Named Entity Recognition for early modern textual documents: a review of capabilities and challenges with strategies for the future

Humbel, Marco; Nyhan, Julianne; Vlachidis, Andreas; Sloan, Kim; Ortolja-Baird, Alexandra

Journal of Documentation (10.1108/JD-02-2021-0032)

Purpose:
Named Entity Recognition (NER) can enhance the (re)search capabilities of digitised documents and infrastructure; it can also open new possibilities for the interlinking of digitised documents with wider knowledge domains and resources. We map out the current capabilities, challenges and limitations of NER and establish the state of the art of the technique in the context of digital early-modern research.

Design/methodology/approach:
We survey the application of NER to early modern documents through a systematic review of the literature (2002 to 2019). Given the ongoing reliance on rule-based NER among digital early-modern projects, we map the landscape of authority files. Furthermore, we present a new case study of NER research undertaken by Enlightenment Architectures: Sir Hans Sloane's Catalogues of his Collections (2016-21), a Leverhulme funded research project and collaboration between the British Museum and University College London, with contributing expertise from the British Library and the Natural History Museum.

Findings:
Currently it is not possible to benchmark the capabilities of NER applied to documents of the early-modern period, because more robust reporting of NER approaches is required. We highlight open questions around the ethical and socio-cultural import of NER and authority files, and propose future directions that might be followed to push forward the state of the art.

Originality:

This paper brings together previously fragmented academic and grey literature on the technical elaboration and application of NER. We set out a comprehensive summary of digital tools and resources to apply NER to early-modern materials.

Keywords:
Artificial Intelligence; Information Extraction; Data Criticism; Data Ethics; Named Entity Recognition; Digital Humanities

1. Introduction and Research Context

In this review article we ask: what is the state of the art of Named Entity Recognition (NER) as applied to early modern documents (1400-1800)? To what extent can the early-modern knowledge base of NER systems be enriched with extant authority files? Which NER systems are currently being used by digital cultural heritage and digital humanities projects? What accuracy rates are they reporting? To answer these questions, we bring together previously fragmented academic and grey literature on the technical elaboration and application of NER, and present a survey of the digital projects and tools that are currently using it. In doing so, we compile a detailed synthesis of this technique, which will assist researchers who seek to apply NER to the abundant early-modern materials held in Museums, Galleries, Archives, Libraries and private collections, and, indeed, to the abundance of digitised documents and collections of the period that are available online. This article also highlights open questions around the ethical and socio-cultural import of NER and authority files, and proposes future directions that might be followed to push forward the state of the art.

NER is an information extraction technique for identifying, segmenting and labelling named entities of common interest, like those of people, organizations, places, currencies, time and percentage expressions (Grishman and Sundheim, 1996: 467). Typical named entities permit the investigation of "who did what when and where with whom" (Ehrmann et al., 2016: 98). The many digitisation projects that have been undertaken by private, public and private-public bodies, individuals and communities in recent years have resulted in a significant section of Humanities' analogue record being remediated as binary data. Yet, simple digitisation is not usually sufficient to realise the possibilities of conceptualising and operationalising previously self-contained historical documents as "open sets of data". Named Entity Recognition (NER) is one of the approaches that can play a pertinent role in this by unlocking digitised outputs (Kaplan, 2015: 3; Hoekstra and Koolen, 2019: 90; Piskorski and Yangarber, 2013: 23–24).

NER can allow the early modern period's materials, which include a mass of handwritten and printed resources including, inter multa alia, books, pamphlets, letters, collection catalogues, birth registers and encyclopaedias, to be enriched for professional researchers, information professionals and the interested general public (Southall et al., 2011: 133–34). Accurate NER can thus facilitate the study of, for example, people and places, and their social and spatial interconnections and representations as manifested in historical corpora. (Liu, 2018: 135–36, Gelling, 2011: 1–2). With regard to the early modern period, the opportunities of studying and linking person entities is particularly promising. Nelson has observed that NER is "[...] a sweet spot for prosopography, the study of large data sets about people and their relationships within a well-defined group or

network." (2014: 3). It can allow documentation to be discovered and interlinked in new ways: "people, individuals as well as organisations, are primary entities that serve to interconnect information across repositories, collections and systems and, more broadly, in the open Web" (Angjeli et al., 2014: 2). It is also notable that the efforts of academic publishers to monetize the text analysis of digitised cultural heritage materials have included NER among the other foundational computational routines offered via their gated platforms (see, for example, Gale Digital Scholars lab: https://www.gale.com/intl/primary-sources/digital-scholar-lab and ProQuest TDM Studio: https://about.proquest.com/products-services/TDM-Studio.html).

Accordingly, a number of digital humanities projects have sought to annotate the named entities in their datasets (Stork et al., 2018: 1), including the 'ASCH project' which digitized archival resources related to Georg Thomas von Asch (1729-1807). The linking of entities across this dispersed collection allowed the provenance of collection items to be specified (ASCH, 2018). The Archaeology of Reading in Early Modern Europe (AOR) facilitates the investigation of the handwritten annotations of John Dee and Gabriel Harvey in early printed books, which contain, among other entities, people and places (AOR, [n.d.]). Date, person and place entities are also crucial to the contextualization of objects held in early-modern museum collections, which the Digital Ark project seeks to document (DigitalArk Project, [n.d.]). Yet, as far as these projects report on their methods,[1] they did not use automated NER, but annotated entities manually. This is actually not surprising. For all the benefits that NER can deliver, significant difficulties attend efforts to bring automated NER to bear on digitised archival documents.

This can be due to a range of factors that may be internal and/or external to the respective corpora and documents. Relevant factors range from the rich textual content of early modern documents to the traces of the contexts and temporalities of when and how documents were initially made or subsequently digitised. Multilinguality is a common feature of such documents (Ravenek et al., 2017: 319), as is unstandardized spelling (Nevalainen, 2006: 1–9), which can be further complicated by the inadvertent transcription errors that were made by scribes (Piotrowski 2012: 11–13) and modern transcribers. Moreover, the accuracy of automated processes like Optical Character Recognition (OCR), and Handwritten Text Recognition (HTR), which are used to extract text from digitised images of printed and handwritten documents, vary (Prell, 2018: 13; Smith and Cordell, 2019; Tanner et al., 2009). Research has shown that inaccurate automated transcriptions can have a significant impact on the results of research (Prescott, 2018: 63–66) and, of course, on the accuracy of automated approaches to NER.

[1] We contacted the mentioned projects in April 2020 to verify that this is indeed the case.

The inherently dynamic nature of named entities can also prove problematic for NER precision. Place names can change and so can the spaces they represent. The genre of a given document may also be decisive: place names may also be used in documents in figurative or imaginary senses (McDonough and Camp, van de, 2017: 2; Southall et al., 2011: 128). This can make the use of contemporary longitude and latitude values, and general authority files, redundant yet specialized authority files may not be available. Named entities may contain inherent ambiguities that require disambiguation (though it should not be assumed that this is always possible. The identities of some individuals have been irretrievably lost to history). A personal name can, for instance, also be a place name or an occupation (like 'Miller'). Entities are also often ambiguous, raising questions about to which 'Newscastle' or 'Richmond' the tagged entity refers (Bontcheva et al., 2002: 618; Gregory and Hardie, 2011: 302).

The early modern period is fertile ground for a study of NER because of the complexity that characterises much of its documentation. The application of NER to this documentation has the potential to advance digital approaches to the study of the period, while the lessons of applying this technique to complex early-modern documentation will be applicable, at least in part, to handwritten and printed texts of other periods too. NER is, for instance, applied to newspaper archives (Mac Kim and Cassidy, 2015; Kettunen and Ruokolainen, 2017; Rochat et al., 2016), modern literature and writing (Sprugnoli, 2018; Borin et al., 2007), and medieval manuscripts (Aguilar et al., 2016).

Thus, while NER can play an important role in digital workflows, the practical aspects of its implementation remain challenging, as we will now set out.

2 Evaluative matrix for review of NER of early-modern text

NER systems can be broadly classified into systems based on hand-crafted rules and statistical techniques. The latter includes machine learning approaches such as supervised, semi-supervised, and unsupervised (Nadeau and Sekine, 2009: 7–11), and most recently deep learning systems (Yadav and Bethard, 2019). External knowledge bases have a greater role in rule-based techniques than supervised and unsupervised machine learning techniques, which mainly rely on training sets and models. In the sections below, we review state of the art implementations of these approaches on early-modern materials and summarise our findings according to the following matrix:

- Name of the project or corpus, plus a short description of the project itself.
- Approach: the knowledge-based resources (authority files, training sets or corpora) used and how entities were marked-up through NER.

- Results: The effectiveness of the NER system as reported by the project. The standard approach to assess the effectiveness of NER systems is to account on the 'Precision' (P) and 'Recall' (R) (Grishman and Sundheim, 1996: 466). P = the percentage of correctly identified entities in relation to all identified entities. R = the percentage of the correctly identified entities in relation to the absolute number of entities with the corpus. The harmonic mean between Precision and Recall is expressed in the F1-score (Manning et al., 2008: 142–44).

For the survey below, we searched databases and journals that cover the field of Digital Humanities, broadly conceived (IEEE, Springer Link, CLARINE, ACM Digital Library, Google Scholar, TEI Journal, International Journal of Humanities and Arts Computing, Frontiers in Digital Humanities, Digital Humanities Quarterly, Journal of Digital Humanities, Digital Scholarship in the Humanities). Query terms used include: Named Entity Recognition; Information Extraction; Natural Language Processing; Early Modern, Enlightenment; Eighteenth-Century; Seventeenth-Century; Museum Catalogues; Inventories; Cabinets Of Curiosities; and Collections. In February 2019, from a further environmental scan (or review of online, digital projects rather than publications about projects), we identified 9 projects that cover a time span from 2002 to 2019. Given the mass of digitized resources from the early-modern period, the number of identified projects seemed to us to be too scarce to be complete. Yet, we argue that this is likely due to the difficulties of applying NER to early-modern documents and the tendency to under-report technical methods used in the Humanities. Links to project websites, tools and authority files (as of February 2021) are given in the appendix.

