

# Accessing the Republic. Entity Extraction from the Resolutions of the Dutch States-General

Marijn Koolen<sup>1,2</sup>, Esger Renkema<sup>1,2</sup>, Nienke Groskamp<sup>1</sup>, Frank Smit<sup>1</sup>, Jirsi Reinders<sup>1</sup>, Ger Dijkstra<sup>1</sup>, Ronald Sluiter<sup>1</sup>, Rik Hoekstra<sup>1,2</sup>, and Joris Oddens<sup>1</sup>

<sup>1</sup>Huygens Institute, Amsterdam, Netherlands

<sup>2</sup>DHLLab - KNAW Humanities Cluster, Amsterdam, Netherlands

## 1. Introduction

In the REPUBLIC project, we are making digitally accessible the resolutions of the States General (SG) of the Dutch Republic (1576-1796). The SG were the central ruling body of the Dutch Republic, with representatives from each province in the republic, deciding on matters of interest common to all seven provinces, such as foreign policy, defence, trade and religion. The resolutions are transcripts of the decisions made by the SG in their daily meetings. The archive consists of almost 500,000 handwritten and printed pages and around 700,000-900,000 resolutions in total. Each resolution consists of at least two parts, a proposition, introducing the issue at stake, and a decision (Thomassen, 2019a).

We have made machine-readable transcriptions and segmented the text into individual resolution using OCR, HTR and structure extraction (Koolen et al., 2020, 2023b) and made them available and full-text searchable via a public search application<sup>1</sup> as well as downloadable datasets on Zenodo Koolen et al. (2025). Beyond full-text search, we offer additional access points for researchers to navigate and comprehend this large and complex resource. These access points are based on several categories of named entities, including person names, institutions and geographical names and various domain- and collection-specific entity types, which are operationalised in the application both as search facets and as links in the resolution text to further information.

Named Entity Recognition (NER) on historic documents has come along in leaps and bounds in the last decade (Ehrmann et al., 2023). Improvements are partially due to the rapid increase in the availability of large historic document collections (Kaplan and Di Lenardo, 2017; Terras, 2022, 2011), the improvement of easily trainable Automatic Text Recognition (ATR) models (Colutto et al., 2019; Kahle et al., 2017; van Koert et al., 2024), sequence tagging models and NLP frameworks (Akbik et al., 2019),

---

<sup>1</sup> <https://app.goetgevonden.nl>

and the development of language-specific Large Language Models (LLM) for historic languages (Manjavacas and Fonteyn, 2021, 2022).

The domain of historic administrative governmental documents brings its own challenges for NER. Beyond the fact that historic languages can differ from modern variants in spelling, vocabulary, morphology and syntax, there are additional challenges in dealing with highly specific and formulaic terminology (Koolen et al., 2023a), many abbreviations, and long and complex nested entity references (Aguilar et al., 2016; Prada Ziegler, 2024). Aguilar et al. (2016) describes a strategy for training a NER tagger for identifying person names and locations in Medieval Latin charters, and the challenge of single entity mentions that reference a person, a location and an organisation. Prada Ziegler (2024) focuses on historic land registers, where nesting is particularly challenging problem, and presents an approach for recursively handling the multiple levels of nesting. Orasmaa et al. (2022) created a test collection of annotated named entities in a corpus of 19th century Estonian parish records, and found that NER models perform well on person names but struggle with locations. NER has also been conducted on parliamentary debates (Hyvönen et al., 2024; Puren et al., 2025) and directories (Abadie et al., 2022), where the semi-structured nature of speeches and directory listing present their own challenges and opportunities for recognising and resolving named entities.

Evaluating and curating NER output should be informed by how the output will be used. In our case, the recognised entity references serve two main purposes. First, recognised entities are highlighted in the text presented to users of the REPUBLIC search application,<sup>2</sup> which helps these users to spot these entities and quickly grasp which persons, organisations and locations are involved in a resolution. Second, the entity references are curated and offered as search facets with which users can easily select subsets of the resolutions and get an overview of the entities that occur in the resolutions that are returned as results to a search query.

In this paper we report on our approach to extracting entities from the REPUBLIC corpus, including evaluation of NER taggers for different types of entities, and our findings from curating three entity types. Based on findings by Akbik et al. (2019), we experiment with taggers that combine different types of embedding models, e.g. character embeddings, static word embeddings and BERT-based embeddings.

We address the following research questions:

- How well can we identify named entities in the resolutions?
- How can we curate entity mentions to make them useful for information access?
- What can we learn from the curation of entities about the corpus of resolutions and the operation of the States General?

Although the NER output is far from perfect, our analysis shows that it can help users in drilling down into the corpus and get insights in the relationships between (types of) entities and the issues discussed in the resolutions.

Our approach to curating the entity references is not unique, but we describe the underlying model in the hope that it offers a generalisable process that has several advantages over traditional approaches to curation. Most important is that our approach makes curation decision explicit and reusable, allowing the process to be repeated when new versions of NER output are created.

---

<sup>2</sup> <https://app.goetgevonden.nl>

## 2. The Resolutions as a Corpus

The corpus of the resolutions of the States General of the Dutch Republic consists of 278,872 scans and 537,032 pages, divided across 657 books.<sup>3</sup> The resolutions were recorded as minutes by the greffier, who was present during the daily meetings and who was responsible for recording and archiving decisions, and providing previous resolutions to the meeting that were relevant for a matter at hand. The minutes were extended to full resolutions before being read at the start of the next meeting for approval. More information on the recording process is provided by Riemsdijk (1885); Thomassen (2019a,b), and on the digitisation and structure extraction conducted in the REPUBLIC project by Koolen et al. (2020, 2023a,b). In the project, we work with the fully extended resolutions. There are an estimated 700,000-900,000 resolutions<sup>4</sup> that together contain around 130 million words. The text of the scans has been automatically transcribed using Loghi, (van Koert et al., 2024) which was trained on ground truth scans and transcriptions created with a team of volunteers. The 107 scans of printed transcriptions and 515 scans on handwritten transcriptions are available on Zenodo (Sluijter et al., 2023; van Koert, 2023). There are handwritten and printed books of resolutions and each book contains either only *ordinary* or only *secret* resolutions. Most books contain handwritten text, but for the *ordinary* resolutions of 1703-1796, we used the printed versions that were available. The secret resolutions were, for reasons of security, never printed, and cover the period 1592-1796 (see Appendix A for more details on our selection and the complex nature of the corpus). Text recognition quality is high, with a Character Error Rate (CER) below 1% for the printed resolutions and around 2-3% for the handwritten resolutions (van Koert et al., 2024).

Each resolution contains a proposition and a decision.<sup>5</sup> For most propositions, the resolution mentions who introduced that proposition to the SG, and if the proposition was made via a document (letter, report, petition, bill, etc.), the date and location of sending it. The resolutions often mention organisations there are involved in the matters discussed, such as government, religious, military and naval institutions, as well as committees of SG members who were tasked to investigate the matter at hand, before coming to a final decision.

The corpus of resolutions is therefore connected to a larger archive of correspondence and decision making of the SG and is considered a key resource for studying the political history of the Dutch Republic and its interactions with the rest of the world.

## 3. Annotating Entities in the Resolutions

For digital access, the standard entity types of person, organisation and location are useful, but there are additional entity types that we think are valuable in the context of the resolutions and that make it easier to study aspects of the decision-making process, such as the committees that were tasked with investigating a proposed matter further, and references to earlier resolutions. We identify eight types of entities:

---

<sup>3</sup> This is the set of books used in the REPUBLIC project. The full archives of the States General contains many more books, some with copies of the resolutions, resolutions in minute form or with indexes to the resolutions that were created by the greffier's office, as well as all the incoming and correspondence and reports by committees.

<sup>4</sup> The exact number is unknown, as segmenting the transcribed text into individual resolutions contains errors.

<sup>5</sup> Although in the early years, up to roughly 1637, this model was used inconsistently, with a substantial set of resolutions containing only a decision paragraph.

- Person (PER): a person identified by name and identifying attributions or qualifications.
- Person attributions/qualifications (ATT): the attributions or qualifications, such as title, job, legal status or relationship to the SG that is used to identify a person.
- Committees (COM): the committees of the SG that are tasked with investigation matters raised in discussing a proposition.
- Organisations (ORG): organisations including the governing bodies of regions (e.g. the court of the Kingdom of France)
- Locations (LOC): geographical locations, including as part of the names of organisations or person attributions.
- Dates (DAT): explicit date reference, absolute (e.g. 15 April 1678) or relative (the 15th of last month).
- Resolutions references (RES): explicit references to an earlier resolution.
- Other names (OTH): any other names.

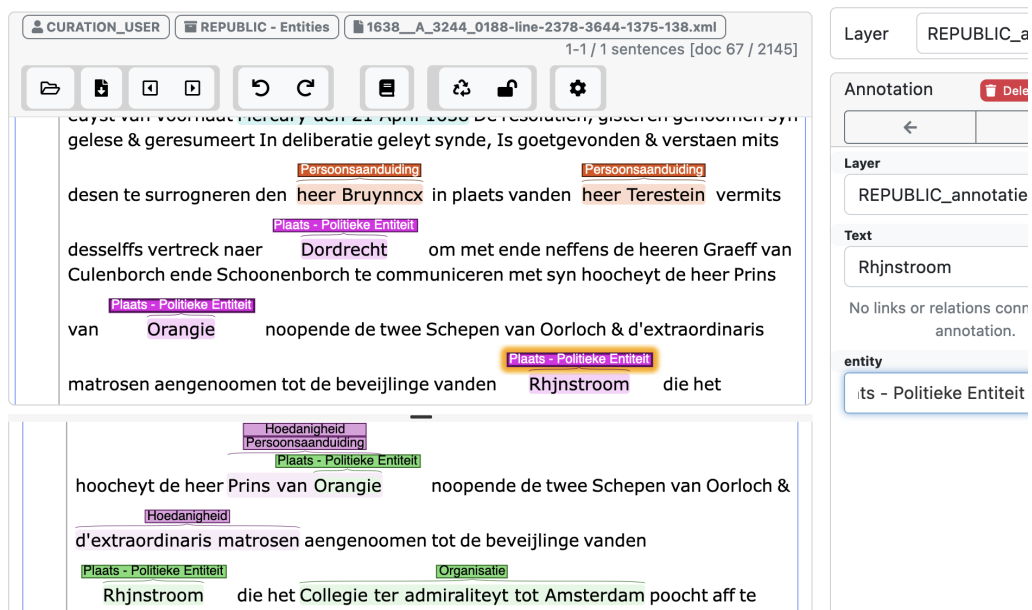


Figure 1: Tagging entities using INCEpTION.

