THE JOURNAL OF
DIGITAL FORENSICS,
SECURITY AND LAW

Journal of Digital Forensics,
Security and Law

Volume 17

Article 5

3-7-2022

The Digital Detective's Discourse - A toolset for forensically sound collaborative dark web content annotation and collection

Jesper Bergman
Department of Computer and Systems Sciences, Stockholm University, jesperbe@dsv.su.se

Oliver B. Popov
Department of Computer and Systems Sciences, Stockholm University, popov@dsv.su.se

Recommended Citation

Bergman, Jesper and Popov, Oliver B. (2022) "The Digital Detective's Discourse - A toolset for forensically sound collaborative dark web content annotation and collection," *Journal of Digital Forensics, Security and Law*: Vol. 17 , Article 5.
DOI: https://doi.org/10.15394/jdfsl.2022.1740
Available at: https://commons.erau.edu/jdfsl/vol17/iss1/5

The Digital Detective's Discourse - A toolset for forensically sound collaborative dark web content annotation and collection

Cover Page Footnote

The work is partially supported by the NordForsk Grant No. 80512 for the project "Police Detectives on the TOR-network". Additionally, this project was also a part of the COST network action DigForAsp, project number CA17124.

THE DIGITAL DETECTIVE'S DISCOURSE: A TOOLSET FOR FORENSICALLY SOUND COLLABORATIVE DARK WEB CONTENT ANNOTATION AND COLLECTION

Jesper Bergman, Oliver B. Popov
Stockholm University
Department of Computer and Systems Sciences
Kista, 164 07, Sweden
{jesperbe, popov}@dsv.su.se

ABSTRACT

In the last decade, the proliferation of machine learning (ML) algorithms and their application on big data sets have benefited many researchers and practitioners in different scientific areas. Consequently, the research in cybercrime and digital forensics has relied on ML techniques and methods for analyzing large quantities of data such as text, graphics, images, videos, and network traffic scans to support criminal investigations. Complete and accurate training data sets are indispensable for efficient and effective machine learning models. An essential part of creating complete and accurate data sets is annotating or labelling data. We present a method for law enforcement agency investigators to annotate and store specific dark web content. Using a design science strategy, we design and develop tools to enable and extend web content annotation. The annotation tool was implemented as a plugin for the Tor browser. It can store web content, thus automatically creating a dataset of dark web data pertinent to criminal investigations. Combined with a central storage management server, enabling annotation sharing and collaboration, and a web scraping program, the dataset becomes multifold, dynamic, and extensive while maintaining the forensic soundness of the data saved and transmitted. To manifest our toolset's fitness of purpose, we used our dataset as training data for ML based classification models. A five cross-fold validation technique was used to evaluate the classifiers, which reported an accuracy score of 85 - 96%. In the concluding sections, we discuss the possible use-cases of the proposed method in real-life cybercrime investigations, along with ethical concerns and future extensions.

Keywords: digital forensics, dark web, annotation, cybercrime, Tor

1. INTRODUCTION

Strong encryption algorithms and anonymous routing protocols have turned the Internet into a safe haven for two types of communities on the opposite sides of the societal spectrum. On one side are primarily journalists, whistle-blowers, and free speech activists who challenge diverse undemocratic and hybrid regimes; on the other side, there are miscreants behind a range of illicit activities. The later ones' favorite places are the so-called "dark markets," where the criminals trade mainly with illegal goods. The privilege of anonymity, among other things, has stimulated the growth of cybercrime and cyber-dependent crime which are increasingly becoming difficult to manage and police (identification, investigation, prevention, and eradication) for law enforcement agencies (LEAs) (Europol, 2017). Based on a survey with 270 law enforcement agency respondents from 30 countries, it was reported that 46% of the child sexual abusers were active on the Tor network, while 54% on the clear web in 2018 (Netclean, 2019).

In 2020, a follow up study with 470 respondents from 39 countries reported that the use of darknets in CSA cases had increased "considerably" (17%) or "moderately" (33%) in the year of the Covid-19 pandemic (Netclean, 2021).

Standard methods for investigating cybercrimes are text retrieval and text analysis. Previous studies have successfully deployed these methods for authorship attribution of dark web fora data (Spitters, Klaver, Koot, & van Staalduinen, 2015), prediction of criminal activities on the dark web (Kwon, Priniski, Sakar, Shakarian, & Shakarian, 2017), and their classification (Nunes et al., 2016).

Zhang et al. (2010) proposed a dark web collector that indexed Islamic extremist websites on the dark web and classified them related to the specific content and the extremist group behind it. Kalpakis et al. (2016) presented a clear and dark web crawler capable of identifying home-made explosives using a link-based classification technique. Other studies have provided similar methods of classification of cybercrimes (Dalins, Wilson, & Carman, 2018; Ghosh, Das, Porras, Yegneswaran, & Ghehani, 2017).

Accurate and well-structured data sets are fundamental for correct and reliable text analysis. Data collected in web content needs to be parsed and structured for effective analyses. Bad input equals bad output. In other words, if "your data is bad, the machine learning tools may be rendered useless". The effective and efficient use of supervised learning algorithms on data collected from the dark web requires generating a training data set that fits the purpose and represents the working environment. The precision of classification models is proportional to the reliability of the ground-truth data set.

A tool tailored for manual dark web data collection would fit a three-fold purpose. Namely, it would

- allow its users to manually select and collect both clear and dark web data and automatically structure and archive it in a forensically sound manner.
- aid the creation of a ground-truth training data set of dark web data.
- train classification models to classify previously unseen websites based on the ground-truth data set.

To the best of our knowledge, such a tool does not exist yet. Hence, the research presents a proof-of-concept dark web collection and annotation tool for classifying previously unknown web sites based on the ground-truth data set it creates.

2. RELATED WORK

Hitherto published research articles in digital forensics and cybercrime investigations have focused on text-based classification and cybercrime web content prediction. This section elaborates and extends on the research corpus enumerated in the introductory chapter.

2.1 Dark Web Cybercrime Investigations

The infamous Silk Road market place entry on the Tor network served almost as an inaugural promotion of the dark web becoming first-page news. However, the trade of illicit goods on the clear web has been around for a long time, along with some investigation techniques. While the world wide web and the visible Internet provide LEAs with better traceability than ACNs like Tor and I2P, there are still various challenges, such as cross-border legislation between the internet exchange points (Ghappour, 2017). The inability to collect evidence due to encryption and anonymous communication protocols, despite the legal permissions by the US Federal Bureau of Investigations (FBI), has been termed as the problem of “going dark" (Comey, 2015). One technique of identifying criminals despite the anonymous protocols is automated analysis of textual content and authorship attribution, as presented in the early 2000s by (Qin, Huang, & Chen, 2003) and (de Vel, Anderson, Corney, & Mohay, 2001).

With the enlargement of the user-base and continuous improvement of ACNs, text analysis thrives since textual content is often readily available. A number of studies have successfully applied text analysis to attribute authorship (Spitters et al., 2015; Arabnezhad, La Morgia, Mei, Nemmi, & Stefa, 2020) and automate classification tasks (Tai, Soska, & Christin, 2019; Hayes, Cappa, & Cardon, 2018; Nunes et al., 2016; Sabbah, Selamat, Selamat, Ibrahim, & Fujita, 2016).

In the last several years, a couple of EU financed research projects have delved into the evolving cybercrime threat on the dark web. The most prominent ones are Titanium and Tensor. The former dealt with virtual currencies connected to cybercrime and terrorism (Titaniu, n.d.), while the latter focused on intelligence gathering of terrorism content on both the clear web and the dark web (Tensor, n.d.). Complementary to text analysis and crypto-currency transaction analysis, are (dark) web crawling and scraping which retrieve the foundational raw data.

Cybercrime research on the dark web has proven to be possible and potentially very beneficial. With all the genuine concerns about privacy and hence the quest for anonymity, it has also become indispensable for policing and guarding the inherent creativity of the global Net. Three already cited articles that utilize machine learning algorithms to classify content, also employed techniques for scraping web content from the Tor hidden services (Kwon et al., 2017; Nunes et al., 2016; Dalianis, 2018). Previous research by Popov et al. (2018), where the concept of forensic soundness permeates every step in the digital forensic investigation process, presented a framework for scraping, archiving, and analyzing dark web content. Moreover, this kind of crawler is envisioned as an automated “powerhouse" for the digital investigation that should also in the realm of scraping (identification, acquisition) have the ability for data reduction and hence induce additional efficacy in the processes of archiving and analysis. By using various affordances from artificial intelligence, including machine learning, the system should be able to provide a reconstruction of the digital crime scene, and thus an opportunity for better interpretation of the analysis (Popov et al., 2018).

2.2 Annotation Tools

According to van Beek et al. (2015), possibilities for collaborative annotation are crucial to any criminal investigation to speed up the entire process. Detectives should be able to annotate or tag traces that are interesting or unclear to them. Moreover, the authors argue that “other detectives and digital investigators must have access to the annotation so that they can act on it." (van Beek et al., 2015)

Creating quality data sets is essential for valuable and accurate machine learning models, such as classifiers and predictors, and more sophisticated models based on artificial intelligence. Experts, who are knowledgeable in the domain of the specific data, are central in the production of data sets. Multiple annotators often cross-verify each other’s annotations to ensure the data set’s accuracy and quality. Once the data set is complete, it is referred to as the “gold standard" or the definitive data set regarding the data’s quality and accuracy and its labels. If multiple domain experts are involved and cannot agree on all the annotations, one can create a “silver standard" (Dalianis, 2018). Since annotation is an important step in text analysis, there are many digital annotation tools available.

