The Datafication of Early Modern Ordinances

C. Annemieke Romein[1,2,3], Sara Veldhoen[1], and Michel de Gruijter[1]

[1]KB, National Library of the Netherlands
[2]Ghent University
[3]Erasmus University Rotterdam

Keywords: Early Modern Printed Ordinances; Text recognition; Text segmentation; Categorisation; Machine Learning; Annif; Transkribus; Dutch Gothic Print.

The project *Entangled Histories* used early modern printed normative texts. The computer used to have significant problems being able to read Dutch Gothic print, which is used in the vast majority of the sources. Using the Handwritten Text Recognition suite Transkribus (v.1.07-v.1.10), we reprocessed the original scans that had poor quality OCR, obtaining a Character Error Rate (CER) much lower than our initial expectations of <5% CER. This result is a significant improvement that enables the searching through 75,000 pages of printed normative texts from the seventeen provinces, also known as the Low Countries.

The books of ordinances are compilations; thus, segmentation is essential to retrace the individual norms. We have applied – and compared – four different methods: ABBYY, P2PaLA, NLE Document Recognition and a custom rule-based tool that combines lexical features with font recognition.

Each text (norm) in the books concerns one or more topics or *categories*. A selection of normative texts was manually labelled with internationally used (hierarchical) categories. Using Annif, a tool for automatic subject indexing, the computer was trained to apply the categories by itself. Automatic metadata makes it easier to search relevant texts and allows further analysis.

Text recognition, segmentation and categorisation of norms together constitute the datafication of the Early Modern Ordinances. Our experiments for automating these steps have resulted in a provisional process for datafication of this and similar collections.

1 Introduction

Normative rules - from any era - provide very versatile insights into society, as there are many (hidden) layers within such texts. They can, for instance, shed light on ideas on and interpretations of the organisation of society; indicate what troubles society faced; provide insights into communication patterns. In the early modern period (±1500 - 1800) rules were announced by a city crier. He walked through the city or rode a horse to rural villages in order to visit contractually-indicated locations to proclaim new rules and repeat older ones to the local residents. After reading them aloud, the printed texts were fixed to 'well-known places' (e.g. at the church door, trading places (see Figure 1), or at the market square) for people to be able to reread them. Sometimes there was merely a duty to affix the rules, an obligation that was carried out by the city's affixer (Dut. *stadsaanplakker*) (Der Weduwen, 2018). In the mid-seventeenth century about 50% of all urban residents could read in the Republic, so for the remainder of the residents having the new rules read aloud was still very important (Hoftijzer, 2015). The rules had to make sense, so people could remember them by heart. Hence, there is a repetitiveness in the texts – which makes sense given that the 16th and 17th century had an important oral tradition.

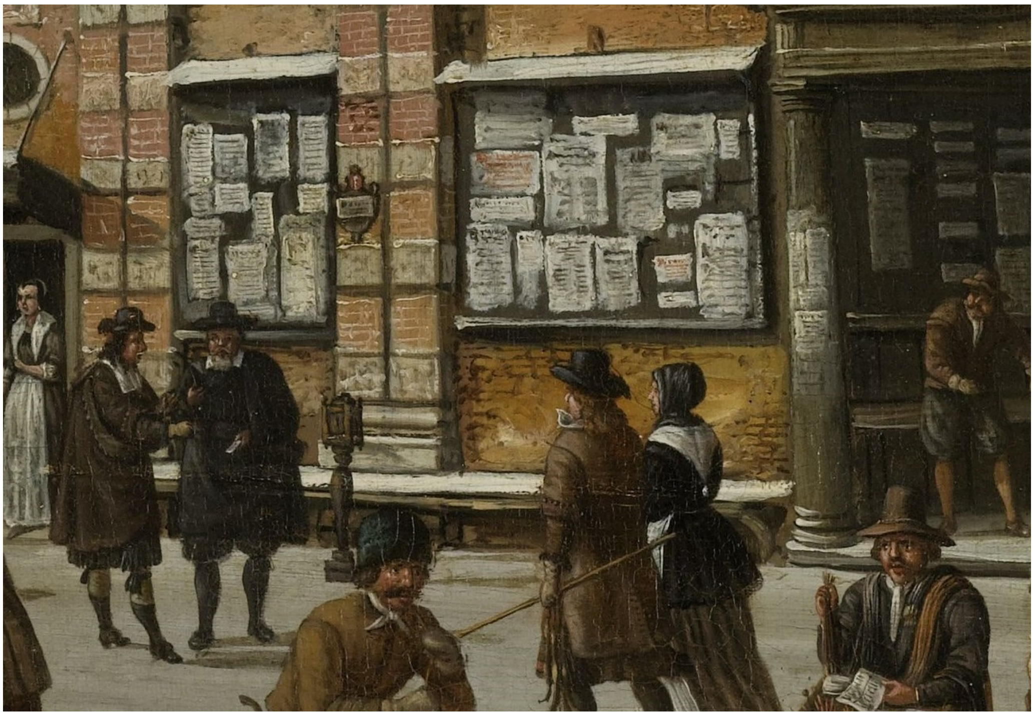


Figure 1: Detail from The Paalhuis and the New Bridge (Amsterdam) during the winter, by Jan Abrahamsz. Beerstraten, c. 1640 – 1666. Oil Painting, 84 × 100 cm. Source: `http://hdl.handle.net/10934/RM0001.COLLECT.5966`

The affixation of ordinances, or placards, to known places made them official, for if a rule remained unknown to the public, it could (and would) not be obeyed. The federation-states (Dutch: *gewestelijke staten* or 'provincial' estates[1]) considered it to be essential to also print a selection of their agreed-upon texts in books of ordinances (Dut. *plakkaatboeken*). These volumes formed, e.g. a source for lawyers as a reference work, but cannot be considered a complete overview. In most cases, they merely provide an indication of what government officials deemed essential rules. These folio books were much more manageable to handle than the original offprints that could size around

[1] The English translation *province* for the Dutch word *gewest* can be misleading, as it has the connotation of being subordinate to another entity – while the Low Countries' provinces were basically federation-states which is why we use this latter term.

