Accessing the Republic. Entity extraction from the resolutions of the Dutch States-General

Marijn Koolen[1,2], Esger Renkema[1,2], Nienke Groskamp[1], Frank Smit[1], Jirsi Reinders[1], Ronald Sluijter[1], Rik Hoekstra[1,2], and Joris Oddens[1]

[1]Huygens Institute, Amsterdam, Netherlands
[2]DHLab - 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 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 an estimated one million resolutions in total. Each resolution consists of at least two parts, a proposition and a decision (Thomassen, 2019).

Beyond making the text of the resolutions available and full-text searchable via OCR, HTR and structure extraction (Koolen et al., 2020, 2023), we want to offer additional access points for researchers to navigate and comprehend this large and complex resource. These access points will be based on several categories of named entities, including person names, institutions and geographical names and various domain- and collection-specific entity types.

Named Entity Recognition (NER) on historic documents has come along in leaps and bounds in the last decade (Ehrmann et al., 2023). This is 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 sequence tagging models and NLP frameworks (Akbik et al., 2019), and the development of language-specific Large Language Models (LLM) for historic languages (Manjavacas and Fonteyn, 2021, 2022).

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.

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?

2 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:

- Person (PER): a person identified by name and identiyfing 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): expliciet 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.

To train NER taggers, we created a ground truth dataset of 1631 full resolutions randomly sampled from the printed resolutions (1705-1796) and 513 paragraphs from resolutions of 1597-1704. The entities were manually tagged using INCEpTION Klie et al. (2018) (see Figure 1), with each resolution being tagged by three annotators and curated by a single curator to get consistent tags.

3 Training and Evaluating NER Taggers

We trained multiple NER taggers, one per entity type, to make it easy to deal with nested entity, e.g. a person entity containing an attribution, which contains an organisation, which includes a geographical location.

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 character embeddings 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 contextual embeddings model trained on historic Dutch.

We trained NER taggers using all possible combinations of embeddings and selected the best one per entity type. In line with earlier work (Boros et al. (2020), Ghannay

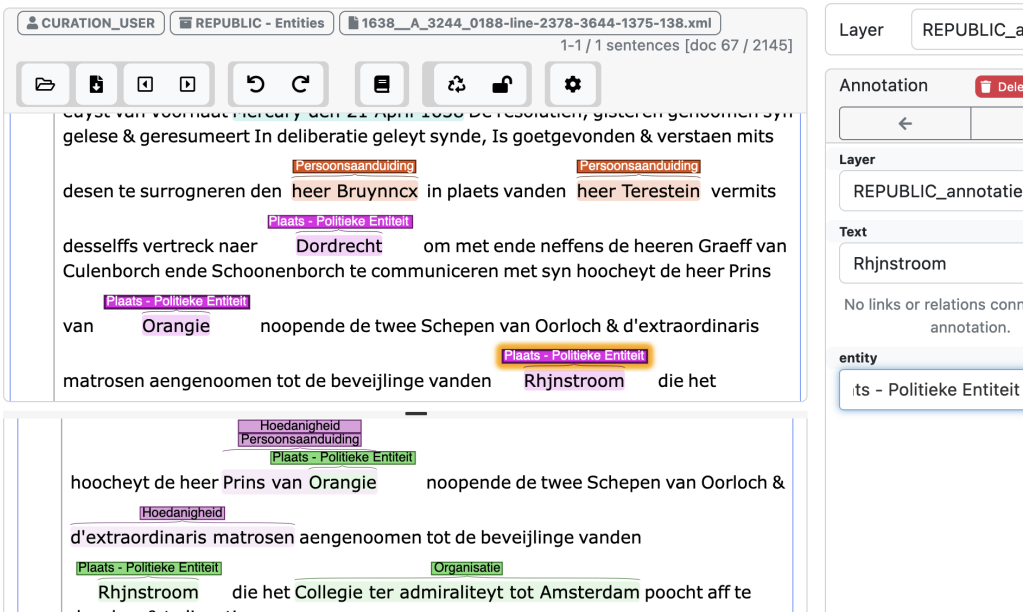

Figure 1: Tagging entities using INCEpTION.

