The Semantics of Structure in Large Historical Corpora

Marijn Koolen - KNAW Humanities Cluster
Rik Hoekstra - KNAW Humanities Cluster

Structuring large historical corpora that are too big to be processed manually can take two approaches. The first is an inductive method extracting implicit entities and meaning from textual (and sometimes visual) content. With the help of AI or manually compiled (existing) lists of entities, the entities are converted into information. The second, that Colavizza (2019) calls referential information systems, takes existing reference systems (like archival indexes) and uses them to contextualize individual documents. Both methods are used to turn corpora into computer accessible information systems. Ideally a more complete information system would result from combining both approaches, but in practice they are hard to bridge because of a number of different problems. This paper presents an approach that addresses those problems and combines inductive methods of automated text analysis and information extraction techniques with knowledge of the referential information systems to add rich semantic layers of information to large historical corpora.

Making large historical corpora accessible for research usually involves a pipeline of processing steps, ranging from text recognition to entity and event spotting, disambiguation, identification and ideally contextualization (Meroño-Peñuela et al. 2015). In many projects much effort is spent on producing a close-to-perfect text by transcribing, or by a mixed procedure of automatic transcription by Optical Character Recognition (OCR) or Handwritten Text Recognition (HTR) and manual correction of the results, as many of the later elements in the pipeline require high-quality text to work well. There are ways to partially solve OCR or HTR (Handwritten Text Recognition) errors automatically through post-correction (see e.g. Reynaert 2014, Reynaert 2016), or to use word embeddings to overcome matching problems, (e.g. Egense 2017). The most important limitation of this approach is that full-text alone is not enough to make a corpus available for research that is not primarily directed at the text but rather at its information (Hoekstra and Koolen 2018, Upward 2018). Extracting and contextualizing information has many issues such as OCR and HTR errors that make it difficult to use standard Natural Language Processing (NLP) tools like Named Entity Recognition (NER), topic modelling, Part Of Speech (POS) tagging and sentiment analysis, which has been common knowledge for a long time (Lopresti 2008, Traub et al. 2015, Mutuvi et al. 2018, Hill & Hengchen 2019, van Strien et al. 2020). However, solutions for such issues are scarce and badly documented, as argued by, amongst others, Piersma and Ribbens (2013), van Eijnatten et al. (2013) and Leemans et al. (2017).

Many archives and libraries have experimented with giving access to their collections by means of their digitized inventories and some have gone a step further, using existing indexes of serial collections (Jeurgens 2016, Colavizza 2019, Head 2003). But these archival referential systems are too coarse for access beyond the document level. However, 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 (Upward 2018, Opitz 2018).