In 1998, an annotation feature built into Microsoft Word 6.0 was used study on the effects of collaborative annotation interfaces. It was discovered that position of the annotation window on the screen did not affect the time for completing the annotation task (Wojahn, Neuwirth, & Bullock, 1998). Ten years later Farzan and Brusilovsky (2008) developed a web social navigation support enabled web browser annotation tool, arguing that social navigation helps students navigate to relevant educational resources. In traditional, or analogue forensics, Neto, Pinto, Proença, Amorim, and Conde-Sousa (2021) presented a tool for annotating reference DNA sequence data sets in accordance with the forensic standards to reproduce and verify the results.

According to Sorokin and Forsyth (2008), data set annotation can be outsourced with high quality results, using online services such as Amazon Mechanical Turk. Neves and Leser (2014) did a comparative study of 13 free annotation tools for biomedical literature, concluding that there is no one perfect tool for annotating biomedical text; researchers need to find one that fits their specific task and purpose.

2.3 Text Analysis in Cybercrime Investigations

Text analysis has been a broad interest research topic in cyber security for over ten years. Chen et al. (2008) analyzed 94.326 clear web pages’ textual content and classified it based on the type of jihad extremism. There are newer studies that employ the same techniques to identify criminal activity and criminal adversaries.

Nunes et al. (2016) led similar research in which web pages were fetched from ten dark marketplaces and then parsed to be used as training data for classification models. By using a support vector machine (SVM) classifier, the authors classified drug ads in the correct category with a precision and recall of 85% and 87%, respectively.

Portnoff et al. (2017) developed a tool for automatic analysis of criminal activity on dark markets - both on anonymity networks (or darknets) and the world wide web (or clearnets). Using eight different data sets with web content involving criminal activity, the researchers built three different classifiers from these data sets. The first classifier identified whether or not the forum post is related to buying or selling, the second classifier for recognizing the illegal product in question, and the final classifier for determining the price of the product. The authors concluded that the SVM was the most accurate algorithm for all three tasks.

Spitters et al. (2015) successfully attributed textual drug trafficking-related content retrieved from the Tor web to authors using an SVM based text analysis method with an accuracy of 88% based on 177 users' posts on a no longer available Tor .onion service forum called “Black Market Reloaded". The authors also managed to classify aliases of two different user accounts with a precision and recall of 91% and 25%, respectively, for 177 of the user accounts examined.

Kwon et al. (2017) collected data from a dark web forum called “X," where discussions of different dark marketplaces (and similar topics) took place before the some of the marketplaces were closed down. The authors carried out a thorough content analysis of communication that had taken place on “X" when law enforcement agencies seized the dark marketplaces Utopia and SilkRoad 2.0. The authors used a CIPS-based analysis and concluded that dark marketplace “crises" could be an opportunity to induce distrust among cyber criminals.

3. RESEARCH METHOD

A design science research methodology was chosen to solve the research problem. There are many instances and variations of design science model, which is basically a modification of the water-flow paradigm. Specifically, the design research primer by Johannesson and Perjons (2014) was deemed the most suitable for this research study due to its flexibility and comprehensiveness. The method applied consists of five activities (or stages) such as (1) Problem Explication, (2) Requirements Elicitation, (3) Artifact Development, (4) Artifact Demonstration, and (5) Artifact Evaluation. The activity Problem Explication is equivalent to the research problem definition placed in the Introduction section. The activity Artifact Demonstration is presented under the Artifact Demonstration section as well as in the Results section. The remaining three stages, Requirements Elicitation, Artifact Development, and Artifact Evaluation are given in the remaining subsections of this section under each respective heading.

3.1 Requirements Elicitation

Requirements for the artifact were elicited from the literature review combined with the own experience and knowledge by the researchers. Considering that no previous research has examined the annotation phenomena in dark web content, the requirements were extracted from research related to the utilization of various annotation tools in other scientific areas. Hence, the requirements stipulated for designing the artifact presented in the paper are:

1. RQ1: Forensically sound annotation and preservation of web content relevant to a cybercrime investigation.
2. RQ2: Automatically archive and preserve an annotated web page and its metadata.
3. RQ3: Automatically synchronize data to a central server.
4. RQ4: Enable for non-technical investigators to use the annotation tool.
5. RQ5: Support for multiple annotators.
6. RQ6: Web browser based monitoring panel for data visualization.
7. RQ8: Inter-annotator or consensus agreement calculation of annotations.

8. RQ9: Intelligent classification models based on annotations.

The requirements RQ1, RQ2, and RQ3 stem from the two general principles of forensics, such as preserving the integrity of digital evidence and the sustainability of the chain of custody that states that every interaction with the evidence should be monitored and documented. Previous research that necessitates this requirement include (McKemmish, 2008) and (Casey, 2011). Furthermore, the requirements RQ6 and RQ8 are motivated by the previously developed annotation tools in (Wojahn et al., 1998) and (Neto et al., 2021). Requirement RQ5 reflects the fundamental design principle of Hansken - a centralized, collaborative digital forensic service engine presented in (van Beek et al., 2015) and (van Baar, van Beek, & van Eijk, 2014) and used by the Dutch Police (Hansken.nl, 2020). The requirements RQ5, RQ7, and RQ9 are based on the experience and cognizance of the authors, cross-fertilized with the continuous formal and informal discourse with the investigators from the Swedish Police.

3.2 Artifact Development

Several items comprise the artifact, termed as “The Digital Detective’s Discourse" (with the acronym D3). The source code can be found on the researchers Git repository[1] and the topology is found in Figure 1:

1. **D3-Annotator**: An annotation (including categorization and text excerpting) program written in Javascript for Firefox and the Tor Browser.

2. **D3-Centraliser**: A central NodeJS and SQLite3 database management program that uses representational state transfer (REST) API to receive and transmit data.

3. **D3-Collector**: A fetch-and-archive script that downloads .onion web pages as soon as the URL is received by D3-Centraliser and stores them and their metadata in the database. Written Python, uses Shell script to create a Tor socket.

4. **D3-Analyser**: Two analysis scripts that (1) classifies web content using machine learning algorithms, and (2) calculates the inter-annotator agreement scores between annotators’ categorizations. Written in Python using Scikit-Learn.

5. **D3-Visualiser**: A simple web browser dashboard that displays graphs based on data and statistics from the annotation database. It is written in Python using Flask and Dash.

3.2.1 D3-Annotator

A Firefox add-on was developed using Mozilla Firefox’s software development kit (SDK). The add-on was produced in JavaScript using an SQLite database to store the web content marked and selected for the collection. The add-on was tested in Firefox 89 and Tor Browser 10.5.6 (based on Mozilla Firefox 78.14.0esr) under Arch Linux 5.11.6.

1. Whenever a tab is activated, its URL is stored in the database with the SHA256 sum of the URL as the ID. The URL is temporarily stored in the local storage together with the current timestamp.

2. The user can right click and choose to create a new category to place the active web page in (see Figure 2).

3. The user can right click the web page to save the current URL to any category previously created.

[1] `https://gitea.dsv.su.se/jebe8883/D3`

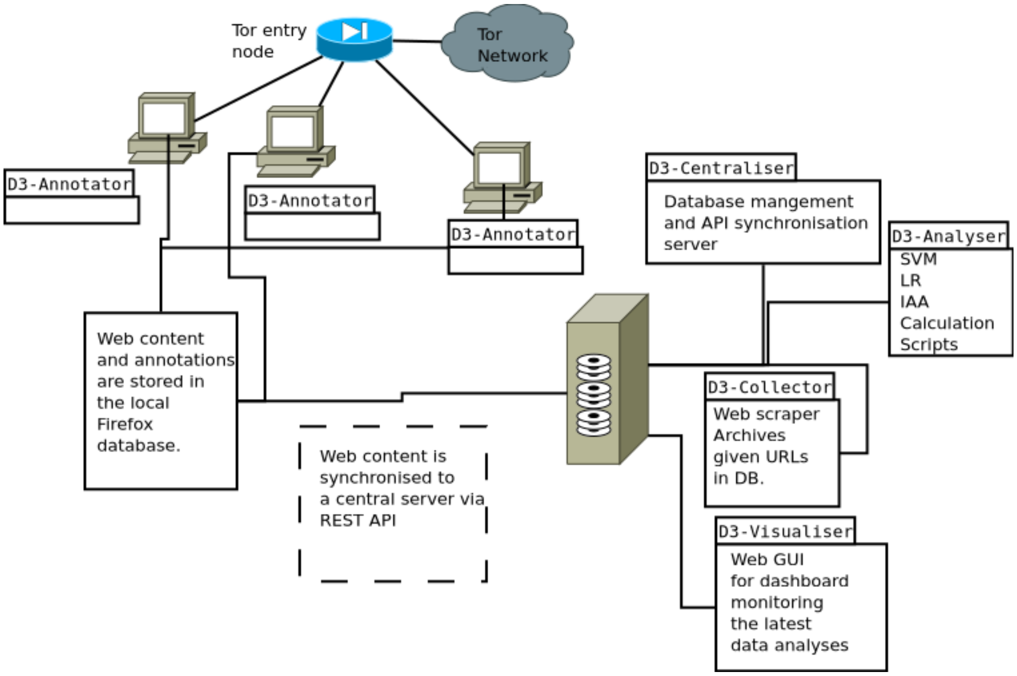
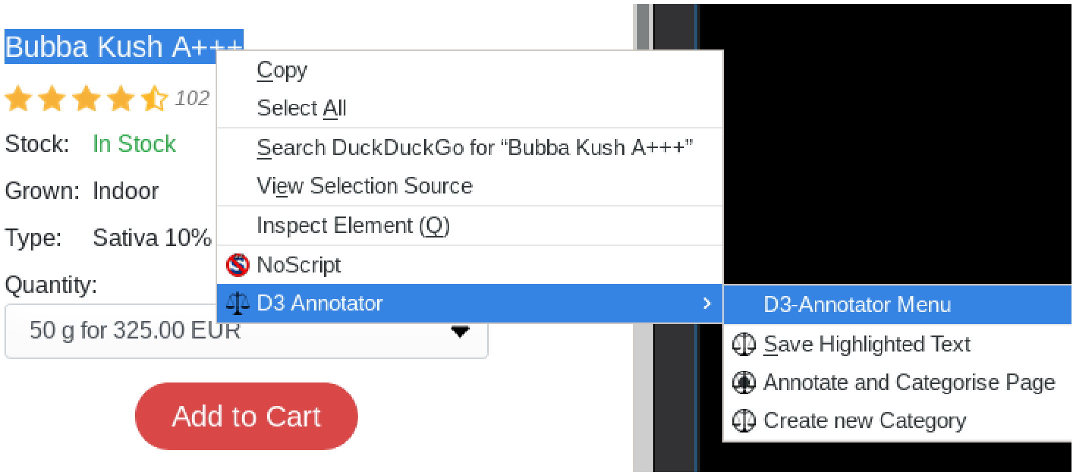