Persons are rarely mentioned just by name, but almost always with attributions or qualifications such as their job (*ambassador, captain, carpenter*), title (*duke, princess*) or legal status (*widow, minor, heir*).

A difficulty arises in tagging locations that are a part of a reference to a person or organisation, as the named location can represent both the geographical location, the political organisation or the persons in that location. The ACE coding schema introduced geopolitical entities (GPEs) (Doddington et al., 2004)), but we opted to annotate multiple types separately.

We note that person attributions differ from the rest in that they do not refer to specific entities, but to generic groups of persons with specific attributes (therefore,

are strictly speaking not named entities). We decided to include them and tag them separately, because we think they provide a valuable additional access point to the resolutions, as they allow researchers to zoom in on all resolutions dealing with certain groups of people.

We iteratively developed annotation instructions, starting from a draft set of instructions, and updating them while several project members annotated a set of resolutions and discussed any disagreements, difficult cases and uncertainties.<sup>6</sup> The annotations of this development process were then discarded and a different set of resolutions was chosen for the annotation phase.

To train NER taggers, we worked with a group of 24 volunteers<sup>7</sup> who annotated the resolutions in two phases. In the first phase, which ran from September until April 2023, we created a ground truth dataset of 1,631 full resolutions (370,560 word tokens) randomly sampled from the printed resolutions of 1705-1796. In the second phase, which ran from July until September 2023, 513 paragraphs (28,387 word tokens) from the handwritten resolutions of 1597-1702 were annotated.<sup>8</sup> The entities were manually tagged using INCEpTION Klie et al. (2018) (see Figure 1). Although the length of the annotated paragraphs varies strongly (ranging from a single word to 6,269 words), the vast majority of annotated paragraphs range between 27 words (5th percentile) and 650 words (95th percentile). More details are available in Appendix A.1. Each resolution was annotated by three annotators. Agreement was high, with Cohen’s  $\kappa$  between pairs of annotators ranging between 0.7 and 0.9. Afterwards, a curator checked all annotations and resolved disagreements (mostly regarding the exact boundaries of entity references).

### 3.1. Nested Entities

One challenge is that entities can be nested, with e.g. a person entity containing an attribution, in turn containing an organisation, which can contain a geographical location. Prada Ziegler (2024) reports on the problem of nested entities in historical land registers of Basel, where around 60% of entity references are part of larger entity references (compared to 12% in the ACE 2005 data set<sup>9</sup>). Our data has many similarities with the Basel land registers, as 17,481 out of all 27,886 (63%) entity references contain a nested entity.

The fraction of entity overlap between pairs of types is shown in Table 1, indicating what fraction of entities of the type in the row overlap with the entity type in the column. Thus, of all committee references (COM), 25% contain at least one person attribution (ATT), 35% contain a location (LOC), 1% contain an organisation (ORG) and 91% contain a person name (PER). Locations are frequently part of person names (42%) person attributions (39%) and committees (35%), and the majority of person name references (69%) contain attribution information as well. Resolution references (RES) almost always contain a date (98%), but we found that only 21% of all dates are part of resolution references (in other words, most date mentions are not part of

---

<sup>6</sup> The annotation instructions are published with the entity data (Dijkstra et al., 2025)

<sup>7</sup> The REPUBLIC project worked with a large set of volunteers during the project, mostly for making and correcting ground truth transcription for the HTR and OCR step. The volunteers who signed up for the entity annotation task are a subset of this. All had developed knowledge and expertise regarding the nature of this corpus during the earlier transcription phase.

<sup>8</sup> The handwritten resolutions were done in a second phase because their transcriptions were not available for the first phase. The set is smaller because we had limited time and had to prioritise other tasks with the volunteers of our project. Again, see Appendix A for more details.

<sup>9</sup> <https://catalog.ldc.upenn.edu/LDC2006T06>

	ATT	COM	DAT	LOC	NAM	ORG	PER	RES
ATT	0.08	0.00	0.00	<b>0.39</b>	0.01	0.08	<b>0.15</b>	0.00
COM	<b>0.25</b>	0.00	0.00	<b>0.35</b>	0.00	0.01	<b>0.91</b>	0.00
DAT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LOC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NAM	0.00	0.00	0.01	0.04	0.00	0.01	0.00	0.00
ORG	<b>0.10</b>	0.00	0.00	<b>0.32</b>	0.00	0.01	0.08	0.00
PER	<b>0.69</b>	0.00	0.00	<b>0.42</b>	0.01	0.09	<b>0.10</b>	0.00
RES	0.00	0.00	<b>0.98</b>	0.00	0.00	0.00	0.00	0.01

Table 1: Fraction of entities of one type (row) that contain at least one entity of another type (column). In bold are fractions above 0.10.

larger entity mentions).

The way we model this has consequences for information access and the ease with which research questions can be studied. Consider the following example:

Henricus Gerhardus de Beveren Esveld, Predikant in de Gereformeerde Gemeente te Schoondyke onder het Classis van Walcheren, be roepen zynde tot Predikant in de Gemeente te Enkhuisen (EN: *Henricus Gerhardus de Beveren Esveld, Pastor in the Reformed Municipality at Schoondyke, under the Classis of Walcheren, named Pastor in the Municipality of Enkhuisen*)

This names a specific person, but the reference also contains entities of other types, e.g. the locations Schoondyke and Enkhuisen and their respective municipalities and the classis of Walcheren<sup>10</sup>, which contains the location of Walcheren.

Users may be looking for this specific person, but also for any of the entities nested inside this person name. Therefore, we want to identify and give access to all entities contained in a complex reference.

The way we model this is with overlapping annotations. In XML this would correspond to the following structure (text-centric view):

```
<PER>Henricus Gerhardus de Beveren Esveld, <ATT>Predikant in de
<ORG>Gereformeerde Gemeente te <LOC>Schoondyke</LOC></ORG>
<ORG>onder het Classis van <LOC>Walcheren</LOC></ORG></ATT>,
<ATT>be roepen zynde tot Predikant in de <ORG>Gemeente te
<LOC>Enkhuisen</LOC></ORG></ATT></PER>
```

With indentation per level (data-centric view), the structure is clearer:

<sup>10</sup> A *classis* is a supra-municipal meeting of multiple churches.

```

<PER>Henricus Gerhardus de Beveren Esveld,
  <ATT>Predikant
    <ORG>in de Gereformeerde Gemeente te
      <LOC>Schoondyke</LOC>
    </ORG>
    <ORG>onder het Classis van
      <LOC>Walcheren</LOC>
    </ORG>
  </ATT>,
  <ATT>be roepen zynde tot Predikant in de
    <ORG>Gemeente te
      <LOC>Enkhuisen</LOC>
    </ORG>
  </ATT>
</PER>

```

It is possible that an entity of a specific type contains another entity of the same type. For instance, a person name reference may contain an attribution that contains another person name. This is demonstrated in the following example:

```

Jan Barthold Wenthold, minderjaige Soon van Johan Ludolph ten Bhem Wenthold ; ... (EN:
Jan Barthold Wenthold, underage Son of Johan Ludolph ten Bhem Wenthold ; ...)

```

We consider the full phrase to be the most precise identifier for this underage son. This is tagged in INCEpTION with three entities as shown in Figure 1. We create two types of representations for training taggers. In one representation, a separate tagger is trained for each type of entity, in the other, a single tagger is trained to recognise all types of entities. Regardless of which representation we use, the flat representation required for training a sequence tagger lacks information on nesting. From (partial) overlap between entities of different types, we can reconstruct the nesting to some extent. In the above example, the person attribution partially overlaps with both the longest person name reference as the person name nested inside it. But this will not correctly identify all nesting, as in a flat structure, it is not always possible to distinguish between a nesting of two entities of the same type and a sequence of two entities of the same type. There are techniques that address these issues (Wang et al., 2022), such as the ‘second-best’ strategy by Shibuya and Hovy (2020) or the recursive strategy developed by Prada Ziegler (2024). As Table 1 shows, nesting of entities of the same type mainly occurs with person references and person attributions. In Section 5 we explain how we curate the recognised entities. Since we deconstruct composite attributions to give access to resolutions via individual terms, we do not need correct nesting information. For person references, we lack the resources to resolve the majority of them. In future work, we want to properly extract nested person references. Given the similarity to the complex person references in the Basel land registers, we consider using the recursive approach by Prada Ziegler (2024).

### 3.2. Entity Length

The references to entities have a high variation in length, from a single words to refer to a e.g. location, to a long phrase of dozens of words to describe a person or organisation as shown above. This variation in length is relevant for how well taggers can identify the correct boundaries. For the long person reference to *Henricus Gerhardus de Beveren*

Jan	Barthold	Wenthold ,	minderjarige	Soon	van	Johan	Ludolph	ten	Bhem	Wenthold
<i>tagged in inception</i>										
							B-PER	I-PER	I-PER	I-PER
			B-ATT	I-ATT	I-ATT	I-ATT	I-ATT	I-ATT	I-ATT	I-ATT
B-PER	I-PER	I-PER	I-PER	I-PER	I-PER	I-PER	I-PER	I-PER	I-PER	I-PER
<i>per-type training</i>										
			B-ATT	I-ATT	I-ATT	I-ATT	I-ATT	I-ATT	I-ATT	I-ATT
B-PER	I-PER	I-PER	I-PER	I-PER	I-PER	B-PER	I-PER	I-PER	I-PER	I-PER
<i>combined types training</i>										
B-PER	I-PER	I-PER	I-PER	B-ATT	I-ATT	I-ATT	B-PER	I-PER	I-PER	I-PER

Table 2: Tagging of nested entities and the representations used for training individual taggers for single types or a single tagger for all entities

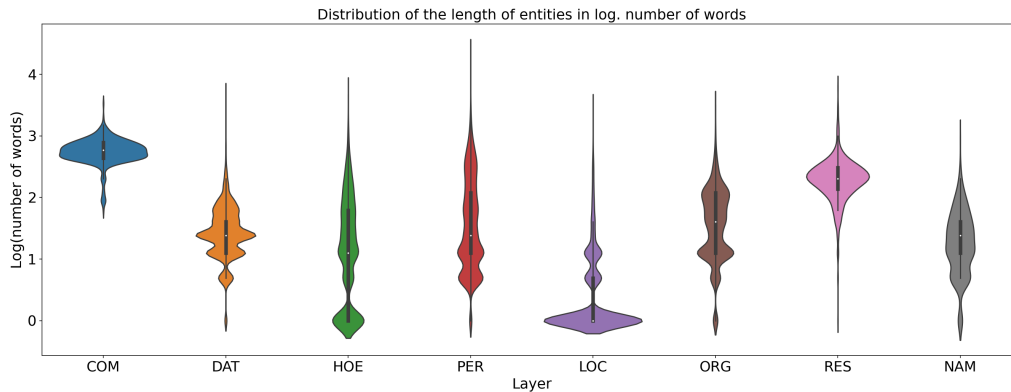


Figure 2: Distribution of the length of entities per type, in the logarithm of the number of words per entity.

