Unsilencing colonial archives via automated entity recognition

UNSILENCING COLONIAL ARCHIVES VIA AUTOMATED ENTITY RECOGNITION

Mrinalini Luthra, Konstantin Todorov, Charles Jeurgens and Giovanni Colavizza

Universiteit van Amsterdam, The Netherlands

ABSTRACT

Colonial archives are at the center of increased interest from a variety of perspectives, as they contain traces of historically marginalized people. Unfortunately, like most archives, they remain difficult to access due to significant persisting barriers. We focus here on one of them: the biases to be found in historical findings aids, such as indexes of person names, which remain in use to this day. In colonial archives, indexes can perpetuate silences by omitting to include mentions of historically marginalized persons. In order to overcome such limitations and pluralize the scope of existing finding aids, we propose using automated entity recognition. To this end, we contribute a fit-for-purpose annotation typology and apply it on the colonial archive of the Dutch East India Company (VOC). We release a corpus of nearly 70,000 annotations as a shared task, for which we provide baselines using state-of-the-art neural network models. Our work intends to stimulate further contributions in the direction of broadening access to (colonial) archives, integrating automation as a possible means to this end.

Keywords Archives · Natural Language Processing · Digital Humanities · Named Entity Recognition and Classification · Accessibility · Dutch East India Company

1 Introduction

While archives serve as important resources for researching the past, we must also recognize that they are flawed building blocks for understanding it. As Verne Harris aptly put it, archives only offer "a sliver of a sliver of a sliver" [Harris, 2002]. These snatches of the past usually hide more than contemporary researchers would like to see. The focus of this paper concerns records created in a colonial context, at a time when slavery was rampant. These records mirror the 17th and 18th century hierarchies in terms of religion, race, gender and class and echo the voices and views of the creators of the records, the colonial administrators, the traders, and slave owners. They were the ones who made observations and determined what was worth noting. They did this —how could it be otherwise— from a European point of view, in their own interest and not always with a thorough knowledge of the local situation, let alone of the indigenous or enslaved people they encountered. The archives that have been created in this way are therefore not so much a faithful representation of the reality of what went on in the colonial areas but a constructed and selective image of the social order [Spivak, 1985, Jeurgens and Karabinos, 2020].

In recent years, the colonial archive has attracted a great deal of scholarly interest; not only from the perspective of archives as a source for historical research but above all as an object of study in itself [Ballantyne, 2004, Burton et al., 2003, Hamilton et al., 2002, Bastian, 2003, Stoler, 2010, Lowry, 2017, Namhila, 2017]. This transformation of research interest has led to numerous publications in which power aspects of archiving and the archive are addressed [Schwartz and Cook, 2002, Jimerson, 2009, Stoler, 2010] and has resulted in a shift in perception whereby archives are no longer seen as neutral repositories of sources but as contested spaces of knowledge production. It has also led to more critical reflection on the role archivists play in how they describe and contextualize these records in terms of perpetuating or adjusting one-sided perspectives and power imbalances [Caswell and Cifor, 2016, Ghaddar and Caswell, 2019]. Scholars who take the archive as the object of research are not only interested in what archives contain, but as much in what archives do not reveal. After all, the archive "conceals, distorts and silences as much as it reveals" [Fuentes, 2016, 48]. David Thomas asserts that archival silences have now become a proper subject for research and are no longer

seen as just passive and annoying gaps in our knowledge [Thomas et al., 2017]. Archival silences are of many types. It was Michelle Rolph Trouillot's [Trouillot, 2015] groundbreaking book *Silencing the Past*, in which he dissected mechanisms of silencing in the production of history through an analysis of the Haitian revolution and the attention this revolution received in historiography, which gave an impulse to investigate the mechanisms of silencing in the archives. Trouillot distinguished four moments in which silences could enter the process of historical production: 1) the moment of fact creation (which refers to the recording of information: who decides what is recorded, which words are used etc.); 2) the moment of fact assembly (which refers to which recorded information is admitted to become part of the archives); 3) the moment of fact retrieval (which refers to what information is used from the archives to compile a narrative); and 4) the moment of what he called retrospective significance, or the making of history. What interests us most here is the role archivists and archival institutions have played in creating or perpetuating silences. Archivists not only determine to a significant extent which records are admitted in the archives, but also how retrieval of information from the documents is facilitated via archival descriptions, indexes and other technologies designed and used to expedite access to the records from the past. In this paper our focus is on the latter.

If we look at how archivists have facilitated access to the contents of the records, we may conclude that their work was always focused on a very formal way of describing archives. Not the content of documents, but the content of the archive in terms of document types and the function of these documents from the perspective of the formal archive creator were the points of reference for making descriptions. Archivists have for long presented this method as neutral and impartial. The dominant methods of description, Verne Harris and Wendy Duff argued, "facilitate the needs of certain types of users, but give short to others" [Duff and Harris, 2002, 279]. In some cases, archivists chose to improve access to the contents of the documents, for instance by generating indexes on the names mentioned in the documents. To make the labor-intensive indexing manageable, choices were made, for example by indexing only the names of certain categories of persons that appeared in the documents. Those choices mirror what was considered important and what was less important, thus reinforcing the mechanisms of privilege and marginalization by spotlighting certain groups and obscuring others. Only recently these problematic mechanisms have started to receive fundamental attention in the archival discipline [Yeo, 2017].

The application of machine learning methods to improve the means of accessing historical records is a growing trend [Colavizza et al., 2022]. A key task in information extraction from texts is that of *named entity recognition and classification*: the automatic "detection of named entities, i.e., elements in texts which act as a rigid designator for a referent, and their categorisation according to a set of predefined semantic categories" [Ehrmann et al., 2021]. Canonical examples include persons, places and organizations. Named entity tasks underpin the automatic creation and enrichment of information systems, allowing to search and browse an archive using as anchors the entities and relations among them established in the very contents of the records. What is more, even indirect (e.g., unnamed) references to entities can be detected, offering a means to surface previously ignored traces. This approach constitutes a radical expansion of the traditional means of access to archival records, as previously described: it allows to navigate the archives more fluidly, across individual records and groups [Ranade, 2016], all the while not discarding archival context to follow the allure of full-text search [Winters and Prescott, 2019]. *Content-based indexing offers new possibilities in surfacing and foregrounding mentions of people who are at present hidden in the documents and historical access tools.*

In this work, our goals are two-fold:

1. in general, to propose using automated entity recognition on historical archival records as a means to expand and pluralize access to archives;
2. specifically, to advance the application of entity recognition to colonial archives, where its use seems most urgent in order to complement critically lacking access tools.

Our contributions include a typology of entities and their attributes designed for automated entity recognition in colonial archives and further adapted to our case study: the Dutch East India Company testament records. We release a shared task in the form of a high-quality corpus of 68,429 annotations. Furthermore, we establish baselines relying on modern machine learning methods. Lastly, our work intends to call for further contributions in the direction of broadening access to colonial archives, taking the form of models, annotations, applications, and their assessment.

Ethical Considerations

Before turning to the work of archivists in their efforts to provide access to archives, we would like to briefly reflect on one of the most fundamental problems of colonial archives and archives of slavery: the deep and unbridgeable chasm between what these archives offer and the needs of contemporary users seeking traces of their history. In most cases, searching these records by the descendants of colonized and/or enslaved people in the hope of detecting some direct voices from their ancestors leads to disappointment, frustration, and anger after being confronted with the violent categorizations from the past in which enslaved people are described as saleable commodities which turned them into

"nonpersons" [Hartman, 2008, Patterson, 2018, Fuentes, 2016, Zijlstra, 2021]. There is no doubt that the colonial archive is inherently problematic for those who want to hear snatches of the voices of the enslaved and colonized population. The fact that the colonized and enslaved have left hardly any traces in the colonial archive produced by themselves, does not mean that they do not feature frequently in the archive. The archive is full of records about them. But even then, it is not easy to gain access to these indirect traces of their presence. Feminist historian Durba Ghosh called on historians to also pay attention to the history of the people who have been made nameless and blames them for not looking critically enough at different naming practices since "incompletely named and renamed subjects have histories that are waiting to be told" [Ghosh, 2004, 316]. She argues that the colonial archives are uneven in their erasure and although the renaming of enslaved people may have led to a sort of archival death, because it is not possible anymore to know where they came from, this does not release researchers from the obligation to ask how they may "decipher women from the archives when they are unnamed?" [Ghosh, 2004, 299]. She mentions several examples of individuals who can be followed in their life course including the changes in social status after they were given a Christian name. Historian Marjoleine Kars [2020] has shown in her recently published book on the massive slave revolt in the Dutch colony of Berbice (which just like the Haitian slave revolt, has received scant attention in historiography) how valuable the indirect traces of renamed enslaved can be. She used 900 testimonies of enslaved persons which were recorded in the interrogation reports after the revolt, which started in 1763 and lasted for one year, was put down. Kars argues that even though the testimonies have been given under pressure, they do provide information about the personal experiences of the enslaved.

2 Related Work

Our research ties in with two major debates and developments in the humanities: the use of digital methods and artificial intelligence to make extensive historical collections more accessible and, under the heading of 'decolonization of the archives', dismantling coloniality in archival infrastructures, for example by developing inclusive and people-oriented search infrastructures.

2.1 Named Entity Recognition in Historical Documents

The digitization and extraction of information from existing, historical indexing tools can serve to bootstrap the creation of searchable archival information systems [Colavizza et al., 2019, Koolen et al., 2020]. Nevertheless, it is by considering information extraction from the full contents of archival records that a broader, more systematic pluralization of access can be achieved. This is becoming a possibility largely thanks to growing efforts in digitization [Terras, 2011] and progress in the automatic extraction of text from handwritten records [Muehlberger et al., 2019]. *Named entity recognition and classification* (NER for short) constitutes a key step in information extraction pipelines [Ehrmann et al., 2021], and rests at the core of our approach here. In recent years, this task has seen rapid improvements thanks to neural networks [Lample et al., 2016] and pre-trained language models such as BERT [Devlin et al., 2019]. NER is also related, and typically followed up by the tasks of *disambiguation or linking* an entity mention to a knowledge base, and of *relation extraction* among named entities.