Tor entry node
Tor Network
D3-Annotator
D3-Annotator
D3-Annotator
D3-Centraliser
Database mangement and API synchronisation server
D3-Analyser
SVM
LR
IAA
Calculation Scripts
Web content and annotations are stored in the local Firefox database.
Web content is synchronised to a central server via REST API
D3-Collector
Web scraper Archives given URLs in DB.
D3-Visualiser
Web GUI for dashboard monitoring the latest data analyses

Figure 1. A topology of the D3 toolset for cybercrime web content annotation, archiving, and analysis.

4. The user can also choose to annotate the web page with a free text string, by right clicking “Annotate and Categorize" page.
5. If any text on the active web page is highlighted, it is saved together with the corresponding category and annotation.
6. The user can also choose to just save the highlighted text by clicking “Save Highlighted Text" from the right-click menu.
7. If any of the above options have been clicked, D3-Annotator immediately sends the data to the central server via REST API.

The add-on was developed in JavaScript, and Mozilla's API was used to fulfill all the functional requirements. The exceptions are the crypto and hashing algorithms, which were not implementable since Tor browser's does not include the “service workers" that comprise the crypto library in Firefox and Firefox add-ons (Tor-Project, n.d.). The JavaScript algorithms employed for MD5 and SHA256 hashing were from (Johnston, n.d.) and (Webtoolkit, n.d.) respectively. The MD5 algorithm is prone to hash collision attacks and could lead to two different files ending up with the same MD5 hash sum. However, this only mildly affects the field of digital forensics and hard drive image verification since different files with the same hash on two different drives will still generate different hard drive MD5 sums (Kessler, 2016).

In contrast to hard drives, in this study short strings and short sequences of bytes are received and transmitted over IP. To eliminate any possibility of hash collisions we decided to use the hitherto unbroken hashing algorithm SHA256. With the exception of .onion URLs, only SHA256 was for integrity verification. The reason for using MD5 for calculating the hash sum of .onion URLs was



Bubba Kush A+++

102

Stock: In Stock

Grown: Indoor

Type: Sativa 10%

Quantity:

50 g for 325.00 EUR

Figure 2. The user menu options of D3-Annotator in Tor Browser (v. 10.5.6).

to enable comparison with Ahmia's black list[2] which contains MD5 hashed URLs of child abuse material.

D3-Annotator used Mozilla's built in local storage[3] for temporary storing annotation data until it is synchronized with the D3-Centraliser database server. The data is synchronized in JSON format via HTTP REST API requests, typically constructed as in Listing 1.

```
{/* SHA256 of URL is UID */
601
    F929EE4AE739422270327F12A542922B633
    :
    {url: ``http://loremipsum.onion/
    cat/weed/product/37",
    category: [``dark marketplace"],
    domain: "http://loremipsum.onion",
    md5: ``0260922a36845db58f9a827cf9
    ",
    timestamp:1594732929013,
    annotation:[``New .onion V3 URL
    .","Suspect #g55 relating to case
    UID20210311-113 "],
    highlightedText: [``Please save
    our mirror URLs!!"]
  }
}
```

Listing 1. Example of locally stored web page annotations in D3-Annotator.

The JSON message is then synchronized with the D3-Centraliser server. Once the server has received the JSON message, a SHA256 verification hash sum is sent as a response back with REST API to the D3-Annotator. The code for verifying the integrity of the API request and response conversation is presented in Listing 2.

```
/* Synchronise with remote database
    server. */
async function syncWithServer(
    itemType, itemObject){
     let itemJSON = ``";
     let xhr = new XMLHttpRequest();
     let urlQuery = 'http://SERVER-IP
    :8080/api/command?uuid=' + uuid.
    annotatorUUID;
     xhr.open(``POST", urlQuery);
     let t2 = null;
     let t1 = performance.now();
     let sent_hashsum = ``";
     let received_hashsum = ``";
     let verified_hashsum = 0;

     /* Send the proper header
    information along with the request
     */
```

[2] https://ahmia.fi/banned/

[3] https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/ WebExtensions/API/storage/local

```
    xhr.setRequestHeader(‘‘Content-Type", ‘‘application/json; charset=utf-8");

    xhr.onreadystatechange = function(){ /* Call a function when the state changes. */
        if (this.readyState === XMLHttpRequest.DONE && this.status === 200){
            /* Request finished. Do processing here. */
            t2 = performance.now();
            received_hashsum = xhr.response;
            /* Verify the hash sum of the JSON message */
            if(received_hashsum == sent_hashsum){
                verified_hashsum = 1;
            }
        }
    }
    if(itemType == ‘‘webpage" || itemType == ‘‘annotation"){
        itemJSON = JSON.stringify({‘‘webpage" : itemObject});
    }
    if(itemType == ‘‘category"){
        itemJSON = JSON.stringify({‘‘category": itemObject});
    }

    /* Calculate hash sum before sync */
    sent_hashsum = calculateHashsum(itemJSON);

    /* Sync JSON message with DB */
    xhr.send(itemJSON);

    /* Calculate time */
    let sync_time = t1-t2;
}
```

Listing 2. Code Excerpt of database synchronization and message verification in D3-Annotator.

3.2.2 D3-Centraliser

The D3-Centraliser, as the name implies, is the central server that the D3-Annotator clients connect to to synchronize their annotations. D3-Centraliser was written in JavaScript and runs under NodeJS as a REST API server. The D3-Centraliser writes all synchronized data to an SQLite3 database. When a URL is received, D3-Centraliser calls a Shell script that starts a Torsocks[4] instance in which D3-Collector is run to fetch and archive the web page of that URL in the database.

The reason for selecting SQLite3 as the primary central database for this toolset was its minimalist design. Moreover, it proved to be the appropriate choice for this study since (1) it could handle many concurrent read/write operations per second (SQLite.org, n.d.), and (2) the particular use-case required neither remote access to the database nor multiple user account authentication.

3.2.3 D3-Collector

A web page collector named D3-Collector was created to preserve the forensic soundness and the chain of custody of the content being annotated, categorized, or excerpted using D3-Annotator. D3-Collector is invoked by D3-Centraliser once a new URL is received. D3-Collector fetches the raw web page and archives it in the database table “Raw Page" (see Figure 3). The Python libraries Urllib, BeutifulSoup, and Scrapy were used to to make the HTTP request and retrieve the body HTML, respectively (see code excerpt in Listing 1). To reach .onion sites, the script is run through Torsocks.

```
def download_onion_page(url):
    req = request.Request(url,
        data=None,
        headers={'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; rv:78.0) Gecko/20100101 Firefox/78.0'}
    )
```

