The Value of Preexisting Structures for Digital Access: Modelling the Resolutions of the Dutch States General

MARIJN KOOLEN, RIK HOEKSTRA, JORIS ODDENS, and RONALD SLUIJTER,
Huygens Institute for the History of the Netherlands, Netherlands
RUTGER VAN KOERT, GIJSJAN BROUWER, and HENNIE BRUGMAN, KNAW Humanities Cluster - Department of Digital Infrastructure, Netherlands

The Resolutions of the Dutch States General (1576–1796) is an archive covering over two centuries of decision making and consists of a heterogeneous series of handwritten and printed documents. The archive, which has recently been digitised, is a rich source for historical research. However, owing to the archive's heterogeneity and dispersion of information, historians and other researchers find it hard to use the archive for their research.

In this article, we describe how we deal with the challenges of structuring and connecting the information in this archive. We focus on identifying the existing structural elements, to turn the archive from a set of pages into a set of meeting dates and individual resolutions, with rich metadata for each resolution. To deal with the challenges of historical language change, spelling variation, and text recognition mistakes, we exploit the repetitive nature of the language of the resolutions and use fuzzy string searching to identify structural elements by the formulaic expressions that signal their boundaries. We also discuss and provide an analysis of the value of extracting different types of entities from the text and argue that the choice of which types of entities to focus on should be made based on how they support relevant research questions and methods. In the resolutions, we choose to prioritise person qualifications such as profession, legal status, or title, over person names. Qualifications allow users to select certain groups of people and to meaningfully combine with other layers of metadata, whereas person names lack contextual information to disambiguate them, making it unclear which and how many persons are referred to by selecting a specific person name. We show how our methodology results in a computational platform that allows users to explore and analyse the archive through many connected layers of metadata.

CCS Concepts: • **Applied computing → Digital libraries and archives**; **Document metadata**; **Annotation;**

Additional Key Words and Phrases: Information extraction, digital history, text recognition, data modelling

ACM Reference format:
Marijn Koolen, Rik Hoekstra, Joris Oddens, Ronald Sluijter, Rutger van Koert, Gijsjan Brouwer, and Hennie Brugman. 2023. The Value of Preexisting Structures for Digital Access: Modelling the Resolutions of the Dutch States General. *ACM J. Comput. Cult. Herit.* 16, 1, Article 1 (June 2023), 24 pages.
https://doi.org/10.1145/3575864

This research is funded by the Dutch Research Council (NWO) through the NWO Groot project REPUBLIC (REsolutions PUBLIshed in a Computational environment) 2019–2024.

Authors' addresses: M. Koolen, R. Hoekstra, J. Oddens, and R. Sluijter, Huygens Institute for the History of the Netherlands, Amsterdam, Netherlands; emails: {marijn.koolen, rik.hoekstra}@di.huc.knaw.nl, {joris.oddens, ronald.sluijter}@huygens.knaw.nl; R. van Koert, G. Brouwer, and H. Brugman, KNAW HUmanities Cluster - Department of Digital Infrastructure, Amsterdam, Netherlands; emails: {rutger.van. koert, gijsjan.brouwer, henniebrugman}@di.huc.knaw.nl.

1556-4673/2023/06-ART1 $15.00
https://doi.org/10.1145/3575864

1 INTRODUCTION

The Resolutions of the Dutch States General (1576–1796) constitute an archival series that covers more than two centuries of continuous decision making. The archive consists of more than 500,000 pages of handwritten and printed resolutions, in separate, chronologically ordered series. The Resolutions of the **States General (SG)** in the Dutch Republic are a key resource to the political history of the period, as they contain all decisions made by the SG, the central ruling body of the Republic, from the 16th to 18th Century. Each resolution is a formal statement and record of the proposition submitted to the daily meetings of the SG and of the corresponding decision that was reached. The archive was designated as a key resource when in 1905 the work of publishing the resolutions started [24]. The manual editing resulted in two series of print publications of (a selection of) the resolutions—divided in an old series (14 volumes running from 1576–1609), a new series (7 volumes, 1610–1625)—and a digital edition (1626–1630).[1] The complete archive has recently been digitised, resulting in a set of around 250,000 scans. The resolutions reveal the decision making during and after the daily meetings of the SG, and are relevant for researchers interested in the politics of the Republic. Owing to their enormous richness, they allow researchers to answer many different research questions about politics—and more than that, see below—in the Dutch Republic and its position in the world. The resolutions are also a key to all the other records of the SG (taking about 1,500 m of shelf space) and form a hub to which these other records can be connected and with which they can be contextualised.

Many research questions require researchers to work through hundreds of large volumes of text without adequate indices. The relevant information is scattered across millions of paragraphs of dense and repetitive text (see Figure 1). Different research questions require different selections, reorganisations and re-orderings of the records to bring together and connect dispersed information. Many archives and libraries have experimented with giving access to their collections by means of digitised inventories and some have gone a step further, using existing indices of serial collections [9, 19, 25]. However, these archival referential systems are too coarse for access below the document level. Here, we note that the existing structure of the archive consists of many more reference systems and tools that can also be put to good use. Centuries of dealing with the complications of access to gradually expanding paper archives have led to a number of convenient and often-employed structures that are part of the printed culture—such as indices, registers and the use of different layout and fonts to help users locate elements of interest—but are often ignored in the translation to digital access [31, 44]. Therefore, it is crucial to extract high-quality metadata from the corpus of resolutions on various levels, including the meetings, dates, attendants, the individual resolutions and their topics. In this article, we describe the REPUBLIC project,[2] which aims to publish the resolutions in an online computational environment.

We focus on identifying and extracting several of these reference systems and structures as layers of data and metadata. These layers enhance access to all the resolutions in an online computational environment that supports a broad range of digital historical research. We combine established information extraction techniques with a workflow in which we iteratively build models of the structure of the corpus and of many of the standard phrases used in the resolutions. With these phrase models, we can exploit both expert knowledge and the fact that the resolutions contain a highly repetitive language to improve the extraction process.

Our approach starts from the important condition that the extracted information should reflect the structure of the resource and support a broad range of research questions. We illustrate this with a research problem for the following question: *Do the resolutions reflect changes in the petitioning of the SG by the citizens of the Republic over time?* To investigate such a question, we need different types of information, including (1) what types of proposals and requests were submitted that led to resolutions, (2) when each was put forward, (3) who submitted them, and (4) what decisions were reached. In addition, answering the question whether access to

[1] http://resources.huygens.knaw.nl/retroboeken/statengeneraal.

[2] REPUBLIC is an acronym for REsolutions PUBlished In a Computational Environment (2019–2023), a project funded by the Netherlands Organisation for Scientific Research (NWO). A prototype is available at https://demo.docere.diginfra.net/projects/republic.

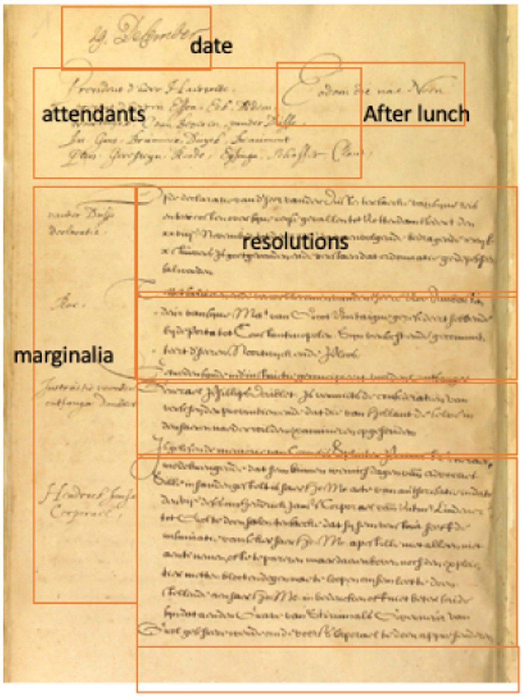

date
attendants
After lunch
resolutions
marginalia

Fig. 1. Structure of the handwritten resolutions. Image courtesy of Nationaal Archief.

the SG for the general public changed over time requires that the set of resolutions is either *complete* or *at least representative* [34]. Just digitising the archive without extracting information does *not* support answering this question. A digitised archive typically consists only of a combination of images and (roughly) digitised text—the so-called *physical* structure of pages, with an individual page or pairs of pages as a single image—and only very limited metadata. However, it is usually without any identification and operationalisation of the *logical* structure of the content, such as the chapters, resolutions, or notarial deeds that make up the logical organisation of the digitised archives.[3]

The logical organisation of the text—which does not align with the physical structure—includes the meaningful ordering of the content like temporal and geographical ordering, as well as templated text and repetitive textual characteristics of text that the authors and printers of the resolutions created to be able to easily find back information. Once we have identified these elements, we continue with **Named Entity Recognition (NER)**.

We prioritise the logical structure, because these elements can be operationalised as connected information layers and facets that allow meaningful navigation and selection. They can be identified with high accuracy but with relatively little effort, as we will demonstrate in this article. Moreover, by focusing on repetitive contextual structures in which person and place names are mentioned, we can reduce part of the challenge of recognising names by using the repetitive structures surrounding these names as templates with variable elements to be filled by person or place names. Generic NER on unstructured text written in historic language tends to have low accuracy [14, 15, 32] and results in long lists of mis-recognised names and names that mostly occur once or

[3] See, e.g., https://babel.hathitrust.org/ and our institution's resources, http://resources.huygens.knaw.nl. We note that in cases where the corpora are heterogeneous, there may be few or no logical and textual structures used consistently across the corpus, but there are usually coherent subsets that do.

twice, thereby providing little support as information access points. The biggest advantage of templates is that they provide contextual information on why a specific person name is mentioned, thereby increasing its value as information access point.

Many types of documents contain the structures mentioned above, such as notarial deeds, court files, ordinances, charters, missives, reports, procès-verbaux (authorised statements of acts or proceedings in the exercise of duty), but they also appear in early modern newspaper articles and advertisements [26, Ch.2]. The specifics of these structures tend to differ across corpora, so extracting and operationalising them require corpus specific approaches. Here, we argue that several generic techniques can be developed that can be easily adjusted to specific contexts.

The main challenge of our research is how to incorporate our knowledge of these structures into the recognition and extraction algorithms, in such a way that this expert knowledge can improve performance and lead to highly accurate information layers.

In the rest of the article, we elaborate on the techniques we used and developed, and we discuss how we deal with four challenges to transform a digitised archive into an online publication that supports the structured analysis required for historical research. These challenges are interconnected, as errors in text recognition influence the quality of information extraction, which in turn influences the accuracy of linking and the possibilities for users to interact with the information system [45]. In Section 2, we start with describing the corpus of resolutions and how its structure relates to different types of research questions that can be addressed by it. In this section, we also briefly describe our approach to text recognition. In Section 3, we provide a more detailed description and discussion of how we extract various structural elements. Connecting the various layers and linking them across different volumes of resolutions is described in Section 4. Finally, we draw conclusions in Section 5.

2 THE RESOLUTIONS AS A HISTORICAL RESOURCE

In this section, we start with a description of the content and structure of the archive (Section 2.1), then discuss what kinds of research questions the REPUBLIC platform should support (Section 2.2), and the application of text recognition (Section 2.3), and then how we have modelled the metadata structure (Section 2.4).

