Untapped data resources

Applying NER for historical archival records of state authorities

Venla Poso[1], Tanja Välisalo[1], Ida Toivanen[1], Antero Holmila[1] and Jari Ojala[1]

[1]*University of Jyväskylä, Seminaarinkatu 15, Jyväskylä, Finland*

Abstract

Archives around the world are digitising their material at a growing speed. The National Archives of Finland launched a mass digitisation process in 2019 aiming to digitise vast amounts of state authority archives. In order to improve the access and use of this data by researchers, we present the data transfer process of state authority data and the development of named entity recognition (NER) for enriching and using archival data from state authorities. In this process, we have developed two new named entities that are not included in published NER models for the Finnish language. This work is conducted as part of the DARIAH-FI infrastructure.

Keywords

named entity recognition, archival records, state authority archives, tool development

1. Introduction

Archives around the world are digitising their material at a vastly growing speed. This means that massive amounts of records will be made available to researchers in various fields of study. This opens up a wide range of possibilities for researchers. For historical research in particular, this kind of mass digitisation is important in helping prevent the risk of 'source myopia', which can result from very limited types of data being available in digital format[1].

The National Archives of Finland launched a mass digitisation project in 2019 ultimately aiming to digitise 135 shelf kilometres of state authority records with the intent of destroying the original documents [2]. The mass digitisation project includes various areas of development such as improving the quality of optical character recognition (OCR) and segmentation detection [3]. The aim of mass digitisation is not only to make the archives more accessible for state authorities but also to advance the possibilities of using archival material in various fields of research. Similar large-scale digitisation has been underway or forthcoming in several other national archives as well, such as National Archives of the Netherlands, State Archives of Belgium, The Swedish National Archives, and US National Archives and Records Administration (e.g., [4]), which makes developing the usability of archival data from the research perspective of particular importance.

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

venla.s.m.poso@jyu.fi (V. Poso); tanja.valisalo@jyu.fi (T. Välisalo); ida.m.toivanen@jyu.fi (I. Toivanen); antero.holmila@jyu.fi (A. Holmila); jari.ojala@jyu.fi (J. Ojala)

0000-0001-8678-4683 (T. Välisalo); 0000-0003-2456-7223 (A. Holmila); 0000-0002-4348-8857 (J. Ojala)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

Advancing the use of archival data for research is the objective of the FIN-CLARIAH[1] infrastructure project, which focuses on finding the best practices for enriching and accessing the recently digitised data in the National Archives of Finland. Making the digitised archives more accessible and usable for research purposes demands enriching the data in various ways. One way to make the data more usable for researchers is to utilise a natural language processing (NLP) task called Named Entity Recognition (NER). NER is an information extraction method, which is used to identify different types of entities, such as persons, organisations, places, dates, times, or events, from unstructured text. The development of NER for state authority archives in this project will lead to the deployment of NER models as open-source tools for researchers, as well as integrated in the services of the National Archives of Finland.

Before starting our project, some of the digitised data could be accessed through the National Archives' online service ASTIA[2]. ASTIA is meant for browsing and accessing digitised documents through a web browser. ASTIA interface consists of a split screen with the original document image on the left and the text file on the right. Individual documents can be downloaded as 1) JPG, 2) PDF, and/or 3) ALTO XML files, and larger document collections as separate JPG and ALTO XML files. These options are certainly diverse and sufficient enough for most current needs with openly available documents, especially those with traditional qualitative research approaches. However, for many digital humanities and social sciences methods (SSDH), particularly those identified as big data methods, this is not the most convenient or efficient technique for accessing the data. Additionally, sensitive data needs a secure method for accessing and browsing it. Therefore, there is a need for other forms of data transfer.

In this paper, we will describe the design of the NER process for digitised state authority archives and consider the potential benefits and challenges of using NER in historical research with this type of archival data. As part of the design process, we created a survey directed at researchers within the fields of humanities and social sciences in order to bring new perspectives to the possibilities of using NER. We will also report the results of this survey and how they were incorporated in the process.

2. Digital history

Digital humanities (DH) is an interdisciplinary field where computational studies and humanities meet [5]. It is debatable if fields such as digital history, corpus linguistics and other digitally oriented research factions fall under DH or whether they are co-managing the field. Nevertheless, in this paper we concentrate on the field of digital history. Similar to the definition of DH, digital history is a diverse field which can be determined from the perspective of the subject of study or the methodological approach [6, 7]. Hannu Salmi [6] has defined digital history as "an approach to examining and representing the past; it uses new communication technologies and media applications and experiments with computational methods for the analysis, production

[1]FIN-CLARIAH infrastructure project 2022-2023, funded by the Academy of Finland, comprises FIN-CLARIN and DARIAH-FI, which are part of European research infrastructures CLARIN ERIC and DARIAH-EU. FIN-CLARIAH aims to develop processes, methods and tools for processing unstructured text in social sciences and humanities research.

