Modelling Resolutions of the Dutch States General for Digital Historical Research

Marijn Koolen[1,2], Rik Hoekstra[1,2], Ida Nijenhuis[1], Ronald Sluijter[1],
Esther van Gelder[1], Rutger van Koert[2], Gijsjan Brouwer[2], and Hennie
Brugman[2]

[1] Huygens Institute for the History of the Netherlands, Amsterdam, Netherlands
`ida.nijenhuis,ronald.sluijter,esther.van.gelder@huygens.knaw.nl`
`https://www.huygens.knaw.nl`
[2] KNAW Humanities Cluster, Amsterdam, Netherlands
{`marijn.koolen,rik.hoekstra,`
`rutger.van.koert,hennie.brugman,gijsjan.brouwer`}`@di.huc.knaw.nl`
`https://huc.knaw.nl`

Abstract. 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. This archive has rich potential for historical research, but the heterogeneity and dispersion of information makes using it for research a challenge. In this paper we describe how we deal with the challenges of structuring and connecting the information contained within this archive, and how this results in a computational platform that allows users to explore and analyse the archive through many connected layers of metadata.

Keywords: Information Extraction · Digital History · Data Modelling

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 and consists of more than 500,000 pages, handwritten and printed resolutions, in separate, chronologically ordered series. The Resolutions of the States General in the Dutch Republic are a key resource to the political history of this period as they contain all decisions made by the States General (SG), the central ruling body in the Republic. It was designated as a key resource when in 1905 the work of publishing the resolutions started [8]. The manual editing resulted in two series of analogue 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).[3]. The resolutions reveal the decision making process and are relevant for both high and low politics. They allow researchers to answer many different research questions about politics -

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

but not only politics - in the Dutch Republic and its position in the world. The resolutions are also key to all the other records of the SG (about 1 mile) and form a backbone with which these other records can be connected and contextualised.

Many research questions require working your way through hundreds of large volumes of text without adequate indexes. The relevant data are hidden and scattered across millions of paragraphs of dense and repetitive text. Moreover, different research questions require different selections, reorganisations and reorderings of the records to bring together and connect the dispersed information. It is therefore 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. Many archives and libraries have experimented with giving access to their collections by means of their digitised inventories and some have gone a step further, using existing indexes of serial collections [9,2,5]. But these archival referential systems are too coarse for access beyond the document level. Moreover, the existing scholarly apparatus consists of many more reference systems and tools that can be put to good use. Centuries of dealing with these complications have led to a number of convenient and often-employed structures that are part of the printed culture but are often ignored in the translation to digital access [16,12].

In this paper, we describe our work on identifying and extracting several of these reference systems and structure as layers of data and metadata to 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 the many standard phrases used in the resolutions. With these phrase models, we can exploit both expert knowledge and the fact that the resolutions contain highly repetitive language to improve the extraction process.

Our approach departs from the important condition that the extracted information should both reflect the structure of the resource and the type of research questions it is supposed to support. We illustrate this with a relatively simple research problem for the following question:

> Do the resolutions over time reflect an increasing possibility for citizens to put forward their concerns to the SG?

To investigate this, we need to know what types of proposals and requests were submitted that led to resolutions, when each was put forward, by whom and what decision was reached. Moreover, answering whether access to the general public changed over time requires that the set of resolutions is either complete or at least representative. Just digitising the archive does *not* support answering this question. A digitised archive consists only of a combination of images and (roughly) digitised text, without structure and only very limited metadata.

This paper discusses the way we deal with a number of challenges to transform a digitised archive into an online publication which supports the structured analysis required for such historical research:

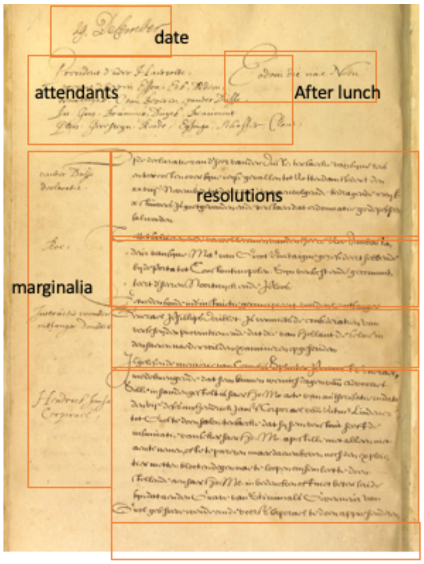

date
attendants
After lunch
resolutions
marginalia

Fig. 1. Structure of the handwritten resolutions.