Type	GT Layers	Embeddings	Prec.	Recall	F_1	support
PER	PER	Char, GysBERT	0.81	0.69	0.75	405.00
ATT	HOE	Char, FastText, GysBERT	0.57	0.56	0.56	573.00
COM	COM	Char	1.00	0.73	0.85	41.00
ORG	ORG	Char, FastText, GysBERT	0.82	0.71	0.76	283.00
LOC	LOC	FastText, GysBERT	0.79	0.76	0.77	570.00
DAT	DAT	Char	0.90	0.88	0.89	249.00
RES	All	FastText, GysBERT	0.82	0.70	0.75	57.00
OTH	All	Char, FastText, GysBERT	0.63	0.26	0.36	47.00

Table 1: Precision, recall and F_1 scores of the best trained models per entity type.

et al. (2020), Rodrigues Alves et al. (2018), Yang et al. (2018)), we find that all best models use RNNs instead of linear layers and a CRF for the prediction layer to capture dependencies between sequences of tags (introduced by Huang et al. (2015)).

The evaluation results for the best model per entity type are shown in Table 1. The GT Layers columns indicates whether the model was trained on only the annotations of the given entity type, or on the annotations of all types. The Embeddings columns indicates which embedding types were used.

For most entity types, the best tagger is trained on only the entities of that type. The two exceptions are resolution references (RES) and other names (OTH), which are better identified when the tagger also tags other entities.

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 Person attributions and Other names. Since we do not plan to use the latter category for information access, the low performance causes little problems. For attribtuions the difficulty is partly in correctly identifying the boundaries (leading to low recall) and partly in tagging too much, i.e. mentions of attributions without being an explicit referent (leading to low precision).

Entity type	Total tags	Distinct tags Exact	Regularised	% Reduction
Person	1,929,235	983542	949078	3%
Person attribution	1,743,086	763133	713275	6%
Organisation	743,860	187661	155481	17%
Committee	135,198	57969	37215	35%
Location	2,551,180	336402	310165	7%
Date	873,202	255044	230950	9%
Resolution reference	189,865	13285	9255	30%
Other names				
Total	8,165,626	2,597,036	2,405,419	7%

Table 2: Total and distinct number of tags per entity type over all resolution text. Regularised is after algorithmically normalising spelling. Reduction is based on regularised w.r.t. exact.

For six types, it is best to include the GysBERT model, which suggests they require some *understanding* of the context to identify entities. For Committees and Dates, using only a character-based embeddings model leads to better performance than including word-level or contextual embeddings, suggesting that they merely require *recognising* the right context, without getting distracted by deeper semantics.

4 Curating Entities

Running the NER taggers on all resolutions results in over 8 million entities. The number of recognised entities per type are shown in Table 2. The number of Person attributions is close to the number of Person names, which suggests that persons are almost always mentioned in a combination of their name and some attribution.

Though the details of processing every recognised entity class are highly specific to their respective domains, all share a common approach. The NER output is a list of potential *references to* entities. In order to identify these with the referred entities, we have to account for the various ways in which specific entities can be described. For this variation, we identify three causes. The first two are the variations in spelling and phrasing inherent to any large corpus spanning centuries, and the third cause is errors in the text recognition. We find that the interplay between recognition errors and orthographical variation poses a particularly tough challenge. This challenge must be overcome before we can address the variation in phrasing, which is the only of the three causes where complex domain-specific knowledge is required to identify different descriptions as referring to the same entities. Following this analysis, our method of identifying entities consists of three stages, each addressing a different cause of variation.

Of the three stages, the variation in spelling seems easiest to address, since, at least on a conceptual level, simple rewriting rules can be given to attain a normalised form for every word.