[4] https://support.torproject.org/glossary/torsocks/

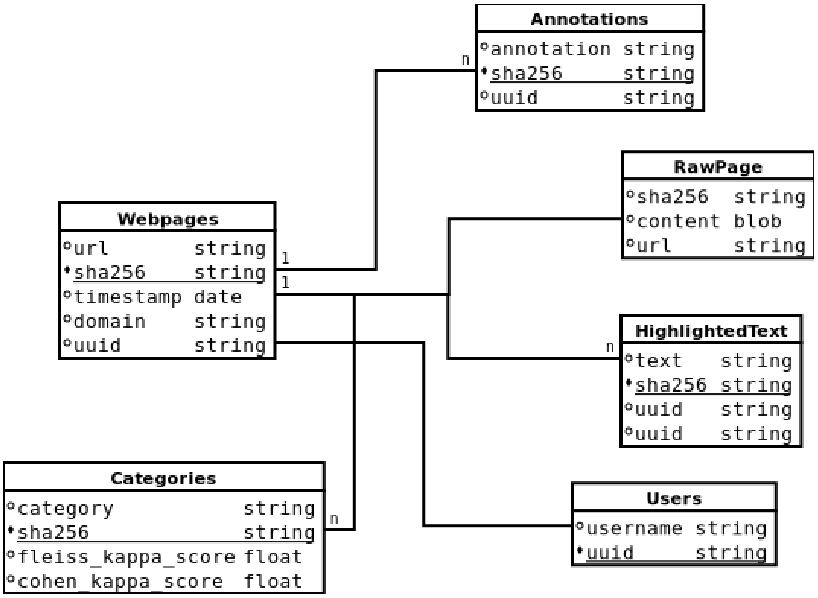


Figure 3. Database schema of the annotation database managed by the module D3-Centraliser.

```
    # Initiate array to store the web page content in.
    onion_page_array = []

    try:
        html = urlopen(req, timeout=300)
        bs = BeautifulSoup(html,'html.parser')
        raw = html.read().decode('utf-8')
        parsed = bs.get_text()
        raw = str(raw)
        onion_page_array = [raw, parsed]
    except URLError as e:
        logging.debug(‘‘Could not download. URL error. ")
        print(‘‘URL error.", e)
    except HTTPError as e:
        logging.debug(‘‘Could not download ")
        print(‘‘HTTP error.", e)

    return onion_page_array
```

Listing 3. Code excerpt from the web page download function in D3-Collector

3.2.4 D3-Analyser

To utilize the data collected, and to demonstrate its analysis potential, a tool named D3-Analyser was developed. This tool consisted of two Python scripts:

1. `d3_classifier.py` that builds different machine learning classifiers that are trained on annotated data and downloaded web pages.

2. `d3_iaa.py` that calculates inter-annotator agreements (IAA) between two annotators' categorizations.

The D3-Analyser, reads annotations, categorizations, highlighted text excerpts, and web page content from URLs in the database and then creates a bag of words of them. The bag of words is a matrix of token counts of texts used to create a training data for a classification model. The training data was split into 70% training and 30% testing data set. The classification algorithms used in `d3_classifier.py` were developed using the Scikit-Learn API (Pedregosa et al., 2011). The algorithms implemented for classification were: Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and Multinomial Naïve Bayes (NB). The algorithms use annotations, text excerpts, and web page content from the database as features for training the models to make classifications of previously unseen web pages.

The purpose of the D3-Analyser was merely to verify that the collected data, i.e. the annotations, categorizations, and text excerpts from D3-Annotator, fit their purpose and to demonstrate usefulness of them. The performance of the classifiers was evaluated using Scikit-Learn's five-fold cross validation and calculating precision and recall, as well as accuracy scores.

Inter-annotator agreement (IAA), or "consensus score," is a practical and highly relevant component of an annotation system. In D3-Analyser, the script `d3_iaa.py` runs the IAA calculations of the categorizations in the database. One of the most commonly used IAA scores is Cohen's kappa. In D3-Analyser, Cohen's kappa was implemented using the Scikit-Learn library. One disadvantage with this IAA calculation is that it allows only two annotators, and it requires the input data sets to be of the same size. In the controlled environment used to demonstrate this feature, these limitations were not a problem. Cohen's kappa (κ) is calculated by number of agreements (P_o) subtracted by number of possible agreements due to chance (P_c) divided by the total number of samples (1) minus the agreements due to chance. The kappa score ranges from -1 to 1, where 1 is an absolute agreement and -1 no agreement at all (Cohen, 1960) (see Formula 1).

$$\kappa = \frac{P_o - P_c}{1 - P_c} \tag{1}$$

3.2.5 D3-Visualiser

In an investigation, it could be useful to have an overarching control panel or dashboard to enlarge the details. To visualize and generalize the content in the annotation database, D3-Visualiser was developed. D3-Visualiser is a Python Flask[5] web service with a Dash[6] based interface that presents the latest URLs with their respective category, notes, and highlighted text that has been added to the database.

Moreover, D3-Visualiser presents diagrams of descriptive statistics of the database and its content. The different classifiers' scores are presented in this web interface. Similarly, the inter-annotator agreement scores on the categorizations of URLs in the database are shown in the D3-Visualiser interface.

3.3 Artifact Demonstration

The purpose of the demonstration is to show how the artifact can be used to addresses the research problem. This is done by designing a use case – either a fictitious one, literature-based or a real-life case. The latter is both the better choice and the most expensive one (Johannesson & Perjons, 2014, p. 133).

In this research study, we chose to combine an experiment-based demonstration with a fictitious use case demonstration. The reason is two-fold: first, to demonstrate that the infrastructure and the developed tool set worked in an experimental setting, and second, to evoke its deployment in a real-life situation by going through a fictitious scenario.

3.3.1 Experiment Based Artifact Demonstration

The controlled environment consisted of two computers: a client running D3-Annotator as a web extension in the Tor browser, and a server computer running the tools D3-Centraliser, D3-Analyser, D3-Visualiser, and D3-Collector. The two computers were put on different networks, with a switch and

[5] https://flask.palletsprojects.com

[6] https://plotly.dash.org

Table 1. Example from database entries of web pages annotated using D3-Annotator. These entries also serve as training data machine for learning classifiers.

SHA256	URL	Annotation	Highlighted Text	Category	User
a2a...13	http://eeyovrly7 charuku.onion/	“CaseID: 202103312 possibly related to gangUID55"	“MDMA CRYSTAL PURE! MARQUIS TESTED: 1.0 G(30),3.0G(90), 5.0G (120)"	“Dark marketplace"	U1
06b...4c	http://xdsa5xcrrrxxxolc .onion/?product= sputnik-v-10-shots	“possible scam"	“Sputnik V is the world's first registered vaccine based on a well-studied human adenoviral vector-based platform."	“Covid-19"	U2
ea2...1e	http://bepig5bcjdhtlwpg eh3w42hffftcqmg 7b77vzu7ponty52k iey5ec4ad.onion/	“Kamagra seller c.f. Case 2021-331:3"	“Kamagra 100mg Generic Viagra Tablets are a very popular, successful, and widely accepted treatment for erectile dysfunction."	“Other medical"	U2
e2c...8e	http://oscarn4se6ji4 leq.onion/10g-81- purity-metham phetamine	“suspect 11"	“10g-81% PURITY METHAM PHETAMINE"	“Dark marketplace"	U2

router in between them. All network interfaces and Ethernet cables used gigabit standard (IEEE 802.3ab). The specifications for each machine can be found in Table 2.

The authors themselves used D3-Annotator to annotate, categorize, and save highlighted excerpts from different .onion web pages, both legitimate and illicit ones. The web pages were found by using the search engine Ahmia[7] and the search words: “dark marketplace," “drug market," “steroids," and “covid-19." In addition, The Hidden Wiki[8] was used to find .onion links. The result was a demonstration dataset consisting of a list of 95 URLs in six different categories. Naturally, within the 85% of illegal content, there were pointers to web sites that afford certain criminal activities such as the purchase of narcotics, medical supplements, stolen credit cards, counterfeited goods, and falsified documents. All links were checked against Ahmia's black listed URLs[9], to verify that no URLs referred to child abuse material.

The list of .onion URLs can be found on the researchers' Git repository[10]. In total, 95 web pages were categorized in categories such as “dark marketplace," “steroids," “covid-19," “gambling," “hacking," “credit cards," “seized," and “legitimate." The URL list comprised circa 85% cybercriminal (illicit) content and circa 15% legitimate content.

In addition to categorizing web pages, the two authors wrote annotations, and marked any web page text found relevant to the categorization (see example in Figure 2 and 4). The structure of the database entries is presented in Table 1. The categorizations, annotations, and text excerpts made with D3-Annotator were confirmed and validated to be intact after being synchronized to the database via D3-Centraliser. In the same manner, manual inspection was done to con-

[7] `https://ahmia.fi`

[8] `https://thehiddenwiki.com/`

[9] `https://ahmia.fi/banned/`

[10] `https://gitea.dsv.su.se/jebe8883/D3/onion_urls.txt`

Table 2. Specifications of machines used in the controlled experiment for artifact demonstration and evaluation.

Machine	Operating System	CPU	RAM	HDD	File System
Client	Arch Linux (Linux 5.11.6)	Intel Core i5 (2.20 GHZ) x86_64	16 GB (DDR3)	HGST HTS725050A7E635 300 MB/s (7200RPM)	BTRFS
Server	FreeBSD 11.04p5 (jail)	AMD Athlon 5150 x86_64	8GB (DDR3)	Western Digital WD2500BEVT 300MB/s (5400 RPM)	ZFS
Router	PfSense 2.5 (FreeBSD 12.02 (stable))	AMD Athlon 64 2.2 GHz x86_64	3GB (DDR2)	Western Digital WD800JD-75MSA1 300MB/s (7200 PRM)	UFS

firm that the annotated URLs were automatically fetched and archived in the database by D3-Collector. The entire D3-Annotator database along with annotations, text excerpts, categories, web page content, and metadata is available on the aforementioned code repository site.

3.3.2 Artifact Demonstration use case

The research problem stated there was no annotation tool bespoke for dark web cybercrime investigations. To demonstrate that we have developed such a tool, a list of URLs has been defined for the experiment and simulated cybercrime investigation environment. While this section focuses on the collection of data from the user, the next section deals with the data analysis and evaluation. To concretize the use of the developed tool by an investigator, we present a fictitious cybercrime investigation scenario from a user perspective as follows.

