TITLE PAGE

NETANOS - Named entity-based Text Anonymization for Open Science

AUTHORS:

Bennett Kleinberg¹, Maximilian Mozes^{1,2}, Yaloe van der Toolen¹, Bruno Verschuere¹

AFFILIATIONS:

¹ University of Amsterdam, Department of Psychology, Amsterdam, The Netherlands

² Technical University of Munich, Department of Informatics, Munich, Germany

KEYWORDS: open science; data transparency; text anonymization; named entity recognition

WORD COUNT ABSTRACT: 245

WORD COUNT MANUSCRIPT: 5577

ABSTRACT

Background: The shift towards open science, implies that researchers should share their data. Often there is a dilemma between publicly sharing data and protecting their subjects' confidentiality.

Moreover, the case of unstructured text data (e.g. stories) poses an additional dilemma: anonymizing texts without deteriorating their content for secondary research. Existing text anonymization systems either deteriorate the content of the original or have not been tested empirically. We propose and empirically evaluate NETANOS: named entity-based text anonymization for open science.

NETANOS is an open-source context-preserving anonymization system that identifies and modifies named entities (e.g. persons, locations, times, dates). The aim is to assist researchers in sharing their raw text data.

Method & Results: NETANOS anonymizes critical, contextual information through a stepwise named entity recognition (NER) implementation: it identifies contextual information (e.g. "Munich") and then replaces them with a context-preserving category label (e.g. "Location_1"). We assessed how good participants were in re-identifying several travel stories (e.g. locations, names) that were presented in the original ("Max"), human anonymized ("Max" → "Person1"), NETANOS ("Max" → "Person1"), and in a context-deteriorating state ("Max" → "XXX"). Bayesian testing revealed that the NETANOS anonymization was practically equivalent to the human baseline anonymization.

Conclusions: Named entity recognition can be applied to the anonymization of critical, identifiable information in text data. The proposed stepwise anonymization procedure provides a fully automated, fast system for text anonymization. NETANOS might be an important step to address researchers' dilemmas when sharing text data within the open science movement.

Introduction

The behavioral sciences are moving towards a transparent, open and reproducible manner of conducting science. One of the key conclusions of the Reproducibility Project Psychology (Open Science Collaboration, 2015) was that we need open research practices for science to become reproducible (i.e. that others can redo the experiments and understand the procedure in detail), controllable (i.e. that others can test the claims made in peer-reviewed papers) and to make better use of the cumulative character of knowledge generation of science (e.g., conduct secondary analyses). Open science is the umbrella term for various initiatives and practices (for an overview see Spellman et al., 2017). For example, to allow other researchers to replicate an experiment or learn from the procedure in the original experiment, the material of the original experiment (e.g. a computerized reaction-time task, or a pen-and-paper personality test) should be accessible to other researchers. Likewise, to prevent so-called questionable research practices (i.e. procedural and statistical techniques that allow the researcher to yield desirable results independent of the originally collected data, John et al., 2012) and to enable re-analyses, researchers are advised to share their data with others.

Suggested best open research practices were published under the Transparency and Openness Promotion (TOP) guidelines (Nosek et al., 2015). Although more than 2,900 journals have adopted the TOP framework for their publication process, the daily research practice still seems to lag behind. For example, when the authors of more than 200 studies published in four top APA journals were contacted to share their data, more than 70% of the authors did not comply with that request (Wicherts et al., 2006). One of the reasons for not sharing data is that it can be time-consuming.

Dilemma 1: Data sharing versus subject confidentiality

Leading scientific organizations like the American Psychological Association's (APA, 2015) and the National Science Foundation (2014) strongly encourage researchers to share data on request. Moreover, the APA's code of conduct states that "[a]fter research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis" (APA, 2010). At the same time, however, there are measures in place to protect the anonymity of the subjects that participated in a

study (“[...] provided that the confidentiality of the participants can be protected”; APA, 2017).

Whereas it is relatively easy to provide raw reaction time data with autobiographical stimuli words as the only identifying information, other data have a higher density and complexity of identifiable information. At the end of the spectrum of difficult-to-share data are verbal statements. Consider the field of verbal deception detection research where a typical procedure is to interview subjects, transcribe their statement, and then analyze the verbal content (e.g. Warmelink et al., 2012). To share such data, researchers often need to go to great lengths of manual preparation of the data to comply with the confidentiality agreements between researcher and subject or decide not to share their data at all.

To solve this "data sharing versus confidentiality dilemma", a straightforward solution is to anonymize the text data. Ideally, in doing so, there is no reason not to share data because the ethical requirements of subject protection are met. Although a simple, coarse anonymization such as blacking out all identifiable information (e.g. replacing "Charlotte" with "XXX") is in principle sufficient to solve the dilemma, this will often make the data unusable for secondary analyses and irreproducible in content-based research.