2.1 Structure of the Archive

The volumes of the resolutions embody a continuous series summarising and recording the decision made during the day-to-day meetings of the Dutch States General as the central assembly of delegates from (ultimately) seven sovereign provinces. The core elements are *sessions* and *resolutions*. Each *session* starts with a *date*, the *president* of the day, and a list of *attendants* ordered by province, followed by summaries of a varying number of *resolutions*. They are proposals and requests (see Figure 1) about a wide variety of subjects of both high and low politics, such as foreign policy, finance, army and navy, pensions and patents, administrative and cultural issues. All summarised issues consist of at least the following two parts, the *proposition* and the *conclusion* or decision.[4] The *proposition* refers to both the actual proposal submitted to the SG in oral or written form, and the written summary of it in the resolution [41, p. 188]. The *conclusion* contains a decision (resolution) of acceptance, rejection or deferral, pending further investigation or requests for information [37, 41, 42]. Both unresolved and decided issues led to trails of resolutions and all of them had to be traceable for the SG and other governing bodies, as the decisions had the force of law. Sometimes the SG wanted to know what had been decided previously in the same case or a similar case, and had the clerks refer to the archived decisions, called *retroacta* [21]. In addition, some summaries contain copies of letters or other important incoming documents as insertions.

The resolutions were summarised and archived by the *griffier* (EN: secretary) and read aloud for approval at the beginning of the next day's session. The *griffier* also indexed the resolutions, to support access to them at a later date. A relatively fixed set of index terms were used across the years, and per year an index was created

[4] Occasionally, resolutions included the advise given per province [37, pp. 18–19].

with references to resolutions of that year, as a means to find back those that were relevant as input for new resolutions. The nature and potential of these indices are discussed in Section 4.

2.2 Supporting Research Questions

The types of data layers we want to make operable are related to the types of research questions and themes that we want to support. Below, we discuss five types of analysis with examples of research questions:

Narrative analysis: A narrative analysis is conducted to offer an explanation of the temporal sequence of events [4]. For instance, the role of the SG in developments and shifts, both internationally and within Dutch society, and the causes and effects of these developments. E.g., how did the competition between navy and army develop, how did the SG deal with different religious groups? And what caused these developments?

Thematic analysis: Thematic analyses trace the development of certain themes/topics over time. For instance, do the resolutions over time reflect an increasing possibility for citizens to put forward their concerns to the SG? Do we see changes in the treatment of petitions and can they be used to analyse the SG's accessibility?

Content analysis: Content analysis is a quantitative analysis with regard to what or who was discussed, when and how often. For instance, how many resolutions deal with financial and economic policy, nominations for officeholders or army positions, or with petitions by citizens?

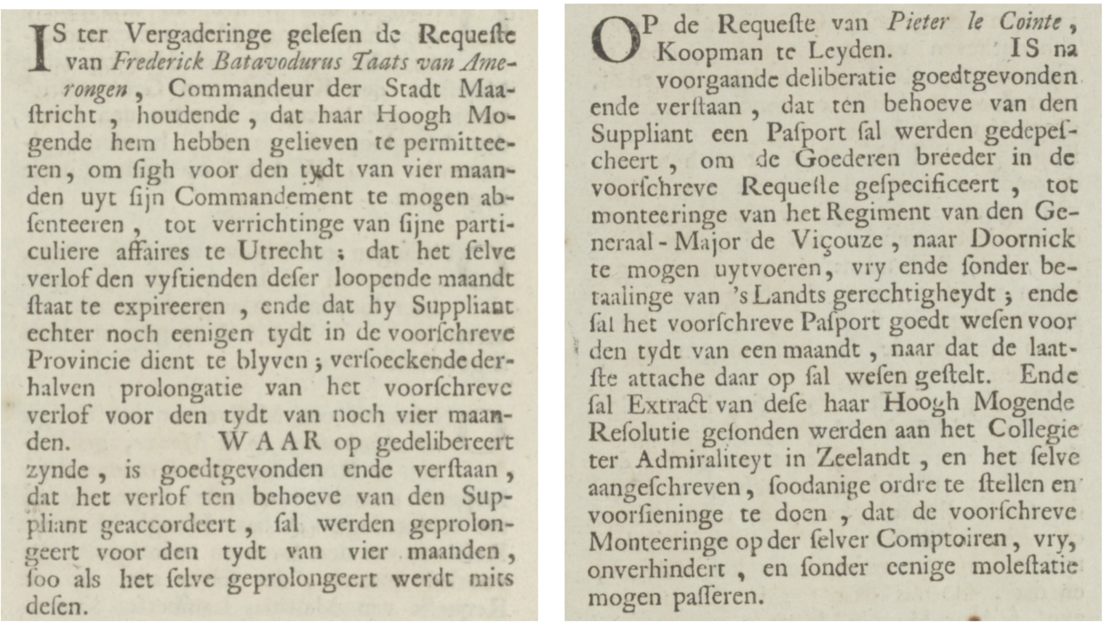
Network analysis: Network analyses focus on the relationships between actors and how they interact and influence each other. Performing serial research into the attendance at meetings and in committees can address questions such as: Can we identify where formal politics turns into informal politics, and politics behind the scenes? Who worked together? Who were involved in decisions around specific topics and in larger policy issues and how were these persons related?

Linguistic analysis: An analysis of the (development of the) language used in the SG. How did the language of decision-making develop and is it possible to link transformation of the SG's language with its growing administrative competence?

Addressing all these questions requires operationalising several elements from the logical structure of the written and printed texts, into multiple layers of metadata. The goal is to organise and classify the resolutions and make useful selections. For instance, to study changes in how petitions of citizens were treated, a researcher needs to select *resolutions* related to *petitions*, categorise them according to what group a *petitioner* belonged to and order them *temporally*. To study the network of actors who attended the meeting and those who were involved in committees requires operationalising the *attendance lists* and identifying the *committees* who were instigated following a *decision* by the SG, and who submitted *reports* that were the subject of later *resolutions*.

Next, we need to consider how these questions can be related to the structure of the archive and subsequently, how this can be translated to queries to the information system. Citizens of the Dutch Republic could put forward their concerns via petitions or requests (referred to as *"requesten"* in the resolutions). Each resolution based on a request states who submitted the request, at what date, from which location and what the main concern was about.

Requests that were discussed in the SG appear in the resolutions with one of two opening formulas, examples of which are shown in Figure 2. The one on the left contains the formula "IS ter Vergaderinge gelesen de Requeste van..." (EN: *was read the request of...*), followed by the name of the person who submitted the request, *Frederik Batavodorus Taats van Amerongen*, and a qualification or attribute, such as a title, occupation or their legal status. In this case, the proposer states he is "Commandeur der Stadt Maastricht" (EN: *commander of the city of Maastricht*). The request on the right uses the other formula, "OP de Requeste van..." (EN: *On the request of...*), followed again by the name of the person submitting the request, "Pieter le Cointe," and a qualification, namely, "Koopman" (EN: *merchant*), and the city where he operates, "Leyden." There are tens of thousands of



IS ter Vergaderinge geleſen de Requeſte van *Frederick Batavodurus Taats van Amerongen*, Commandeur der Stadt Maaſtricht, houdende, dat haar Hoogh Mogende hem hebben gelieven te permitteeren, om ſigh voor den tydt van vier maanden uyt ſijn Commandement te mogen abſenteeren, tot verrichtinge van ſijne particuliere affaires te Utrecht; dat het ſelve verlof den vyftienden deſer loopende maandt ſtaat te expireeren, ende dat hy Suppliant echter noch eenigen tydt in de voorſchreve Provincie dient te blyven; verſoeckende derhalven prolongatie van het voorſchreve verlof voor den tydt van noch vier maanden. WAAR op gedelibereert zynde, is goedtgevonden ende verſtaan, dat het verlof ten behoeve van den Suppliant geaccordeert, ſal werden geprolongeert voor den tydt van vier maanden, ſoo als het ſelve geprolongeert werdt mits deſen.

OP de Requeſte van *Pieter le Cointe*, Koopman te Leyden. IS na voorgaande deliberatie goedtgevonden ende verſtaan, dat ten behoeve van den Suppliant een Paſport ſal werden gedepeſcheert, om de Goederen breeder in de voorſchreve Requeſle geſpecificeert, tot monteeringe van het Regiment van den Generaal-Major de Viçouze, naar Doornick te mogen uytvoeren, vry ende ſonder betaalinge van 's Landts gerechtigheydt; ende ſal het voorſchreve Paſport goedt weſen voor den tydt van een maandt, naar dat de laatſte attache daar op ſal weſen geſtelt. Ende ſal Extract van deſe haar Hoogh Mogende Reſolutie geſonden werden aan het Collegie ter Admiraliteyt in Zeelandt, en het ſelve aangeſchreven, ſoodanige ordre te ſtellen en voorſieninge te doen, dat de voorſchreve Monteeringe op der ſelver Comptoiren, vry, onverhindert, en ſonder eenige moleſtatie mogen paſſeren.

Fig. 2. Two resolutions for requests submitted to the SG on 8 and 9 February 1725, respectively, with different opening formulas. Images courtesy of Nationaal Archief.

resolutions that use one of these two formulas, with some spelling variation and spelling changes over time. For example, from the second half of the 18th Century "geleesen" is used instead of "gelesen" (EN: *has been read*).

Adding metadata for these different aspects requires extracting the relevant information from an estimated one million resolutions, which requires an information extraction process that is automated where possible, but which needs to be informed by expert knowledge and a human in the loop. On top of that, each selection or reorganisation made by users creates a different view on the data, the interpretation of which is influenced by our decision in creating metadata layers, so it is important that the processing we do is transparent and visible to the user [22].

We return to the example question about requests in Sections 3 and 4 to discuss how the extraction and linking processes allow for these kinds of interactions.

2.3 Text Recognition

The physical state of the records is heterogeneous. The corpus has a mix of formats. All resolutions were written by hand, and in addition, the resolutions from the 18th Century were also published as printed volumes. The collection spans 220 years and has a large variety of handwriting, caused by differences between the successive clerks that were employed by the SG as well as by changes in fashions of handwriting in general, that changed from 16th Century Gothic to 18th Century roman script. Some printed volumes have single column pages, but others have double column pages and there are more complex column splits, insertions of letters and extracts, marginalia, tables (including multi-column and multi-page tables) and indices organised by main terms (referred to as *respecten* in the corpus). For our project, we use the clean copies of the handwritten resolutions made by the clerks until 1703, and the printed copies after that date.

The automatic recognition of the texts of the resolutions is performed by **Optical Character Recognition (OCR)** and **Handwritten Text Recognition (HTR)**. We use a typical pipeline consisting of Layout Analysis and detection of baselines of text in the images. During the project, we continuously update the OCR and HTR models using ground truth data sets and feedback from the information extraction process. The current OCR model achieves a **Character Error Rate (CER)** of 2, that is, 2 of every 100 characters are incorrectly recognised. At the level of words, the error rate is 6%. That is, 6 of every 100 words is not recognised correctly. HTR is more difficult because of the irregularities of handwriting. Recognising the 425,000 pages of handwritten resolutions requires several steps in a pipeline. We used the **P2PaLA (Page to PAGE Layout Analysis)** tool [36] on the scans without ground truth for layout analysis of text regions and text (base)lines. Next, we created a manual transcription of ground truth data set of 1,000 pages. Through iterative recognition and correction of batches of pages in Transkribus,[5] we currently achieve a CER of 2.99 on a 100 page evaluation set in which the identified text regions and baselines were manually corrected. The model is fine-tuned by corrections made by volunteers in the Vele Handen crowd-sourcing platform[6] that uses the web version of Transkribus.

2.4 Operationalising the Logical Structure

In many digitisation projects of, e.g., newspapers or books, the text of individual scans is recognised using a model trained on ground truth. After this process, metadata about the scan is added to organise the text of the scans. For newspapers these are typically the name of the newspaper, the date and the page number. For books, this is more difficult. The bibliographic metadata from a library catalogue can be added, but information on the structure of books is often not available as metadata. The text is accessible only as a sequence of plain text pages and maybe paragraphs, making it hard to figure out whether the book has chapters or ordinances or resolutions, where these start and end, and whether they have titles or headings.