30x40 cm, as the used font is much smaller too.

1.1 Hypothesis

Both the Dutch Republic and the Habsburg Netherlands were federations of autonomous states. In the Republic, the Estates-General held sovereign powers, and in each of the federation-states, the estates held the highest power. The Republic's Stadtholders were officially civil servants. The Habsburg Netherlands differed from the Republic as they had a sovereign prince (the King of Spain), though the federation-states did have a certain amount of freedom. They had to verify that new rules did not jeopardise traditions and customs. However, when one looks in many history books, the early modern European-scene is depicted as a conflict among the noble dynasties; in other words, there is a strong focus on monarchies.

This focus results in poorly studied political-institutional constellations of multi-layered republics - the Dutch Republic and Switzerland alike. We know too little to say something concrete about the rule of federation-states. While Belgium has a long tradition of republishing the rules through the Royal Commission for the Publication of Ancient Laws and Ordinances[2] (since 1846), the Netherlands do not hold such an institute. This knowledge-gap resulted in a study between Holland and Flanders – two federation-states that are trade-oriented – indicating that the differences in the Republic's and Habsburg's legislation were not that significant (Romein, 2019). Hence, the following hypothesis arose:

> Early Modern European states struggled for survival, making it impossible to 'reinvent the wheel' each time a problem arose. Hence, it was of tremendous importance to copy, adapt and implement normative rules (often understood as legislation) that were already proven successful elsewhere.

In order to be able to study this hypothesis, a massive amount of data needs to be generated, categorised and analysed. When that data is available, it will be possible to create a topical subset which will then allow textual comparisons. The first Digital Humanities building-block in this puzzle became the KB Researcher-in-residence[3] project *Entangled Histories*. In this article, we present research informed by this hypothesis as a preliminary analysis; however, a more rigorous analysis will be done in the future. In this article, we explain the process of datafication of the Early Modern Ordinances that consisted of text recognition, segmentation and automated topic classification of the sources. We elaborate on the challenges and prospects of this (type of) research.

1.2 Digitisation of Books of Ordinances

The digitisation of books of ordinances may not have been a priority of libraries, as most of these books are reasonably readable by researchers. However, this close reading requires reliance on the provided indexes - if any is available. The number of pages per book in our set ranges from 34 to 1921 pages and a lot of normative rules are contained in one volume, making it a challenge to read all and everything. Linguistic challenges such as spelling variation and the way of referencing to specific problems pose another challenge.

[2] See `https://justitie.belgium.be/nl/informatie/bibliotheek/koninklijke_commissie_uitgave_belgische_oude_wetten_en_verordeningen`

[3] See `https://www.kb.nl/en/organisation/research-expertise/researcher-in-residence`

For example, the word ‘gipsy’ (Dut. *zigeuner* or Fr. *manouche*) as a reference to the Roma or Sinti people is not in the texts, but the words *heiden* or *heidens* (Eng. *infidels*) are. As language - and thus, references - change over time, a book from the 1600s could have chosen a specific word in the index, whereas a book from the 1700s could have chosen another word - which complicates (automatic) searches. Full-text searches thus require complete access to the sources.

To be able to start the datafication-process, an overview of the available books of ordinances is required, which did not exist. Hence, as a by-product of *Entangled Histories*, we have published a list of 108 digitised books of ordinances.[4] The list is presumably incomplete, primarily since it is unknown which books of ordinances were printed. Hence, this should also serve as an invitation to inform us about excluded books.

The available digitised books can be distinguished into various groups, not just per publisher or per federation-state. Eighty-eight of these books were printed in a Roman-type font, twenty in a Dutch Gothic font - which differs from the German Gothic font (*Fraktur*). The language in the books varies among the regions, but the main languages are Dutch (67), French (26), Latin (1) and a mix of those (14). The total amount of pages is approximately 75,000, resulting in an estimate of 550 million characters to process. All books have been published between 1532 and 1789, within the seventeen federation-states of the Habsburg Netherlands and the Dutch Republic.[5]

1.3 Datafication

In this paper, we describe the datafication of the collected books of ordinances. We discern three phases in the datafication: text recognition, segmentation and categorisation, which provide the structure of this document.

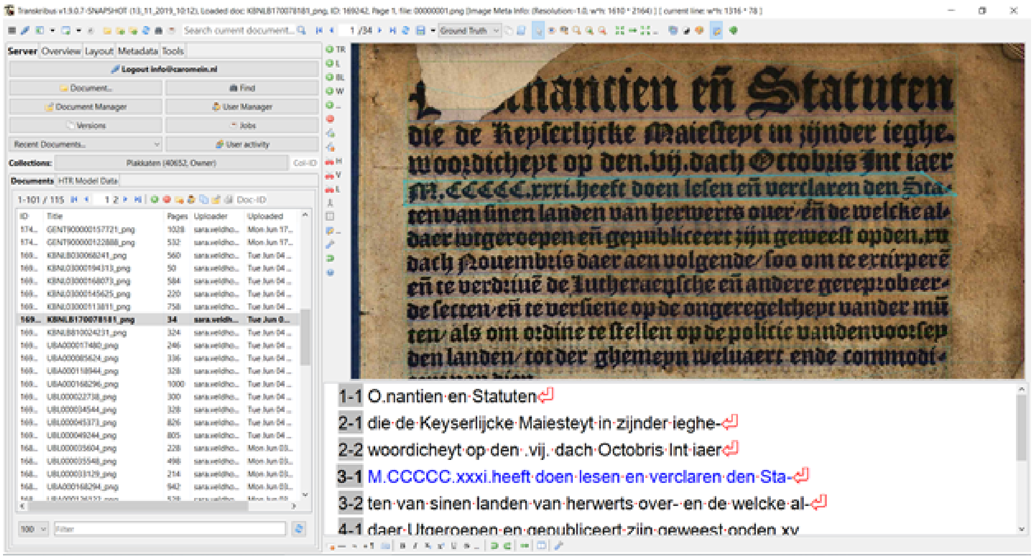
We used Transkribus for text-recognition; which yielded good results for the entire corpus. Due to time limitations, we have chosen to work with a Proof of Concept for the next two phases. For this purpose we selected the *Groot Gelders Placaet-boeck, Volume 2*. The first author had previously made a list of titles with topic annotations for this volume, that could serve as a basis for our experiments. We experimented with several tools for segmentation, and applied Annif for the categorisation.