Dilemma 2: Anonymization versus content deterioration

The costs involved in anonymizing hundreds of statements manually aside, it is critical that some key information in the verbal statements be preserved to be useful for other researchers. For example, an important finding from verbal credibility assessment is that truthful statements tend to contain more contextual embeddings (e.g. specific places, events, persons) than deceptive statements (Köhnken, 2004). An anonymization procedure that makes it unrecoverable whether a word (or sequence of words) referred to a person, a location, or a date, makes it difficult to perform secondary analyses on the anonymized document.

As a solution to the dilemma between anonymizing data and deteriorating valuable content information, an anonymization procedure should find a compromise between meeting ethical standards and preserving sufficient content. For example, whereas "Charlotte" = "XXX" deteriorates the content of the original information, a context-preserving method such as "Charlotte" = "Person 1" would provide useable, anonymized text. In the latter, follow-up analysis could still assess how many different persons were mentioned in a text. We propose a text anonymization procedure (i) that retains essential text properties, ii) anonymizes a document sufficiently to comply with subjects' privacy

privileges, iii) and is easy to use for researchers. The global aim is to provide a tool that minimizes the costs for researchers to share their text data.

Computer-automated text anonymization

To date, research text data anonymization is predominantly manual work (e.g. patient reports in clinical psychology, interview transcripts from deception research). The time and costs of manual anonymization might be the principal impediment for researchers to share their data. There exists a variety of approaches to computer-automated text anonymization in different scientific areas (see Table 1).¹ The system proposed by Vico and Calejari (2015) is the most relevant to the current investigation: it is based on the identification of named entities within a document and the subsequent anonymization while preserving critical contextual information. Once identified the named entities are replaced with a generic term. However, Vico and Calejari's system is not available openly and relies partly on proprietary third-party software. Most importantly, although they provided a case study on legal documents, they did not experimentally validate the anonymization of anonymized documents. Contrary to the validation of the accuracy of the proposed system itself (e.g. how many of the critical information is identified), we argue that it is essential to examine how valid any anonymization system is (e.g. how well a text is anonymized for readers).

Despite the promising features of the text anonymization tools (Table 1), none of them meets all of the criteria essential for the current investigation of i) being validated, ii) being available as open-source software, as well as iii) preserving context needed for secondary research aims.

Validating information anonymization

Several variations of data anonymization have already been proposed (Table 1), but none of these tools have been validated. To validate an anonymization system, one can subject anonymized texts to a test of the re-identification of previously anonymized information (i.e. can readers identify the original content). A level of information anonymization that is appropriate for many research settings is that someone not knowing the content of a story should not be able to identify the original content. This working definition deviates from a conservative definition that only the subjects (here: the authors) themselves should be able to recognize themselves (Rock, 1999). Corti et al. (2000) argue that the sufficient level of anonymization depends on the nature of the data, and we would add that it

¹ We limit this discussion to anonymization tools for unstructured text data (e.g. free texts) rather than the encryption of tabular data (see Prasser et al., 2014).

also depends on the purpose of the data (for an elaborate discussion, see Thomson et al., 2005). In the current investigation, we focus specifically on the case of facilitating sharing of unstructured text data in agreement with ethical research guidelines.

Table 1. *Summary of existing text anonymization tools*

Name of tool/reference	Original purpose	Method	Example ²	Open-source?	Content-preserving?	Experimental validation?
“Scrub” Sweeney (1999)	Anonymization of medical documents	- identification of named entities (e.g. locations, names, countries, medical terms) - replacing of sensitive information with a pseudo-value	“Peter lives in Berlin.” -> “[pseudo_value] lives in [pseudo_value].”	No	Yes	No
Neamatullah et al. (2008)	Anonymization of free-text medical records for research purposes	- identification of protected health information with lexicon-based and context-related checks - replacing of sensitive information with a non-indexed category value	“Peter lives in Berlin.” -> “[**Name**] lives in [**Location**].”	No	Yes	No
Mottwani and Nabar (2008)	Anonymization of unstructured data in which individuals are explicitly characterized by their distinct properties	- identification of distinctive properties of identifying information - removal of sensitive information	"Doraville animal shelter" -> “— animal shelter”	No	No	No
The United Kingdom Data	General anonymization of qualitative data	- identification and highlighting numbers and words starting with a capital letter in texts	“Peter lives in Berlin.” -> “XXX lives in XXX.”	Yes	No	No

² In cases where the authors did not provide a specific example for a pseudo-value used in their anonymization tool, we stayed as close as possible to the description in the original paper.

