	70,217	137	2,439	JSON					■					■	■	
MKV	7,530	5,000	81	6,830	1,616	906	MKV	-							■		■		
MP3	4,650	5,000	415	1,206	995	230	MP3	-				■	■	■			■		
MP4	7,540	5,000	81	6,933	1,619	906	MP4	-				■	■	■			■		
	7,110	5,000	16	61,335	1,528	2,250	MP4-FROM-WEB	-				■	■	■			■		
ODS	1,500	5,000	9	48,620	322	1,513	ODS	-	■			■					■		
OXPS	1,320	5,000	20	13,822	282	746	OXPS		■								■		
PDF	4,510	5,000	3	56,986	968	2,799	PDF	-	■		■	■	■	■	■	■	■		■
	4,510	5,000	3	56,986	968	2,799	PDF-NOMAGIC	-	■		■	■	■	■	■	■	■		■
	4,220	5,000	2	56,903	906	2,724	PDF-PASSWORD	-	■		■	■	■	■	■	■	■		■
PNG	175,000	5,000	10,941	47,713	37,792	5,664	LOSSLESS	COMPRESSION-0	■	■		■	■					■	■
	93,900	5,000	4,155	32,767	20,181	3,792		COMPRESSION-3	■	■		■	■					■	■
	93,200	5,000	3,766	35,879	20,021	3,942		COMPRESSION-5	■	■		■	■					■	■
	90,800	5,000	3,585	35,731	19,517	3,986		COMPRESSION-7	■	■		■	■					■	■
	90,000	5,000	3,428	35,731	19,517	3,986		COMPRESSION-9	■	■		■	■					■	■
	1,560	5,000	1	1,049	274	327	PNG-FROM-WEB	-	■	■		■	■					■	■
PPT	5,050	2,105	15	20,087	2,576	3,056	PPT	-				■	■	■		■	■		■
	5,050	2,105	15	20,057	2,576	3,056	PPT-NOMAGIC	-				■	■	■		■	■		■
	4,820	2,105	42	20,009	2,460	2,955	PPT-PASSWORD	-				■	■	■		■	■		■
PPTX	10,300	4,215	39	118,293	2,634	4,005	PPTX	-				■	■	■		■	■		■
	10,300	4,215	39	118,293	2,634	4,005	PPTX-NOMAGIC	-				■	■	■		■	■		■
	10,100	4,215	46	119,254	2,576	3,993	PPTX-PASSWORD	-				■	■	■		■	■		■
Powershell	20	5,000	1	459	4	12	PS		■			■	■	■			■		■
RANDOM	17	5,000	1	2	1	0	PSEUDO	-								■			
	17	5,000	1	640	1	9	PURE	-								■			
RAR	17,600	5,000	6	59,321	3,793	4,750	RAR	-	■						■	■	■		■
SVG	122,000	5,000	4,860	48,285	26,379	5,385	SVG	-		■	■	■	■	■				■	
	401	5,000	1	1,049	78	164	SVG-FROM-WEB	-		■	■	■	■	■				■	
TAR	20,500	5,000	20	61,542	4,421	5,339	TAR	-							■		■		
TIF	127,000	5,000	4,149	42,664	22,900	5,019	TIF	-				■		■			■		■
	8,490	5,000	470	3,423	1,823	466	TIF-RESIZED	-				■		■			■		■
TXT	1	5,000	1	231,094	5,109	28,617	TXT	-	■		■	■	■				■	■	■
WEBP	348	5,000	1	2,4605	1,218	283	WEBP-FROM-WEB	-		■						■			
	5,650	5,000	344	2,460	1,214	281	LOSSLESS	COMPRESSION-0		■						■			
	5,540	5,000	336	2,450	1,188	280		COMPRESSION-2		■						■			
	5,540	5,000	327	2,450	1,188	281		COMPRESSION-4		■						■			
	5,490	5,000	327	2,450	1,179	280		COMPRESSION-6		■						■			
	555	5,000	9	331	114	56	50% QUALITY	COMPRESSION-0		■						■			
	511	5,000	8	333	105	55		COMPRESSION-2		■						■			
	451	5,000	6	316	93	52		COMPRESSION-4		■						■			
	423	5,000	6	310	87	50		COMPRESSION-6		■						■			
XLS	2,370	5,000	5	55,651	509	1,632	XLS	-				■	■	■		■	■		■
	2,370	5,000	5	55,651	509	1,632	XLS-NOMAGIC	-				■	■	■		■	■		■
	2,420	5,000	20	55,620	518	1,646	XLS-PASSWORD	-				■	■	■		■	■		■
XLSX	2,660	5,000	8	93,930	570	2,769	XLSX	-				■	■	■		■	■		■
	2,660	5,000	8	93,930	570	2,769	XLSX-NOMAGIC	-				■	■	■		■	■		■
	2,700	5,000	15	96,190	578	2,815	XLSX-PASSWORD	-				■	■	■		■	■		■
XML	320	5,000	1	16,920	75	708	XML	-	■			■	■				■	■	■
ZIP	18,200	5,000	7	59,865	3,912	4,852	BZIP2	-	■			■		■	■		■		■
	18,200	5,000	6	59,762	3,907	4,855	DEFLATE	-	■			■		■	■		■		■
	17,600	5,000	6	59,352	3,790	4,740	LZMA	-	■			■		■	■		■		■
	18,000	5,000	6	60,660	3,878	4,831	PPMD	-	■			■		■	■		■		■
	18,100	5,000	6	59,695	3,896	4,845	HIGH-COMPRESSION	-	■			■		■	■		■		■
	18,200	5,000	7	59,762	3,907	4,855	ENCRYPTED	-	■			■		■	■		■		■
ZLIB	18,200	5,000	6	59,762	3,912	4,860	ZLIB	-	■			■		■	■		■		■
7Zip	18,200	5,000	7	59,864	3,915	4,852	BZIP2	-	■					■		■	■		■
	17,600	5,000	6	59,351	3,785	4,738	LZMA	-	■					■		■	■		■
	17,500	5,000	6	59,015	3,776	4,724	LZMA2	-	■					■		■	■		■
	18,000	5,000	6	60,626	3,881	4,830	PPMD	-	■					■		■	■		■
	17,500	5,000	6	58,790	3,775	4,721	HIGH-COMPRESSION	-	■					■		■	■		■
	17,500	5,000	6	59,016	3,776	4,724	ENCRYPTED	-	■					■		■	■		■
EXTRA	4,410	5,000	7	32,156	945	2,086	RANSOMWARE	NOTPETYA								■			
	3,740	5,000	1	32,156	802	1,951		SODINOKIBI								■			
	3,730	5,000	1	32,156	801	1,951		MAZE								■			
	4,190	5,000	1	44,381	954	2,283		PHOBOS								■			
	3,740	5,000	1	32,156	802	1,951		NETWALKER								■			
	3,902	5,000	1	32,257	804	1,951		DHARMA								■			
	3,994	5,000	1	31,580	802	1,951		RYUK								■			

