Reusing Existing Structures for Access to Large Historical Corpora

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

Introduction

Making large historical corpora accessible for research usually involves a pipeline of processing steps, ranging from text recognition to entity spotting, disambiguation, identification and ideally contextualization (Meroño-Peñuela et al. 2015). In many projects much effort is spent on producing a 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. Apart from limited scalability, 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). Extracting and contextualizing information has many issues. Issues such as OCR and HTR errors that make it difficult to use standard NLP tools like NER, topic modelling, POS tagging and sentiment analysis, have been common knowledge for a long time. However, solutions for such issues are scarce and badly documented (e.g. Piersma and Ribbens 2013, van Eijnatten et al. 2013, Leemans et al. 2017).

An additional but rarely studied set of issues relates to the fact that many information units that are typically extracted to improve access, like persons, locations, events, sentiments and topics, are unevenly distributed in corpora (Hoekstra and Koolen 2018). In the case of named entities, highly frequent entities tend to have representations in external knowledge bases, with which they can be disambiguated and issues with spelling variation can to some extent be tackled (Ilievski et al. 2018), but for low frequency entities this is often not possible. At the same time, there tend to be only few high frequency entities, while the vast majority occur only once or a few times, that is, they represent the long tail of the distribution in a corpus (Ilievski et al. 2018, Limpert et al. 2001, Postma et al. 2016). 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.

Existing structure in a corpus comprise text structuring of any kind (division of books, volumes, chapters etc), table of contents, indexes, and visual aids like italics, bold text, newlines, capitals and lines of any kind. They all are expression of the intentions of the creators and/or editors of the corpus (or its edition) and often contain a lot of implicit knowledge about the text. For instance, the meetings of the States General of the Dutch Republic were not only meticulously transcribed, but also indexed at the level of persons, locations and organizations, as this archive would otherwise quickly become inaccessible as it grew in size.

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. We find it remarkable that many digitization programmes take no advantage of the structures and sometimes do not even digitize them, extracting only the main textual body as plain text. There are initiatives to develop generic layout extraction tools[1] but knowledge of individual corpora is needed to understand the semantics of these structures. Our main argument is that creating good access to historical corpora requires solutions that are specific to the individual corpus, that exploit any available structure as much as possible.

Projects and Examples

In this paper, we describe our work on exploiting structure in a number of finished and ongoing projects around large historical corpora, like the General Missives of the Dutch East India Company (VOC)[2] (Hoekstra 2017), the resolutions of the States General of the Dutch Republic[3] (Hoekstra en Nijenhuis 2012, Sluijter et. al. 2016, Toljamo, 2017), the archival records around Dutch migration to Australia[4] (van Faassen 2017, Haentjens Dekker et al. 2016, van Faassen en Hoekstra 2017) and the medieval charter books of Holland and Zeeland.[5]

As a concrete and simple example, the charter books of Holland and Zeeland contain some 3500 charters that were written between the 7th century and 1299. They have a painstakingly compiled index of persons and place names mentioned in the charters, with historical name and spelling variants and references to page and line numbers where they are mentioned. We used layout information from the OCR output to identify the individual charters, which start with a header containing the date it was written, the charter number and where it was found (see Figure 1).[6]

[1] See e.g. the Document and Analysis Recognition competitions, http://icdar2019.org/competitions-2/