The task of named entity recognition and classification on historical documents poses a distinct set of challenges [Ehrmann et al., 2021]: document type and domain variety, dynamics of language, noisy input, and lack of resources. While common to the broader application of natural language processing to historical texts [Piotrowski, 2012], these challenges emerge in full force in the context of historical archives. The variety of record typologies (which entail different formats, layouts, and diplomatic characteristics) and the domains of their contents is broad, and only mitigated by a relative uniformity within homogeneous record groups. Language variety and the dynamic change of language use over time are also related issues. All of this has direct implications for the named entities to be extracted, and their typologies. Noise in the inputs is primarily, yet not exclusively, due to the preceding task of automatic text recognition. Such noise can have a significant impact on downstream tasks, including named entity recognition [Chiron et al., 2017, Hill and Hengchen, 2019, van Strien et al., 2020], and can only partially be mitigated by post-hoc correction [Rigaud et al., 2019, Nguyen et al., 2021]. Lastly, NER is a resource-intense task, specifically requiring named entity typologies, lexicons, corpora and more recently pre-trained language embeddings [Ehrmann et al., 2016]. All these often lack in a historical setting, or are not immediately re-usable due to the previously mentioned challenges.

Named entity recognition on historical documents is dominated by neural network-based approaches [Ehrmann et al., 2021]. Architectures relying on strong, pre-trained embeddings and BiLSTM-CRF layers [Todorov and Colavizza, 2020a] or transformer-based models [Boros et al., 2020] typically outperform alternatives. Perhaps as expected, the availability of larger corpora, lower degrees of noise and pre-cleaning texts via heuristics all contribute towards achieving better results [Ehrmann et al., 2020a]. What is more, embedding models working at the sub-word level

(e.g., character or sequences of characters) mitigate issues due to noise and out of vocabulary forms. Fine-tuning large pre-trained embeddings also helps in general [Gururangan et al., 2020], and specifically for historical texts [Konle and Jannidis, 2020]. Related to our work, Hendriks et al. [2020] focus on person NER and entity linkage from notarial archives in Dutch, considering a similar period to ours and records which, at least in part, mention VOC sailors. They start by applying the spaCy [1] and BERTje [Vries et al., 2019] models, yielding limited out-of-the-box results. Further fine-tuning spaCy allows them to achieve a maximum precision, recall and F1-score of 0.731, 0.756, 0.743 respectively, in their best fold using fuzzy matching evaluation of at least 90% between recognized and ground truth mention spans. Their results provide us with an indicative, albeit not fully compatible, comparison in what follows.

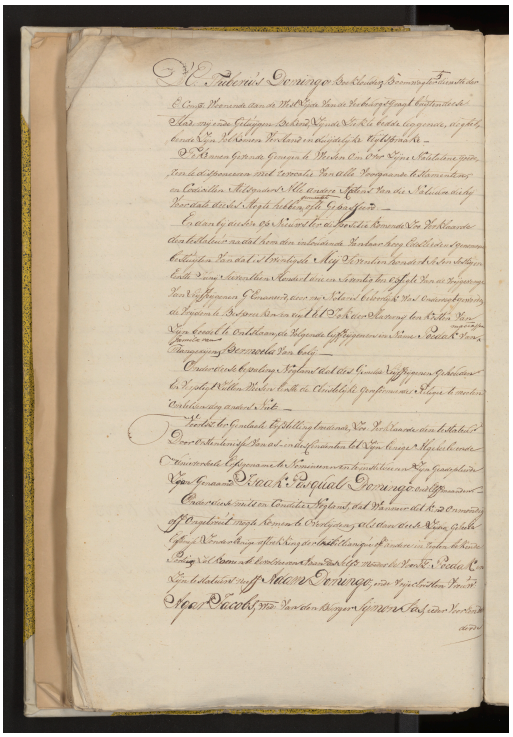
2.2 VOC Testaments

The establishment of the Dutch United East India Company (Verenigde Oost-Indische Compagnie, VOC), founded in 1602 as a merger of six small local companies, marked the beginning of Dutch expansion in Asia. In 1619, Batavia was founded after the conquest of the Javanese town Jacatra and became the administrative center of Dutch trade and power in Asia. The supreme authority was in the hands of the Governor General and Council of the Indies and in the areas which were under immediate control of the company, local administrative bodies were set up which fell under the authority of the Governor General and Council. The VOC settlements, such as Batavia, were complex microcosms, controlled by a very small group of Europeans (in 1680, about 7 percent of the population in Batavia was European), far away from the Dutch Republic, in a society that consisted of many different, mainly Asian populations. In the multi-ethnic and multi-religious settlements, the policy of the VOC was to maintain the status quo by keeping the different groups separate from each other: the free separated from the unfree, Christians separated from non-Christians [Brandon et al., 2020, 197]. Christian slave owners were not allowed to resell their slaves to non-Christians and within the VOC logic, Christian slaves were more likely to be released than non-Christians. Some local institutions that were established reflect this division. The “College van Weesmeesters" [Board of Governors of the Orphan Chamber] was an important facility for Europeans, as it ensured that the estates of deceased Europeans were handled. The counterpart of this board was the “College van Boedelmeesters van Chinese en andere Onchristen Sterfhuizen” [Trustees for the Deceased estates of Chinese and other non-Christian Bereaved]. Notaries, although not employed by the Company, played an important role in the larger VOC settlements. Between 1620 and 1650 there were always two public notaries in the city of Batavia and after that three or four. Notaries were obliged to offer their services to everyone, so not only to the Europeans, but also to the Asian population and even to the enslaved people who made up about half of the population of Batavia in the 17th and 18th century. The notaries thereby documented the lives of ordinary people and most of the individuals appearing in these archives belong to the Asian underclass [Brandon et al., 2020, 201-208]. In his book *Batavia*, Niemeijer [2012] gives many details of the lives of enslaved people, thus showing the richness of these notarial archives that offer the opportunity to get very close to the local society and its inhabitants. The Indonesian National Archives holds the archives of 111 notaries who resided in Batavia between 1620 and early 19th century, and some archives from notaries who worked in other settlements. These notary archives consist of almost 9000 inventory numbers with documents, deeds and registers with a total size of approximately 1160 linear metres.

A specific genre of notarial deeds are the testaments (or wills). Due to the high risk of death during their stay in Asia, personnel sent out by the VOC had to draw up a will. The number of personnel employed by the VOC was approximately 11,000 in the 17th century but rose to more than 25,000 in the company’s heyday. Testaments facilitated the settlement of estates after one’s death which was also in the interest of the company. Such a will could be drawn up in various ways: in the Republic, on the VOC ships and in the VOC settlements. Copies of wills made up in the settlements were sent to the VOC headquarters in the Republic after an employee passed away. A relatively small number of 10,000 of these wills (mainly from the 18th century) compiled in 51 bundles (‘banden’ in Dutch) consisting of a total of 53,370 pages survived and are nowadays in custody of the Dutch National Archives. A much larger set of copies of wills are kept in the archive of the Orphan Chamber [Weeskamer] in Batavia in custody of the National Archives in Jakarta. Wills drawn up by notaries are an example of a document type that has relatively unambiguous and well-structured diplomatic features. There is a fixed order in the text structure and wills have recurring elements such as the name of the notary, the name or names of the testator(s), the name or names of the beneficiaries (persons or organizations) and the names of witnesses. Dutch archivists compiled a name index of the 10,000 testators of these wills in the nineteenth century and even though in many cases married couples had a will drawn up together, only the names of the European male testators are indexed. It is a clear example of how the work of archivists contributes to silencing already marginalized people. The names of the female co-testators have been left out of the index, as have all the other names appearing in the wills: those of male and female locals of different ethnic backgrounds and the enslaved who appear in many different roles in the wills such as beneficiary, housemate, concubine, creditor, debtor, or property [Jeurgens and Karabinos, 2020]. These wills have recently been digitized and made machine-readable with the

[1] `https://spacy.io`.

HTR software *Transkribus* [Kahle et al., 2017] as part of the Dutch National Archive's[2] digitization project: *IJsberg Zichtbaar Maken*[3] [Making the Iceberg Visible]. An example of a scan of a VOC testament[4] and its corresponding HTR output is provided in Figure 1.



(a) Digital scan of a VOC testament page.

M: Thiberius Domingo: Boekhouder Boomwagten diensteder
E: Comp: Woonende aan de West zijde van de Verbeurgs gragt buijtendeese.
Stad: mij ende Getuijgen Bekend, zijnde ziek te bedde leggende, dogheb„
„bende zijn volkomen Verstand en duijdelijke uijtspraake.
—
Te kennen Gevende Genegen te Weesen Om over zijne Natelatene Goede„
„ren te disponeeren met revocatie van alle voorgaande testamenten,
en Codicillen Mitsgaders Alle andere Actens van die Natuure die hij
ternaakt
Voor dato deeses Mogte hebben, ofte Gepasseerd
En dan bij deesen op Nieuwster dispositie komende zoo verklaarde
den testateur na dat hem den inhoudende van haar hoog Edelheedens genomene
besluijten van dat is twintigste Meij Seventien hondert sesen sestig en
Eerste lunij Seeventtien Hondert drie en seventig ten opsigte van de vrygevinge
Van Leijffeijgenen G'Emaneerd, door mij Notaris behoorlijk was onderregt
geworden
de Vrijdom te Bespreeken en uijt het lok der Slavernij ten kosten van
maccassen
zijn boedel te ontslaan, de volgende lijffeijgenen in Name Poedak van
Jamela van
Mangerijen, Bermoela van balij:
-
Onder deese bepaling Nogtans dat des Gemelde Leijsfeijgenen gehouden
En Verpligt zullen Weesen Eerste de Christelijke gereformeerde Religie te moeten
Omhelsen dog anders Niet
I
voortsz: ter Genelaale Erfstelling treedende, zoo: Verklaarde den testateur
1
Door onstentenisse van as-en desCondenten tot zijn Eenige Algeheele ende
Punieverseele Erfsgename te Nomineeren en te institueeren zijn Geadopteerde
Zijn Genaamd Isaak Pasqual Ddomingo: oudElff maanden.
Onder deese mits en Conditie Nogtans, dat Wanneer dit kind Onmondig
off Ongetrout mogte komen te Overlijden, als dan deese zijno Geheele
Erffenise zonder Eenige aftrekking der tribillianique of andere in regten bekende
Portien zal kamen te devolveeren Aaan desselfs moeder de Voorsz: Poedak en
zijn testateurs neeff Adam Domingo:, ende vrijechristen Vrouw:
Agar Iacobs, wed: van den Burger Sijmon Pas, ieder Voorlende
derde

(b) HTR output from Transkribus.