– **NapierOne-Small** This archive contains a copy of all the *Small* archives mentioned in the bullet point above.

– **NapierOne-Total** This archive contains a copy of all the *Total* archives mentioned in the bullet point above. This archive is very large as it contains all the data from the *NapierOne* data set.

Each data set archive is accompanied by documentation providing the user with a detailed description of that particular archive including attribution and licensing instructions. Using this structured approach, researchers will be able to tune, what and how much they download, to their specific requirements. Password protected versions of document and archive files all use the same password, *napierone*.

When accessing the website some simple registration may be required. This was implemented to allow the authors to gather some basic usage information regarding the data set. Information gathered during the registration will be handled in accordance with GDPR regulations [60]. The anonymised information gathered during the registration will allow the authors to monitor the use, application and adoption of the data set.

6. Conclusion

Inspired by the work performed by Garfinkel et. al.[30], this paper describes the design, development and distribution of the *NapierOne* [19] mixed file data set. This data set creation was guided by the approaches and techniques used in the development of the Govdocs1 [30] and attempts to complement the original data set with examples of file types that have appeared, or increased in popularity, since its original release. There is no duplication of data between to original Govdocs1 and *NapierOne* data sets.

The data set attempts to address a reoccurring theme in the literature where it has been shown that there exists a lack of useful data sets available for researchers and how the default behaviour of many researchers is to generate their own data sets and not distribute them. A consequence of this behaviour is that research validation and reproducibility is impacted. Many papers highlight the opinion that increased accessibility to high-quality data sets would undoubtedly improve the speed and direction of research and in so doing facilitate high-quality research, that is easier to peer review, producing more robust solutions to the issues being investigated.

