



<https://helda.helsinki.fi>

Helda

Ethically Archiving a Hard-to-Access Massive Research Data Set in the Language Bank of Finland : The Finnish Dark Web Marketplace Corpus (FINDarC)

Lindén, Krister

2023

Lindén, K, Ruokolainen, T, Hämäläinen, L & Harviainen, J T 2023, Ethically Archiving a Hard-to-Access Massive Research Data Set in the Language Bank of Finland : The Finnish Dark Web Marketplace Corpus (FINDarC). in M M Rantanen , S Westerstrand, O Sahlgren & J Koskinen (eds), Proceedings of the Conference on Technology Ethics 2023 - Tethics 2023. CEUR Workshop Proceedings, vol. 3582, CEUR-WS.org, Aachen, pp. 114-131, Conference on Technology Ethics, Tethics , Turku, Finland, 18/10/2023.

<http://hdl.handle.net/10138/569220>

cc_by

publishedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

Ethically Archiving a Hard-to-Access Massive Research Data Set in the Language Bank of Finland: The Finnish Dark Web Marketplace Corpus (FINDarC)

Krister Lindén^{1,*}, Teemu Ruokolainen^{2,†}, Lasse Hämäläinen^{2,†} and J. Tuomas Harviainen^{2,*}

¹*Department of Digital Humanities, University of Helsinki, Yliopistokatu 4, 00014 University of Helsinki, Finland*

²*Faculty of Information Technology and Communication Sciences, Tampere University, Kalevantie 4, Tampere, 33014, Finland*

Abstract

We discuss the archiving procedure of a corpus comprising posts submitted to Torilauta, a Finnish dark web marketplace website. The site was active from 2017 to 2021 and during this time one of the most prominent online illegal narcotics markets in Finland. As a result of the presented work, a reduced version of the corpus, Finnish Dark Web Marketplace Corpus (FINDarC), has been archived in the Language Bank of Finland. Researchers can apply for access rights to the corpus under the CLARIN RES licence. The discussion presented in this paper addresses the archiving process, including assessment of the risk and impact of data subject re-identification, assessment and implementation of viable data anonymization/reduction approaches, assessment of privacy and security measures implemented by the Language Bank of Finland, and future corpus management plan coordinated by the Language Bank of Finland.

Keywords

dark web, illegal narcotics, online marketplace, data sharing

1. Introduction

We discuss the archiving procedure of a corpus comprising posts submitted to Torilauta, a dark web marketplace website, for research purposes. The site was active from 2017 to 2021 and during this time one of the most prominent online illegal narcotics markets in Finland. Functionally, the site consisted of discussion imageboards where vendors and customers were able to set up instances of face-to-face trading, typically with the assistance of instant messaging software such as Wickr or Telegram. The original, unmodified data set comprising 3,104,976

Conference on Technology Ethics – Tethics, October 18–19, 2023, Turku, Finland

*Corresponding author.

†These authors contributed equally.

✉ krister.linden@helsinki.fi (K. Lindén); teemu.ruokolainen@tuni.fi (T. Ruokolainen); lasse.hamalainen@tuni.fi (L. Hämäläinen); tuomas.harviainen@tuni.fi (J. T. Harviainen)

🆔 0000-0003-2337-303X (K. Lindén)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

📄 CEUR Workshop Proceedings (CEUR-WS.org)

posts was collected and handed over to the ENNCODE consortium¹ by the site administration to be archived and shared for research purposes, as permitted by the site's Terms of Service. As a result of the presented work, a reduced version of the corpus comprising 3,104,515 posts, referred to as the Finnish Dark Web Marketplace Corpus (FINDarC), has been deposited in the Language Bank of Finland, a language resource service coordinated by the national FIN-CLARIN consortium formed by Finnish universities and other research organizations. Researchers can contact the Language Bank and apply for permission to access the corpus under the CLARIN RES license.²

While the dark web online market places, including Torilauta, emphasize user anonymity, the posts submitted to such sites can nevertheless contain personal information, such as unique usernames and personal names, enabling data subject re-identification. Therefore, as described in this paper, we have made our best effort to assess and identify the type and amount of personal information in the original unmodified data set, to assess and implement viable data anonymization/reduction approaches, to assess privacy and security measures implemented by the Language Bank of Finland, and to put in place a future corpus management plan coordinated by the Language Bank of Finland. Moreover, any future research based on the corpus is encouraged to implement appropriate ethical proofreading measures (see e.g. [19]) in order to further mitigate any potential harm from access to the material, to both the researchers and the studied populations.

The rest of the paper is organized as follows. We first discuss previous studies on Torilauta and existing related corpora in Section 2. We then provide an overview of the data set in Section 3. In Section 4, we present our experimental work, the purpose of which was to assess the amount of personal and sensitive data in the corpus. Subsequently, given the experimental results, we discuss in detail our data release plan in Section 5. Finally, we provide conclusions on the work in Section 6.

2. Related Work

In this section, we discuss previously published studies using the Torilauta site as a data source and related corpora in Sections 2.1 and 2.2, respectively.