- **Text recognition challenges** The corpus contains handwritten and printed materials, and many different types of information (paragraphs, lists, tables, marginalia). Moreover, the language use and spelling changed throughout the covered period.
- **Information extraction challenges** The text recognition process results in textual representations, but without handles for using the rich structure in the analogue documents. Rich layers of metadata need to be derived from the text. This requires a combination of text mining expertise and domain knowledge to identify, extract and connect the dispersed information.
- **Linking challenges** The structure of the records is complex, with various types of materials that are linked to each other, and identifying this structure and making it inter-operable is crucial to making the online publication usable.
- **Presentation challenges** Users approach the digital archive with different aims and background knowledge and with different tasks of reading, searching, combining and extracting information. The information system should support a broad range of methods to select, organise and order the records.

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 [15]. In the rest of this paper we discuss how we deal with these challenges.

2 The Resolutions as a Historical Resource

The volumes of the resolutions embody a continuous series of day-to-day meetings of the Dutch States General as the central assembly of delegates from (ultimately) seven sovereign provinces. Each session starts with a date and a list of attendants and the president of the day, followed by summaries of a varying number of 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 contain a decision (resolution) of acceptance, rejection or deferral, pending further investigation or requests for information. 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. Some summaries contain copies of letters or other important incoming documents as insertions.

2.1 Supporting Research Questions

The types of data and metadata layers we want to make operable are related to the types of research questions and themes that we want to support:

Narrative analysis Long-term developments and shifts, both internationally and within Dutch society (e.g. what was the relative position and wealth of the province, how did the competition between navy and army develop, how important were the different colonies, how did the SG deal with different religious groups).

Thematic analysis the development of certain themes/topics over time. Do the resolutions over time reflect an increasing possibility for citizens to put forward their concerns to the SG?

Content analysis Quantitative analysis with regard to what or who was discussed, when and how often (e.g. financial and economic policy, nominations for officeholders or army positions, petitions by citizens).

Network analysis How can we trace the dividing line between formal and informal politics, and politics behind the scenes? Performing serial research into the attendance at meetings and in committees can answer questions like: who worked together? Who were involved in decisions around specific topics and in larger policy issues and how were these persons related?

Linguistic analysis 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 these questions requires operationalising several structural elements of the written and printed texts, into multiple layers of metadata 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 proposer belonged to and *order* them temporally. For a network analysis, a researcher needs to *select* the relevant resolutions, *extract* and *classify* the types of parties involved (e.g. proposers, the attending SG members, committees investigating and reporting on issues, and financiers providing budget) and the relationships between them. Adding metadata for these different aspects requires extracting the relevant information from an estimated one million resolutions, which requires an information extraction that is automated where possible, but which needs to be informed by export knowledge and a human in the loop. On top of that, each selection or reorganisation made by users creates a different view of the data, the interpretation of which is influenced by our decision in creating the metadata layers, so it is important that our processing is transparent and visible to the user [7].

2.2 Operationalising the Structure

In many digitisation projects for e.g. newspapers or books, the text of individual scans is recognised using a model trained on ground truth, then metadata about the scan is added to give ways to organise the text. For newspapers these are typically the name of the newspaper and the date and page number. For books, this is more difficult. The bibliographic metadata from a library catalogue can be added, but 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 parts or chapters or sections, where these start and end, and whether they have titles or headings.

For the volumes of resolutions, text recognition results in one document per scan with the recognised words and their pixel coordinates. 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 1000 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 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. This also allows scaling the analysis of resolutions. E.g. how many resolutions 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 elements of resolutions as metadata layers:

Meetings and meeting dates The specific date on which a proposition discussed and a decision reached.