2.1 Rule or pattern based NER techniques

Typical rule-based NER systems consist of a core engine and a domain specific knowledge-based resource. Such knowledge-based resources can range from relatively 'simple' indexes to sophisticated thesauri or ontologies that make relationships between knowledge-objects explicit (see below). Most rule-based NER systems use a hand-crafted rules executed in a pattern-matching mechanism, in order to identify and tag entities with respect to knowledge-based resources (Piskorski and Yangarber, 2013: 35). Exemplary cases of digital humanities projects that have used this technique on early-modern materials are discussed below.

'The Old Bailey online'[2] provides online access to the proceedings of the Old Bailey court in London from 1674 to 1913 (Hitchcock et al., 2012). Approach: 'The Old Bailey online' used GATE's (General Architecture for Text Engineering) NER engine ANNIE to tag person name

[2] The links to this and all resources mentioned subsequently are given in the Appendix

entities automatically (Shoemaker, 2005: 300–01). Bontcheva et al. (2002: 615–19) reported on the creation of pattern-matching rules for person and place entities, and person occupation and status, which were then implemented with GATE. This resulted in the enriching of the existing knowledge-base of ANNIE with additional, specialized vocabulary (for example: titles like 'Governor', or London specific locations like 'Addington Basin'). Further enrichments were added for the characteristics of early modern English (spelling, capitalisation and punctuation variations), noisy input and for entity disambiguation ('Baker' as person or occupation). Accuracy: GATE's NER engine annotated 85% of the entities within the Old Bailey's corpus (Bontcheva et al. 2002: 619). The NER task was more successful in recognizing dates and money amounts. Also, Shoemaker (2005: 301) reports that 80-90% of the names in the corpus were successfully annotated.

The focus of 'BOPCRIS'[3] is the Journals of the House of Lords (1688-1854). The BOPCRIS corpus consists of 12 volumes from 1688-1741 and one volume from 1814-1817 (Grover et al., 2010: 3875–77). Approach: The corpus served as a test case to evaluate the Edinburgh Geoparser's capabilities in identifying place names and person names.[4] 'BOPCRIS' is written mainly in English, with sections in Latin and French. NER was applied to English text only, which was identified through the tool 'TextCat. The Alexandria Digital Library Project Gazetteer' and 'GeoCrossWalk' were used as look-up lists for place names. A lexicon of forenames was used for person names, and variations of entities were saved in an 'on-the-fly' lexicon. Textual features like person titles, and the fact that person names were written in italics (the OCR text contained the font information), were used to build additional rules to support the NER system (Grover et al., 2010: 3879–81). Results: The gold-standard (manually annotated test set) contained 1181 place and 4909 person entities. On place names, the Edinburgh Geoparser achieved a precision of 55.92%, a recall of 61.56% and a F1-score of 58.61%. For person entities a precision of 81.83%, a recall of 82.57% and a F1 score of 82.20% (Grover et al., 2010: 3884–85).

'The Lancaster Newsbook Corpus' consists of surviving news periodicals of the Lancaster Newsbook from 1653 to 1654. Gregory and Hardie (2011: 297–301) investigated spatial phenomena within a sub-section of this corpus, which consisted of approximately 870,000 words. Approach: In order to extract all place entities, the authors first extracted all proper nouns that were identified through the part-of-speech tagger CLAWS and the semantic tagger USAS. The proper nouns list was then mapped against the World Gazetteer. To achieve better results, the proper nouns list was manually purged of entities that were obviously not place-names. Further manual

[3] The project is currently accessible via the Internet Archive only (see Appendix)

[4] It was easier to write rules that also identified place names that occur in person names (e.g. Earl of Essex) (Grover et al., 2010: 3875–77).

intervention was necessary due to polysemy and spelling variations of place names. The scope of the Gazetteer was similarly reduced to place name entities that were outside of Europe (Gregory and Hardie, 2011: 301–02). Results: The authors did not test the accuracy of their NER technique against a gold-standard. From the 870,000 words in the corpus, 8,430 provisional place-entities were extracted. The cleaned list of place names consisted of 6,297 entities (Gregory and Hardie, 2011: 301–05).

The objective of 'Johann Friedrich Blumenbach – Online' was to create a digital edition of the writings and natural history collections of Johann Friedrich Blumenbach (1752-1840) (Blumenbach-online, [n.d.]). A case-study of the sub-project 'Semantic Blumenbach' reports on the application of NER to twelve TEI-encoded German editions of Blumenbach's *Handbuch der Naturgeschichte*, published between 1779 and 1830. The goal of the project was to align the identified entities with the semantics of the CIDOC Conceptual Reference Model (Wettlaufer et al., 2015: 187–89). Approach: The NER task involved the marking-up of persons, places, objects, dates and domain-specific terms (Wettlaufer and Thotempudi, 2013). For an automated mark-up of named entities, some preparations were necessary: the special vocabulary of the documents required the creation of project-specific authority files, based on contemporary resources and the authority files CERL, GND and the Getty Thesaurus of Geographic Names. However, challenges were posed not only by the specialized vocabulary, but also the hierarchical structure of TEI that had already been applied to the documents in an earlier iteration. Adaptations in the NER tool were for instance necessary to accommodate line breaks in the documents (Wettlaufer and Thotempudi, 2013; Wettlaufer et al., 2015: 189–90). Unfortunately, the link to the parser that the project developed is not available through the Internet Archive or 'Blumenbach-online'.[5] A tool the authors used for a preliminary NER was SynCoPe, developed by the Deutsches Textarchiv. Date entities were identified through regular expressions (Wettlaufer et al., 2015: 189–90). Results: In a single edition of the *Handbuch*, with approximately 10,000 domain-specific terms, 1,000 persons, 1,2000 places and 1,300 references the NER tagger was able to reach a precision and recall of above 90%. Manual corrections were required (Wettlaufer et al., 2015: 189–90).

'Circulation of Knowledge and Learned Practices in the 17th-Century Dutch Republic' (CKCC) focussed on the correspondence of various 17th-century scholars, like Caspar Barlaeus or Hugo Grotius. The corpus contains over 20,000 letters. One of CKCC's outcomes includes the 'ePistolarium' research environment, which offers several tools to analyse the letters (Ravenek et al., 2017: 317–18). Approach: For 'ePistolarium' the letters were encoded in line with TEI. In order

[5] http://dhfv-ent2.gcdh.de/Tagging/ (17.01.20)

to facilitate co-citation analysis of the correspondents in the epistolary network, person names were tagged through NER. A knowledge-base was created by adding previously annotated names and indexes excerpted from the hardcopy editions of the letters. Latinised versions of names were generated automatically (e.g. Grotius → Grotii). The authority files were normalized and letters mapped to modern orthography (e.g. v → u). A rule-based algorithm[6] was then applied to mark-up person names using normalized authority files and the letter texts as an input (Ravenek et al., 2017: 318; 320). Results: the 'ePistolarium' NER tool was able to identify 94,553 of the 124959 person entities in the CKCC. An overview of the results[7] shows that the accuracy of NER can vary significantly between different correspondences (Ravenek et al., 2017: 320).

The Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers par une Société de Gens de lettres (1751–1772) served as a case-study to evaluate the effectiveness of the Edinburgh Geoparser and Perdido (McDonough et al., 2019; McDonough and Camp, van de, 2017). Approach: 100 geography-related entries of the *Encyclopédie* were randomly selected to create a gold standard set, containing 2151 place-entities. The authors did not make any adaptations to Perdido and the Edinburgh Geoparser for handling French text, besides that the latter received a French Part-of-Speech Tagger. The Edinburgh Geoparser's NER rules were thus specialized for the English language. Perdido comes in contrast to the Edinburgh Geoparser with a full French pipeline. The gazetteer Geonames was used in both cases to map the place-entities. (McDonough et al., 2019: 11–15). Results: For the Edinburgh Geoparser the authors report on a recall of 9.20%, a precision of 94.64% and a F1-score of 16.78%. Perdido achieved a recall of 55.58%, a precision of 77.71% and a F1-score of 64.10%. The results must be viewed under the light that Perdido was built for the French language and that the author's aim of their experiment was to show-case the limitations of non-specialized NER-systems and authority files. (McDonough et al., 2019: 2;15-17).

2.2 Machine learning NER techniques

Machine learning NER systems start from a large annotated training corpus, and automatically induce rules from learning algorithms (such as Hidden Markov Models, or Naïve Bayes Classifier) to identify named entities. The effectiveness of this NER approach does not depend on whether a knowledge base covers a specific domain exhaustively, but on the quality and amount of available training data (Nadeau and Sekine, 2009: 7; Piskorski and Yangarber, 2013: 36–38).

[6] Cf. Aho, Alfred V., and Margaret J. Corasick. 1975. 'Efficient String Matching: An Aid to Bibliographic Search'. *Communications of the ACM* 18 (6): 333–40. https://doi.org/10.1145/360825.360855.

[7] E.g. From 3,074 entities in correspondence 'Barlaeus' 1,460 were identified. From 4,096 entities in correspondence 'Descartes' 3,917 were identified (Ravenek et al., 2017: 320).

Systems that are based on semi-supervised learning techniques (also known as 'bootstrapping') start from a small training set. The NER system then exploits additional contextual information from the given examples to identify entities within a similar context. Through succeeding iterations, the system repeats the expansion process to increase its knowledge base (Nadeau and Sekine, 2009: 8; Piskorski and Yangarber, 2013: 38–39). Recent NER systems that build up on deep learning feed character or words into Recurrent Neural Networks (Yadav and Bethard, 2019).

Sluijter et al. (2016) evaluated different NER techniques based on 100 pages (containing 366 resolutions from the year 1725) of the 'Resolutions of the Dutch States-General from 1576 to 1795'. Approach: The resolutions are highly structured documents and the text block 'session' reports, for instance, the information on the session's date, the name of the chairman and list of the attendees and the province they represent. The research-team enhanced a rule-based tool from the Stanford Natural Language Processing Group to identify the dates of the sessions. To identify the provinces, names of attendants and the chairman, a tool based on a Naïve Bayes Classifier was used (Sluijter et al., 2016). The paper does not link to the created tool, or report what tool exactly was created, or how it was trained. Results: The authors report only on the NER task of identifying the dates. Here a precision of 91,1% was reached. This paper reports on a then ongoing project, and the authors suggest, for instance, the future incorporation of the 'Biography Portal of the Netherlands' and the 'Compendium of Office-Holders and Civil Servants 1428-1861' authority files to improve NER accuracy further (Sluijter et al., 2016).