2.1. Previous Studies on Torilauta

Prior to the work presented here, Torilauta was utilized as a data source in multiple linguistic and social science studies [15, 18, 16, 17, 21]. In particular, Haasio et al. [15] examined information needs of drug users using a sample of 9,300 posts.³ Harviainen et al. [18] studied cultural and socioeconomic aspects of drug traders using the same 9,300 post sample. Hämäläinen and Ruokolainen [17] studied narcotic substance vocabulary based on a sample of 3,000 posts. Hämäläinen et al. [16] studied a sample of 1,654 usernames extracted from posts submitted to the site. Karjalainen et al. [21] examined the availability of illegal narcotics during the first wave of the COVID-19 pandemic using a sample of 535 posts.

¹Consortium website: <https://research.tuni.fi/enncode/>

²Permanent link to the corpus: <http://urn.fi/urn:nbn:fi:lb-2022062221>

³In their paper, Haasio et al. [15] refer to *Torilauta* using its other commonly used name *Sipulitori*.

It is notable that none of the previous studies attempted to share their data sets with the research community in a systematic manner. This practice negatively impacts the replication and verification of the published studies and potentially discourages further research on the topic. On the other hand, given the sensitive and potentially incriminating nature of the data, not releasing the data is an understandable approach since preparing and managing such a resource gives rise to multiple technical, ethical, and legal challenges. The purpose of this paper is to describe and discuss these challenges and how we approached them.

2.2. Related Corpora

To the best of our knowledge, there exist relatively few published dark web corpora or text data sets. Three notable exceptions include the Dark Net Market archives (2013–2015) [5], a collection covering 89 dark net markets and over 37 related forums (1.6TB uncompressed) scraped during 2013-2015, DUTA [2], a set of 7,000 text samples formed by sampling the Tor network for two months, and CoDa [20], a set of 10,000 web documents tailored towards text-based dark web analysis. All three corpora comprise primarily English texts and are either publicly downloadable (Dark Net Market archives) or available to researchers upon request (DUTA, CoDa).

Existing Finnish web forum corpora include texts collected from the Ylilauta imageboard [35] and Suomi24 social networking site [6]. While emphasizing user anonymity, both Ylilauta and Suomi24 forums operate on the clear web and strictly forbid illegal content. Both corpora are available for research purposes via the Language Bank of Finland under a Creative Commons (CC BY-NC 4.0) license.

The Finnish Internet Parsebank [22] is a large-scale syntactically analyzed text collection created using plain text webpage data made available by the Common-Crawl2 Internet crawl project. Due to the employed web crawling approach to data collection, the corpus is likely to contain web forum content.

Finally, in a recent study, Leedham et al. [23] discussed their work on archiving a hard-to-access WiSP corpus consisting of texts written by social work professionals describing their work practices. Due to the potentially sensitive nature of the texts, Leedham et al. [23] created two versions of the corpus: one for the research project and an anonymized/reduced version for archiving. In a similar vein, our work presented here aims to provide an extensive discussion on the process of preparing a corpus of potentially sensitive texts for archiving and sharing.

3. Data

In this section, we describe the original, unmodified corpus received by the ENNCODE consortium. We provide an overview of the data as a whole and discuss in more detail the contained data fields, post lifespans, and thread and message lengths in Sections 3.1, 3.2, 3.3, and 3.4, respectively.

3.1. Overview

The original data set received by the consortium included all posts submitted to Torilauta between 2019-09-11 and 2020-05-20 (1,863,639 posts in 251 days) and 2020-06-17 and 2020-10-31 (1,099,710 posts in 136 days). In addition to the posts collected during these active collection periods, the data contained “residue” posts submitted between 2017-11-02 and 2019-09-11 (141,627 posts in 678 days). Meanwhile, posts submitted between 2020-05-20 and 2020-06-17 were missing completely. Therefore, the original unmodified corpus consisted of 3,104,976 posts in total.

Table 1 presents an overview of the post and thread frequencies in the data grouped by boards along with brief topic descriptions. Of the 32 boards, the board with the highest activity measured by the total number of submitted posts and threads was the market board dedicated to narcotics transactions within the city of Helsinki (*/hki*). The total number of posts submitted to this board was 787,459 corresponding to 25.4% of all posts in the data. Meanwhile, in total 96.5% (2,997,624) of all posts were submitted to the 16 boards dedicated to transactions (denoted with (*market*) in the topic column of Table 1).

3.2. Data Fields

Each post in the corpus is represented as a data structure with 8 fields as shown in Table 2. Each field belongs to one of the following three types: a string, an integer, or a date. All dates are in timezone UTC+0 (GMT+0). Note that throughout this paper, we refer to a set of these 8 data fields as *post*, whereas the content of the text data field within a single post is referred to as *message* or *text body*. Missing values have different meanings depending on the field. For deletions, a missing value means the post was never deleted. The first post of a thread, referred to as the original post (OP), always has a missing *postId* value and is instead identified by its (*boardUri*, *threadId*) pair. The *subject* field is missing for 46.2% of all posts since it was common practice to omit the subject. Similarly, the *poster name* field is missing for 54.1% of all posts which is in line with the anonymous nature of the site and since any optional contact information, such as an instant messenger username, was often included in the message text body instead.

Finally, the posts submitted to Torilauta optionally contained an attached image. However, no images were included in the original data set. Moreover, the data fields comprising a post did not include information on whether the post contained an image or not.