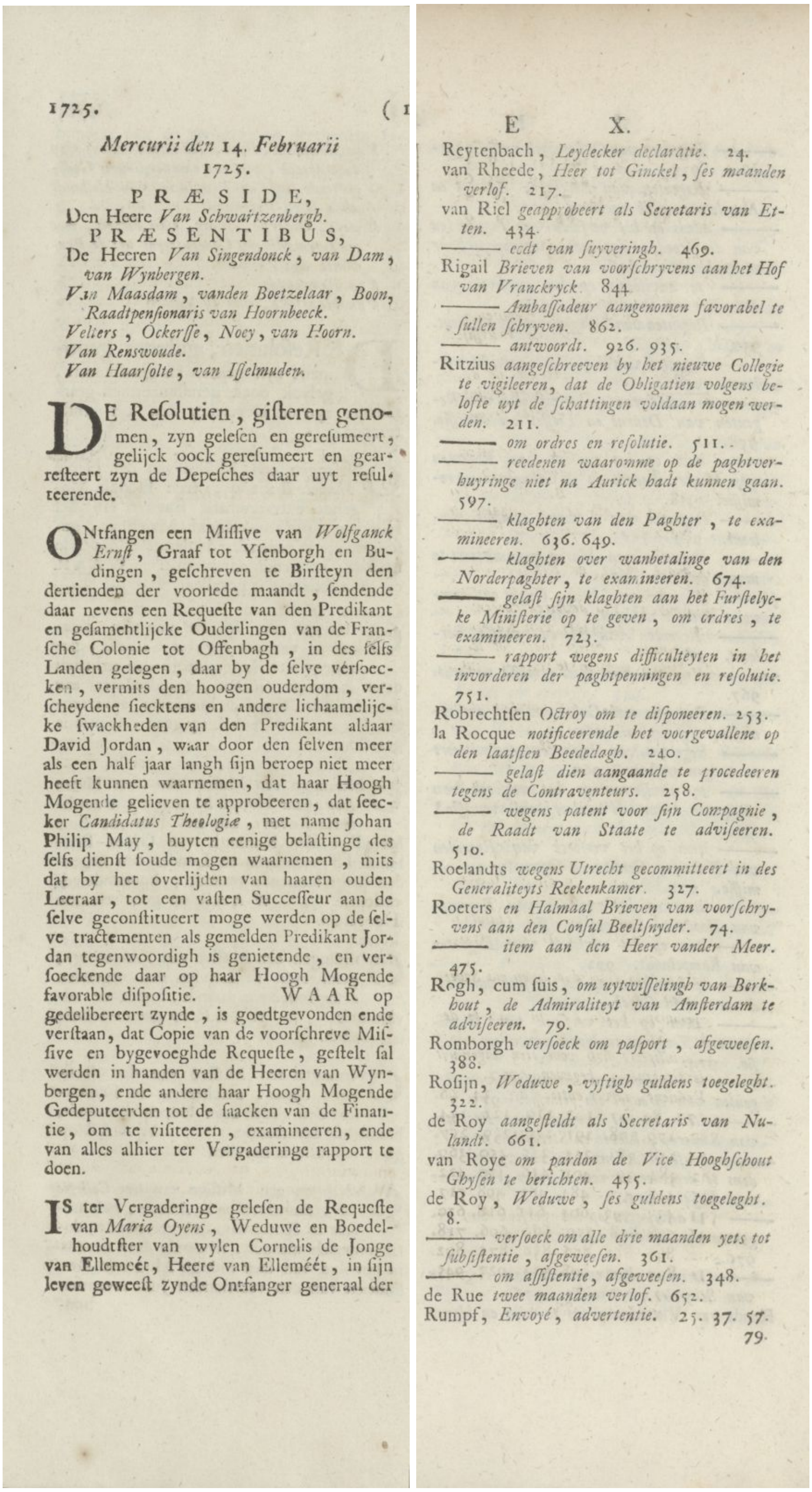
Exploiting Referential Information Systems and Repetitive Phrases

Instead of trying to find latent semantic structures through full-text analysis, these explicit structures allow for finding intended semantic information that is likely not available in another form. Remarkably, many digitization programmes take no advantage of these structures and sometimes do not even digitize them, extracting only the main textual body as plain text.

A concrete and relatively simple example, the Resolutions of the Dutch States General is a collection of all resolutions (decisions) of the Dutch Republic from 1576 until 1796 and contains around 440,000 pages. Roughly half of the pages, up to 1703, are handwritten. From 1703 onwards, the States General printed yearly editions for easier access to previous resolutions. The States General met six days a week and kept a list of who was present on what date, followed by a summary of each resolution. The left side of Figure 1 shows one column from the printed edition of 1725 with the meeting date, attendants list and some of the resolutions. On the right side is a column from an index page with indexed terms and page references.

Typical steps in digitizing such a large resource are the identification of columns on the page and OCR of the text to make it full-text searchable. Most typically NER would be applied to automatically identify named entities, even if its success depends highly on the OCR and HTR quality. With an estimated 1.5 to 2 million mentions of person names and close to half a million geographic names and organisations, NER would result in a huge list of names, with many false positives (incorrect names) and false negatives (unrecognised names). One issue is that NER tools tend to use uppercase initial letters as a signal that a word is a name, but in early modern texts, uppercase was used for many nouns as well. The most frequently recognized person names then tend to be terms that were often mentioned, like Money, Passport and Meeting. The main problem is that there is no way to predict which names are correct and which are not so users cannot easily filter out the obvious errors from the hundreds of thousands of distinct names. Using reference tools with lists of known historical persons for identification has the limitation that they cover only a small selection of the most famous names, which are not necessarily the most relevant names.

We argue that projects should focus more energy on extracting and operationalizing the existing structure of such corpora. In the case of the resolutions, there is a lot of value in recognizing the meeting dates, attendants lists and index terms.



1725. (1

Mercurii den 14. *Februarii* 1725.