As part of the 'Reassembling the Republic of Letters' project, Won et al. (2018: 2–3) evaluated the effectiveness of 5 NER tools on place entities in the 'Mary Hamilton Papers' (1756-1816) and 'Samuel Hartlib' (1600-1662) early-modern letter collections. The evaluation focussed on 161 letters of the Mary Hamilton collection and 54 letters of the Hartlib corpus. Approach: Pre-processing was necessary for both datasets; including the removal of features such as hyphenations or editorial information in the pre-existing mark-up. Problems in automating the cleaning process for the Hartlib corpus led the authors to perform a full clean process (the closest possible version to the original text) and a fast-clean process (text with missing words). The Hartlib letters are written in Early Modern English and the Mary Hamilton Papers in Modern English. Thus, the authors performed an automated translation of the Hartlib letters from Early Modern English to Modern English with with MorphAdorner and VARD. For the Hartlib letters, the NER evaluation was performed 4 different times, according to the different pre-processing methods.[8] The NER task was conducted with the Stanford Named Entity Recognizer, the NER-Tagger software package, spaCy, Polyglot, and The Edinburgh Geoparser. Excepting for the latter, these tools are based on supervised learning techniques. However, the authors did not re-train the systems, but used only the models

that came out of the box with the tools. The evaluation also included testing an ensemble method, where the 5 NER systems were combined to a voting system. In this system each place entity candidate was checked against the Early Modern Letters Online (EMLO) gazetteer, where an additional vote for an entity could be obtained (Won et al., 2018: 4–7). Results: Polyglot achieved the best results on the Hamilton corpus with a F1 score of 61.6%. The NER-Tagger delivered the lowest result with a F1 score of 51.2%. The Stanford NER tool delivered the best F1 score (with 70.8% the highest and 67.8% the lowest) on all 4 Hartlib test cases. However, the same tool achieved on the Hamilton papers a significant lower F1 score of 59.5%. Overall the best results were obtained through the voting system, where a potential place entity had to receive at minimum 2 votes (Won et al., 2018: 8–10).

Toledo et al. (2019) present an information extraction architecture that is capable of identifying named entities, the relation between the entities and automatically transcribes handwritten text. For their experiment the authors used Catalan marriage records, written between 1617 and 1619. The records are provided by the 'Esposalles' database (Romero et al., 2013: 1661–62). Approach: The authors propose two approaches for the information extraction task (Toledo et al., 2019: 30–31). Both approaches are based on a neural network method (Convolutional Neural Networks) that classifies word images into semantic categories (Toledo et al., 2016: 545–48). For the first approach this method is enhanced through a Bigram inspired language model (a collocation). The second variant makes use of a Bidirectional Long Short-Term Memory Network (BLSTM), where a Recurrent Neural Network captures a variable word sequence. For the experiment the authors used 125 pages of the Esposalles corpus, which contained 1221 marriage records. 968 of these records with 31,501 words were used for the training of the neural networks, and 253 with 8026 words were put aside for a test set. For the evaluation the accuracy score is based on the Character Error Rate (CER) of the automated transcription. Only if the semantic label was correctly identified the score is calculated by 1-CER (Toledo et al., 2019: 30–32). Results: The two approaches are evaluated in relation to two different tasks. In the first task a transcription plus the semantic category (surname, name, location, occupation, state and other) must be provided. The second tasks require additionally the relation between two the person entities (e.g. husband and wife). The Bigram approach achieved in the first task a score of 87.89 and in the second 79.68. The BLSTM method resulted in a score of 94.63 for the first task and 94.02 in the second task (Toledo et al., 2019: 33–34).

2.3 Performance of NER systems

What, then, is the ideal accuracy of an NER system and which accuracy rates are projects currently reporting? The inter-annotator agreement, that measures the consistency of human annotators, provides a target of what can be expected from automated annotation systems (Sperberg-McQueen, 2016: 386). While the appropriate inter-annotator agreement depends on the task and purpose (Artstein and Poesio, 2008: 576), human annotators within the cultural heritage domain reach scores of up to 95% and more (e.g. McDonough and Camp, van de, 2017: 5–6; Erdmann et al., 2016: 87). An accurate comparison of this with NER approaches is difficult due to the heterogeneity of historical resources of the early-modern period, the obstacles for applying NER to such documents, and also the different measurements that projects use to report on the effectiveness of their methods. However, if we stay with the two most recent projects that reported in the most comparable way (Won et al., 2018; McDonough et al., 2019) it becomes evident that NER of early modern material has not reached this level of accuracy. In the best cases we encounter F1-scores of around 70% (Won et al., 2018: 8–9). An accuracy of this rate makes significant post-processing efforts unavoidable. However, only few projects in our survey reported on the necessity of cleaning NER output (Gregory and Hardie, 2011: 105; Wettlaufer et al., 2015: 190). The amount of labour that went into data pre-processing, modelling and evaluation efforts also remains unclear; an issue that is prevalent in the whole area of AI development (Dyer-Witheford et al., 2019: 75–79). As long as the human labour required for pre- and post-processing efforts is invisible, it remains unclear at which point NER can be applied in a way that reduces human effort. While NER might be leveraged at some point to abandon repetitive mark-up tasks, it will not make human expertise superfluous in the near future.

The difficulties of applying NER to historical documents can go beyond the limits of accuracy measures like those above. What ultimately constitutes an entity or not cannot always be reduced to a binary yes or no. Sluijter (2016) reports, for instance, on complex date entities like “the resolution taken yesterday”. Ontological questions also arose during our attempts to model objects in Sloane's catalogues. One of the insights was that the ‘boundary’ of an entity is constituted by the domain knowledge of the data modeller and the potential use of the extracted data. Our present-day failure to model certain entities ultimately permitted insight into the early-modern construction of the catalogues and, ultimately, their processes of knowledge production (Ortolja-Baird et al., 2019: 21–28). Yet, because the definition of a Named Entity varies from corpora, project and even NER tool, there is currently no way to make absolute claims about the general performance of NER tools (Marrero et al., 2013: 484–86). The honest answer to “what is the state of the art of NER applied to early modern documents?” is that we only know how particular NER tools perform on designated

corpora. The transferability of individual NER project outcomes and findings remain limited. The success of such approaches may increase as methodological practices for data transformation in historical research are articulated and formalised (Hoekstra and Koolen, 2019: 80; 92–93).

3 The landscape of early-modern authority files

Early modern documents teem with occurrences of named entities. To automate the significant task of segmenting, disambiguating and annotating legions of entities, NER systems are dependent on external vocabulary resources such as lexicons (Bernhard et al., 2018: 35), gazetteers (lookup-lists for geographical entities), name authority lists or training data. Given our conclusions above about the ongoing reliance on rule-based NER among digital early-modern projects, in this section we turn our attention to the current landscape of authority files.

A Name Authority File or List is an authoritative compilation of named entities. Authority files usually list canonical forms of named entities, assign or associate them with a unique identifier and map between canonical and variant names. So, authority files can be used to enrich early modern NER and provide reliable and persistent access to bibliographic and other supplementary resources (The Library of Congress, n.d.). Aside from incompleteness, other known inadequacies of Name Authority files for the early modern period include the hegemonic and colonial world views they can encode, for example, through the overwriting of indigenous place names with colonial placenames, usually in the English language (MacDonough et al. 2019: 2504).

Below we set out the main extant categories of authority file, and give examples of authority files that are representative of the respective categories. To do so, we draw especially on our experiences with authority files during the Enlightenment Architectures project. We ask: What are the main categories of authority files that may be used for early modern NER? What is the content of those authority files? What information about how Name Authority Files are compiled, and about the circumstances of their elaboration, is packaged with the files? Questions of source, context, voice and representation are often asked of early-modern documents. As has recently been shown, digital tools and resources are neither neutral nor impervious to the particularities of the contexts in which they were made (see, for example, Fyfe 2016), and it is appropriate similar questions of them too.

3.1 General purpose name authority files

VIAF

The Virtual International Authority File (VIAF), which is hosted by Online Computer Library Center, Inc (OCLC), aggregates name authority information from national library, archives and museum catalogues, and union catalogues, across c.20 countries (http://viaf.org/). Despite this

apparent abundance, it does not, however, include much British data except from the British Library (BL). Moreover, VIAF doesn't visually distinguish BL data with flag or icons as it does others. Given that the respective authority files have mostly been created by libraries, archives and museums, the individuals included in them are usually those who published outputs in forms recognised by those institutions. Additional data is also included from institutions outside of the cultural heritage sector proper, such as that created or derived by Wikidata. As such, "VIAF is positioning itself at the crossroad of the library data and the broad cultural heritage field, with a special value in increasing the visibility of authority data in the long tail of the Web" (Angjeli et al. 2014). Positioned as being of general applicability, from the homepage of VIAF, at the time of writing, no summary statistics are provided about the temporal span, scope or extent of the aggregated collection, such as would aid a researcher interested in establishing the likely applicability of the authority data to their specialist area. Rather, one must navigate to sub-pages of the main site that have been populated by contributors, though the information given on those sub-pages is sometimes quite out of date, as is the case with the information provided from the Library of Congress, for example (see: http://viaf.org/viaf/partnerpages/LC.html)

Library of Congress Names (http://id.loc.gov/authorities/names.html)

Known as LC Names or NACO Authority File, this covers "persons, organizations, events, places, and titles" (http://id.loc.gov/authorities/names.html). Again, given the origins of the list in the work of the Library of Congress, it is overwhelmingly authors who are listed. Similar to VIAF (with which it is aggregated), limited information about the shape and history of the list, or its applicability to different subject areas can be discovered on the landing page of the resource. There it is simply observed that: "Library of Congress Names includes over 8 million descriptions created over many decades and according to different cataloging policies".