3.3. Post Lifespans

Submitted posts were deleted from the site for three main reasons. First, the site hosted a fixed number of threads on each board at a given time and so inactive threads were regularly removed by an automatic pruning mechanism to make room for new, active threads. Second, posts which violated the site rules (e.g. spam) were removed by the site administration. Third, the site interface did not provide users with means to edit messages and, therefore, the only way to correct erroneous message content (e.g. typos, updates) was to delete the post and resubmit. Lastly, a small portion of posts were “pinned” by the site administration, that is, they were meant to stay available on the site indefinitely.

Table 1

An overview of the number of submitted posts by board. Board topics marked with *(market)* indicate that the board was dedicated to transactions.

board	topic	# posts	posts (%)	# threads	threads (%)
hki	City of Helsinki (market)	787,459	25.4	239,750	25.9
tre	City of Tampere (market)	435,463	14.0	152,410	16.5
vnt	City of Vantaa (market)	242,848	7.8	74,173	8.0
oulu	City of Oulu (market)	242,240	7.8	86,114	9.3
muut	Other areas (market)	232,513	7.5	29,293	3.2
tku	City of Turku (market)	232,225	7.5	75,121	8.1
esp	City of Espoo (market)	224,707	7.2	57,229	6.2
kpo	City of Kuopio (market)	155,304	5.0	27,903	3.0
jkl	City of Jyväskylä (market)	145,890	4.7	49,079	5.3
lhti	City of Lahti (market)	110,358	3.6	35,860	3.9
bulk	Bulk transactions (market)	65,220	2.1	17,070	1.8
vsa	City of Vaasa (market)	60,137	1.9	23,795	2.6
t	Dates	33,688	1.1	13,694	1.5
roi	City of Rovaniemi (market)	27,850	0.9	11,460	1.2
seka	Miscellaneous (market)	22,208	0.7	9,566	1.0
hm	Narcotics markets	18,280	0.6	2,204	0.2
b	Random	9,068	0.3	1,249	0.1
h	Narcotics	7,942	0.3	1,753	0.2
y	Jobs	7,579	0.2	2,532	0.3
pm	Mail orders (market)	7,023	0.2	2,074	0.2
a	Everyday/mundane	6,772	0.2	1,351	0.1
hox	Hormones (market)	6,180	0.2	2,137	0.2
hax	Hacking	4,197	0.1	1,203	0.1
kkk	Cultivation	3,725	0.1	1,349	0.1
meta	Meta discussion	3,646	0.1	827	0.1
test	Testing	2,927	0.1	2,240	0.2
spam	Spamming	2,495	0.1	2,495	0.3
rotta	Vendor feedback	2,220	0.1	314	0.0
tt	Health	1,838	0.1	248	0.0
k	Getting and staying sober	1,562	0.1	247	0.0
fap	Porn	1,323	0.0	343	0.0
pgp	PGP public keys	89	0.0	2	0.0
total		3,104,976	100.0	925,085	100.0

There are two known caveats related to the deletion timestamps. First, while the data set included the creation and deletion times of posts, it unfortunately did not include information about the reason for the deletion. Second, all posts deleted during the pause in collection 2020-05-20 - 2020-06-17 had their deletion value marked as missing and, therefore, appeared as if they were not deleted. The amount of these potentially erroneous missing values was, however, relatively small and 97.22% of all posts (3,104,976) in the data had reliable deletion time information.

Table 2

Data fields comprising a single post. The column titled *missing (%)* indicates the portion of all posts where the field value is not available.

	description	example	missing (%)
boardUri	board identifier	roi	0.0
creation	post creation datetime (UTC)	2020-01-14T17:51:24.714Z	0.0
deletion	post deletion datetime (UTC)	2020-01-27T16:49:03.663Z	2.8
threadId	thread identifier	27961	0.0
postId	post identifier	28069	29.8
name	poster name	example-name	54.1
subject	message subject	Example message subject	46.2
message	message text body	Example message text body	0.0

Finally, we estimated the median lifespans of submissions to the market and non-market boards to be 23 and 238 hours, respectively. The difference was mainly due to the lower posting frequency and consequently lower thread pruning frequency of the non-market boards. For noise filtering purposes, we are mostly interested in the messages with short lifespans. To this end, we note that 5% of all messages had a lifespan of less than 32 minutes. Since posts with such short lifespans were likely removed by the user and resubmitted after minor modifications, they may be discarded as noise.⁴

3.4. Thread and Message Lengths

By and large, the threads in the market boards were rather short as the majority (57.9%) consisted of a single post (the thread has no reply and/or follow-up messages to the original post) and 99% of the threads have 23 posts or less. Similarly, on the non-market boards, 64.18% of the threads consisted of a single post and 99% of the threads have 35 posts or less.

In order to examine individual message lengths, we tokenized the message bodies using the tokenizer included in the Finnish Tagtools (v1.5) software developed at the University of Helsinki.⁵ The tokenizer splits running text into sentences and extracts punctuation and special characters from word bodies into separate tokens. The median length of original posts and reply/follow-up posts to the market boards were 36 and 6 word tokens, respectively. In other words, the original posts were typically short (a few sentences) and reply/follow-up posts even shorter (a few words). As for the reply/follow-up posts, this was because of the prevalence of short update posts, the purpose of which was to keep the thread alive and close to the front page. These types of posts comprised roughly one fifth of all market board reply/follow-up posts. Therefore, one can reduce the amount of noise in the FINDarC considerably by ignoring reply/follow-up posts with, e.g. less than 5 word tokens (or less than 20 characters) in the message body.⁶

⁴This is in agreement with the recommendation of the site administrator.