For the volumes of resolutions, which are our case study in this article, text recognition results in one document per scan with the recognised words and their pixel coordinates. At this point, the text reflects the physical structure. To access the resolutions of a specific date, the recognised text gives you few handles to go to the right set of pages. Using string matching to locate dates is extremely error prone because of the combination of recognition errors, linguistic variation and frequent references to previous dates within a resolution. You have to select the book with resolutions for the desired year, then browse through the roughly 1,000 pages of dense text to identify the pages corresponding to the desired date. For systematic analysis across longer periods of time, this effort is multiplied.

To improve access, we want to use a more complex model of the resolutions that captures some of the elements of the logical structure of the paper archive that we think are the most helpful for users to navigate through and work with the digital version (what Herbert Stachowiak, as cited in Reference [34], calls the pragmatic property of models). We want to identify the date and start and end point of meetings, to label resolutions with those dates, so users can search and select resolutions by date or period. Knowing which text belongs to which resolution makes it possible to search for individual resolutions that contain multiple search terms. For instance, for the example research question on petitioning by citizens, when searching for resolutions that mention both petitions (NL: "Requeste") and citizens—who are mentioned using a range of terms and phrases, including: "Burger" and "Borger" (EN: citizen) and their plural forms—just having full-text search per page is of limited use. A single page typically contains multiple resolutions, so it can contain all search terms but spread across multiple resolutions. In these cases, the term occurrences are unrelated to each other. Moreover, a resolution can cross page boundaries, and have some of the search terms on the first page and other search terms on the page. The search terms *"requeste"* and at least one of the three terms related to citizens mentioned above result in 6,236 pages as hits, but in many of these, the search terms do not co-occur in individual resolutions (pages contain on average 3.5 resolutions). These page-level results also fail to capture relevant resolutions that contain one of the search terms on one page and other search terms on the next page. The effort of identifying the text boundaries of individual

[5] https://transkribus.eu/lite/.
[6] https://velehanden.nl.

resolution pays off, with the same search returning 5,560 resolutions that contain petition and citizen terms. This also allows scaling the analysis of resolutions. For instance, how many petitions were discussed on each day? How many were accepted, rejected or postponed? When or on how many dates were certain topics discussed? There are no standard NLP tools to help with this.

To enable the various types of research methods and questions, we extract and operationalise the following six elements of resolutions as metadata layers:

Meetings and meeting dates: The specific date on which a proposition was discussed and a decision was reached, including the day of the week.

Attendance lists and president: The persons who were present and involved in the decision making process of each resolution, as well as the person presiding over the meeting. The latter is especially important, since the president set the agenda together with the *griffier*, and thereby determined which propositions were discussed.

Resolutions: The text that belongs to a single resolution, as well as metadata on the proposition and decision, including the *type of proposition* that was submitted, e.g., a request, report or missive, *who* submitted the proposition, whether it was *accepted*, *rejected* or *postponed* for later discussion, and what *action* was decided on.

Insertions: Extracts of earlier resolutions or of resolutions by one of the Provincial States, or of memorandums, letters or requests submitted to the SG. In the case of earlier resolutions by the SG, identifying these offers a way to link trails of related resolutions. Insertions of resolutions by other organisations and other documents provide a starting point for linking the resolutions to other archives. At the moment, we have no concrete plans to do this, as many of these documents are not digitised yet, but we do aim to identify and categorise these references so they can be easily selected for analysis.

Named entities: Persons who submitted propositions and persons receiving instructions in the decision paragraph, committees selected to investigate and report on issues that were discussed, and other named entities such as organisations, geographic locations and ship names. We prioritise identifying person qualifications over proper names, as these are easier to identify with high accuracy, and offer more meaningful search facets. These qualifications offer ways to group person entities and link resolutions that are for instance based on propositions submitted by persons with the same qualification. Moreover, once we have identified them in the text, we expect that they help us spot the proper names with higher accuracy.

Index terms: The topic of individual resolutions. This is partly provided by the *contemporary indices*, *lists of index terms* and *marginalia*. The terms are a combination of person names, organisations, geographic locations and topics like finance or military and maritime matters. Some form of key phrase extraction or topic modelling could provide alternative (and differently biased) topical perspectives.

In the next section, we describe how we use this model to extract these different layers of information and that provides both handles for navigating the corpus as well as for analysing how information is distributed across the corpus.

3 INFORMATION EXTRACTION

A typical step in extracting information from historical texts is to use general approaches like NER, part-of-speech tagging, and lemmatisation to identify entities and topical words and phrases [30]. This step is thwarted by both text recognition errors and the lack of good NLP-resources for historical spelling and vocabulary in early modern corpora [14, 15, 20, 28, 32, 45].

On English texts these generic approaches work to some extent. English orthography has not changed much since the 18th Century, therefore resources for modern English can be effective [20, 45]. In Dutch, changes are larger, making generic approaches less useful. To investigate the value of generic NER techniques, we annotated named entities in 200 pages of manually transcribed resolutions and retrained the Spacy NER tagger[7] with 90% of the pages; the remaining 10% were used for testing. This led to a precision of 0.49 and recall of 0.19. Such low recall is typical in NER on historical documents [2, 10, 14, 38] Although performance is likely to improve by annotating more pages, there are two real hurdles. First, the upper bound for precision and recall remains low, because in the resolution texts, many nouns have uppercase initials, which makes it hard to algorithmically distinguish them from named entities. Second, precision and recall will be significantly lower on the vast majority of pages that are not manually but automatically transcribed.

In addition, such techniques do not alleviate the problem of identifying at least five aspects, viz. the start and end of daily sessions in the text, the precise date of each session, the attendance lists, the start and end of individual resolutions and the type of decision reached. Generic approaches of layout analysis can detect standard structures like tables, figures, footnotes, headers and tables of content with varying levels of success [8, 11, 35, 39], but they cannot interpret specific semantics such as temporal orderings of meeting dates and the geographical ordering in the attendants lists.

Therefore, we decided on an alternative approach that is based on a combination of (1) exploiting repetitive structural elements such as the layout and ordering used for indices and attendants lists, similar to Colavizza et al. [9], (2) explicitly modelling domain knowledge in lists of formulaic textual phrases, and (3) approximate string searching and matching. These are discussed below.

First, the sessions have a fixed structure and layout, with the opening of a next session and the attendance list represented in a different font and text alignment than the resolution summaries.

Second, the resolution texts are extremely repetitive, using the same phrasings with little variation across decades of meetings. As mentioned in Section 2.2, resolutions based on requests used one of two standard opening formulas. Formulaic phrases can be short or long, e.g., a single word or an entire sentence including punctuation. We use phrase models that contain lists of these fixed formulas and assign each phrase to a metadata category and one or more labels, in such a way that an approximate match in the text can be tied to a metadata layer. For instance, phrases for the opening sentence of a proposition have the label *proposition opening*, and the phrase "OP de Requeste van" (EN: *ON the request of...*) has the additional label of *proposition_type:request*. Phrases can also have known variants, e.g., alternative phrasings that we have encountered. The phrase models represent knowledge of the domain and the corpus, what information we expect to find, where, and in what order.

Third, our approach exploits the fact that, even with a relatively high *Word Error Rate*, the majority of characters in frequently occurring phrases and names are correct and in the right order, such that the string distance between the recognised text and its corresponding phrase in our model is small. We have developed a fuzzy searching algorithm that accepts one or more phrase models to find approximate matches, and uses configurable string distance thresholds to control how much textual variation is accepted.[8]

The phrase model for opening formulas contains 32 different formulas, each with a list of variant phrasings and a label for the type of proposition. For instance, the opening formula "Is ter vergaderinge gelesen de Requeste van" (EN: *Was read during the meeting the petition of*) indicates the proposition is a petition. For the resolutions of 1705–1796, 56,713 matches are found with 3,242 different OCR strings (Table 1). The most frequent OCR string occurs 13,767 times, and identifies only 24% of the resolutions with that formula found through fuzzy searching. Other proposition types include missives, reports, memos (diplomatic notes) and (previous) resolutions. The evaluation of this approach is described below.

We exploit domain knowledge differently across multiple iterations of information extraction. In the first iteration, we focus on extracting information with very high precision, by using high thresholds for approximate string searching, to build lists of, e.g., the starting point of meeting sessions and names of attendants. In later iterations, we use additional domain knowledge. For instance, if we have found the starting points of the meeting

[7] https://spacy.io/models/nl.

[8] See https://github.com/marijnkoolen/fuzzy-search.

Table 1. OCR Text String Matches Found for the Opening Formula "Is ter Vergadering gelesen de Requeste van" in the Resolutions for the Years 1705–1796

Text string found	Frequency	Fraction
IS ter Vergaderinge geleesen de Requeste van	13,767	0.24
IS ter Vergaderinge gelesen de Requeste van	10,854	0.19
Is ter Vergaderinge gelesen de Requeste van	5,942	0.10
IS ter Vergaderinge geleezen de Requeste van	2,956	0.05
ls ter Vergaderinge gelesen de Requeste van	2,396	0.04
15 ter Vergaderinge gelesen de Requeste van	1,299	0.02
18 ter Vergaderinge geleesen de Requeste van	1,273	0.02
IS ter Vergaderinge geleesen de Requestevan	1,177	0.02
15 ter Vergaderinge geleesen de Requeste van	1,171	0.02
3333 other variations	15,878	0.28
Total	56,713	1.00