Figure 1: Example of VOC testaments and their HTR processing.

3 Annotation Typology

To 'unsilence' colonial archives by broadening access, more inclusive finding aids are required, encompassing all persons mentioned in the archive with emphasis on marginalized ones. Existing generic typologies for named entity recognition and classification tasks such as CoNLL [Tjong Kim Sang and De Meulder, 2003] or ACE [Doddington et al., 2004] mainly focus on the high-level 'universal' or 'ubiquitous' triad *Person*, *Organization* and *Location* [Ehrmann et al., 2016]. These entity typologies alone are insufficient to overcome the challenges we face, for two main reasons. Firstly, they focus exclusively on *named* entities, while colonial archives also contain traces of unnamed persons: we need to broaden the scope of the typology to incorporate mentions of unnamed persons too. Secondly, colonial archives, and the VOC testaments archive more specifically, are rich in further information about entities of interest, for example their role, gender, legal status. This information is also important in view of enriching finding aids, and can be captured via entity attributes. To address these needs, we propose a fit for purpose NER annotation typology. Our custom typology extends the universal triad to encompass all mentions of entities, both named and unnamed, and further qualifies them by gender, legal status, notarial roles and other relevant attributes. What is more, an initial exploratory pilot was conducted in order to acquaint the annotators with the corpus and task at hand, and to consolidate the annotation typology and guidelines. The final typology is illustrated in Figure 2.

Typically, in named entity recognition and classification "the word 'Named' aims to restrict [Named Entities] to only those entities for which one or many rigid designators [..] stand for the referent" [Nadeau and Sekine, 2007, 130]. In our custom typology, only the entity type *Proper name* corresponds to this definition of the named entity. It thus separates the name of an entity (always tagged separately as *Proper Name*) from a generic reference to an entity type (*Person*,

[2] `https://www.nationaalarchief.nl/en`.

[3] `https://noord-hollandsarchief.nl/ontdekken/nhalab/ijsberg-zichtbaar-maken`.

[4] `https://www.nationaalarchief.nl/onderzoeken/archief/1.04.02/invnr/6848/file/NL-HaNA_1.04.02_6848_0150`.

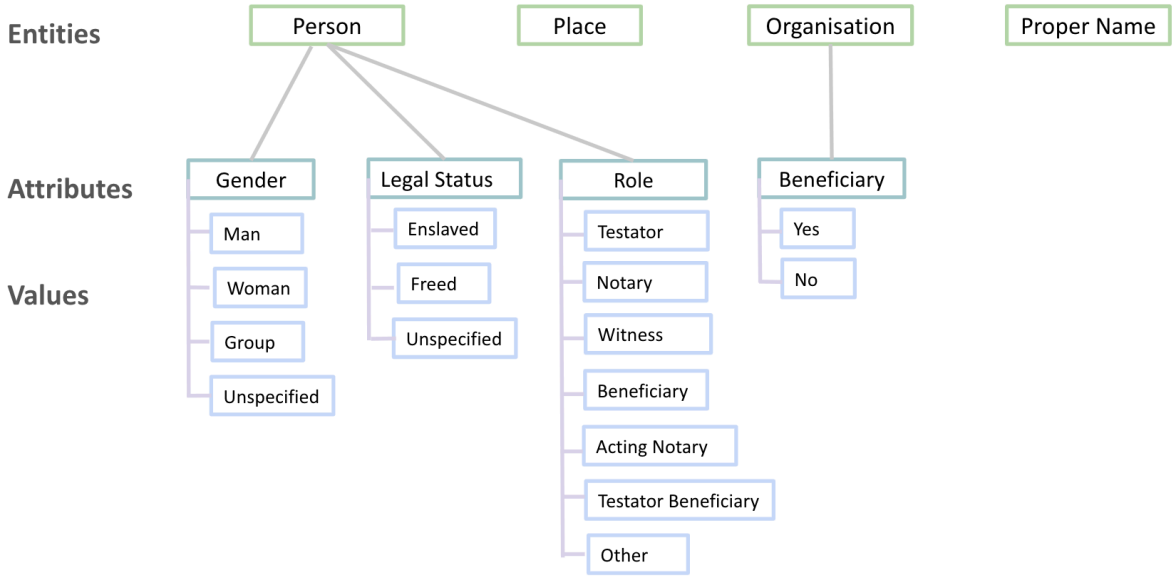


Figure 2: Proposed annotation typology.

Place or *Organization*). We introduce this distinction primarily because marginalized persons are frequently mentioned in the VOC testaments, and in colonial archives more generally, without name. Instead they are referred to by a variety of terms such as "slaaf" [slave], "leiffeigenen" [serf] and "inlandse burger" [indigenous person with some privileges]. *As a result, what we propose is an Entity Recognition and Classification task, expanding on the scope of classic Named Entity Recognition and Classification.* The detection of entities, either named or unnamed, introduces a further degree of complexity in the task since it might make annotation guidelines less precise with respect to the exact boundaries of annotations.

In what follows, we present the annotation typology and mention related annotation guidelines, with examples from the corpus.

3.1 Person

The entity type *Person* may refer to individuals or groups of people. When annotating a text span as a person, the span should include the proper name and/or available contextual "trigger words" [Ehrmann et al., 2021]. Trigger words in this typology also include words or phrases which provide information on the gender, legal status or notarial role of the person(s). Accordingly, the entity type person has three attributes: *Gender*, *Role* and *Legal Status*. When a person is mentioned multiple times across a testament (with or without trigger words), they are annotated with the same attribute which was inferred from the presence of the trigger words. An example is provided in Figure 9.

3.1.1 Gender

When the mention of a person is followed or preceded by trigger words which reveal their gender, the text is annotated as a *Person* with the appropriate value of the attribute *Gender*. Figures 4, 5 and 6 are examples of annotations for the gender attribute. For each entity person, the attribute gender takes exactly one of the values from the legend in Figure 3.

Figure 3: Legend for labeling person-gender attribute values.

When a person is mentioned without a gender trigger word, their gender is marked as *Unspecified*. This approach restricts possible 'annotator bias' due to unfounded inferences.

Excerpt from VOC testaments:

… zal komen te devolveeren Aan desselfs moeder de Voorsz: Poedak en zijn testateurs neeff Adam Domingo:, ende vrijechristen Vrouw: Agar Jacobs, wed: van den Burger Sijmon Sas, ieder Voorende derde …

English translation:

… will belong to the mother of the aforementioned Poedak and the testator's nephew Adam Domingo, and the free Christian woman Agar Jacobs, widow of the citizen Sijmon Sas, each for one third …

Figure 4: Instance of annotations of genders of persons, with and without leading qualifiers.

Excerpt from VOC testaments:

…Onder deese bepaling Nogtans dat des Gemelde Leijffeijgenen gehouden en verplicht zullen Weesen Eerste de Christelijke gereformeerde religie te moeten omhelsen …

English translation:

… Under the stipulation, however, that the reported serfs shall be held and obliged first to embrace the Christian reformed religion …

Figure 5: Instance of an annotation of a group of persons.

Persons are annotated by trigger words alone, in the absence of a proper name and in the case marginalised persons such as enslaved and formerly enslaved persons. This is because such persons are often mentioned without name and are of particular interest to our research. An example of a mention of an enslaved man without name is given in Figure 6.

Excerpt from VOC testaments:

… Wie den man slaaf is dewelke met driehonderd rds is gerelageert …

English translation:

… who is the male slave who was released with 300 riksdollar …

Figure 6: Instance of an annotation of a person mentioned without name.

3.1.2 Legal Status

For each entity *Person*, the attribute *Legal Status* takes exactly one of the values explained using the legend in Figure 7. Figures 8 and 9 contain examples of annotations for the legal status attribute.

The attribute legal status takes the value *Enslaved* when the trigger words clearly identify the individual(s) to be enslaved (see Figure 8). The attribute value *Free(d)* is often triggered by the word 'vrije' [free]. It refers to persons who were set free (for different reasons such as when they bought themselves free, as an act of benevolence, or for economic reasons) sometimes on the condition that they adopted the Christian religion. It could also refer to children of the manumitted slaves who, although born free, kept carrying the adjective 'vrije' [free], or if they were Christian they were labelled as

Figure 7: Legend for tagging person-legal status attribute values.

Excerpt from VOC testaments:

…Onder deese bepaling Nogtans dat des Gemelde Leijffeijgenen gehouden en verplicht zullen Weesen Eerste de Christelijke gereformeerde religie te moeten omhelsen …

English translation:

… Under the stipulation, however, that the reported serfs shall be held and obliged first to embrace the Christian reformed religion..

Figure 8: Instance of an annotation of persons with legal status of enslavement.

'free Christian'. Finally, the adjective 'free' was also used for groups of free indigenous (who were never enslaved) labelled for instance as 'vrije inlander' [free native]. The attribute value *Free(d)* captures these three different senses of the word 'vrije', for which there is no clear way to clearly disambiguate among. When no trigger words are used, legal status is instead annotated as *Unspecified* (see Figure 9).

Excerpt from VOC testaments:

… zal komen te devolveeren Aan desselfs moeder de Voorsz: Poedak en zijn testateurs neeff Adam Domingo:, ende vrijechristen Vrouw: Agar Jacobs, wed: van den Burger Sijmon Sas, ieder Voorende derde …

English translation:

… will belong to the mother of the aforementioned Poedak and the testator's nephew Adam Domingo, and the free Christian woman Agar Jacobs, widow of the citizen Sijmon Sas, each for one third …

Figure 9: Instance of annotations of the legal status of persons.