*Esveld* above, a tagger could place the end boundary of the entity at various places in the long string, based on different levels of nesting. If the tagger leaves out the last attribution, organisation and location it has made a mistake according to the ground truth data, but not completely failed. Moreover, for curating the tagged entity references (see Section 5), the length has consequences for the possible amount of variation found in references to the same entity.

The distribution of entity lengths in the human-annotated data is shown in Figure 2, with the length of entities in number of words on a logarithmic scale. This scale is chosen because for some entity types the length ranges between a single words and dozens of words. Descriptive statistics are given in Table 3, with mean, standard deviation and 25th, 50th (median) and 75th percentiles.

There is a large difference in the length of entities across entity types. Locations are the shortest, with the majority of locations being at most 1 word, while the longest location name is 32 words. Attributions are slightly longer, with a median length of 3 words, but with high variation ( $Std = 3.8$ ), and some very long attributions.

Person names have the highest variation, with a median of 4 words but 25% of all person names are 8 words or longer. Long person names tend to contain attributions, locations and organisations.

Long organisation entities, i.e. 20 words or more, often contain lists of organisations, e.g. a single reference to the deputies of a list of provinces (“Heeren Gedeputeerden van de Provinciën van Gelderland , van Zeeland , van Vriesland , van Overijssel en

Type	Count	Mean	Std	Min	25%	50%	75%	Max
COM	391	15.9	3.0	6	14	16	18	34
DAT	3,034	4.5	2.1	1	3	4	5	40
ATT	7,743	4.2	3.8	1	1	3	6	39
LOC	6,803	1.9	2.0	1	1	1	2	32
NAM	360	4.0	2.1	1	3	4	5	19
ORG	3,437	5.5	3.2	1	3	5	8	33
PER	5,422	6.3	5.3	1	3	4	8	74
RES	639	10.3	3.6	1	8.5	10	12	44
Total	27,829	4.5	4.2	1	1	3	6	74

Table 3: Descriptive statistics of entity lengths in the human-annotated data in number of words.

van Stad en Lande”). This is a reference to a single group, the Gentlemen deputies of five of the seven provinces.

Committees are always multi-word phrases of at least six words long and the majority are 16 words or more. Resolution references are also longer phrases with 50% of all occurrences being at least 10 words long.

## 4. Training and Evaluating NER Taggers

To identify and classify entities in the resolutions, we trained multiple NER taggers, one per entity type.<sup>11</sup> The reason for this is that nested entities can more easily be extracted. Each tagger identifies a specific entity type, and by computing overlap (in terms of ranges in the running text of a resolution) between entities, we can determine which entities are contained within larger entities.

The resolutions were written over a period of 220 years, in which spelling and vocabulary changed, and words had no fixed spelling (partly due to there no being no official spelling rules for Dutch until 1804). On top of that, the OCR and HTR processes introduced recognition errors, which further increased the degree of orthographic variation. Therefore, the NER taggers should be able to deal with this.

For training taggers, we split the ground truth data into sets for training, validation and testing, using a 80/10/10 split. We used the Python Flair library (Akbik et al., 2019), which allowed us to combine multiple types of embeddings, including contextual character embeddings (Akbik et al., 2018) and word-level Fasttext embeddings (Bojanowski et al., 2017), both developed from scratch based on the texts of the resolutions, and GysBERT (Manjavacas and Fonteyn, 2022), which is a BERT-based model trained on historic Dutch.

We trained NER taggers using all possible combinations of embeddings and selected the best one per entity type. In our experimental setup, we trained models with any possible combination of five different binary parameters and the three types of embeddings (*Character*, *FastText* and *GysBERT*), using as ground truth either all entity types or each of the eight entity types separately (see Appendix B for details on the training parameters and links to code).

To evaluate and compare the models, we use set-based metrics *precision*, *recall* and

<sup>11</sup> The full `REPUBLIC` code base, including all code for NER and curation and the trained NER models have been made available. See Appendix B.

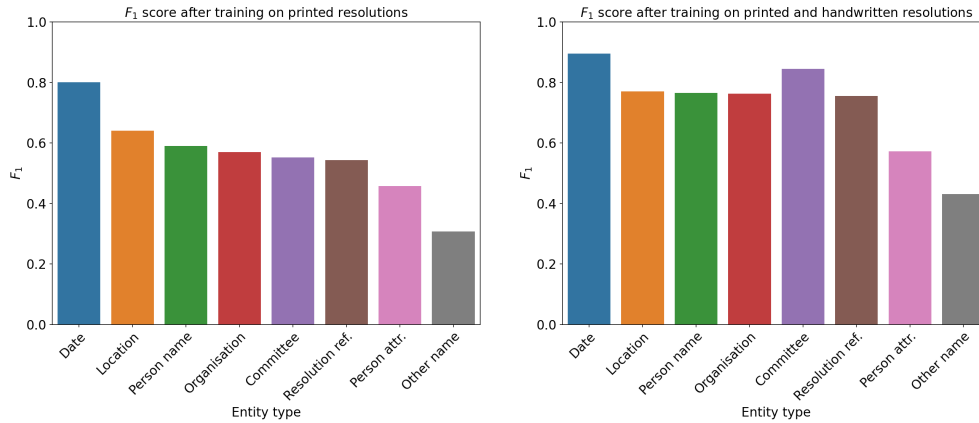


Figure 3:  $F_1$  score of best models per entity type after training only on the printed resolutions of 1705-1796 (left) and on both printed resolutions and the handwritten resolutions of 1597-1702 (right).

$F_1$  (the  $F$ -score that gives equal weight to precision and recall. In determining the correctness of a tagged entity, the default, strict, interpretation is that an entity is correctly tagged if the boundaries (start and end offsets) of the tagged entity exactly align with the entity in the ground truth. Especially for longer entity, it can be insightful to also include a more lenient interpretation that considers a tagged entity correct if it partially overlaps with the ground truth entity. For instance, if the ground truth entity is “the deputies of the provinces of Gelderland and of Vriesland” and the model tagged only “the deputies of the provinces of Gelderland”. Arguably, the tagged entity is still useful for users to quickly locate entities in the text (e.g. when entities are highlighted) and to select resolutions that mention the deputies of the province of Gelderland.

For model comparison, we only consider the strict interpretation. For the more detailed analysis in Section 4.2 we also consider the lenient interpretation.

As described in Section 4, we started with annotating entities in 1633 ordinary resolutions from the printed volumes of 1705-1796. Once we managed to extract decent quality paragraphs from the handwritten volumes of ordinary resolutions of 1597-1702<sup>12</sup> we also annotated entities in 513 handwritten paragraphs. Adding the ground truth annotations for the handwritten material made the taggers more robust, as can be seen in Figure 3, where the  $F_1$  scores of the best models per entity type are shown when using only the ground truth of the printed resolutions (left side) and all ground truth (right side).

The evaluation results for the best performing model per entity type (in terms of  $F_1$ ) are shown in Table 4. To understand the relative strengths of the different types of embeddings, we also include a comparative evaluation of models trained on a single type of embeddings. The Test column indicates which entity type the model was tested on, the Train column indicates whether the model was trained on only the annotations of the given entity type, or on the annotations of all types. In the latter case, the model is trained to tag all entity types, but we use only the tagged entities of the specific type for evaluation. The *support* column indicates the number of entities of each type in the test set.

In line with earlier work (Boros et al. (2020); Ghannay et al. (2020); Rodrigues Alves et al. (2018); Yang et al. (2018)), we find that all best models use Recurrent Neural

<sup>12</sup> For the secret resolutions and for the ordinary resolutions of the years 1576-1596 we had no good quality text yet at the time of ground truth creation. See Appendix A for more details.

Entity type		Embeddings	Prec.	Recall	F1	Support
Test	Train					
COM	COM	Char	1.00	0.73	0.85	41
	COM	FT	0.77	0.73	0.75	41
	COM	Gys	0.96	0.59	0.73	41
	COM	Best=Char+Gys	1.00	0.73	0.85	41
DAT	DAT	Char	0.90	0.88	0.89	249
	DAT	FT	0.87	0.86	0.87	249
	DAT	Gys	0.88	0.82	0.85	249
	DAT	Best=Char	0.90	0.88	0.89	249
ATT	ATT	Char	0.46	0.35	0.39	573
	ATT	FT	0.54	0.49	0.52	573
	ATT	Gys	0.61	0.46	0.53	573
	ATT	Best=Char+FT+Gys	0.57	0.56	0.56	573
LOC	LOC	Char	0.70	0.65	0.67	570
	LOC	FT	0.75	0.74	0.74	570
	LOC	Gys	0.79	0.60	0.68	570
	LOC	Best=FT+Gys	0.79	0.76	0.77	570
ORG	ORG	Char	0.75	0.60	0.66	283
	ORG	FT	0.80	0.71	0.75	283
	ORG	Gys	0.75	0.56	0.63	283
	ORG	Best=Char+FT+Gys	0.82	0.71	0.76	283
PER	PER	Char	0.72	0.58	0.64	405
	PER	FT	0.75	0.72	0.74	405
	PER	Gys	0.75	0.68	0.71	405
	PER	Best=Char+Gys	0.82	0.69	0.75	405
RES	RES	Char	0.81	0.62	0.70	57
	all	FT	0.83	0.67	0.74	57
	all	Gys	0.83	0.58	0.68	57
	all	Best=FT+Gys	0.82	0.70	0.76	57
OTH	All	Char	0.00	0.00	0.00	47
	All	FT	0.67	0.21	0.32	47
	All	Gys	0.30	0.21	0.25	47
	All	Best=Char+FT+Gys	0.63	0.26	0.36	47

Table 4: Evaluation results of NER taggers, trained on different types and combinations of embeddings, on the test set (*ordinary* resolutions of the periods 1597-1702 and 1705-1796).

Networks (RNNs) instead of linear layers and a Conditional Random Fields (CRF) model for the prediction layer to capture dependencies between sequences of tags (introduced by Huang et al. (2015)).<sup>13</sup>

The Train column indicates that most entity types are best served by training a tagger only on entities of that type. The two exceptions are resolution references (RES) and other names (OTH), which are better identified when the tagger is trained to also tag other types of entities. We speculate that for resolution references, it is beneficial to train a tagger that also learns to tag dates, because references to previous resolution always contain references to a specific date. Because the task of recognising dates seems easier (many models achieve high accuracy for tagging dates), we expect that models that separately tag the date part of resolution references, make it easier to correctly recognise the rest of the reference as a resolution reference.