2 Text Recognition

108 books have been digitised either by the library’s initiatives to digitise books (University Utrecht, Bodleian Library) or through the Google Books project[6]. These books have been processed for Optical Character Recognition (OCR) by a version of ABBYY FineReader. In the OCR-technology “[...] scanned images of printed text are converted into machine-encoded text, generally by comparing individual characters with existing templates”(Muehlberger et al., 2019, p. 955). Manual inspection of the OCR-output for some of the books in our study revealed they were completely incomprehensible. Unfortunately, the older the books are, the more problematic OCR is – especially when printed in Gothic script. Even the Roman-type font is a challenge with the long s (ſ) a character that resembles the f to a great extent.

[4] See `https://lab.kb.nl/dataset/entangled-histories-ordinances-low-countries` and archived in Zenodo: `https://doi.org/10.5281/zenodo.3567844`.

[5] It thus excludes the Prince-Bishopric of Liège and the principality of Stavelot, which were not



ID	Title	Pages	Uploader	Uploaded
174...	GENT900000157721_png	1028	sara.veldho...	Mon Jun 17...
174...	GENT900000122888_png	532	sara.veldho...	Mon Jun 17...
169...	KBNLB030068241_png	560	sara.veldho...	Tue Jun 04 ...
169...	KBNL03000194313_png	50	sara.veldho...	Tue Jun 04 ...
169...	KBNL03000168073_png	584	sara.veldho...	Tue Jun 04 ...
169...	KBNL03000145625_png	220	sara.veldho...	Tue Jun 04 ...
169...	KBNL03000113811_png	758	sara.veldho...	Tue Jun 04 ...
169...	**KBNLB170078181_png**	**34**	**sara.veldh...**	**Tue Jun 0...**
169...	KBNLB810024231_png	324	sara.veldho...	Tue Jun 04 ...
169...	UBA000017480_png	246	sara.veldho...	Tue Jun 04 ...
169...	UBA000085624_png	336	sara.veldho...	Tue Jun 04 ...
169...	UBA000118944_png	328	sara.veldho...	Tue Jun 04 ...
169...	UBA000168296_png	1000	sara.veldho...	Tue Jun 04 ...
169...	UBL000022738_png	300	sara.veldho...	Tue Jun 04 ...
169...	UBL000034544_png	328	sara.veldho...	Tue Jun 04 ...
169...	UBL000045373_png	826	sara.veldho...	Tue Jun 04 ...
169...	UBL000049244_png	805	sara.veldho...	Tue Jun 04 ...
168...	UBL000035604_png	228	sara.veldho...	Mon Jun 03...
168...	UBL000035548_png	498	sara.veldho...	Mon Jun 03...
168...	UBL000033129_png	214	sara.veldho...	Mon Jun 03...
168...	UBA000168294_png	942	sara.veldho...	Mon Jun 03...

1-1 O.nantien·en·Statuten
2-1 die·de·Keyserlijcke·Maiesteyt·in·zijnder·ieghe-
2-2 woordicheyt·op·den·.vij.·dach·Octobris·Int·iaer
3-1 M.CCCCC.xxxi.heeft·doen·lesen·en·verclaren·den·Sta-
3-2 ten·van·sinen·landen·van·herwerts·over-·en·de·welcke·al-
4-1 daer·Utgeroepen·en·gepubliceert·zijn·geweest·opden.xv

Figure 2: Screenshot of Transkribus (v. 1.9.0.7) - showing a 16th century Dutch Gothic ordinance and the transcription.

2.1 Method: Automatic Text Recognition (ATR)

Properly training OCR tools such as Kraken, Tesseract or, indeed, ABBYY could lead to more recognition of gothic script (Tafti et al., 2016)[7]. However, we wanted to test another potential method: Handwritten Text Recognition (HTR). Within the Recognition and Enrichment of Archival Documents-project (READ) Transkribus applies this HTR-method.

Given the complexity of handwriting a combination of techniques is used: advanced pattern recognition is combined with artificial intelligence, and recurrent neural networks. These three are employed to recognise various hands after training (Muehlberger et al., 2019). This technique can also be applied to complex printed texts, such as Sanskrit, Cyrillic, or, indeed, early modern printed texts from Western Europe, such as Gothic (see Figure 3). As the technique is used for handwriting as well as print, the term Automatic Text Recognition (ATR) is applied. This tool has a graphic user interface, so it is a tremendous asset for the traditionally trained humanities-scholars allowing a transformation of sources into searchable data.

When applying ATR to printed texts, the user trains the computer to regard the characters as 'impeccable handwriting'. With that in mind, we expected to be able to recognise texts with a Character Error Rate (CER) of less than 5%.

The texts were uploaded into Transkribus in the PNG-format[8]. Then, the follow-

part of these entities.

[6] See https://books.google.nl/

[7] OCR is being developed further and other interesting tests could be made - but were not included in this project:, e.g. Konstantin Baierer, Rui Dong, and Clemens Neudecker. 2019. Okralact - a multi-engine Open Source OCR training system. In Proceedings of the 5th International Workshop on Historical Document Imaging and Processing (HIP '19). Association for Computing Machinery, New York, NY, USA, 25–30, DOI:https://doi.org/10.1145/3352631.3352638; Reul, C.; Christ, D.; Hartelt, A.; Balbach, N.; Wehner, M.; Springmann, U.; Wick, C.; Grundig, C.; Büttner, A.; Puppe, F. 'OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings'. Appl. Sci. 2019, 9, 4853.https://doi.org/10.3390/app9224853; or at https://ocr-d.de/.