⁵Available at <http://urn.fi/urn:nbn:fi:lb-2021042102>

⁶This is in agreement with the recommendation of the site administrator.

4. Assessing the Frequency and Nature of Personal Data

In this section, we present our experimental work, the purpose of which was to assess the amount and types of personal and sensitive data in the FINDarC corpus. As the size of the corpus exceeded over 3,000,000 messages, it was deemed infeasible to curate all the posts manually given the consortium resources. In Section 4.2, we discuss experiments utilizing full-text search using hand-crafted regular expressions. Finally, the implications of the results on data reduction and data release are discussed in Section 5.

4.1. Manual Annotation

In this part of the experiments, we examined the types of personal data contained in the message bodies of posts using a manual annotation approach. In particular, we were interested in types of personal information found in the market board submissions which contained the majority of posts in the data set (97%) and represent the primary function of the site as a narcotics marketplace.

4.1.1. Definition of Personal Data

We adopted the Article 4(1) of the GDPR⁷ which defines personal data as

any information relating to an identified or identifiable natural person (data subject); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;

and divided potential identifiers into the following 5 classes

1. person_name (e.g. first name, family name)
2. id_number (e.g. the Finnish social security number)
3. location (e.g. street address, city, city district)
4. online_id (e.g. instant messaging username, email address, IP address)
5. other (e.g. phone number, identifying physical appearance)

Moreover, GDPR Article 9 (1)⁸ further defines the *special categories of personal data* as

Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.

For cases belonging to this group, we added a sixth potential identifier class

⁷ <https://gdpr-info.eu/art-4-gdpr/>

⁸ <https://gdpr-info.eu/art-9-gdpr/>

6. special_category

Given the personal data class specifications, the annotation task then comprised two subtasks: entity span detection and class assignment on a word token level to continuous, non-overlapping sequences of tokens.⁹ For example, consider

```
priimaa kukkaa [Helsingin keskustassa]location ! yhteydenotot Wickerillä W //  
[example-name]online_id  
  
excellent bud [in the centre of Helsinki]location ! contact Wicker W // [example-  
name]online_id
```

4.1.2. Personal Data versus Named Entities

The definition of personal data and the personal data classes described in the previous section are closely related to named entities [14, 30], in particular, person and location names. However, while there is an overlap between the personal data classes and the named entities, they are not equivalent nor is one a subset of the other. This is because of the following two main reasons. First, a person/location name mentioned in running text is almost invariably identified as a named entity independent of the context. For example, consider the sentence fragment

Sauli Niinistö on Suomen presidentti.

Sauli Niinistö is the President of Finland.

where *Sauli Niinistö* and *Suomen (= Finland)* are defined as person and location named entities, respectively, according to the Finnish NER guidelines [29, 25]. In contrast, neither of these should be interpreted as personal data according to the GDPR as stating the name and occupation of a well-known public figure does not breach their privacy. Second, not all pieces of personal name or location information found in a message will be classified as named entities. This is because the definition of personal data takes into account a wider context than what is directly available in the message. In other words, any piece of information found in text relating to an individual can be considered personal data if it is reasonable to assume that the information in question could be combined with other information to identify that individual. For example, consider the modified version of a post in corpus presented in Table 3. From the information contained in this exemplary post, one can infer that

1. at 12:12 (UTC) on January 1st 2019 (according to the post creation timestamp)
2. a person using the Wicker username *example-name* (according to the contact information included in the message)
3. was selling illegal narcotics (according to subject and message) in Helsinki area (according to the post boardUri) and
4. more specifically in the Eastern part of Helsinki and
5. with a high likelihood near the metro track

In consequence, we would regard the mentions of East (*idässä*) and metro (*metro*) in the message body as personal location data. In contrast, these mentions would not be considered named location entities according to the Finnish NER annotation guidelines.

⁹The tokenization of messages was acquired using the Finnish Tagtools (ver. 1.5) toolkit.

Table 3

A modified example post containing personal data.

field	value
boardUri	hki
creation	2019-01-01T12:12:12.123Z
deletion	2019-02-02T13:13:13.345Z
threadId	12345
postId	23456
name	[missing]
subject	kukkaa myynnissä (<i>selling bud</i>)
message	kukkaa myynnissä idässä. kuljetus metron lähelle. Wicker // example-name (<i>selling bud in the east. delivery nearby metro. Wicker // example-name</i>)

4.1.3. Manual NER annotation

The results of a preliminary manual annotation showed, rather unsurprisingly, that the majority of personal information cases consisted of instant messenger usernames and areas posted as potential settings for face-to-face transactions. On the other hand, the preliminary experiment also revealed that even native Finnish speakers can struggle with the text domain and that manually curating large parts of the corpus would be an excessively tedious effort. On the other hand, this means that automatic text processing methods are also likely to struggle with the text domain.