In the excerpt in Figure 9, "Poedak" has the attribute value *Enslaved* in the absence of a trigger word that indicates enslavement. The reason is that Poedak is mentioned earlier (aforementioned) on the same testament as "lijffeijgenen in name Poedak van Macassar" [serf with name Poedak of Macassar]. Accordingly, all mentions of "Poedak" in the testament are labelled as a person with legal status enslaved.

3.1.3 Role

The attribute *Role* refers to roles specific to testaments in notarial archives, which may take exactly one of the following values illustrated in Figure 2: *Testator*, *Notary*, *Witness*, *Beneficiary*, *Acting Notary*, *Testator Beneficiary* or *Other*.

An *Acting Notary* is a role taken on by a person who, in the absence of an officially recognized notary, performs the notarial deed as can be inferred from the extract in Figure 10. The role *Testator Beneficiary* is attributed to those people who are both testator and beneficiary in the context of the testament. For instance, when man and wife collectively write down their testaments, each of them is a testator and often both of them are also each-other's beneficiaries. The role

Other is attributed to those persons whose role does not correspond to any of the six roles (for instance the annotation in orange in Figure 10) or is when their role is not clearly mentioned.

Excerpt from VOC testaments:

… Compareerde voor mij Jan van Zeijst boekhouder indienst
der E Comp: alhier, tot 't passeren deses bij absentie van den
gesw: scriba geauthoriseerd door den wel Edele heer Rob„
bert Hendrik Armenault opperkoopman, oud sabandhaer …

English translation:

…appeared before me Jan van Zeijst bookkeeper in the service
of the noble company here, for the passing of this, in the absence of the
sworn scribe authorized by the noble lord Rob,,
bert Hendrik Armenault chief merchant, former shahbandar…

Figure 10: Instance of annotations of an acting notary and a person with role: other.

3.2 Place

The entity *Place* is used to annotate places or locations. This entity is often called *Location* in other typologies such as CoNLL [Tjong Kim Sang and De Meulder, 2003]. Consider Figure 11 where place is annotated in yellow.

Excerpt from VOC testaments:

… bij de Edele Hooge regeeringe van Nederlands India geadmit,,
teerd, binnen de stad Batavia residerende..

English translation:

… admitted to the noble high government of the Dutch East Indies,
residing within the city of Batavia …

Figure 11: Instances of annotations of places.

3.3 Organization

This entity, as the name suggests, refers to organizations such as companies, orphanages, religious institutions and other branches of the church. Organizations have the attribute *Beneficiary* which can take the value *Yes* or *No* depending on whether the testator decides an organization to be their beneficiary. Figure 12 is an instance of the latter.

3.4 Proper Name

The entity *Proper name* refers to names (proper nouns) of the other entities in this typology: *Person*, *Place* and *Organization*. In Figure 13, proper names are annotated in pink, which can be compared with Figure 4 and Figure 9 where the same excerpt is labeled using the entity person and attributes gender and legal status respectively. In our dataset, annotations overlap.

4 Results

4.1 Annotated Corpus

HTR Quality The VOC testament texts were extracted via Handwritten Text Recognition (HTR) by the Dutch National Archives, using a model combining ground truth from 17th and 18th-century VOC records. The ground truth

Excerpt from VOC testaments:

… Middelen te versoekende en Constitueeren het Eerwaarde Collegie van heeren Weesmeesteeren deeser steede gevende aan hun Eerwaardens zoodanige Breedvoerige last magt …

English translation:

… resources to request and constitute the honorable board of the chamber of orphans of this city, giving the honorables such broad burden and power to…

Figure 12: Instance of an annotation of an organization.

Excerpt from VOC testaments:

… zal komen te devolveeren Aan desselfs moeder de Voorsz: Poedak en zijn testateurs neeff Adam Domingo:, ende vrijechristen Vrouw: Agar Jacobs, wed: van den Burger Sijmon Sas, ieder Voorende derde …

English translation:

… will belong to the mother of the aforementioned Poedak and the testator's nephew Adam Domingo, and the free Chirstian woman Agar Jacobs, widow of the civilian Sijmon Sas, each for one third …

Figure 13: Instances of annotations of proper names.

for this combined model consists of 4810 manually transcribed pages from the VOC archives. The Dutch National Archives report an HTR Character Error Rate (CER) of 5.3 on a test set and 7.3 on a held-out sample set[5]. The CER for a given page is calculated as:

$$\text{CER} = \frac{i + s + d}{n} \times 100$$

where n is the number of characters inclusive of spaces; i, s and d are the minimum number of insertions, substitutions and deletions respectively, required to attain the ground truth result from the HTR text.

Corpus Selection At the onset of the project, it was unclear by which criteria testaments are grouped into a bundle and also whether there is a logic in the order of the 51 extant bundles. Given this ambiguity and in the attempt to capture as much variation in content and transcription quality as possible, 13 non-consecutive and equally spaced (i.e., every fourth) bundles have been selected for annotation. From each of these, a range between 15-50% pages have been annotated (as much as possible). Each bundle contains on average 1200 pages, thus 180-600 pages have been annotated per bundle. Each annotator was allocated a fixed number of pages per bundle. Furthermore, for each resulting bundle-annotator pair, 10% of the pages have been duplicated into the sample allocated to two other annotators, in order to calculate their inter-annotator agreement. During the pilot, we established that this overlap was sufficient for the stable calculation of inter-annotator agreement. Given that different annotators advanced at different speeds and different pages contain varying amounts of text, the resulting overlap between each set of annotators might be less than 10%, as shown in Figure 14.

The Brat annotation tool [Stenetorp et al., 2012] was used for manually annotating the corpus. While working on HTR texts in Brat, the annotators were invited to compared the texts with the scans provided on the website of the Dutch National Archives. This way of working proved instrumental in overcoming the limitations of HTR quality. The corpus

[5] `https://noord-hollandsarchief.nl/ontdekken/nhalab/project-transkribus-2`.

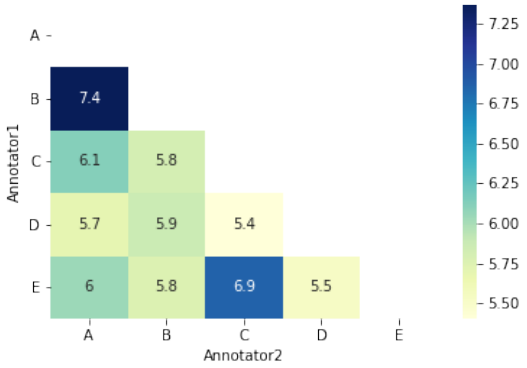


Figure 14: Overlap of annotated pages between each pair of annotators, calculated via Jaccard's distance. The overlapping pages are used to calculate the inter-annotator agreement.

is also provided in machine-readable IOB format (inside-outside-beginning); further details on the export from Brat to IOB are provided in the accompanying repository.

Corpus Characteristics The corpus consists of 2193 unique pages, plus 307 duplicated ones for calculating the inter-annotator agreement, resulting in a total of 2500. This corresponds to roughly 4% of the entire VOC testaments archive. The total number of annotations is 68,429, of which 32,203 at entity level (47%) and 36,226 at attribute level (53%); more details are given in Table 1. The total number of annotated tokens is 79,797. We divide the corpus of annotations into three splits: training (70%), validation (10%), and test (20%). We randomly sample annotated pages into splits by applying stratified sampling over annotation typologies and annotators, to maintain the overall data distribution over splits.

Entity type	Number of annotations	Percentage over total
Person	11,715	36.3 %
Place	4510	14.0 %
Organization	1080	3.5 %
Proper name	14,898	46.2 %
Total	**33,203**	100 %

Table 1: Number and share of annotations per entity type.

Inter-Annotator Agreement We use the Cohen's kappa score to evaluate the inter-annotator agreement [McHugh, 2012]. We measure it both exactly and using a *fuzzy matching offset*. This we define as the character offset that can exist between the same annotation given by two different annotators. Using an offset of 0 is equivalent to requiring an exact match, whereas an offset of 5 characters would entail considering two annotations to be the same if they overlap with a discrepancy of 5 characters at most. The inter-annotator agreement results between all pairs of annotators are shown in Figure 15 (a), while the average scores per entity are in Figure 15 (b). While with exact comparisons the kappa scores are only of moderate quality (0.5-0.6), with a modicum of fuzziness they converge to acceptable or strong values of 0.7-0.8 (at the 10 character offset mark).

4.2 Entity Recognition Baselines

In order to provide for a strong baseline to our proposed task, we make use of the best model configuration established in recent work on NER for historical documents [Todorov and Colavizza, 2020b]. This model text representation layer combines a variety of embeddings, including character-level embeddings and those produced from trained BERT models. We here use BERTje [Vries et al., 2019], a state-of-the-art Dutch version of BERT. All embeddings are concatenated and followed by a Bi-LSTM-CRF layer [Huang et al., 2015]. We use and compare single-task and multi-task approaches. The former focuses on learning one entity type at a time, whereas the latter combines these tasks into a single model. Finally, we include results from the same baseline scorer used in the CLEF-HIPE-2020 challenge [Ehrmann et al., 2020a]: a Conditional Random Fields model [Lafferty et al., 2001] based on the CFRsuite implementation [Okazaki, 2007] and exposed via the `sklearn_crfsuite` package.