Geonames

Covers geographical names in particular, and at the time of writing, comprises eleven million place names from 379 datasources (https://www.geonames.org/datasources/). Data sources include government departments, like the City of Austin; open platforms like the Humanitarian Data Exchange; and projects like that entitled 'zetashapes[:] crodsourced [sic] shapes, base tiger and flickr', whose given url (http://zetashapes.com) resolves to an advertisement for a domain that is currently available for purchase. It is noted in the relevant files that the data it supplies is "provided "as is" without warranty or any representation of accuracy, timeliness or completeness" (http://download.geonames.org/export/dump/readme.txt). Again, the researcher who seeks summary

information about the extent of the dataset will not find it on the landing page but is required to review ancillary data sources in search of this. The limitations of Geonames for historical data have been noted: “It does not have the ability to deal with location boundaries changing over time, changing hierarchy (for example places that changed national affiliation as country boundaries moved) or granularity, in terms of houses, properties or streets that no longer exist” (Callaghan 2018)

3.2 Specialist name authority files

Getty Thesaurus of Geographic Names

This is a specialist list of place names aimed especially at those working in art history and cognate disciplines. Historical place names are a particular focus of the list. Many place names are accompanied by approximate coordinates. Other kinds of information typically included is: “multilingual, multicultural, historical, archival, inscribed, and other types of names and information for a place; dates for names, dates for relationships, and dates for place types may be included”. Quite comprehensive information about the scope and origins of the list are included on subpages which are easily discoverable by researchers (see https://www.getty.edu/research/tools/vocabularies/tgn/about.html)

Pleiades

This historical information resource on the ancient world is aimed at scholars, students and interested parties. At the time of writing it lists 37,108 Ancient places, 32,947 Ancient Names and 40,153 Ancient Locations. It is noted that “Pleiades has extensive coverage for the Greek and Roman world, and is expanding into Ancient Near Eastern, Byzantine, Celtic, and Early Medieval geography” (https://pleiades.stoa.org/home). Though somewhat dispersed, a range of documents and reports are presented on project pages which can be reviewed for summary statistics and information about the scope of the list.

Other resources that are particularly rich for the ancient world include Pelagios Commons, which is a “community & infrastructure for Linked Open Geodata in the Humanities” (https://pelagios.org/). Other sources that are likely to be relevant to early modern digital research projects include the Historical Gazetteer of England’s Place Names, however the geographically-limited range of this can be immediately surmised from the title and so it is likely that this resource would be used in combination with others.

3.2 Semi-public

The British Museum's authority lists were developed in-house as no existing terminology resources covered the museum's requirements for chronological, geographical and cultural breadth. Both the Person and Place Authorities were generated in 2000, from terms used within the museum's database at that time. They have since been continually improved through the addition of terms as required for cataloguing, and the enhancement of records with additional information (e.g. dates, profession, biography, alternative names), though levels of detail remain variable and work is ongoing. The head of cataloguing in the British Museum has written:

> The wide range of British Museum terminologies include relatively stable drop-down lists, polyhierarchical thesauri, and sophisticated authorities, most especially the Biographical Authority for recording the names of people and institutions. Candidate terms can be created by the users, and are then vetted to ascertain whether and how they should be incorporated (Szrajber n.D.).

The authorities do not aim to be comprehensive, but to meet the needs of the British Museum's collection. The Person Authority currently includes the names of 223,217 people and organisations associated with the British Museum collection (e.g. producers or artists, collectors, subjects). The Place Authority currently includes 49,823 terms, organized hierarchically. While developed as an internal resource, the terms are viewable online via the British Museum Collection Online web pages. The authorities cannot at present be downloaded directly but are made available for the use of other institutions on request.[8]

3.3 Private

Individual humanities scholars are known to draw up their own personal authority lists to aid them in their research. Though not necessarily prepared for wider dissemination, and usually discoverable through personal networks and gift economies, such lists can include a wealth of information that can be leveraged by projects. Examples used by the Enlightenment Architectures project in a private capacity include a list of names occurring in Sloane's catalogue of Miscellanies and Antiquities compiled by a volunteer, a list of obscure people mentioned by Fellows of the Royal Society compiled by a researcher while going through correspondence and the authority lists created for the Sloane Letters project and for the Royal Society digitisation of their early Minute books.

4 New case study: Enlightenment Architectures

The Enlightenment Architectures: Sir Hans Sloane's Catalogues of his Collections (2016-21), is a Leverhulme funded research project involving <redacted for anonymity/>. One of the most voracious collectors and cataloguers of the early modern period, Sir Hans Sloane's (1660-1753) original collection is now mostly dispersed or lost (Caygill, 2012: 120). To better study the richness of information held in his catalogues, the project sought to annotate the entities contained in his catalogues.

Approach: The NER task focused on the automatic detection of person and place name instances in the transcribed body of the manuscripts and their consequent annotation (tagging) with the corresponding TEI tags for person and place respectively, including, when appropriate, specialised attributes to enrich their meaning. The inherent ambiguities of the Sloane's writings with regard to language, and other issues, meant that absolute accuracy of NER could not be expected. Challenges originated, for example, from the mixed language of the textual body (English and Latin), the breviloquent style of catalogue entries, replete with abbreviations, and the presence of morpho-syntactical markers that arose during the double keying of the documents, such as sentence splitting hyphenation and extra spacing between letters which increased variation. Most importantly, the presence of existing TEI tags in the transcribed manuscript text preceding the NER automation increased the technical complexity of task, which had to provide a mechanism for by-passing existing tags from the matching process. The steps taken towards easing this particular complication are discussed below. The combination of an automated approach with a meticulous manual TEI tagging process the text ensured that any imperfections introduced by automation could be refined by the human annotators.

4.1 The NER method

A rule-based information extraction pipeline was constructed for detecting person and placename mentions. The lack of a training corpus to support a supervised machine learning method was decisive for the adoption of a rule-based approach, which was benefited by the availability of the extensive domain knowledge contained in the British Museum's name authority files. The rule-based pipeline employed a range of domain-independent modules, such as tokenizer, sentence splitter and Part-of-Speech Tagger (PoS) and incorporated the domain vocabulary originating from the authority files, which was converted into parametrised lists (Gazeteers) that equipped the hand-crafted rules with a matching dossier.

The pipeline was deployed in the General Architecture for Text Engineering (GATE), a popular and mature open source JAVA-based suite of tools with over twenty years of continuous development from the Natural Language Processing research group at the University of Sheffield (Cunningham et al., 2013). The GATE platform has been employed for the construction of NER pipelines in the broader digital humanities domain, including processing of 18th century court proceedings (Bontcheva et al. 2002), historical newspapers (Allen et al. 2007), and archaeological grey-literature report (Vlachidis & Tudhope, 2016).

The NER process was arranged in a cascading order a) non-domain specific tools (NLP modules), b) input from domain vocabulary and 3) bespoke handcrafted rules that utilised input from the cascading order to construct pattern-matching rules expressed in JAPE[9] grammars. In detail, the pipeline incorporated a tokenizer, a PoS, a GATE Gazetteer, a set of rules for matching entities of interest, a set of rules for applying specialised attributes to entities and a final matching validation phase, as seen in Figure 1. The Sloane manuscripts are written in Latin and English (and other languages too) but the focus of the NER remained in English, hence, the part-of-speech phase did not engage a Latin-based input.

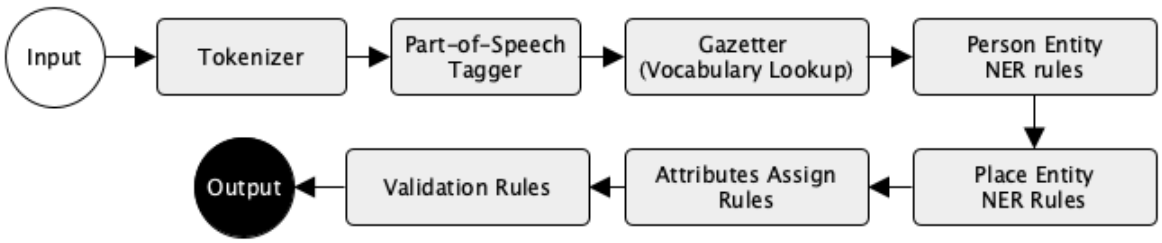

Figure 1 The NER pipeline modules in a cascading order.

The authority lists contain a vast base of place name and person name instances but have not been originally designed to support automated Natural Language Processing (NLP) tasks. Therefore, it was necessary to convert the original files to parametrised listings capable of supporting the NLP aims of the pipeline. The authority lists were converted to CSV format and contained approximately 50,000 place name entries, 215,000 person name entries and a range of data attributes. They included 80 distinct attribute types for person entries and 55 distinct attributes for place entries, varying from unique identifiers to notes, modification dates and classification values. The conversion focused on the field of actual name values normalising the entries for optimum NLP use, addressing the following issues: a) non-atomic values such "George John James Gordon, 5th

[9] https://gate.ac.uk/sale/thakker-jape-tutorial/GATE JAPE manual.pdf

Earl of Aberdeen", b) de-coupling of honorifics such as "Sir Ralph Abercromby", c) re-arrangement of first name – surname order such as "Baehr, Johann" and d) removing non-person-name entries such as "Aboriginal Children's Advancement Society" (organisation and institution names were included in the original lists). The conversion aimed at compensating the lists with entries of person names in the form of name-surname (e.g "Johann Baehr"), which is more likely to appear in the manuscripts. In the case of honorifics (e.g, Ms, Mr, Sir, etc.), specialised rules were constructed to allow their optional and complementary matching to person name instances. Another issue related to the appearance of multiple unique identifiers in the vocabulary under a single person name. It was not clear whether the separate identifiers reflected cases of synonymy or legacy attribution, and the decision was made to maintain all identifiers in the parametrised listings. In contrast, the place name list contained clear atomic values and did not present any normalisation issues. The list held a vast number of entries, ranging from city, county and country names to more specific geographic locations such as rivers, valleys and monument places (e.g. bridges and buildings). The decision was made to use the list in its entirety, as specialised and rare geographical entries and monuments are assumed to have no influence on the precision performance of the pipeline and potentially can benefit recall rates.