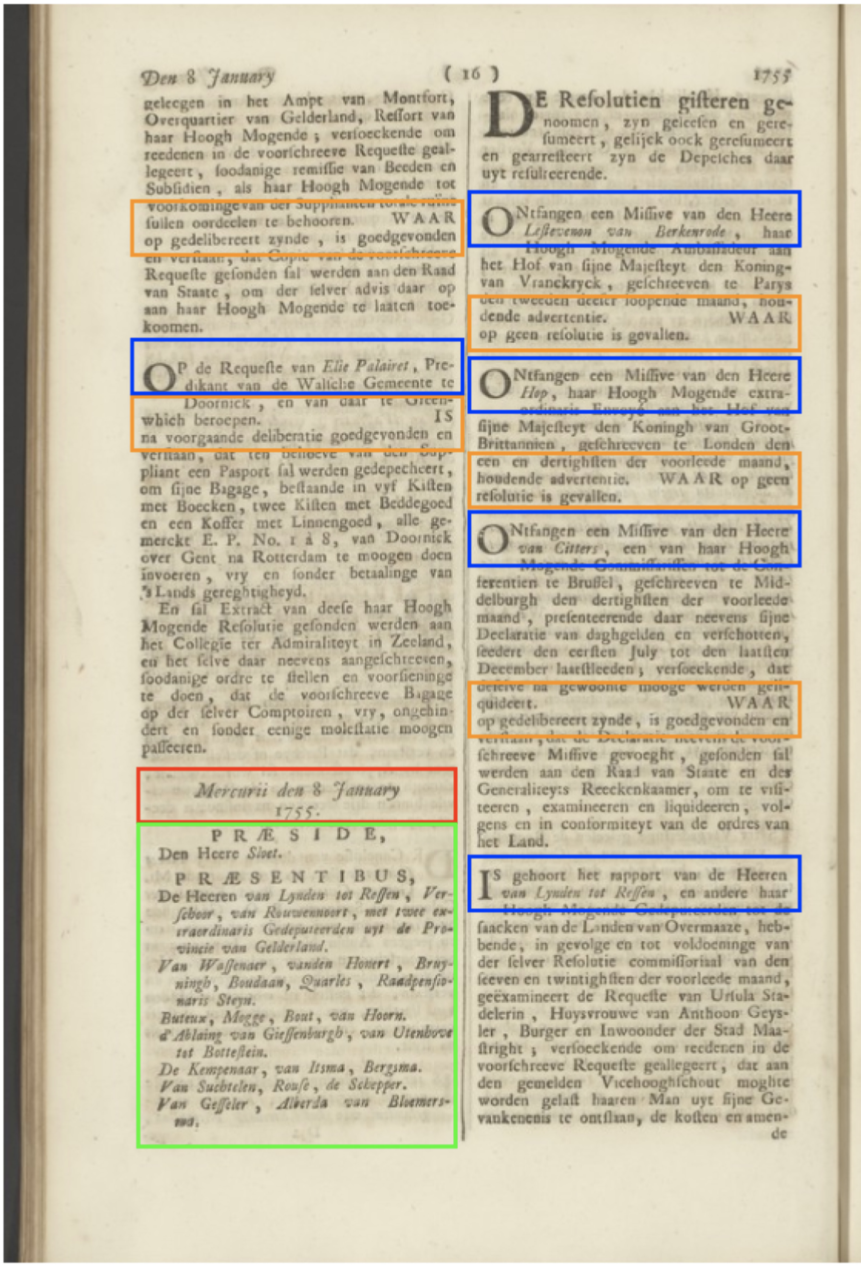
sessions for 12 and 15 January 1725 in the first iteration, then we exploit our knowledge that sessions are chronologically ordered. Therefore, we know that the sessions for 13 and 14 January should be in between these starting points. In the second iteration, we can search specifically for 13 and 14 January in a much smaller amount of text using lower fuzzy matching thresholds, with a much higher chance of success. For the names of attendants, we exploit our knowledge that the president of the meeting, being a representative of one of the provinces, rotated every week between the provinces, and that these persons were regular attendants during the other weeks. Once we know some of the names of presidents, we use approximate searching to find them in the lists either as president (taking care of recognition errors), or as regular attendants, reducing in this way the number of unknown and uncertain names in the list. For the individual resolutions, we use the textual formulas for the start of a proposition and of a decision (see Figure 3). For all elements that we extract, we store them with the fuzzy matches of phrases as evidence to explain how our metadata was created. Together with the explicit phrase models that we publish in our GitHub repository,[9] it makes the process of generating metadata transparent and repeatable.

As mentioned above, there are as yet no good NLP resources for syntactical analysis of historic Dutch, which limits the possibilities for research questions related to linguistic analysis (see Section 2.2). The text of the resolutions cannot be queried at the level of part-of-speech tags or lemmas. But with fuzzy search techniques—such as implemented in our fuzzy search algorithm but also in most modern search engines—it is possible to search for variants of words and phrases, and study the context in which they are used or track their usage over time.

3.1 Evaluation

Although all extraction steps require more manual effort than using generic, off-the-shelf text analysis tools, the gains are high. The reasons are: (1) Complex search queries can target resolutions instead of pages, (2) search facets can show how many resolutions match a certain facet, and (3) timelines can display how resolutions that match a query are distributed over time. Moreover, by extracting from the attendance lists the names of the president of the meeting and of the other attendants, users can select resolutions based on who, or which province, set the agenda for the day, who were present when a certain topic was discussed, and on which days of week certain types of propositions were discussed. All that stands or falls with the quality of the identification and extraction of these types of information. We therefore evaluated our approach with various ground truth data sets, along five identifications:

[9] See https://github.com/HuygensING/republic-project.



Den 8 *January* (16) 1755

geleegen in het Ampt van Montfort, Overquartier van Gelderland, Reſſort van haar Hoogh Mogende; verſoeckende om reedenen in de voorſchreeve Requeſte geallegeert, ſoodanige remiſſie van Beeden en Subſidien, als haar Hoogh Mogende tot voorkominge van der Supplianten totale ruïne ſullen oordeelen te behooren. WAAR op gedelibereert zynde, is goedgevonden en verſtaan, dat Copie van de voorſchreeve Requeſte geſonden ſal werden aan den Raad van Staate, om der ſelver advis daar op aan haar Hoogh Mogende te laaten toekoomen.

OP de Requeſte van *Elie Palairet*, Predikant van de Waliche Gemeente te Doornick, en van daar te Greenwhich beroepen. IS na voorgaande deliberatie goedgevonden en verſtaan, dat ten behoeve van den Suppliant een Pasport ſal werden gedepecheert, om ſijne Bagage, beſtaande in vyf Kiſten met Boecken, twee Kiſten met Beddegoed en een Koffer met Linnengoed, alle gemerckt E. P. No. 1 à 8, van Doornick over Gent na Rotterdam te moogen doen invoeren, vry en ſonder betaalinge van 's Lands gereghtigheyd.

En ſal Extract van deeſe haar Hoogh Mogende Reſolutie geſonden werden aan het Collegie ter Admiraliteyt in Zeeland, en het ſelve daar neevens aangeſchreeven, ſoodanige ordre te ſtellen en voorſieninge te doen, dat de voorſchreeve Bagage op der ſelver Comptoiren, vry, ongehindert en ſonder eenige moleſtatie moogen paſſeeren.

Mercurii den 8 *January* 1755.

PRÆSIDE,

Den Heere *Sloet*.

PRÆSENTIBUS,

De Heeren *van Lynden tot Reſſen*, *Verſchoor*, *van Rouwenoort*, *met twee extraordinaris Gedeputeerden uyt de Provincie van Gelderland*.

Van Waſſenaer, *vanden Honert*, *Bruyningh*, *Boudaan*, *Quarles*, *Raadpenſionaris Steyn*.

Buteux, *Mogge*, *Bout*, *van Hoorn*.

d'Ablaing van Gieſſenburgh, *van Utenhove tot Botteſtein*.

De Kempenaar, *van Itsma*, *Bergsma*.

Van Suchtelen, *Rouſe*, *de Schepper*.

Van Geſſeler, *Alberda van Bloemersma*.

DE Reſolutien giſteren genoomen, zyn geleeſen en gereſumeert, gelijck oock gereſumeert en gearreſteert zyn de Depeſches daar uyt reſulteerende.

ONtfangen een Miſſive van den Heere *Leſtevenon van Berkenrode*, haar Hoogh Mogende Ambaſſadeur aan het Hof van ſijne Majeſteyt den Koning van Vranckryck, geſchreeven te Parys den tweeden deeser loopende maand, houdende advertentie. WAAR op geen reſolutie is gevallen.

ONtfangen een Miſſive van den Heere *Hop*, haar Hoogh Mogende extraordinaris Envoyé aan het Hof van ſijne Majeſteyt den Koningh van Groot-Brittannien, geſchreeven te Londen den een en dertighſten der voorleede maand, houdende advertentie. WAAR op geen reſolutie is gevallen.

ONtfangen een Miſſive van den Heere *van Citters*, een van haar Hoogh Mogende Commiſſariſſen tot de Conferentien te Bruſſel, geſchreeven te Middelburgh den dertighſten der voorleede maand, preſenteerende daar neevens ſijne Declaratie van daghgelden en verſchotten, ſeedert den eerſten July tot den laatſten December laatſtleeden; verſoeckende, dat deſelve na gewoonte mooge werden geliquideert. WAAR op gedelibereert zynde, is goedgevonden en verſtaan, dat de Declaratie neevens de voorſchreeve Miſſive gevoeght, geſonden ſal werden aan den Raad van Staate en des Generaliteyts Reeckenkaamer, om te viſiteeren, examineeren en liquideeren, volgens en in conformiteyt van de ordres van het Land.

IS gehoort het rapport van de Heeren *van Lynden tot Reſſen*, en andere haar Hoogh Mogende Gedeputeerden tot de ſaacken van de Landen van Overmaaze, hebbende, in gevolge en tot voldoeninge van der ſelver Reſolutie commiſſoriaal van den ſeeven en twintighſten der voorleede maand, geëxamineert de Requeſte van Urſula Stadelerin, Huysvrouwe van Anthoon Geysler, Burger en Inwoonder der Stad Maaſtright; verſoeckende om reedenen in de voorſchreeve Requeſte geallegeert, dat aan den gemelden Vicehooghſchout moghte worden gelaſt haaren Man uyt ſijne Gevankenenis te ontſlaan, de koſten en amende

Fig. 3. Elements in the printed resolutions of 1755. A meeting starts with a date (red box) and attendants list (green box), followed by individual resolutions with opening propositions (blue boxes) and decision formulas (orange boxes). Original image courtesy of Nationaal Archief.

Page type identification: Identifying whether a page contains resolutions, index entries or *respecten* (lists of index terms), and whether a page is the title page of a section (and therefore the start of a section). Our model uses a combination of layout and textual evidence and has an accuracy of >0.99 on a test set of 3376 manually annotated pages. The printed volumes contain 91,302 pages with resolutions, 10,698 index pages and 828 pages with *respecten*.

Meeting date identification: Identifying the start of a meeting and the date of that meeting. We created ground truth data for 500 randomly selected meeting dates between 1705 and 1796, and annotated the starting point in the text as well as the exact date and day of the week. We evaluated and updated the phrase model in two iterations, using a batch of 100 meeting dates per iteration. Our current phrase model, after a third iteration of updating the phrase model, leads to a precision of 0.96 and recall of 0.99 on the test set consisting of the remaining 300 meetings. The extraction algorithm detects the correct start for 100% of the extracted meetings, but in 3 of 300 cases the identified date is incorrect. For several other dates in the test set, no meeting start is found. The 91,302 pages of printed resolutions are thereby transformed into a new layer consisting of 25,639 meetings. Of the meeting dates that were not correctly identified (7% of the total), the algorithm signals for 65% of them that they are appended to the previous meeting, so we know where to focus manual effort to correct them.

Resolution identification: Identifying the individual resolutions, including their opening proposition, the decision reached and the closing summary. On the ground truth test set of 311 resolutions in 90 randomly selected pages, our phrase model currently achieves a precision of 0.99 and recall of 0.93 in identifying the opening phrases. The relatively low recall signals that our phrase model is incomplete. In future iterations, we will create additional ground truth for testing, so as not to overfit our model on the initial ground truth data.

Attendants identification: Identify the names of the attendants and link recurring names to the correct entities. Current recall of attendant names is at 0.89, with precision at 1, due to the structured nature of the attendance lists. We are currently developing further consistency checks to link more attendants' names.

Index entry and reference identification: Identify the lemma and page reference of an entry, and link the lemma to the correct resolution. We have a first version of a model, but have not finished the ground truth data yet.

Large historical resources all have their own textual characteristics and structural features, which require the modelling of expert knowledge of these resources and incorporating these into generic NLP techniques. To know if this phrase model and fuzzy search approach generalises to other collections, we have experimented successfully with extraction of the dates and finding locations of medieval charters, such that over 17,000 extracted mentions of place names can be treated as historically dated attestations. Our goal is to continue to develop this open and reusable toolkit as an approach for structure-driven information extraction of digitised resources for historical research.

3.2 Named Entities, Document Types, and Person Qualifications

We take an unconventional route to (named) entity recognition and extraction. Again, we are guided by both the nature of the corpus and by the kinds of research questions we wish to support. The types of entities considered valuable to identify warrant some discussion. The resolutions mention document types that are the sources of propositions (missives, reports, requests, and letters) but also types of document that are requested or discussed, such as passports, placards, and pamphlets. These document types are important indicators of what type of proposition was submitted and what the resolution is about. Many resolutions mention passports to allow persons to travel, sometimes in combination with certain possessions or merchandise, to destinations within and outside the Dutch Republic. Sometimes the passport is requested, sometimes it is only mentioned in the decision paragraph. But its occurrence is a strong indicator that the resolution is about permitting access or passage. Although the term “Paspoort” is not typically considered a named entity, within this corpus and domain it is a valuable term for information access.