Attendance lists and president The persons who were present and involved in the decision making process of each resolution.

Resolutions 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.

Named entities Persons who submitted propositions and persons, committees selected to investigate and report on issues that were discussed, and other named entities such as organisations, geographic locations and ship names.

Topics The topic of individual resolutions. This is partly provided by the *contemporary indices*, *lists of index terms* and *marginalia*. But some form of key phrase extraction or topic modelling could provide alternative (and differently biased) topical perspectives.

Before we describe how we use this model for extract information (Section 4), it is necessary to discuss the numerous text recognition challenges this resource puts forward, because the quality of text recognition has a big impact on the of information extraction process that follows it.

3 Text Recognition Challenges

The physical state of the records is heterogeneous. The corpus has a mix of formats, with an ongoing record of handwritten resolutions, that from the 18^{th} 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 modes of handwriting in general, that changed from 16^{th}-century Gothic to 18^{th}-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).

The automatic recognition of the texts of the resolutions is performed by OCR (Optical Character Recognition) and HTR (Handwritten Text Recognition). 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.

OCR is standard technology, but requires training because of diverging printing formats and especially orthographic difficulties. The current model achieves a Character Error Rate (CER) of 8, that is, 8 out every 100 character are incorrectly recognised. HTR is more difficult because of the irregularities of handwriting. Recognising the 425,000 pages of handwritten resolutions requires several steps in a pipe line. We used the P2PaLA tool [14] 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, 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 that uses the web version of Transkribus.

On handwritten pages with no layout corrections, we observed that misidentified regions result in much more variation in the number of errors per page and a higher CER. We continue to expand the ground truth set and plan to train specific models for pages diverging from a 'standard' resolution page (e.g. indices and tables) and for handwriting on which the HTR+-model performs worse than the average.

In the next section we discuss the results of the information extraction process on the OCR output of the printed resolutions.

4 Information Extraction Challenges

A typical step in extracting information from historical texts is to use general approaches like Named Entity Recognition (NER), part-of-speech tagging, and lemmatisation to identify entities and topical words and phrases [11]. 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 [4,13,10].

On English texts these generic approaches work to some extent. English orthography has not changed much since the 18^{th} century, therefore resources for modern English can be effective [6,15]. In Dutch, changes are larger, making generic approaches less useful. We annotated named entities in 200 pages of manually transcribed resolutions and retrained the Spacy NER tagger[4] with 90% of the pages and tested on the remaining 10%. This led to a precision of 0.49 and recall of 0.19. Although this is likely to improve by annotating more pages, there are two 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.

Moreover, such techniques do not alleviate the problem of identifying 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 [3,1], but cannot interpret specific semantics such as temporal orderings of meeting dates and the geographical ordering in the attendants lists.

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. [2], 2) explicitly modelling domain knowledge in lists of formulaic textual phrases, and 3) approximate string

[4] `https://spacy.io/models/nl`

Text string found	Frequency
Ontfangen een Miflive van	286
Ontfangen cen Miflive van	101
Ontfangen een Miffive van	65
Ontfangen cen Miffive van	31
Ontfangen een Mifive van	27
Ontfangen een Miílive van	11
Ontfangen een Millive van	10

Table 1. OCR text string matches found for the opening formula 'Ontvangen een Missive van' in the resolutions for the year 1705.