PRÆSIDE,

Den Heere *Van Schwartzenbergh.*

PRÆSENTIBUS,

De Heeren *Van Singendonck*, *van Dam*, *van Wynbergen.*
Van Maasdam, *vanden Boetzelaar*, *Boon*, *Raadtpenſionaris van Hoornbeeck.*
Velters, *Ockerſſe*, *Noey*, *van Hoorn.*
Van Renswoude.
Van Haarſolte, *van Iſſelmuden.*

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

ONtfangen een Miſſive van *Wolfganck Ernſt*, Graaf tot Yſenborgh en Budingen, geſchreven te Birſteyn den dertienden der voorlede maandt, ſendende daar nevens een Requeſte van den Predikant en geſamentlijcke Ouderlingen van de Franſche Colonie tot Offenbagh, in des ſelfs Landen gelegen, daar by de ſelve verſoecken, vermits den hoogen ouderdom, verſcheydene ſiecktens en andere lichaamelijcke ſwackheden van den Predikant aldaar David Jordan, waar door den ſelven meer als een half jaar langh ſijn beroep niet meer heeft kunnen waarnemen, dat haar Hoogh Mogende gelieven te approbeeren, dat ſeecker *Candidatus Theologiæ*, met name Johan Philip May, buyten eenige belaſtinge des ſelfs dienſt ſoude mogen waarnemen, mits dat by het overlijden van haaren ouden Leeraar, tot een vaſten Succeſſeur aan de ſelve geconſtitueert moge werden op de ſelve tractementen als gemelden Predikant Jordan tegenwoordigh is genietende, en verſoeckende daar op haar Hoogh Mogende favorable diſpoſitie. WAAR op gedelibereert zynde, is goedtgevonden ende verſtaan, dat Copie van de voorſchreve Miſſive en bygevoeghde Requeſte, geſtelt ſal werden in handen van de Heeren van Wynbergen, ende andere haar Hoogh Mogende Gedeputeerden tot de ſaacken van de Finantie, om te viſiteeren, examineeren, ende van alles alhier ter Vergaderinge rapport te doen.

IS ter Vergaderinge geleſen de Requeſte van *Maria Oyens*, Weduwe en Boedelhoudtſter van wylen Cornelis de Jonge van Ellemeét, Heere van Elleméét, in ſijn leven geweeſt zynde Ontfanger generaal der

E X.

Reytenbach, *Leydecker declaratie.* 24.
van Rheede, *Heer tot Ginckel*, *ſes maanden verlof.* 217.
van Riel *geapprobeert als Secretaris van Etten.* 434.
——— *eedt van ſuyveringh.* 469.
Rigail *Brieven van voorſchryvens aan het Hof van Vranckryck.* 844.
——— *Ambaſſadeur aangenomen favorabel te ſullen ſchryven.* 862.
——— *antwoordt.* 926. 935.
Ritzius *aangeſchreeven by het nieuwe Collegie te vigileeren, dat de Obligatien volgens belofte uyt de ſchattingen voldaan mogen werden.* 211.
——— *om ordres en reſolutie.* 511.
——— *reedenen waaromme op de paghtverhuyringe niet na Aurick hadt kunnen gaan.* 597.
——— *klaghten van den Paghter, te examineeren.* 636. 649.
——— *klaghten over wanbetalinge van den Norderpaghter, te examineeren.* 674.
——— *gelaſt ſijn klaghten aan het Furſtelycke Miniſterie op te geven, om ordres, te examineeren.* 723.
——— *rapport wegens difficulteyten in het invorderen der paghtpenningen en reſolutie.* 751.
Robrechtſen *Octroy om te diſponeeren.* 253.
la Rocque *notificeerende het voorgevallene op den laatſten Beededagh.* 240.
——— *gelaſt dien aangaande te procedeeren tegens de Contraventeurs.* 258.
——— *wegens patent voor ſijn Compagnie, de Raadt van Staate te adviſeeren.* 510.
Roelandts *wegens Utrecht gecommitteert in des Generaliteyts Reekenkamer.* 327.
Roeters *en Halmaal Brieven van voorſchryvens aan den Conſul Beeltſnyder.* 74.
——— *item aan den Heer vander Meer.* 475.
Rogh, cum ſuis, *om uytwiſſelingh van Berkhout, de Admiraliteyt van Amſterdam te adviſeeren.* 79.
Romborgh *verſoeck om paſport, afgeweeſen.* 388.
Roſijn, *Weduwe*, *vyftigh guldens toegeleght.* 322.
de Roy *aangeſteldt als Secretaris van Nulandt.* 661.
van Roye *om pardon de Vice Hooghſchout Ghyſen te berichten.* 455.
de Roy, *Weduwe*, *ſes guldens toegeleght.* 8.
——— *verſoeck om alle drie maanden yets tot ſubſiſtentie, afgeweeſen.* 361.
——— *om aſſiſtentie, afgeweeſen.* 348.
de Rue *twee maanden verlof.* 652.
Rumpf, *Envoyé*, *advertentie.* 25. 37. 57. 79.