[2]https://astia.narc.fi/uusiastia/

and dissemination of historical knowledge." The ambiguousness of the field has been captured by Seefeldt and Thomas[7]:

> On one level, digital history is an open arena of scholarly production and communication, encompassing the development of new course materials and scholarly data collections. On another, it is a methodological approach framed by the hypertextual power of these technologies to make, define, query, and annotate associations in the human record of the past.

We emphasise the methodological approach and centre our attention to the development and possibilities of the computer-assisted methods and tools. Digitised data holds a multitude of possibilities for historians, in addition to remote access in itself [8]. Within the field of digital history researchers have found new angles to old subjects. The multitude of methodological paths to choose from is forever expanding. Although the digital history approaches might offer some objectivity to the process and reveal unseen patterns from the used data, the researchers' choices are still on the lead. Within topic modelling, sentiment analysis, text network analysis or other data mining options there are various choices to make which affect the outcome [9]. Using a computer-assisted approach helps the researcher to avoid 'cherry picking', which means that researcher finds parts of the data to support their preliminary hypothesis and disregard other viewpoints [10, 11]. Also, the major promise for digital history in the historian's perspective is the possibility of examining vastly more data than has been the typical practice among historians. As the amount of data has exploded over the last couple of decades, it is evident that in the future historians will need a totally new toolset for the practice of their discipline. According to Thaller[12]:

> The Humanities have since their earliest inception always been focusing on the ability to draw a maximum of conclusions from a rather limited amount of information they could access physically. The [sic] only start to notice that this barrier has broken down. The primary qualification of a Humanities' researcher of the year 2050 will not be, how to lovingly extract insights from a few isolated bits of information, but how to meaningfully integrate the information contained in the largest possible set of data.

However, despite the promise of DH, digital history is not a short-cut for new insights and ideas. As David Blei [13] has written:

> . . . statistical models are meant to help interpret and understand texts; it is still the scholar's job to do the actual interpreting and understanding [. . .] the hope is that the model helps point us to [. . .] evidence. Using humanist texts to do humanist scholarship is the job of a humanist.

The field of digital history is not without its problems. Researchers must understand the assisting technologies and what the information extracted with them actually tells. In addition to that, we are aware of the possible pitfalls, such as the "virtual dismemberment" [8] of archives, when single documents can become separate from the collection they are part of, and thus, lose some of their interpretative potential. As Lara Putnam has reminded, for the first time, historians

(and other humanists), can find vast amounts of data without understanding the genealogy of the data - that is, who has created it and why; what is its place in the larger hierarchy of the archive and so on: "Web-based full-text search decouples data from place [...] for the first time, historians can find without knowing where to look.[14]"

Part of conquering these difficulties with digital history is choosing the right computer-assisted methods and understanding the limits and possibilities that they offer. As the DH field is booming and new methods, analysis tools and datasets are mushrooming, the challenge for newcomers is simply where to start. In this paper, we concentrate on the NER and its applications in the field of digital history. Locating the different entities from the archival material offers valuable information for the researchers in various stages of the research process, and enriching the archival data with NER can be useful to the archive itself.

3. Named Entity Recognition

Named entity recognition (NER) was originally developed as a form of information extraction (IE)[15]. Current use of NER exceeds the original purpose, and it has been utilised in a wide range of different NLP tasks. The core task of NER has remained the same, as locating and naming predefined entities[16], but the development of new applications continues to be a popular field in natural language processing (e.g., [17, 18, 19, 20, 21, 22, 23, 24]).

Nordic and Baltic languages have quite sufficient NER models and corpora (see [25, 26, 27]). For Finnish NER, there are three notable corpora: 1) The FiNER[24], 2) Turku NER[28], and 3) TurkuONE[29]. The FiNER corpus is mainly based on single-domain text, technology news from the magazine Digitoday, so its contribution is limited when moving on to a different domain. In FiNER, there are six entity groups (organisation, location, person, product, event, date). The Turku NER corpus took this into account, by being constructed from various domains and text types. The Turku NLP group has created TurkuONE, a new fine-grained NER corpus, which combines and extends the two previously published corpora. The most notable difference is in the used NER categories, which have been revised to match international standards. This means that the number of categories has been changed from six to eighteen different entity groups (based on OntoNotes 5.0, see [30]). It is important to note that the definitions for entity groups differ from the older NER corpora to TurkuONE, since the fine-grained version divides some of the categories, such as location, to smaller sections, such as facility, geopolitical location, and other locations.

