Anonymization and anonymized text data in statistical production

Matti Kokkonen, Statistics Finland, matti.kokkonen@stat.fi

Katja Löytynoja, Statistics Finland, katja.loytynoja@stat.fi

Henna Ylimaa, Statistics Finland, henna.ylimaa@stat.fi

Abstract

According to the EU General Data Protection Regulation (GDPR) data minimization principle, personal data should be limited to what is necessary in relation to the processing purposes. Current process of statistics on road traffic accidents partly relies on humans reading text documents. Anonymizing these documents while keeping them logical is a challenge. In this paper we:

1. *present briefly two tools for anonymization of free text fields*
2. *describe the results of testing of the anonymization tools*
3. *examine the effect of the anonymized data to statistical production*

The tested tools were Anoppi, a Ministry of Justice tool using language technology -based artificial intelligence and operated through RESTful web service, and NameFinder, a tool created in Statistics Finland, that uses a combination of machine learning, morphological analyzer, and name lists.

NameFiinder was tested with original statistical data. Anoppi was tested with simulated data based on investigation reports of Safety Investigation Authority of Finland (SIAF) and names from Population Information system. The SIAF data was selected for its similarity with the road traffic accident data. Tools produced confusion matrices.

NameFinder anonymized data was tested by simulating the steps of the statistical production process. Simulated steps include 1. checking the geo-positions of the accidents, 2. completing the tabular data with the data from the free text, and 3. controlling the tabular information by making free text queries. Anoppi could not yet be tested in statistical production because of data protection rules as the tool is physically not in Statistics Finland premises.

Correct anonymization percentage was nearly same in the tools. However, Anoppi produced less false positives while keeping the documents more logical and human readable as opposed to NameFinder. Anonymization had a small effect on geo-positioning since keywords were sometimes anonymized. However, the effect on the final statistics would be insignificant.

Keywords: *Anonymization, free text, General Data Protection Regulation, Machine learning, Natural language processing,*

1. Introduction

According to the EU General Data Protection Regulation (EU2016/679) data minimization principle, personal data should be limited to what is necessary in relation to the processing purposes. In this paper we present two different anonymization tools and compare their anonymization capabilities. We also present how we anonymized text documents that are used for production of statistics on road traffic accidents in Finland and what effect the anonymization had on the statistics.

Current process of statistics on road traffic accidents in Finland partly relies on humans reading text documents. These text documents contain a lot of personal data and at least part of it is sensitive personal information. Personal data per se should not be needed in statistical production and should not be needed in this use case either.

What is needed though is the possibility to connect the references of the individuals within the text data and outside to the existing tabular data. For the current statistics production it is also important to be able to connect the individuals to the information on the vehicles so that drivers and passengers of each vehicle can be placed in the correct vehicle in the tabular data.

2. Production of statistics on road traffic accidents

2.1. *Structure of the dataset*

Road traffic accident statistics in Finland uses the accident data from police as the main data source. Data from the police accident statistics consists of tabular data and text documents that describe the accident (figure 1.). These two datasets are connected by the case number and by the identity numbers of the individuals and registration number of the vehicles in accidents.

Text data is used in several production stages to supplement and correct the tabular data provided by the police. Other sources are also used to supplement the existing tabular data. The production stages where the text data is mainly used can be divided into two main stages: 1. geo-positioning and 2. supplementing and correcting the tabular data. In both stages the data handler will try to interpret the description written by the police and turn it into verified information of the accident. References to individuals can be names or part of names such as first name or last name which are coupled with an identity number in the beginning of the text document.