The pipeline was configured to execute the recognition process into two separate modes, addressing the specific extraction requirements of the manuscripts. Prior to the automatic process, all manuscripts had been manually annotated (tagged) with TEI elements for line break, label, division, page break and underlined text etc. A subset of the manuscripts contained, in addition to the TEI tags, manually applied annotations of place names. The first mode of the pipeline aimed at automatically recognising place name and person name instances and assigning the respective "place" and "person" tags, including the appropriate unique identifiers originating from the authority lists. In respect to vocabulary entries that declared more than a single unique reference value, the pipeline engaged a subsequent stage for assigning the multiple references. The unique references for each vocabulary entry were concatenated into a single attribute value using the tilde (~) symbol, as seen in Figure 2.

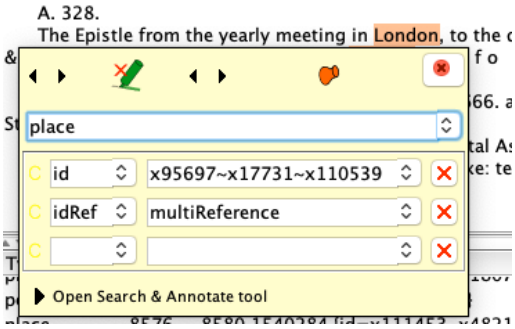

Figure 2 Multi-refence Assignment for instance "London"

The second mode of the pipeline aimed at assigning latitude and longitude values to place name instances that have already been manually annotated, prior to the automated NER phase. A set of dedicated rules matched the pipeline vocabulary and the existing place name instances, assigning a comma concatenated value to the latitude and longitude attribute (lat_long). The unique vocabulary reference was assigned to the dedicated attribute (viaf), as seen in Figure 3. A final pipeline stage common to both modes produced the output of the manuscripts, compiling the pre-existing manually defined TEI tags and the auto-generated annotations into a single XML (TEI compliant) structure.

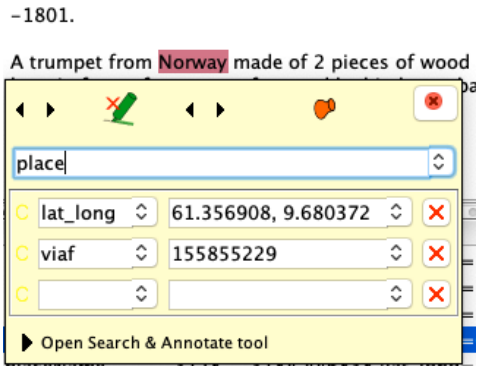

Figure 3 Latitude and longitude values and unique vocabulary reference for the instance "Norway"

4.2 Results of the NER Process

The output files containing both the manually produced TEI tags and the auto-generated NER annotations were delivered to cataloguing experts. The experts were asked to visually inspect the manuscripts and to assess whether the pre-existing TEI mark-up had been damaged in the process. In the absence of a gold-standard which could have been used to evaluate the performance of the NER process, the experts were asked to judge the accuracy of the automatically produced tags from a qualitative (visual-inspection) prospective. The process did not affect the original TEI mark-up which remained valid and intact. The overall outcome of the automated process was judged as reasonably satisfactory, but a number of cases of inaccurate or partially correct annotations were

identified that suggested immediate improvements of the extraction rules. The cases highlighted the following areas of improvement:

The original rules included matching of honorifics (e.g. Sir, Dr, Mr, Mrs, etc) for complementing person name annotations, but the results revealed missed instances that carried extra spacing within the letters of a honorific. The extra space was a result of the transcription process, which delivered instances in the form of "M r" instead of "Mr". The rules were refined to address the extra spacing and also were expanded to include additional types and abbreviations such as "Captain, Capt.,", not previously addressed by the rules.

The performance of some rules was also affected by the appearance of the pre-existing TEI tags, intercepting whole matching of person and place name instances. For example, the pre-existing tagging of the instance "Basing Castle" which included the TEI element <lb>. Ideally, the rule should have delivered an annotation spanning throughout the length of the place name (Table 1).

Table 1 Problematic NER annotation of a place name due to pre-existing TEI tags

Pre-existing TEI	<p>M Peter Kolben Reise an das Cape<lb></lb> du bonne Esperance
NER output	<p>M Peter Kolben Reise an das Cape<lb></lb> du bonne <place id="x69751">Esperance</place>.
Correct NER output	<place id="x69751">Cape<lb></lb> du bonne Esperance</place>

Addressing the above issue is particularly challenging because vocabulary matching can be configured against whole vocabulary entries,[10] or against partial matching. The latter configuration can result in a vast number of matches, introducing a significant amount of noise. The decision was made to refine such cases manually during a succeeding TEI tagging phase of the project.

Another issue was related to the well-known problem of polysemy that affects NER performance against place name and person name instances. The cascading order of the rules is important for addressing the polysemy behaviour, giving priority to person instances by incorporating contextual evidence with the rules, for example, by detecting a first name instance before a polysemous surname. However, the TEI tags undermined the cascading arrangement and delivered false annotations. For example, in the case of "Lorenzo de Mendoza" the pipeline delivered the result Lorenzo<lb></lb> de <place id="x67376">Mendoza</place>, ignoring the name instance as a

[10] Note that "Esperance" existed as a single vocabulary entry and delivered by the whole match configuration, not as a partial match of the "Cape du bonne Esperance" entry.

whole and delivering a partial place name annotation. Another case concerned “China”, which was used in the context of porcelain material and not in the sense of a place name, which generated a large number of false positive matches. The decision was made to silence the particular vocabulary instance from matching due to the limited contextual evidence available in the manuscripts that could have been used to enable the definition of word-sense disambiguation rules. The vocabulary was also enhanced with Latin versions of commonly occurring place names, such as “Hamburgi”, “Antwerpiæ” and “Londini” to improve the place name matching performance. The final validation stage was also modified to remove erroneous matches by filtering out any instances that did not qualify as being nouns or proper nouns and had a length shorter than 4 characters long.

The introduction of an automated NER process to support the TEI annotation of Sloane’s catalogues brought reasonable benefits. It helped to accelerate the process of manual tagging by delivering a large amount of place name and person name annotations equipped with unique references. Despite false positives, the process enabled human annotators to review instances that could potentially have been missed out due to the extended length of the manuscripts. Most importantly, it lifted the weight of assigning unique references to tags, which is a tedious and error prone task when done manually. In the absence of a training corpus, the rule-based approaches were capable of delivering reasonably satisfactory results when supported by an extensive domain vocabulary. However, significant cleansing and adaptation of this vocabulary for the NLP task is required for increasing the accuracy of the rules. The most important lesson learnt during the engagement of the automated phase concerned the methodological order in which such automated tasks should be conducted. We learned that automated tagging processes should precede the human-led TEI annotation phase. This allows automations to pave their way in context, undisturbed by pre-existing tags and increases the chances of refining any erroneous annotations during a full-scale TEI manual tagging phase.

5 Synthesis and discussion: Rule or pattern-based approaches to NER

The picture that emerges from the review above is of the proclivity of historical, information studies and digital humanities researchers towards rule or pattern-based approaches to NER. Automatic annotation of place names and person names is at the core of many projects focused on the study of early-modern manuscripts; there is apparently no limitation to the kind of textual dataset to which rule-based NER can be applied. Many of the projects reviewed above focused on identifying place names, followed by person entities. Only Johann Friedrich Blumenbach – Online moved beyond the automated mark-up of place- and person names to domain-specific terms.

Despite their common focus on entities, however, we have seen that each project deals with characteristics that are special to their particular domain. Examples of general-purpose tools that

have been adapted for historical research include GATE (as discussed in the Enlightenment Architectures case study above), CLAWS and USAS. Yet, 'out-of-the-box' NER tools do not tend to accommodate the needs of specialized disciplines without significant customisation, and there is a demand for tools that are easily adoptable for specific domains (Alex et al., 2015: 32–33).

The projects surveyed above are largely pursuing bespoke solutions to the difficulties of implementing rule- or pattern-based NER. Solutions include the enrichment or adaptation of knowledge-bases, including general purpose ones, algorithmic interventions to modify NER platforms like GATE, and the manual annotation of entities. The transferability and generalizable qualities of these approaches is often limited. It is not only that methods are tailored to a particular dataset, it is also that the detail of these solutions is not necessarily made publicly available in publications or other venues. This may reflect the disciplinary culture of the Humanities, which has traditionally pursued situated research questions, using a range of different approaches, and often circulated findings in the form of monographs (Whitley 2000). The pursuit, and reporting of technical work in the Humanities is not always valued. Scholars and projects have been known, for some time, to underreport the more technical aspects of their projects so as to secure publication in more traditional Humanities publications. An example of this is seen in the comments of Goldfield on the decisions that fellow scholars had made to relegate the detail of their computational techniques and data in formal publications:

> At this juncture I therefore accept Paul Fortier's politically wise approach in his study on Gide's *L'immoraliste*: statistical sophistication in stylometric and thematic analysis, as well as statistical details implicit in the interpretation, are relegated to appendices or simply not included in the publication (Goldfield 1993, 370).[11]

While the Humanities may not necessary count reproducibility among their chief concerns, within the context of NER this approach is potentially detrimental to research trajectories.

This tendency may also reflect a certain "technological optimism", which ranks questions about sustainability as secondary, or lower (Smithies et al., 2019). Project's bespoke modifications may be hampered by the problems of sustainability due to the lack of long-term funding for most digital projects. NER findings, moreover, are published in a multiplicity of venues, and so the levels of communication ongoing between projects, and the ability of individuals to establish a synthesis of the field is impaired. The application of automated NER to early-modern texts is not only a technical matter, issues of disciplinary culture and the fragmentation of knowledge are apparently

[11] Joel D. Goldfield, 'An Argument for Single-Author and Similar Studies Using Quantitative Methods: Is There Safety in Numbers?', *Computers and the Humanities*, 27.5–6 (1993), 365–74 (p. 370) <https://doi.org/10.1007/BF01829387>.

also in play.