The paper describes how the specific file types within this data set were selected, how examples were sourced and how researchers are able to gain free, unlimited access to the data. Actions performed to ensure that the data contained within the data set is of the highest quality are also described. The breadth and depth of the *NapierOne* data set should also remove the need for other researchers to complement the data set with additional files as has previously been the case [47, 48, 51].

The *NapierOne* data set is an ongoing project and the authors intend to employ the developed data set as part of their research in to ransomware detection techniques. Researchers are strongly encouraged to leverage the data set in their own research. The authors are also open to any comments or discussion concerning any aspect of this data set.

7. Postscript

The authors consider NapierOne data set to be a starting point data set and act as a foundation for the creation of a much larger corpus.

The authors have attempted to validate the content of the data set to the best of their ability. This included visual inspection of image files, submission to malware scanning services, local virus checking as well as programmatically checking the format of each file to confirm that it complies with that specific format and can be opened by a program designed to handle that format.

It is hoped that over time other researchers will provide additional data sets that can be incorporated into NapierOne allowing it to grow and develop in the spirit of Open Source.

References

[1] , . File Formats Assessments - wiki.dpconline.org. URL: https://wiki.dpconline.org/index.php?title=File_Formats_Assessments, (Last Accessed: 2021-04-21).

[2] , . Format Description Categories - Sustainability of Digital Formats | Library of Congress. URL: https://www.loc.gov/preservation/digital/formats/fdd/descriptions.shtml, (Last Accessed: 2021-04-21).

[3] , . Recommended Preservation Formats for Electronic Records | Smithsonian Institution Archives. URL: https://siarchives.si.edu/what-we-do/digital-curation/recommended-preservation-formats-electronic-records, (Last Accessed: 2021-04-21).

[4] Abt, S., Baier, H., 2016. Are We Missing Labels? A Study of the Availability of Ground-Truth in Network Security Research. Proceedings - 3rd International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, BADGERS 2014 , 40–55doi:10.1109/BADGERS.2014.11.

[5] Al-rimy, B.A.S., Maarof, M.A., Shaid, S.Z.M., 2018a. Ransomware threat success factors, taxonomy, and countermeasures: A survey and research directions. Computers and Security 74, 144–166. URL: https://doi.org/10.1016/j.cose.2018.01.001, doi:10.1016/j.cose.2018.01.001.

[6] Al-rimy, B.A.S., Maarof, M.A., Shaid, S.Z.M., 2018b. Ransomware threat success factors, taxonomy, and countermeasures: A survey and research directions. Computers and Security 74, 144–166. URL: https://doi.org/10.1016/j.cose.2018.01.001, doi:10.1016/j.cose.2018.01.001.

[7] Antonio Franco, . segmentation-fault/random-iCal-events-generator: Generates random events in iCal ics format. URL: https://github.com/segmentation-fault/random-iCal-events-generator.

[8] Baggili, I., Breitinger, F., 2015. Data sources for advancing cyber forensics: What the social world has to offer. AAAI Spring Symposium - Technical Report SS, 6–9.

[9] Berrueta, E., Morato, D., Magana, E., Izal, M., 2019. A Survey on Detection Techniques for Cryptographic Ransomware. IEEE Access 7, 144925–144944. doi:10.1109/ACCESS.2019.2945839.

[10] Calderdale, . Dataset Search - Calderdale Data Works. URL: https://dataworks.calderdale.gov.uk/dataset?format=spreadsheet&page=10, (Last Accessed: 2021-03-30).

[11] Carreira, J., Noland, E., Hillier, C., Zisserman, A., 2020. A short note on the kinetics-700 human action dataset. arXiv arXiv:2010.10864.

[12] Chang, D., Singh, M., Verma, R., 2015. A Collision Attack on Sdhash Similarity Hashing, in: Proceedings of 10th Intl. Conference on Systematic Approaches to Digital Forensic Engineering., pp. 36–48. URL: http://www.sadfe.org/pastsadfe/sadfe2015/layoutimg/SADFE-2015-Proceedings.pdf, doi:ISBN:978-84-608-2068-0.