4.1.4. Automated NER annotation

Text anonymization/reduction approaches proposed in literature commonly utilize automatic NER as a part of the processing pipelines to varying extents [8, 32, 1, 27, 7, 11, 13, 12]. Ideally, NER tools would also be useful when processing FINDarC as they could, in principle, be employed to automatically detect some direct personal identifiers, such as names and addresses. However, we found this approach problematic to implement as the number of posts marked with person and location entities was hundreds of thousands. While substantially smaller than the original data set of over 3,000,000 posts, this set was still too large to be curated manually given the available consortium resources. Examining the predictions on the manually annotated data set, suggested that the available tools suffered from a domain mismatch in addition to the inherent mismatch between personal data and named entity classes. This was not a completely surprising outcome since the text domain could also cause problems for human annotators. Because the tools tended to miss entities of interest (low recall) but also be incorrect when detecting entities (low precision), we did not consider them efficient curating tools for FINDarC in their current states and instead continued to the full-text search experiments presented in Section 4.2.

4.2. Full-Text Search

In this section, we assess the amount and types of personal data in the data set using a full-text search approach. In particular, we are interested in finding common personal identifiers with relatively rigid formats, such as social security numbers and phone numbers. We describe the experiment setup in Section 4.2.1, the applied textual patterns (regular expressions) in Section 4.2.2, and the obtained results in Section 4.2.3.

4.2.1. Setup

We define a target set of textual patterns (regular expressions), search for matches in message bodies. Specifically, we are interested in finding expressions matching

1. (Finnish) social security numbers
2. (Finnish) phone numbers
3. Email addresses
4. IBAN bank accounts
5. IP addresses

all of which have relatively rigid formats. The employed regular expressions are presented in Section 4.2.2. We apply the search to all posts in the data and assign the matches manually to personal data and non-personal data according to post context. The pattern matching is performed using MongoDB (v.5.0) full text search.¹⁰ In contrast to the manual annotation experiment discussed in Section 4.1, we do not filter out noise from the data and instead apply the search to all 3,104,976 posts in the original corpus.

4.2.2. Regular Expressions

In what follows, we provide brief descriptions of the applied regular expressions.

Social security numbers The Finnish social security number (SSN) is a sequence of 11 characters assigned to individuals by the Finnish government based on their date of birth and gender. The first 10 characters of the sequence are 6 numbers (date of birth) followed by a hyphen or A, followed by 3 numbers. The last character is alphanumeric, i.e., a number or a letter. Valid sequences likely have, therefore, format "121212-1234" and "121212-123A". We detect the sequences using the regular expression `\d\d\d\d\d\d\d\d\d\d[-\d\d\d[a-zA-Z0-9]]`. Persons born in the 2000s, who would have an "A" instead of hyphen, were not found in the sample.

Phone numbers According to the specification of the Finnish telephone network numbering, Finnish mobile phone numbers begin with a routing number (04-, 050, or 059) and are followed by a subscriber number, such as, "040 1234567", "059 1234567", and so forth.¹¹ The first zero ("0") of the number can optionally be replaced by the country code of Finland +358 (e.g. "+358

¹⁰<https://www.mongodb.com/>

¹¹Specification of numbers in the Finnish phone network is available at: <https://www.finlex.fi/viranomaiset/normi/480001/47180>

40 1234123”, ”+358 59 4321432”, etc.). Based on a preliminary examination of the data set, we detect common phone number formats using two regular expressions: `[\+]?358[\-\s]?0[45][\-\s]?[0-9]{7}` which matches numbers starting with the country code and `0[45][\-\s]?[0-9]{7}` which detects numbers with the country code omitted. Moreover, the expressions detect most commonly used grouping patterns using hyphens (e.g. ”+358-40-12345-567”) and whitespaces (e.g. ”059 123 4567”). While the subscriber part of the number can, in principle, vary in length, the patterns match the most common length of 7 digits. Landline numbers would be shorter but follow the same principles; none were however found in the data.

Email addresses According to the RFC 5322 standard¹², an email address as an identifier which contains a locally interpreted string followed by the at-character (“@”) followed by an internet domain, such as ”name@domain.com”, ”firstname.surname@subdomain.domain.com”, and ”underscore-hyphen-plus+sign@domain.com”. We detect the addresses using a regular expression `\S+@\S+\.\S+` which successfully detects all the above examples from a running text.

IBAN bank accounts We search for bank account numbers matching the International Bank Account Number (IBAN) structure specified by the ISO 13616-1:2020 standard¹³. The IBAN formatted numbers consist of the Finnish bank account number (14 digits) preceded by a two letter country code (“FI” for Finland) and two check digits (e.g. ”FI72 1234 5678 1234 12”). We detect the pattern using the regular expression `[Ff][Ii][0-9]{14}` which takes into consideration the letter case of the country code and the commonly used grouping whitespaces.

IP addresses IP (internet protocol) addresses are unique addresses which identify devices on the internet and local networks. We search for IP addresses using the following regular expression `(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)3(25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)` which matches patterns such as 88.777.66.555 and so forth.