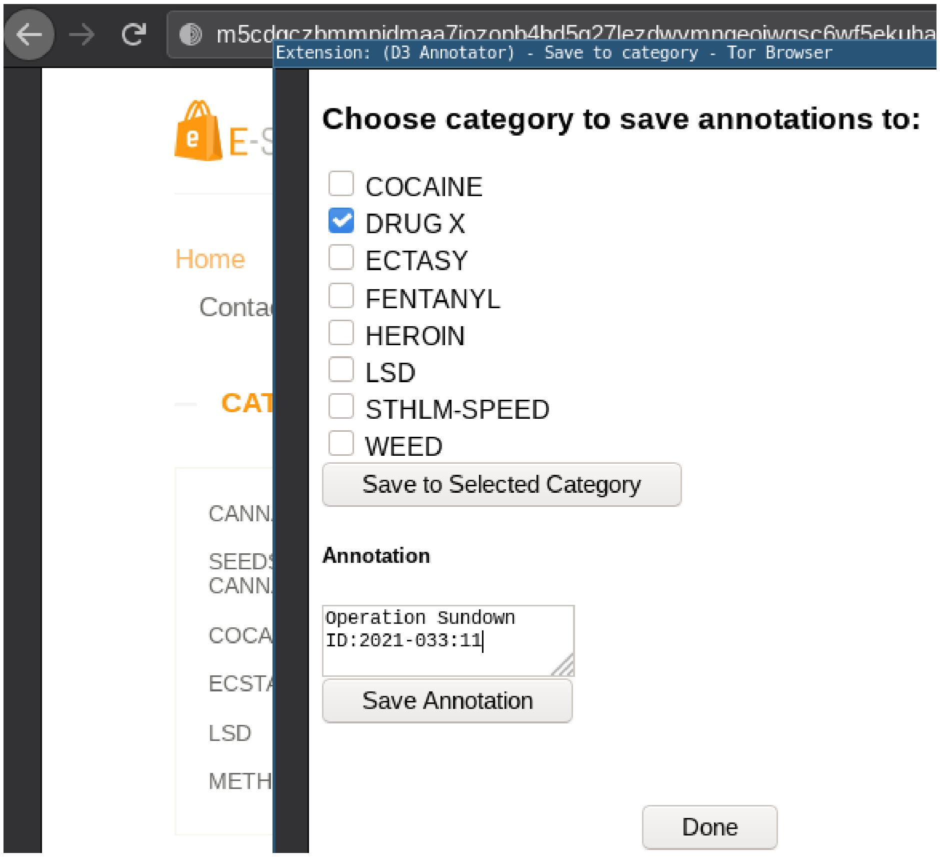
1. Investigator Watson has been given the task to collect information relating to a specific type of drug X being sold on different Tor websites by, possibly, one and the same user.

2. The investigator starts by searching for URLs to different drug marketplaces on the Tor network using a clear web service, e.g., `https://darkweblive.net/markets/`

3. Watson uses D3-Annotator to make annotations, mark any text of interest, and categorize the web page as “Drug-X" if it possesses the characteristics of drug X. See the examples in Figure 2 and Figure 4.

4. In the background, D3-Annotator automatically transmits the annotations, highlighted text excerpts, timestamps, and URLs, to D3-Centraliser on the synchronization server. The transmission and database insertion preserve the data’s forensic integrity by verifying hash sums before and after transmission and logging timestamps.

5. Once D3-Centraliser has retrieved and stored the data from D3-Annotator, it invokes D3-Collector to fetch the web page of the last received URL and automatically download and archive that web page to the database.


m5cdqczhmmpidmaa7iozoph4bd5g27lezdwymngeoiwqsc6wf5ekuha
E-S
Home
Conta
CAT
CANN
SEEDS
CANN
COCA
ECSTA
LSD
METH

Extension: (D3 Annotator) - Save to category - Tor Browser

Choose category to save annotations to:

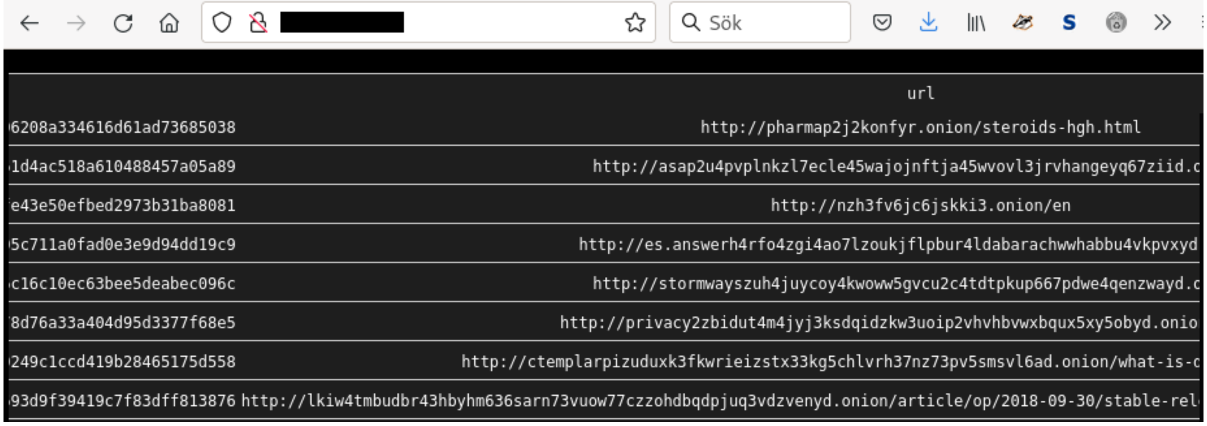
- ☐ COCAINE
- ☑ DRUG X
- ☐ ECTASY
- ☐ FENTANYL
- ☐ HEROIN
- ☐ LSD
- ☐ STHLM-SPEED
- ☐ WEED

Save to Selected Category

Annotation

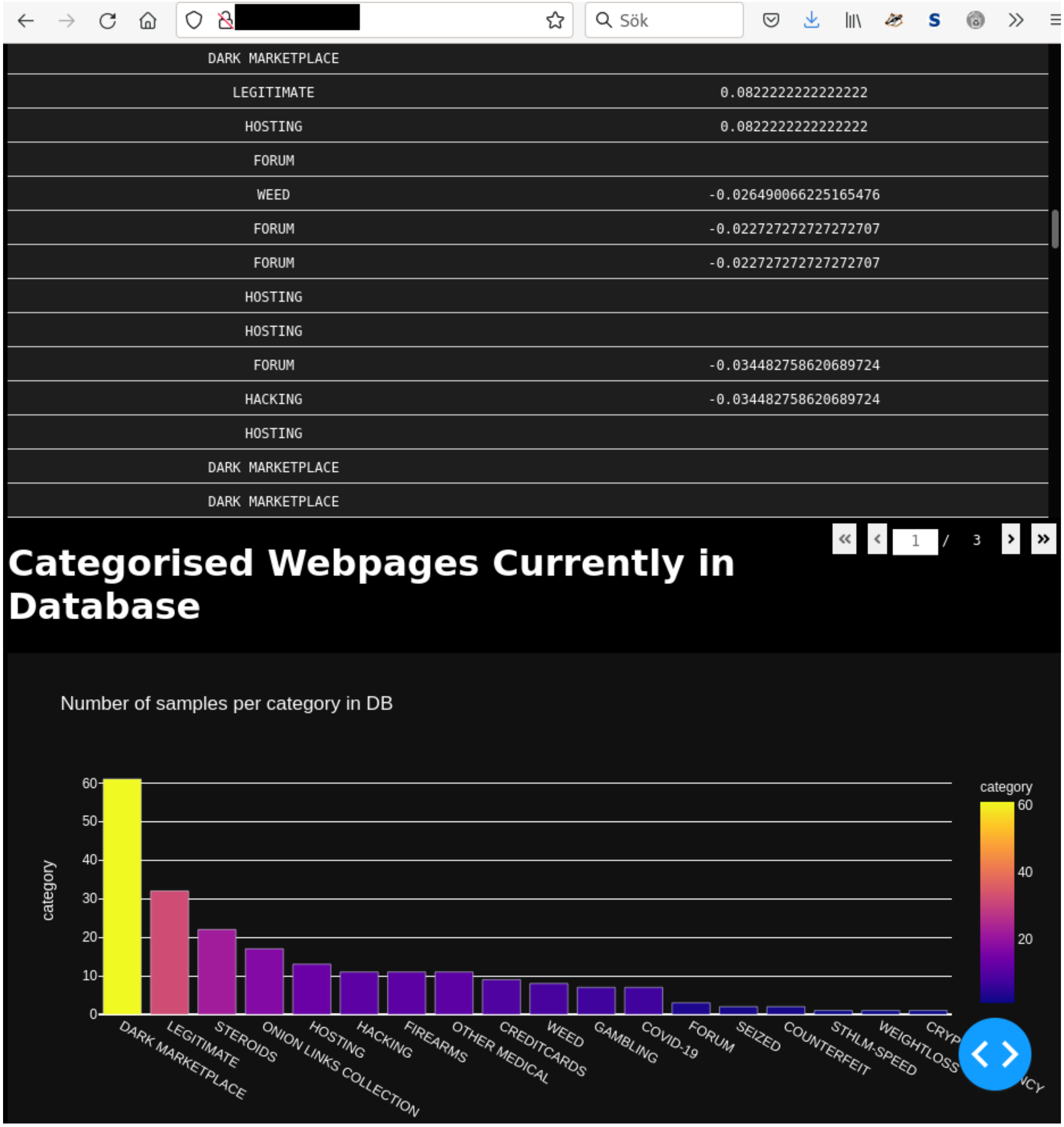
```
Operation Sundown
ID:2021-033:11
```