[13] Ckan, . Welcome - data.gov.uk. URL: https://ckan.publishing.service.gov.uk/, (Last Accessed: 2021-03-30).

[14] Cleary, G., 2018. Digital Evidence Detection using Bytewise Approximate Matching Gabrielle Cleary RD5 Report Edinburgh Napier University School of Computing. Technical Report January. Edinburgh Napier University.

[15] CommonCrawl, 2021. CommonCrawl. URL: https://commoncrawl.github.io/cc-crawl-statistics/plots/mimetypes, (Last Accessed: 2021-04-08).

[16] Dang-Nguyen, D.T., Pasquini, C., Conotter, V., Boato, G., 2015. RAISE - A raw images dataset for digital image forensics. Proceedings of the 6th ACM Multimedia Systems Conference, MM-Sys 2015 , 219–224URL: https://doi.org/10.1145/2713168.2713194, doi:10.1145/2713168.2713194.

[17] Dargahi, T., Dehghantanha, A., Bahrami, P.N., Conti, M., Bianchi, G., Benedetto, L., 2019. A Cyber-Kill-Chain based taxonomy of crypto-ransomware features. Journal of Computer Virology and Hacking Techniques 15, 277–305. URL: https://doi.org/10.1007/s11416-019-00338-7, doi:10.1007/s11416-019-00338-7.

[18] Data.gov.uk, . Find open data - data.gov.uk. URL: https://data.gov.uk/, (Last Accessed: 2021-03-30).

[19] Davies, S.R., Macfarlane, R., Buchanan, W.J., 2021. NapierOne. URL: www.napierone.com.

[20] De Gaspari, F., Hitaj, D., Pagnotta, G., De Carli, L., Mancini, L.V., 2020. EnCoD: Distinguishing Compressed and Encrypted File Fragments, in: 14th International Conference, Network and System Security, pp. 1–19. URL: https://dx.doi.org/10.1007/978-3-030-65745-1_3, doi:10.1007/978-3-030-65745-1_3, arXiv:2010.07754.

[21] Deepmind, 2017. Kinetics Dataset. URL: https://deepmind.com/research/open-source/kinetics.

[22] Defferrard, M., Benzi, K., Vandergheynst, P., Bresson, X., 2017. FMA: A dataset for music analysis. Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017 , 316–323URL: https://arxiv.org/abs/1612.01840, arXiv:1612.01840.

[23] Digital Corpora, 2014. Digital Corpora ż Govdocs1. URL: https://digitalcorpora.org/corpora/files, (Last Accessed: 2021-03-22).

[24] File-Extension, 2021. File-Extension.org - Solve your problems with file extensions. URL: https://www.file-extension.org/, (Last Accessed: 2021-03-24).

[25] File-extensions.org, 2021. https://www.file-extension.info/top. URL: https://www.file-extension.info/top, (Last Accessed: 2021-04-09).

[26] FileInfo, . FileInfo - The File Information Database. URL: https://fileinfo.com/, (Last Accessed: 2021-03-24).

[27] Fitzgerald, S., Mathews, G., Morris, C., Zhulyn, O., 2012. Using NLP techniques for file fragment classification. Digital Investigation 9, 44–49. URL: https://doi.org/10.1016/j.diin.2012.05.008, doi:10.1016/j.diin.2012.05.008.

[28] Freed, N., Borenstein, N., 1996. Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types. Technical Report. Internet Society. URL: https://www.rfc-editor.org/info/rfc2046https://www.doi.org/10.17487/RFC2046, doi:10.17487/rfc2046.

[29] Freed, N., Klensin, J., Hansen, T., 2013. Media Type Specifications and Registration Procedures. Technical Report. Internet Society. URL: https://www.rfc-editor.org/info/rfc6838https://www.doi.org/10.17487/RFC6838, doi:10.17487/rfc6838.

[30] Garfinkel, S., Farrell, P., Roussev, V., Dinolt, G., 2009. Bringing science to digital forensics with standardized forensic corpora. Digital Investigation 6, 2–11. URL: http://dx.doi.org/10.1016/j.diin.2009.06.016, doi:10.1016/j.diin.2009.06.016.