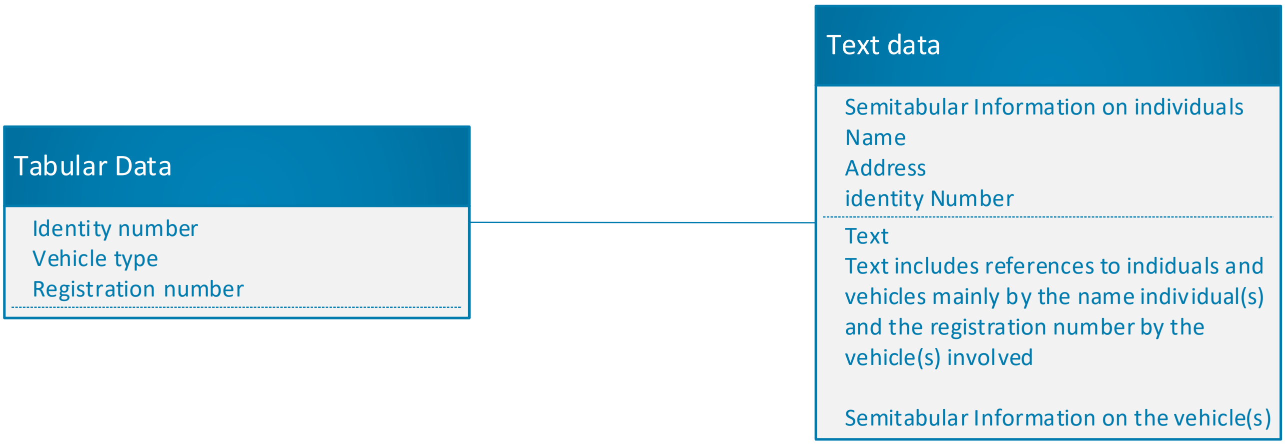

Tabular Data
Identity number
Vehicle type
Registration number
Text data
Semitabular Information on individuals
Name
Address
identity Number
Text
Text includes references to indiduals and vehicles mainly by the name individual(s) and the registration number by the vehicle(s) involved
Semitabular Information on the vehicle(s)

Figure 1. Structure of the data sets from police accident statistics.

2.2. *Geo-positioning of the road traffic accidents*

In the geo-positioning, the accident is first geo-positioned in an automated step. The handler's role is to verify and correct the geo-positions that weren't verified by the automatic process. The handler uses both the tabular and text data to determine the place of the accident. Manual positioning is done in a GIS-application (QGIS) case by case, and the handler benefits from all the geospatial information provided such as addresses, business names, known places, and landmarks. In this stage it is important that none or at least very little of the geospatial information is anonymized.

2.3. *Supplementing and correcting the tabular data*

In the supplementation and correction stage the data is supplemented and corrected manually by interpreting the text data and inputting it to a tabular form in a viewing program. The cases to be corrected are selected to a work queue by established rules. These rules include logical rules and rules that are derived from the values of other variables. Text data is also examined with selected keywords and the handler is prompted to go through the text if these keywords are found.

During this stage it is important to establish and maintain a connection between references to an individual within a document so that there is a connection between the semi tabular part and the text part for supplementation of the tabular data. In this step it is important to determine the vehicles and individuals that have been in the accident and place the individuals in the correct vehicles. There might also be information on individuals and companies that have experienced material damage to the vehicles or structures that they own, and these parties should be excluded from the tabular data concerning the accident.