As the next section shows, although rule-based systems are gradually being superseded by machine learning techniques within the field of computer science, this approach is still not being commonly used by digital humanities and digital cultural heritage projects. The main reason for this is that supervised machine learning approaches are dependent on large corpora of annotated data, which at can only be created through human labour (Yu, 2017; Gray and Suri, 2017). The advances of unsupervised methods can overcome some of the hindrances on the development of training set but this promise is still unexplored in the humanities domain. Such methods require large volumes of data for their effective (unsupervised) training which might not always be available within the historical (early modern) corpora. Section 2.3 shows that such approaches were only recently pursued with a reasonable amount of training data.

5.1 Supervised NER techniques: summary and synthesis

From our second review it becomes evident that while rule-based NER systems still dominate, projects devoted to the early-modern period are experimenting and developing NER techniques that are based on supervised learning. The number of projects is too scarce to draw firm conclusions, yet three trends clearly emerge even from this limited sample. First, where projects have developed their own NER tools there are no detailed accounts about how the tools worked or where they might be found on the web. Second is the use of general-purpose tools, or tools that come from domains other than that of historical research (Sluijter 2016 and Won et al. 2018), that are then applied to the cultural heritage domain. Third, the training data for the models is in most cases based on military messages (MUC), news corpora (e.g.: ConLL, OntoNotes), or web resources such as Wikipedia (see: Grishman and Sundheim, 1996: 466; Explosion AI, 2020; Al-Rfou, 2015; Stanford NLP Group, [n.d.]). The results from the experiments by Toledo et al. (2019) on the 'Esposalles' database show interesting results when a NER tool is trained on a specific corpus. However, it must be kept in mind that marriage records are a highly structured documents. Only future research will reveal the capability of machine learning NER approaches on less structured and more heterogeneous early modern corpora, where we encounter for instance letters of different authors in various languages.

5.2 Name Authority Lists

A number of general-purpose and specialist named entity authority files exist and many are available under open licences that support reuse. The authority files surveyed above have been devised or aggregated by a range of actors, from heritage organisations to individual scholars.

Whether created in public, semi-public or private contexts, authority files can be difficult to discover. There is no central registry of authority files that can be consulted to identify resources especially applicable to early-modern texts. Researchers apparently encounter such lists through idiosyncratic combinations of specialist domain knowledge, literature reviews that include environmental scans of available resources, and through intrapersonal contacts.

While some geographical locations and historical epochs, like the ancient world, are quite well covered by extant files, this is less true for the early-modern period. Moreover, the use of general-purpose authority files can hold the potential of being inaccurate and at worst insensitive to past and present local languages and communities (McDonough et al., 2019: 6–7). It may be desirable to combine multiple authority lists for early-modern NER, yet the ways that information is formalised in authority lists can serve to constrain and complicate the recombination of individual lists.

Our review echoes and reinforces Nelson's and McDonough et al.'s identification of a dearth of specialist authority files for the early-modern period (Nelson, 2014: 8–11; McDonough et al., 2019: 20–21). Our review also raises questions about the level of critical interrogation that the supplementary documentation provided with name authority lists currently does or does not promote. It is not only that critical frameworks that could guide the appropriate interweaving of name authority lists are currently underdeveloped, as we have seen, information that is relevant to researchers who wish to interrogate the commensurability of lists is often difficult to uncover. In many cases above, core information about, for example, the selection rationale that underpinned the lists is not readily available from the landing pages of the resources we consulted and neither is it bundled with the lists.

It seems that there has been a tendency to view name authority lists as merely a tool for allowing documentary sources to be wrangled. Yet authority files can be conceptualised not only as aids to study but also as proper objects of study themselves. Dictionary research has served to identify the many ways that dictionaries are implicitly and explicitly interwoven with one another, and the languages and societies that they sought to reflect and sometimes influence (Considine 2011). Their close relatives, authority lists, many also, then, be understood as documents that record preserve, and potentially canonise decisions about which actors were or were not worthy of attribution. At stake, then, going forward, should not only be the role of technology in structuring authority lists and making them interoperable, but also the qualitative decisions that underpinned their elaboration and how those decisions are recorded, and made accessible for future users, both human and machine.

6 Conclusion and recommendations

By mapping-out the capabilities, challenges and limitations of NER, this article has aimed to synthesise the state of the art of NER in the context of early-modern research resources and to inform discussions about the kind of resources, methods and directions that may be pursued to enrich the application of the technique going forward. Moreover, we have drawn attention to the situated nature of authority files, and current conceptualisations of NER, leading us to the conclusion that more robust reporting of NER approaches and findings are required.

Currently it is not possible to benchmark the capabilities of NER applied to documents of the early modern period. Thus, a forum is needed where tools are evaluated according to community standards. Within the wider NER community, the MUC and ConLL corpora are used for such experimental set-ups and are accompanied by a conference series. The ultimate nature of such a forum must be discussed with the whole research community of the early modern domain. However, there are according to Marrero et al. certain requirements that should be met. They are:

- Content validity: the experimental set-up reflects community needs. People who work with documents of the early-modern period decide on the desired goals of the forum, the entities of interest, their semantics and documents used for the experiments.
- External validity: ideally the corpus is heterogeneous and contains a mix of various sub-genres of early-modern archival documents (e.g. letters, court proceedings, encyclopedia, library and museum catalogues). Results are then more likely to be generalizable to the whole domain.
- Convergent validity: precision, Recall, and the F1-score are established measurements to report on the effectiveness of NER systems and make results comparable. The inter-annotator agreement stipulates a desired level of NER accuracy for early-modern documents
- Conclusion validity: the criteria used to evaluate NER for the early-modern period must be understandable, consistent and in alignment to evaluation standards established within the wider field of NER development. This allows a judgement on the validity of findings (2013: 486–87).

The authority files and documents created by the projects we discussed in this review could present a starting point for such a shared experimental forum.

Yet, as we have argued above, the issue is not technical only. As numerous studies have shown, “automation has an ideological function as well as a technological dimension” (Taylor, 2018). The creation, use and promotion of algorithmic technologies like NER is thus not a neutral process, and neither is their output (Risam, 2018: 123–32). As this article has attested, NER is an algorithmic intervention that transforms data according to certain rules-, patterns- or training data and ultimately

affects how we interpret the results (see: Hoekstra and Koolen, 2019: 381). A digital tool criticism, as proposed by Koolen et al., aims to support detailed accounts about why a certain tool was selected, who made it, and about its limitations (Koolen et al., 2019: 382–83). The application of frameworks like Koolen et al's could foster a more critical understanding of the role and impact of NER on early-modern documents and research and focalize some of the data- and human-centric aspects of NER routines that are currently overlooked. The utilization of such a framework would call for the provenance of sources, and authority files, that come into shared corpora to be made explicit, together with their limitations and biases. An account of the labour that is involved in data processing, modelling and evaluation efforts should also be necessary. Alongside established measurements above, NER should also be evaluated through cost measures, that is how much additional human labour is required to reach adequate results (Marrero et al., 2013: 488). In short, the imbrication of frameworks like that of Koolen et al. could lead to more holistic, and transferable ways of articulating the uses of NER in early-modern research.

Next to this, ways of encouraging the Digital Humanities and related fields to continue to follow, and critique new technical developments in AI and machine learning may be useful. Emerging (or stalled) approaches may assist with automating named entity recognition e.g. information extraction with context-aware neural models in addition to more traditional approaches like gazetteers and name authority lists.

Finally, we will note ongoing work in the area of Digital Ethics:

> the level of detail that can be quickly gleaned about individuals from the past, particularly when multiple digital archives are accessed, raises ethical questions. For example, when reporting findings researchers could be disclosing personal information that is unknown to descendants, and if it relates to a sensitive topic then there is the potential for the researcher to cause distress (Crossen-White 2015:108).

NER researchers are also well positioned to contribute to, and learn from work in the area of Digital Ethics, that reflects on what may be gained, as well as lost, through the digitization and unlocking of historical documents that mention human subjects.

7. References

Aguilar, S. T., Tannier, X. and Chastang, P. (2016). Named entity recognition applied on a database of Medieval Latin charters. The case of chartae burgundiae. Proceedings of the 3rd HistoInformatics Workshop on Computational History (HistoInformatics 2016) Co-Located with Digital Humanities 2016 Conference (DH 2016). Krakow, pp. 67–71 http://ceur-ws.org/Vol-1632/paper_9.pdf (accessed 11 June 2020).

Alex, B., Byrne, K., Grover, C. and Tobin, R. (2015). Adapting the Edinburgh Geoparser for Historical Georeferencing. International Journal of Humanities and Arts Computing, 9(1): 15–35 doi:10.3366/ijhac.2015.0136.

Allen, R.B., Japzon, A., Achananuparp, P. and Lee, K.J. (2007). A framework for text processing and supporting access to collections of digitized historical newspapers. In Symposium on Human Interface and the Management of Information (pp. 235-244). Springer, Berlin, Heidelberg.

Al-Rfou, R. (2015). Named Entity Extraction: Languages Coverage — Polyglot 16.07.04 Documentation https://polyglot.readthedocs.io/en/latest/NamedEntityRecognition.html#languages-coverage (accessed 17 February 2020).

Angjeli, Anila, Andrew Mac Ewan, and Vincent Boulet. 2014. 'ISNI and VIAF – Transforming Ways of Trustfully Consolidating Identities'. In IFLA 2014 Lyon, 19. Lyon.

AOR ([n.d.]). Archaeology of Reading https://archaeologyofreading.org/ (accessed 12 February 2020).

Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. Computational Linguistics, 34(4): 555–96 doi:10.1162/coli.07-034-R2.

ASCH (2018). Developing interoperable metadata standards for contextualising heterogeneous objects, exemplified by objects of the provenance von Asch http://asch.wiki.gwdg.de/index.php/asch:About (accessed 12 February 2020).