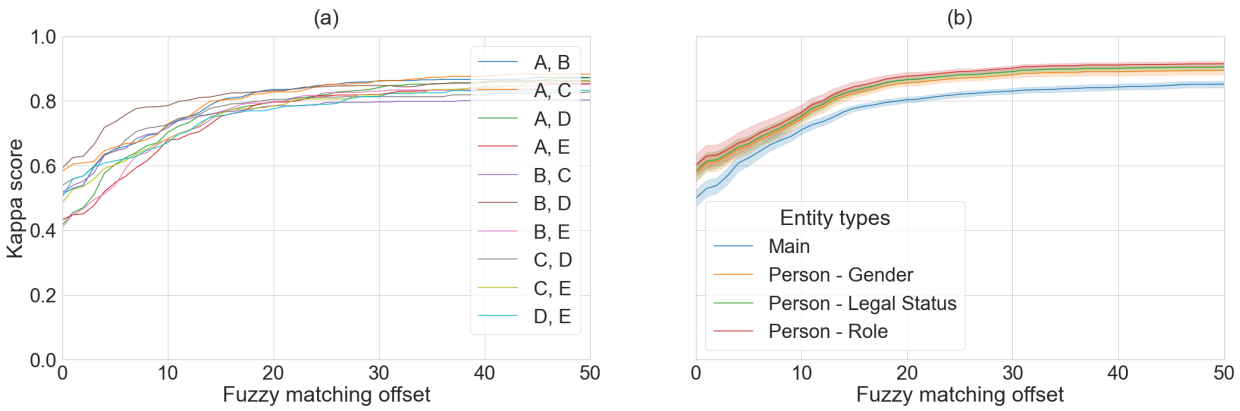


Figure 15: Inter-annotator agreement evaluation, considering the Person, Place and Organization entities. (a) Scores between annotators. (b) Averages per entity, with 95% bootstrapped confidence intervals.

We use the data splits established as described above, and execute three runs with different random seeds and identical model configuration, reporting results on the run achieving best results over the validation set. Our scoring approach follows best practices from CLEF-HIPE-2020 [Ehrmann et al., 2020a,b]: all metrics are calculated as micro averages at annotation level (not the token level). The document-level macro average is given in two different flavors: as the average of micro scores across pages, and as the overall macro average at annotation level. We remind the reader that a document corresponds to a transcribed page in our corpus; in this section only, we use 'document' in a machine learning sense. Furthermore, we use *strict* and *fuzzy* scoring. The former only considers exact boundary matching, whereas the latter includes overlapping boundaries. More specifically, fuzzy scoring considers predictions as correct if they have *at least* one token overlap with the ground truth (and the predicted tag is the correct one). We report precision (P), recall (R) and F1 Score (F).

Entity-level results for *Person*, *Place* and *Organization* ('Main entities') are shown in Table 2. Micro average scores for the entity type *Person name* and all attributes are shown in Table 3, except those for *Organization beneficiary* which are given separately in Table 4. The best scores for each metric and tag are in bold. What emerges is that the task is clearly difficult with results remaining low in particular for recall, while faring better for precision. Some tags are more difficult to detect than others, specifically *Organization beneficiary*. While a direct comparison is not possible, our results are not too distant from those achieved by Hendriks et al. [2020] on similar archival records and focusing on the *Person* entity type only. Lastly, we underline how the CRF baseline remains a strong option, even when compared with neural network-base approaches. The CRF baselines appears to be more precise, while the neural network architecture usually achieved better recall. This result is partially different from the conclusions drawn by the CLEF-HIPE-2020 challenge [Ehrmann et al., 2020a], and warrants further study.

Finally, in Figure 16 we provide confusion matrices showing how predictions compare against the ground truth at the entity type level (*Person*, *Place*, *Organization*). In general, we see how all models tend to over-predict the 'O' tag (Outside of an annotation): a common issue in NER models, in part caused by the overabundance of 'O' tokens. Mitigating this issue and thus improving recall, while also not loosing in terms of precision, constitutes a promising avenue for future work.

5 Conclusion

We started our work from the question of how it would be possible to find information on people that are hidden in archives. Extant indexes of historical records often embed systematic omission biases. The issue is particularly pressing for colonial archives, whose indexes ignore the presence of the enslaved and colonized. We have shown that it is possible to automatically and systematically surface information on marginalized groups even from these records. By considering the archives of testaments from the Dutch East India Company (VOC) as a case study, we have proposed an entity recognition shared task comprising an annotation typology, a corpus of annotations, and baseline results. Our main contribution entails expanding the task of named entity recognition to encompass mentions of unnamed historical entities, in particular persons. We found that, while challenging, the task appears doable even on such complex historical archival records.

Nevertheless, the proposed approach to 'unsilence' archival records might also be regarded as problematic: a form of disclosure leading to new, possibly dubious forms of categorization. Awareness of this issue has led us to avoid

Model	Main						Main - Macro					
	Fuzzy			Strict			Fuzzy			Strict		
	P	R	F	P	R	F	P	R	F	P	R	F
CRF baseline	**.73**	.56	**.63**	**.53**	.41	**.46**	**.69**	.43	.51	.37	.28	.32
BERTje + Bi-LSTM-CRF	.71	.57	**.63**	.51	.41	**.46**	.68	**.53**	**.59**	**.47**	**.37**	**.41**
+ multi-task	.69	**.59**	**.63**	.49	**.42**	.45	.67	**.53**	.58	.45	.35	.39

(a) Micro and macro averages.

Model	Main - Macro, Document					
	Fuzzy			Strict		
	P	R	F	P	R	F
CRF baseline	**.73** $\pm$ **.18**	.56 $\pm$ .22	**.63** $\pm$ **.19**	**.53** $\pm$ **.28**	.41 $\pm$ .26	**.46** $\pm$ **.26**
BERTje + Bi-LSTM-CRF	.71 $\pm$ .18	.57 $\pm$ .21	**.63** $\pm$ **.18**	.51 $\pm$ .28	**.42** $\pm$ **.25**	.45 $\pm$ .25
+ multi-task	.68 $\pm$ .19	**.59** $\pm$ **.22**	**.63** $\pm$ **.18**	.48 $\pm$ .28	**.42** $\pm$ **.26**	.45 $\pm$ .26

(b) Macro averages, document level aggregation.

Table 2: Results for the entity types Person, Place and Organization.

Model	Person name						Gender					
	Fuzzy			Strict			Fuzzy			Strict		
	P	R	F	P	R	F	P	R	F	P	R	F
CRF baseline	**.8**	.61	**.69**	**.71**	.54	**.61**	**.77**	.58	.66	**.47**	**.35**	**.4**
BERTje + Bi-LSTM-CRF	.72	**.65**	**.69**	.62	**.56**	.59	.73	.59	.65	.37	.3	.33
+ multi-task	.7	.64	.67	.58	.54	.56	.72	**.64**	**.67**	.39	**.35**	.37

(a) Entity Person name and attribute Gender.

Model	Legal status						Role					
	Fuzzy			Strict			Fuzzy			Strict		
	P	R	F	P	R	F	P	R	F	P	R	F
CRF baseline	**.77**	.6	**.68**	**.58**	.45	**.51**	**.76**	.53	.63	**.43**	**.3**	**.35**
BERTje + Bi-LSTM-CRF	.7	.63	.66	.5	.44	.47	.71	.6	.65	.33	.28	.3
+ multi-task	.7	**.66**	**.68**	.52	**.48**	.5	.71	**.61**	**.66**	.33	.28	.3

(b) Attributes Legal status and Role.

Table 3: Results for the entity type Person name and attributes; micro averages.

interpretations about which group people belonged to (e.g., by gender or legal status), unless this information was clearly stated in the sources. We also refrained from attempting to detect a persons' origin, since it would immediately result in problematic questions regarding ethnicity and its portrayal, in turn often erroneous and not helpful for modern-day finding aids. Finding the right balance for a respectful broadening of access to colonial records will require, going forward, a constant dialogue among archivists, scholars, governments, and the public.

We conclude by suggesting some directions for future work. Most immediately, the results from our baselines can be improved upon. After having reached satisfactory results, the proposed approach could be used to enrich the existing finding aids at the Dutch National Archives, turning this research into an application. In general, we deem important for more use cases from other colonial archives to be conducted, in such a way that more typologies, annotated corpora, models and other resources can be produced in the near future. These efforts should gradually be consolidated, primarily by devising a general typology and guidelines for entity recognition in colonial archives, providing for a shared conceptual ground all the while maintaining flexibility to accommodate the specificities of every archive. All efforts should not remain at the research stage but empower user-facing applications, in such a way that colonial archives can become more accessible and pluralized over time. To this end, the development of an ethical framework for the appropriate application of automation constitutes another key direction of future work.

Model	Organization beneficiary					
	Fuzzy			Strict		
	P	R	F	P	R	F
CRF baseline	**.27**	.12	**.17**	**.21**	**.09**	**.13**
BERTje + Bi-LSTM-CRF	.07	**.32**	.11	.0	.0	.0
+ multi-task	.26	.11	.15	.16	.07	.1

Table 4: Results for the attribute Organization beneficiary; micro averages.

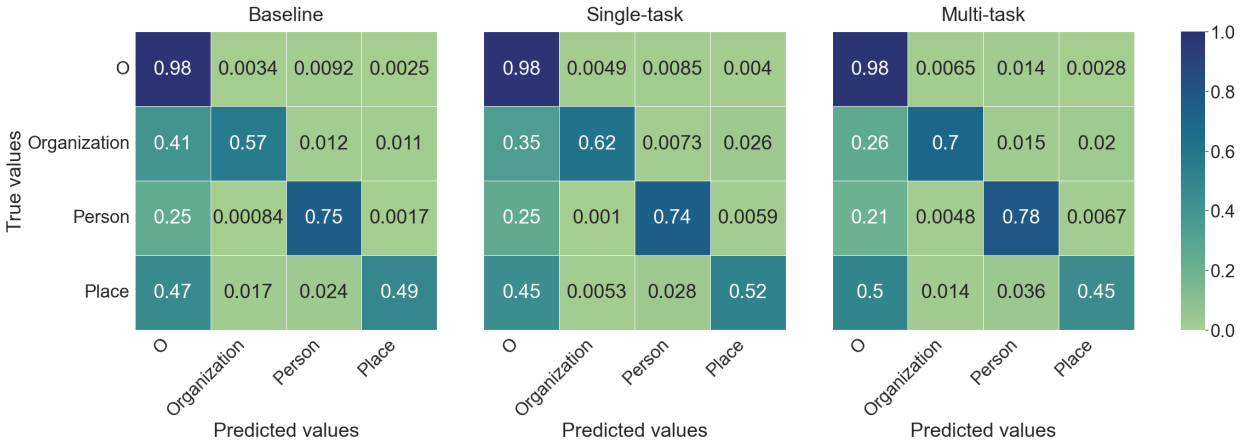


Figure 16: Normalized confusion matrices comparing ground truth and predictions for entities Person, Place and Organization.