For most entity types, performance in terms of  $F_1$  is in the range of 0.75-0.90, with high precision and recall around or above 0.70, indicating that the majority of entities are identified correctly. The two exceptions are Persons attributions (ATT) and Other names (OTH). Since we did not plan to use the latter category for information access, and most other types can be identified without training on all types by a single model, the low performance causes little problems. For attributions the difficulty is partly in correctly identifying the boundaries, leading to low recall (under the lenient interpretation, performance is much higher, see Section 4.2) and partly in tagging too much, i.e. mentions of attributions that do not refer to a specific entity.

#### 4.1. Comparing embedding types

For seven types, it is best to include the GysBERT model, which suggests that it helps to have some word- and phrase-level context to identify these types of entities. For Committees and Dates, using only a character-based embeddings model leads to better or at least equal performance compared to including word-level or contextual embeddings. We speculate that this is because they have a very standardised surface form, that merely require recognising the right character context, without interpreting deeper word-level semantics.

FastText outperforms Character embeddings for all types except committees and dates, and outperforms GysBERT for all types except attributions. FastText represents each words by sets of ngrams of length 3-6, whereas GysBERT represents words by one or a few sub-tokens. The better performance of FastText could mean that explicitly representing words by multiple levels of partially overlapping sub-tokens better captures the range of spelling variations found in long serial publications like the resolutions.

GysBERT performs better than Character embeddings for attributions, locations and persons entity types. The difference between the best combination model and the best model of the individual embedding types is relatively small.

#### 4.2. Period-specific evaluation

Koolen et al. (2023a) found that the spelling and use of formulaic language in the resolutions changes over the 220 year period. Therefore, it is possible that the performance of taggers differs across resolutions written in different periods. Based on the stylistic periodisation introduced in (Koolen et al., 2023a), we split the test data into

---

<sup>13</sup> We further experimented with various parameters of the PyTorch transformer models and FLAIR Stacked Embeddings, but leave out these details because of limited space.

label	Support			$F_1$ strict			$F_1$ lenient		
	1597	1705	1754	1597	1705	1754	1597	1705	1754
	1702	1753	1796	1702	1753	1796	1702	1753	1796
COM	2	17	22	1.00	0.97	0.74	1.00	0.97	0.74
DAT	27	132	90	0.88	0.87	0.88	0.96	0.97	0.94
ATT	82	225	266	0.60	0.62	0.50	0.85	0.84	0.77
LOC	74	242	254	0.85	0.81	0.75	0.93	0.88	0.89
OTH	4	9	34	0.00	0.75	0.33	0.00	0.94	0.54
ORG	30	117	136	0.73	0.76	0.70	0.81	0.83	0.85
PER	51	181	173	0.74	0.79	0.61	0.93	0.93	0.89
RES	4	23	30	0.67	0.71	0.75	0.89	0.81	0.89

Table 5: Evaluation of the NER taggers for resolutions written in different periods, with support per period (left), and the scores using strict scoring (center) and lenient scoring (right).

smaller test sets per period, so that we can compare performance of the taggers on resolutions from different periods. The available amount of ground truth per period differs strongly, with certain entity types having no or only a few positive examples for certain periods, we grouped the periods into three larger periods that each have enough positive examples to do a meaningful comparison: 1597-1702, 1705-1753, and 1754-1796.<sup>14</sup>

The strict and lenient evaluation scores per period are shown in Table 5. Columns 2-4 contain the number of positive examples per entity type and period. Particularly in the period 1597-1702 there is amount of test data is limited, meaning there is more uncertainty about the scores. Focusing on the strict scoring first (columns 5-7), the differences between 1597-1702 and 1705-1753 are small. For most entity types, occurrences are harder to recognise precisely in resolutions in the period 1754-1796. The only exceptions are dates (where there is almost no difference between the three periods) and resolution references, where performance is highest in the most recent period. With the lenient interpretation, the differences between the first two and the last period are smaller, suggesting that the difficulty in the last period is partially due to finding the correct boundaries. This is contrary to what we expected, as earlier work has established that the resolutions become increasingly formulaic, and we expect more formulaic phrasings of entities to be more easily recognisable. We have not yet been able to establish why this happens, but it serves as a warning that entity information in the latter period is less reliable.

## 5. Operationalising NER output

This section details the process of resolving textual references to named entities so that they can be operationalised in the search application as search facets and links in the text. A first version of the curated entity data has been published on Zenodo Dijkstra et al. (2025).

Operationalising entities found by automated processes is a problem of data linking, for which there at least two broad strands of research. We describe our approach using

<sup>14</sup> ground truth for the years 1703 and 1704 are missing, since we do not yet have resolution texts for these years.

the term *curation* because what we describe below differs in various ways from what is commonly used in the literature on recognition (NER), detection, (NED) and linking (EL) of entities. We do more than just link entity mentions and what we do differs slightly per entity type. We also categorise entities in a bottom-up data-driven fashion. Moreover, these processes contain many manual steps, and for this combination of steps, curation is a more common term.

### 5.1. Approaches to Linking Entity Mentions

One strand focuses on linking individual entity mentions to external knowledge bases Guellil et al. (2024); Shen et al. (2014). Often used external knowledge bases are DBpedia, Wikidata and Wikipedia (Hachey et al., 2013; Möller et al., 2022). However, these resources have a bias in their coverage of persons, locations and organisations towards the more famous or well-researched entities (Ilievski et al., 2018). This problem is particularly relevant for early modern texts, where coverage in external knowledge bases of the persons, committees and organisations mentioned in the text is even lower (Fabo and Poibeau, 2019; Munnely and Lawless, 2018; Pontes et al., 2020). Munnely and Lawless (2018) analysed a sample of 16 Irish depositions written in 1641, and found 480 reference to 283 entities distinct entities, of which only 64 (23%) had a representation in DBpedia. There are historic knowledge bases that are more relevant for historic documents of specific periods, languages and domains, but the examples we found focus on texts and knowledge bases of more recent periods. Brando et al. (2016) used the Linked Data repository of the French National Library for linking to 19th and 20th century literary authors, while Heino et al. (2017) developed WarSampo for linking to Finnish persons and geolocations for the period 1920-1950, to deal with the many changes due to Finland ceding large areas of land to Russia at the end of the Second World War.

The other strand of research is on linking historical records (see Bailey et al. (2020) and Abramitzky et al. (2021) for overviews), where entity mentions are not linked to external resources, but to other mentions of the same entity within the dataset (internal linking). In this strand there is a strong focus on linking census records (Abramitzky et al., 2020; Bailey et al., 2020; Ruggles et al., 2018) which are typically centred on persons and locations, although it has also been explored for linking cultural heritage collections (Van Hooland and Verborgh, 2014). The record linking approaches featured in the literature share many commonalities, such as using logical criteria to rule out certain matches, using edit distance and other forms of fuzzy matching to reduce variation and using weights or probabilities to rank or filter candidate links.

Just as with *recognising* named entities in text, *linking* them is made more challenging in texts that are the result of automatic text recognition (Ehrmann, 2008; Linhares Pontes et al., 2019; Pontes et al., 2020; Stern et al., 2012; van Strien et al., 2020). Text recognition errors increase the variation in spelling, which makes it more difficult both to recognise names and to then reduce the variation to a single form.

### 5.2. Linking Entities in the Resolutions

The corpus of resolutions is a serial publication written in the context of daily meetings of the same governing body, so it is likely that a substantial set of entities will be mentioned in multiple resolutions. Given the low coverage of external knowledge bases for persons, committees and organisations for early modern texts (Munnely and Lawless, 2018; Pontes et al., 2020), we focus on internal linking of entities, and

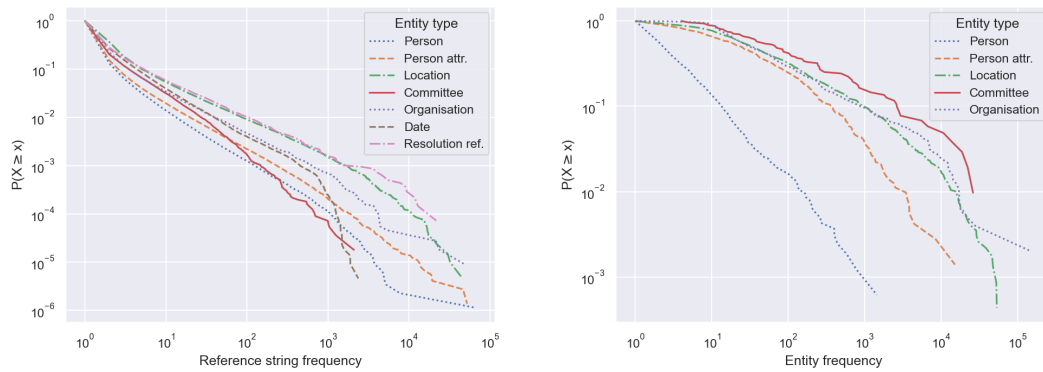


Figure 4: Complementary Cumulative Distribution Function of the frequency of entity reference strings (left) and resolved or curated entities (right). Both X and Y axis are on a logarithmic scale to show straight lines typical of long-tail distributions.

only link the resolved entities to external knowledge bases for a subset of locations, persons and organisations (Section 5.5). The internal links give users of the digital publication and search application a way to navigate and select resolutions via the different types of entities.

The resolutions lack the kind of structured elements that census records have, at least at the level that we could exploit for this curation phase. For each entity mention we therefore have limited contextual information.

