Anonymization Service for Finnish Case Law: Opening Data without Sacrificing Data Protection and Privacy of Citizens

Minna Tamper[1,2], Arttu Oksanen[1,2,3],
Jouni Tuominen[1,2], Eero Hyvönen[1,2], and Aki Hietanen[4]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland
http://seco.cs.aalto.fi, firstname.lastname@aalto.fi
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
http://heldig.fi
[3] Edita Publishing Ltd.
http://www.editapublishing.fi
[4] Ministry of Justice, Finland
http://oikeusministerio.fi, firstname.lastname@om.fi

1 Introduction

The Finnish public sector produces vast amounts of valuable data that concern private citizens of Finland. The Finnish court case data, for example, could be useful in governmental decision making and research if opened for the public. However, due to issues of data protection and privacy it is not possible to share the data openly. This challenge is especially urgent now that the General Data Protection Regulation (GDPR) came into effect on May 25, 2018. After this, the court cases need to be sealed from the public, unless anonymized.

The GDPR (EU) 2016/679[5] is a regulation in EU law on data protection and privacy for all individuals within the European Union. Prior to the GDPR, thousands of selected court cases have been made available to the public via the Finlex service[6], and the related Semantic Finlex Linked Data service[7] managed by the Ministry of Justice. The court cases in Finland are available and can be obtained in print for a request from courts, but their publication on the web is restricted.

In order to release court case documents to the public it is required that the documents are anonymized. Anonymization is the process that removes explicitly or implicitly identifying details of persons and companies from text. Edita Publishing Ltd. estimated that it takes roughly 40 minutes to manually anonymize one decision. The proceedings of the decision making in courts have been highly interesting for the public but few decisions have been anonymized and made public. In order to publish the decisions anonymized more human resources are required, but these may not be available.

In this paper, we propose an automatic anonymization method and tool for the court decisions. Currently, according to Back & Keränen [2], the Finnish public sector utilizes poorly automatic anonymization tools because of the difficulty of evaluating the sufficiency of the anonymization for different kinds of data and needs. In addition, it is hard to find a service or tool that can handle both Finnish and Swedish language texts properly. For these reasons, we are in the process of creating a semi-automatic tool for Finnish and Swedish languages.

2 Background

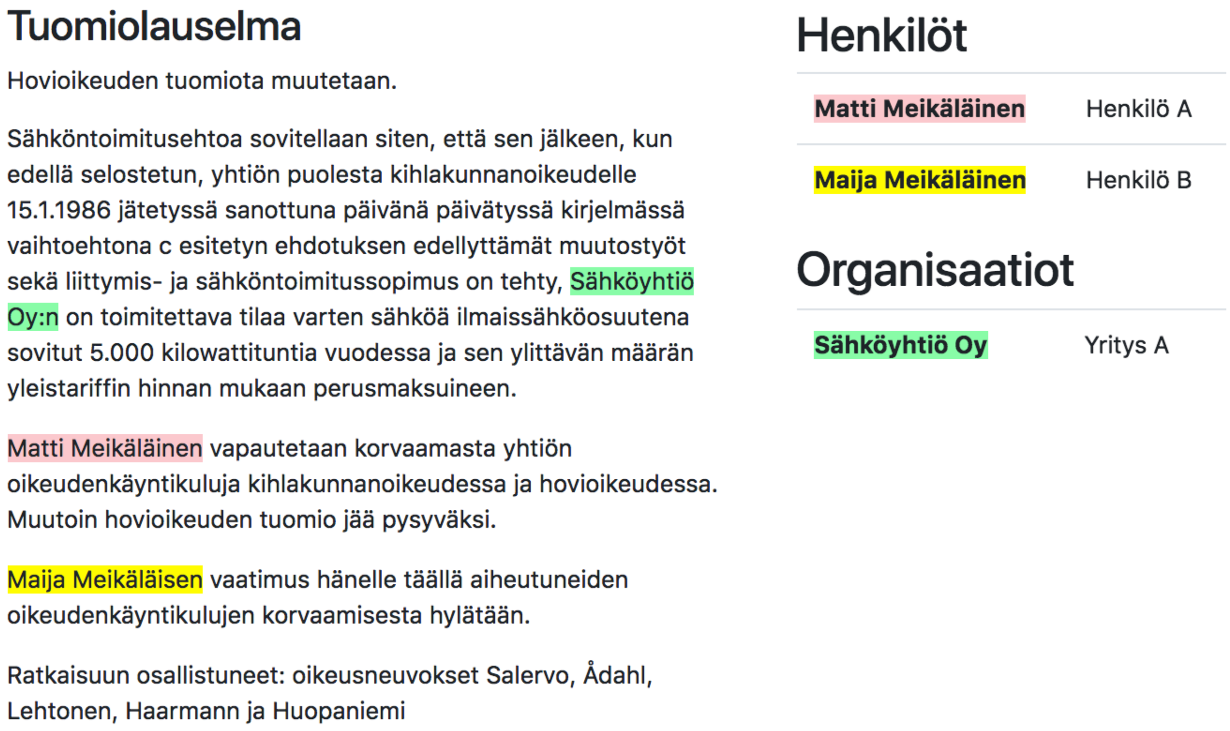
A study [2] in 2017 about the anonymization of Finnish public sector documents highlights that there is a growing need for anonymization but also a number of challenges. Due to these challenges, the current state of anonymization tools is that each actor of the Finnish public sector takes care of the anonymization