[8] Transkribus allows several formats, including JPG, TIFF, PNG, PDF. However, it does not allow JPEG2000, the much compressed format in which most Google Books are saved. We therefore converted

ing steps were taken: (1) the automatic layout analysis, and (2) the partial manual transcription of the books to develop a Ground Truth to develop models in order to transcribe the rest of the texts automatically. It entailed manual transcription of a minimum of 50-75 pages per font-type/ per period.

PLACCAET
ENDE
Ordonnantie op 't Stuck
van de Generale ende Gemeene
Middelen / by de Staten van de Ver-
eenighde Provincien bewilliget tot onder-
houdt van de gemeene Saecke.
Inhoudende mede 't Verbodt met de Vyanden
ofte op der Vyanden Landen te handelen.

De Staten
Generael der Ver-
eenichde Nederlanden /
Allen den ghenen die
deſe tegenwoordige ge-
toont ſullen worden /
Saluyt: Alſoo omme
metten gevoegelijckſten /
gelijckmatichſten / en-
de minſtquetſelijckſ
Middelen / te
loge / tot beſ

Figure 3: A Gothic printed book of Ordinances (KB National Library of the Netherlands).

2.2 Results: ATR

Within *Entangled Histories*, we chose to combine the material into groups related to the font and language. Hence, it has resulted in three self-created models: *Gothic_Dutch_Print*; *French_18thC_Print*; and *Romantype_Dutch_Print*.[9] For our Latin book from Artois, we applied the publicly available Latin model Noscemus GM v1 created by Stefan Zathammer.[10]

Each of these models has been created with the use of both a train and a test set. The test set was used to test the model and predict its ability to work for unseen material. It is crucial to prevent the overfitting of a model for a specific type of text. As a rule of thumb deep-learning-expert Gundram Leifert (CITlab University of Rostock/ Planet AI GmbH), who develops the HTR-component of Transkribus, advices to use a minimum of 1000 lines of text of which 10% is entered as validation.

The results as presented in Table 1, with CER's of 1.71% (Dutch_Gothic_Print), 0.65% (French_18thC_Print) and 1.17% (Dutch_Romantype_Print), exceeded our expectations as well as goals for the Entangled Histories-project tremendously. With those excellent CER-results, the datafication of the original source-material - the books of ordinances -

them to PNG using imagemagick (`https://imagemagick.org/`).

[9] The option to train models is not standard in the GUI, it needs to be requested by email (email@transkribus.eu). HTR+ is an additional feature, allowing the Recurrent Neural Networks to run more than the standard 40 epochs (in HTR) to 200 epochs (standard), although it can be raised to a manually altered number of epochs of over 1200.

[10] The model Noscemus GM v1 comprises 170658 words and 27296 lines, it shows a CER of 0.87% on the training set and 0.92% on the test set. This HTR+-model is tailored towards transcribing (Neo-)Latin texts set in Antigua-based typefaces, but it also, to a certain degree, able to handle Greek words and words set in (German) Fraktur.

Table 1: Results per created model (CER).

Model name [ID]	Training (CER)	Test (CER)	# Words (training)	# Lines (training)
Dutch_Gothic_Print [Model ID18944]	0.22%	1.71%	51143	7143
French_18thC_printed [Model ID19166]	0.33%	0.65%	38487	3883
Romantype_Dutch_Print [Model ID19423]	1.26%	1.17%	88105	13013

is even much better than expected. In other words, the texts will be well human- and machine-readable.

3 Segmentation

How can the computer segment a text? The human eye can easily spot different parts on a page, such as headers, footers, marginalia, titles and paragraphs in contrast to a computer. Within *Entangled Histories*, several tools were explored: *ABBYY FineReader v.11*, *P2PaLA*, *NLE Document Understanding* and a rule-based approach we created ad hoc as a backup. Initially, the expectation was that assigning layout structures within Transkribus would be regarded as text-enrichment, but it turned out to be the first step in the analysis process. Although we tested several segmentation-tools within *Entangled Histories*, we could not rely on them as they became available too late or have not yet reached their full potential. However, P2PaLA and NLE Document Understanding are very promising tools-in-development. We therefore created a tool for segmenting the proof of concept ad hoc.

3.1 Method

Several segmentation options were investigated on various volumes within the collection.

ABBYY FineReader v.11

As an OCR-engine, built into Transkribus, we initially used ABBYY to recognise columns, whereas the Transkribus automatic layout analysis (LA) could not recognise them properly.[11] Interestingly, ABBYY also adds information regarding the size of fonts to the XML-data that is provided. Which - at least for printed texts – could be used to discern different parameters such as e.g. columns, font-size. However, this is not foolproof as ABBYY regularly fails or adds information where it should not.

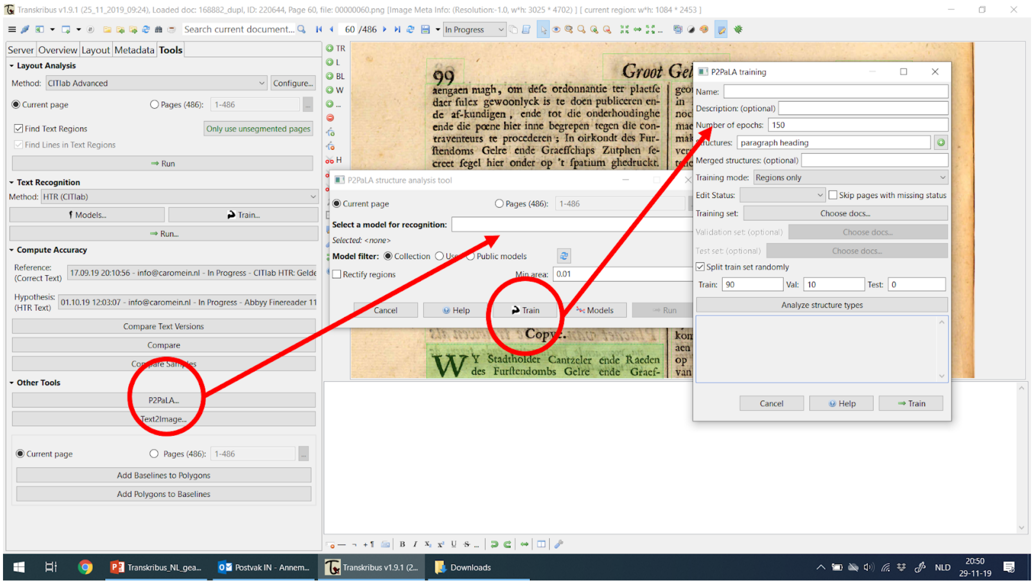
P2PaLA (Alpha-version)