Figure 1. A column from a resolution page (left side) with meeting date, attendants list and resolutions, and a column from an index page (right side) with indexed terms and page references.

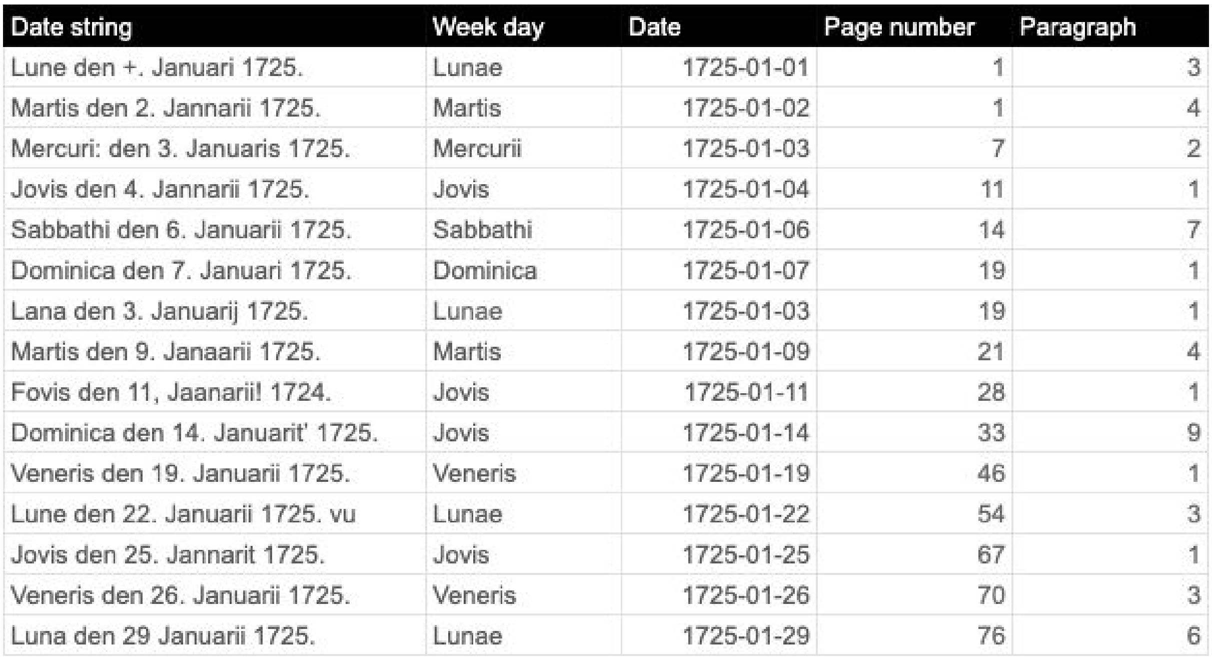
As the resolutions have a fixed format, this allows for segmenting the printed pages using a lot of fixed formulas indicating 1) the start of a new session, 2) a list of representatives attending the session, 3) a number of propositions, that end with an indication of the state of decision (accepted, rejected, or on hold). Finding these parts of text entailed using the paragraphs of the OCRed text and their visual characteristics to identify them. These visual clues were meant to guide contemporary users through the text and find important information quickly. Two examples illustrate this:

1. For the attendance list the capitalized words “PRAESIDE” for the president and “PRAESENTIBUS” for the common members. While these words often did not come out correctly in the OCR process, their place and capitalization made it easier to find them using approximate heuristic tools and in this way identify the attendance lists.
2. Similarly, to determine the chronologically ordered sessions across the pages, we used an iterative approach in which first a rough list of dates for sessions was identified (Table 1), which in a second pass could be completed for the missing sessions, that we then knew to be within a specific range of pages.

We treat the extraction of information as a text collation problem (Gilbert 1973): the resolutions have textual overlap when it comes to the start of a meeting day, the introduction of each individual resolution and its decision paragraph, with some unknown amount of textual variation (including variation in spelling, phrasing and text recognition errors), and the aim is to identify and align the textual repetitions and variations.