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679
[6] http://www.finlex.fi
[7] http://data.finlex.fi

of their own documents. The challenges in the anonymization of the Finnish public sector documents are due to the nature of the data, different needs of different actors, and the difficulty of evaluating the level of anonymization for different types of data for different actors and their stakeholders. Based on these different needs and challenges, the study proposed three different approaches to handling these challenges: a) a service that collects all related data and in return offers anonymized versions of the data collections, b) expert services to support and guide for anonymization done in companies and organizations, and c) a service that provides users a query facility that produces anonymized answers [2]. In this paper, we propose a type of service where users can get anonymized data in return for a request.

The service needs to be able to anonymize the court orders. According to Bäck & Keränen the general practice of anonymization in Finland is that all corporate and personal details must be removed and that entails information about the vocation, industry, residency, and job location as well [2]. This information has to be de-identified, that is the practice of replacing identifiable information with a corresponding referent identifier [7]. The procedure of substituting identifiable information with neutral names, i.e. pseudonyms (such as ”person A”), is called pseudonymization. In order to build a working de-identification system, language technology tools specialized in Finnish and Swedish languages, such as part-of-speech taggers, named entity recognizers, and coreference recognizers, are needed [4,3]. These tools help in automatic identification different entities from text, such as names, organizations, and places. In addition, there will be functions for metadata annotation in the user interface (UI). [1]

3 Application



Tuomiolauselma

Hovioikeuden tuomiota muutetaan.

Sähköntoimitusehtoa sovitellaan siten, että sen jälkeen, kun edellä selostetun, yhtiön puolesta kihlakunnanoikeudelle 15.1.1986 jätetyssä sanottuna päivänä päivätyssä kirjelmässä vaihtoehtona c esitetyn ehdotuksen edellyttämät muutostyöt sekä liittymis- ja sähköntoimitussopimus on tehty, Sähköyhtiö Oy:n on toimitettava tilaa varten sähköä ilmaissähköosuutena sovitut 5.000 kilowattituntia vuodessa ja sen ylittävän määrän yleistariffin hinnan mukaan perusmaksuineen.

Matti Meikäläinen vapautetaan korvaamasta yhtiön oikeudenkäyntikuluja kihlakunnanoikeudessa ja hovioikeudessa. Muutoin hovioikeuden tuomio jää pysyväksi.

Maija Meikäläisen vaatimus hänelle täällä aiheutuneiden oikeudenkäyntikulujen korvaamisesta hylätään.

Ratkaisuun osallistuneet: oikeusneuvokset Salervo, Ådahl, Lehtonen, Haarmann ja Huopaniemi

Henkilöt

Matti Meikäläinen	Henkilö A
Maija Meikäläinen	Henkilö B

Organisaatiot

Sähköyhtiö Oy	Yritys A

Fig. 1. Application user interface.

In order to make court cases publicly available, we have started to build an application that anonymizes these documents. The anonymization application consists of two separate software components, more specifically a web service and a user interface.

The web service comprises a functionality that takes text as input and produces as output the same text annotated with special tags that mark the occurrences of named entities in the text. To produce the annotations the web service utilizes a variety of different natural language processing tools, such as Omorfi [6] and Stanford NER [8]. These special tags also contain additional metadata about the occurrences, such as a category (person, place, organization etc.), grammatical base form and grammatical case. The category and base form are required so that the named entities can be correctly transformed into their corresponding pseudonymized forms such as ”person A” or ”company B”. The information about the grammatical case is needed so that the named entities appearing in the text can be replaced with their pseudonyms in correctly inflected forms.