There are also many references to persons. On the one hand, we see many recurring terms for person qualifications or attributes, such as occupations (NL: “Koopman”; EN: *merchant*), legal status (NL: “Weduwe”; EN: *widow*), functions or roles (NL: “Ambassadeur”; EN: *ambassador*), that indicate what kind of person is mentioned. These kinds of terms are often not considered named entities, but they are useful attributes to understand for instance what type of person submitted a proposition to the SG. This is used in some fine-grained NER approaches, where person-role labels can be used to contextualise the entity [10]. However, these approaches typically use external knowledge bases as sources of knowledge about the entities [7, 16, 29, 47], whereas few of the persons in this corpus occur in any knowledge base.

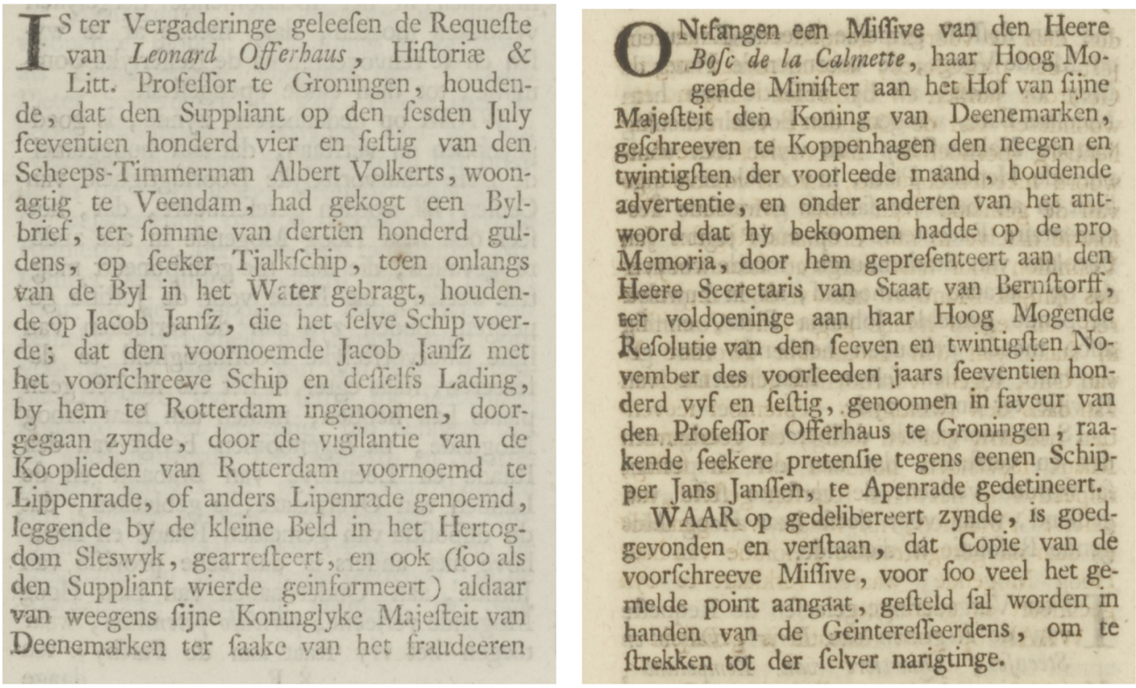
Finding terms for person qualifications is relatively straightforward using our fuzzy search module, as we can make term frequency lists from, e.g., the terms in the opening sentences of propositions, and build lists of a few hundred roles and titles in a matter of hours.

On the other hand, we see many person proper names. Person name is a standard type of named entity and is almost by default included in generic NER techniques [14]. However, person names provide little information about the nature of a resolution, unless it is the name of a well-studied person known to be involved in specific events or topics. But in the latter case, the name is probably included in the contemporary indices with page numbers referring to the pages with the relevant resolutions, and those resolutions have most likely been studied many times already. The vast majority of the estimated 250,000 distinct person names[10] occur only once or twice and refer to persons of whom little is known and which give no indication of why they are mentioned in the resolution, nor what it is about. As mentioned above, on such historic texts with OCR errors and spelling variation and change (including changes in the use of capitalisation), the quality of generic NER is poor [3, 14, 33, 40], resulting in long lists of names of which many are incorrect and from which most person names occurring the resolutions are missing. Providing person names as a search facet also creates a problem for presenting this to the users. The list of names is far too long to scroll through and almost all names lack context to provide users with useful information to select names from them. Finally, many persons have common names, so users are confronted with the problem of identifying which of the potentially dozens or hundreds of individuals is referred to.

By focusing on person qualifications and attributes first, we get much shorter lists of perhaps a few hundred or thousand qualifications, like *Professor*, *Metselaar* (EN: *brick layer*) and *Ambassadeur* (EN: *ambassador*), which can be curated with relatively little effort into a domain-related hierarchy (grouping qualifications related to artisans, maritime, military, church, etc.), which offers researchers a handle to navigate through these lists of roles [10]. Relating this to our example research question, by categorising qualifications hierarchically, a researcher can select resolutions based on propositions submitted by different groups of citizens. As mentioned in Section 2.4, there are 5,560 resolutions containing search terms for petition and citizen, but only 4,274 where the proposition document is a petition (in the others, a petition is mentioned, but is not the source of the proposition). In only 2,696 of these petitions does the citizen term refer to the proposer.

Identifying qualifications also allows us to improve the spotting and recognition of person proper names, as these names tend to occur just before or after such qualifications. The qualifications function as so-called *trigger words* that can be used as contextual features to improve NER [14]. Postponing the recognition of person proper names until after the recognition of qualifications avoids many mistakes where the qualification is recognised as part of the name. It has the advantage that the name can be connected to and contextualised by the qualification, thereby providing users more information about that person.

[10] Based on an extrapolation from the 1626–1630 resolutions that have been digitised earlier.



IS ter Vergaderinge geleeſen de Requeſte van *Leonard Offerhaus*, Hiſtoriæ & Litt. Profeſſor te Groningen, houdende, dat den Suppliant op den ſesden July ſeeventien honderd vier en ſeſtig van den Scheeps-Timmerman Albert Volkerts, woonagtig te Veendam, had gekogt een Bylbrief, ter ſomme van dertien honderd guldens, op ſeeker Tjalkſchip, toen onlangs van de Byl in het Water gebragt, houdende op Jacob Janſz, die het ſelve Schip voerde; dat den voornoemde Jacob Janſz met het voorſchreeve Schip en deſſelfs Lading, by hem te Rotterdam ingenoomen, doorgegaan zynde, door de vigilantie van de Kooplieden van Rotterdam voornoemd te Lippenrade, of anders Lipenrade genoemd, leggende by de kleine Beld in het Hertogdom Sleswyk, gearreſteert, en ook (ſoo als den Suppliant wierde geinformeert) aldaar van weegens ſijne Koninglyke Majeſteit van Deenemarken ter ſaake van het fraudeeren

ONtfangen een Miſſive van den Heere *Boſc de la Calmette*, haar Hoog Mogende Miniſter aan het Hof van ſijne Majeſteit den Koning van Deenemarken, geſchreeven te Koppenhagen den neegen en twintigſten der voorleede maand, houdende advertentie, en onder anderen van het antwoord dat hy bekoomen hadde op de pro Memoria, door hem gepreſenteert aan den Heere Secretaris van Staat van Bernſtorff, ter voldoeninge aan haar Hoog Mogende Reſolutie van den ſeeven en twintigſten November des voorleeden jaars ſeeventien honderd vyf en ſeſtig, genoomen in faveur van den Profeſſor Offerhaus te Groningen, raakende ſeekere pretenſie tegens eenen Schipper Jans Janſſen, te Apenrade gedetineert.

WAAR op gedelibereert zynde, is goedgevonden en verſtaan, dat Copie van de voorſchreeve Miſſive, voor ſoo veel het gemelde point aangaat, geſteld ſal worden in handen van de Geintereſſeerdens, om te ſtrekken tot der ſelver narigtinge.

Fig. 4. Fragment of a resolution from 22 November 1765 based on a request by a professor Leonard Offerhaus. Fragment of a resolution from 7 April 1766 based on a missive by a minister in Denmark, Bosc de la Calmette. Images courtesy of Nationaal Archief.

The same approach can be taken with geographic locations, which are contextualised through frequently used signal words and phrases like "woonende te" (EN: *living or residing in*), that tell us that what follows is a geographic location, as well as that the geographic location is the place of residence of the person mentioned just before the signal phrase "woonende te." Our approach is a classic example of template-based information extraction [18, 27, 46], except that the historic language use and OCR errors create hurdles to benefit from advanced syntax parsing that can be done on more recent texts with modern language use [3, 40, 43].

We illustrate the issue with person names through a fragment of a resolution mentioning three person names (the left side of Figure 4). This resolution fragment contains the person names *Leonard Offerhaus*, *Albert Volkerts*, and *Jacob Jansz*. Recognising these names and making them available as person names via a search facet in the graphical user interface is of limited value. Users who know these names can use full-text search to find mentions, but the vast majority of users do neither know who these names refer to, nor what kinds of resolutions they are involved in.

The name *Leonard Offerhaus* is listed in the contemporary index, with a reference to this resolution, but the other two names do not appear in the index. That is in itself an indication that *Albert Volkerts* and *Jacob Jansz* were not deemed of enough interest or relevance to back this resolution. It is perhaps also an indication that they are of limited value for current readers.

Of course, there is value in identifying whether the few occurrences of distinctive names are likely referring to the same person, such as the name *Leonard Offerhaus*, which appears, as far as we can identify, in three different resolutions. In two of these, in 1765 and 1766, Leonard Offerhaus is mentioned as a professor in Groningen. In the third resolution, in 1725, as professor in Lingen. Disambiguation is a problem regardless of whether we get this name from NER or from the indices. The resolution in April 1766 refers to the one from November 1765, so

Table 2. Distribution of the Different Types of Source (Documents) of the Propositions

Source type	# Resolutions	(%)
total	277,301	(100%)
Missive	160,816	(57%)
Request	75,634	(27%)
Report	12,958	(4%)
Oral	6,913	(2%)
Memo	6,259	(2%)
Resolution	5,394	(1%)
Unknown	3,664	(1%)
Other	5,663	(2%)

from the text, we can infer they refer to the same *Leonard Offerhaus*, but it is difficult to determine whether they also refer to the same person mentioned in the resolution from 1725.

3.3 Analysis of Extracted Information

To demonstrate how the extraction layers connect to each other to allow multi-faceted access to and analysis of the resolutions, we provide preliminary results for the 277,301 resolutions identified in the printed resolutions in the period 1705–1796. First, the opening formulas identify the type of source of the proposition, either a document that was submitted or an issue that was presented orally by one of the attendants. The distribution of proposition source types is shown in Table 2.

The majority of propositions come from missives (57%). These are documents sent by ambassadors and other diplomats reporting on the situation and events elsewhere. In some cases, these missives required no decision, in which case the resolution is short and closes with the phrase “WAAR op geen resolutie is gevallen” (EN: *ON WHICH no resolution was reached*).

The other main sources of propositions are requests or petitions (27%) and reports (4%). Over the 92-year period, 75,634 petitions were discussed, which corresponds to 822 petitions per year and almost 16 per week. Other types of sources include previous resolutions (by the SG or by one of the seven individual states or provinces), memos, and orally presented propositions, bills, advisory notes, and passports.