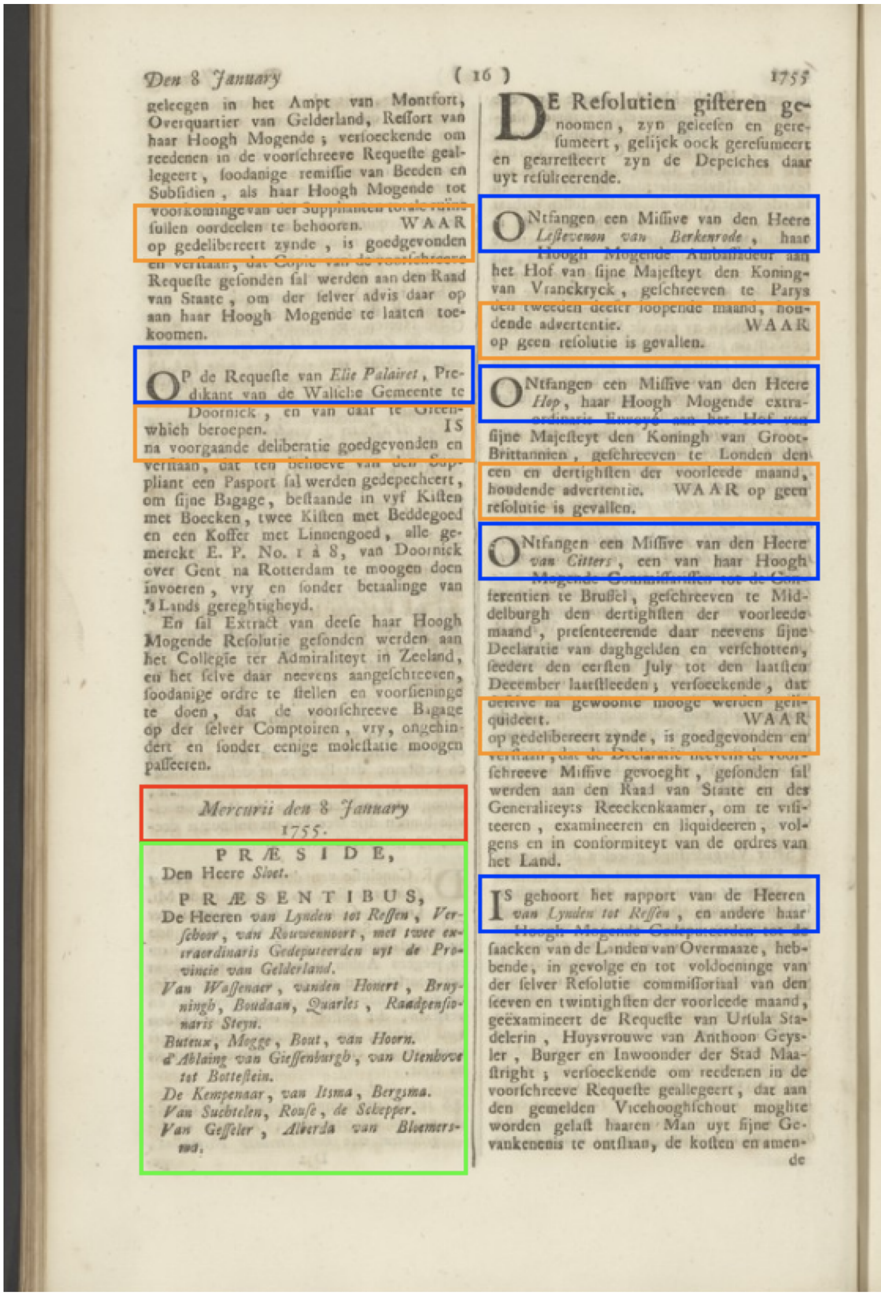
searching and matching. First, the sessions have a fixed structure and layout, with the opening of a next session and the attendance list represented in a different font and text alignment than the resolution summaries. Second, the resolution texts are extremely repetitive, using the same phrasings with little variation across decades of resolution. We use phrase models that contain lists of frequently occurring phrases, such as the standard formulas used to announce the next resolution or its decision, or the start of a new meeting (see Figure 2). Formulaic phrases can be short or long, e.g. a single word or an entire sentence including punctuation. Each phrase has a metadata category and label so that an approximate match in the text can be tied to a metadata layer. 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 CER, 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.[5]

The phrase model for opening formulas contains eight different formulas, each with a list of variant phrasings and a label for the type of proposition. For instance, the opening formula 'Ontvangen een Missive van' (EN: *Received a Missive of*) indicates the proposition is a missive. For the resolutions of year 1705, 950 matches are found with 315 different OCR strings (Table 1). The most frequent OCR string occurs 286 times, so identifies only 30% of the resolutions for missives. Other proposition types include reports, requests, memories 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