What we do have is frequency information on the entity mentions, and their distribution provides us with some metrics. In a natural text like the resolutions, the entity distribution is marked by their importance for the organisation that created the texts. For all entity types, important entities for the States General appear more often in the text of the resolutions and also have more orthographic variations. Here we distinguish between the frequency of distinct *reference strings* and *entity mentions*. The reference strings ‘Raad van State’ and ‘Raed van Staatte’ each occur thousands of times. They are distinct strings, but refer to the same entity. The entity mentions are the sum of the occurrences of the reference strings. The frequency distribution of the reference strings is highly skewed, with most reference strings occurring only once and relatively few occurring (tens of) thousands of times. We show the Complementary Cumulative Distribution Function (CCDF) of this distribution in Figure 4 for the reference strings (left side) and the curated entities (right side). The CCDF shows the probability that a string or entity occurs at least  $x$  times. That is, for the frequency  $x = 100$ , the Y axis shows the probability of a string or entity occurring at least 100 times ( $X \geq 100$ ). The CCDF visualisation shows smooth curves that are easier to distinguish from and compare with each other at all points of the frequency range. All entity types have a distribution with a long tail.<sup>15</sup> Although most *distinct* reference strings occur only once, they represent only a small fraction of the total number of strings, which are dominated by the tens or hundreds of reference strings that each occur (tens of) thousands of times. The distribution of person references (left side plot) falls fastest, meaning there are relatively many low-frequency references compared to e.g. the distribution of location references. The CCDF shows that a person reference string has a 0.1%

<sup>15</sup> From a statistical modelling perspective, the tail consist of the small number of highly frequent entities, from a information extraction perspective, the tail is the opposite side of the distribution, consisting of the thousands of entities that occur only once or twice.

probability of occurring 100 times or more, while for a location reference string this is around 1%. For the curated entities (right side plot), the distribution of persons against falls fastest, having a relatively high proportion of entities that occur rarely. For the other entity types, the curves are much flatter (falling less fast) than for the reference strings. This means that they do not have a long tail of entities occurring only once or twice. The vast majority of committees, organisations, locations and attributions occur at least 10 times (at  $10^1$  on the X axis, their values on the Y axis are close to  $10^0$  or 100%). The curation process described below has thus strongly reduced variation for these entity types.

The most relevant entities from the perspective of the SG occur most frequently, therefore we consider their frequency to be an important clue for identification. The general procedure consists of finding the most frequent entities. In general we start with those entity reference strings that add up to 80-85 percent of all references of a certain type. Time permitting, this may further be extended to up to 90 percent. Then we try to reduce all variations to a standard form. Using the assumption that more frequent entities have more variations and some variations are more common than others, we choose the most frequent reference string as the standard form and map other variations to that form. These steps can be done using internal criteria. Further identification is possible by making a cross-section of persons and attributions. In the resolutions, people are often designated by a further attribution, if only for identification in the States General; 69% of person tags contain an attribution tag. Based on manual inspection, the person tags that contain no nested attribution tag are either persons mentioned earlier in a resolution, or are so well-known to the States General as to need no further qualification.

All identified entities are made available in the search application in the form of search facets and the resolved entity references are operationalised in the resolution texts as links to more information about them, and with the option to add them as search facets to narrow down the search results.

### 5.3. Two models of data curation

Entity type	Total references	Unique references	Unique entities
Person	1,864,673	882,081	8,108
Person att.	2,160,912	730,747	716
Location	2,187,588	201,265	2,459
Committee	139,787	55,528	103
Organisation	599,239	108,739	504
Date	834,983	217,581	43,208
Resolution ref.	185,210	13,754	?

Table 6: Total and distinct number of entity references, per entity type.

The number of entity references and the many sources of variation make a the usual method of data curation untenable. The process can be described in a schematic way as in figure 5. It is a three-step process: after applying a NER tagger to identify entity references, the text of the references is first cleaned up as far as possible in an automated way (spelling harmonisation and/or fuzzy matching reference strings that

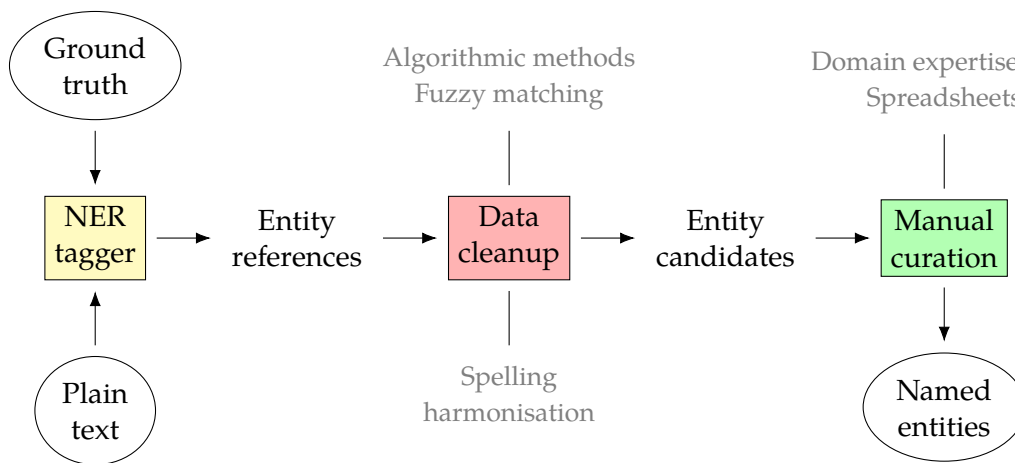


Figure 5: A linear interpretative model of data curation.

are variants), and then passed to a domain expert, usually in the form of a spreadsheet, for manual identification.

One advantage of this *linear interpretative* method is a clear separation of competences: technical and historical know-how can be provided by different people in succession. Some of its limits, on the other hand, are immediately clear. For one, a manual pass through all references becomes impractical when the number of *distinct* references rises above a few thousand. As table 6 shows, the compound variation gives rise to much larger numbers. And while there are various ways of reducing this variation, the sequential nature of the traditional method precludes applying domain-specific knowledge to this step. The gains of doing so, meanwhile, can be significant. Take for example the five admiralties of the Republic:<sup>16</sup> most references include a place name, and because the set of admiralty seats is limited, place names included in an admiralty organisation reference can be identified with certainty even when severely scrambled.

More generally speaking, the clean-up step benefits from the set of entities to-be-recognised being known in advance, which is not the case here. The largest drawback of this model, however, is that the manual step makes this a one-time process that cannot easily be repeated when either the input text, the NER tagger or the clean up process changes. The domain expert could ask the technical expert to add a step in the clean up process, but in the newly resulting spreadsheets, all manual work of the domain expert on the previous version needs to be redone. Since both automated text recognition and NER are rapidly-improving technologies, the need for an easily repeatable, and therefore automated, data curation step is obvious.

From these observations, we can draw two intermediate conclusions.

- One, that the integration of the two tasks of identification and resolution benefits from applying domain expertise throughout all steps of the curation process;
- Two, that any manual labour performed must be done in a way that has reusable results.

Taken together, these must lead to an *integrated process of data curation*, for which we propose a model outlined in figure 6. In this *iterative automated* model, the single

<sup>16</sup> To wit, those of Rotterdam ('on the Meuse'), Amsterdam, West-Frisia (with its seat alternating between Hoorn and Enkhuizen), Middelburg and Frisia (first in Doccum, later in Harlingen).

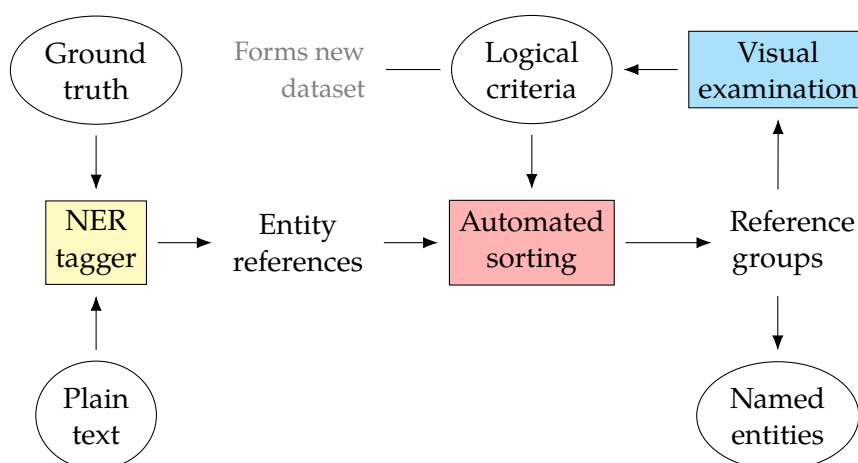


Figure 6: An iterative automated model of data curation.

pipeline is replaced by a two-part process, consisting of an automated pipeline and an iterative development loop containing manual curation.

The pivotal step in this process consists not of assigning references to entities, but of *grouping* similar references in separate buckets on the basis of a growing set of logical criteria. From inspection of these groups, new or improved logical criteria can be formulated leading to finer distinctions on the next iteration. Ultimately, these criteria will grow into a separate dataset of their own, by means of which the data curation can be performed automatically. And when the reference groups have become a sufficiently fine partitions of entity references, a frequency cut-off can be declared—any non-grouped reference strings below the frequency cut-off remain unresolved—after which each group of reference strings represent a named entity.

In the last regard, the difference with the linear interpretative model of data curation is fundamental, even if the result may be very similar. Because the iterative automated process ‘zooms in’ on specific entities, it allows for the gradual discernment of boundaries between entities. These are not always clear in advance. Discerning between similar professions may be desirable when both occur frequently but unhelpful when both are rare. Changes over time may also warrant further divisions.

The iterative automated model has the advantage that identification and resolution are carried out simultaneously, allowing these finer distinctions to be made late in the curation process and only for those entities for which they would be useful. The consequence, on the other hand, is that the most common entities will be resolved with both the greatest certainty and precision, while the least-often occurring entities may not be resolved at all.

An additional advantage making the logical criteria explicit, is that it makes curation decisions transparent for criticism and improvement. At a later point in time, and outside of the scope of the REPUBLIC project, the raw NER output can easily be re-curated for a different research aim by adapting the logical criteria to suit the needs of a specific research question or project.

#### 5.4. Methods of reducing variation

The main product of the data curation method described in the previous section is a set of logical criteria for sorting references to entities into groups of references belonging to the same entity. Remember that even though their ultimate goal is to *distinguish*

entities, we do so by ‘bundling up’ disparate strings, that is to say: by reducing variation. Consequently, the first part of the curation consists of criteria for addressing sources of variation common to all entity types: spelling and language change, scribal mistakes, and text recognition errors. Some of these sources are more-or-less regular, while others amount to noise; these sources are hard to address together. After all, strategies for addressing text recognition errors by matching terms within a certain edit distance may suffer from different spellings being too far apart, while a spelling harmonisation rule may be obstructed by a single misrecognised letter.

Month	Expression	Variants	Keywords	Variants
January	januar[ijy]+	5	2	161
February	februar[ijy]+	6	2	186
March	m[ae]+rt[eijy]*	13	9	130
April	april+(is)?	4	3	105
May	m[ae][yij]+e?	14	10	14
June	jun[ijy]+(us)?	8	6	71
July	jul[ijy]+(us)?	6	5	56
August	august([ijy]+ us)	7	2	81
September	septemb(e?r? ris)	5	3	166
October	octob(e?r? ris)	4	3	107
November	no[uv]emb(e?r? ris)	7	4	149
December	decemb(e?r? ris)	4	3	149
Total		83	52	1375

Table 7: A two-stage method for recognising month names in date strings.