The formulaic expressions in the resolutions makes them good candidates for fuzzy searching. In this paper we discuss the methods we developed[1] and the evaluation results of applying them on 100,000 pages and roughly 50 million words of resolution text of the printed resolutions from 1703 until 1796. This same principle was used on the entities in the resolutions. The important persons, institutions and subjects figuring in the resolutions were summarized in the index in the front of the book, with references to the pages. They give both the terms and the location where to find them in the text, which makes it much easier to find them using approximate matching.

[1] See https://github.com/marijnkoolen/fuzzy-search for the fuzzy search library we developed.



Date string	Week day	Date	Page number	Paragraph
Lune den +. Januari 1725.	Lunae	1725-01-01	1	3
Martis den 2. Jannarii 1725.	Martis	1725-01-02	1	4
Mercuri: den 3. Januaris 1725.	Mercurii	1725-01-03	7	2
Jovis den 4. Jannarii 1725.	Jovis	1725-01-04	11	1
Sabbathi den 6. Januarii 1725.	Sabbathi	1725-01-06	14	7
Dominica den 7. Januari 1725.	Dominica	1725-01-07	19	1
Lana den 3. Januarij 1725.	Lunae	1725-01-03	19	1
Martis den 9. Janaarii 1725.	Martis	1725-01-09	21	4
Fovis den 11, Jaanarii! 1724.	Jovis	1725-01-11	28	1
Dominica den 14. Januarit' 1725.	Jovis	1725-01-14	33	9
Veneris den 19. Januarii 1725.	Veneris	1725-01-19	46	1
Lune den 22. Januarii 1725. vu	Lunae	1725-01-22	54	3
Jovis den 25. Jannarit 1725.	Jovis	1725-01-25	67	1
Veneris den 26. Januarii 1725.	Veneris	1725-01-26	70	3
Luna den 29 Januarii 1725.	Lunae	1725-01-29	76	6

Figure 2. Extraction of meeting date metadata and the page and paragraph number where the resolutions of that day start. Column 1 contains the meeting date as found in the OCR output.

Adding Domain-Specific Semantics to Extracted Information

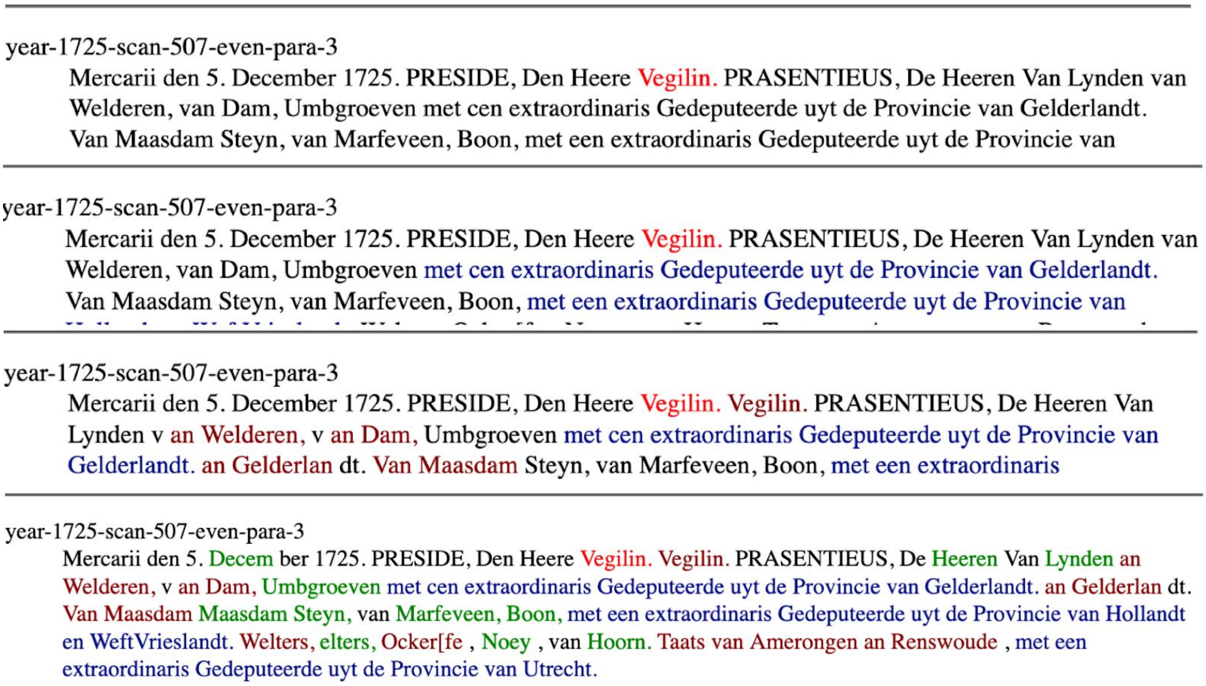
The fuzzy searching algorithm takes as input a manually created phrase model, which is a list of keywords and phrases of interest, where each entry can have an optional list of alternative phrases or spellings, and uses character skip-grams to find candidate strings in the text for each phrase in the model. The searcher can be configured with different thresholds for edit distance and length variations (candidates may be shorter or longer than the phrase in the model).