The aforementioned NER corpora have been used in the training and testing of different NER models. For example, the first two have been used to train and test FiNER tagger, which is a rule-based NER model[31, 24]. As seen in Table 1, the FiNER tagger performed reasonably well for FiNER corpus, but when tested with Turku NER corpus, the F1 score, which is used to measure the accuracy of machine learning models, dropped remarkably. There are also differences in performance of FiNER tagger when it comes to different entity groups. For example, when FiNER tagger was tested with a Wikipedia test set, the overall F1 score dropped to 79.91, while the scores for PRO, ORG and EVENT classes were close to or under 60[24].

Recently the FiNER tagger has been outperformed by other models. Development of the BERT model[32] has brought new possibilities to the NLP field. BERT can be used as a backbone in

Table 1
FiNER tagger scores, tested with Turku NER corpus[28] and FiNER corpus/Digitoday test set[24]

NER corpus	F1-score	Prec.	Rec.
Turku NER corpus	74.08	77.16	71.24
FiNER corpus	85.20	90.79	80.25

Table 2
FinBERT scores, tested with TurkuONE corpus and Turku NER corpus[28, 29].

NER corpus	F1-score	Prec.	Rec.
TurkuONE corpus	92.99	92.58	93.41
Turku NER corpus	91.65	90.87	92.44

tasks such as NER. The Finnish version of BERT, the FinBERT model, is pre-trained from scratch on Finnish data[33]. FinBERT has been used in testing the most recent NER corpora[28, 29] and performed well in both cases. As seen in Table 2, FinBERT seems to perform better on the test set from the TurkuONE corpus, implicating that the difference in performance is explained by the extended number of entity groups. Drawing from these results, we hope to build a NER tool for state authority archives using FinBERT.

4. Applications of NER in historical research

"...historian exhausting the records before they exhaust the historian.[34]"

Researchers can rarely control the processes that lead to the formation of the archives which they choose to examine. Independent of the type of the archive, decisions need to be made on what is preserved and in what quantity. Additionally, when it comes to the state authority and administrative processes, the amount of documentation has increased over the course of history [35]. The variation in archival practices over the years and in different institutions have added their own twist to the overabundance of archival material[36]. This has led to a situation where serendipity might play an important role in researchers' work[36]. NER can assist in improving information retrieval and making the process of selecting the relevant archival material for research purposes a more traceable process[37]. Common search functions enable researchers to find information with a specific and controlled vocabulary, which most times gives exactly what the researcher is looking for. However, using a search based on NER, the search produces a wider set of results, which might reveal something beyond what the researcher expected. Nevertheless, the NER-based search results still need to be analysed by the researcher and all results do not automatically hold meaning for the research. Further analysis can include

determining what is the meaning of each entity in a passage or document.

Improving information retrieval for archival data is and will be an important part of NER, but entity recognition offers a wide set of other possibilities as well. It can work as a basis or in connection with other digital humanities and digital history methods. As the three fundamental entity types, person, organisation and location (MUC-6 competition[38]) often are most frequently mentioned across the document types, the recognition of them also seems to be the most trustworthy and most often used as a basis for research. For example, [39] studied canonisation of cultural memory in the online audiovisual archives of the Finnish Broadcasting Company using most frequently appearing entities of names, places and events in the archival metadata. As another example, Erik Edoff used frequency of place entities in newspapers from different eras in the late 1900s to see if new technologies really made the world smaller[40]. Place entities revealed that contrary to what is generally perceived, newspapers included more mentions of places in the local region than of far-away places, which would not speak for a smaller world, but a tighter and constant connection with the neighbouring cities. Place entities can be further explored using geomapping[41], where the place entities are combined with map visualisations. In historical research, geomapping has often been used to depict temporal changes. For example, Clifford[42] studied the development of an industrial ecosystem with the aid of geomapping.

In addition to locations, the entity category "person" has been a valuable source in historical research. The frequency counts of specific persons can be used to detect their cultural meaning when the data supports this kind of hypothesis[39]. A more complex NLP task compilation where NER has been used as a basis for the analysis, is presented in a study by Fields et al.[43] on the Ottoman-Iraqi personal diaries. There, named entity recognition is used alongside network analysis to map a person's daily life, community structure, and social relations.

5. Survey on named entities

In order to gain a deeper understanding of potential needs for NER based tool development, we conducted an online survey targeted at researchers interested in using state authority archives. The aim of the survey was to support sustainable NER development, where diverse research perspectives would be taken into consideration starting from the beginning of the process. The survey was distributed to researchers in the fields of history and social sciences in particular, through universities, research societies and conferences between November 2022 and February 2023.

The survey gathered 57 responses from Finnish researchers. The respondents represented multiple research areas with an emphasis on history (see Table 3). The survey described briefly what named entity recognition is, and the respondents were given a list of named entities and asked which named entities they felt were most useful for them. The results give a general idea about the entity types that the researchers would prefer, although it is important to note that the number of survey respondents was fairly low.