Lorenzo Quirós Díaz is developing Page to Page Layout Analysis (P2PaLA) at the UPVLC (Universidad Politécnica de Valencia), a READ-COOP within the research-centre: Pattern Recognition and Human Language Technology (PRHLT).[12] The idea(l)

[11] See `https://www.abbyy.com/media/10433/what-is-new-in-finereader-engine-en.pdf`

[12] See `https://www.prhlt.upv.es/wp/`

is to train several pages by labelling the text regions (without baselines) and telling the computer what is so special about the fields (Quirós, 2018). The models operate pixel-based, and (text) regions could be any shape. This flexibility comes with a price: many data points are needed per region type. At the moment of testing, we were advised to annotate about 200 pages per (type of) book, with at most five different region types to be tagged.


Transkribus v1.9.1 (25_11_2019_09:24), Loaded doc: 168882_dupl, ID: 220644, Page 60, file: 00000060.png [Image Meta Info: (Resolution:-1.0, w*h: 3025 * 4702)] [current region: w*h: 1084 * 2453]

Layout Analysis
Method: CITlab Advanced
● Current page ○ Pages (486): 1-486
☑ Find Text Regions
☑ Find Lines in Text Regions

Text Recognition
Method: HTR (CITlab)

Compute Accuracy
Reference: (Correct Text) 17.09.19 20:10:56 - info@caromein.nl - In Progress - CITlab HTR: Gelde
Hypothesis: (HTR Text) 01.10.19 12:03:07 - info@caromein.nl - In Progress - Abbyy Finereader 11

Other Tools
● Current page ○ Pages (486): 1-486

P2PaLA structure analysis tool
● Current page ○ Pages (486): 1-486
Select a model for recognition:
Selected: <none>
Model filter: ● Collection ○ User ○ Public models
□ Rectify regions
Min area: 0.01

P2PaLA training
Name:
Description: (optional)
Number of epochs: 150
Structures: paragraph heading
Merged structures: (optional)
Training mode: Regions only
Edit Status:
□ Skip pages with missing status
Training set:
Validation set: (optional)
Test set: (optional)
☑ Split train set randomly
Train: 90 Val: 10 Test: 0

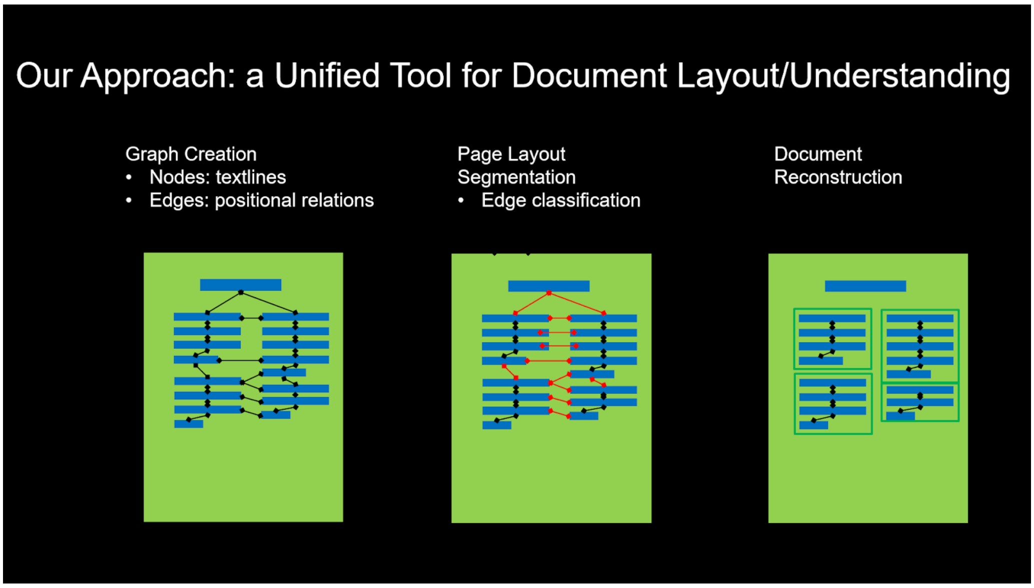
Figure 4: Screenshot of Transkribus (v. 1.9.1) - showing the training-screen for P2PaLA. In the right pop-up screen, several 'structures' (fourth field) can be selected to train an LA-model.

At the moment that we tested P2PaLA, it still required external involvement from Innsbruck to set up the training and implement the trained model into the P2PaLA-module. As of December 2019, a selected group of alpha-users can train their own P2PaLA-models within Transkribus and use this straight-away (as shown in Figure 4).

NLE Document Understanding

Like P2PaLA, NLE Document Understanding is a tool under development. NLE stands for Naver Labs Europe, which is one of the READ-COOP partners to develop tools to process texts better. At this point, NLE Document Understanding still requires external help to process the analysis, but it is expected to be incorporated into Transkribus in 2020. Using Artificial Intelligence, the page-layout is processed into nodes and edges and consequently classified in order to reconstruct the role and position of the text within the document (Clinchant et al., 2018, Prasad et al., 2019) (Figure 5). Here a crucial element, addressed by Koolen and Hoekstra (2019) at the DHBenelux 2019 conference, is obeyed: text was not placed at a specific spot by accident, there was a deliberate thought behind it.

Naver Labs' approach can be applied to tables as well as other document structures. Although very promising, the last-minute availability of this approach prevented us from exploring the options further.