4.2.3. Results

The frequencies of matched social security numbers, phone numbers, email addresses, bank account numbers, and IP addresses are presented in Table 4. As shown, the most and least frequent matched types were email addresses and bank account numbers with 1,840 and 12 regular expression matches, respectively. Due to the sufficiently low number of original matches, we were able to perform manual verification of all the cases, presented also in Table 4. According to this inspection, the phone numbers and email addresses occurred in two contexts. First, similarly to the instant messaging usernames, 491 out of 858 and 1,622 out of 1,837 of the phone numbers and email addresses, respectively, were posted as contact information by individuals themselves. The remaining cases were posted as a means of targeting people. In

¹²The RFC 5322 specification is available at: <https://datatracker.ietf.org/doc/html/rfc5322>

¹³<https://www.iso.org/standard/81090.html>

Table 4

Matched regular expression frequencies. The columns titled *matches* and *verified* denote the number of found regular expression matches and the number of manually verified cases, respectively. The columns titled *posts* and *threads* denote the number of distinct posts and threads where the verified cases occurred.

	matches	verified	posts
phone	875	858	699
hetu	91	73	65
email	1,840	1,837	1,707
iban	12	12	12
ip_address	121	16	14
total	2,939	2,796	2,261

such cases, personal details (e.g., name, relationship information, area of residence) were shared in connection with one or more usernames, in order to paint the person as a potential target for violence. Bank account numbers occurred similarly in two contexts. Out of the 16 IP addresses, 10 cases were included as a means of targeting, while the remaining 6 were provided as a type of contact information. Finally, all 73 and 12 found cases of social security numbers and bank account numbers were posted with a purpose of targeting. Thus, we identified in total 667 cases of targeting in 295 posts using this method.

Finally, we created a second regular expression list using words and prefixes related to the personal information contained in the identified 295 targeting posts. This list consisted of 77 keywords and parts of person names and addresses.¹⁴ After performing a second search with these patterns and a subsequent manual inspection, we identified an additional set of 166 posts submitted as a means of targeting.

4.2.4. Discussion

Similarly to automatic NER, rule-based search for text patterns is a commonly used part of text anonymization pipelines to varying extents [8, 32, 1, 27, 7, 11, 13, 12]. For multiple cases, such as email addresses and social security numbers, it is rather straightforward to write the patterns as regular expressions. Moreover, one can efficiently find pattern matches from large data sets using suitable databases, such as the MongoDB, or search engines. However, the caveat of examining the results is, of course, that manually verifying the matches only provides us with an estimate of the precision of the search while neglecting recall. In other words, one can not estimate the number of cases missed by the search in the whole corpus without manually curating thousands or, preferably, tens of thousands, of posts. Unfortunately, an annotation effort of this scale was not feasible given the available resources of the consortium. Nevertheless, using the search combined with manual examination of the search results, we were able to uncover a set of 461 posts submitted with a purpose of targeting individuals. While extremely rare compared to the total number of posts in the corpus, these posts are particularly interesting

¹⁴We do not present the list here due to obvious privacy issues.

from the point of view of data reduction and will be discussed further in Section 5.

5. Data Release

In this section, we draw on the discussion and experimental results presented in the previous sections and present an outline of the FINDarC data release.

Data Reduction Conventionally, the most direct approach to protect data subjects from re-identification has been to anonymize the data by removing/obscuring the parts containing personal information. [26] However, it appears evident that, if implemented successfully, this type of processing would have a profound impact on the usefulness of FINDarC for research purposes. For example, subsequent to removing usernames from their post contexts or from the data altogether, one would not be able to replicate the study of Hämäläinen et al. [16] who examined how sellers and buyers of illegal drugs represent themselves in their usernames. In turn, subsequent to removing location and/or timestamp data, one would no longer be able to replicate the study of Karjalainen et al. [21] who studied the availability of drugs specifically in the city of Tampere during the COVID-19 epidemic in the spring of 2020. From a utility point of view, therefore, it could be argued that reducing personal information from the buy/sell post threads would quickly degrade, or destroy, the usefulness of the corpus as a data source for research. This problem is generally referred to as the privacy-utility trade-off within the data privacy literature [24, 3].

Due to the problematic privacy-utility trade-off, we posit here that reducing the FINDarC extensively would not be appropriate even if sufficient resources could be allocated for domain-specific tool development and manual labour. Furthermore, we note that Torilauta and other drug trading sites have also been under observation by other parties, including both criminals and law enforcement agencies. Therefore, it is our assessment that leaving the sell/buy posts, which form the majority of the FINDarC, largely intact poses few additional risks to the studied populations. However, as discussed in 4.2, in addition to the sell/buy posts, the data also contains posts with the intention of doxxing/targeting individuals. Here, our position is that removing these submissions is warranted from an ethical point of view while not decreasing the value of the corpus as a data source significantly. This is because these posts are not directly related to the main functionality of the site as an online marketplace. Accordingly, we removed from the corpus all 461 posts containing identified doxxing/targeting information described in Section 4. The reduced corpus, therefore, comprises 3,104,515 posts. A summary of the removed posts by board is presented in Table 5.

Finally, as per the Terms of Service of Torilauta, the site users gave consent to data collection for academic use by using the site. Consequently, site users could opt out of the data collection by not submitting new posts and/or contacting the site administration about previously submitted posts. However, it could be argued that by removing a previously submitted post, a user has withdrawn the permission to use the data. Unfortunately, as discussed in Section 3.2, the original data set received from the site administration did not include information about the reasons behind post deletions. Therefore, we were not able to exclude any posts from the corpus based on the deletion status.

