Protective Measures for Sharing the Finnish Dark Web Marketplace Corpus (FINDarC)

Krister Lindén
Department of Digital Humanities
University of Helsinki, Finland
krister.linden@helsinki.fi

Teemu Ruokolainen
Faculty of Information Technology
and Communication Sciences
Tampere University
Tampere, Finland
teemu.ruokolainen@tuni.fi

Lasse Hämäläinen
Faculty of Information Technology
and Communication Sciences
Tampere University
Tampere, Finland
lasse.hamalainen@tuni.fi

J. Tuomas Harviainen
Faculty of Information Technology
and Communication Sciences
Tampere University
Tampere, Finland
tuomas.harviainen@tuni.fi

Martin Matthiesen
CSC - IT Center for Science
Espoo, Finland
martin.matthiesen@csc.fi

Mietta Lennes
Department of Digital Humanities
University of Helsinki, Finland
mietta.lennes@csc.fi

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. A reduced version of the corpus, Finnish Dark Web Marketplace Corpus (FINDarC), has been archived in the Language Bank of Finland. In the current work, we focus on the protective measures for storing the data and how researchers can apply for access rights to the corpus under the CLARIN RES licence.

1 Introduction and Background

Torilauta was a 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. 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 posts was collected and handed over to the ENNCODE consortium[1] by the site administration to be archived and shared for research purposes, as permitted by the site's Terms of Service. To promote the FAIR data principles, 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 offering a time-restricted personal license to re-use the data according to an approved research plan.[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 previous papers (Lindén et al., 2023a, 2023b), 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

[1] Consortium website: https://research.tuni.fi/enncode/

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

of Finland, and to put in place a future corpus management plan coordinated by the Language Bank of Finland. In this paper, we add information on the protective measures that are applied when storing the data in the Language Bank of Finland and what researchers need to do to get access to the data.

Those carrying out future research based on the corpus are encouraged to implement appropriate ethical proofreading measures (see e.g. (Harviainen et al., 2021)) in order to further mitigate any potential harm from access to the material, to both the researchers and the studied populations.

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. For a discussion of the privacy-utility trade-off within the data privacy literature, see Alvim et al. (2011) and Li and Li (2009).

2.1 Previous Studies on Torilauta

Prior to the data publication presented here, Torilauta was utilized in the ENNCODE project as a data source in multiple linguistic and social science studies (Haasio et al., 2020; Hämäläinen & Ruokolainen, 2021; Hämäläinen et al., 2021; Harviainen et al., 2020; Karjalainen et al., 2021). In particular, Haasio et al. (2020) examined information needs of drug users using a sample of 9,300 posts.[3] Harviainen et al. (2020) studied cultural and socioeconomic aspects of drug traders using the same 9,300 post sample. Hämäläinen and Ruokolainen (2021) studied narcotic substance vocabulary based on a sample of 3,000 posts. Hämäläinen et al. (2021) studied a sample of 1,654 usernames extracted from posts submitted to the site. Karjalainen et al. (2021) examined the availability of illegal narcotics during the first wave of the COVID-19 pandemic using a sample of 535 posts.

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) (Branwen et al., 2015), a collection covering 89 dark net markets and over 37 related forums (1.6TB uncompressed) scraped during 2013-2015, DUTA (Nabki et al., 2017), a set of 7,000 text samples formed by sampling the Tor network for two months, and CoDa (Jin et al., 2022), 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 (Ylilauta, 2016) and Suomi24 social networking site (City Digital Group, 2021). 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 (Laippala & Ginter, 2014) 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.

In a recent study, Leedham et al. (2021) 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. (2021) created two versions of the corpus:

[3] In their paper, Haasio et al. (2020) refer to *Torilauta* using its other commonly used name *Sipulitori*.

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 of the process of preparing a corpus of potentially sensitive texts for archiving and sharing.

3 Data Set Description

This section provides an overview of the data as a whole and discusses the post lifespans in more detail.

3.1 Overview

Table 1: 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

Each post in the corpus is represented as a data structure with 8 fields as shown in Table 1. 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*.