Figure 4. The right click menu that D3-Annotator presents to the user when text is highlighted on a web page.



	url
6208a334616d61ad73685038	http://pharmap2j2konfyr.onion/steroids-hgh.html
1d4ac518a610488457a05a89	http://asap2u4pvplnkzl7ecle45wajojnftja45wvovl3jrvhangeyq67ziid.c
e43e50efbed2973b31ba8081	http://nzh3fv6jc6jskki3.onion/en
5c711a0fad0e3e9d94dd19c9	http://es.answerh4rfo4zgi4ao7lzoukjflpbur4ldabarachwwhabbu4vkpvxyd
c16c10ec63bee5deabec096c	http://stormwayszuh4juycoy4kwoww5gvcu2c4tdtpkup667pdwe4qenzwayd.c
8d76a33a404d95d3377f68e5	http://privacy2zbidut4m4jyj3ksdqidzkw3uoip2vhvhbvwxbqux5xy5obyd.onio
249c1ccd419b28465175d558	http://ctemplarpizuduxk3fkwrieizstx33kg5chlvrh37nz73pv5smsvl6ad.onion/what-is-c
93d9f39419c7f83dff813876	http://lkiw4tmbudbr43hbyhm636sarn73vuow77czzohdbqdpjuq3vdzvenyd.onion/article/op/2018-09-30/stable-rel

Figure 5. D3-Visualiser's homepage showing the latest added URLs and their UUID in form of a SHA256 sum



DARK MARKETPLACE	
LEGITIMATE	0.0822222222222222
HOSTING	0.0822222222222222
FORUM	
WEED	-0.026490066225165476
FORUM	-0.022727272727272707
FORUM	-0.022727272727272707
HOSTING	
HOSTING	
FORUM	-0.034482758620689724
HACKING	-0.034482758620689724
HOSTING	
DARK MARKETPLACE	
DARK MARKETPLACE	

Categorised Webpages Currently in Database

Number of samples per category in DB

Figure 6. D3-Visualiser's homepage displaying a table of two annotator's agreement scores on different categories for a certain URL. Below, a bar chart of number of web pages in each different category is shown.

6. Watson has now shared his annotations, categorizations, and highlighted text excerpts from his visited .onion web pages with the other colleagues who also have access to the synchronization server. On the server, D3-Analyser continuously uses that data to train classifiers and produce statistics. To summarize, Watson made a contribution to the training data for classification models.

7. Watson and his colleagues can go to the synchronization server's website to find visualizations automatically uploaded by the D3-Visualiser script to get an overview of the latest URLs, annotations, annotators' agreement scores, and statistics of categorized URLs that have been analyzed in the background by the D3-Analyser scripts (see Figure 5 and 6 for an example).

8. When lists of new URLs, from an external source for example, are fed into the database, D3-Collector fetches the web content from the URLs and D3-Analyzer classifies them based on the already existing training data. If the new URLs point to a web page relating to drug X, the classification models will be able to spot it, so neither Watson, nor his colleagues need to manually visit them to make the classification.

3.4 Artifact Evaluation

An artifact evaluation defines how well the artifact solves the problem it was developed to solve, i.e. how well it fulfills the requirements. The former section demonstrates the collection of data from the user and synchronized with the central server. In this section, the artifact proceeds with the analyzes of the collected data and the respective evaluation of the wellness of the analysis. The outcome of the evaluation is presented in the Results section.

To provide an indication of how well our toolset works, we made controlled experiments that measured how quickly annotations are transferred to the database. We also examined if the integrity of the annotations had been compromised in any way in transit. Furthermore, the precision, recall, F2, and accuracy scores of the classification models were assessed. Finally, the inter-annotator agreements were calculated to evaluate how well the annotators agree on each categorization, and to demonstrate that the feature works as intended.

3.4.1 Experiment Setup for Artifact Evaluation

The hardware used in the experiment setup in previous section, was used for the artifact evaluation as well. The performance of the D3-Annotation and D3-Centraliser symbiosis, i.e. the times of read, write, transmit, and store operations were registered using the JavaScript Performance API[11]. A mean of ten transactions between D3-Annotator and D3-Centraliser were calculated. Likewise, was the transmission time between D3-Centraliser retrieving a URL and D3-Collector inserting it into the database calculated on an average of ten transactions.

The forensic soundness was evaluated by a coded routine that confirms the SHA256 hash sums of the messages sent from D3-Annotator matched the hash sums of the messages received on the D3-Centraliser server. The level of forensic soundness is in this case a binary measure; either the hash sums match or they do not.

The evaluation of the classification models in D3-Analyser consisted of an evalua-

[11] `https://developer.mozilla.org/en-US/docs/Web/API/Performance/`

tion data set of 150 .onion URLs[12] categorized by the authors using D3-Annotator. The URLs belonged to five different categories: “covid-19," “steroids," “cocaine," “cannabis," or “firearms." The URLs were found, like in the demonstration section, via `www.ahmia.fi` and `hiddenwiki.org`. As designed, these 150 pages were automatically retrieved and stored in the database by D3-Collector. After parsing the HTML from the pages the content was used to train the classifiers with the categories as the class label.

The only requirement which was not possible to fully evaluate in a simulated environment or with numerical measures was requirement RQ-4: “Enable for non-technical investigators to use the annotation tool." This apparently requires a more resource intense evaluation in form of an action research or case study. However, a description of how this works and why it works was presented in the previous section. To summarize the artifact assessment, each of the requirements were evaluated in the following way:

1. *RQ1: Forensically sound annotation and preservation of web content relevant to a (cybercrime) investigation.* - Evaluated in experiment by verifying integrity of all data stored using hash sums.

2. *RQ2: Automatically archive and preserve an annotated web page and its metadata.* - The performance of the fetch and archive functions in D3-Collector were evaluated using Javascript's timer API. Manual inspections confirmed that the data was stored as intended.

3. *RQ3: Automatically synchronize data to a central server.* - Operational performance measured by Javascript's timer API. Programmatic verification using hash check sums between client and server.

4. *RQ4: Enable for non-technical investigators to use the annotation tool.* - Demonstrated to work only.

5. *RQ5: Support for multiple annotators.* - Demonstrated to work only.

6. *RQ6: Web browser based monitoring panel for data visualization.* - Demonstrated to work via D3-Visualiser only.

7. *RQ7: Inter-annotator or consensus agreement calculation of annotations.* - Demonstrated to work by using a common statistics APIs for calculating Cohen's kappa in D3-Analyser. The scores were based on 95 .onion pages annotated by the authors.

8. *RQ8: Intelligent classification models based on annotations.* - Evaluated with Scikit-Learn's 5-fold cross validation method. D3-Analyser constructed classifiers based on five different algorithms. The evaluation was done on 150 .onion pages categorized in five different categories with a 70/30 split on training and test data set respectively.

4. RESULTS

This section elucidates the results of the toolset evaluation to answer the research

[12] `https://gitea.dsv.su.se/jebe8883/D3/onion_urls_150.txt`

question. Firstly, we could confirm that all requests sent to the server from the Tor Browser annotation tool, D3-Annotator, to the central server management script, D3-Centraliser, in our test case were correctly received, verified with SHA256 hash sums, and stored in the database. The performance evaluation of the API HTTP transmissions provided an indication of how swift the tools were when used together in the experiment setting (see Table 3 for details).

Secondly, to demonstrate that the categorization worked as intended, and to evaluate the annotators' agreement over the categories, D3-Analyser calculated the Cohen kappa scores of each of the two annotators' categorizations. The annotators did not make the same categorization of five out of 21 items in the category "Dark Marketplace" and two out of six web pages in the category "Steroids", thus lower score in these categories as can be seen in Table 4. The total agreement of all 95 web pages categorizations was 0.84.

Another component of D3-Analyser is the classifier that uses data from the annotation database to build different classification models. The classification algorithms implemented in D3-Analyser were: Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and Naïve Bayes (NB). Based on the manually verified list of URLs, categorized using the D3-Annotator, the classifiers predicted the categories of unseen web pages at an accuracy of 85% to 96%. The random forest classifier was able to classify 48 web pages with a balanced accuracy score of more than 85% with a precision of 0.86 and a recall of 1.00. The precision, recall, and F1 scores for each of the classifiers SVM, LR, RF, and NB can be found in Tables 5, 6, 7, and 8 respectively. The SVM and LR classifiers reached the same cross validation score and balanced accuracy score of 0.9399 and 0.9642 respectively. The naïve Bayes classifier reached over 0.9 for both scores as well.

5. DISCUSSION

The results show that the prototype presented in this paper can be used for automatically categorizing unknown websites based on sets of data that have been manually annotated, excerpted, or categorized with the same prototype. With the intention of being a law enforcement investigative toolset, it is not designed to intrude on the Tor user's privacy. Although the tool was specifically created and evaluated for the Tor browser and the Tor network, it can effortlessly be configured for the clear web using a regular Firefox browser as well.

It should be noted that D3-Annotator is a tool that registers its user's interactions and does not regard the privacy of the user when exerting the Tor browser. Add-ons can, given an insecure Tor configuration, be identified by .onion services' web servers and possibly block access to the website. However, in this prototype setting, it is presumed that the configurations are set to secure defaults.