[31] Garfinkel, S.L., 2009. Automating disk forensic processing with SleuthKit, XML and python. 4th International Workshop on Systematic Approaches to Digital Forensic Engineering, SADFE 2009 , 73–84doi:10.1109/SADFE.2009.12.

[32] Garfinkel, S.L., 2010. Digital forensics research: The next 10 years. Digital Investigation 7, 64–73. URL: http://dx.doi.org/10.1016/j.diin.2010.05.009, doi:10.1016/j.diin.2010.05.009.

[33] Garfinkel, S.L., 2020. Govdocs1. URL: http://downloads.digitalcorpora.org/corpora/files/govdocs1/, (Last Accessed: 2020-09-14).

[34] Gitumarkk, 2019. kinetics-datasets-downloader. URL: https://github.com/dancelogue/kinetics-datasets-downloader.

[35] Google, 2015. File types indexable by Google. URL: https://support.google.com/webmasters/answer/35287?hl=en, (Last Accessed: 2021-03-24).

[36] Google, 2016. Google Search Appliance Adaptors. URL: http://googlegsa.github.io/adaptor/index.html.

[37] Grajeda, C., Breitinger, F., Baggili, I., 2017. Availability of datasets for digital forensics And what is missing. Digital Investigation 22, S94–S105. URL: https://dx.doi.org/10.1016/j.diin.2017.06.004, doi:10.1016/j.diin.2017.06.004.

[38] Haar, M., 1998. RANDOM.ORG - True Random Number Service. URL: https://www.random.org/, (Last Accessed: 2020-10-14).

[39] Hart, M., 2021. Free ebooks - Project Gutenberg. URL: https://www.gutenberg.org/, (Last Accessed: 2021-03-22).

[40] Hiestand, R., 2021. Advanced Batch Image Compression. URL: http://abic.sourceforge.net/, (Last Accessed: 2021-04-09).

[41] Jackson, A.N., 2012. Formats over Time: Exploring UK Web History. iPRES 2012: Proceedings of the 9th International Conference on Preservation of Digital Objects URL: http://arxiv.org/abs/1210.1714, arXiv:arXiv:1210.1714v1.

[42] Jung, S., Won, Y., 2018. Ransomware detection method based on context-aware entropy analysis. Soft Computing 22, 6731–6740. URL: https://doi.org/10.1007/s00500-018-3257-z, doi:10.1007/s00500-018-3257-z.

[43] Kharraz, A., Arshad, S., Mulliner, C., Robertson, W., Kirda, E., 2016. UNVEIL: A Large-Scale, Automated Approach to Detecting Ransomware, in: Usenix 25th USENIX Security Symposium, Austin, TX. pp. 757–772. URL: https://doi.ieeecomputersociety.org/10.1109/SANER.2017.7884603, doi:10.1109/SANER.2017.7884603.

[44] Kolodenker, E., Koch, W., Stringhini, G., Egele, M., 2017. PayBreak : Defense against cryptographic ransomware. ASIA CCS 2017 - Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security , 599–611URL: 10.1145/3052973.3053035, doi:10.1145/3052973.3053035.

[45] Maigida, A.M., Abdulhamid, S.M., Olalere, M., Alhassan, J.K., Chiroma, H., Dada, E.G., 2019. Systematic literature review and metadata analysis of ransomware attacks and detection mechanisms. Journal of Reliable Intelligent Environments 5, 67–89. URL: https://doi.org/10.1007/s40860-019-00080-3, doi:10.1007/s40860-019-00080-3.

[46] McCarrin, M.R., 2013. Exploration and validation of the sdhash parameter space. Ph.D. thesis. Naval Postgraduate School, Monterey, California. URL: http://calhoun.nps.edu/handle/10945/34704http://hdl.handle.net/10945/34704.

[47] Nguyen, K., Tran, D., Ma, W., Sharma, D., 2014. A proposed approach to compound file fragment identification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8792, 493–500. URL: https://doi.org/10.1007/978-3-319-11698-3_38, doi:10.1007/978-3-319-11698-3_38.

[48] Penrose, P., MacFarlane, R., Buchanan, W.J., 2013. Approaches to the classification of high entropy file fragments. Digital Investigation 10, 372–384. URL: http://dx.doi.org/10.1016/j.diin.2013.08.004, doi:10.1016/j.diin.2013.08.004.