The survey respondents were also asked to self-evaluate their previous experiences with digital research methods on a given scale (Table 4). Majority of respondents (91.2%) described themselves as having at least little knowledge of digital research methods. When asked to

Table 3
Participants according to their academic discipline (n=57).

Academic discipline	f	%
History	31	53.4
Other Humanities	12	20.7
Social Sciences	11	19.0
Other disciplines	2	3.4
Multiple disciplines	1	1.7
Total	57	100.0

Table 4
Previous experience with digital research methods (n=57).

Previous experience with digital research methods	N	%
None	5	8.8
Little	13	22.8
Some	27	47.3
Much	7	12.3
Very much	5	8.8
Total	57	100.0

elaborate, half of the respondents (27; 49.1%) gave a more detailed description. Most common were mentions of using digital data (e.g., digitised archival data) with 10 respondents mentioning only this form of digital research methods. Other common responses described using digital analysis tools (9 mentions) or digital search tools (8 mentions). There were also mentions of using digital databases, digital tools for collecting data (e.g., survey tools), and digital data management tools.

The main question in the survey was on the perceived usefulness of different named entities in regard to the state authority archives. The respondents were presented with 19 different entities along with a few examples, and they were asked to evaluate how useful these entities were on a 4-point scale with options 'very useful', 'useful', 'possibly useful', and 'not useful'. All respondents answered this question but not all entities. The response rate per entity differed between 89.5% and 100%. The entities considered 'very useful' were Journal number, Date, Nationality, religious or political group, Geopolitical location, Organisation, and Person (see Figure 1).

Respondents were also asked what other things or entities might be useful to recognize in the survey data. The responses (n = 18) were quite diverse and some of the things mentioned can already be solved through existing categories; for instance, different joint municipalities, when directly mentioned, would fall under the organisation (ORG) entity. Many suggestions were also so specific that they would be more easily attained by traditional search functions (e.g., nuclear power, inflation). One suggestion made by more than one respondent was profession or professional title. However, this question also yielded an interesting result pertaining to researcher needs: several respondents mentioned needs pertaining to the metadata, such as the

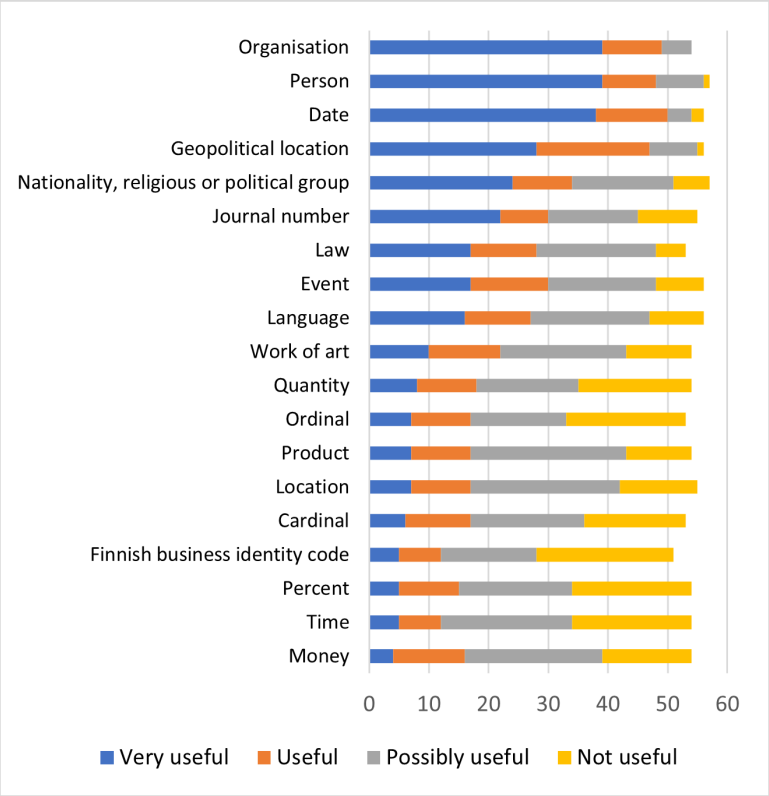

Figure 1: The perceived usefulness of different entities by survey respondents.

need to know what type of documents is in question or whether the named entity (NE) is in a heading or body text.