Archive (UK DA)		- replacing of identified information with “XXX” (using an MS Word macro)				
Vico and Calegari (2015)	General domain-independent context-preserving anonymization of unstructured texts	- identification of named entities with a pipeline of multiple natural language processing tools (FreeLing, Apache OpenNLP, OpenCalais, and LingPipe) - replacing of sensitive information with a generic value	“Peter lives in Berlin.” -> “[generic_term] lives in [generic_term].” ³	No	Yes	No

Named entity recognition

The concept of named entity recognition (NER) is widely used in computational linguistics. In the early stages, the detection of named entities was predominantly based on hand-crafted, rule-based techniques and included lexicon-based approaches to identify named entities. More recent approaches utilize algorithms based on supervised machine learning models such as Hidden Markov Models, Decision Trees or Support Vector Machines (Nadeau & Sekine, 2007).

The main aim of NER is the identification (i.e. "Is it a named entity?") and classification (i.e. "What kind of entity is it?") of named entities such as names, persons, locations, and numbers within a document. For example, the sentence "Steve and Bill met in New York City 25 years ago" contains the following named entities: Steve (name), Bill (name), New York City (location), and 25 years (date). With identifying ("Steve"), classifying ("Steve" = person entity) and replacing that information (e.g. "Steve and ..." = "[Person_1] and ...") in a text, we hypothesize to be able to anonymize the critical content of the text. At the same time, an NER-based anonymization procedure offers a solution to the content deterioration dilemma: by replacing the actual entity, for example, "New York City", with its corresponding entity type [LOCATION_1], critical information is preserved for linguistic analyses.

³ The authors do not provide a specific description of the generic replacement terms used in their anonymization tool. The described transformation is therefore based on the theoretical description of the software functionality.

Aim of this investigation

The current investigation has as the primary objective of introducing and empirically validating a computer-automated anonymization procedure of unstructured text data. We test how well different anonymization procedures make the original core elements of travel stories non-identifiable. We propose **NETANOS** (Named Entity-based Text ANonymization for Open Science), which we evaluate against a standard, human anonymization baseline, as well as the UK Data-Archive anonymization tool.⁴ Our core hypothesis is that the NETANOS procedure results in a similar identification accuracy as the human anonymization baseline. Our control hypothesis is that the highest identification accuracy is achieved for the original (i.e. non-anonymized texts). Exploratorily we test how the different text states affected the plausibility and readability. We further investigate which strategies participants apply to re-identify information from processed text.

Method

Materials

This experiment was pre-registered before data collection on [aspredicted.org](https://aspredicted.org/#3629) (#3629 <https://aspredicted.org/ja2qw.pdf>). All data collected for this study as well as the analysis scripts are publicly available on the Open Science Framework.⁵ The original experimental task is accessible via <http://newlylabs.net/wp-content/research/anonym/f/html/main.html>. The source code for NETANOS is available on GitHub at <https://github.com/ben-aaron188/anonymiseme>. The method, results, and discussion of a pre-registered pilot experiment for the development of NETANOS can be found on the OSF repository corresponding to this paper.

Statements

Since stories about traveling naturally contain contextual information (people travel to certain *destinations*, for a certain amount of *time* with other *persons*), we decided to collect travel stories as documents illustrative for texts containing references to contextual and recognizable information. Eight different, unique combinations of contextual information units (information about persons,

⁴ We choose for the UK Data Archive method as a comparison since it is the only openly available system for automated text anonymization and highlights the difference between context-preserving and context-deteriorating anonymization.

⁵ Link to the data: https://osf.io/e365p/?view_only=4f33a25d54504b3884eb67bf9f851aea

locations and time) formed the building blocks for different travel stories. Each combination consisted of the following ten pieces of contextual information: two main persons, four cities, three dates and one additional person (e.g. Daniel, Harry; London, Manchester, Liverpool, Leeds; December 2016, 1 week, 2 days; Peter). For all combinations of contextual information, it was ensured that this information would lead to a plausible, realistic travel story.⁶ The resulting combinations of contextual information were sent to volunteers who were asked to write short travel stories in approximately 100 words based on the given pieces of information. We selected the four stories that contained all required information and were well-written for use in this study (Online Appendix A).

Manual anonymization

Two independent coders read the four travel stories. They were instructed to identify all entities belonging to one of the following categories: person references (e.g. “Ben”, or “my sister”), specific location references (e.g. “London”, or “15th Park Avenue”, but not generic examples such as “at home” or “in a restaurant”), date or time references (e.g. “the third of May”, or “early in the morning”), and number references (e.g. “one dollar”, or “3”). After identifying the words that belonged to one of these categories, the coders were instructed to replace all selected entities with an anonymous reference. For instance, a person entity such as “Ben” was de-identified into “PERSON_1”, “Lucy” got replaced with “PERSON_2” and a time reference such as “early in the morning” was changed into “DATETIME_1”. After the individual coding process, the two coders compared their anonymized statements and resolved differences (initial agreement: 94%).