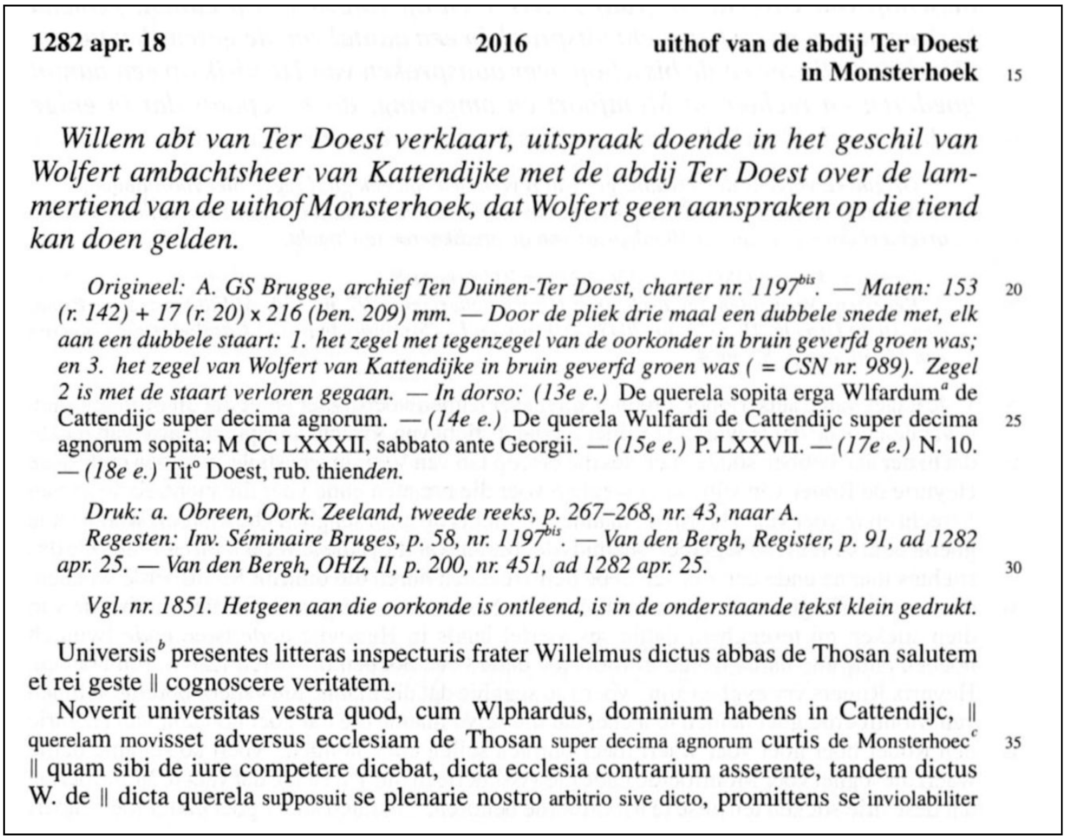
[2] See http://resources.huygens.knaw.nl/retroboeken/generalemissiven/

[3] https://www.huygens.knaw.nl/resoluties-staten-generaal-1576-1796-vroegmoderne-politieke-besluitvorming-en-politieke-dialoog

[4] See https://www.huygens.knaw.nl/migrant-mobilities-and-connection/

[5] See http://resources.huygens.knaw.nl/retroboeken/ohz/

[6] The code for this is available on GitHub: https://github.com/marijnkoolen/digital-history-charter-books



1282 apr. 18 **2016** **uithof van de abdij Ter Doest in Monsterhoek**

Willem abt van Ter Doest verklaart, uitspraak doende in het geschil van Wolfert ambachtsheer van Kattendijke met de abdij Ter Doest over de lammertiend van de uithof Monsterhoek, dat Wolfert geen aanspraken op die tiend kan doen gelden.

Origineel: A. GS Brugge, archief Ten Duinen-Ter Doest, charter nr. 1197bis. — Maten: 153 (r. 142) + 17 (r. 20) x 216 (ben. 209) mm. — Door de pliek drie maal een dubbele snede met, elk aan een dubbele staart: 1. het zegel met tegenzegel van de oorkonder in bruin geverfd groen was; en 3. het zegel van Wolfert van Kattendijke in bruin geverfd groen was (= CSN nr. 989). Zegel 2 is met de staart verloren gegaan. — In dorso: (13e e.) De querela sopita erga Wlfardum[a] de Cattendijc super decima agnorum. *— (14e e.)* De querela Wulfardi de Cattendijc super decima agnorum sopita; M CC LXXXII, sabbato ante Georgii. *— (15e e.)* P. LXXVII. *— (17e e.)* N. 10. *— (18e e.)* Tit^{o} Doest, rub. thien*den.*

Druk: a. Obreen, Oork. Zeeland, tweede reeks, p. 267–268, nr. 43, naar A.
Regesten: Inv. Séminaire Bruges, p. 58, nr. 1197bis. — Van den Bergh, Register, p. 91, ad 1282 apr. 25. — Van den Bergh, OHZ, II, p. 200, nr. 451, ad 1282 apr. 25.

Vgl. nr. 1851. Hetgeen aan die oorkonde is ontleend, is in de onderstaande tekst klein gedrukt.