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.

The data is grouped by boards. 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.

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.

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

3.2 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.

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.

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]

4 Data release

Text anonymization approaches proposed in the literature commonly utilize automatic named-entity recognition (NER) as a part of the processing pipelines to varying extents (Adams et al., 2019; Csányi et al., 2021; F. & Trabelsi, 2018; Francopoulo & Schaub, 2020; Garat & Wonsever, 2022; Glaser et al., 2021; Oksanen et al., 2019; Tamper et al., n.d.). Ideally, NER tools would also have been useful when processing FINDarC. However, examining the prediction quality on a manually annotated test section of the 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 completely surprising since the text domain also caused problems for human annotators when creating the test section. Because the available tools tended to miss entities of interest (low recall) and be incorrect when detecting entities (low precision), we did not consider them efficient pre-processing tools for FINDarC in their current state.

4.1 Common Personal Identifiers

Instead of using NER tools, we decided to use full-text search, to find common personal identifiers with relatively rigid formats, such as social security numbers and phone numbers. We defined a target set of textual patterns (regular expressions) and searched for matches in message bodies. Specifically, we were 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. We applied the search to all posts in the data and assigned the matches manually to personal data and non-personal data according to post context. We did not filter out noise from the data and instead applied the search to all 3,104,976 posts in the original corpus.

4.2 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,

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

therefore, format ”121212-1234” and ”121212-123A”. We detect the sequences using the regular expression `\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.[5] The first zero (”0”) of the number can optionally be replaced by the country code of Finland +358 (e.g. ”+358 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]?\d\d\d[\-\s]?\d[\-\s]?\d\d\d` which matches numbers starting with the country code and `0[45]\d[\s\-]?\d\d\d[\-\s]?\d[\-\s]?\d\d\d` 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 standard[6], 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 standard [7]. 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]\d\d[\s\-]?\d\d\d\d[\s\-]?\d\d\d\d[\s\-]?\d\d\d\d[\s\-]?\d\d` 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 identity 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.

4.3 Search Results

The frequencies of matched social security numbers, phone numbers, email addresses, bank account numbers, and IP addresses are presented in Table 2, which shows that 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.

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 the individuals themselves. The remaining cases were posted as a means of targeting people. In 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.

[5] Specification of numbers in the Finnish phone network is available at: https://www.finlex.fi/fi/viranomaiset/normi/480001/47180

[6] The RFC 5322 specification is available at: https://datatracker.ietf.org/doc/html/rfc5322

[7] //www.iso.org/standard/81090.html

Table 2: 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

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 and removed in total 667 cases of targeting by removing 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.[8] After performing a second search with these patterns and a subsequent manual inspection, we identified and removed the additional 166 posts submitted as a means of targeting. In conclusion, posts with personal information concerning the submitting individual were kept while 461 posts aimed at targeting others were removed.

5 Protective measures

In the following, we discuss the consequences of anonymizing or pseudonymizing the data, before we explore an alternative approach.

5.1 Consequences of Anonymization

Conventionally, the most direct approach to protecting data subjects from re-identification has been to anonymize the data by removing/obscuring the parts containing personal information (Ohm, 2009). This process aims at potentially being able to release the data to the public. 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. (2021) 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. (2021) 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 (Alvim et al., 2011; Li & Li, 2009). However, in case anonymisation is not an option, there are other means to protect the data which minimize the risk of additional exposure for the data subjects and which can therefore justify the use of the data. The efficacy of the protective measures can be evaluated with the help of a data protection impact assessment.

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

[8] We do not present the list here due to obvious privacy issues.

also been under observation by other parties, including both criminals and law enforcement agencies. Therefore, it is our assessment that, if restricting access to academic research, leaving the sell/buy posts, which form the majority of the FINDarC, largely intact poses few additional risks to the studied populations as they had entered their data for public use. However, 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 significantly decreasing the value of the corpus as a data source. This is because these posts are not directly related to the main functionality of the site as an online marketplace. Accordingly, we removed all 461 posts containing identified doxxing/targeting information from the corpus The reduced corpus, therefore, comprises 3,104,515 posts.