UK Data Archive anonymization

This method is inspired by the UK Data Archive (UK DA) text anonymization approach and therefore identifies all words starting with a capital letter and all numeric values and replaces each with “XXX”. Using this approach, “Steve and Bill met in New York City 25 years ago” would be transformed to “XXX and XXX met in XXX XXX XXX XXX years ago”. The code used for this approach is available in the NETANOS tool as well.

NETANOS

⁶ We specifically picked well-known names for the three characters in the combinations (e.g. by checking which baby names in the United States were most popular in the past five years). Moreover, we carefully matched the cities and dates in every combination so that they could easily resemble real travel stories.

To de-identify continuous text, we wrote a software tool capable of identifying and modifying named entities such as names, locations, dates, and organizations within a given unstructured text input. The software was written in JavaScript and its back-end library Node.js, respectively. Specifically, NETANOS' primary anonymization method replaces the identified named entities with an indexed generic replacement of the corresponding entity type. For instance, "Steve and Bill met in New York City 25 years ago" would be transformed to "[PERSON_1] and [PERSON_2] met in [LOCATION_1] [DATE/TIME_1]" The "_1" is a unique number denoting the X-th occurrence of the corresponding entity type while keeping the index consistent for recurring entities.

Named entity recognition is the backbone of NETANOS and consists of a stepwise system of two algorithm interfaces. First, we used the Stanford Named Entity Recognizer (Finkel et al., 2005), a named entity recognition software tool written in Java (implemented in Node.js). Stanford's NER follows a probabilistic approach for identifying named entities by using Conditional Random Fields (CRF). Second, we applied the named entity recognition tool of NLP Compromise (Kelly, 2016), a Natural Language Processing library written in JavaScript. Contrary to the Stanford NER tool, NLP Compromise employs a lexicon-based approach, that is, the algorithm compares tokenized terms of the given text with an underlying lexicon to detect and classify named entities.

To successfully anonymize texts, we adopt both of them successively. The text is firstly analyzed by Stanford's Named Entity Recognizer, extracting organizations, persons, and dates. After that, we replaced all identified entities within our text. To further enhance the quality of our data anonymization, we subsequently executed the NLP-compromise tool attempting to identify remaining unrecognized entities.

The following anonymization features are built into NETANOS:

- name and person anonymization (e.g. Steve -> [PERSON_1], Bill -> [PERSON_2])
- location anonymization (e.g. New York City -> [LOCATION_1], Madrid -> [LOCATION_2])
- date and time anonymization (e.g. 25 years -> [DATE/TIME_1], 2 days -> [DATE/TIME_2])
- gendered pronoun anonymization (e.g. His -> [HIS/HER_1], He -> [HE/SHE])
- anonymization of numeric values (e.g. 42 -> [NUMERIC_1], 1337 -> [NUMERIC_2])
- anonymization of written numeric values (e.g. one -> [NUMBER_1], two -> [NUMBER_2])
- anonymization of remaining, potentially non-recognized entities: this feature aims to

anonymize potential entities that have not been identified by the above features; therefore, the feature scans through the text after all the other features have been applied and replaces all words starting with a capital letter with [OTHER_X]

Procedure

Participants were recruited via the crowdsourcing platform Prolific Academic. After giving informed consent, participants were told that they were about to read several travel stories and that some of them would be presented in an anonymized form. All participants were randomly allocated to one of four between-subjects conditions (original, NETANOS, human, UK DA). The order of the travel stories was randomized, and the anonymization state was counterbalanced across participants. Travel stories were presented one by one. Besides, each participant read one control story provided in the original version to ensure participants' attentiveness (and those failing to match it were excluded, see Results).

Each statement was accompanied by four possible scenarios: one correct scenario (i.e., the ten information units on which the original travel story had been based) and three distractors (i.e. ten similar information units that had no relation to any of the travel stories used in this experiment, Online Appendix B)⁷. Participants were instructed to read the presented statement carefully and could only proceed after 20 seconds. Their task was to select the scenario which had been the basis for the travel story. They were asked to indicate the certainty of their choice, and the readability and plausibility of the text on a scale from 0 (not certain/readable/plausible) to 100 (very certain/readable/plausible).⁸

After reading and identifying the four statements, participants were asked to mention at least three strategies that they had used to identify the correct scenario. After this, participants provided demographic information (age, gender, education, country of origin, native language) and were debriefed. The reward to GBP 0.50 as a base payment plus a performance-based bonus payment. Each participant was rewarded with GBP 0.20 for each correct answer, resulting in payments ranging from GBP 0.50 to 1.50.

⁷ The distractor combinations were matched to the correct option on contextual information (i.e. weather, distance between locations, gendered names).