6. Developing NER for digitised state authority archives

In applying named entity recognition for digitised state authority data, our objective is to pave the way for researchers who wish to use archival data in novel ways. The National Archives of Finland is currently digitising masses of archival documents, which means there is a need for rapid and consistent metadata generation. Our process for named entity recognition follows similar steps as the previous work done on Finnish NER. Existing Finnish NER models, however, are not tailored to archival data, so developing NER further is crucial. We aim to document, evaluate and report the whole development process for faster deployment of mass digitised archival data in research. We approach the process from the perspective of creating tools for the end-user who would like to better utilise archival data in their research.

In the following, we will first describe the process of accessing state authority records in the National Archives, and then the annotation of named entities. We will describe the particular demands that state authority archives make on the process.

6.1. Accessing data from the National Archives

As a pilot data for our project, we used the mass digitised archives of Finland's Ministry of Economic Affairs and Employment. In order to access the digitised data, the research institution makes a data transfer agreement with all the parties involved - NARC, data owner, and CSC (see Figure 2). CSC, or IT Center for Science, is a Finnish government owned company, which provides higher education and research institutions research infrastructure services such as supercomputers and servers. As this project was aimed at research infrastructure development, particularly developing methodological tools, it was not a typical research project where agreement and licensing processes between state authorities and archives are already established. Additionally, the large quantities of data made it necessary to use third party services for data transfer, which was another departure from the more common forms of cooperation between these organisations. These novel features of this project introduced new demands also to the formal agreements and licences.

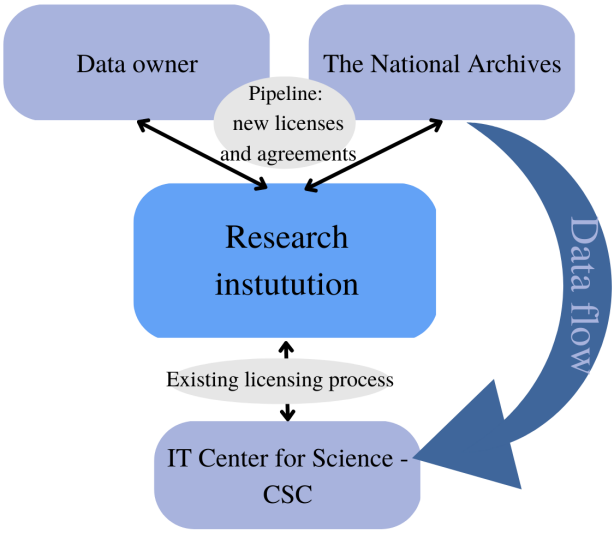

Figure 2: The agreements and data flow between different actors in making the mass digitised state authority archives available to researchers.

The pilot data contains data from 1999-2017 that consists mostly of typewritten materials. The data includes documents in several languages, but we focus only on data in Finnish. The pilot data entails personal details and confidential data, which places particular demands on the agreements and the security of transferring and processing data. The pilot data is transformed from image of text into machine-readable text using OCR technology. The quality of digitised material often varies regarding the OCR, metadata and data structure[44]. The document type can also heavily influence the results yielded by applying NER. For example, documents containing an abundance of tables have been found to affect NER results with archival data (e.g., [41]).

Currently, the quality of the OCR is generally fairly sufficient for research purposes. Particular challenges of applying OCR for the pilot data at this stage include hand-written data, non-

conventional document layout, headings, and special characters. Handwritten text recognition (HTR) is not yet included in the mass digitisation process. Non-conventional document layout, for example, a 2-column layout, is not recognized in the OCR process as separate text areas, which results in the output text rows consisting of fragments from both columns, breaking the original sentences and paragraphs. Headings in the text are not consistently recognized as such, but are sometimes combined with the adjacent paragraph sign, and are also inconsistently recognized as text. In addition to these difficulties, similar to the OCR errors reported in the literature[45], in the pilot data there also may be misrecognized, missing or extra characters, as well as nearby words grouped together, or division of a word into several subwords. In previous studies, it is suggested that OCR quality should be adequate in order to develop state-of-the-art NER[46]. In this context, the data having OCR generated errors makes it noisy - that is, the data is at least partially corrupted. The difficulties with noisy data may seem impenetrable, but as Fridlund et al.[1] (2020) have argued: "If we limited our research to clean datasets, very little would be accomplished." Part of this process is to document the effects of OCR issues on NER in this particular data type.

6.2. Annotation process

Based on our NER survey, researchers found Journal number, Date, Nationality, religious or political group, Geopolitical location, Organisation, and Person as the most useful named entities. NARC executed a similar survey directed to seven different authorities in which Journal number and Finnish business identity code were considered 'very useful'. Building from both survey results, we ended up with ten entity categories: person (PER), organisation (ORG), location (LOC), geopolitical location (GPE), product (PRO), event (EVENT), date (DATE), journal number (JON), Finnish business identity code (FIBC), and nationality, religious and political group (NOPR). Journal number (JON) and Finnish business identity code (FIBC) are new entities created and defined specifically in this project. In addition to the survey results, we based our entity categories on previous work on Finnish NER[24, 28, 29].