Many entity references have a formulaic structure, which allow effective fuzzy matching against a curated list of relevant keywords. Common elements of forms of address (e.g. 'majesty') are good candidates for inclusion in this list, as well as frequent domain-specific terms (e.g. *gedeputeerden*, 'delegates' or *admiraliteijt*, 'admiralty'). Of course, by restricting these keyword queries to the results of our NER tagger, we

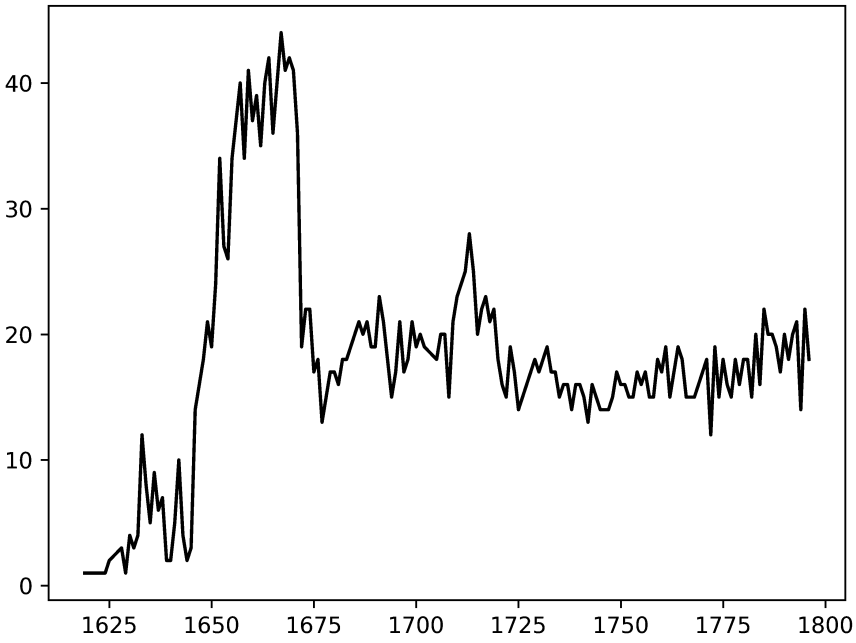

Figure 2: The number of different committees active per year.

are operating on information inaccessible to the text recognition at the time. Note that matching against external keywords, we also indirectly improve text recognition quality.

After the first two stages, the methods used for each entity class begin to diverge. The general approach, however is the following. First, all identical references (that is, exact matches) are grouped together (column 3 in Table 2). From this list, starting with the most-frequent entity descriptions, a manual curator selects keywords (mostly, these will be the identifying parts of the reference itself) and associates them with the referred entities. These keyword are queried against the list of entity references, and (near) matches will be removed. This process is repeated until a cutoff point is reached. Then, we are left with a definitive list of entities present in the text, each entity having one or more criteria to match against.

4.1 Insights from Curation

The curation of the committee entities allows us to validate the claims that committees were used more and more from around 1650 to investigate matters submitted by petitioners and prepare decisions (Thomassen, 2019, 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, 2019, pp.122-123), (Riemsdijk, 1885, pp.268-272).

The distribution of the number of different committees active per year is shown in Figure 3, and shows a rapid increase in the number of committees per year from 1650 to around 40 per year in 1655-1670, which decreases and stabilises to around 20 from 1672.

Figure 3: Extracting personal names proper from person entities: excluding the attribution leaves 'Heere van Borsselen vander Hooge'. From this we can infer that the location name 'Borsselen' does not in fact refer to the town itself.

By comparing the results of curation against earlier findings, we both sanity check the NER output and corroborate these earlier findings.

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).

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

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.

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.

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.

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.

Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

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, 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, 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, 2023.

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.

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.

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, 2019. ISBN 9789088908798.

Jie Yang, Shuailong Liang, and Yue Zhang. Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3879–3889, 2018.