⁸ Readability was defined as “The travel story is easy to read”, plausibility as “The travel story could have happened as described”.

Analysis plan

As per the pre-registration, we examined the hypotheses with both null hypothesis significance testing (control hypothesis) as well as Bayesian testing (core hypothesis). In contrast to null hypothesis significance testing, Bayesian testing allows to quantify evidence for the null hypothesis (here: no accuracy difference between the NETANOS, human, and UK DA anonymization). Specifically, we use Kruschke's (2013) *Bayesian Estimation Supersedes the t-test* (BEST) approach whereby we specify a Region of Practical Equivalence (ROPE), which represents the interval of the difference that is deemed practically irrelevant. That is, if the difference between two accuracies A and B falls into the ROPE, we conclude that there is no practically relevant difference between A and B. In BEST, the difference is resampled in a Markov chain Monte Carlo (MCMC) procedure to derive the 95% Highest Density Interval (HDI). If the 95% HDI falls into the ROPE, this is evidence for the null hypothesis that there is no difference between A and B greater than the ROPE.

Since the ROPE depends on the context of the investigation, we defined two ROPEs: the conservative ROPE has a lower limit of -0.05 and an upper boundary of 0.05, whereas the liberal ROPE is defined as [-0.10; +0.10]. These values imply that, for the conservative ROPE, any accuracy difference that lies within +/-5 percent is deemed irrelevant. Hence the two anonymization versions are considered to be equivalent. Substantial support for the equivalence of two anonymization types would stem from a 95% HDI in the conservative ROPE, whereas strong support would come from the HDI falling into the liberal ROPE.

Results

Participants

We collected data of 555 due to simultaneous starting times.⁹ Participants were excluded if (1) their IP was recorded more than once ($n = 126$; a conservative criterion that avoids multiple participation see also Kleinberg & Verschuere, 2015), (2) their data were not recorded ($n = 0$), and (3) they did not correctly identify the non-anonymized control story ($n = 31$). The final sample consisted of 398 participants ($M_{age} = 31.37$ years, $SD = 9.75$, 47.49% female). The random condition allocation resulted in 110 participants in the original version condition ($M_{age} = 31.65$, $SD = 9.69$, 47.27% female), 104 in the NETANOS condition ($M_{age} = 32.33$, $SD = 9.79$, 50.96% female), 99 in the human-

⁹ We aimed to collect data from 492 participants based on the preregistered a priori power analysis for a power of 0.80, Cohen's f of 0.15, and a significance level of 0.05.

anonymized condition ($M_{age} = 31.31$, $SD = 9.60$, 51.52% female), and 85 in the UK DA condition ($M_{age} = 29.93$, $SD = 9.93$, 38.82% female). There were no differences in gender, $\chi^2(3) = 3.71$, $p = .295$, or age, $F(3, 394) = 0.98$, $p = .400$, $f = 0.05$.

Null hypothesis significance testing

We predicted that the percentage of correct identification is highest in the original version. There was a significant main effect of the Travel Story Version on the accuracy of choosing the correct option, $F(3, 394) = 152.80$, $p < .001$, $f = 0.47$. The accuracy in the original version condition was higher than that of the NETANOS condition, $t(212) = 18.05$, $p < .001$, $d_{between} = 2.47$ [2.11; 2.82]; higher than that of the human-anonymized condition, $t(207) = 18.20$, $p < .001$, $d_{between} = 2.52$ [2.16; 2.88]; and higher than that of the UK DA condition, $t(193) = 16.58$, $p < .001$, $d_{between} = 2.39$ [2.02; 2.76]. Figure 1 shows the average accuracies in percentages.

Figure 1. Percentage of correct identifications per travel story version in Exp.2.

Bayesian analyses

Our core hypotheses predicted that there were no differences in the accuracies of correct identifications between the human-anonymized version and the NETANOS version and UK DA version. We used the BEST approach (Kruschke, 2013) with both the conservative [-0.05; +0.05] and the liberal ROPE [-0.10; +0.10].

NETANOS versus human baseline

The 95% HDI of the difference between the NETANOS version ($M = 38.22$; $SD = 24.37$; HDI: [33.39; 42.97]) and the human-anonymized version ($M = 38.64$; $SD = 23.22$; HDI: [33.83; 43.19]), was [-6.81; +6.67]. The HDI of the difference fell completely into the liberal ROPE, whereas only 86% fell into the conservative ROPE. These findings support the notion that the accuracy of the NETANOS anonymization is practically equivalent to the human-anonymized version under the liberal ROPE.

UK Data Archive versus human baseline

The 95% HDI of the difference between the UK DA version ($M = 42.94$; $SD = 22.03$; HDI: [37.64; 47.26]) and the human-anonymized version was [-10.81; +2.58]. 96% of the HDI fell into the liberal ROPE with the actual interval of the HDI providing strong support for the notion that the human-anonymized version is better than the non-context preserving one. The upper boundary of the difference is +2.58%, suggesting that any inferiority of the human-anonymized version is well within the conservative ROPE.