3.3.1 Fine-grained Analysis. Since we know the exact dates of each resolution, we can do more fine-grained analyses of these types of proposition sources. In Figure 5 the number of propositions of the three main types per year is shown.

The blue line at the top shows that the number of resolutions per year, regardless of the type of proposition source, was between 6,000 and 7,000 in the beginning of the 18th Century, but after about 1715 it dropped to around 3,000 per year. A breakdown over proposition sources shows that particularly the number of requests dropped over time. One possible explanation is that the length of the daily meetings decreased substantially, from 4–6 h in the 16th and early 17th Century to 2 h in the second half of the 17th Century, and from early in 18th Century to 1 h. With less time per session, it could be that fewer propositions were discussed. Another possible explanation is that the meetings were increasingly standardised, focusing the meeting mainly on propositions dealing with standard matters and standard decisions, while moving much of the work on other propositions to committees that reported to the SG, and making important decisions in secret [41, p. 183]. So the drop in numbers of resolutions per year fits with and corroborates what is known from other sources.

Next, we analyse the extraction of proposition type. It allows us to compare the number of missives and petitions discussed on each of the weekdays (Figure 6). In the first half of the 18th Century, the SG met six

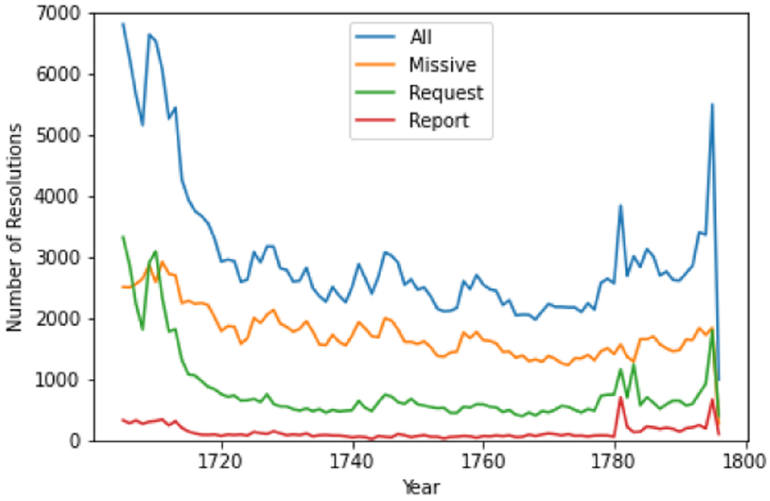


Fig. 5. Distribution of the number of resolutions per year, overall, and for the three most common proposition document types.

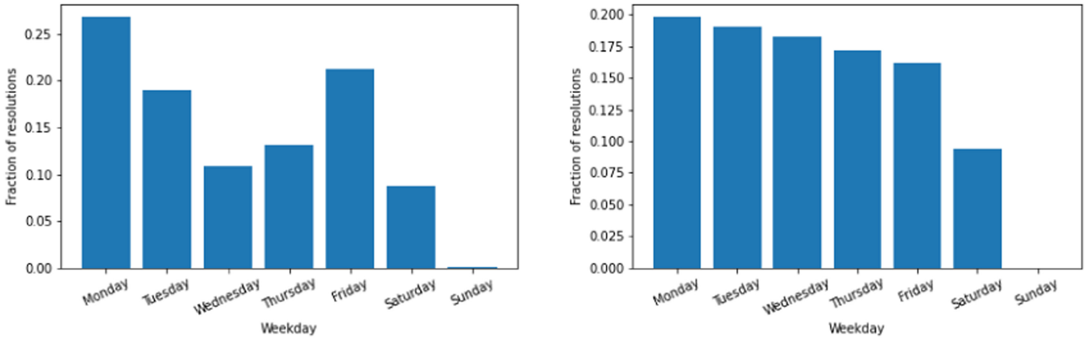


Fig. 6. Distribution of the number of missives (left) and requests (right) across the different days of the week.

days a week (Monday to Saturday), but from the 1754 the meetings were only five days a week (with only occasional meetings on Saturdays or Sundays in case of emergencies). There are clear differences between the two types of proposition sources. Many more missives were discussed on Mondays, Tuesdays, and Fridays than on Wednesdays and Thursdays. It is not immediately clear why this would be the case. A manual check revealed that this is not an error in assigning the right date to resolutions. One speculative explanation is that the easy-to-handle missives were discussed early in the week, after which the ones concerning more complex issues were discussed, for which the president of that week wanted to wrap things up by the end of the week, resulting in more decisions reached on missives on the Friday. The distribution of propositions handled each day of the week shows a steady but small decline throughout the week. It is possible that petitions built up over the Saturday and Sunday, in such a way that more were available early in the week. Or that, in line with the speculative explanation above, petitions were deemed standard matters that were preferably handled early in the week, leaving more time in the rest of the week for more complex issues.

We can also see what person roles are associated with the people who submitted requests (Figure 7). The average number of propositions submitted in each of the 12 months (aggregated over the entire 92-year period)

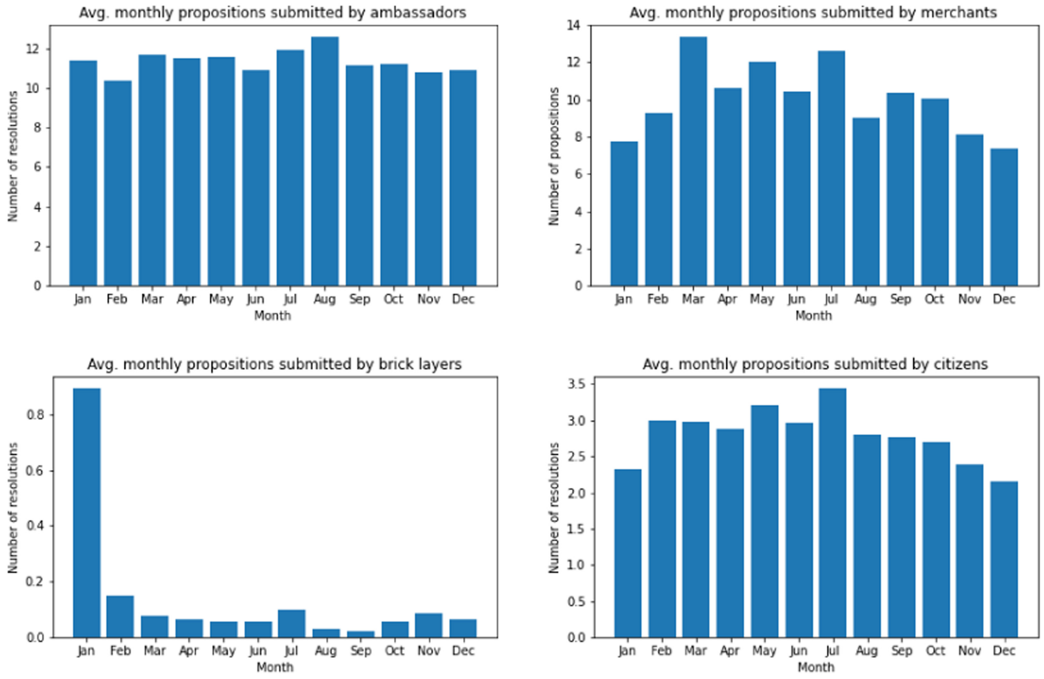


Fig. 7. Distribution of the number of requests submitted in each month by ambassadors (left), merchants (middle), and brick layers (right).

by different groups is shown, with, respectively, ambassadors (top left), merchants (top right), bricklayers (bottom left), and proposers referred to as citizens[11] (bottom right). Two things jump out from their comparison.

First, ambassadors and merchants submit roughly equal numbers of propositions to the SG (around 11 or 12 per month), but many more than brick layers and citizens. For ambassadors this is not surprising given the nature of their roles. One of the tasks of ambassadors representing other sovereign bodies was sending missives to the SG to keep them informed and request a response regarding matters in which the involvement of the Dutch Republic was deemed necessary. The number of propositions they sent per months is stable between 10, suggesting that such requests were equally likely throughout the year.

Second, the numbers of propositions submitted by merchants, brick layers, and citizens differs more strongly per month. Merchants and citizens submitted more propositions in early spring to early summer than in the winter months. Brick layers almost exclusively submit propositions in January. Close reading selections of their requests from various months makes clear that in January, the brick layers who were commissioned by the SG submitted a petition for a new year's gift in early January, or the settling of bills accumulated over the previous year.

Focusing on the person qualifications and other entity types allows us to group resolutions in a meaningful way. This offers handles for various research questions, including our example question about the petitioning

[11] We use the terms "Burger," "Borger," "Inwoonder" and their plurals "Burgers," "Borgers," "Inwoonders" as qualifications indicating that someone is a citizen. We admit that this is too simplistic for properly addressing the example research question—many citizens are mentioned only by occupation, such as brick layers, so citizenry could be modelled by excluding nobility, administrators and diplomats, and "inwoonder" was used to refer to people residing in a place without the legal status of citizen—but, we use this shortcut merely to illustrate the potential of operationalising qualifications.

by citizens, but also for questions related to network analysis. One can make networks of (groups of) persons involved in resolutions (proposers, attendants, presidents, and persons mentioned in the proposition and the decision parts) and to capture some of their interactions.

Similarly, identifying the geographic locations from which proposers submitted their documents allows for selecting resolutions by place of origin. Note that here the identification of a place name or proposer is contextualised by our knowledge that they appear in syntactically interpretable parts of the opening formulas (despite the language variation and change and the OCR/HTR errors). So, we can label them as *person name* and *place name*, and more specifically as the *proposer* and *place of origin* of a proposition. We can then link them to the content of the resolution, as well as to many other characteristics, such as the type of proposition document, the *date* of the meeting, the *attendants* and the *president* (and thereby the province they represent) of the meeting. This procedure creates valuable, connected information layers that support the types of research questions discussed in Section 2.

Making the same kinds of aggregations and temporal distributions with person names instead of qualifications is problematic for two reasons. The first reason is surmountable with significant effort: the accuracy of NER results for person names is much lower than the accuracy of the identification of qualifications, but could, at least in theory, be increased by creating more training data and manually correcting errors. The second reason, which points to a more fundamental problem, is that for the vast majority of person names, there is no data we can use to disambiguate individual persons. We can aggregate resolutions by person name, but we cannot know how many persons are included in the aggregation, or which name variants should be included because they refer to the same person. As a result, if we make aggregations by person name, then we do not know which or how many persons are included in the aggregations.

4 LINKING INFORMATION ACROSS RESOLUTIONS

Information extraction gives us a way to navigate, select and order the meetings and resolutions in individual volumes through the different layers of metadata. The next step is linking information. There are three types of links. The first type of link is obvious, as the metadata layers are connected to the resolutions so can be connected to each other via the resolutions. This step enables queries such as (a) when was financing of the military discussed, (b) who were involved in the decisions made around this topic, and (c) what kind of decisions were reached. Here, connections between the topics from the indices and the resolutions they refer to, as well as to the correct dates of the resolutions are indispensable. The second type of link connects metadata elements within a single layer across different years and meetings found in different volumes. For instance, to enable a good layer of topical metadata, we need to connect the indices and marginalia across the entire period, so that all resolutions regarding financing of the military can be retrieved for all 220 years. The marginalia and contemporary indices were made by different people at different times, resulting in an incoherent system over the years. Marginalia differ in level of extensiveness, indices in level of completeness, and their individual terms in level of scope and interpretation. This makes connecting them into a coherent layer of information to access the entire archive a challenge.