Bernhard, D., Magistry, P., Ligozat, A.-L. and Rosset, S. (2018). Resources and Methods for the Automatic Recognition of Place Names in Alsatian. Proceedings of the Second Workshop on Corpus-Based Research in the Humanities. Vienna: Gerastree Proceedings, pp. 35–44 https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/CRH2.pdf (accessed 16 May 2019).

Bontcheva, K., Maynard, D., Cunningham, H. and Saggion, H. (2002). Using Human Language Technology for Automatic Annotation and Indexing of Digital Library Content. In Agosti, M. and Thanos, C. (eds), Research and Advanced Technology for Digital Libraries, vol. 2458. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 613–25 doi:10.1007/3-540-45747-X_46. http://link.springer.com/10.1007/3-540-45747-X_46 (accessed 13 June 2019).

Borin, L., Kokkinakis, D. and Olsson, L.-J. (2007). Naming the Past: Named Entity and Animacy Recognition in 19th Century Swedish Literature. Proceedings of LaTeCH. Prague, pp. 1–8.

Callaghan, Samantha. 2018. 'Workshop Reflections: Early Modern Collection Catalogues, British Museum'. Georgian Papers Programme. 26 March 2018. https://georgianpapers.com/2018/03/26/workshop-reflections-early-modern-collections-catalogues-british-museum/.

Caygill, M. (2012). Sloane's catalogues and the arrangement of his collections. From Books to Bezoars: Sir Hans Sloane and His Collections. London: British Library, pp. 120–36.

Considine, John. 2011. Dictionaries in Early Modern Europe: Lexicography and the Making of Heritage. Reissue. Cambridge University Press.

Crossen-White, Holly L. 2015. 'Using Digital Archives in Historical Research: What Are the Ethical Concerns for a "Forgotten" Individual?' Research Ethics 11 (2): 108–19. https://doi.org/10.1177/1747016115581724.

Cunningham, H., Tablan, V., Roberts, A., Bontcheva, K. (2013) Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoS Computational Biology 9(2): e1002854. doi:10.1371/journal.pcbi.1002854

DigitalArk Project ([n.d.]). DigitalArk Project Blog: Project Documentation DigitalArk Project Blog https://digitalarkproject.blogspot.com/p/guidelines-for-transcriptions.html (accessed 12 June 2020).

Dyer-Witheford, N., Kjøsen, A. M. and Steinhoff, J. (2019). Inhuman Power: Artificial Intelligence and the Future of Capitalism.

Ehrmann, M., Colavizza, G., Rochat, Y., Kaplan, F., Dipper, S., Neubarth, F. and Zinsmeister, H. (2016). Diachronic Evaluation of NER Systems on Old Newspapers. Proceedings of the 13th Conference on Natural Language Processing. Bochum: Bochumer Linguistische Arbeitsberichte, pp. 97–107 https://infoscience.epfl.ch/record/221391/files/13_konvensproc.pdf (accessed 14 June 2019).

Erdmann, A., Brown, C., Joseph, B., Janse, M., Ajaka, P., Elsner, M. and Marneffe, M.-C. de (2016). Challenges and Solutions for Latin Named Entity Recognition. Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH). Osaka, Japan: The COLING 2016 Organizing Committee, pp. 85–93

Explosion AI (2020). English: spaCy Models Documentation https://spacy.io/models/en (accessed 9 June 2020).

Fyfe, Paul. 2016. 'An Archaeology of Victorian Newspapers'. Victorian Periodicals Review 49 (4): 546–77. https://doi.org/10.1353/vpr.2016.0039.

Gelling, M. (2011). Place-Names and Archaeology. In Hinton, D. A., Crawford, S. and Hamerow, H. (eds), The Oxford Handbook of Anglo-Saxon Archaeology. Oxford University Press doi:10.1093/oxfordhb/9780199212149.013.0050. http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780199212149.001.0001/oxfordhb-9780199212149-e-50 (accessed 7 January 2020).

Goldfield, Joel D. (1993). 'An Argument for Single-Author and Similar Studies Using Quantitative Methods: Is There Safety in Numbers?', *Computers and the Humanities*, 27.5–6, 365–74. https://doi.org/10.1007/BF01829387.

Gray, M. L. and Suri, S. (2017). The Humans Working Behind the AI Curtain. Harvard Business Review https://hbr.org/2017/01/the-humans-working-behind-the-ai-curtain (accessed 6 June 2020).

Gregory, I. N. and Hardie, A. (2011). Visual GISting: bringing together corpus linguistics and Geographical Information Systems. Literary and Linguistic Computing, 26(3): 297–314 doi:10.1093/llc/fqr022.

Grishman, R. and Sundheim, B. (1996). Message Understanding Conference-6: a brief history. Proceedings of the 16th Conference on Computational Linguistics -, vol. 1. Copenhagen, Denmark: Association for Computational Linguistics, p. 466 doi:10.3115/992628.992709. http://portal.acm.org/citation.cfm?doid=992628.992709 (accessed 15 May 2019).

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S. and Ball, J. (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 368(1925): 3875–89 doi:10.1098/rsta.2010.0149.

Hitchcock, T., Shoemaker, R., Emsley, C., Howard, S. and McLaughlin, J. (2012). The Old Bailey Proceedings Online, 1674-1913 www.oldbaileyonline.org (accessed 16 January 2020).

Hoekstra, R. and Koolen, M. (2019). Data scopes for digital history research. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 52(2): 79–94 doi:10.1080/01615440.2018.1484676.

Kaplan, F. (2015). A Map for Big Data Research in Digital Humanities. Frontiers in Digital Humanities, 2 doi:10.3389/fdigh.2015.00001. http://www.frontiersin.org/Digital_Humanities/10.3389/fdigh.2015.00001/full (accessed 11 January 2020).

Kettunen, K. and Ruokolainen, T. (2017). Names, Right or Wrong: Named Entities in an OCRed Historical Finnish Newspaper Collection. Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2017. Göttingen, Germany: ACM Press, pp. 181–86 doi:10.1145/3078081.3078084. http://dl.acm.org/citation.cfm?doid=3078081.3078084 (accessed 11 June 2020).

Koolen, M., Gorp, J. van and Ossenbruggen, J. van (2019). Toward a model for digital tool criticism: Reflection as integrative practice. Digital Scholarship in the Humanities, 34(2): 368–85 doi:10.1093/llc/fqy048.

The Library of Congress. n.D. 'LC Linked Data Service: Authorities and Vocabularies (Library of Congress)'. Webpage. https://id.loc.gov/authorities/names.html.

Liu, A. (2018). Friending the Past: The Sense of History in the Digital Age. Chicago; London: The University of Chicago Press.

Mac Kim, S. and Cassidy, S. (2015). Finding Names in Trove: Named Entity Recognition for Australian Historical Newspapers http://www.aclweb.org/anthology/U15-1007 (accessed 31 May 2018).

Manning, C. D., Raghavan, P. and Schütze, H. (2008). Introduction to Information Retrieval. New York: Cambridge University Press.

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J. and Gómez-Berbís, J. M. (2013). Named Entity Recognition: Fallacies, challenges and opportunities. Computer Standards & Interfaces, 35(5): 482–89 doi:10.1016/j.csi.2012.09.004.

McDonough, K. and Camp, M. van de (2017). Mapping the Encyclopédie: Working Towards an Early Modern Digital Gazetteer. Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities - GeoHumanities'17. Redondo Beach, CA, USA: ACM Press, pp. 16–22 doi:10.1145/3149858.3149861. http://dl.acm.org/citation.cfm?doid=3149858.3149861 (accessed 24 January 2019).

McDonough, K., Moncla, L. and Camp, M. van de (2019). Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora. International Journal of Geographical Information Science: 1–25 doi:10.1080/13658816.2019.1620235.

Nadeau, D. and Sekine, S. (2009). A survey of named entity recognition and classification. Named Entities: Recognition, Classification and Use, vol. 19. Amsterdam: John Benjamins Publishing Company, pp. 3–27 https://benjamins.com/catalog/li.30.1.03nad (accessed 9 May 2019).

Nelson, B. (2014). From Index to Interoperability: The Desideratum of Authority Files in Large-Scale Digital Projects. Scholarly and Research Communication, 5(4) doi:10.22230/src.2014v5n4a192. http://src-online.ca/index.php/src/article/view/192 (accessed 21 February 2019).

Nevalainen, T. (2006). The Early Modern English period. An Introduction to Early Modern English. Edinburgh: Edinburgh University Press, pp. 1–11 www.jstor.org/stable/10.3366/j.ctt1g09z3p.5 (accessed 19 January 2020).

Ortolja-Baird, A., Pickering, V., Nyhan, J., Sloan, K. and Fleming, M. (2019). Digital Humanities in the Memory Institution: The Challenges of Encoding Sir Hans Sloane's Early Modern Catalogues of His Collections. Open Library of Humanities, 5(1): 44 doi:10.16995/olh.409.
Piotrowski, M. (2012). Natural Language Processing for Historical Texts. Morgan & Claypool. Vol. 17. (Synthesis Lectures on Human Language Technologies) http://www.morganclaypool.com/doi/abs/10.2200/S00436ED1V01Y201207HLT017 (accessed 4 June 2018).
Piskorski, J. and Yangarber, R. (2013). Information Extraction: Past, Present and Future. In Poibeau, T., Saggion, H., Piskorski, J. and Yangarber, R. (eds), Multi-Source, Multilingual Information Extraction and Summarization. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 23–49 doi:10.1007/978-3-642-28569-1_2. http://link.springer.com/10.1007/978-3-642-28569-1_2 (accessed 9 May 2019).

Prell, M. (2018). Frühneuzeitliche Briefe als Herausforderung automatisierter Handschriftenerkennung: Ein Transkribus-Projektbericht. doi:10.22032/dbt.34849. https://www.db-thueringen.de/servlets/MCRFileNodeServlet/dbt_derivate_00041045/Transkribusbericht_2018_06_02.pdf (accessed 27 March 2019).
Prescott, A. (2018). Searching for Dr. Johnson: The Digitisation of the Burney Newspaper Collection. In Brandtzæg, S. G., Goring, P. and Watson, C. (eds), Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century. BRILL, pp. 51–71 doi:10.1163/9789004362871_004. https://brill.com/view/book/edcoll/9789004362871/B9789004362871_006.xml (accessed 20 February 2020).