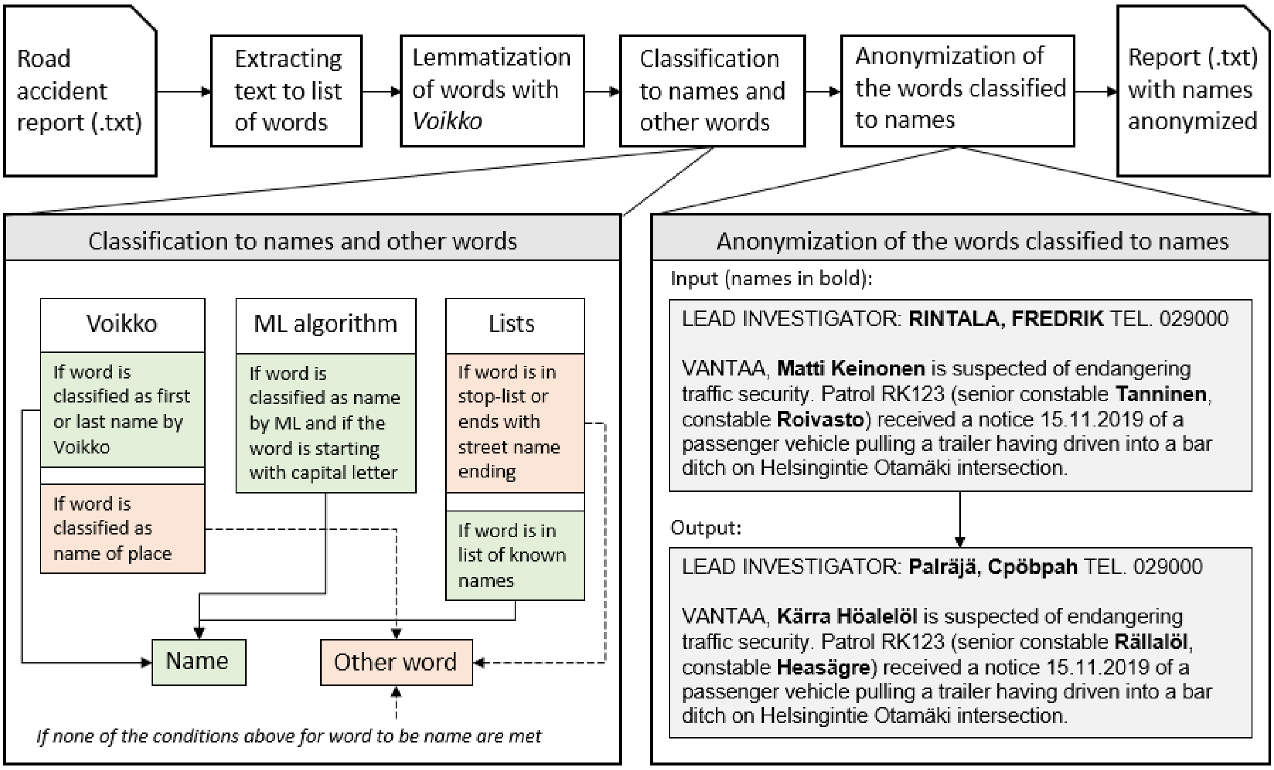
3. **Materials and methods**

Our initial goal was to anonymize the accident report data which is sensitive and contains a lot of personal information. For that purpose we developed our own in-house tool NameFinder which we were able to use without data protection concerns for the real dataset. The main purpose was to examine the effect on the road traffic accident statistics and the effect on the production process.

We later gained access to a externally developed anonymization tool Anoppi that is used through RESTful API, which meant we weren’t able to use it with the real production data as the statistical data cannot be handed over to an external party. We were able to compare the performance of the NameFinder and Anoppi with a simulated data set where the names were added retrospectively. This also meant that we were able to record the places of the names and accurately calculate confusion matrices.

3.1. *NameFinder*

Internally developed NameFinder -tool is in its final form a four step program (Figure 2.). It utilizes classification machine learning algorithms, externally developed NLP-tool Voikko for classification and lemmatization and two types of word-lists: one stop-list for words that are not names in this context and one acceptance-list for words that are considered names. The basic idea is to classify all the words given to it in two categories: names and other words. After the classification the tool will anonymize the words in the name category. We developed a few strategies for anonymization to best suit the needs of the handlers of the accident-data. NameFinder considers every word separately and does not have the ability to interpret the context of the word. NameFinder also does not form any connection between the found name-entities.


Road accident report (.txt)
Extracting text to list of words
Lemmatization of words with *Voikko*
Classification to names and other words
Anonymization of the words classified to names
Report (.txt) with names anonymized
Classification to names and other words
Voikko
If word is classified as first or last name by Voikko
If word is classified as name of place
ML algorithm
If word is classified as name by ML and if the word is starting with capital letter
Lists
If word is in stop-list or ends with street name ending
If word is in list of known names
Name
Other word
If none of the conditions above for word to be name are met
Anonymization of the words classified to names
Input (names in bold):
Output:

LEAD INVESTIGATOR: **RINTALA, FREDRIK** TEL. 029000

VANTAA, **Matti Keinonen** is suspected of endangering traffic security. Patrol RK123 (senior constable **Tanninen**, constable **Roivasto**) received a notice 15.11.2019 of a passenger vehicle pulling a trailer having driven into a bar ditch on Helsingintie Otamäki intersection.

LEAD INVESTIGATOR: **Palräjä, Cpöbpah** TEL. 029000

VANTAA, **Kärra Höalelöl** is suspected of endangering traffic security. Patrol RK123 (senior constable **Rällalöl**, constable **Heasägre**) received a notice 15.11.2019 of a passenger vehicle pulling a trailer having driven into a bar ditch on Helsingintie Otamäki intersection.

Figure 2. NameFinder workflow.