Table 5

Number of posts removed from FINDarC prior to archiving due to personal information used for targeting.

board	# posts
hm	113
b	110
esp	62
muut	41
hki	23
t	19
tre	14
a	10
rotta	10
tku	9
jkl	8
kpo	7
bulk	5
vnt	5
spam	4
seka	4
vsa	3
oulu	3
roi	2
hox	2
pm	1
meta	1
k	1
hax	1
h	1
fap	1
lti	1
total	461

Access Restrictions Due to the limited applicability of data reduction as a means of protecting data subjects from reidentification, we next discuss restricting the access to the corpus. In general, the Language Bank recommends sharing resources using standard CC-BY or other open source licenses, in which case only the metadata of the resource needs to be registered with the Language Bank, although the Language Bank also hosts openly available and publicly accessible resources (PUB) in its Language Bank Download service. However, since the FINDarC resource in its current form contains personal data, both copyright and personal data legislations apply and the corpus cannot be published with open access. Instead, FINDarC is provided via protected access under the CLARIN RES licence which means that permission to download and use the corpus is only granted to researchers based on written applications reviewed by the data controller (principal investigator of the ENNCODE consortium) including a data protection

impact assessment. The purpose of this limitation is to ensure that the material is accessed only by verified researchers for legitimate research purposes. It also lessens sharing-related risks to both the researchers and the subjects of study, as mandated by the consortium's data management policy. Restricting access to the corpus as described here is in line with the current literature on data sharing [26, 28, 31, 10, 9] which also acknowledges the limitations of data anonymization/reduction and encourages the use of user group limitations.

Corpus Version Control Resources deposited in the Language Bank may have several different variants (i.e. versions) which form a resource group. Typically, a resource group consists of different annotations (raw data or preprocessed data for a single corpus), accumulated data (the content is almost identical but one version has more or newer content), or repaired data (flaws or necessary modifications, e.g. justified requests to remove or stop processing data, which have been identified and fixed manually or automatically). As for FINDarC, we emphasize the importance of the third point and note that if the Language Bank receives a notification and/or request for removal of content on grounds of sensitive data from a user of the corpus, these requests will be reviewed and acted upon. In particular, the Language Bank may update the corpus by reducing the data further and only store and share the most recent version.

6. Conclusions

We discussed the archiving procedure of FINDarC, a Finnish dark web marketplace corpus, in the Language Bank of Finland. The discussion included an overview of the data, assessment of the risk and impact of data subject re-identification, assessment and implementation of viable data reduction approaches using manual and automatic text processing, assessment of privacy and security measures implemented by the Language Bank of Finland, and a future corpus management plan implemented and coordinated by the Language Bank of Finland. As a result of the presented work, a reduced version of the corpus has been archived in the Language Bank of Finland. Researchers can apply for access to the corpus under the CLARIN RES licence.

As this article shows, the data set was cleared using best practices for ethical proofreading, which consistently sought to prioritize the protection of the posters on the forum. Given how even indirect identifiers could be utilized against the site's users by either law enforcement or other members of the drug-using community, it was necessary to opt for maximal removal efficiency whenever possible. Nevertheless, the amount of data is so high that no clearing can be considered ethically sufficient for the purpose of releasing the data openly, so we opted for a gatekeeping approach in addition to clearing everything that could be found.

Acknowledgments

References

- [1] Adams, A., E. Aili, D. Aioanei, R. Jonsson, L. Mickelsson, D. Mikmekova, F. Roberts, J.F. Valencia, and R. Wechsler 2019. AnonyMate: A toolkit for anonymizing unstructured chat data. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pp. 1–7.

- [2] Al Nabki, M.W., E. Fidalgo, E. Alegre, and I. De Paz 2017. Classifying illegal activities on tor network based on web textual contents. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 35–43.
- [3] Alvim, M.S., M.E. Andrés, K. Chatzikokolakis, P. Degano, and C. Palamidessi 2011. Differential privacy: on the trade-off between utility and information leakage. In *International Workshop on Formal Aspects in Security and Trust*, pp. 39–54. Springer.
- [4] Artstein, R. 2017. Inter-annotator Agreement. *Handbook of Linguistic Annotation*, pp. 297–313. Springer.
- [5] Branwen, G., N. Christin, D. Décary-Hétu, R.M. Andersen, StExo, E. Presidente, Anonymous, D. Lau, D.K. Sohlzl, V. Cakic, V. Buskirk, Whom, M. McKenna, and S. Goode. 2015, July. Dark net market archives, 2011-2015. <https://www.gwern.net/DNM-archives>. Accessed: 2022-06-28.
- [6] City Digital Group. 2021. Suomi24 virkkeet -korpus 2001-2020, Korp-versio [tekstikorpus]. Kielipankki. Available at <http://urn.fi/urn:nbn:fi:lb-2021101525>.
- [7] Csányi, G.M., D. Nagy, R. Vági, J.P. Vadász, and T. Orosz. 2021. Challenges and Open Problems of Legal Document Anonymization. *Symmetry* 13(8): 1490.
- [8] Di Cerbo, F. and S. Trabelsi 2018. Towards personal data identification and anonymization using machine learning techniques. In *European Conference on Advances in Databases and Information Systems*, pp. 118–126. Springer.
- [9] Elliot, M., E. Mackey, and K. O’Hara. 2020. The anonymisation decision-making framework 2nd Edition: European practitioners’ guide.
- [10] Elliot, M., K. O’hara, C. Raab, C.M. O’Keefe, E. Mackey, C. Dibben, H. Gowans, K. Purdam, and K. McCullagh. 2018. Functional anonymisation: Personal data and the data environment. *Computer Law & Security Review* 34(2): 204–221.
- [11] Francopoulo, G. and L.P. Schaub 2020. Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP. In *workshop on Legal and Ethical Issues (Legal2020)*, pp. 9–14. ELRA.
- [12] Garat, D. and D. Wonsever. 2022, jan. Automatic Curation of Court Documents: Anonymizing Personal Data. *Information* 2022, Vol. 13, Page 27 13(1): 27. <https://doi.org/10.3390/INFO13010027>.
- [13] Glaser, I., T. Schamberger, and F. Matthes. 2021, jun. Anonymization of German legal court rulings. *Proceedings of the 18th International Conference on Artificial Intelligence and Law, ICAIL 2021*: 205–209. <https://doi.org/10.1145/3462757.3466087>.
- [14] Grishman, R. and B.M. Sundheim 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.
- [15] Haasio, A., J.T. Harviainen, and R. Savolainen. 2020, mar. Information needs of drug users on a local dark Web marketplace. *Information Processing and Management* 57(2): 102080. <https://doi.org/10.1016/j.ipm.2019.102080>.
- [16] Hämäläinen, L., A. Haasio, and J.T. Harviainen. 2021, aug. Usernames on a Finnish Online Marketplace for Illegal Drugs. *Names A Journal of Onomastics* 69(3). <https://doi.org/10.5195/NAMES.2021.2234>.
- [17] Hämäläinen, L. and T. Ruokolainen. 2021. Kukkaa, amfea, subua ja essoja: Huumausaineiden slanginimitykset Tor-verkon suomalaisella kauppapaikalla. *Sananjalka* 63: 130–153.