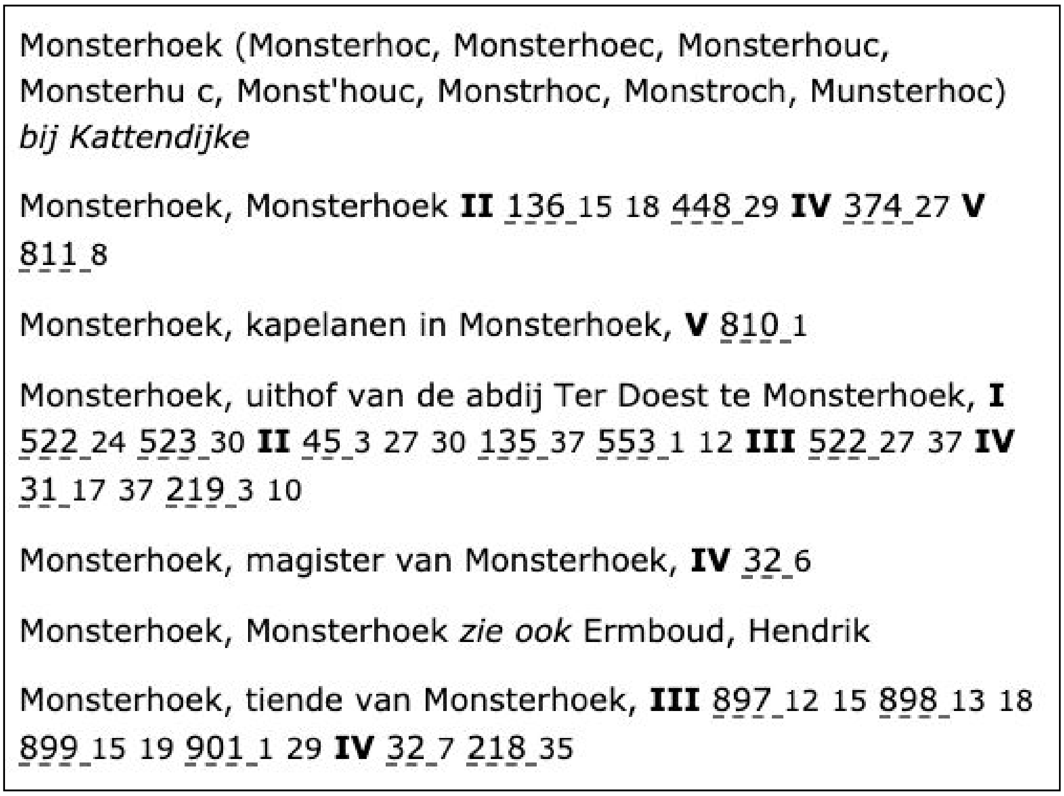
Universis[b] presentes litteras inspecturis frater Willelmus dictus abbas de Thosan salutem et rei geste || cognoscere veritatem.

Noverit universitas vestra quod, cum Wlphardus, dominium habens in Cattendijc, || querelam movisset adversus ecclesiam de Thosan super decima agnorum curtis de Monsterhoec[c] || quam sibi de iure competere dicebat, dicta ecclesia contrarium asserente, tandem dictus W. de || dicta querela supposuit se plenarie nostro arbitrio sive dicto, promittens se inviolabiliter

Figure 1. Charter number 2016 in Oorkondeboeken van Holland en Zeeland tot 1299 mentioning Monsterhoek on line 35, page 218 of part IV.

Because the OCR has errors, the date and find place were not always correctly recognized, but exploiting the fact that the charters are printed in chronological order, and that the left-hand side of the header specifies a date, we could use regular expressions to automatically identify many of the cases that were incorrect, which made the manual correction process manageable. Knowing on which page(s) and lines a charter starts and ends (Figure 2), we linked place names from the index to the charter they are mentioned in and thereby to the date and find- place of that charter.

The result is a list of over 17,000 historical place name attestations with evidence of when specific places existed and which name variants were used. Doing this without the index would be much more difficult, as it would provide no starting point for which historical places and name variants to expect, nor a way to distinguish place names mentioned in the charters or in the commentary, the latter not being historical attestations. The index does not cover all places and persons mentioned and has a few mistakes, and should therefore not be the only way to provide access and extract information, but it captures many low frequency places and offers clear advantages over approaches that treat the entire resource as ‘unstructured’ text but require a much lower OCR error rate.