Data annotation is necessary to train deep learning models for the NER task. There are several steps in the annotation process. The annotation scheme we use is BIO/IOB2 format. First, we preprocess the data. This includes changing the format from AltoXML to CSV, as well as filtering out the data that is in another language than Finnish. Then, we add pseudo-labels for entities using a previously trained NER model by TurkuNLP. After pseudo-labelling, there are still two entity groups (i.e., journal number and Finnish business identity code) that are not pre-annotated. After the first round of model-aided annotations, we manually examine the data and pseudo-labels and make necessary additions and corrections. For the whole annotation process, CSC's computational services are used in order to handle the data in a secure environment. CSC offers computational services for sensitive data (e.g., SD Desktop and SD Connect), which makes it possible to develop the methodological tools needed.

While some researchers have measured inter-annotator agreement when using manual annotation as part of the NER process (e.g., [47]), we opted for creating precise NER annotation guidelines before the annotation began and refining the guidelines through close communication between annotators as the annotation was taking place. Similar decision was made when making the TurkuONE corpus, where only one annotator was involved in the process[29]).

After the manual annotation phase, the annotated data is used for model training. The annotation tool is then tested with other data in cooperation with researchers to define the need for improvements in the process. Re-training and testing are executed if needed. The final aim is to distribute the NER tool for researchers to use.

7. Conclusions and discussion

Mass digitisation produces previously unseen quantities of archival data in a uniform digital format. However, the methods and tools for using this data are still under development. In this paper, we have reported the two main contributions of our work thus far. First, we have defined the process of accessing and using the state authority records from the National Archives in order to make the process available as a benchmark for future researchers. Second, we have applied new named entities to the NER annotation process for Finnish language text. In our exploration of using NER for historical research and state authority data in particular, we started by mapping the existing needs of researchers within the field. Based on the survey and the survey by NARC, we identified the need for two named entities, Journal number (JON) and Finnish business identity code (FIBC), which were not included in existing NER models.

State authority archival records are often arranged based on the types of content (letters, minutes etc.), rather than by the topics or themes present inside the documents. This makes named entity recognition particularly useful for researchers as it can help recognize the documents that are useful with a particular research topic. NER provides a multitude of possibilities for researchers. For example, it can help identify different actors (e.g., advocacy groups) affecting different processes. NER also enables tracing policy trends and effects of local/world events in different processes. It can also help identify regional variations, which can further be examined using visualisations. For example, researchers could explore whether certain areas are emphasised when implementing certain policies or distributing funds, or which particular foreign countries or cities are present as points of comparison or partners in certain areas. Furthermore, NER can open up new perspectives on the state authority practices when used for open exploration of archive contents.

Next steps in the process include testing the developed NER model with a wide variety of state authority data as well as other types of archival data. Applications of the NER model should include combining NER results with metadata extracted using other techniques, such as identifying document structures. As state authority data also includes data and documents in other languages, especially Swedish, a multilingual NER development is one possible direction for further tool development. Future work on named entity recognition and analysis tools based on NER should entail actively utilising new technological advancements emerging in the fields of natural language processing and machine learning.

References

[1] M. O. Fridlund, Mats, P. Paju, Digital Histories: Emergent Approaches within the New Digital History, Helsinki University Press, 2020. URL: https://doi.org/10.2307/j.ctv1c9hpt8.

[2] T. Hölttä, V. Kajanne, No more new archive buildings – mass digitisation and retroactive digitisation improve the accessibility of material, in: J. Nuorteva, P. Happonen (Eds.), The National Archives of Finland Strategy 2025: Perspectives for the future, The National Archives of Finland, 2020, pp. 14–15. URL: "https://kansallisarkisto.fi/documents/141232930/153230445/KA_Strategy_2025_eng.pdf".

[3] T. N. A. of Finland, Dalai - using artificial intelligence to improve the quality and usability of digital records, ???? URL: https://kansallisarkisto.fi/en/dalai-en.

[4] L. Hirvonen, Survey of Digitization in Archives, The National Archives of Finland, 2017. URL: https://kansallisarkisto.fi/documents/141232930/150411434/Liite_2_Digitization_Survey_2017.pdf.

[5] T. Schwandt (Ed.), Digital Methods in the Humanities: Challenges, Ideas, Perspectives, Bielefeld University Press, 2021.

[6] H. Salmi, What is Digital History?, Polity Press, Cambridge, UK, 2021.

[7] D. Seefeldt, W. G. Thomas, What is Digital History? A Look at Some Exemplar Projects, Perspectives on History (2009). URL: https://www.historians.org/research-and-publications/perspectives-on-history/may-2009/what-is-digital-history.