The user interface is a web-based WYSIWYG editor. Fig. 1 shows its basic design. The anonymization process starts by importing an unanonymized document into the editor. The document is first pre-processed by the web service producing an initial set of named entity candidates. After that the document text with the suggested named entities highlighted is shown in the left column of the application window and a list of the suggested named entities along with their pseudonyms is shown in the right column. The user is able to edit these suggestions, add new ones, and delete them. When editing is finished, an anonymized version of the document can be exported with all of the occurrences of selected named entities pseudonymized.

In addition to names that must be pseudonymized, the documents may contain names that need not be obfuscated. For example, the last paragraph of the document in the Figure 1 contains the names of the judges of the Supreme Court that have worked on the case in question. However, those names are not highlighted by the application because they contain information that is public and therefore need not be pseudonymized. Additional logic is built into the application to automatically take into account these special cases.

4 Discussion

The anonymization of all the court decision materials is costly and a time consuming process when done manually. Based on the estimates of Ministry of Justice, on a yearly basis, the district courts make approximately half a million decisions and the courts of appeal 9 500 decisions of which most decisions do not change. A fraction of these decisions have become publicly available prior GDPR. From the prejudicate made by the supreme court approximately 30% of the decisions are also manually anonymized. The administrative courts have anonymized approximately 50 from the total of 20 000 decisions per year. Based on the estimate done by Edita Publishing Ltd. on 40 randomly selected decisions, it takes nearly 38 minutes to anonymize one document. This estimate includes familiarization which takes roughly 20 minutes per document.

In order to measure the functionality of the service, its performance will be compared with the estimates of Edita Publishing Ltd. to see if it fares better. The success of the tool depends on the applicability of the UI in addition to the precision and recall of the language technology tools. The recall of the tools is more crucial than the precision [9,7], as it is more important to hide all critical information than to have some non-critical information incorrectly hidden. The UI of the application is pivotal to the efficiency of the tool. Therefore, incorporating the users and carrying out usability tests [5] are vital to the success of the semi-automatic anonymization tool.

It is yet to be seen whether or not the service will be helpful in anonymization of the decision materials. However, the tool is expected to perform faster than a person doing manual anonymization. The number of mistakes and the time it takes to fix the mistakes will decide the usefulness of the service in the case of anonymizing Finnish court decision materials.

Acknowledgements The work is a joint effort between Aalto University, HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, the Ministry of Justice, the Ministry of Finance, Edita Publishing Ltd., and Business Finland (formerly Tekes). We thank Risto Talo, Jari Linhala and Sari Korhonen of Edita Publishing Ltd. for collaboration.

References

1. Ahrenberg, L., Blomstrand, N., Fogde, M., Nilsson, A.: Anonymization of personal stories. In: The second national Swe-Clarin workshop: Research collaborations for the digital age. Ümeå, Sweden (November 2016)
2. Bäck, A., Keränen, J.: Anonymisointipalvelut. tarve ja toteutusvaihtoehdot (2017), `http://urn.fi/URN:ISBN:978-952-243-503-3`, Liikenne- ja viestintäministeriön julkaisuja 7/2017
3. Jurafsky, D., Martin, J.H.: Speech and language processing, vol. 3. Pearson London: (2014)
4. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1), 3–26 (2007)
5. Nielsen, J.: Usability inspection methods. In: Conference companion on Human factors in computing systems. pp. 413–414. ACM (1994)
6. Pirinen, T.A.: Omorfi—free and open source morphological lexical database for finnish. In: Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania. pp. 313–315. No. 109, Linköping University Electronic Press, Linköpings universitet (2015)
7. Povlsen, C., Jongejan, B., Hansen, D.H., Simonsen, B.K.: Anonymization of court orders. In: 11th Iberian Conference on Information Systems and Technologies (CISTI). IEEE, Las Palmas, Spain (June 2016)
8. Rose Finkel, J., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. pp. 363–370 (January 2005)
9. Szarvas, G., Farkas, R., Busa-Fekete, R.: State-of-the-art anonymization of medical records using an iterative machine learning framework. Journal of the American Medical Informatics Association 14(5), 574–580 (2007)