Our Approach: a Unified Tool for Document Layout/Understanding

Graph Creation
- Nodes: textlines
- Edges: positional relations

Page Layout Segmentation
- Edge classification

Document Reconstruction

Figure 5: Slide Naver Labs Europe DevView 2019 (booth presentation) *READ: Recognition and Enrichment of Archival Documents. Digital preservation and discovery of the past* (slide 11/23). By Jean-Luc Meunier and Hervé Dejean.

Rule-based approach

Using the *Groot Gelders Placaet-boeck, volume 2* as the book to get a proof of concept on had one significant benefit: the 470 laws in this volume had been manually annotated with metadata including the titles and an indication on which page the title could be found. Using the information on font size from ABBYY whenever available, together with keyword matching on common title words, we developed a script[13] to trace the titles - and thus paragraphs - within the document. Although the wider applicability of this approach is quite limited, it enabled us to pursue our proof of concept.

3.2 Results

The layout of original documents secures much information. Not being able to recognise this structure with a computer and having to find means to re-implement the structure afterwards is a waste of energy, especially when one needs to reprocess HTR-models after applying lay-out analysis. When the returned results are adequate though, it will be worth the effort.

P2PaLA

We provided two books with structural tags indicating left/right paragraph, heading, header, page number and marginalia to be tested within P2PaLA. For each of the books, between 150 and 200 pages were marked. The results were ambiguous: one trained model gave ambiguous results. As this tool is still much in development, not much can be said about the results. This inconclusiveness - in the pre-alpha phase - left us no choice but to abandon this route for now.

[13] The code (XSLT transformation sheets and a Python notebook) for this rule-based tool can be found on github: `https://github.com/KBNLresearch/EntangledHistories`

NLE Document Understanding

The initial training on a handful of pages through NLE Document Understanding resulted in - at that moment - an accuracy of 85% correctly performed layout analysis. They claim to be able to reach a 95% correct performed layout analysis: providing that a training set of pages - including ATR-transcriptions - is representative for the document's structure.[14]

Rule-based

Based on a combination of typographic information derived from ABBYY and matching common title-words, we were able to recognise titles with 95% accuracy. We were able to segment the book into individual laws under the admissible hypothesis that all text following one title until the next title belongs to the same law.

4 Categorisation

Classifying documents as belonging to topics generally improves searchability. Moreover, topics or categories can inform one about the relations between texts from different books (provinces) without requiring close reading. As such, automatic categorisation could help fine-tuning a selection which would in turn allow the primary hypothesis to be answered. For now, we applied it to the single book in our proof of concept. We manually annotated the laws with topics from a controlled vocabulary and used the annotations to train an automatic subject indexing tool called Annif.

4.1 Method

We applied a controlled subject vocabulary, in which the norms were labelled with subjects from a categorisation created by the German Max-Planck-Institute for European Legal History (MPIeR). The same categories have been applied internationally, in over 15 early modern European states. It was developed in the projects Repertorium der Policeyordnungen[15], and Gute Policey und Policeywissenschaft[16] ran by Karl Härter and Michael Stolleis (Kotkas, 2014, Stolleis et al., 1996).

Within the MPIeR-project a four-level deep hierarchical categorisation considering police ordinances (public law) was designed. The books of ordinances also contain international laws that are out of the scope of these categories. For this reason, we added another level (level 1) to distinguish 'Police Legislation' and 'International Law'. In Figure 6, the five categories (at level 2) in Police legislation are displayed, together with subcategories that occur in our dataset. For levels 3-5, the number of available categories are 25, 163 and 1584 respectively. Note that the deepest level (level 5), is open for adding extra terms.

[14] Naver Labs Europe DevView 2019 (booth presentation) READ: Recognition and Enrichment of Archival Documents. Digital preservation and discovery of the past (slide 15 and 16/23). By Jean-Luc Meunier and Hervé Dejean. Data from the Passau archives and personal communications.

[15] See https://www.rg.mpg.de/forschungsprojekt/repertorium-der-policeyordnungen?c=2124983

[16] See https://www.rg.mpg.de/1928092/gute-policey-administrative-law-and-the-science-of-public-affairs

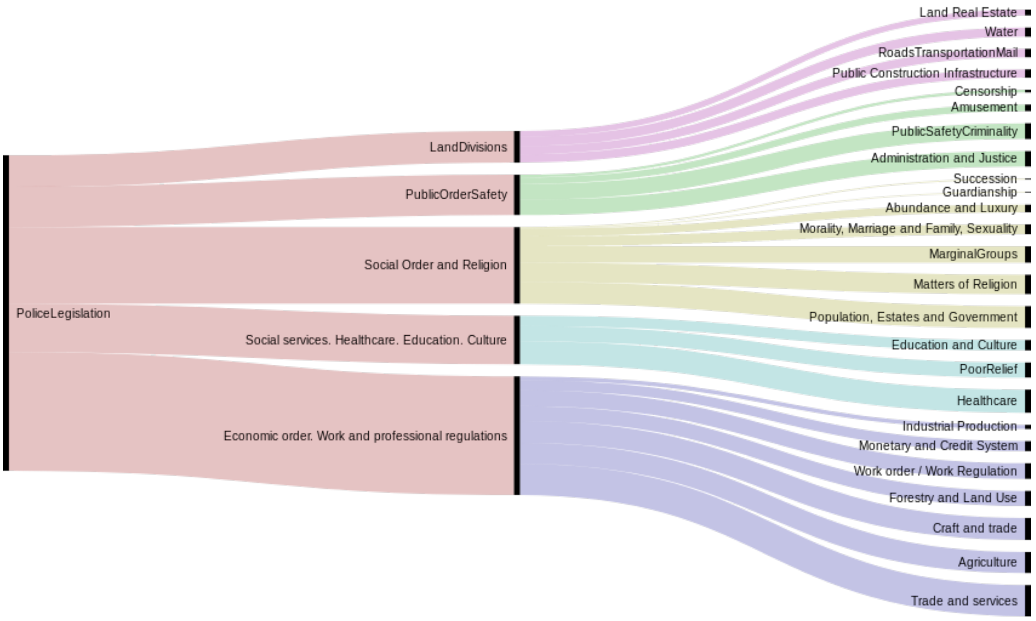


Figure 6: The first three levels for the top category 'Police legislation'. Bar width indicates the number of texts in each category. Through: Alluvial Diagram - `https://app.rawgraphs.io/`