[8] B. Ogilvie, Scientific Archives in the Age of Digitization, Isis 107 (2016) 77–85. URL: https://www.journals.uchicago.edu/doi/full/10.1086/686075.

[9] S. Ramsay, Databases, in: S. Schreibman, R. Siemens, J. Unsworth (Eds.), A Companion to Digital Humanities, Blackwell Publishing Ltd, 2004, pp. 177–197. doi:10.1002/9780470999875.ch15.

[10] P. Baker, C. Gabrielatos, M. KhosraviNik, M. Krzyżanowski, T. McEnery, R. Wodak, A useful methodological synergy? combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the uk press, Discourse & Society 19 (2008) 273–306. doi:10.1177/0957926508088962.

[11] V. Koller, G. Mautner, Computer applications in critical discourse analysis, in: C. Coffin, A. Hewings, K. O'Halloran (Eds.), Applying English grammar: functional and corpus approaches, Routledge, London, 2020, pp. 216–228.

[12] M. Thaller, The humanities are about research, first and foremost; their interaction with computer science should be too, in: C. Biemann, G. R. Crane, C. D. Fellbaum, A. Mehler (Eds.), Computational Humanities - bridging the gap between Computer Science and Digital Humanities (Dagstuhl Seminar 14301), volume 4 of *Dagstuhl Reports*, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2014, pp. 80–111. doi:10.4230/DagRep.4.7.80.

[13] D. M. Blei, Topic modeling and digital humanities, Journal of Digital Humanities 2 (2012). URL: https://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/.

[14] L. Putnam, The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast, The American Historical Review 121 (2016) 377–402. doi:10.1093/ahr/121.2.377.

[15] D. D. Palmer, D. S. Day, A statistical profile of the named entity task, in: Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics, Washington, DC, USA, 1997, pp. 190–193. doi:10.3115/974557.974585.

[16] A. V. K. S. A. O.-B. Marco Humbel, Julianne Nyhan, The effect of morphology in named

entity recognition with sequence tagging, Journal of Documentation 77 (2021) 1223–1247. doi:10.1108/JD-02-2021-0032.

[17] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 260–270. URL: https://aclanthology.org/N16-1030. doi:10.18653/v1/N16-1030.

[18] X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1064–1074. URL: https://aclanthology.org/P16-1101. doi:10.18653/v1/P16-1101.

[19] O. Güngör, S. Uskudarli, T. Güngör, Improving named entity recognition by jointly learning to disambiguate morphological tags, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 2082–2092. URL: https://aclanthology.org/C18-1177. doi:99.9999/woot07-S422.

[20] O. Güngör, T. Güngör, S. Üsküdarli, The effect of morphology in named entity recognition with sequence tagging, Natural Language Engineering 25 (2019) 147–169. doi:10.1017/S1351324918000281.

[21] A. Katiyar, C. Cardie, Nested named entity recognition revisited, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 861–871. URL: https://aclanthology.org/N18-1079. doi:10.18653/v1/N18-1079.

[22] M. G. Sohrab, M. Miwa, Deep exhaustive model for nested named entity recognition, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2843–2849. URL: https://aclanthology.org/D18-1309. doi:10.18653/v1/D18-1309.

[23] A. Goyal, V. Gupta, M. Kumar, Recent named entity recognition and classification techniques: A systematic review, Computer Science Review 29 (2018) 21–43. doi:https://doi.org/10.1016/j.cosrev.2018.06.001.

[24] T. Ruokolainen, P. Kauppinen, M. Silfverberg, K. Lindén, A Finnish news corpus for named entity recognition, Language Resources and Evaluation 54 (2019) 247–272. doi:10.1007/s10579-019-09471-7.

[25] S. Almgren, S. Pavlov, O. Mogren, Named entity recognition in Swedish health records with character-based deep bidirectional LSTMs, in: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 30–39. URL: https://aclanthology.org/W16-5104.

[26] B. Johansen, Named-entity recognition for Norwegian, in: Proceedings of the 22nd Nordic Conference on Computational Linguistics, Linköping University Electronic Press, Santa Fe, New Mexico, USA, 2019, pp. 222–231. URL: https://aclanthology.org/W19-6123.

[27] L. Derczynski, C. V. Field, K. S. Bøgh, DKIE: Open source information extraction for Danish, in: S. Wintner, M. Tadić, B. Babych (Eds.), Proceedings of the Demonstrations

at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 61–64. doi:10.3115/v1/E14-2016.

[28] J. Luoma, M. Oinonen, M. Pyykönen, V. Laippala, S. Pyysalo, A broad-coverage corpus for finnish named entity recognition, in: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), European Language Resources Association (ELRA), Marseille, France, 2020, pp. 4615–4624. URL: https://aclanthology.org/2020.lrec-1.567.