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



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 der 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. 2. 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)

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 sessions for 12 and 15 January 1725 in the first iteration, we exploit our knowledge that sessions are chronologically ordered, therefore know that the sessions for 13 and 14 January should be in between these starting points. In the second iteration, we can search with lower thresholds in a much smaller amount of text and much higher chances of success. For the names of attendants, we exploit our knowledge that the president, 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 as either president but with recognition errors, or as regular attendants, reducing the number of unknown and uncertain names in the list. For the individual resolutions, we use the fact that they use extremely repetitive language, with typical textual formulas for the opening proposition (see Figure 2), e.g. 'Ontvangen een missive van ...', (English: 'Received the missive of ...') and the decision, e.g. 'WAAR op gedelibereert zynde, is goedgevonden ende verstaan ...' (English: 'WHICH after deliberation has been accepted ...'). Through multiple iterations of extracting resolutions, we gradually build lists of standard formulas, which we subsequently use in approximate searching to find variations of these formulas that may contain both text recognition errors and slightly different spellings. For all elements 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,[6] this makes the decision process transparent and repeatable.

4.1 Evaluation

We evaluated our approach with various ground truth data sets:

Page type identification Identifying whether a page contains resolutions, index entries or respect, 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 term lists.

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 1703 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 0.99 and recall of 0.93 on the test set consisting of the remaining 300 meetings. The extraction

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

algorithm detects the correct start for 100% of the extracted meetings, but in three out 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 23,605 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. Based on the annotated ground-truth, there are an estimated 311,586 resolutions in the printed volumes (13.2 per meeting). On the ground truth test set of 198 resolutions, our phrase model currently achieves a precision of 0.94 and recall of 0.76 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. We are currently developing ground truth data for this.

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.

5 Linking Challenges

Information extraction gives us a way to navigate, select and order the the meetings and resolutions in individual volumes through the different layers of metadata. The next step is to connect those layers to each other and to connect them across the hundreds of volumes. Connecting the metadata across different layers enables queries like 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 other dimension is connecting metadata elements within a single layer but 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.

We manually transcribed and merged the lists of index entries for seven volumes of resolutions between 1742 and 1785. These turned out to contain 3934 distinct entries, with individual volumes containing between 673 and 994 entries. Geographic locations and organisations have a higher overlap between years and are more stable across time than persons. Subjects seem more stable as well, but it is challenging to establish whether, for instance, the index entry on 'declaration' refers to the same thing over the years. 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 logically 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.

Our preliminary analysis of the overlap in index terms across volumes shows 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.

6 Presentation Challenges

The size and prominence of the corpus offers research opportunities for many years to come, but its potential can only be reached if it is presented to end users with relevant and easy-to-use functionality. The Resolutions need a reliable yet flexible platform that stays in active development to publish information when it becomes available. One problem of custom build publication environments is the risk of stalling maintenance and further development when project finances dry up. When development stops, stability and flexibility decrease. We use a generic publication environment we call Docere. Several other projects are currently being developed on the platform. Development is mostly mutually beneficial, a generic feature realised for one project, will be instantly available to all the others. On the one hand, this means the cost of development is split over projects. On the other, individual projects have less influence on development decisions, as a decision for one project will directly affect all other projects.

To offer flexibility per project some components of the user interface can be custom built, for instance the page header and footer, which search facets are available, or how search results are shown. Full customisability is available for the rendering of the text. The OCR and HTR output are mapped to custom components, which means all extracted text elements can have custom features attached to them. A meeting or resolution is shown in relation to previous and next meetings or resolutions. Related meetings or resolutions can be found via topical metadata or using full-text search, giving the user extra possibilities to navigate the set. Docere leverages a search index to create a faceted search. A facet can represent metadata (date, president, attendee and others) or entity (person, place, topic and others) from a meeting or resolution.