Code and Data

The corpus of annotations, codebase and the corresponding data card[6] necessary to replicate our experiments can be found as a Zenodo archive: `https://doi.org/10.5281/zenodo.6958430`.

Acknowledgments

We thank Saskia Virgina Noot, Thijs Vorstenburg and Clare Shutt for conducting the pilot[7] for this project. This work was made possible by the digitization efforts of the Dutch Nationaal Archief and we thank Milo van de Pol, Liesbeth Keijser and Diederick Kortlang for providing us with context on the testaments. Nadia F. Dwiandari from the Indonesian Arsip Nasional was helpful in providing some background information on the VOC-notary archives in Indonesia. We express our gratitude for the integral feedback of the participants of our workshop at The Critical Visitor[8] Field Lab which have been crucial in helping us in thinking about the politics of categories which led us to revise our typology. The dataset was made possible by the annotations created by researchers: Roos Bijleveld, Silja de Vilder Coombs, Emma Louise van der Hage, Jonas Guigonnat, Yolien Mulder and Bert van Splunter. Sincere thanks to Leon van Wissen for setting up the annotation software infrastructure and his insightful feedback at numerous points during this project. We thank our reviewers for their feedback. Finally, we thank the digital humanities research group CREATE[9] at the University of Amsterdam and the Dutch Research Council (NWO, project number NWA.1228.192.108), for providing financial support.

References

Verne Harris. The archival sliver: power, memory, and archives in south africa. *Archival science*, 2(1):63–86, 2002. ISSN 1573-7519. doi:10.1007/BF02435631. URL `https://doi.org/10.1007/BF02435631`.

[6]The data card is based on and extends the accountability frameworks proposed by Gebru et al. [2021] (2021), Bender and Friedman [2018] (2018) and Pushkarna et al. [2022] (2022) to explicate assumptions and possible biases in dataset creation.

[7]`https://www.nationaalarchief.nl/innovatie-in-archiefonderzoek-prijs`.

[8]`https://www.universiteitleiden.nl/en/research/research-projects/humanities/critical-visitor`.

[9]`https://www.create.humanities.uva.nl`.

Gayatri Chakravorty Spivak. The rani of sirmur: An essay in reading the archives. *History and theory*, 24(3):247–272, 1985. ISSN 00182656, 14682303. URL http://www.jstor.org/stable/2505169.

Charles Jeurgens and Michael Karabinos. Paradoxes of curating colonial memory. *Archival Science*, 20(3): 199–220, 2020. ISSN 1573-7500. doi:10.1007/s10502-020-09334-z. URL https://doi.org/10.1007/s10502-020-09334-z.

Tony Ballantyne. Archives, empires and histories of colonialism. *Archifacts: The Journal of the Archives and Records Association of New Zealand*, 2004.

Antoinette M Burton et al. *Dwelling in the archive: women writing house, home, and history in late colonial India*. Oxford University Press on Demand, 2003.

Carolyn Hamilton, Verne Harris, Michèle Pickover, Graeme Reid, Jane Taylor, and Razia Saleh. *Refiguring the archive*. Springer Science & Business Media, 2002.

Jeannette A Bastian. *Owning memory: how a Caribbean community lost its archives and found its history*. Number 99. Libraries Unlimited, 2003.

Ann Laura Stoler. *Along the archival grain*. Princeton University Press, 2010.

James Lowry. *Displaced archives*. Taylor & Francis, 2017.

Ndeshi Namhila. *" Little Research Value": African Estate Records and Colonial Gaps in a Post-Colonial National Archive*. African Books Collective, 2017.

Joan M Schwartz and Terry Cook. Archives, records, and power: The making of modern memory. *Archival science*, 2 (1):1–19, 2002. ISSN 1573-7519. doi:10.1007/BF02435628. URL https://doi.org/10.1007/BF02435628.

Randall C Jimerson. *Archives power: memory, accountability, and social justice*. Society of American Archivists, 2009.

Michelle Caswell and Marika Cifor. From human rights to feminist ethics: radical empathy in the archives. *Archivaria*, 81(1):23–43, 2016. URL muse.jhu.edu/article/687705.

Jamila J Ghaddar and Michelle Caswell. "to go beyond": towards a decolonial archival praxis, 2019. ISSN 1573-7500. URL https://doi.org/10.1007/s10502-019-09311-1.

Marisa J Fuentes. *Dispossessed lives*. University of Pennsylvania Press, 2016.

David Thomas, Simon Fowler, and Valerie Johnson. *The silence of the archive*. Facet Publishing, 2017.

Michel-Rolph Trouillot. *Silencing the past: Power and the production of history*. Beacon Press, 2015.

Wendy M Duff and Verne Harris. Stories and names: archival description as narrating records and constructing meanings. *Archival Science*, 2(3-4):263–285, 2002. ISSN 1573-7519. doi:10.1007/BF02435625. URL https://doi.org/10.1007/BF02435625.

Geoffrey Yeo. Continuing debates about description. *Currents of Archival Thinking*, pages 163–192, 2017.

Giovanni Colavizza, Tobias Blanke, Charles Jeurgens, and Julia Noordegraaf. Archives and AI: An Overview of Current Debates and Future Perspectives. *Journal on Computing and Cultural Heritage*, 15(1):1–15, February 2022. ISSN 1556-4673, 1556-4711. doi:10.1145/3479010. URL https://dl.acm.org/doi/10.1145/3479010.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. Named Entity Recognition and Classification on Historical Documents: A Survey. *arXiv:2109.11406 [cs]*, September 2021. URL http://arxiv.org/abs/2109.11406. arXiv: 2109.11406.

Sonia Ranade. Traces through time: A probabilistic approach to connected archival data. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 3260–3265, 2016. doi:10.1109/BigData.2016.7840983.

Jane Winters and Andrew Prescott. Negotiating the born-digital: a problem of search. *Archives and Manuscripts*, 47 (3):391–403, 2019. doi:10.1080/01576895.2019.1640753. URL https://doi.org/10.1080/01576895.2019.164073.

Saidiya Hartman. Venus in two acts. *Small Axe*, 2008. ISSN 0799-0537. doi:10.1215/-12-2-1. URL https://doi.org/10.1215/-12-2-1.

Orlando Patterson. *Slavery and social death: A comparative study, with a new preface*. Harvard University Press, 2018.

Suze Zijlstra. *De Voormoeders. Een verborgen Nederlands-Indische familiegeschiedenis*. Ambo Anthos, 2021.

Durba Ghosh. Decoding the nameless: gender, subjectivity, and historical methodologies in reading the archives of colonial india. *A New Imperial History: Culture, Identity, and Modernity in Britain and the Empire*, pages 1660–1840, 2004.

Marjoleine Kars. *Blood on the river: A chronicle of mutiny and freedom on the Wild Coast*. The New Press, 2020.

Giovanni Colavizza, Maud Ehrmann, and Fabio Bortoluzzi. Index-Driven Digitization and Indexation of Historical Archives. *Frontiers in Digital Humanities*, 6, March 2019. ISSN 2297-2668. doi:10.3389/fdigh.2019.00004. URL https://www.frontiersin.org/article/10.3389/fdigh.2019.00004/full.

Marijn Koolen, Rik Hoekstra, Ida Nijenhuis, Ronald Sluijter, Esther van Gelder, Rutger van Koert, and Gijsjan Brouwer. Modelling Resolutions of the Dutch States General for Digital Historical Research. In *Proceedings of the International Conference Collect and Connect: Archives and Collections in a Digital Age*, 2020.

Melissa M Terras. The Rise of Digitization. In Ruth Rikowski, editor, *Digitisation Perspectives*, volume 39, pages 3–20. SensePublishers, Rotterdam, 2011. ISBN 978-94-6091-299-3. doi:10.1007/978-94-6091-299-3_1. URL http://dx.doi.org/10.1007/978-94-6091-299-3_1http://www.emeraldinsight.com.ezproxy.lancs.ac.uk/doi/full/10.1108/OIR-06-2015-0193.

Guenter Muehlberger, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, Hervé Déjean, Markus Diem, Stefan Fiel, Basilis Gatos, Albert Greinoecker, Tobias Grüning, Guenter Hackl, Vili Haukkovaara, Gerhard Heyer, Lauri Hirvonen, Tobias Hodel, Matti Jokinen, Philip Kahle, Mario Kallio, Frederic Kaplan, Florian Kleber, Roger Labahn, Eva Maria Lang, Sören Laube, Gundram Leifert, Georgios Louloudis, Rory McNicholl, Jean-Luc Meunier, Johannes Michael, Elena Mühlbauer, Nathanael Philipp, Ioannis Pratikakis, Joan Puigcerver Pérez, Hannelore Putz, George Retsinas, Verónica Romero, Robert Sablatnig, Joan Andreu Sánchez, Philip Schofield, Giorgos Sfikas, Christian Sieber, Nikolaos Stamatopoulos, Tobias Strauß, Tamara Terbul, Alejandro Héctor Toselli, Berthold Ulreich, Mauricio Villegas, Enrique Vidal, Johanna Walcher, Max Weidemann, Herbert Wurster, and Konstantinos Zagoris. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5):954–976, September 2019. ISSN 0022-0418. doi:10.1108/JD-07-2018-0114. URL https://www.emerald.com/insight/content/doi/10.1108/JD-07-2018-0114/full/html.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June 2016. Association for Computational Linguistics. doi:10.18653/v1/N16-1030. URL https://aclanthology.org/N16-1030.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Michael Piotrowski. *Natural language processing for historical texts*. Number 17 in Synthesis lectures on human language technologies. Morgan & Claypool, San Rafael, CA, 2012. ISBN 978-1-60845-946-9. OCLC: 812510472.