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, 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.

5.2 Technical and Organisational Measures to Protect the Data

Due to the limited applicability of data reduction as a means of protecting data subjects from re-identification, we instead need to restrict access to the corpus. Since the FINDarC resource in its current form contains personal data, both copyright and personal data legislation apply and the corpus cannot be published with open access. Instead, FINDarC has 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) when including a data protection impact assessment of the intended research. 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.

Whenever researchers wish to deposit a resource in the Language Bank of Finland, the Language Bank (formally represented by the University of Helsinki) negotiates a deposition license agreement with the researcher and/or their home organization, unless the resource to be deposited already has an open license. The agreement defines, e.g., the end-user license in the CLARIN framework, including the resource-specific data protection terms and conditions for resources that contain personal data. In the case of the FINDarC corpus, the original right holder and data controller authorized the Language Bank as the data controller responsible for the redistribution of the resource. Thus, the Language Bank has the right to independently maintain the resource and to process the applications from users without consulting the original right holder. In order for the Language Bank to accept the resource, the original research project was required to assess the potential risks involved in processing the data according to the instructions from their home organisation, and to present documentation of the extent and the rationale behind their process for minimizing personal data, which influenced the need for additional protective measures at the Language Bank.

Extra Security Measures to Protect the Data. In addition to the security measures described in the Language Bank's Core Trust Seal Certificate (see Core Trust Seal, 2022, R16), extra measures to ensure the security of the data have been taken. The data is encrypted at-rest according to the Language Bank encryption guidelines[9]. Each authorized person needs to provide a self-generated public key, the corresponding private key is secured by a password only known to the staff member. At present, a minimum of five members of FIN-CLARIN staff have access to the corpus. A sixth emergency access key is stored in clear text without a password in a physical safe at CSC. A change in personnel requires re-encryption to reach the minimum of authorized staff to be able to access the data, so that a former member of staff

[9]Language Bank Encryption Guideline: https://urn.fi/urn:nbn:fi:lb-202401191

no longer can access the re-encrypted data set. The emergency key and a minimum of five authorized staff ensures that the data is accessible even in case of staff absence and personnel changes.

Security Measures to Ensure Authorized Access. Restricting access to the FinDARC corpus as described in this paper is in line with the current literature on personal data sharing (Elliot et al., 2018, 2020; Ohm, 2009; Rubinstein & Hartzog, 2016; Stalla-Bourdillon & Knight, 2016) which adheres to the FAIR principles, while acknowledging the limitations of data anonymization/reduction and encouraging the use of user group limitations.

At the Language Bank we have implemented restricted access to corpora using the Language Bank Rights service[10], which is based on CSC's REMS service[11]. For FinDARC, an applicant is required to provide an application according to our policy for corpora containing personal data, where she explains the need for the corpus and the associated research project. Since the resource contains personal data, the applicant must also submit a public link to the openly available privacy notice about the personal data processing regarding the research purpose in question. In addition, the applicant needs to supply a self-generated public key via the same Language Bank Rights application in order to be able to receive a copy of the corpus in encrypted form[12]. The application is then evaluated by the data owners who approve or deny access. If access is granted, an authorized member of staff decrypts a copy of the corpus using his private key and re-encrypts it using the key provided by the applicant. The encrypted data set is then sent to the applicant. This method ensures that only the holder of the corresponding private key, the applicant, can open the delivered copy.

Limitations of the Approach. While it is unlikely that all six access keys are lost, it is not impossible. In that case the data is not recoverable from within the Language Bank. Furthermore, we effectively prevent leaking of the data at-rest within the Language Bank and during transit to the researcher, but we cannot guarantee the secure handling of the data after the researcher has decrypted it. However, delivering the data set in encrypted form makes it easier for researchers to keep the data encrypted at-rest since they needed to set up the necessary tools and keys to receive the data in the first place.