NETANOS versus UK Data Archive

The 95% HDI of the difference between the NER version and the non-context preserving version was [-11.17; +2.35]. Although only 95% of the HDI fell into the liberal ROPE, the actual interval of the HDI provides strong support for the notion that the NER version is superior to the UK DA version (the upper boundary below that of the conservative ROPE at +2.21%). 57% of the HDI of the difference fell into the conservative ROPE.

Exploratory analysis

Perceived text quality: Readability

A one-way ANOVA with Travel Story Version (original, NETANOS, human-anonymized, UK DA)

on the travel stories' readability revealed a significant main effect of Travel Story Version, $F(3, 394) = 79.21, p < .001, f = 0.38$. The readability was higher in the original version ($M = 85.10; SD = 15.38$) than in the NETANOS version ($M = 43.32; SD = 26.58$), $t(212) = 14.17, p < .001, d_{between} = 1.94$ [1.61; 2.26], the human-anonymized version ($M = 48.42; SD = 26.46$), $t(207) = 12.40, p < .001, d_{between} = 1.72$ [1.40; 2.03], and the UK DA version ($M = 42.54; SD = 24.39$), $t(193) = 14.88, p < .001, d_{between} = 2.15$ [1.79; 2.50].

There were no differences in readability between the NETANOS version and the UK DA version, $t(187) = 0.21, p = .837, d_{between} = 0.03$ [-0.26; 0.32], nor between the human-anonymized version and the UK DA version, $t(182) = 1.56, p = .121, d_{between} = -0.23$ [-0.52; 0.06]. Neither was there difference between the NETANOS and human-anonymized version, $t(201) = 1.37, p = .172, d_{between} = -0.19$ [-0.47; 0.08].

Perceived text quality: Plausibility

A one-way ANOVA on the travel story plausibility revealed a significant main effect of Travel Story Version, $F(3, 394) = 28.06, p < .001, f = 0.25$. The plausibility was higher in the original version ($M = 75.14; SD = 19.24$) than in the NETANOS version ($M = 54.07; SD = 21.30$), $t(212) = 7.60, p < .001, d_{between} = 1.04$ [0.75; 1.32], the human-anonymized version ($M = 53.33; SD = 23.14$), $t(207) = 7.43, p < .001, d_{between} = 1.03$ [0.74; 1.32], and the UK DA version ($M = 52.21; SD = 21.65$), $t(193) = 7.81, p < .001, d_{between} = 1.13$ [0.82; 1.43].

There were no differences in plausibility between the NETANOS version and the UK DA version, $t(187) = 0.60, p = .552, d_{between} = 0.09$ [-0.20; 0.37], nor between the human-anonymized version and the UK DA version, $t(182) = 0.34, p = .735, d_{between} = -0.05$ [-0.34; 0.24]. Neither was there difference between the NETANOS and human-anonymized version, $t(201) = 0.24, p = .813, d_{between} = 0.03$ [-0.24; 0.31].

Self-reported de-anonymization strategies

We were also interested in the strategies used by participants to re-identify the content of the travel stories. We randomly sampled 35 participants from each condition (140 in total). Table 2 displays the results for a qualitative analysis of the reported strategies (see Online Appendix C) The strategy most often used was "comparing unique features of locations to the answer options". In the date category, participants indicated that they most often looked at the plausibility related to the time frame for the

answer options.

Table 2. *Percentage of strategies used to identify the correct answer option (per condition)*

	Date/ Time	Locations	Persons	Weather	Other	Simple matching	No (clear) strategy
Original	0	5.71	2.86	0	5.71	100.00	0
Human	34.29	57.14	20.00	17.14	17.14	0	5.71
NETANOS	42.86	42.86	22.86	11.43	17.14	0	8.57
UK DA	25.71	45.71	25.71	8.57	22.86	0	14.29

Note. Since participants could have used more than one strategy, the percentages do not necessarily add up to 100%.

General Discussion

This paper set out to introduce and empirically test an anonymization system for unstructured text data. Motivated by the need for more data transparency, the primary aim was to compare a novel anonymization algorithm to a human baseline. Since a key requirement for text anonymization useful for secondary research purposes is the preservation of context, we used a stepwise named entity recognition (NER) system as the engine for the proposed anonymization system (called NETANOS). We also assessed the difference between the context-preserving NETANOS system and an existing, recommended anonymization tool that deteriorates context (UK Data Archive).

By accounting for contextual information in the anonymized texts, the NETANOS anonymization algorithm was found to be practically equivalent to the human baseline.