Monsterhoek (Monsterhoc, Monsterhoec, Monsterhouc, Monsterhu c, Monst'houc, Monstrhoc, Monstroch, Munsterhoc) *bij Kattendijke*

Monsterhoek, Monsterhoek **II** 136 15 18 448 29 **IV** 374 27 **V** 811 8

Monsterhoek, kapelanen in Monsterhoek, **V** 810 1

Monsterhoek, uithof van de abdij Ter Doest te Monsterhoek, **I** 522 24 523 30 **II** 45 3 27 30 135 37 553 1 12 **III** 522 27 37 **IV** 31 17 37 219 3 10

Monsterhoek, magister van Monsterhoek, **IV** 32 6

Monsterhoek, Monsterhoek *zie ook* Ermboud, Hendrik

Monsterhoek, tiende van Monsterhoek, **III** 897 12 15 898 13 18 899 15 19 901 1 29 **IV** 32 7 218 35

Figure 2. Index of the charter books Oorkondeboeken van Holland en Zeeland tot 1299 with an entry for Monsterhoek stating it is mentioned in part IV, page 218, line 35.

Presentation

In the presentation we will also elaborate on more complex examples of exploiting structure to improve access to unevenly distributed entities, such as using the amount of handwritten text on 50,000 migration cards for classification, and using structural conventions for linking low frequency person names to information about their roles and professions. By comparing our findings across these projects and corpora, we draw insights and lessons for digital historical research. Following the advice of Koolen et al. (2018), we discuss biases in information extraction tools, reusable solutions for some of these issues, and provide recommendations for improving how we evaluate their performance.

References

Haentjens Dekker, R., Hoekstra, R. and van Faassen, M., 2016. Bringing Migration Data Into Context Using Digital Computational Methods. In DH (pp. 213-215).

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.

van Faassen, M., 2017. The whereabouts of Migrants: A comparison of Dutch migrant registration systems. International Conference on Information and Power in History.

van Faassen, M. and Hoekstra, R., 2017. Modelling Society through Migration Management. Exploring the role of (Dutch) experts in 20th century international migration policy. Conference paper. Government by Expertise: Technocrats and Technocracy in Western Europe, 1914-1973. Panel 3. Global Expertise.

Hoekstra, F.G., 2017. Rapport Ontsluiting Generale Missiven VOC. https://www.researchgate.net/publication/314232346_Rapport_Ontsluiting_Generale_Missiven_VOC [DOI: 10.13140/RG.2.2.19896.96008]

Hoekstra, R. and Koolen, M., 2018. Data Scopes for Digital History Research. Historical Methods: A Journal of Quantitative and Interdisciplinary History, pp.1-16.

Hoekstra, R., and Nijenhuis, I., 2012. “Enhanced Access for a Paper World.” Paper presented at the European Society for Textual Scholarship 2012, KNAW, Amsterdam, 22–24 November 2012. https://www.researchgate.net/publication/283725866_Enhanced_Access_for_a_Paper_World_Enhanced_Access_for_a_Paper_World [DOI: 10.13140/RG.2.1.2074.6323]

Ilievski, F., Vossen, P. and Schlobach, S., 2018, August. Systematic Study of Long Tail Phenomena in Entity Linking. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 664-674).

Koolen, M., van Gorp, J. and van Ossenbruggen, J., 2018. Toward a Model for Digital Tool Criticism: Reflection as integrative practice. *Digital Scholarship in the Humanities*.

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

Limpert, E., Werner A. Stahel, Markus Abbt; Log-normal Distributions across the Sciences: Keys and Clues: On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question, BioScience, Volume 51, Issue 5, 1 May 2001, Pages 341–352, https://doi.org/10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2

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.

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

Postma, M., Ilievski, F. and Vossen, P., 2018, June. Semeval-2018 task 5: Counting Events and Participants in the long tail. In *Proceedings of The 12th International Workshop on Semantic Evaluation* (pp. 70-80).

Sluijter, Ronald, Marielle Scherer, Sebastiaan Derks, Ida Nijenhuis, Walter Ravenek, and Rik Hoekstra. 2016. “From Handwritten Text to Structured Data: Alternatives to Editing Large Archival Series.” Paper presented at the Digital Humanities 2016, Kraków, 11–16 July 2016. http://dh2016.adho.org/abstracts/36.

Toljamo, Tuomo. 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