[29] J. Luoma, L.-H. Chang, F. Ginter, S. Pyysalo, Fine-grained named entity annotation for Finnish, in: Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), Linköping University Electronic Press, Sweden, Reykjavik, Iceland (Online), 2021, pp. 135–144. URL: https://aclanthology.org/2021.nodalida-main.14.

[30] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, M. El-Bachouti, A. H. Robert Belvin, Ontonotes release 5.0, 2013. doi:doi.org/10.35111/xmhb-2b84.

[31] K. Kettunen, L. Löfberg, Tagging named entities in 19th century and Modern Finnish newspaper material with a Finnish semantic tagger, in: Proceedings of the 21st Nordic Conference on Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2017, pp. 29–36. URL: https://aclanthology.org/W17-0204.

[32] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: CoRR, volume abs/1810.04805, 2018. URL: http://arxiv.org/abs/1810.04805.

[33] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, S. Pyysalo, Multilingual is not enough: BERT for finnish, CoRR abs/1912.07076 (2019). doi:10.48550/arXiv.1912.07076. arXiv:1912.07076.

[34] W. H. McNeill, Mythistory, or Truth, Myth, History, and Historians, The American Historical Review 91 (1968) 1–10. doi:10.2307/1867232.

[35] S. Myllyniemi, Suomen historian asiakirjalähteet, 2nd. ed., Kansallisarkisto; WSOY, Helsinki, 1994.

[36] S. Decker, The silence of the archives: business history, post-colonialism and archival ethnography, Management & Organizational History 8 (2013) 155–173. doi:10.1080/17449359.2012.761491.

[37] D. Colla, A. Goy, M. Leontino, D. Magro, C. Picardi, Bringing semantics into historical archives with computer-aided rich metadata generation, J. Comput. Cult. Herit. 15 (2022). doi:10.1145/3484398.

[38] R. Grishman, B. Sundheim, Message Understanding Conference- 6: A brief history, in: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, 1996. URL: https://aclanthology.org/C96-1079.

[39] M. Kannisto, P. Kauppinen, Digital Histories: Emergent Approaches within the New Digital History, Helsinki University Press, 2020, pp. 165–180. doi:10.2307/j.ctv1c9hpt8.15.

[40] J. Jarlbrink, All the work that makes it work: Digital methods and manual labour, in: M. O. Fridlund, Mats, P. Paju (Eds.), Digital Histories: Emergent Approaches within the New Digital History, Helsinki University Press, 2020, pp. 113–126. doi:10.2307/j.ctv1c9hpt8.12.

[41] J. Clifford, B. Alex, C. M. Coates, E. Klein, A. Watson, Geoparsing history: Locating

commodities in ten million pages of nineteenth-century sources, Historical Methods: A Journal of Quantitative and Interdisciplinary History 49 (2016) 115–131. doi:10.1080/01615440.2015.1116419.

[42] J. Clifford, West Ham and the River Lea: A Social and Environmental History of London's Industrialized Marshland, 1839-1914, University of British Columbia Press, Vancouver, 2017.

[43] S. Fields, C. L. Cole, C. Oei, A. T. Chen, Using named entity recognition and network analysis to distinguish personal networks from the social milieu in nineteenth-century Ottoman–Iraqi personal diaries, Digital Scholarship in the Humanities (2022). doi:10.1093/llc/fqac047, fqac047.

[44] P. Ihalainen, B. Janssen, J. Marjanen, V. Vaara, Building and testing a comparative interface on northwest european historical parliamentary debates: Relative term frequency analysis of british representative democracy, in: CEUR Workshop Proceedings, volume 3133 of *Digital Parliamentary Data in Action (DiPaDA 2022) workshop*, Uppsala, Sweden, 2022, pp. 52–68. URL: http://ceur-ws.org/Vol-3133/paper04.pdf.

[45] E. Soper, S. Fujimoto, Y.-Y. Yu, BART for post-correction of OCR newspaper text, in: Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), Association for Computational Linguistics, Online, 2021, pp. 284–290. URL: https://aclanthology.org/2021.wnut-1.31. doi:10.18653/v1/2021.wnut-1.31.

[46] A. Hamdi, A. Jean-Caurant, N. Sidere, M. Coustaty, A. Doucet, An analysis of the performance of named entity recognition over ocred documents, in: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Champaign, IL, USA, 2019, pp. 333–334. doi:10.1109/JCDL.2019.00057.

[47] S. Orasmaa, K. Muischnek, K. Poska, A. Edela, Named entity recognition in Estonian 19th century parish court records, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 5304–5313. URL: https://aclanthology.org/2022.lrec-1.568.