We used fuzzy string searching to identify formulaic expressions and iteratively built a corpus-specific phrase model with which we identify:

1. the date and attendance list of each meeting, which are followed by all the resolutions of that day,
2. resolution boundaries, e.g. where they start and stop in the running text, so we know which text belongs to which resolution,
3. different types of opening phrases that correspond to different types of sources (e.g. requests, missives, reports, etc., see Figure 2), and
4. the decision paragraphs that state what decision, if any, was reached.

By recognizing the opening of a meeting, we know that the mentioned date is the meeting date, which allows us to not only recognize the date, but also adding a label for what the date refers to.

The attendants lists follow the opening of the meeting, and are structured. The first person mentioned is the president, so the recognized name can be labelled with the role of president. The function of president rotated by week among the provinces and presidents therefore also appear as common delegates in other weeks, where they can be found. This leaves a number of unidentified names to spot. All attendants are grouped by province, so can be labelled with their role as attendant of the meeting and with the province they represent. We have lists of all delegates in the meetings of the States General including years and province, that can be used to identify delegates. As a bonus, many delegates also appear in the resolutions in other roles as commissioners or representatives of the States.



year-1725-scan-507-even-para-3

Mercarii den 5. December 1725. PRESIDE, Den Heere Vegilin. PRASENTIEUS, De Heeren Van Lynden van Welderen, van Dam, Umbgroeven met cen extraordinaris Gedeputeerde uyt de Provincie van Gelderlandt. Van Maasdam Steyn, van Marfeveen, Boon, met een extraordinaris Gedeputeerde uyt de Provincie van

year-1725-scan-507-even-para-3

Mercarii den 5. December 1725. PRESIDE, Den Heere Vegilin. PRASENTIEUS, De Heeren Van Lynden van Welderen, van Dam, Umbgroeven met cen extraordinaris Gedeputeerde uyt de Provincie van Gelderlandt. Van Maasdam Steyn, van Marfeveen, Boon, met een extraordinaris Gedeputeerde uyt de Provincie van

year-1725-scan-507-even-para-3

Mercarii den 5. December 1725. PRESIDE, Den Heere Vegilin. Vegilin. PRASENTIEUS, De Heeren Van Lynden v an Welderen, v an Dam, Umbgroeven met cen extraordinaris Gedeputeerde uyt de Provincie van Gelderlandt. an Gelderlan dt. Van Maasdam Steyn, van Marfeveen, Boon, met een extraordinaris

year-1725-scan-507-even-para-3

Mercarii den 5. Decem ber 1725. PRESIDE, Den Heere Vegilin. Vegilin. PRASENTIEUS, De Heeren Van Lynden an Welderen, v an Dam, Umbgroeven met cen extraordinaris Gedeputeerde uyt de Provincie van Gelderlandt. an Gelderlan dt. Van Maasdam Maasdam Steyn, van Marfeveen, Boon, met een extraordinaris Gedeputeerde uyt de Provincie van Hollandt en WeftVrieslandt. Welters, elters, Ocker[fe , Noey , van Hoorn. Taats van Amerongen an Renswoude , met een extraordinaris Gedeputeerde uyt de Provincie van Utrecht.

Figure 3. Subsequent stages of spotting delegates in the attendance lists

We can add these elements as meaningful metadata labels to individual resolutions, which aids subsequent information access and analysis. Moreover, we can use the consistency of the structure of the resolutions to spot errors. For instance, we can check for cases where an opening formula was found but no decision formula. Or check for missing meeting dates by temporally ordering the ones we did find, or check for OCR errors in delegate names that occur often. In other words, the 'messy' output of the OCR process is turned into what Christoph Schöch (2013) calls 'smart data', i.e. cleaned, structured and semantically explicit layers of metadata.