One approach we found fruitful is building up a lexicon of *intentional* variants of relevant keywords, and then perform a fuzzy (edit distance-based) match against those. See table 7 for an example of a lexicon built in a semi-automatic way: for every month of the year, a regular expression is constructed covering all common variants. Matches to these expression are sorted in descending order of occurrence, and then added to the lexicon if their edit distance to earlier-selected keywords exceeds some threshold. The resulting 52 keywords are matched against the corpus again with a lower edit-distance threshold, resulting in 1375 variants that each can be mapped to one of the months. Manually-created lexicons used in other entity types include lists of abbreviations, synonyms and vocabulary specific to the entity type. A small excerpt from the lexicon used for organisation (ORG) references are the three similar words ‘Bataillon’, ‘Bataillons’ and ‘Bataille’. Since a word can only be corrected to a single keyword, the first two are present to preserve the distinction between singular and plural, while the third prevents battles from being turned into battalions. (Battles are no organisations, but some are erroneously marked as such by the NER tagger. They are discarded by a later criterium.) By fuzzy matching words in entity references to entries in the lexicon, and replacing the matching words with the matching lexicon entry, the variation in reference strings is gradually reduced. For instance, “tweede Bathillon” becomes “tweede Bataillon” (EN: second Bataillon). Organisation references are relatively long, averaging 5.5 words or 35 characters; with the steps described up to this point, their variation can be reduced by a factor of five. And thanks to the iterative automated model described above, the lexicon used for this step can be improved at any later point without disturbing subsequent steps.

The next set of criteria can be constructed by taking the largest reference groupings and picking out keywords. The purpose of these keywords may of course be selecting entities directly, but can also be grouping entities together within a certain domain. Selecting entities directly is done for personal names (PER), which form a structurally homogeneous entity type. There, the regular expression `/[uv]?a?n? *olden?b[ea]rn\w*/` identifies a certain grand pensionary of the Holland: Van Oldenbarnevelt, and all misspellings of this frequent (and therefore oft hastily-written) name. Organisations, on the other hand, are an entity type of much greater diversity, and here entity references are first sorted in separate domains (see table 8). Most domains have a distinct structure and logic of their own. Subsequent criteria may then exploit domain knowledge or characteristics of formulaic language for making refinements within those groups that would be impossible on the total set of entity references. The largest domain consists of a single organisation, the Council of State, from which only some false positives had to be removed. The example of identifying seats of admiralties has already been mentioned above. A more complex case is formed by the regiments of the States’ army, which are usually referred to by their captain. Since the military rank of these captains may change over time, so does the name of the regiment. Hence, a targeted additional set of criteria must be applied for merging references to regiments together.

Category	Total	Unique	Category	Total	Unique
Council of State	139365	3635	Countries	8524	3591
Regions	139124	29191	Offices	5207	1941
Admiralties	97359	9228	Invalid	3016	1212
Princely courts	92174	12132	Mints	2063	774
Generality	58938	11051	Representatives	2000	596
Councils	45514	13405	Internal affairs	1232	755
Other	32648	16280	Secretariats	1128	518
Local governments	27958	12878	Diplomacy	780	218
Companies	26737	9376	Parliaments	667	234
Audit institutions	25984	1006	Advocates	521	51
Military	19258	12097	Congresses	105	62
Religious	11982	7739			

Table 8: Subdomains of entity references, with total and unique number of references. The unique number is after the normalisation steps described above.

Not all criteria involve lists of keywords. Some make use of the interdependence and nesting of different entity types (See subsection 3.1). Many personal names, for instance, contain names of towns. Yet we must not conclude that one of the towns named Rijswijk occur in a resolution on a petition by a merchant named ‘Theodorus Ryswijck’! Thanks to the nested structure of the different entity types, however, we can exclude altogether all place names that occur in personal names. The identification of local governments offers an example of the interaction of *three* different entity types. Local governments are sometimes addressed as such (‘the magistrate of Amsterdam’), but more commonly by the titles of their government offices: ‘burgomasters and regents of Amsterdam’, ‘syndic and council of Geneva’, ‘high bailiff and aldermen of Ghent’, etc. etc. In the latter cases, the reference may be resolved to the proper local government whenever a person attribution (ATT) belonging to *any* local government function co-occurs with its corresponding location (LOC) within the same organisation

(ORG) reference.

In the above examples, a great strength of the iterative automated model is that it uses different sources of expertise throughout the process of data curation. Curation strategies and criteria can be tailored to (parts of) different entity types in a straightforward way, and worked on in parallel: changes made to earlier criteria do not invalidate subsequent steps. Note that while entity references are bundled together (and sometimes split again), they are not *identified* with each-other until the very last moment. This means that the context of the individual references is not lost, but remains preserved for later criteria to use.

## 5.5. External identification

We used HisGIS<sup>17</sup> and Geonames<sup>18</sup> to link location entities to geographical coordinates. Of the 2,459 location entities, 2,327 could be mapped to coordinates, representing 70% of all location references.

We are still working on further identification for organisations and persons by using external databases, some of which are available at the Huygens Institute. There is a limited number of descriptions of institutions. There are several person databases available that may be used for external identification. The most important of these are the database of officeholders<sup>19</sup> and the overview of foreign representatives both to and from the Republic,<sup>20</sup> supplemented by more generally available overviews of European rulers and by indexes from a number of published relevant correspondences. These are a great help as both office holders and diplomats are central persons in the resolutions.

Of course, identification to these databases is somewhat more involved. Databases are not complete as they are compiled from other sources than the resolutions and based on their own criteria. Names as they appear in the databases are often spelt differently from the variations in the resolutions, and in the resolutions not all persons are designated by their full name. Some examples:

- in the 1620s Nicolaas van den Bouchorst was Lord of Noortwijk, and usually designated as Noortwijck, although in the 18th century Noortwijck was shorthand for Wigbold van der Does van Noortwijck or his son Wigbold jr.
- depending on the context, members from the prominent families like the Van Wassenaers were either simply designated as Wassenaer, or with an extension like Graaf (EN: Count) (van) Wassenaer, Van Wassenaer Twickelo, van Wassenaer Alkemade, Van Wassenaer Obdam, van Wassenaer van Duivenvoorde, Grave van Wassenaer tot Wassenaer, Van Wassenaer tot Warmond, Wassenaer van Sternburgh etc.

Fortunately, these prominent families are not very numerous because often these types of confusion can only be resolved by hand, if at all. The external databases provide dates of birth and death and dates of residence in an office that can be used, even if these are not always accurate, to merge or separate entity mentions. For prominent persons, death dates are often known, but birth dates are often missing, especially for

---

<sup>17</sup> A Dutch project to map historic placenames and geographical maps to geographic coordinates, <https://hisgis.nl>.

<sup>18</sup> <https://www.geonames.org>

<sup>19</sup> <https://resources.huygens.knaw.nl/repertoriumambtsdragersambtenaren1428-1861/>

<sup>20</sup> <https://resources.huygens.knaw.nl/retroboeken/schutte/>



Figure 7: The number of different committees active per year.

persons in the earliest period. In those cases, we used estimated birth dates based on a combination of 1) the date they started their office, and 2) the average age of people who started office. All together, it is usually possible to use semi-automatic methods to identify the most prominent persons and also to overcome the variety in spelling. However, a manual check for the most prominent families remains necessary.

## 6. Insights from Curation

We can use the curated entities to analyse the corpus in various ways. First, using knowledge of the domain and corpus, experts can do temporal analyses of when and how often certain types of entities occur as a sanity check. If the results are different from what is expected, this can be a signal that there are problems in the recognition process, or that the resolutions call for revising expectations. Second, different types of entities can now easily be combined in a quantitative analysis to investigate how the resolutions relate to certain historical events. One may expect that the number of resolutions that mention military organisations and persons increases in times of war and decreases once war is over, but this need not necessarily be the case. Third, entity references can be used to investigate previous historiographic claims.

For instance, the curation of the committee entities allows us to validate the claims that the States General made more and more use of committees from around 1650 to investigate matters submitted by petitioners and to prepare decisions (Thomassen, 2019a, p.162), and that from 1672, many ad hoc committees were subsumed under a smaller number of permanent committees, each related to a fixed topic (Thomassen, 2019a, pp.122-123), (Riemsdijk, 1885, pp.268-272).

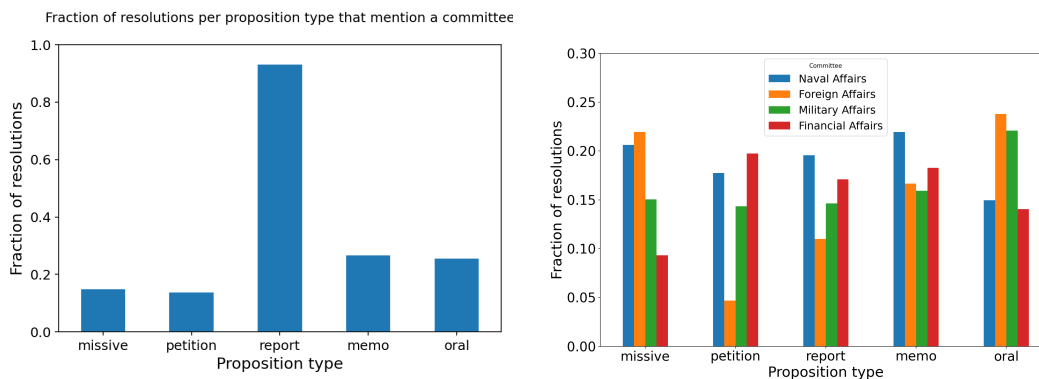


Figure 8: The fraction of resolutions that contain a reference to a committee reference per proposition type (left), and a break down of the committee mentioned in those resolutions (right).

The 97,860 committee references recognised by the COM tagger are mapped in the curation process to about 120 distinct committees. Figure 7 shows the distribution of the number of different committees active per year, and indicates a rapid increase in the number of committees starting from 1650. The number of active committees reaches its zenith of about 40 per year in 1655–1670, then decreases suddenly and stabilises to around 20 after 1672. Inspection of the number of resolutions that were passed on to each committee reveals that from the beginning of the 18th century, the vast majority of resolutions were handled by less than a dozen committees. This analysis corroborates the earlier findings and provides evidence that the output of the recognition and curation steps makes sense.

