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)	**INDUSTRY SECTOR**	**DATASET CURATORS**
Nationaal Archief, University van Amsterdam (UvA), Emerald Publishing	Academic	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	**FUNDING TYPE**	**DATASET CONTACTS**
Dutch Science Foundation (NWO), grant number: NWA.1228.192.108, CREATE, UvA funds	Public Research Funding	mrinalini.luthra@gmail.com g.colavizza@uva.nl

Motivations

DATASET PURPOSE(S)	**KEY APPLICATIONS**	**PROBLEM SPACE**
Research Purposes Machine Learning Training, testing and validation	Machine Learning Entity Recognition	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	**INTENDED AND/OR SUITABLE USE CASE(S)**
	Provide ground truth for training entity recognition models on colonial archives	ML Model Evaluation & ML Model Training for: - Entity detection - Attribute detection

Uses of Dataset

SAFETY OF USE	**CONJUNCTIONAL USE**	**KNOWN CONJUNCTIONAL USES AND DATASETS**
Research Use	Safe to use with other datasets	-
METHOD	**SUMMARY**	**KNOWN CAVEATS**
Entity Recognition	An entity recognition and classification model can be trained	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)

Training
Testing
Validation

DATASET SPLIT(S)

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.

USAGE GUIDELINES OR POLICIES

CRF baseline is a strong option compared with neural network-based approaches. For further information, refer to the paper.

Description of Annotators & Curators

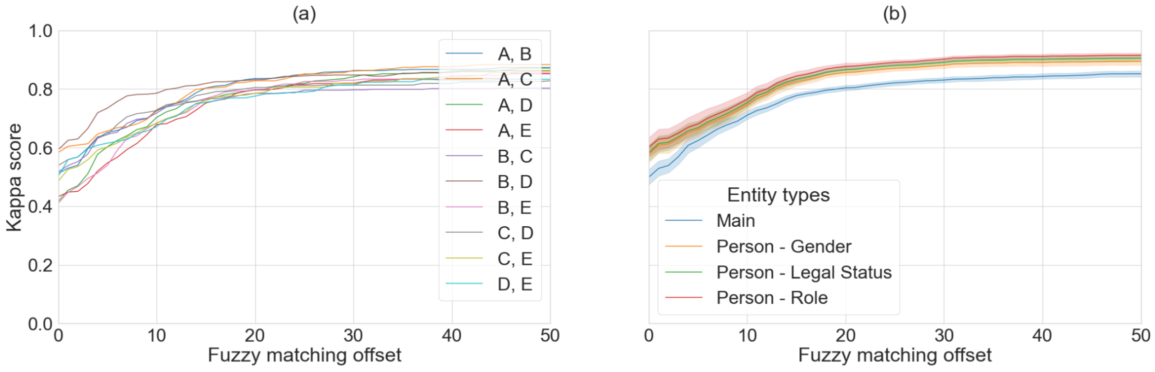
CURATORS

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	CITATION GUIDELINE(S)
Dataset	https://github.com/budh333/UnSilence_VOC	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.
Research Paper	**RESEARCH PAPER** Paper to be published in the Journal of Documentation’s special issue on *Artificial Intelligence for Cultural Heritage Materials.* Postprint on arxiv.	**CITATION GUIDELINE(S)** -

Versioning & Maintenance

VERSION STATUS	DATASET STATUS	MAINTENANCE PLAN
Limited Maintenance This data will not be updated, but any technical issues will be addressed		- No refreshes planned - Dataset may be updated to incorporate feedback

Version	1.2
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.