3.2. *Classification machine learning algorithm*

Machine learning algorithm that we used for this purpose was a MultinomialNB-model and it was trained with a two-part dataset containing most common words that are not names from the accident data and words that are names from the population statistics in such a way that one last name and one first name occurs once. The data set was balanced in a way that the dataset contained an equal number of name-words and word-words.

3.3. *Voikko NLP-tool*

Voikko, a free linguistic software and data for Finnish language, originally developed as spell checker for Open Office/LibreOffice, was also used in the tool. Voikko does morphological analysis, spelling and grammar checking and hyphenation, and includes a collection of linguistic data for Finnish. Voikko was used for lemmatization and also for name-word – word-word classification.

3.4. *Name-endings, name-lists and stop-lists*

It is essential for the process of road traffic accident statistics that street names and other positioning related names such as place names are not anonymized. For this reason words ending with ‘road’, ‘street’, or any word that refers to direction were not anonymized. The accident specific place names were also selected and used accident-by-accident bases from the structured part of the accident report

Furthermore, a stop-word list was added to the tool, that contains words that will be not anonymized. The list includes for example animal words that are common in accidents (moose, deer) and also as last names, that would be anonymized when in the beginning of a sentence where they have upper case first letter. The list also includes one hundred most common words in accident reports that are not names.

We also used a name-list with most common names that will always be anonymized and are not used as place names and are not names of any animals.

3.5. *Anonymization*

Anonymization of word strings recognized as names was first done by changing a letter to another systematically using ascii values and moving randomly generated number of letters to the left or right in the alphabet (figure 2). Later the letters were changed to shift from vowels to vowels and from consonants to consonants so that the anonymized names would be easier to read and remember. Which letters were chosen changed from one document to another. Names were anonymized from their lemmatized version. This is to ensure that a name is always the same regardless of the original declination. Characters that did not belong to Finnish alphabet were not changed. The first letters of the anonymized names were changed to be upper case for easier recognition by the handler.

3.6. *Anoppi*

Anoppi is a Ministry of Justice led project in collaboration with Semantic Computing Research Group (SeCo) at Aalto University, Centre for Digital Humanities (HELDIG) at University of Helsinki, and Edita Publishing (Oksanen et al. 2022). The project has created a tool for automated court decision anonymization using ML and NLP. The tool accepts multiple electronic formats of documents and is available as RESTful web service.

3.7. *Data used in testing Anoppi and NameFinder performance*

To test the performance of Anoppi and NameFInder in name anonymization we created a tailor made test data that consists of text in Finnish with names added in grammatically appropriate places. Investigation reports from Safety Investigation Authority of Finland (SIAF) were used as the base text. Investigation reports do not originally contain any names just references to people for example with pronouns or job titles.

The names used were handpicked from the Population Information System open data set so that both common and rare names were selected. Attention was also paid to that the selected names included both traditional Finnish names and foreign names as well as names with two parts and names with common meaning, for example Karhu (bear). Also a few very rare names outside the Population Information System data were added. The final collection of names, from which random sampling was made to add names to the report text, consisted of 75 last names and 50 forenames half of which were names for women and half names for men.

To place names in grammatically appropriate places in the text, words and suffixes were identified that can be replaced by names or after which names can be placed. Such places include, for example, personal pronouns and other words referring to persons, such as job titles. Combinations of names from the name collection were selected randomly and added to the suitable places found in the text. One name could be selected multiple times.

All necessary information about the added names was stored, such as the number of individual names and the name combinations added. Since Anoppi considers the first and last name combinations as one name, it must also be taken into account when examining the anonymization of the names.

4. **Results**

4.1. *Statistical production simulation using anonymized data*

In the simulation the statistical production phases were executed using anonymized data and the effect of anonymization to the results were reported. Of the statistical production phases, 1. geo-positioning and 2. supplementing and correcting the tabular data. were performed.

The effect of anonymization was tested twice. First round was performed with the base version of the NameFinder and the second round with the improved version of the NameFinder that was more adjusted for the text documents of the road traffic accident statistics.

4.2. *The effect of the anonymized data on statistical production*

In the anonymized data from the first version of the anonymization tool the most common Finnish names were identified well but more rare names produced more errors. Lemmatization with Voikko improved notably the identification of common declined names. However, Voikko's erroneous lemmatization influenced the anonymization of extra words. In addition, names common for both humans and streets resulted in street names being anonymized. Anonymization of animal species names that are also common as last names made the differentiation between humans and animals more complicated in the accident reports. These factors were taken into account when making road traffic accident statistics specific changes to the anonymization tool.