Evaluation

We built ground truth datasets to evaluate the different elements of our extraction pipeline: 1) traditional NER using only the resolution text to demonstrate the problems encountered with conventional NLP tools, 2) the identification of page types (index page, resolution page, title page), and 3) the phrase model and the fuzzy searching and extraction process for identifying the meeting dates and the individual resolutions and decisions.

For NER, we labelled all named entities in 200 manually transcribed resolution pages. We re-trained the Spacy NER tagger for Dutch[2] using 10-fold cross validation, taking 90% of the pages for training in each fold, and the remaining 10% for testing. Even though these pages have no OCR errors (but no doubt some transcription errors), the best training run resulted in an F_1 score no higher than 0.28 (with a precision of 0.49 and recall of 0.19). Although this can probably be improved with more training data, we conclude that this traditional NLP approach is not suitable for providing high-quality information access.

For the page type identification, we randomly sampled 1696 scans with 3376 pages (most scans contain two pages, some only a single front or back cover) and manually annotated the page type of each page. In the first step, we used layout analysis and fuzzy matching of our phrase models, which achieved an accuracy of 0.91. In the second step, we used the title pages to identify different parts of each book (e.g. the section of index pages, the section of resolution pages) and in each section chose the most frequent page type as the correct page type for all pages in that section. This increased the accuracy to 0.99.

For the ground truth for meeting dates and resolutions, we randomly sampled 300 historical dates in the period 1703-1796, locating the pages containing the resolutions for those dates, and manually annotated the opening of the meeting, the attendance list and all resolution openings and decisions. For the meeting dates and attendance lists, our domain model and fuzzy searching approach reached a precision of 99% and recall of 93%. For the resolution openings and decisions, we reached a precision of 90% and recall of 68%. We have not yet modelled insertions of extracts and letters, and therefore have not evaluated those aspects yet. We are in the process of building a ground truth dataset for the attendants lists.

Reusability and Generalisability

Extracting semantic information requires heuristics for specific digitized resources. Generic approaches of layout analysis can detect standard structures like tables, figures, footnotes, headers and tables of content (Doermann and Tombre 2014, Clausner 2019), but cannot interpret specific semantics such as temporal orderings of meeting dates and the geographical ordering in the attendants lists.

[2] See https://spacy.io/models/nl

The page identification rules and phrase model we developed in the Republic project are specific for the Resolutions of the States General, but the general method is reusable. Many large-scale resources have an internal structure that was created to enhance accessibility. Often, they were part of an information system designed for an analog, or ‘paper’ age, but it pays off to devise ways to transfer this to a digital equivalent. Devices like indexes contain a wealth of information about a resource, its context and its use. In many official resources corpus-specific phrases are used that may be categorized and used for searching. As there is often a degree of variation, the extraction process requires fuzzy searching and matching. Our fuzzy searching module is configurable, with matching thresholds that can be adapted to the peculiarities of the corpus or even to specific parts of the corpus. Fuzzy searching can also help to overcome the limitations of inaccurate OCR and HTR.

Beyond the Republic project, we have used this approach for a number of projects. For example, we have used it for books with medieval charters, in which the standard charter structure allows us to date the 17,000 historical place names mentioned in those charters and use them as place name attestations in historical gazetteers. Another example is the standard structure of advertisements of auctions in 18th century Dutch newspapers, in which brokers, auction date and venue and the auctioned goods can be extracted with a simple phrase model despite high character error rates.

A limitation of our method is that it does not work out-of-the-box but requires a corpus-specific phrase model, as existing information systems and structures are specific to the resource and require information extraction techniques to be adapted to corpus. This requires an iterative approach in which the models will be evaluated and refined. Standard phrases occur in many corpora, but they are usually corpus-specific. Furthermore, even this approach requires a minimum quality of text recognition to have enough contextual information for recognizing phrases.

Our method is still in development, but we think it is possible to make a toolkit with components that are relevant and reusable for many corpora. For different corpora, using the toolkit will be a matter of tuning the components. However, this must be done programmatically; we think that using an approach using just graphical user interfaces would obscure too many of the features and the implications of tuning. For evaluation of intermediate results in the iterative process, however, existing tools can be used.