The findings from this pre-registered study support the proposed NETANOS anonymization system. Most importantly, not only was it practically equivalent to the human baseline but it also outperformed a non-context preserving anonymization system. Whereas the non-context preserving computer-automated system (i.e. replacing "Ben" with "XXX") outperforms manual efforts in the time and resources need for the "anonymization", too much contextual information is lost for any secondary linguistic analyses. NETANOS was intended to capture the key advantages of both worlds: the identification of fine-grained contextual information feasible in manual, human anonymization efforts, on the one hand, and the fast and automated procedure needed for the anonymization of large

numbers documents feasible with computer-automated tools. Our findings lend support to the novel named entity-based anonymization system. This paper offered the first empirical evaluation of a text anonymization system that outputs data usable for secondary analysis and provides an open-source tool useable by any researcher to overcome privacy-related restrictions in data sharing efforts.

Limitations and outlook

Despite the promising findings reported here, some limitations merit attention in future work. (1)

Although the NER-based anonymization is practically equivalent to manual anonymization, the average accuracy for the NETANOS system (38.22%) and the human baseline (38.64%) are still higher than chance level. It could be that the method of anonymization is too shallow, although a similar procedure is proposed for throughout the literature (see Table 1) and is used in clinical research. While one can argue that the accuracies are too high, there is little evidence to the accuracy of re-identification with existing manual anonymization procedures. Further research could compare existing anonymized texts with the NER-based system with regard to their re-identifiability.

(2) Since the proposed context-preserving anonymization system relies on the recognition of named entities, this system is inherently limited by the accuracy of NER engine. Software tools that extract entities do not guarantee the identification of every named entity that occurs within this text. Although modern named entity recognition tools exhibit promising performances (Jiang et al., 2016), they are not yet capable of identifying 100% of the named entities within a given text. Jiang et al. (2016) presented a comparison of four widely-used natural language processing software tools including the Stanford Named Entity Recognizer, a named entity recognition tool that we integrated into our anonymization software. Their research shows that of the four considered NLP tools (i.e. Stanford NER, LingPipe, spaCy and NLTK), Stanford NER has the best overall performance in recognizing named entities. The tool performs with a weighted accuracy of 70.75% (Online Appendix E). However, this evaluation shows that there is room for improvement in named entity recognition. To account for this accuracy limitation, we used a stepwise approach with another named entity recognition tool (NLP Compromise, Kelly, 2016) which follows a lexicon-based identification approach, meaning that it compares tokenized phrases with built-in libraries to detect named entities rather than considering only probabilistic or syntactic approaches. Moreover, by adding a generic non-named entity detector (i.e. all words starting with a capital letter), we will have captured, at least for the English language, a large proportion of potentially undetected named entities. The current

findings suggest that despite the accuracy constraints, the proposed approach and software tool is suitable to mask the real information contained in text data. We anticipate that future accuracy improvements in NER engines go hand in hand with improvements in the NETANOS system. As such, we argue that the accuracy limitations imply that our system is more likely the lower boundary of anonymization performance rather than an overestimation. With better NER engines, we expect future work to show even better anonymization results.

(3) Furthermore, the results suggested that while the anonymization of NETANOS is practically equal to the human baseline as standard, we also see a decline in perceived text quality (plausibility and readability). It might be that for both text properties, it does not matter whether “Steve and Bill” is replaced with “XXX and XXX” or “[Person_1] and [Person_2]”. In both cases, the readability and plausibility are rated as equally low. Future developments of NER-based anonymization could move towards semantic replacements that ideally make the text quality indistinguishable from the original (e.g. “Steve and Bill” = “Paul and Peter”). This requires an additional level the preservation, namely that of semantic relationships. To provide meaningful semantically correct replacements, the system could be trained to learn that “Moscow is the capital of Russia” can only be replaced with a capital/country pair (e.g. “Berlin ... Germany”).

(4) Related to semantic relationships, NETANOS does not include the anonymization of qualitative information. For example, for the sentence "Ben likes working with Max", the preservation of "likes working with" is highly desirable for research purposes, but might be a source for potential re-identification of persons. Future work could also look at how much of this qualitative information is needed for meaningful secondary linguistic analysis.

(5) The test of the proposed anonymization system with travel stories was conducive to the purpose of testing the potential re-identification of information in a reasonably standardized manner. However, there are several other areas in which document anonymization might be useful such as the double-blind peer-review process or fair procedure in job applications (e.g. anonymizing the applicant's origin). Future work could use the approach outlined as well as the software provided by the paper in different contexts.

Conclusion

This paper presented evidence for a freely available, fully automated text anonymization system that preserves critical contextual information needed for secondary linguistic analyses. By using named

entity recognition as the backbone of the proposed system, we were able to develop an anonymization algorithm that is practically equivalent to manual anonymization while being fast and scalable to hundreds of documents. We provide the anonymization tool, NETANOS, as free, open-source software and encourage others to use and further improve our proposed system and challenge the findings made in the current investigation.