Most of the problems with accident positioning with anonymized data concerned misanonymization of street and location names and travel direction. Anonymization of street and location names made it harder to correctly locate the accident. Information of travel direction, e.g. compass point or starting place of travel are needed for example for identifying the correct line for the accident on a four lane road. If travel direction is unknown, it complicates the choice of correct lane and may affect location identification in other ways as well.

In some cases we could not be certain whether the accident report contained misanonymization that would affect geopositioning. The first version of the tool handled 508 accidents of which 42 (circa 8.3 %) seemed to have locating problems due to anonymization. Of these, 15 accidents had misanonymized address or location name, 17 did not have travel direction information, and 10 were likely missing relevant information.

Participant supplementation was affected by anonymization in reports where there were multiple participants, and subsequently multiple anonymizations. On the other hand, multiple participants are difficult to deal with with un-anonymized data as well. Error list checking was very little affected by anonymization. Also word searches were unaffected by anonymization

For the second simulation the tool was improved especially regarding place name identification. Street names and other positioning related names such as place names were intended to make more visible to users. Improved tool improved the impact of anonymization on positioning related names, but not fully eliminating the problems. Simulation was done using one 10 day accident data that was different from the first simulation. Altogether positioning was simulated in 220 cases and of these, in 10 cases (circa 4.5 %)

anonymization prevented exact positioning, and in 14 cases (6.4 %) anonymization could have affected the positioning results.

In the first simulation the main problem were words depicting directions, and road related words such as road names and for example the word ‘ditch’ being anonymized.

In the second simulation problems were related to more precise positioning, and for example anonymization of business names, slip road locations, or land mark names that were close to the accident.

There were no notable differences between simulations in participant information supplementation and word searches between the two versions of the NameFinder. The anonymization tool is currently coded to deal with mainly Finnish texts but due to it working on a word by word basis it performed well enough with the accident reports in Swedish.

4.3. *ANOPPI and NameFinder performance testing*

The anonymization performances of Anoppi was first tested with both inflected names and non-inflected names. In the test datasets there were approximately 150 added names and name combinations and the only difference between them were in the inflections. There were no major differences between the results when comparing the lists of found and anonymized names and their counts from Anoppi. Therefore testing could be done with non-inflected names which enables the use of larger test data if needed.

The final comparison for results was done between the stored information about non-inflected names added in the test data and the results gotten from Anoppi about the same data. There were 152 added names and name combinations from which Anoppi identified and anonymized 136 (89,5 %) correctly (Table 1). In total, Anoppi identified 141 names or name combinations from the test data, but five of them were false positive words. A total of 16 names remained unidentified. Most of the unidentified names were located in parts of the text where names are not common. In other words, in those parts the test data creation process failed to add the name in appropriate place. In most of those cases the same name was identified correctly somewhere else in the text. However, some of the names were unidentified regardless of where they were placed. Those were mostly names that are also used as words with other meanings, but still a couple of rare names were unidentified without any of the reasons previously presented. There are some examples of these presented in table 2.

NameFinder was then tested with the data put through Anoppi for comparison. NameFinder found all 152 names and thus outperformed Anoppi in finding true positives (Table 1). The downside was that NameFinder was less sensitive to false positives i.e. labelled more words as names (n=157) than Anoppi.

Table 1. Confusion matrix of name identification results for Anoppi and NameFinder. True negatives are not reported.

			Predicted Condition	
Anoppi			**Positive**	**Negative**
		n	141	-
Actual Condition	**Positive**	152	136	16
	Negative	0	5	-
NameFinder			**Positive**	**Negative**
		n	309	-
Actual Condition	**Positive**	152	152	0
	Negative	0	157	-

Table 2. Examples of names that were completely or partly unidentified or incorrectly identified, and some more difficult to identify names correctly identified by Anoppi. The number of names indicate frequency in population (Digital and Population Data Service Agency's Name Service), except for Incorrectly identified that are the number of false positives (FP) in the data. Last names are current names of living persons, fist names in red since birth year of 1899 both genders together.