[49] Pont, J., 2020. stats-tools-research/README.md at master · anti-ransomware/stats-tools-research · GitHub. URL: https://github.com/anti-ransomware/stats-tools-research/blob/master/README.md, (Last Accessed: 2020-11-10).

[50] Pont, J., Abu Oun, O., Brierley, C., Arief, B., Hernandez-Castro, J., 2019. A Roadmap for Improving the Impact of Anti-ransomware Research, in: Nordic Conference on Secure IT Systems, Springer Inter-

national Publishing. pp. 137–154. URL: http://dx.doi.org/10.1007/978-3-030-35055-0_9, doi:10.1007/978-3-030-35055-0-9.

[51] Pont, J., Hernandez-Castro, J., 2020. Why Current Statistical Approaches to Ransomware Detection Fail, in: International Conference on Information Security, 23rd Information Security Conference,. pp. 199–218. URL: https://dx.doi.org/10.1007/978-3-030-62974-8_12, doi:10.1007/978-3-030-62974-8_12.

[52] Roussev, V., 2011a. An evaluation of forensic similarity hashes. Digital Investigation 8, S34–S41. URL: http://dx.doi.org/10.1016/j.diin.2011.05.005, doi:10.1016/j.diin.2011.05.005.

[53] Roussev, V., 2011b. the t5 corpus. URL: http://roussev.net/t5/t5.html, (Last Accessed: 2020-11-10).

[54] Roussev, V., Garfinkel, S.L., 2009. File fragment classification - The case for specialized approaches. 4th International Workshop on Systematic Approaches to Digital Forensic Engineering, SADFE 2009 , 3–14URL: https://dx.doi.org/10.1109/SADFE.2009.21, doi:10.1109/SADFE.2009.21.

[55] Roussev, V., Quates, C., Martell, R., 2013. Real-time digital forensics and triage. Digital Investigation 10, 158–167. URL: http://dx.doi.org/10.1016/j.diin.2013.02.001, doi:10.1016/j.diin.2013.02.001.

[56] Ryan, H., 2014. EVALUATING FILE FORMAT ENDANGERMENT LEVELS AND FACTORS. Ph.D. thesis. University of North Carolina. URL: https://doi.org/10.17615/fhkk-dv90.

[57] Scaife, N., Carter, H., Traynor, P., Butler, K.R., 2016. CryptoLock (and Drop It): Stopping Ransomware Attacks on User Data. 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS) , 303–312URL: https://doi.org/10.1109/ICDCS.2016.46, doi:10.1109/ICDCS.2016.46.

[58] Sharpless, K.E., Avila, R.L., Bruce, S.S., Chang, W.L., Covahey, V., Dimeo, R.M., Dohne, A.K., Evans, H.M., Fein, A.P., Kimball, D.J., Medina-Smith, A.N., Munter, A.E., Olthoff, J.K., Poster, D.L., Roberts, K.M., Schiller, S.B., Scott, J.H.J., Silcox, B.P., St. Pierre, J.A., Stiles, M.D., 2015. The NIST Plan for Providing Public Access to Results of Federally Funded Research. Technical Report. National Institute of Standards and Technology. Gaithersburg, MD. URL: https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8084.pdf, doi:10.6028/NIST.IR.8084.

[59] The National Archives, 2021. Open Government Licence. URL: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/, (Last Accessed: 2021-05-04).

[60] Union, E., 2017. REGULATION(EU)2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. URL: https://advisera.com/eugdpracademy/gdpr/, doi:10.1515/9783866537118.131.

[61] Université du Luxembourg, 2021. AndroZoo. URL: https://androzoo.uni.lu/, (Last Accessed: 2021-04-08).

[62] University of New Haven, 2021. Datasets For Cyber Forensics. URL: https://datasets.fbreitinger.de/, (Last Accessed: 2021-04-07).

[63] VirusTotal, 2019. Statistics - VirusTotal. URL: https://www.virustotal.com/en/statistics/, (Last Accessed: 2021-03-24).

[64] W3Techs, 2021. Usage Statistics of Image File Formats for Websites, November 2014. URL: http://w3techs.com/technologies/overview/image_format/all, (Last Accessed: 2021-03-24).

[65] Wayback Machine, 2021. Internet Archive Wayback Machine. URL: https://archive.org/web/, (Last Accessed: 2021-03-24).