The tool developed for collecting data from the (dark) web was a prototype that uses a local SQLite3 database to store the data collected. It might raise some concerns apropos the forensic soundness, veracity, integrity, credibility, and reliability of the tool to investigate cybercrime under a judicial system, which are of interest in the refining process of the tool.

Most likely, law enforcement agencies already have established solutions for database management that have been accredited as storage for digital forensic investigations. Concerning the D3 toolset, SQLite3 could be replaced by any other database management system supported by NodeJS, that runs the synchronization service, D3-Centraliser.

Table 3. Performance metrics of transmissions over an average of ten transactions between the different programs in the developed D3 toolset.

Operation	**Nodes**	**Time**
HTTP Rest API transmit web page array	Client (D3-Annotator) to Server (D3-Centraliser)	1900 milliseconds
Insert web page array to SQLite3 DB	Server (D3-Centraliser)	27 milliseconds.
Insert category array to SQLite3 DB	Server (D3-Centraliser)	18 milliseconds.
Insert annotation array to SQLite3 DB	Server (D3-Centraliser)	18 milliseconds.
Insert highlightedText array to SQLite3 DB	Server (D3-Centraliser)	17 milliseconds.
Fetch and archive web page in SQLite3 DB	Server (D3-Centraliser to D3-Collector)	29.33 seconds.

Table 4. Inter-annotator agreement scores of 95 web pages categorized with D3-Annotator.

Category	**Cohen's Kappa**
COVID-19	1.0
CREDITCARDS	1.0
DARK MARKETPLACE	0.25
LEGITIMATE	1.0
SEIZED	1.0
STEROIDS	0.80
Total	0.8449

The results indicate that the performance of the network transactions between the annotation tool client and the synchronization server and the database were satisfactory in the environment of our experiment. The transactions that took the longest were of course the D3-Collector fetching and archiving of .onion web pages. Since D3-Collector was designed to start a Tor HTTP proxy connection using a Torsocks shell for every new URL that was sent to the server, these requests could take up to five minutes (which was the request timeout). The time between annotation in the browser and archiving a web page on the server could therefore be more than five minutes. Potentially, in this time window, the web host could go down or the webmaster could take the web page down or it can be modified in some way. Therefore it would be suitable to try to reduce this time window.

The primary purpose of the D3 toolset was not machine learning-based classification tasks. Nevertheless, the results indicate that based on previously categorized and annotated web pages, all four classification algorithms managed to correctly classify previously unseen web pages with an accuracy rate of between 85% and 95%. In general, this is considered an acceptable number in data mining and data analysis. Though, it should be stressed that these figures were achieved from training on a small set of 150 parsed HTML web pages, mostly dark marketplaces. None of the sites required a login or a CAPTCHA to be solved, which is common on many of the .onion sites that enable and engage in criminal activity. Automating scraping and crawling of .onion sites

Table 5. Precision, recall, and F1 scores of support vector machine (SVM) classifications trained on D3-Annotator data.

Category	**Precision**	**Recall**	**F1-score**	**Support**
CANNABIS	1.00	1.00	1.00	18
COCAINE	0.67	1.00	0.80	2
COVID-19	1.00	1.00	1.00	10
CREDIT CARDS	1.00	1.00	1.00	7
FIREARMS	1.00	0.75	0.86	4
STEROID	1.00	1.00	1.00	7

Table 6. Precision, recall, and F1 scores of logistic regression (LR) classifications trained on D3-Annotator data.

Category	**Precision**	**Recall**	**F1-score**	**Support**
CANNABIS	1.00	1.00	1.00	18
COCAINE	0.67	1.00	0.80	2
COVID-19	1.00	1.00	1.00	10
CREDIT CARDS	1.00	1.00	1.00	7
FIREARMS	1.00	0.75	0.86	4
STEROID	1.00	1.00	1.00	7

Table 7. Precision, recall, and F1 scores of random forest (RF) classifications trained on D3-Annotator data.

Category	**Precision**	**Recall**	**F1-score**	**Support**
CANNABIS	0.86	1.00	0.92	18
COCAINE	1.00	1.00	1.00	2
COVID-19	1.00	0.90	0.95	10
CREDIT CARDS	1.00	1.00	1.00	7
FIREARMS	1.00	0.25	0.40	4
STEROID	0.88	1.00	0.93	7

Table 8. Precision, recall, and F1 scores of multinomial naïve Bayes (NB) classifications trained on D3-Annotator data.

Category	**Precision**	**Recall**	**F1-score**	**Support**
CANNABIS	1.00	0.86	0.92	21
COCAINE	1.00	1.00	1.00	2
COVID-19	1.00	0.92	0.96	12
CREDIT CARDS	0.70	1.00	0.82	7
FIREARMS	1.00	1.00	1.00	3
STEROID	0.75	1.00	0.86	3

Table 9. Performance evaluation metrics of classification models based on D3-Annotator input data.

Algorithm	Balanced Accuracy	Cross Validation Mean Score
Support Vector Machine	0.9642	0.9399
Logistic Regression	0.9642	0.9399
Random Forest	0.8583	0.9166
Naïve Bayes	0.9166	0.9552

protected by CAPTCHA supported authentication to build a training data set requires additional effort and is a research project in its own. Furthermore, the division of categories was not evenly spread, so a better balanced and bigger data set would produce a more reliable result, yet the demonstration and evaluation indicate that the toolset fits its purpose.

In our experiment setting, the free text annotations were reasonably minimal; thus, the training data is not optimal for achieving high precision classification predictions or text analysis. However, the inter-annotator agreement calculation proved the categorization feature in D3-Annotator worked as intended and also indicated that Cohen's kappa has its limitations when used in the this setting.

Cohen's kappa is a commonly used agreement score calculation. However, it is not as dynamic as a dark web annotation tool requires. It requires two annotators, annotating the same number of samples in order to calculate an agreement score. This worked in a controlled experiment setting, however, in real-life annotators will most likely be (1) more than two, and (2) the number of annotations and annotated samples will most likely differ between each annotator. Even though this agreement measure was not optimal for the use case of (dark) web page annotations designed for criminal investigators, they still indicate when there is a considerable agreement or a significant disagreement among annotations.

5.1 Limitations

The confidentiality and the forensic soundness of the SQLite3 database and the HTTP API request and response data in transit is dependent on the infrastructure in which the system is operating. In this study we limited our research to a closed environment network with only a client, a router, and a server present. The idea with the D3 toolset is to have a central server hosted internally to avoid any unnecessary exposure to external parties who should not be privy to the server's data.

As mentioned, the reliability and validity of some of the tool's evaluation are affected by the small data sets that were used in the experiments. The primary purpose of the data sets is to serve as a demonstration of functionality which implies acceptability of the sample's modest size.

In an ideal case, a naturalistic ex-ante and ex-post evaluation in an action research model of the D3 toolset study should be very beneficial. The experiment-based performance evaluation of the artifact in this study is a proper one to demonstrate that it works and works well.

6. CONCLUSIONS AND FUTURE RESEARCH

The paper presents a web content annotation tool developed as a Tor Browser (and Firefox) plugin, favoring law enforcement agency investigators working with cybercrime. The plugin works as a manual web content annotation, categorization, and collection tool that serves as an input for supervised machine learning algorithms. The algorithms learn to classify web content relevant to cybercrime investigations from a user generated set of web contents. The annotation tool automatically synchronizes the annotations and web page content with a central server, enabling for colleagues to share material with each other and extend the database.

There is a need to induce a continuous development and to test the proposed data collection and analysis tools. As indicated throughout the article some of the venues for future research include assessing improved forensic soundness of the tool examining the resilience of the chain of custody, and preserving the integrity of the collected data.

In addition, the analysis of graphic content is almost on the forefront of the work that follows. Currently, the developed collection tool, D3-Collector, is capable of storing graphic content but is not programmed to analyze it. This could be a rather convenient feature to include in the classifier's training and testing data sets. D3-Collector, is a primitive tool, and any extension of this toolset would benefit from developing a more powerful and sophisticated web page crawler and collector that could be configured to handle authentication pages, session handling, cookies, CAPTCHAs, and two-factor authentication - in addition to speeding up the scraping process in order to reduce the time window between annotation and collection.

Furthermore, an inquiry scrutinizing the most adequate inter-annotator agreement algorithm for cybercrime investigators is certainly deemed necessary based on the result from this research study. A future research topic would thus include an inter-annotator's agreement or consensus algorithm that allows multiple annotators and is capable of managing differently sized inputs.

Finally, free text annotations produced by human intervention are challenged by a few “traditional" yet persistent problems such as subtle differences in the knowledge domains, priorities, perceptions due to the different languages and their semantic reflections in the processes of understanding and consensus building.

Recent advances in artificial intelligence, including data science and natural language processing, generate compelling reasoning paradigms and inference engines that may address numerous challenges ranging from multi-model problem solving to respective ethical dimensions. Morphed in modules such as knowledge elicitation, extraction, interpretations, and explanations could lead to standardized platforms for tackling large and complex cases involving many international investigators working on cross-border cybercrimes in a global context.

ACKNOWLEDGEMENT

The work is partially supported by the NordForsk Grant No. 80512 for the project “Police Detectives on the TOR-network." Additionally, the project has participated in the COST network action DigForAsp, project number CA17124.