Since most Universities use the same access credentials for university email which are used to access Language Bank Rights, identity theft is possible. In such a scenario, the attacker would need to monitor the applicants email account closely to remove automatic messages from the Language Bank Rights application and the Language Bank staff, as the messages might otherwise seem odd to the legitimate researcher. Two factor authentication would further mitigate the risk of identity theft.

6 Conclusions

We have discussed the archiving procedure of FINDarC, a Finnish dark web marketplace corpus, in the Language Bank of Finland. It was unlikely that the corpus could be fully anonymized to be shared publicly without also compromising its value for research, so instead other protective measures were taken to make it possible to share the data. 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 corpus management plan implemented and coordinated by the Language Bank of Finland outlining the protective measures applied in the Language Bank and the justifications a prospective researcher needs to produce to get access to the data to get access to the corpus under the CLARIN RES licence.

Acknowledgments

We acknowledge the funding for the Language Bank and FIN-CLARIN by the Research Council of Finland.

[10] See instruction for RES corpora in https://www.kielipankki.fi/support/access/

[11] For details on REMS see https://urn.fi/urn:nbn:fi:lb-2014120230

[12] See https://urn.fi/urn:nbn:fi:lb-2023051121

References

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

Alvim, M. S., Andrés, M. E., Chatzikokolakis, K., Degano, P., & Palamidessi, C. (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).

Branwen, G., Christin, N., Décary-Hétu, D., Andersen, R. M., StExo, E. P., Anonymous, D. L., Sohhlz, D. K., Cakic, V., Buskirk, V., Whom, M. M., & Goode, S. (2015). *July* [Dark net market archives, 2011-2015.]. https://www.gwern.net/DNM-archives

City Digital Group. (2021). Suomi24 virkkeet -korpus 2001-2020, Korp-versio. http://urn.fi/urn:nbn:fi:lb-2021101525

Core Trust Seal. (2022). *Core trust seal certificate for the language bank of finland* (tech. rep.). Core Trust Seal. https://www.coretrustseal.org/wp-content/uploads/2022/05/20220530-the-language-bank-of-finland_final.pdf

Csányi, G. M., Nagy, D., Vági, R., Vadász, J. P., & Orosz, T. (2021). Challenges and Open Problems of Legal Document Anonymization. Symmetry 13*(8): 1490*.

Elliot, M., Mackey, E., & O'Hara, K. (2020). In *The anonymisation decision-making framework 2nd Edition: European practitioners' guide*.

Elliot, M., O'hara, K., Raab, C., O'Keefe, C. M., Mackey, E., Dibben, C., Gowans, H., Purdam, K., & McCullagh, K. (2018). Functional anonymisation: Personal data and the data environment. *Computer Law & Security Review*, *34(2)*, 204–221.

F., D. C., & Trabelsi, S. (2018). Towards personal data identification and anonymization using machine learning techniques. *In European Conference on Advances in Databases and Information Systems*, 118–126.

Francopoulo, G., & Schaub, L. P. (2020). Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP. *workshop on Legal and Ethical Issues (Legal2020)*, 9–14.

Garat, D., & Wonsever, D. (2022). Jan. Automatic Curation of Court Documents: Anonymizing Personal Data. In *Information 2022, vol. 13* (pp. 1–27). https://doi.org/10.3390/INFO13010027

Glaser, I., Schamberger, T., & Matthes, F. (2021). Anonymization of German legal court rulings. In *Proceedings of the 18th international conference on artificial intelligence and law, icail 2021* (pp. 205–209). https://doi.org/10.1145/3462757.3466087

Haasio, A., Harviainen, J. T., & Savolainen, R. (2020). Information needs of drug users on a local dark Web marketplace. *Information Processing and Management*, *57*(2), 1016. https://doi.org/10.1016/j.ipm.2019.102080

Hämäläinen, L., Haasio, A., & Harviainen, J. T. (2021). Usernames on a Finnish Online Marketplace for Illegal Drugs. *Names - A Journal of Onomastics*. https://doi.org/10.5195/NAMES.2021.2234