References

- American Psychological Association (2010). *Ethical Principles of Psychologists and Code of Conduct* (Including 2010 and 2016 Amendments). Retrieved from <http://www.apa.org/ethics/code/>
- American Psychological Association (June 2015). *Data Sharing: Principles and Considerations for Policy Development*. Retrieved from <https://www.apa.org/science/leadership/bsa/data-sharing-report.pdf>
- Corti, L., Day, A., & Backhouse, G. (2000). Confidentiality and informed consent: Issues for consideration in the preservation of and provision of access to qualitative data archives. *Forum: Qualitative Social Research*, 1(3). Retrieved from <http://www.qualitative-research.net/index.php/fqs/article/viewArticle/1024/2207.gcit>
- Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. *Association for Computational Linguistics: Proceedings of the 43rd annual meeting on association for computational linguistics*, 363-370. Retrieved from http://www.aclweb.org/old_anthology/P/P05/P05-1.pdf#page=391
- Jiang, R., Banchs, R. E., & Haizhou, L. (2016). Evaluating and combining biomedical named entity recognition systems. *Proceedings of the Sixth Named Entities Workshop Joint with 54th ACL*, 21–27. doi: 10.3115/1572392.1572430
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5), 524–532. doi: 10.1177/0956797611430953
- Kelly, S. NLP Compromise (2016), GitHub repository, <https://github.com/nlp-compromise/compromise>
- Kleinberg, B., & Verschuere, B. (2015). Memory detection 2.0: The first web-based memory detection test. *PloS one*, 10(4), e0118715.
- Köhnken, G. (2004). Statement validity analysis and the detection of the truth. In Granhag & Strömwall (Eds.). *The Detection of Deception in Forensic Contexts* (pp. 41-63). Cambridge: Cambridge University Press
- Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2), 573–603. doi: 10.1037/a0029146
- Motwani, R., & Nabar, S. U. (2008). Anonymizing unstructured data. *arXiv:0810.5582*. Retrieved

- from <https://arxiv.org/pdf/0810.5582.pdf>
- Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1), 3–26. doi: 10.1075/li.30.1.03nad
- National Science Foundation (February 2014). Grant General Conditions. Retrieved from <https://www.nsf.gov/pubs/policydocs/gc1/feb14.pdf>
- Neamatullah, I., Douglass, M. M., Lehman, L. H., Reisner, A., Villarroel, M., Long, W. J., ... Clifford, G. D. (2008). Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1), 32. doi: 10.1186/1472-6947-8-32
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. doi: 10.1126/science.aab2374
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), 943-952. doi: 10.1126/science.aac4716
- Prasser, F., Kohlmayer, F., Lautenschläger, R., & Kuhn, K. A. (2014). Arx-a comprehensive tool for anonymizing biomedical data. In *AMIA Annual Symposium Proceedings* (Vol. 2014, p. 984). American Medical Informatics Association.
- Rock, F. (2001). Policy and practice in the anonymisation of linguistic data. *International Journal of Corpus Linguistics*, 6(1), 1–26. doi: 10.1075/ijcl.6.1.01roc
- Spellman, B. A., Gilbert, E. A., & Corker, K. S. (2017, April 19). Open science: what, why, and how. Retrieved from osf.io/preprints/psyarxiv/ak6jr
- Sweeney, L. (1996). Replacing personally-identifying information in medical records, the Scrub system. *AMIA Annu Symp Proc*, 333–7. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2233179&tool=pmcentrez&rendertype=abstract>
- Thomson, D., Bzdel, L., Golden-Biddle, K., Reay, T., & Estabrooks, C. A. (2005). Central questions of anonymization: A case study of secondary use of qualitative data. *Forum: Qualitative Social Research* 6(1). Retrieved from <http://www.qualitative-research.net/index.php/fqs/article/view/511/1102>
- UK Data Service (no date) ukds.tools.textAnonHelper / Home [BitBucket Wiki]. Retrieved February 25, 2017, from <https://bitbucket.org/ukda/ukds.tools.textanonhelper/wiki/Home>
- Vico, H., & Calegari, D. (2015). Software Architecture for Document Anonymization. *Electronic*

Notes in Theoretical Computer Science, 314, 83–100. doi: 10.1016/j.entcs.2015.05.006

Warmelink, L., Vrij, A., Mann, S., Jundi, S., & Granhag, P. A. (2012). The effect of question expectedness and experience on lying about intentions. *Acta Psychologica*, 141(2), 178–183. doi: 10.1016/j.actpsy.2012.07.011

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, 61(7), 726–728. doi: 10.1037/0003-066X.61.7.726