These metadata layers can be seen as creating links between sets of resolutions. Within most of these layers, the sets of related resolutions have no inherent order apart from chronology. The metadata labels merely group them.

A third type of link is represented by causal links between many resolutions. These links will be harder to operationalise. Frequently, the decision on a proposition called for a committee to investigate and report back on an issue, or for dispatching a letter to another political entity that required a reply. Once the report or reply came in and was discussed in the SG, this lead to another resolution, creating a causal chain of resolutions. We will investigate these causal links in future work, and focus on the grouping links in the rest of this section. An example of causally linked resolutions is shown in Figure 4. The proposition of the resolution on the left is a petition by Leonard Offerhaus stating he had been defrauded of his recently bought ship by Jacob Jansz. who had been arrested in Denmark. The decision paragraph of the resolution on the left (not visible in the fragment) states that a copy of the petition will be sent to Bosc de la Calmette, Minister at the Court of the King of Denmark. The resolution on the right contains a missive in which Bosc de la Calmette replies to the petition by Offerhaus.

A fourth set of links is to external sources. The resolutions explicitly mention all manner of documents, like the missives, requests, memos, bills, and reports that led to the propositions, as well as reports and letters commissioned in the decision paragraphs. Most of these have been archived as well, in a separate archive called the Appendices. At the moment, there is no explicit link to a document submitted to the SG, and the resolution in which it is discussed. But through our analysis of the opening formula of the propositions, we have identified the type of source document for that resolution, the name of the proposer and their qualification, when they sent it and where from. Together, these offer many data points to identify them in the archive of Appendices, if and when these are digitised.

In the rest of this section, we focus on the indices. We conducted two analyses to establish which types of information layers can be extracted from the indices (Section 4.1), what possibilities they offer for linking and grouping resolutions, and thereby, which layers should be prioritised in the project (Section 4.2).

4.1 Types of Index Terms and Temporal Stability

In the first analysis, we manually transcribed and merged the *respecten* or lists of index terms for seven volumes of resolutions between 1742 and 1785. Together these contain 3934 distinct index terms of lemmas, with individual volumes containing between 673 and 994 terms. The index terms include names of persons and organisations (e.g., the *admiralty of Zeeland*), place names (*England*, *Amsterdam*, *Gelre*), other entity types (*letters*, *convoys*, *declarations*, person qualifications like *captain general* or *merchants*) and topics (*commerce*, *finance*, *infectious disease*, *sugar*). The majority of terms are person names (64%), with place names as the second largest group (19%) and the remaining categories together covering 15% of the terms. Although the person names are the largest categories, they are the least stable. Among the 243 terms that occur in at least four of seven volumes, only 36% are names but 43% are place names and 21% of terms fall within the remaining categories. Names of places and organisations as well as topics have a higher overlap between years and are more stable across time than person names. Individuals are only mentioned in the index when they were deemed relevant to finding back a certain resolution or set of resolutions. However, they are only mentioned by surname, which causes certain family names to occur in many indices as children of people in relevant positions often ended up in the same or other important positions. This makes it difficult to identify the individuals who are referred to. The overlap of entries between subsequent years is around 33%, but drops to 10–20% for indices that are decades apart. This is mainly caused by the individual person entries, which may recur in subsequent years but disappear from the index over time. Yet subject terms also change, which creates a challenge for longitudinal analysis of topics such as long-term developments of economic policies or shifts in the accessibility of the meetings and decisions for different classes in society.

The observations above suggest that there are two challenges in creating a single index across the whole series. One is to link recurring terms that might have different spellings or OCR representations, the other is to link orthographically different but semantically similar terms. For the former, we use our fuzzy search strategy. For the latter, some combination of manual categorisation and automatic clustering will be used.

4.2 Analysing Linking Potential of Index Terms

For the second analysis, we started from the manually transcribed indices of the resolutions of 1610–1625 and created a structured version, with index terms and the pages they reference in the digital version of the resolutions. We categorised the terms under one of six categories: *person name*, *geographical name*, *person qualification*, *institute*, *topic*, and *other*. The *other* category contains terms for various entity types like “Brieven aan/van” (EN: letters to/from), contributions, declarations, as well as terms that are hard to categorise. The *person qualifications* in the index are often tied to specific place names or organisations, like ‘Raad van Brabant’ (EN: council of the

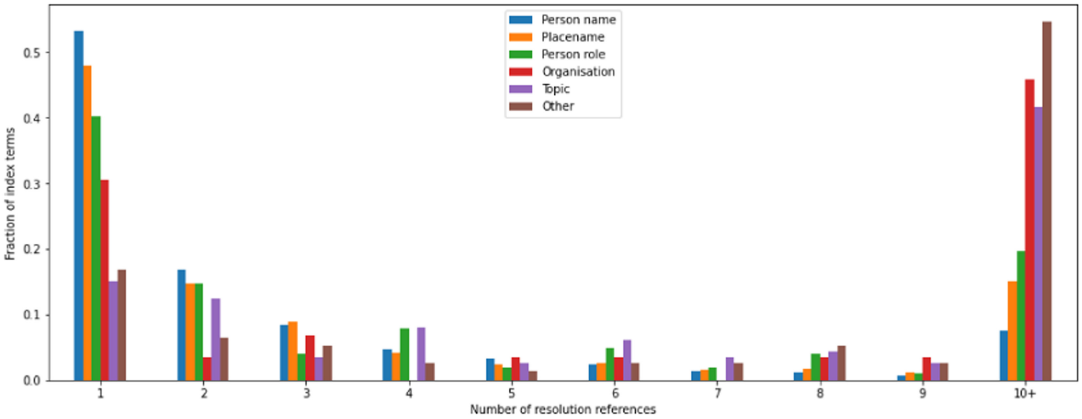


Fig. 8. Distribution of categories of index terms over the number of resolutions they reference, for index terms from the digitised index of 1610–1625.

region Brabant) or "Baron van Breda" (EN: Baron of the city Breda). There are no entries for generic person qualification referencing e.g., all resolutions related to barons or councils. Of the 3,499 index terms, 2,394 (68%) are *person names*, and 754 (22%) are *geographical names*. The other four categories together cover the remaining 351 (10%) of terms. This corresponds well to the distribution in our earlier analysis of the index terms in the 18th Century. So they suggest this is a stable distribution over the entire period. For the non-name categories, the list of terms is small enough to be curated manually. Hence, identifying those terms in the index pages using fuzzy search and matching can be done with high accuracy. The person and place name categories have many more terms, making manual curation even harder, as these lists become much longer when we zoom out to the full 220 years of indices.

Next, we look at how many resolutions are referenced by each term to get an insight into how each category of terms is capable of grouping and linking resolutions. In the context of viewing a specific resolution associated with an index term, the information system can allow the user to jump to other resolutions with the same index terms. Terms with a single reference have no linking or grouping effect. Terms with two references link only two resolutions, but some terms reference hundreds of resolutions in the period 1610–1625 alone.

The distribution of index terms over the number of resolutions they reference, for each of the six categories, is shown in Figure 8. The Y-axis shows the fraction of terms per category and the X-axis shows the number of resolutions that is referenced by an index term in that category. The majority of index terms in the person name category reference only a single resolution (53%), therefore do not create a link between resolutions and do not provide a grouping mechanism. For place names, this is only slightly lower (48%). Together with *person qualification*, these categories of index terms tend to offer little value for linking and grouping. Note that in Section 3.3, we showed that *person qualification* terms can be identified and contextualised in the opening formulas of resolutions. Hence, the resolutions they group and link via the indices is complemented by the resolutions they group and link via their extraction from the resolutions themselves. Among the *organisation*, *topic*, and *other* index terms, only a small fraction reference one or two resolutions, and a much larger fraction reference 10 or more resolutions. They offer more ways to gradually narrow down selections of resolutions, for instance as search facets.

4.3 Linking and Accessibility

How do we make digitised historical corpora accessible to researchers? We stress that recognising and indexing person and place names have value for making large historical corpora accessible, but argue that projects should consider prioritising other types of entities and topics that enable approaching person and geographical name accessibility in a different way. Person and place names are used by many users to search through archives [1, 5, 12, 13, 17, 23], and it is easy to see the value of indexing them in paper archives, as the lack of full-text search makes locating such specific information extremely time consuming and error prone. But with digitised access, the value of such names for faceted search and navigation is diminished and the role of name recognition has changed. Most names are unknown to most users, and once they encounter a name they find relevant, or they come to the corpus looking for a specific person, they can use full-text search to find other documents with that name. Of course, the errors introduced through the OCR process and the spelling variation in historical texts together form hurdles to identifying name occurrences via keyword search but also for linking documents via person names [6]. Finding orthographically similar names can be automated, but modern search engines offer these techniques as well. Searching names has therefore become more a matter of skill than a hard problem.

In prioritising which information layers to operationalise first, there are at least two criteria to take into consideration. First, choices of what type of information to identify and extract first, should be informed by the types of research questions and information needs of expected users, and how these relate to the logical structure of the corpus and the types of entities and topics that are recognisable and extractable. Second, the amount of effort required should be traded off against the quality of the recognition and extraction. Although applying generic NER techniques is straightforward, the output needs to meet a certain accuracy threshold to avoid misleading and/or frustrating users. They might rely on the provided information layer without noticing that either precision or recall is low or both. Or they might notice that most of the listed entities are nonsensical or return bad results, in which case they get frustrated or lose trust in the information system. We have witnessed many occasions where humanities scholars severely criticised or completely turned away from resources that were made accessible by NER for exactly this reason.

We argue that a prime purpose of operationalising entities as information layers should be supporting structured navigation with clear semantics that are relevant to the information needs of users. Identifying the document types of propositions and person qualifications of proposers operationalises two dimensions that communicate clearly to the user what effect they have on selection. We note that this in itself does not automatically make these layers transparent, as that requires careful design of the user interface to communicate potential limitations to users. The reasons we prioritise person qualifications and document types are that they can be extracted with high accuracy, and that they relate to various types of research questions around the interaction of the SG with different groups of society, as mentioned in Section 2. Moreover, they can be used as trigger words that offer clear signals that person and place names can be expected in syntactically predictable positions [10]. This can make person and place name recognition more precise and has the added advantage that they provide context for why and in what function, the name is mentioned.

5 CONCLUSIONS

A main challenge in making large digital archives accessible is to provide different dimensions and levels of access, as their potential users come with many different questions, different background knowledge, and different needs to explore the connections between the records. There are often several structural elements in the physical archive that support these different information access needs, but operationalising the structures and connecting them in the digital version requires techniques that are adapted to the specifics of the corpus.

The main argument underlying the approach we described is that, to make large historical corpora digitally accessible, we should (a) consider the potential information needs of users and (b) how to prioritise the extraction and operationalising of generic elements such as named entities and specific elements. Here, we stress in particular the logical structure of the corpus, the genre of documents, and the recurring attributes of entities that are mentioned. These elements give each other context and meaning that is impossible to extract from the text alone.

Projects that digitise historical corpora often focus on recognising names of persons and places, for which generic NER tools can be used. While these tools can give results with little effort, we argued that these results are less valuable than is often assumed. Recognising names in historic documents is often challenging and even with training on corpus-specific materials, the quality is often low, resulting in an output with many incorrectly recognised names, and many names that are not recognised. Moreover, most names occur only rarely, thereby adding little value for information access over users typing those names in the search box. Among the more commonly occurring names, there is often not enough information to disambiguate individuals, which leads entity-based access to give misleading results. The low accuracy problem can be remedied by creating a large set of corpus-specific training data, but not without significantly increasing the required effort. More training data also cannot solve the problem of disambiguation and the loss of context. Finally, generic NER results give users few handles to get an overview of what is in the corpus. In sum, we think that generic NER should not be a default for digitisation projects. Instead, there should be a careful consideration of (1) what kinds of entities (named and unnamed) are useful for (potential, future) users of the corpus that is being made accessible, (2) which of these types of entities can be reliably identified, and (3) how these entities can be transparently and meaningfully operationalised.