Hämäläinen, L., & Ruokolainen, T. (2021). Kukkaa, amfea, subua ja essoja: Huumausaineiden slanginimitykset Tor-verkon suomalaisella kauppapaikalla. *Sananjalka*, *63*, 130–153. https://doi.org/10.30673/sja.106615

Harviainen, J. T., Haasio, A., & Hämäläinen, L. (2020). Drug traders on a local dark web marketplace. *ACM International Conference Proceeding Series*, 20–26. https://doi.org/10.1145/3377290.3377293

Harviainen, J. T., Haasio, A., Ruokolainen, T., Hassan, L., Siuda, P., & Hamari, J. (2021). Information protection in dark web drug markets research. *Hawaii International Conference on System Sciences*.

Jin, Y., Jang, E., Lee, Y., Shin, S., & Chung, J. W. (2022). *Shedding new light on the language of the dark web* [*arXiv preprint (To appear in NAACL 2022)*].

Karjalainen, K., Nyrhinen, R., Gunnar, T., Ylöstalo, T., & Ståhl, T. (2021). Huumeiden saatavuus, käyttö ja huumausainerikollisuus Tampereella koronakeväänä 2020. *Yhteiskuntapolitiikka*, *86*(2), 80–90.

Laippala, V., & Ginter, F. (2014). Syntactic n-gram collection from a large-scale corpus of internet finnish. *Human Language Technologies-The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT*.

Leedham, M., Lillis, T., & Twiner, A. (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). https://doi.org/10.1016/j.acorp.2021.100011

Li, T., & Li, N. (2009). On the tradeoff between privacy and utility in data publishing. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 517–526.

Lindén, K., Ruokolainen, T., Hämäläinen, L., & Harviainen, J. (2023a). Ethically archiving a hard-to-access massive research data set in the language bank of finland: The finnish dark web marketplace corpus (findarc) [Publisher Copyright: © 2023 Copyright for this paper by its authors.; Conference on Technology Ethics, Tethics ; Conference date: 18-10-2023 Through 19-10-2023]. In M. Rantanen, S. Westerstrand, O. Sahlgren, & J. Koskinen (Eds.), *Proceedings of the conference on technology ethics 2023 - tethics 2023* (pp. 114–131). CEUR-WS.org.

Lindén, K., Ruokolainen, T., Hämäläinen, L., & Harviainen, J. (2023b, December). Sharing the finnish dark web marketplace corpus (findarc) [CLARIN Annual Conference ; Conference date: 16-10-2023 Through 18-10-2023]. In K. Lindén, J. Niemi, & T. Kontino (Eds.), *Clarin annual conference proceedings* (pp. 134–139). CLARIN ERIC. https://www.clarin.eu/event/2023/clarin-annual-conference-2023

Nabki, A., M. W., E. F., Alegre, E., & Paz, I. D. (2017). Classifying illegal activities on tor network based on web textual contents. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 35–43.

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, *57*, 1701.

Oksanen, A., Tamper, M., Tuominen, J., Hietanen, A., & Hyvönen, E. (2019). AnoPpi: A pseudonymization service for Finnish court documents. In *Jurix 2019* (pp. 251–254). IOS Press.

Rubinstein, I. S., & Hartzog, W. (2016). Anonymization and risk. *Wash. L. Rev*, *91*, 703.

Stalla-Bourdillon, S., & Knight, A. (2016). Anonymous data v. personal data-false debate: an EU perspective on anonymization, pseudonymization and personal data. *Wis. Int'l LJ*, *34*, 284.

Tamper, M., Oksanen, A., Tuominen, J., Hyvönen, E., & et. al., A. H. (n.d.). Anonymization Service for Finnish Case Law: Opening Data without Sacrificing Data Protection and Privacy of Citizens. *International Conference on Law via the Internet, LVI*.

Ylilauta. (2016). *Ylilauta corpus downloadable version [text corpus]* (tech. rep.). Kielipankki. http://urn.fi/urn:nbn:fi:lb-2016101210