Completely unidentified	**# of names**	**Partly unidentified**	**# of names**	**Correclty identified (examples)**	**# of names**	**Incorrectly identified (examples)**	**# of FP**
Melody	<159	Vorimo	23	Muhudin	13	Pohjanmaa	2
Yli-Liipola	13	Helly	< 431	Muhudin	13	Eija-Leena käsitys	1
Laxman	26	Latva-Kurikka	20	Dawoud	38	Koura	1
Anella	< 70	Karl-Johannes	< 22	Liiban	8		
Korkiakoski	1072	Eenokki	639	Muzaffer	< 5		
Tellervo	< 46698	Nooa	< 3583	Kwame	< 92		
Rauha	< 18288			Thil	152		
				Metso	1079		
				Klami	147		
				de Godzinsky	21		

Also, a stress test was made with the same investigation report data with the difference that names were placed in random places. Largest amount of data that Anoppi was able to process was 42 kb which was about 15 pages of investigation reports converted to text.

5. **Discussion**

5.1. *Anonymized data in statistical production*

According to our simulations the anonymized data works well in the statistical production. Majority of the keywords for manual data processing are accessible and the anonymized parts of the text are easy enough to process and connect.

The main challenges concerned the geopositioning stage when the keyword for exact geopositioning was missing due the anonymization. These challenges could be further tackled with more advanced anonymization tool.

When the anonymization will be adopted it will also affect the other parts of the production process of road traffic accident statistics. In the production process external data sets and databases are used to further improve the quality of the statistics and queries are made based on the personal information gained from the source material.

If all the material would be anonymized these kind of queries would not be possible anymore. We have estimated that it would affect the number of fatalities reported and the impact would roughly be two fatalities per year.

It could also be arqued whether all the manual data processing is necessary and should the text data be part of a source material for the statistics. Are the advantages for more accurate and detailed data greater than the disadvantages that come with the anonymization requirements? Future development with additional datasets might diminish the need for text data but at the moment it is a necessary part of the current statistics production process.

5.2. *Anoppi and NameFinder performance testing*

Main differences between Anoppi and NameFinder were their ability to detect true and false positives. NameFinder performance was better at true positives with the added cost of labelling more words as names than Anoppi. NameFinder also reported no false negatives, i.e. did not misidentify any names as common words, wheras Anoppi mistook some names as words. Anoppi's output can be manually viewed and edited in the web service, which means that optimum balance between correctly identified names and misidentified names and words should be found to minimize manual work. Anoppi outperformend NameFinder in this respect.

The tools use completely different set of methods in identifying names, and the outputs also differ. NameFinder anonymizes both first and last names whereas Anoppi anonymizes the whole name under one word. Anoppi also produces the aliases for the individuals which increases the readability of the documents, whereas NameFinder does rely on character based aliases. One of the reasons for this is that Anoppi has been developed for court document release for public so readability is very important. It is also a nice addition in statistical production when documents have to be manually read but missing flections do not hamper the production significantly.

All the statistics production simulations were performed with the NameFinder-tool and it would be interesting to see how the use of more advanced anonymization tool like Anoppi would perform with the actual data used in statistics production.

5.3. *Conclusions*

Our main goal with this work was to test what effect anonymization has in the statistics production process which relies partly on the usage of text documents and human interpretation of the text. We found that the effect was relatively small and that the text was still human readable and usable in the various stages of the process regardless of it being anonymized. We also found that developing a tool to turn text into anonymized form was feasible with our limited resources, and even though the tool was not perfect and anonymized more text than was necessary, it was good enough to show us that limiting the usage of personal information does not have to mean that we have to compromise too much with the quality of the end product.

We also learned that there are external solutions that are developed for another use case in mind, but could do the actual anonymization work better. These tools could be implement to the current workflow but to test these more comprehensively with real data they should be running on our own servers due to the sensitive nature of the statistical source data. Anonymization tool for text documents should also work together with the anonymization tool for the tabular data since the tabular data and text documents need to work together.

Literature

Digital and Population Data Agency’s Name Serv ice. Available at https://dvv.fi/en/name-service (Accessed: 30.5.2022).

Arttu Oksanen, Minna Tamper, Jouni Tuominen, Aki Hietanen and Eero Hyvönen (2022): A Tool for Pseudonymization of Textual Documents for Digital Humanities Research and Publication. 6th Digital Humanities in Nordic and Baltic Countries Conference (DHNB) 15-18 March Uppsala Sweden, poster paper, book of abtracts, pp. 107-108.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504 (Accessed: 28 June 2022).