We can also combine the referenced committees with the *proposition type* of a resolution, which is the manner in which a proposition is presented to the States General. The proposition types were recognised in the resolution segmentation step, based on formulaic ways to introduce a new resolution, where each formula specifies the way in which a proposition is introduced (Koolen and Hoekstra, 2022; Koolen et al., 2020).<sup>21</sup> The five most common proposition types are *missives*, *petitions*, *reports*, *memos* and *oral presentations*.

The fraction of resolutions that contain a committee reference per proposition type is shown in Figure 8 on the left. Around 20% of resolutions based on missives, reports, memos and oral presentations make reference to a committee, but over 93% of all resolutions based on reports do. The high percentage of the latter is what we would expect, as reports are submitted by committees in response to an earlier resolution, where a committee was tasked to investigate and report back. These resolutions state which committee submitted the report. The fact that not all resolutions based on reports contain a recognised committee reference is most likely due to references missed by the tagger. In this case, the analysis suggests that the recognition and curation of committees leads to results that conform to expectations.

The right side of Figure 8 shows a breakdown of the committees that are referenced in those resolutions. Resolutions based on missives are more likely to reference the committees of Naval Affairs and Foreign Affairs (missives tend to be sent by Dutch diplomats abroad, or foreign diplomats in the Republic), while resolutions based on

<sup>21</sup> Together with the name and attribution of the person introducing the proposition, and the date and location of sending the document, all of which are mentioned at the beginning of each resolutions, these details form a stable referent to the submitted document that is also in the archive of the States General, and in that sense are a named entity reference.

petitions are more likely to reference the committee of Financial Affairs. One common group of petitioners are widows and orphaned children who ask the States General for financial support. In these cases, the person requesting support is mentioned in the resolution with an associated attribution, e.g. *widow* or *orphaned child*, which are tagged by the ATT tagger and (in the curation step) both categorised under *Legal Status and Relations*. This analysis is again not surprising, but increases our confidence in the quality of the curation process, and it shows the value of combining entities of different types and curation entity references into groups. By offering proposition types, committees and person attributions as search facets, users can easily select and drill down into subsets of the resolutions related to e.g. petitions mentioning family relations that involve the committee of Financial Affairs.

## 7. Conclusions

This paper describes the approach taken in the REPUBLIC project to operationalise named entities for information access in the corpus of resolutions of the States General of the Dutch Republic.

Our approach is informed by the purposes of highlighting entities in the text for easy document triage and utilisation as search facets for providing structured ways of making selections and drilling down into specific subsets of the resolutions. The entity types we chose are a mix of common types like persons, locations and organisations, and domain- and corpus-specific entity types such as committees, resolution references and person attributions.

Our first research question was:

- How well can we identify named entities in the resolutions?

We described how we have developed a large set of ground truth annotations for eight types of entities and performed a detailed analysis of these annotations. We found that there are strong differences between types in both the lengths of entity references and in the nature of nesting. References to committees and earlier resolutions tend to be long phrases, and many references to persons, attributions and organisations exhibit a high variance in length. The latter is related to nesting, as persons and their attributions, as well as organisations, often contain nested entities of different types. This has consequences for information access, as we need to make a choice about the level of nesting we use for creating search facets.

We compared several types of embeddings for historic Dutch text on their relative strengths for training NER taggers for different types of entities and found that Character embeddings, FastText embedding and the BERT-based contextual embeddings model GysBERT each have different strengths and weaknesses, and can be effectively combined to improve NER performance. We note that for all entity types, the best individual embedding type is close in performance to the best combination of embedding types, and that FastText embeddings outperform GysBERT on most entity types. We assume this is mostly due to the fact that the FastText embeddings were trained using the resolutions. Although we included fine-tuning of the original embedding weights as one of the options, allowing GysBERT to be fine-tuned to the NER training data, it may have been beneficial to first adapt GysBERT to the domain of resolutions by training on the full text of all resolutions.

Our second research question was:

- How can we curate entity mentions and make them useful for information access?

Next we described our approach to curating entity references, making them useful for information access. For this purpose, we have developed an integrated model of identifying and resolving entities, that makes curation decisions explicit, auditable, repeatable and reusable. Furthermore, it allows the incorporation of domain expertise in the entire curation process. This is particularly useful for large historic corpora for which few external resources of entity information exist and for which full manual curation is infeasible.

Our third research question was:

- What can we learn from the curation of entities about the corpus of resolutions and the operation of the States General?

To address this, we conducted some basic analysis of the curated committee references in relation to the resolutions they appear in. We found corroborating evidence of earlier claims as well as insights in how the SG used committees for handling the different types of documents and requests that were submitted to the SG, and on which the resolutions are based.

In future work, we want to conduct more detailed quality analysis on the output of the curation process, and on the influence of the individual curation steps. We also plan on doing more temporal analysis of how the references to types of entities change over time, both individually and in combination, in order to gain further insights in the workings of the SG, and in the extent the SG standardised their administrative and decision making processes.

## Acknowledgments

This research is funded by the Dutch Research Council (NWO) through the NWO Groot project REPUBLIC (an acronym for RESolutions PUBLISHED In a Computational Environment) 2019-2024 (NWO grant number 175.217.024).

This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-7089 and EINF-10206.

We would like to thank Femke Gordijn, who wrote the tagging instructions and liaised with the volunteers. We thank the volunteers for their invaluable contributions to this project, including the creation and correction of tens of thousands of transcriptions of the resolutions, and annotation the entities in the resolutions.

## References

- Nathalie Abadie, Edwin Carlinet, Joseph Chazalon, and Bertrand Duménieu. A benchmark of named entity recognition approaches in historical documents application to 19 th century french directories. In *International Workshop on Document Analysis Systems*, pages 445–460. Springer, 2022.
- Ran Abramitzky, Roy Mill, and Santiago Pérez. Linking individuals across historical sources: A fully automated approach. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(2):94–111, 2020. doi: 10.1080/01615440.2018.1543034.
- Ran Abramitzky, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. Automated linking of historical data. *Journal of Economic Literature*, 59(3): 865–918, 2021. doi: 10.1257/jel.20201599.

- Sergio Torres Aguilar, Xavier Tannier, and Pierre Chastang. Named entity recognition applied on a data base of medieval latin charters. the case of chartae burgundiae. In *3rd International Workshop on Computational History (HistoInformatics 2016)*, 2016.
- Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649, 2018.
- Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59, 2019.
- Martha J Bailey, Connor Cole, Morgan Henderson, and Catherine Massey. How well do automated linking methods perform? Lessons from us historical data. *Journal of Economic Literature*, 58(4):997–1044, 2020. doi: 10.1257/jel.20191526.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.
- Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, José Moreno, Nicolas Sidère, and Antoine Doucet. Robust named entity recognition and linking on historical multilingual documents. In *Conference and Labs of the Evaluation Forum (CLEF 2020)*, volume 2696, pages 1–17. CEUR-WS Working Notes, 2020.
- Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. Reden: named entity linking in digital literary editions using linked data sets. *Complex Systems Informatics and Modeling Quarterly*, 2016.
- Sebastian Colutto, Philip Kahle, Hackl Guenter, and Günter Mühlberger. Transkribus. a platform for automated text recognition and searching of historical documents. In *2019 15th International Conference on eScience (eScience)*, pages 463–466. IEEE, 2019.
- Ger Dijkstra, Nienke Groskamp, Rik Hoekstra, Marijn Koolen, Esger Renkema, Ronald Sluijter, Frank Smit, and Joris Oddens. Entities recognised in the resolutions of the states general of the dutch republic (1576-1796), May 2025. URL <https://doi.org/10.5281/zenodo.15495712>.
- 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*, 2004.
- Maud Ehrmann. *Named entities, from Linguistics to NLP: Theoretical status and disambiguation methods*. PhD thesis, Paris Diderot University., 2008. URL <https://hal.archives-ouvertes.fr/tel-01639190>.
- Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2):1–47, 2023.
- Pablo Ruiz Fabo and Thierry Poibeau. Mapping the bentham corpus: concept-based navigation. *Journal of Data Mining & Digital Humanities*, 2019.

- Sahar Ghannay, Cyril Grouin, and Thomas Lavergne. Experiments from limsi at the french named entity recognition coarse-grained task. In *Conference and Labs of the Evaluation Forum*, volume 2696, 2020.
- Imane Guellil, Antonio Garcia-Dominguez, Peter R Lewis, Shakeel Hussain, and Geoffrey Smith. Entity linking for english and other languages: a survey. *Knowledge and Information Systems*, pages 1–52, 2024.
- Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R Curran. Evaluating entity linking with wikipedia. *Artificial intelligence*, 194:130–150, 2013.
- Erkki Heino, Minna Tamper, Eetu Mäkelä, Petri Leskinen, Esko Ikkala, Jouni Tuominen, Mikko Koho, and Eero Hyvönen. Named entity linking in a complex domain: Case second world war history. In *Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings 1*, pages 120–133. Springer, 2017.
- Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.
- Eero Hyvönen, Laura Sinikallio, Petri Leskinen, Senka Drobac, Rafael Leal, Matti La Mela, Jouni Tuominen, Henna Poikkimäki, and Heikki Rantala. Publishing and using parliamentary linked data on the semantic web: Parliamentsampo system for parliament of finland. *Semantic Web*, 2024.
- Filip Ilievski, Piek Vossen, and Stefan Schlobach. Systematic study of long tail phenomena in entity linking. In *Proceedings of the 27th international conference on computational linguistics*, pages 664–674, 2018.
- 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.
- Frédéric Kaplan and Isabella Di Lenardo. Big data of the past. *Frontiers in Digital Humanities*, 4:12, 2017.
- Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9, 2018.
- Marijn Koolen and Rik Hoekstra. Detecting formulaic language use in historical administrative corpora. In Folgert Karsdorp and Kristoffer L. Nielbo, editors, *Proceedings of the Computational Humanities Research Conference 2022, CHR 2022, Antwerp, Belgium, December 12-14, 2022*, volume 3290 of *CEUR Workshop Proceedings*, pages 127–151. CEUR-WS.org, 2022. URL [http://ceur-ws.org/Vol-3290/long\\_paper5740.pdf](http://ceur-ws.org/Vol-3290/long_paper5740.pdf).
- Marijn Koolen, Rik Hoekstra, Ida Nijenhuis, Ronald Sluijter, Esther van Gelder, Rutger van Koert, Gijsjan Brouwer, and Hennie Brugman. Modelling resolutions of the dutch states general for digital historical research. In *COLCO*, pages 37–50, 2020.
- Marijn Koolen, Rik Hoekstra, Joris Oddens, and Ronald Sluijter. Formulas and decision-making: the case of the states general of the dutch republic. *Proceedings http://ceur-ws.org ISSN, 1613:0073*, 2023a.