References

Clausner, C., A. Antonacopoulos, S. Pletschacher (2019). "ICDAR2019 Competition on Recognition of Documents with Complex Layouts – RDCL2019", Proceedings of the 15th

International Conference on Document Analysis and Recognition (ICDAR2019), Sydney, Australia, September 2019, pp. 1521-1526

Colavizza G., Ehrmann M., Bortoluzzi F., 2019, Index-Driven Digitization and Indexation of Historical Archives, Frontiers in Digital Humanities, 6, doi:10.3389/fdigh.2019.00004

Doermann, D. and Tombre, K. ed., 2014. *Handbook of document image processing and recognition*. New York Incorporated: Springer.

Egense, T. (2017). Automated improvement of search in low quality OCR using Word2Vec. Digital Humanities in the Nordic Countries 2nd Conference. https://web.archive.org/web/20180613002357/http://dhn2017.eu/abstracts/#_Toc475332345

van Eijnatten, J., Pieters, T. and Verheul, J., 2013. Big Data for Global History: The transformative promise of digital humanities. *BMGN-Low Countries Historical Review*, *128*(4), pp.55-77.

Gilbert, P. (1973). Automatic collation: A technique for medieval texts. Computers and the Humanities, 7(3), 139-147.

Head, R., 2003, Knowing like a state: the transformation of political knowledge in Swiss archives, 1450-1770. *Journal of Modern History* 75, 745–782. doi: 10.1086/383353

Hill, M. J., & Hengchen, S. (2019). 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.

Hoekstra, R. and Koolen, M., 2019. Data scopes for digital history research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *52*(2), pp.79-94.

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

Leemans, I. B., Maks, E., van der Zwaan, J. M., Kuijpers, H. M. E. P., & Steenbergh, K. (2017). Mining Embodied Emotions: A Comparative Analysis of Bodily Emotion Expressions in Dutch Theatre Texts 1600-1800'. *Digital Humanities Quarterly*, *11*(4). https://doi.org/http://digitalhumanities.org:8081/dhq/vol/11/4/000343/000343.html

Lopresti, D. (2008, January). Measuring the impact of character recognition errors on downstream text analysis. In Document Recognition and Retrieval XV (Vol. 6815, p. 68150G). International Society for Optics and Photonics.

Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S. and Van Harmelen, F., 2015. Semantic Technologies for Historical Research: A survey. *Semantic Web*, *6*(6), pp.539-564.

Mutuvi, S., Doucet, A., Odeo, M., & Jatowt, A. (2018, November). Evaluating the impact of OCR errors on topic modeling. In International Conference on Asian Digital Libraries (pp. 3-14). Springer, Cham.

Opitz, J., L. Born, V. Nastase. Induction of a Large-Scale Knowledge Graph from the Regesta Imperii. Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, 2018

Piersma, H. and Ribbens, K., 2013. Digital Historical Research: Context, Concepts and the Need for Reflection. *BMGN - Low Countries Historical Review*, 128(4), pp.78–102. DOI: http://doi.org/10.18352/bmgn-lchr.9352

Reynaert, M. (2016, May). OCR Post-Correction Evaluation of Early Dutch Books Online-Revisited. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 967-974).

Reynaert, M. (2014, August). TICCLops:: Text-Induced Corpus Clean-up as online processing system. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations (pp. 52-56). Dublin City University and Association for Computational Linguistics.

Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the Humanities. Local Programs, Global Audiences, 2.

van Strien, D., Beelen, K., Coll Ardanuy, M., Hosseini, K., McGillivray, B., & Colavizza, G. (2020 Februari). Assessing the Impact of OCR Quality on Downstream NLP Tasks. ICAART 2020.

Toljamo, T. 2017. "A tailored approach to digitally access and prepare the 1740 Dutch Resolutions of the States General." In *Advances in Digital Scholarly Editing: Papers Presented at the DiXiT Conferences in The Hague, Cologne, and Antwerp*, edited by Peter Boot, Anna Cappellotto, Wout Dillen, Franz Fischer, Aodhán Kelly, Andreas Mertgens, Anna-Maria Sichani, Elena Spadini, and Dirk Van Hulle, 351–56. Sidestone Press. https://www.sidestone.com/books/advances-in-digital-scholarly-editing

Traub, M. C., Van Ossenbruggen, J., & Hardman, L. (2015, September). Impact analysis of OCR quality on research tasks in digital archives. In International Conference on Theory and Practice of Digital Libraries (pp. 252-263). Springer, Cham.

Upward, F., Reed, B., Oliver, G. and Evans, J. (2018), *Recordkeeping Informatics for a Networked Age*. Monash