Guillaume Chiron, Antoine Doucet, Mickael Coustaty, Muriel Visani, and Jean-Philippe Moreux. Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–4, Toronto, ON, Canada, June 2017. IEEE. ISBN 978-1-5386-3861-3. doi:10.1109/JCDL.2017.7991582. URL http://ieeexplore.ieee.org/document/7991582/.

Mark J Hill and Simon Hengchen. Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843, December 2019. ISSN 2055-7671, 2055-768X. doi:10.1093/llc/fqz024. URL https://academic.oup.com/dsh/article/34/4/825/5476122.

Daniel van Strien, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pages 484–496, Valletta, Malta, 2020. SCITEPRESS - Science and Technology Publications. ISBN 978-989-758-395-7. doi:10.5220/0009169004840496. URL http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0009169004840496.

Christophe Rigaud, Antoine Doucet, Mickael Coustaty, and Jean-Philippe Moreux. ICDAR 2019 Competition on Post-OCR Text Correction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1588–1593, Sydney, Australia, September 2019. IEEE. ISBN 978-1-72813-014-9. doi:10.1109/ICDAR.2019.00255. URL https://ieeexplore.ieee.org/document/8978127/.

Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. Survey of Post-OCR Processing Approaches. *ACM Computing Surveys*, 54(6):1–37, July 2021. ISSN 0360-0300, 1557-7341. doi:10.1145/3453476. URL https://dl.acm.org/doi/10.1145/3453476.

Maud Ehrmann, Damien Nouvel, and Sophie Rosset. Named Entity Resources - Overview and Outlook. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3349–3356, Portorož,

Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1534.

Konstantin Todorov and Giovanni Colavizza. Transfer Learning for Historical Corpora: An Assessment on Post-OCR Correction and Named Entity Recognition. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*, 2020a.

Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, Jose G Moreno, Nicolas Sidère, and Antoine Doucet. Robust Named Entity Recognition and Linking on Historical Multilingual Documents. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, 2020. URL http://ceur-ws.org/Vol-2696/paper_171.pdf.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névéol, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 12260, pages 288–310. Springer International Publishing, Cham, 2020a. ISBN 978-3-030-58218-0 978-3-030-58219-7. doi:10.1007/978-3-030-58219-7_21. URL https://link.springer.com/10.1007/978-3-030-58219-7_21. Series Title: Lecture Notes in Computer Science.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740.

Leonard Konle and Fotis Jannidis. Domain and Task Adaptive Pretraining for Language Models. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*, 2020.

Barry Hendriks, Paul Groth, and Marieke van Erp. Recognising and Linking Entities in Old Dutch Text: A Case Study on VOC Notary Records. In *Proceedings of the International Conference Collect and Connect: Archives and Collections in a Digital Age*, 2020.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. Bertje: A dutch bert model. *CoRR*, abs/1912.09582, 2019. URL http://arxiv.org/abs/1912.09582. arXiv: 1912.09582.

Pepijn Brandon, Guno Jones, Nancy Jouwe, and Matthias van Rossum. *De slavernij in Oost en West: het Amsterdam-onderzoek*. Spectrum, 2020.

Hendrik E Niemeijer. *Batavia*. Uitgeverij Balans, 2012.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 4, pages 19–24. IEEE, 2017. doi:10.1109/ICDAR.2017.307.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003. URL https://aclanthology.org/W03-0419.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon, 2004.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007. doi:https://doi.org/10.1075/li.30.1.03nad.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, 2012.

Marry L. McHugh. Interrater reliability: The kappa statistic. *Biochemia Medica*, pages 276–282, 2012. ISSN 18467482. doi:10.11613/BM.2012.031. URL http://www.biochemia-medica.com/en/journal/22/3/10.11613/BM.2012.031.

Konstantin Todorov and Giovanni Colavizza. Transfer Learning for Named Entity Recognition in Historical Corpora. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, 2020b. URL http://ceur-ws.org/Vol-2696/paper_168.pdf.

Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv:1508.01991 [cs]*, August 2015. URL http://arxiv.org/abs/1508.01991. arXiv: 1508.01991.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

Naoaki Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007. URL http://www.chokkan.org/software/crfsuite/.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. 2020b. URL http://ceur-ws.org/Vol-2696/paper_255.pdf.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, December 2021. URL https://www.microsoft.com/en-us/research/publication/datasheets-for-datasets/.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi:10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1776–1826, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi:10.1145/3531146.3533231. URL https://doi.org/10.1145/3531146.3533231.

Unsilencing Colonial Archives via Automated Entity Recognition

Dataset Download | Related Publication

Colonial archives remain difficult to access due to significant persisting barriers such as biases to be found in historical findings aids, such as indexes of person names, which perpetrate silences by omitting to include mentions of historically marginalized persons. In order to mitigate such limitations and pluralize the scope of existing finding aids, we propose using automated entity recognition for content based indexing. To this end, we contribute a fit-for-purpose annotation typology and apply it on a specific genre of the colonial archive of the Dutch East India Company (VOC). We release a corpus of nearly 70,000 annotations as a shared task, for which we provide strong baselines using state-of-the-art neural network models.

Authorship

PUBLISHER(S)

Nationaal Archief, University van Amsterdam (UvA), Emerald Publishing

INDUSTRY SECTOR

Academic

DATASET CURATORS

Mrinalini Luthra, UvA, 2022
Konstantin Todorov, UvA, 2022
Leon van Wissen, UvA, 2022
Charles Jeurgens, UvA, 2022
Giovanni Colavizza, UvA, 2022

DATASET ANNOTATORS

Emma Louise van der Hage, UvA, 2022
Jonas Guigonnat, UvA, 2022
Silja de Vilder Coombs, UvA 2022
Yolien Mulder, UvA, 2022
Roos Bijleveld, UvA, 2022

FUNDING

Dutch Science Foundation (NWO), grant number: NWA.1228.192.108, CREATE, UvA funds

FUNDING TYPE

Public Research Funding

DATASET CONTACTS

mrinalini.luthra@gmail.com
g.colavizza@uva.nl

Motivations

DATASET PURPOSE(S)

Research Purposes
Machine Learning
Training, testing and validation

KEY APPLICATIONS

Machine Learning

Entity Recognition

PROBLEM SPACE

This dataset was created for training entity recognition models to create more inclusive content based indexes on the collection of VOC testaments. See accompanying article (in peer review currently).

PRIMARY MOTIVATIONS

Provide ground truth for training entity recognition models on colonial archives

INTENDED AND/OR SUITABLE USE CASE(S)

ML Model Evaluation & ML Model Training for:

- Entity detection
- Attribute detection

Uses of Dataset

SAFETY OF USE

Research Use

CONJUNCTIONAL USE

Safe to use with other datasets

KNOWN CONJUNCTIONAL USES AND DATASETS

-

METHOD

Entity Recognition

SUMMARY

An entity recognition and classification model can be trained

KNOWN CAVEATS

This dataset contains a proportionally low number of organizations because of incomplete annotations.

Unsilencing Colonial Archives via Automated Entity Recognition

Dataset Snapshot

PRIMARY DATA TYPES

Sensitive data about people

Data about places, organizations and proper names

DATASET SNAPSHOT

Total Entities	32,203
Total Attributes	36,226
Total Annotations	68,429
Training	70%
Validation	10%
Testing	20%
Total Tokens Annotated	79,797
Average tokens per label	2.7
Human Annotated Labels	All

DESCRIPTION OF CONTENT

This dataset is based on the digitized collection of the Dutch East India Company (VOC) Testaments under the custody of the Dutch National Archives. These testaments of VOC-servants are mainly from the 18th century, for the most part drawn up in the Asian VOC-settlements and to a lesser extent on the VOC ships and in the Republic. The testaments have a fixed order in the text structure and the language is 18th century Dutch.

The dataset has 68,429 annotations spanning over 79,797 tokens across 2193 unique pages. 47% of the total annotations correspond to entities and 53% to attributes of those entities. Of the 32,203 entity annotations, 11,715 (36.3%) correspond to instances that represent persons with associated attributes of gender, legal status and notarial role, 4,510 (14%) correspond to instances of places, 1,080 (3.5%) correspond to organizations with attribute beneficiary and 14,898 (46.2%) correspond to proper names (of places, organizations and persons).

PRIMARY DATA MODALITY

Labels or Annotations

KNOWN CORRELATIONS

Gender presentation numbers are skewed towards predominantly **man** and **unspecified**;
Legal status numbers are skewed towards **unspecified**

HOW TO INTERPRET DATAPOINT

Each datapoint refers to a central entity that can be a person, place, organization or proper name or their attributes such as gender, legal status and notarial role of a person.

Each entity is represented by a span of characters across single or multiple connected tokens or words.

Datapoint Example

The shared annotation task was performed on the Brat annotation software. For each page of annotations of the testaments corresponding to a .txt file, an annotation file with .ann suffix was created. The general annotation structure is that each line of the .ann file contains one annotation, and each annotation is given an ID that appears first on the line, separated from the rest of the annotation by a single TAB character. The initial ID character 'T' corresponds to text bound annotations whereas 'A' corresponds to an attribute. Consider this example of an annotation from the sentence "Emancipatie van lijfeigenen, en …":

T1	Person 1298 1310	lijfeigenen
A1	Gender T1	Group
A2	LegalStatus T1	Enslaved
A3	Role T1	Beneficiary

Here, the term 'lijfeigenen' [serfs] with characters spanning 1298 to 1310 on that particular page is annotated as entity: Person with attributes A1, A2 and A3 corresponding to that Person's gender, legal status and notarial role.

The dataset is also provided in **machine-readable IOB format**.

Unsilencing Colonial Archives via Automated Entity Recognition

Data Collection & Sources

DATA COLLECTION METHODS

Annotations by paid students and professionals

DATA SOURCE

Digitized collection of the VOC Testaments. The testaments consist of 51 extant bundles consisting of 10,000 wills mainly from the 18th century.