In addition, it is possible to directly query the data set through an Application Programming Interface (API). This is especially beneficial to the growing number of researchers for whom the user interface is not sufficient to answer their research questions. The API will expose the raw and enriched data and will be essential in adhering to the FAIR principles.[7]

7 Conclusions

One of the challenges in digitising large archives is to provide different dimensions and levels of access, as their potential users come with many different questions, different background knowledge and different and different needs to explore the connections between the records. There are often several structural elements in the analogue version that support these different information access needs, but operationalising these structures and connecting them in the digital version requires tackling a cascade of challenges.

This paper describes our approach of modelling, recognising and extraction 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 expert knowledge of the records can be incorporated in the extraction process, and insights from analysing the output can be transparently modelled and fed back into the process. Evaluation of our results shows that this not only leads to highly accurate layers of structured text, annotations and metadata, but also to interpretable models that can be used as provenance to explain how each algorithm came to its decisions.

Acknowledgements

This research is funded by the The Netherlands Organisation for Scientific Research (NWO) through the NWO Groot project REPUBLIC (an acronym for REsolutions PUBlished In a Computational Environment) 2019-2023.

[7] `https://www.go-fair.org/fair-principles/`

References

1. Clausner, C., Antonacopoulos, A., Pletschacher, S.: Icdar2019 competition on recognition of documents with complex layouts-rdcl2019. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1521–1526. IEEE (2019)
2. Colavizza, G., Ehrmann, M., Bortoluzzi, F.: Index-driven digitization and indexation of historical archives. Frontiers in Digital Humanities **6**, 4 (2019)
3. Doermann, D., Tombre, K., et al.: Handbook of document image processing and recognition. Springer (2014)
4. Eijnatten, J.v., Pieters, T., Verheul, J.: Big data for global history: The transformative promise of digital humanities. BMGN-Low Countries Historical Review **128**(4), 55–77 (2013)
5. Head, R.: Knowing like a state: the transformation of political knowledge in swiss archives, 1450–1770. The Journal of Modern History **75**(4), 745–782 (2003)
6. Hill, M.J., Hengchen, S.: Quantifying the impact of dirty ocr on historical text analysis: Eighteenth century collections online as a case study. Digital Scholarship in the Humanities **34**(4), 825–843 (2019)
7. Hoekstra, R., Koolen, M.: Data scopes for digital history research. Historical Methods: A Journal of Quantitative and Interdisciplinary History **52**(2), 79–94 (2019)
8. Japikse, N., et al: Resolutiën der Staten-Generaal 1576-1625. Nijhof, Den Haag (1915-1994)
9. Jeurgens, C.: Schurende systemen: Seriearchieven in de digitale wereld. In: Berende, H., van der Heiden, K., Thomassen, T., Jeurgens, C., van der Ven, C., de Man, H. (eds.) Schetsboek digitale onderzoek-omgeving en dienstverlening: Van vraag naar experiment, pp. 54–61. Stichting Archiefpublicaties, 's-Gravenhage (2016)
10. Leemans, I., Maks, E., van der Zwaan, J., Kuijpers, H., Steenbergh, K.: Mining embodied emotions: A comparative analysis of bodily emotion expressions in dutch theatre texts 1600-1800'. Digital Humanities Quarterly **11**(4) (2017)
11. Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., Van Harmelen, F.: Semantic technologies for historical research: A survey. Semantic Web **6**(6), 539–564 (2015)
12. Opitz, J., Born, L., Nastase, V.: Induction of a large-scale knowledge graph from the regesta imperii. In: Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. pp. 159–168 (2018)
13. Piersma, H., Ribbens, K.: Digital historical research: Context, concepts and the need for reflection. BMGN-Low Countries Historical Review **128**(4), 78–102 (2013)
14. Quirós, L.: P2pala: Page to page layout analysis toolkit. `https://github.com/lquirosd/P2PaLA` (2017), gitHub repository
15. van Strien, D., Beelen, K., Ardanuy, M.C., Hosseini, K., McGillivray, B., Colavizza, G.: Assessing the impact of ocr quality on downstream nlp tasks. In: ICAART (1). pp. 484–496 (2020)
16. Upward, F., Reed, B., Oliver, G., Evans, J.: Recordkeeping informatics for a networked age. Monash University (2018)