Ravenek, W., Heuvel, C. van den and Gerritsen, G. (2017). The ePistolarium: Origins and Techniques. In Utrecht University, NL and Odijk, J. (eds), CLARIN in the Low Countries. Ubiquity Press, pp. 317–23 doi:10.5334/bbi.26. https://www.ubiquitypress.com/site/chapters/10.5334/bbi.26/ (accessed 8 June 2018).
Risam, R. (2018). New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy. Evanston, Illinois: Northwestern University Press.
Rochat, Y., Ehrmann, M., Buntinx, V., Bornet, C. and Kaplan, F. (2016). Navigating through 200 years of historical newspapers. Bern https://infoscience.epfl.ch/record/218707/files/IPRES-2016.pdf (accessed 11 June 2020).

Romero, V., Fornés, A., Serrano, N., Sánchez, J. A., Toselli, A. H., Frinken, V., Vidal, E. and Lladós, J. (2013). The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition. Pattern Recognition, 46(6): 1658–69 doi:10.1016/j.patcog.2012.11.024.

Shoemaker, R. (2005). Digital London: Creating a searchable web of interlinked sources on eighteenth century London. (Ed.) Anderson, I. Program, 39(4): 297–311 doi:10.1108/00330330510627926.

Sluijter, R. G. H., Scherer, M., Derks, S., Nijenhuis, S., Ravenek, W. and Hoekstra, R. (2016). From Handwritten Text to Structured Data: Alternatives to Editing Large Archival Series. Digital Humanities 2016: Conference Abstracts. Krakow: Jagiellonian University & Pedagogical University, pp. 680–82 http://dh2016.adho.org/abstracts/36 (accessed 5 June 2018).

Smith, D. A. and Cordell, R. (2019). A Research Agenda for Historical and Multilingual Optical Character Recognition. Northeastern University https://ocr.northeastern.edu/report/ (accessed 10 March 2019).

Smithies, J., Westling, C., Sichani, A.-M., Mellen, P. and Ciula, A. (2019). Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King's Digital Lab. Digital Humanities Quarterly, 13(1) http://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html#d3876770e516 (accessed 16 February 2020).

Southall, H., Mostern, R. and Berman, M. L. (2011). On historical gazetteers. International Journal of Humanities and Arts Computing, 5(2): 127–45 doi:10.3366/ijhac.2011.0028.

Sperberg-McQueen, C. M. (2016). Classification and its Structures. In Schreibman, S., Siemens, R. G. and Unsworth, J. (eds), A New Companion to Digital Humanities. Chichester, West Sussex, UK: Wiley/Blackwell, pp. 377–93.

Sprugnoli, R. (2018). A Neural Approach to the Identification of Place Names in Historical Texts. Proceedings of the Fifth Italian Conference on Computational Linguistics. Torino, pp. 360–65 http://hdl.handle.net/10807/133038.

Stanford NLP Group ([n.d.]). Stanford NLP: About https://nlp.stanford.edu/software/CRF-NER.html#Citation (accessed 17 February 2020).

Stork, L., Weber, A., Gassó Miracle, E., Verbeek, F., Plaat, A., Herik, J. van den and Wolstencroft, K. (2018). Semantic annotation of natural history collections. Journal of Web Semantics doi: 10.1016/j.websem.2018.06.002. https://linkinghub.elsevier.com/retrieve/pii/S1570826818300283 (accessed 21 July 2018)

Tanya Szrajber (n.D.). The collection catalogue as the core of a modern Museum's purpose and activities.http://cidoc.mini.icom.museum/wp-content/uploads/sites/6/2018/12/Szrajber_CIDOC_2014.final_version.pdf

Tanner, S., Muñoz, T. and Hemy Ros, P. (2009). Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive. D-Lib Magazine, 15(7/8) http://www.dlib.org/dlib/july09/munoz/07munoz.html (accessed 11 January 2020).

Taylor, A. (2018). The Automation Charade. Logic (5) https://logicmag.io/failure/the-automation-charade/ (accessed 18 March 2020).

Thylstrup, N. B. (2018). The Politics of Mass Digitization. Cambridge, MA: The MIT Press.

Toledo, J. I., Carbonell, M., Fornés, A. and Lladós, J. (2019). Information extraction from historical handwritten document images with a context-aware neural model. Pattern Recognition, 86: 27–36 doi:https://doi.org/10.1016/j.patcog.2018.08.020.

Toledo, J. I., Sudholt, S., Fornés, A., Cucurull, J., Fink, G. A. and Lladós, J. (2016). Handwritten Word Image Categorization with Convolutional Neural Networks and Spatial Pyramid Pooling. In Robles-Kelly, A., Loog, M., Biggio, B., Escolano, F. and Wilson, R. (eds), Structural, Syntactic, and Statistical Pattern Recognition, vol. 10029. Cham: Springer International Publishing, pp. 543–52 doi:10.1007/978-3-319-49055-7_48. http://link.springer.com/10.1007/978-3-319-49055-7_48 (accessed 10 February 2020).

Vlachidis, A., Tudhope, D. (2016). A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain. Journal of the association for information science and technology, 67(5), pp.1138-1152.

Wettlaufer, J., Johnson, C., Scholz, M., Fichtner, M. and Thotempudi, S. G. (2015). Semantic Blumenbach: Exploration of Text–Object Relationships with Semantic Web Technology in the History of Science. Digital Scholarship in the Humanities, 30(Issue suppl 1): i187–i198 doi:10.1093/llc/fqv047.

Wettlaufer, J. and Thotempudi, S. G. (2013). Named Entity Recognition in historical texts from the natural history domain. Poster. Berlin doi:www.gcdh.de/index.php/download_file/view/194/538/454/.

https://web.archive.org/web/20160804110151/http://www.gcdh.de/files/2013/6429/9184/Wettlaufer_Thotempudi_2013_NER_final.pdf.
Whitley, Richard. 2000. The Intellectual and Social Organization of the Sciences. Second edition. Oxford England ; New York: Oxford University Press.
Won, M., Murrieta-Flores, P. and Martins, B. (2018). Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. Frontiers in Digital Humanities, 5 doi:10.3389/fdigh.2018.00002. http://journal.frontiersin.org/article/10.3389/fdigh.2018.00002/full (accessed 14 May 2018).
Yadav, V. and Bethard, S. (2019). A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. ArXiv:1910.11470 [Cs] http://arxiv.org/abs/1910.11470 (accessed 5 June 2020).
Yu, M. (2017). The Humans Behind Artificial Intelligence Synced https://syncedreview.com/2017/04/30/the-humans-behind-artificial-intelligence/ (accessed 6 June 2020).

8. Appendix

8.1 Project websites

Name	Link
BOPCRIS	https://web.archive.org/web/20110811020331/http://www.southampton.ac.uk/library/bopcris
Circulation of Knowledge and Learned Practices in the 17th-Century Dutch Republic	http://ckcc.huygens.knaw.nl/
Encyclopédie ou Dictionnaire raisonné des	http://kmcdono.com/enc/

sciences, des arts et des métiers par une Société de Gens de lettres (1751–1772)	
Esposalles' database	http://dag.cvc.uab.es/the-esposalles-database/
Johann Friedrich Blumenbach – Online	https://blumenbach-online.de/index-englisch.php
The Lancaster Newsbook Corpus	https://www.lancaster.ac.uk/fass/projects/newsbooks/default.htm
The Old Bailey online	https://www.oldbaileyonline.org/
Reassembling the Republic of Letters	http://www.republicofletters.net/
Resolutions of the Dutch States-General from 1576 to 1795	http://resources.huygens.knaw.nl/besluitenstatengeneraal1576-1630/index_html_en

8.2 List of NER and NLP tools

Name	Link
CLAWS	http://ucrel.lancs.ac.uk/claws/

Edinburgh Geoparser	https://www.ltg.ed.ac.uk/software/geoparser/
GATE	https://gate.ac.uk/
MorphAdorner	http://morphadorner.northwestern.edu/morphadorner/
NER-Tagger software package	https://github.com/glample/tagger
Perdido	http://erig.univ-pau.fr/PERDIDO/
Polyglot	https://polyglot.readthedocs.io/en/latest/
spaCy	https://spacy.io/
Stanford Named Entity Recognizer	https://nlp.stanford.edu/software/CRF-NER.html
TextCat	http://www.let.rug.nl/vannoord/TextCat/
USAS	http://ucrel.lancs.ac.uk/usas/
VARD	http://ucrel.lancs.ac.uk/vard/about/

8.3 List of authority files

Name	Scope	Link
Alexandria Digital Library Project Gazetteer	Online global placename dictionary. Exists now only on a development server, but research team can be contacted for use	https://www.library.ucsb.edu/map-imagery-lab/alexandria-digital-library-gazetteer
Biography Portal of the Netherlands	Prominent figures from	http://www.biografischportaal.nl/en/

	the Dutch History	
CERL	Authority file for names found in material printed before the middle of the nineteenth century	https://data.cerl.org/thesaurus/_search?lang=en
Compendium of office holders and civil servants 1428-1861	Compendium of office holders and civil servants 1428-1861 on the present-day Dutch territory	http://resources.huygens.knaw.nl/repertoriumambtsdragersambtenaren1428-1861/index_html_en
Early Modern Letters Online	Finding aid for early modern correspondence	http://emlo.bodleian.ox.ac.uk/
GeoCrossWalk	Succeeded in Digimap	https://digimap.edina.ac.uk/
GeoNames	Global	https://www.geonames.org/
Getty Thesaurus	Gazetteer developed	http://www.getty.edu/research/tools/vocabularies/tgn/index.html

of Geographic Names	by the Getty research institute	
GND	Authority file developed by the German National Library	https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html
Library of Congress Names		**http://id.loc.gov/authorities/names.html**
Pleiades	Ancient World	https://pleiades.stoa.org/
VIAF	International	http://viaf.org/
World Gazetteer	Global	https://www.arcgis.com/home/item.html?id=346ce13fa2d4468a9049f71bcc250f37