- <https://doi.org/10.30673/sja.106615>.
- [18] Harviainen, J.T., A. Haasio, and L. Hämäläinen. 2020, jan. Drug traders on a local dark web marketplace. *ACM International Conference Proceeding Series*: 20–26. <https://doi.org/10.1145/3377290.3377293>.
- [19] Harviainen, J.T., A. Haasio, T. Ruokolainen, L. Hassan, P. Siuda, and J. Hamari 2021, 1. Information protection in dark web drug markets research. Hawaii International Conference on System Sciences.
- [20] Jin, Y., E. Jang, Y. Lee, S. Shin, and J.W. Chung. 2022. Shedding new light on the language of the dark web. *arXiv preprint arXiv:2204.06885 (To appear in NAACL 2022)*.
- [21] Karjalainen, K., R. Nyrhinen, T. Gunnar, T. Ylöstalo, and T. Ståhl. 2021. Huumeiden saatavuus, käyttö ja huumausainerikollisuus Tampereella koronakeväänä 2020. *Yhteiskuntapolitiikka* 86(2): 80–90.
- [22] Laippala, V. and F. Ginter 2014. Syntactic n-gram collection from a large-scale corpus of internet finnish. In *Human Language Technologies-The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT*, Volume 268, pp. 184.
- [23] Leedham, M., T. Lillis, and A. Twiner. 2021. Creating a corpus of sensitive and hard-to-access texts: Methodological challenges and ethical concerns in the building of the WiSP Corpus. *Applied Corpus Linguistics* 1(3): 100011. <https://doi.org/https://doi.org/10.1016/j.acorp.2021.100011>.
- [24] Li, T. and N. Li 2009. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 517–526.
- [25] Luoma, J., M. Oinonen, M. Pyykönen, V. Laippala, and S. Pyysalo 2020. A Broad-coverage Corpus for Finnish Named Entity Recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, pp. 4615–4624. European Language Resources Association.
- [26] Ohm, P. 2009. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review* 57: 1701.
- [27] Oksanen, A., M. Tamper, J. Tuominen, A. Hietanen, and E. Hyvönen 2019. AnoPpi: A pseudonymization service for Finnish court documents. In *JURIX 2019*, pp. 251–254. IOS Press.
- [28] Rubinstein, I.S. and W. Hartzog. 2016. Anonymization and risk. *Wash. L. Rev.* 91: 703.
- [29] Ruokolainen, T., P. Kauppinen, M. Silfverberg, and K. Lindén. 2019, aug. A Finnish news corpus for named entity recognition. *Language Resources and Evaluation* 2019 54:1 54(1): 247–272. <https://doi.org/10.1007/S10579-019-09471-7>. arXiv:1908.04212.
- [30] Sang, E.T.K. and F. De Meulder 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147.
- [31] Stalla-Bourdillon, S. and A. Knight. 2016. Anonymous data v. personal data-false debate: an EU perspective on anonymization, pseudonymization and personal data. *Wis. Int'l LJ* 34: 284.
- [32] Tamper, M., A. Oksanen, J. Tuominen, E. Hyvönen, A. Hietanen, and Others 2018. Anonymization Service for Finnish Case Law: Opening Data without Sacrificing Data Protection and Privacy of Citizens. In *International Conference on Law via the Internet*,

LVI.

- [33] Virtanen, A., J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, and S. Pyysalo. 2019. Multilingual is not enough: Bert for Finnish. *arXiv preprint arXiv:1912.07076*.
- [34] Weischedel, R., M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, and Others. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA 23*.
- [35] Ylilauta. 2016. Ylilauta-korpuksen ladattava versio [tekstikorpus]. Kielipankki. Available at <http://urn.fi/urn:nbn:fi:lb-2016101210>.