REFERENCES

Arabnezhad, E., La Morgia, M., Mei, A., Nemmi, E. N., & Stefa, J. (2020). A light in the dark web: Linking dark web aliases to real internet identities. In *2020 ieee 40th international conference on distributed computing systems (icdcs)* (p. 311-321). Singapore. doi: 10.1109/ICDCS47774.2020.00081

Casey, E. (2011). *Digital evidence and computer crime: Forensic science computers and the internet* (Third ed.). USA: Elsevier.

Chen, H., Chung, W., Quin, J., Reid, E., Sageman, M., & Weinmann, G. (2008). Uncovering the dark web: A case study of jihad on the web. *Journal of the American Society for Information Science and Technology*, *59*(8).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. doi: 10.1177/001316446002000104

Comey, J. B. (2015). *Going dark: Encryption, technology, and the balances between public safety and privacy.* Retrieved from `https://www.fbi.gov/news/testimony/going-dark-encryption-technology-and-the-balances-between-public-safety-and-privacy` (Retrieved 2021-03-30)

Dalianis, H. (2018). *Clinical text mining: Secondary use of electronic patient records.* USA: Springer Open.

Dalins, J., Wilson, C., & Carman, M. (2018). Criminal motivation on the dark web: A categorisation model for law enforcement. *Digital Investigation*, *24*, 62-71. doi: 10.1016/j.diin.2017.12.003

de Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. *ACM SIGMOD Record*, *30*(4), 55–64. doi: 10.1145/604264.604272

Europol. (2017). *Drugs and the darknet: Perspectives for enforcement, research and policy.* Retrieved from `https://www.europol.europa.eu/publications-documents/drugs-and-darknet-perspectives-for-enforcement-research-and-policy` (Retrieved 2021-0 1-07)

Farzan, R., & Brusilovsky, P. (2008, January). Annotated: A social navigation and annotation service for web-based educational resources. *New Rev. Hypermedia Multimedia*, *14*(1), 3–32. doi: 10.1080/13614560802357172

Ghappour, A. (2017). Searching places unknown: Law enforcement jurisdiction on the dark web. *Stanford Law Review*, *69*(4).

Ghosh, S., Das, A., Porras, P., Yegneswaran, V., & Ghehani, A. (2017). Automated categorization of onion sites for analyzing the darkweb ecosystem. In *Kdd'17: Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (p. 1793-1802). ACM.

Hansken.nl. (2020). *Dutch investigative services team up to continue hansken development.* Retrieved from `https://www.hansken.nl/latest/news/2020/07/30/dutch-investigative-services-team-up-to-continue-hansken-development`

Hayes, D., Cappa, F., & Cardon, J. (2018). A framework for more effective dark web marketplace investigations. *Information, special issue: Darkweb Cyber Threat Intelligence Mining*, *9*(8:186). doi: 10.3390/info9080186

Johannesson, P., & Perjons, E. (2014). *An introduction to design science.* Springer International Publishing. doi: 10.1007/978-3-319-10632-8

Johnston, P. (n.d.). *Paj's home: Cryptography: Javascript md5: Scripts: md5.js.* Retrieved from `http://pajhome.org.uk/crypt/md5/md5.html` (Retrieved 2021-01-12)

Kalpakis, G., Tsikrika, T., Iliou, C., Mironidis, T., Vrochidis, S., Middleton, J., ... Kompatsiaris, I. (2016). Interactive discovery and retrieval of web resources containing home made explosive recipes. In *Has 2016: Human aspects of information security, privacy, and trust* (p. 221-233). Springer.

Kessler, G. (2016). The impact of sha-1 file hash collisions on digital forensic imaging: A follow-up experiment. *Journal of Digital Forensics, Security and Law*, *11*(10), 129-139. doi: https://doi.org/10.15394/jdfsl.2016.1433

Kwon, K. H., Priniski, J. H., Sakar, S., Shakarian, J., & Shakarian, P. (2017). Crisis and collective problem solving in dark web: An exploration of a black hat forum. In *Proceedings of the 8th international conference on social media & society article no. 45* (p. 1-5). ACM.

McKemmish, R. (2008). When is digital evidence forensically sound? In *Ifip international conference on digital forensics* (p. 3-15). Springer Link.

Netclean. (2019). *Netclean report 2018 - a report on documented sexual abuse against children.* Retrieved from `https://www.netclean.com/netclean-report-2018/` (Retrieved 01/11/2020)

Netclean. (2021). *Netclean report 2020 - covid-19 impact 2020.* Retrieved from `https://www.netclean.com/netclean-report-2020/` (Retrieved 02/02/2021)

Neto, L., Pinto, N., Proença, A., Amorim, A., & Conde-Sousa, E. (2021). 4specid: Reference dna libraries auditing and annotation system for forensic applications. *Genes*, *12*(1). Retrieved from `https://www.mdpi.com/2073-4425/12/1/61` doi: 10.3390/genes12010061

Neves, M., & Leser, U. (2014). The forensic investigation of android private browsing sessions using orweb. *Briefings in Bioinformatics*, *15*(2), 327-340. doi: https://doi.org/10.1093/bib/bbs084

Nunes, E., Diab, A., Gunn, A., Ericsson, M., Vineet, M., Mishra, V., ... Shakarian, P. (2016). Darknet and deepnet mining for proactive cyber treat intelligence. *Intelligence and Security Informatics (ISI)*, 7-12. doi: 10.1109/ISI.2016.7745435

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Popov, O., Bergman, J., & Valassi, C. (2018). A framework for a forensically sound harvesting the dark web. In *Cecc 2018: Proceedings of the central european cybersecurity conference 2018* (p. 1-7). ACM. doi: 10.1145/3277570.3277584

Portnoff, R. S., Afroz, S., Durrett, G., Kummerfeld, J. K., Berg-Kirkpatrick, T., McCoy, D., ... Paxson, V. (2017). Tools for automated analysis of cybercriminal markets. In *Proceedings of the 26th international conference on world wide web* (p. 657–666). Republic and Canton of Geneva, CHE:

International World Wide Web Conferences Steering Committee. doi: 10.1145/3038912.3052600

Qin, R. Z. Y., Huang, Z., & Chen, H. (2003). Authorship analysis in cybercrime investigation. In (p. 59-73). Springer.

Sabbah, T., Selamat, A., Selamat, M. H., Ibrahim, R., & Fujita, H. (2016). Hybridized term-weighting method for dark web classification. *Neurocomputing*, *173*, 1908-1926. doi: 10.1016/j.neucom.2015.09.063

Sorokin, A., & Forsyth, D. (2008). Utility data annotation with amazon mechanical turk. In *Ieee computer society conference on computer vision and pattern recognition workshops* (p. 1-8). Anchorage, AK, USA. doi: 10.1109/CVPRW.2008.4562953

Spitters, M., Klaver, F., Koot, G., & van Staalduinen, M. (2015). Authorship analysis on dark marketplace forums. In *European intelligence and security informatics conference* (p. 631-641). IEEE.

SQLite.org. (n.d.). *35 precent faster than the filesystem.* Retrieved from `https://sqlite.org/fasterthanfs.html#approx` (Retrieved 24/03/2021)

Tai, X. H., Soska, K., & Christin, N. (2019). Adversarial matching of dark net market vendor accounts. In *Kdd '19: Proceedings of the 25th acm sigkdd international conference on knowledge discovery and data mining* (p. 1871-1880). IEEE. doi: 10.1145/3292500.3330763

Tensor. (n.d.). *Titanium: Tools for the investigation of transactions in underground markets.* Retrieved from `https://titanium-project.eu/` (Retrieved 2021-01-30)

Titaniu. (n.d.). *Titanium: Tools for the investigation of transactions in underground markets.* Retrieved from `https://titanium-project.eu/results/` (Retrieved 2021-01-30)

Tor-Project. (n.d.). *index : tor-browser.* Retrieved from `https://gitweb.torproject.org/tor-browser.git/tree/dom?h=tor-browser-24.3.0esr-1` (Retrieved 2021-04-10)

van Baar, R., van Beek, H., & van Eijk, E. (2014). Digital forensics as a service: A game changer. *Digital Investigation*, *11*, S54-S62. (Proceedings of the First Annual DFRWS Europe) doi: 10.1016/j.diin.2014.03.007

van Beek, H., van Eijk, E., van Baar, R., Ugen, M., Bodde, J., & Siemelink, A. (2015). Digital forensics as a service: Game on. *Digital Investigation*, *15*, 20-38. (Special Issue: Big Data and Intelligent Data Analysis) doi: 10.1016/j.diin.2015.07.004

Webtoolkit. (n.d.). *Javascript sha-256 - javascript tutorial with example source code.* Retrieved from `http://www.webtoolkit.info/javascript_sha256.html` (Retrieved 2021-01-03)

Wojahn, P. G., Neuwirth, C. M., & Bullock, B. (1998). Effects of interfaces for annotation on communication in a collaborative task. In *Proceedings of the sigchi conference on human factors in computing systems* (p. 456–463). USA: ACM Press/Addison-Wesley Publishing Co. doi: 10.1145/274644.274706

Zhang, Y., Zeng, S., Huang, C.-N., Fan, L., Yu, X., Dang, Y., . . . Chen, H. (2010). Developing a dark web collection and infrastructure for computational and social sciences. In *2010 ieee international conference on intelligence and security informatics* (p. 59-64). doi: 10.1109/ISI.2010.5484774