DESCRIPTION OF DATA SOURCE

HTR Quality: The testaments were extracted via handwritten text recognition by the Dutch National Archives with a character error rate of 5.3 on a test set and 7.3 on a held out sample.

Speech Situation: The testaments were drawn up in the 17th and 18th centuries and information about which varieties of Dutch are represented is not available.

DATASET TYPE

Static

Data was collected once from a single source

COLLECTION METHODS

Annotations were created as a shared annotation task on the Brat annotation software.

DATA SELECTION CRITERIA

Pages were randomly sampled from 13 non consecutive and equally spaced (every 4th) bundle to capture as much variation in content and transcription quality.

DATA PROCESSING

The data i.e., the collection of annotations were cleaned to remove:

- Incomplete annotations: where a span is labeled as an entity but at least one of the corresponding attributes' value was not chosen by the annotator.
- Duplicate pages: HTR errors sometimes result in duplicate pages, these were labeled by the annotators as duplicates and were excluded from the dataset.

Labeling Process

LABELING METHOD

Manual Annotations

ENTITY TYPES

Entity	#	%
Person	11,715	36.4
Place	4,510	14
Organization	1,080	3.4
ProperName	14,898	46.2

METHOD SUMMARY

Annotations were created by highlighting the relevant span of text and choosing its entity type and where applicable exactly one attribute value through a drop down menu.

To tag the same span as two entities, the span must be selected two times and labeled accordingly. For example: 'Adam Domingo' has been labeled twice as a *Person* and *ProperName*.

ENTITY TYPE

Person

ATTRIBUTE DISTRIBUTION

Gender	#	%
Man	4,303	36.7
Woman	1,232	10.5
Group	420	3.6
Unspecified	5,760	49.2

DESCRIPTIONS & MOTIVATIONS

When the mention of a person is followed or preceded by trigger words which reveal their gender, the text is annotated as a *Person* with the appropriate value of the attribute *Gender*.

When a person is mentioned without a gender trigger word, their gender is marked as *Unspecified*. This approach restricts possible 'annotator bias' due to unfounded inferences. Persons are annotated by trigger words alone, in the absence of a proper name and in the case marginalized persons such as enslaved and formerly enslaved persons. This is because such persons are often mentioned without name and are of particular interest to our research.

iNote Non-binary is not included in set of gender attribute values given that we could not find any instances in the data source.

Unsilencing Colonial Archives via Automated Entity Recognition

Labeling Process

ENTITY TYPE

Person

ATTRIBUTE DISTRIBUTION

Legal Status	#	%
Free(d)	154	1.3
Enslaved	885	7.6
Unspecified	10,676	91.1

DESCRIPTIONS & MOTIVATIONS

The attribute legal status takes the value *Enslaved* when the trigger words clearly identify the individual(s) to be enslaved. The attribute value *Free(d)* is most often triggered by the word 'vrije' [free]. It refers to persons who were set free, children of the manumitted slaves and the groups of free indigenous. The attribute value *Free(d)* captures these three different senses of the word 'vrije', for which there is no clear way to clearly disambiguate among. When no trigger words are used or don't indicate legal status, the legal status is annotated as *Unspecified*.

The motivation to include legal status as a semantic category is because enabling findability of marginalized groups in colonial archives is one of the primary goals of the project.

ENTITY TYPE

Person

ATTRIBUTE DISTRIBUTION

Role	#	%
Testator	1,289	11
Beneficiary	1,830	15.6
Notary	473	4
ActingNotary	801	6.8
Testator Beneficiary	278	2.4
Witness	1,107	9.4
Other	5,937	50.7

DESCRIPTIONS & MOTIVATIONS

In the historic index—used until now— only the male testator was indexed, thus silencing women co-testators, beneficiaries such as enslaved persons, concubines, children, etc. The attribute *Role* was thus created to refer to roles specific to testaments in notarial archives, which may take exactly one of the following values listed in the adjacent table.

An instance of a role is the *Testator beneficiary* which is attributed to those people who are both testator and beneficiary in the context of the testament. For instance, when man and wife collectively write down their testaments, each of them is a testator and often both of them are also each-other's beneficiaries.

ENTITY TYPE

Place

ATTRIBUTE DISTRIBUTION

No attributes

DESCRIPTIONS & MOTIVATIONS

The entity *Place* is used to annotate places or locations. This entity is often called *Location* in other typologies such as CoNLL.

ENTITY TYPE

Proper Name

ATTRIBUTE DISTRIBUTION

No attributes

DESCRIPTIONS & MOTIVATIONS

The entity *Proper name* refers to names (proper nouns) of the other entities in this typology: *Person*, *Place* and *Organization*. In this typology we separate the name of an entity from a generic reference to an entity type because marginalized persons in colonial archives are frequently mentioned without name. For further motivation refer to the paper.

ENTITY TYPE

Organization

ATTRIBUTE DISTRIBUTION

Beneficiary	#	%
Yes	162	15
No	918	85

DESCRIPTIONS & MOTIVATIONS

This entity refers to organizations such as companies, governmental agencies, orphanages, religious institutions and other branches of the church. Organizations have the attribute *Beneficiary* which can take the value *Yes* or *No* depending on whether the testator decides an organization to be their beneficiary.

Unsilencing Dutch Colonial Archives

Use in Machine Learning or AI Systems

DATASET USE(S)	DATASET SPLIT(S)	USAGE GUIDELINES OR POLICIES
Training Testing Validation	We divide the corpus of annotations into three splits: training (70%), validation (10%), and test (20%). We randomly sample annotated pages into splits by applying stratified sampling over annotation typologies and annotators, to maintain the overall data distribution within every split.	CRF baseline is a strong option compared with neural network-based approaches. For further information, refer to the paper.

Description of Annotators & Curators

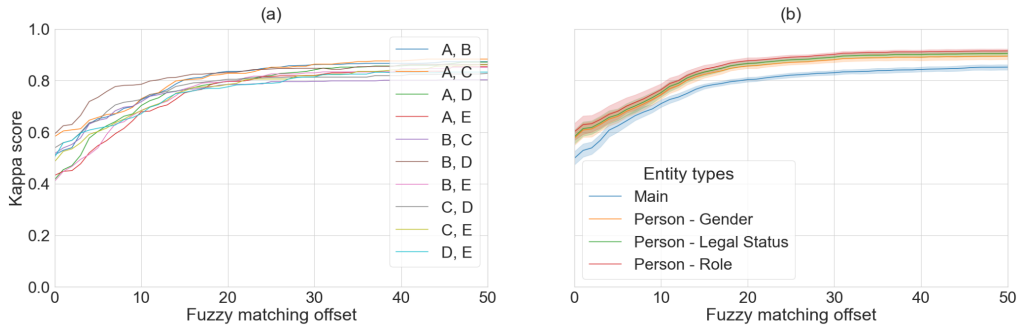
CURATORS

Mrinalini Luthra is responsible for overseeing the project.
Charles Jeurgens is the archival expert, who provided context of the archival records and terms that occur within them.
Giovanni Colavizza is the computer science expert.
Konstantin Todorov is the machine learning expert who set up and trained the baseline models.
Leon van Wissen set up the infrastructure for the collaborative annotation task.

ANNOTATORS

Annotators were recruited specifically for their expertise in 1) reading and understanding historical Dutch and 2) archival and historical knowledge. During the annotation process all annotators were trained to read and understand the original texts by the archival expert and were invited to compare the HTR texts with the scans of the original. This way of working proved instrumental in overcoming limitations of HTR quality.

INTER-ANNOTATOR AGREEMENT



Cohen's kappa score to evaluate the inter-annotator agreement. We measure it both exactly and using a *fuzzy matching offset*. This we define as the character offset that can exist between the same annotation given by two different annotators. Using an offset of 0 is equivalent to requiring an exact match, whereas an offset of 5 characters would entail considering two annotations to be the same if they overlap with a discrepancy of 5 characters at most. The inter-annotator agreement results between all pairs of annotators are shown in the first figure, while the average scores per entity are shown in the second). While with exact comparisons the kappa scores are only of moderate quality (0.5-0.6), with a modicum of fuzziness they converge to acceptable or strong values of 0.7-0.8 (at the 10 character offset mark).

Unsilencing Dutch Colonial Archives

License & Access

LICENSE TYPE(S)	LICENSE BREAKDOWN	LICENSE PERMISSIONS
CC BY 4.0	Annotations are licensed under CC BY 4.0 License. CC BY 4.0	**Share** — copy and distribute the material in any medium or format. **Adapt** — remix, transform, and build upon the material for any purpose, even commercially. **Attribution** —You must give appropriate credit, provide a link to the license, and indicate if changes were made. **No additional restrictions** — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

ACCESS TYPE(S)	ACCESS COST	ACCESS PREREQUISITE(S)
Open Access	N/A - Open Access	-

ACCESS SUPPORT	DATASET WEBSITE	ACCESS DETAILS
Dataset Website	https://github.com/budh333/UnSilence_VOC	-

RESEARCH PAPER	CITATION GUIDELINE(S)
Research Paper Paper currently under review	Mrinalini Luthra, Konstantin Todorov, Charles Jeurgens, Leon van Wissen and Giovanni Colavizza. “Unsilencing Colonial Archives via Automated Entity Recognition”. Zenodo. https://doi.org/10.5281/zenodo.6958430.

Versioning & Maintenance

VERSION STATUS	DATASET STATUS		MAINTENANCE PLAN
Limited Maintenance This data will not be updated, but any technical issues will be addressed	Version	1.2	- No refreshes planned - Dataset may be updated to incorporate feedback
	Last Updated	18/08/2022	
	First Released	18/08/2022	

References:

Bender, Emily M., and Batya Friedman. "Data statements for natural language processing: Toward mitigating system bias and enabling better science." *Transactions of the Association for Computational Linguistics* 6 (2018): 587-604.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. "Datasheets for datasets." *Communications of the ACM* 64, no. 12 (2021): 86-92.

Pushkarna, Mahima, and Andrew Zaldivar. "Data Cards: Purposeful and Transparent Documentation for Responsible AI." In *35th Conference on Neural Information Processing Systems*. 2021.