- Marijn Koolen, Rik Hoekstra, Joris Oddens, Ronald Sluijter, Rutger Van Koert, Gijsjan Brouwer, and Hennie Brugman. The value of preexisting structures for digital access: Modelling the resolutions of the dutch states general. *ACM Journal on Computing and Cultural Heritage*, 16(1):1–24, 2023b.
- Marijn Koolen, Rik Hoekstra, Rutger van Koert, Ronald Sluijter, and Joris Oddens. paragraphs of the resolutions of the states general of the dutch republic (1576-1796), November 2025. URL <https://doi.org/10.5281/zenodo.15074656>.
- Elvys Linhares Pontes, Ahmed Hamdi, Nicolas Sidere, and Antoine Doucet. Impact of ocr quality on named entity linking. In *Digital Libraries at the Crossroads of Digital Information for the Future: 21st International Conference on Asia-Pacific Digital Libraries, ICADL 2019, Kuala Lumpur, Malaysia, November 4–7, 2019, Proceedings 21*, pages 102–115. Springer, 2019.
- Enrique Manjavacas and Lauren Fonteyn. Macberth: Development and evaluation of a historically pre-trained language model for english (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36, 2021.
- Enrique Manjavacas and Lauren Fonteyn. Non-parametric word sense disambiguation for historical languages. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 123–134, 2022.
- Cedric Möller, Jens Lehmann, and Ricardo Usbeck. Survey on english entity linking on wikidata: Datasets and approaches. *Semantic Web*, 13(6):925–966, 2022.
- Gary Munnelly and Séamus Lawless. Investigating entity linking in early english legal documents. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages 59–68, 2018.
- Siim Orasmaa, Kadri Muischnek, Kristjan Poska, and Anna Edela. Named entity recognition in estonian 19th century parish court records. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5304–5313, 2022.
- Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G Moreno, Emanuela Boros, Ahmed Hamdi, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. Entity linking for historical documents: challenges and solutions. In *Digital Libraries at Times of Massive Societal Transition: 22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020, Kyoto, Japan, November 30–December 1, 2020, Proceedings 22*, pages 215–231. Springer, 2020.
- Ismail Prada Ziegler. What’s in an entity? Exploring Nested Named Entity Recognition in the Historical Land Register of Basel (1400-1700)., June 2024. URL <https://doi.org/10.5281/zenodo.11500543>.
- Marie Puren, Fanny Lebreton, Aurélien Pellet, and Pierre Vernus. From parliamentary history to digital and computational history: a nlp-friendly tei model for historical parliamentary proceedings. *Digital Scholarship in the Humanities*, 40(Supplement\_1):i75–i86, 2025.
- Theodorus Helenus Franciscus Riemsdijk. *De griffie van hare hoog mogenden: bijdrage tot de skennis van het archief van de Staten-Generaal der Vereenigde Nederlanden*. M. Nijhoff, 1885.

- Danny Rodrigues Alves, Giovanni Colavizza, and Frédéric Kaplan. Deep reference mining from scholarly literature in the arts and humanities. *Frontiers in Research Metrics and Analytics*, page 21, 2018.
- Steven Ruggles, Catherine A Fitch, and Evan Roberts. Historical census record linkage. *Annual Review of Sociology*, 44:19–37, 2018. doi: 10.1146/annurev-soc-073117-041447.
- Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, 2014.
- Takashi Shibuya and Eduard Hovy. Nested named entity recognition via second-best sequence learning and decoding. *Transactions of the Association for Computational Linguistics*, 8:605–620, 2020.
- Ronald Sluijter, Rutger van Koert, Michael Baars, Marja Swüste, Michel van Gent, Esther van Gelder, Jesse Hollestelle, Ger Ruigrok, Ida Nijenhuis, and Joris Oddens. Republic pagexml ground truth handwritten resolutions states general, March 2023. URL <https://doi.org/10.5281/zenodo.7695131>.
- Rosa Stern, Benoît Sagot, and Frédéric Béchet. A joint named entity recognition and entity linking system. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 52–60, 2012.
- Melissa Terras. Digital humanities and digitized cultural heritage. *The Bloomsbury Handbook to the Digital Humanities*, page 255, 2022.
- Melissa M Terras. The rise of digitization: an overview. *Digitisation perspectives*, pages 1–20, 2011.
- Theo Thomassen. *Onderzoeksgids: Instrumenten van de macht: de Staten-Generaal en hun archieven 1576-1796 (Band 1)*. Sidestone Press, 2019a. ISBN 9789088908798.
- Theo Thomassen. *Onderzoeksgids: Instrumenten van de macht: de Staten-Generaal en hun archieven 1576-1796 (Band 2)*. Sidestone Press, 2019b. ISBN 9789088908828.
- Seth Van Hooland and Ruben Verborgh. *Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata*. Facet publishing, 2014.
- Rutger van Koert. Republic print dataset, May 2023. URL <https://doi.org/10.5281/zenodo.7928973>.
- Rutger van Koert, Stefan Klut, Tim Koornstra, Martijn Maas, and Luke Peters. Loghi: An end-to-end framework for making historical documents machine-readable. In *International Conference on Document Analysis and Recognition*, pages 73–88. Springer, 2024.
- Daniel van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *ICAART (1)*, pages 484–496, 2020.
- Yu Wang, Hanghang Tong, Ziyi Zhu, and Yun Li. Nested named entity recognition: a survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–29, 2022.

## A. Resolution Corpus

The corpus of resolutions is complex in nature because it contains both *ordinary* and *secret* resolutions, and for certain periods, there are multiple versions of extended handwritten resolutions. Furthermore, from 1672 printed versions of the resolutions were made and distributed to the individual states. However, until 1705, the printed versions did not contain all ordinary resolutions.

In the `REPUBLIC` project we made a selection of resolution books that covers all ordinary and secret resolutions for the entire period. For the ordinary resolutions, we use the handwritten versions for the period 1576-1702 and the printed versions of the *ordinary* resolutions of 1703-1796 (that is, the complete set of resolutions from 1705-1796, and incomplete for 1703-1704 because we did not have handwritten versions of those years). From 1592, the SG started a separate series of books for secret resolutions, which were never printed.

Table 9: The selection of resolutions used in the `REPUBLIC` project

Resolution type	Period	Format	Complete	Available for NER training
Ordinary	1576-1702	handwritten	yes	yes
	1703-1704	printed	no	no
	1705-1796	printed	yes	yes
Secret	1592-1796	handwritten	yes	no

The segmentation of the transcriptions into individual resolutions started with the printed resolutions. Because the handwritten resolutions were more challenging to segment, we had only managed to segment the ordinary resolutions of 1597-1702 by the time we had to start the NER ground truth annotation and training. Therefore, NER training data only covers the *ordinary* resolutions, and only of the periods 1597-1702 (handwritten) and 1705-1796 (printed). To clarify the selection of resolutions used in the `REPUBLIC` project and their availability for NER, a structured overview is given in Table 9.

### A.1. Document length

The 18th century printed resolutions and 17th century handwritten paragraphs that were annotated and used for training range considerably in length. The mean length is  $X$  words and the standard deviation is  $Y$ .

Some descriptive statistics, including various percentiles, are presented in Table 10.

The length of resolutions has a bi-modal distribution, that is, there are two peaks in the distribution. One peak is around 40-50 words, and consists of highly formulaic resolutions in which a missive from a diplomat or ambassador has been received that requires no decision. The other peak is around 250 words. The resolutions around this peak are the ones with a decision. In a sense then, the distribution of the length of

resolutions is a mixture of two distinct distributions of the two types of resolutions, those with a decision and those without.

## B. Code and data

The entity data is published on Zenodo: Dijkstra et al. (2025).

The `REPUBLIC` code base is available on GitHub. There are specific entry points to the code for NER and curation:

- Main: <https://github.com/huygensing/republic-project>
- NER: <https://github.com/huygensing/republic-project/tree/main/nlp/>
- Curation: <https://github.com/huygensing/republic-project/tree/main/enrichment>

The trained NER models are available on Huggingface:

- Persons: <https://huggingface.co/marijnkoolen/republic-ner-persons-2023>
- Person attributions: [https://huggingface.co/marijnkoolen/republic-ner-person\\_attributions-2023](https://huggingface.co/marijnkoolen/republic-ner-person_attributions-2023)
- Locations: <https://huggingface.co/marijnkoolen/republic-ner-locations-2023>
- Organisations: <https://huggingface.co/marijnkoolen/republic-ner-organisations-2023>
- Committees: <https://huggingface.co/marijnkoolen/republic-ner-committees-2023>
- Dates: <https://huggingface.co/marijnkoolen/republic-ner-dates-2023>

Handwritten	Paragraphs		
	Printed	All	
Count	514	1631	2145
Mean	97.9	227.2	196.2
Std	99.9	382.0	341.1
Min	1.0	24.0	1.0
1%	2.0	26.0	17.4
5%	20.0	28.0	27.2
25%	44.0	47.0	46.0
50%	73.0	127.0	107.0
75%	120.0	254.5	209.0
95%	233.4	752.5	650.0
99%	435.6	1677.3	1392.2
Max	1199.0	6269.0	6269.0

Table 10: Caption

Model	fine-tune	reprojection	use context	use RNN	use CRF
Persons	yes	yes	no	yes	yes
Person attributions	yes	yes	no	yes	yes
Locations	no	yes	no	yes	yes
Organisations	yes	yes	yes	yes	yes
Committees	yes	yes	no	yes	yes
Dates	no	yes	no	yes	yes
Resolution references	yes	yes	yes	yes	yes
Other names	no	no	no	yes	yes

Table 11: Configuration of the Flair training parameters used for the eight NER models.

- Resolution references: [https://huggingface.co/marijnkoelen/republic-ner-resolution\\_references-2023](https://huggingface.co/marijnkoelen/republic-ner-resolution_references-2023)

### B.1. Training parameters

We used the following training parameters. All models were trained for 10 epochs using a mini-batch size of 8 and a learning rate  $\lambda = 0.05$ .

The Flair library offers a range of parameters for configuring the training layers. The parameters can be on (yes) or off (no):

- *fine tuning*: fine tune the weights of the different types of embeddings (yes) or freeze them (no) during training.
- *reprojection*: reproject the embeddings to a linear layer before connecting to the output (yes) or not (no).
- *using context*: train the input with the document context (yes) or not (no).
- *use CRF*: use a CRF for output layer (yes, the units in the output layer are connected to each other and thus inform each other’s predictions), or a linear layer (no).
- *using RNN*: use an RNN-type (yes, default is LSTM) instead of a linear layer (no).