Annotations

Since we are dealing with normative texts (legislation), the texts are relatively unambiguous compared to many other text genres. Manual annotations were added through close-reading and selecting the appropriate topic categories from the available list. Still, a single text can concern several (up to 10) topics at once. On average, each law was annotated with 3.3 categories, as detailed as possible: 69% and 28% of the annotations concerned categories at level 5 and 4, respectively. The topics are quite distinguishable until at least level three. For example, it is quite apparent to distinguish between several economic professions, or between primary school and university, or indeed, whether a rule applies to marriage or adultery.

Annif

Annif is a toolkit for automated subject indexing, developed at the Finnish National Library.[17] From existing metadata with subject headings from a controlled vocabulary, it can learn how to assign those headings to new, unseen data (Suominen, 2019). The tool comes with a variety of back-ends, ranging from lexical methods to vector-space models. It also offers ensemble learning, allowing one to combine the strengths of trained models from different set-ups.

In our experiments, we focused on TF-IDF first in order to see whether any reasonable categorisation could be found, as it is an accessible back-end that can be used without much adjustment. The terms (words) in every document are weighted by their frequency of occurrence (TF), and compensated by the inverse frequency in the entire corpus (IDF). The term frequencies in new documents are compared to those in existing documents, for which the subjects are known. For this, Annif uses the implementation in Gensim (Řehůřek and Sojka, 2010).

[17] Project description and link to source code can be found at `http://annif.org/`

Most back-ends, including TF-IDF, rely on stemming or lemmatization to unify inflections of content words. We used the Dutch snowball analyser as implemented in nltk[18], although ideally one may want to develop an analyser that is tailored for historical Dutch. Note that remaining character errors, as well as inherent spelling variation in historical texts, may influence both the stemming and the generalising capabilities of Annif models.

The hierarchical nature of the categories could be informative for the automatic categorisation. To allow the hierarchical structure to be imported into Annif, we transformed the vocabulary into a *Simple Knowledge Organization System-format* (SKOS) (Tennis and Sutton, 2008). The provisional SKOS-file is archived in Zenodo.[19] Unfortunately, most back-ends in Annif are currently unaware of hierarchy in the subject vocabulary, except for the Maui Server back-end.

4.2 Results

We were particularly interested in the performance of subject indexing at different levels (depths) of the subject vocabulary. Therefore, we created five versions of the dataset: one for each level, where the document would link to the hierarchical ancestor(s) of its assigned topic(s). In the subsequent analysis, level 1 indicates the distinction between international law vs. police legislation (i.e. public law) and the subsequent levels in the hierarchy of the MPIeR categories.

Due to the limited amount of data in our proof of concept-phase - a mere 470 laws - we ran all the experiments using a 10-fold cross-validation with a 90/10 train/test split. Due to time constraints, the only Annif back-end we have been able to experiment with so far was TF-IDF.

The HPC team of Ghent University (Belgium) has been so kind as to provide access to their infrastructure for running the categorisation experiments. Although Annif is not particularly heavy software, the different back-ends may be more demanding. As such, running experiments in an HPC environment could prove quite useful, especially when testing with more massive datasets.

Precision@1 vs. majority baseline

The majority baseline was determined per hierarchy level as the ratio of documents that were annotated with the most common category in that level. It was computed based on all data (no train/test split). In Figure 7, we present precision@1 to compare the model performance to the majority baseline. Precision@1 indicates the accuracy of the most probable category for every law, as proposed by the model. The test precision shows a lot of variances, which can be attributed to the limited amount of data (10% of 470 documents) on which the scores were based. Train precision is typically higher than test precision, indicating a lack of generalization that is probably due to the limited amount of data as well. For levels 1 and 2, the majority baseline was not even reached by the model. This is probably due to the imbalance in the data: there are relatively few examples to get informed about non-dominant categories. Deeper in the hierarchy, the distribution over topics is more even, and the majority baseline starts to fail, while the model performance remains stable.

[18] See `https://snowballstem.org/algorithms/dutch/stemmer.html`

[19] Find the SKOS ar `https://doi.org/10.5281/zenodo.3564586`

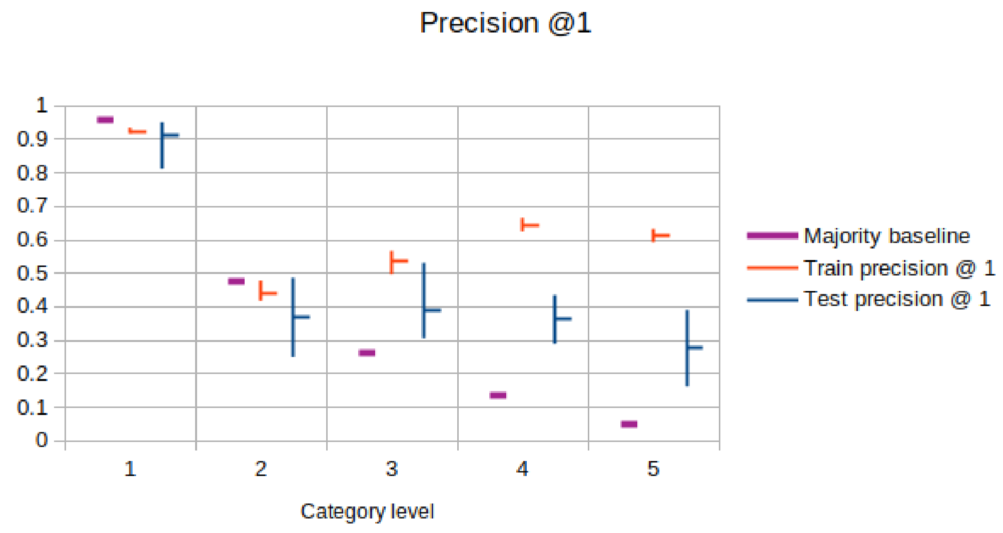


Figure 7: The horizontal axis indicates the hierarchical level of the subject indexing. The figure indicates majority baselines (purple), and precision@1 on the train (red) and test (blue) set. The horizontal bar indicates the average over 10-folds of the data; the vertical bar indicates spread.