Extracting and operationalising corpus-specific elements are time-consuming steps (because tooling needs to be adapted to the specifics of the corpus), but we hope we have been able to demonstrate that they can lead to high-accuracy information layers and leave explainable traces of the decisions that transformed the physical archive to a digitised version. The types of structures that are present in historical corpora range from (a) the logical structures of the content as opposed to the physical structure of the paper sources, (b) ordering of content along temporal, alphabetical, numerical, or geographical axes, and (c) text units that follow a templated set of content elements, and formulaic language use. Many types of documents contain such structures, in particular, in legal documents like ordinances, notarial deeds, charters, and procès-verbaux (authorised statements of acts or proceedings in the exercise of duty), but also often in missives and reports and even early modern newspaper articles and advertisements [26].

We have described our novel approach of modelling, recognising, and extracting these structural elements, dealing with problems of text recognition errors, historical language variation, and the heterogeneity in structure and content found in long serial publications. Instead of using generic Natural Language Processing in a one-shot information extraction pipeline, we developed an iterative approach in which corpus and domain experts can incorporate their knowledge in the extraction process. Our insights from analysing the output are transparently modelled and fed back into the process. By prioritising person qualifications over person names, we have also built a lexicon of trigger words that we can use as contextual features to improve NER for person names, and enrich the output with those person qualifications and possibly relate them to other aspects of the resolutions. Evaluation of our results shows that this leads to highly accurate layers of structured text, annotations and metadata.

We used an example research question about whether and how the Resolutions of the Dutch States General reflect changes in the petitioning by the citizens of the Dutch Republic. The analysis of our results based on this example question shows that the information layer provide meaningful handles to select, re-organise, and re-order the digitised material in relation to different types of analysis that are relevant to the research question at hand. It also shows that our approach arrives at interpretable models that can be used as provenance to explain how each algorithm came to its decisions.

Overall, we hope that the results of our approach lead to reconsideration of the priorities in future projects that aim to digitise and make accessible historical collections.

REFERENCES

[1] Marcia J. Bates. 1996. The Getty end-user online searching project in the humanities; Report No. 6: Overview and conclusions. *Coll. Res. Libraries* 57 (1996).

[2] Lars Borin, Dimitrios Kokkinakis, and Leif-Jöran Olsson. 2007. Naming the past: Named entity and animacy recognition in 19th Century Swedish literature. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH'07)*. 1–8.

[3] Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, José Moreno, Nicolas Sidère, and Antoine Doucet. 2020. Robust named entity recognition and linking on historical multilingual documents. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF'20)*, Vol. 2696. CEUR-WS Working Notes, 1–17.

[4] Alan Bryman. 2016. *Social Research Methods*. Oxford University Press, Oxford, UK.

[5] A. Chardonnens, E. Rizza, M. Coeckelbergs, and S. van Hooland. 2018. Mining user queries with information extraction methods and linked data. *Journal of Documentation* 74, 5 (2018), 936–950. https://doi.org/10.1108/JD-09-2017-0133

[6] Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. Impact of OCR errors on the use of digital libraries: Towards a better access to information. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL'17)*. IEEE, 1–4.

[7] Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. Retrieved from https://arXiv:1807.04905.

[8] Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2019. ICDAR2019 competition on recognition of documents with complex layouts—RDCL2019. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'19)*. IEEE, 1521–1526.

[9] Giovanni Colavizza, Maud Ehrmann, and Fabio Bortoluzzi. 2019. Index-driven digitization and indexation of historical archives. *Front. Dig. Human.* 6 (2019), 4.

[10] Mª Luisa Díez Platas, Salvador Ros Munoz, Elena González-Blanco, Pablo Ruiz Fabo, and Elena Alvarez Mellado. 2021. Medieval Spanish (12th–15th Centuries) named entity recognition and attribute annotation system based on contextual information. *J. Assoc. Info. Sci. Technol.* 72, 2 (2021), 224–238.

[11] David Doermann, Karl Tombre, et al. 2014. *Handbook of Document Image Processing and Recognition*. Springer.

[12] Wendy Duff and Catherine Johnson. 2003. Where is the list with all the names? Information-seeking behavior of genealogists. *Amer. Archiv.* 66, 1 (2003), 79–95.

[13] Wendy M. Duff and Catherine A. Johnson. 2002. Accidentally found on purpose: Information-seeking behavior of historians in archives. *Library Quart.* 72, 4 (2002), 472–496.

[14] Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named entity recognition and classification on historical documents: A survey. Retrieved from https://arXiv:2109.11406.

[15] Joris van Eijnatten, Toine Pieters, and Jaap Verheul. 2013. Big data for global history: The transformative promise of digital humanities. *BMGN-Low Countries Hist. Rev.* 128, 4 (2013), 55–77.

[16] Dan Gillick, Nevena Lazic, Kuzman Ganchev, Jesse Kirchner, and David Huynh. 2014. Context-dependent fine-grained entity type tagging. Retrieved from https://arXiv:1412.1820.

[17] P. Gooding. 2016. Exploring the information behaviour of users of Welsh Newspapers Online through web log analysis. *Journal of Documentation* 72, 2 (2016), 232–246. https://doi.org/10.1108/JD-10-2014-0149

[18] Ralph Grishman. 1997. Information extraction: Techniques and challenges. In *Proceedings of the International Summer School on Information Extraction*. Springer, 10–27.

[19] Randolph Head. 2003. Knowing like a state: The transformation of political knowledge in Swiss archives, 1450–1770. *J. Modern Hist.* 75, 4 (2003), 745–782.

[20] Mark J. Hill and Simon Hengchen. 2019. Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Dig. Scholar. Human.* 34, 4 (2019), 825–843.

[21] Rik Hoekstra. 2017. The griffiers and the keeping of information in the Resolutions of the States General of the United Dutch Provinces, 1576–1796. Retrieved from https://www.researchgate.net/publication/352679752.

[22] Rik Hoekstra and Marijn Koolen. 2019. Data scopes for digital history research. *Hist. Methods: J. Quant. Interdisc. Hist.* 52, 2 (2019), 79–94.

[23] Bouke Huurnink, Laura Hollink, Wietske Van Den Heuvel, and Maarten De Rijke. 2010. Search behavior of media professionals at an audiovisual archive: A transaction log analysis. *J. Amer. Soc. Info. Sci. Technol.* 61, 6 (2010), 1180–1197.

[24] N. Japikse et al. 1915–1994. *Resolutiën der Staten-Generaal 1576–1625*. Nijhof, Den Haag.

[25] Charles Jeurgens. 2016. Schurende systemen: Seriearchieven in de digitale wereld. In *Schetsboek digitale onderzoek-omgeving en dienstverlening: Van vraag naar experiment*, H. Berende, K. van der Heiden, T. Thomassen, C. Jeurgens, C. van der Ven, and H. de Man (Eds.). Stichting Archiefpublicaties, 's-Gravenhage, 54–61.

[26] Wouter Klein. 2018. *New Drugs for the Dutch Republic: The Commodification of Fever Remedies in the Netherlands (c. 1650–1800)*. Ph.D. Dissertation. Utrecht University.

[27] Mitchell Koch, John Gilmer, Stephen Soderland, and Daniel S. Weld. 2014. Type-aware distantly supervised relation extraction with linked arguments. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1891–1901.

[28] I. B. Leemans, E. Maks, J. M. van der Zwaan, H. M. E. P. Kuijpers, and Kristine Steenbergh. 2017. Mining embodied emotions: A comparative analysis of bodily emotion expressions in dutch theatre texts 1600–1800. *Dig. Human. Quart.* 11, 4 (2017).

[29] Xiao Ling and Daniel S. Weld. 2012. Fine-grained entity recognition. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*.

[30] Albert Meroño-Peñuela, Ashkan Ashkpour, Marieke Van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach, and Frank Van Harmelen. 2015. Semantic technologies for historical research: A survey. *Semantic Web* 6, 6 (2015), 539–564.

[31] Juri Opitz, Leo Born, and Vivi Nastase. 2018. Induction of a large-scale knowledge graph from the Regesta Imperii. In *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. 159–168.

[32] Hinke Piersma and Kees Ribbens. 2013. Digital historical research: Context, concepts and the need for reflection. *BMGN-Low Count. Hist. Rev.* 128, 4 (2013), 78–102.

[33] Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Synthesis lectures on human language technologies, Vol. 5. Morgan & Claypool Publishers. 1–157.

[34] Michael Piotrowski. 2019. Historical models and serial sources. *J. Eur. Period. Stud.* 4, 1 (2019), 8–18.

[35] Animesh Prasad, Hervé Déjean, and Jean-Luc Meunier. 2019. Versatile layout understanding via conjugate graph. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'19)*. IEEE, 287–294.

[36] Lorenzo Quirós. 2017. P2PaLA: Page to PAGE Layout Analysis toolkit. Retrieved from https://github.com/lquirosd/P2PaLA.

[37] Theodorus Helenus Franciscus Riemsdijk. 1885. *De griffie van hare hoog mogenden: Bijdrage tot de skennis van het archief van de Staten-Generaal der Vereenigde Nederlanden.* M. Nijhoff.

[38] Kepa Joseba Rodriquez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. 2012. Comparison of named entity recognition tools for raw OCR text. In *Proceedings of the Konferenz zur Verarbeitung natürlicher Sprache/Conference on Natural Language Processing (KONVENS'12)*. 410–414.

[39] C. Annemieke Romein, Michel de Gruijter, and Sara Floor Veldhoen. 2020. The datafication of early modern ordinances. *DH Benelux J.* 2 (2020).

[40] Pedro Javier Ortiz Suárez, Yoann Dupont, Gaël Lejeune, and Tian Tian. 2020. SinNer@Clef-Hipe2020: Sinful adaptation of SotA models for named entity recognition in French and German. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF'20)*.

[41] Theo Thomassen. 2019. *Onderzoeksgids: Instrumenten van de macht: de Staten-Generaal en hun archieven 1576–1796 (Band 1)*. Sidestone Press. 426 pages.

[42] Theo Thomassen. 2019. *Onderzoeksgids: Instrumenten van de macht: de Staten-Generaal en hun archieven 1576–1796 (Band 2)*. Sidestone Press. 426 pages.

[43] Konstantin Todorov and Giovanni Colavizza. 2020. Transfer learning for named entity recognition in historical corpora. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF'20)*.

[44] Frank Upward, Barbara Reed, Gillian Oliver, and Joanne Evans. 2018. *Recordkeeping Informatics for a Networked Age*. Monash University.

[45] Daniel van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the impact of OCR quality on downstream NLP tasks. In *Proceedings of the International Conference on Agents and Artificial Intelligence (ICAART'20)*. 484–496.

[46] Maria Vargas-Vera, John Domingue, Yannis Kalfoglou, Enrico Motta, and Simon Buckingham Shum. 2001. Template driven information extraction for populating ontologies. In *Proceedings of the Workshop on Ontology Learning*.

[47] Mohamed Amir Yosef, Sandro Bauer, Johannes Hoffart, Marc Spaniol, and Gerhard Weikum. 2012. Hyena: Hierarchical type classification for entity names. In *Proceedings of the International Conference on Computational Linguistics (COLING'12)*. 1361–1370.

Received 8 November 2021; revised 25 July 2022; accepted 9 September 2022