Recall and precision with four predicted terms

More indicative of the actual performance of the model are recall and precision measured over all the model predictions. Precision indicates whether the terms suggested by the model are correct according to the manual annotation. Recall measures to what extent manually assigned terms are suggested by the model.

As usual, there is a trade-off between recall and precision: what counts as good performance also depends on the application. If one would use the assigned terms in an information retrieval set-up, high recall means that few relevant documents will be missed. In contrast, high precision will prevent irrelevant documents from showing up. In this stage, we optimised for F1: the harmonic mean of precision and recall. We determined the optimal results on all levels were obtained with a limit of 4 terms to be predicted using a threshold of 0.4, using the hyperparameter optimisation provided by Annif itself.

Figure 8 visualises the recall and precision. We present micro-averaged metrics because those are more robust against imbalanced data. The model is already able to suggest 40% of the relevant detailed terms (level 5) on the test set, which is quite impressive for this task with such a limited amount of data. Again, we see that performance on the train set exceeds that on the test set. This outcome indicates that adding more training data would likely boost performance.

5 Discussion: Facilitating Future Early Modern (Ordinances) Research

The hypothesis posed at the beginning of this article, regarding the cross-border influence of normative texts could not be tested thoroughly due to the limited time this project ran. The results did not go far enough actually to test the hypothesis yet. The hypothesis will be tested in future studies. However, manual verification was done -

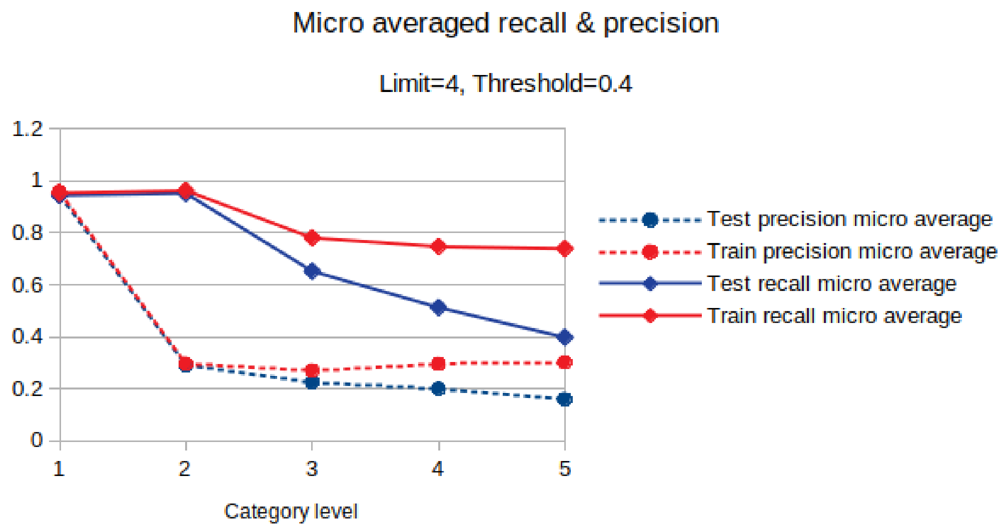


Figure 8: The horizontal axis indicates the hierarchical level of the subject indexing. The figure indicates average precision (dotted) and recall (solid) for the train (red) and test (blue) set.

due to the improved readability - through a full-text search within Transkribus itself. Such an approach obviously neglects the idea that categorisation looks at more context then just the searched key-word. It was already possible to establish a little proof of concept in those few cases that only a full-text search was used. For example, the case of a unification of the axle width to the Holland standard appears in Gelderland and Groningen. The topic of beggars and vagrants appears in multiple federation-states at the same time, though the formulation does differ (*deugniet* ('up to no good') vs. *beggar*). Such a quick search does not provide answers to the fullest extent the hypothesis envisions; hence, further research is needed.

Even if the transcripts are (nearly) perfect, early modern texts tend to contain a lot of spelling variations and dialects, as no 'standard spelling' existed. Furthermore, we know the transcription process is not flawless: contrary to digital-born texts, character errors exist that stem from the digitisation. These may prevent the unification of equivalent words, even after morphological processing. Moreover, remaining character errors may also negatively impact the performance of morphological tools. Within *Entangled Histories* we did not address these issues at this stage, nor did we apply a dedicated stemming algorithm for early modern Dutch. The extent to which this influences further processing, such as categorisation, would be an interesting direction for further research.

That a model is already able to suggest 40% of the relevant categories after being trained with only 90% of 470 texts is promising. Further research will tell us whether it is indeed possible to reach a much higher score by adding more texts and using more sophisticated back-ends. This testing will be done within *A Game of Thrones?!*-project (NWO Veni) at Huygens ING, which will take the *Entangled Histories* knowledge and results as a starting point and continues to work with them.[20] Using the case of Canton Berne (CH)[21] - previously studied in the MPIeR-project - Annif's ability to classify 5500 manually categorised texts will be tested, and the results now obtained

[20] See `https://www.huygens.knaw.nl/projecten/game-of-thrones/`.

[21] See `https://www.rg.mpg.de/2172198/volume007`

with 470 normative texts used in *Entangled Histories* verified. It should then become possible to automatically categorise sources from another Swiss Canton in the future. Furthermore, results from the federation-states Holland and Gelderland are a future basis to categorise Dutch ordinances of other federation-states automatically in the future.

In *A Game of Thrones*, other tools to segment texts will be considered.[22] The previously described tools will be reconsidered if improvements in technique have been made. In the case of Berne, matching with the existing list of ordinances from the MPIeR-project will be possible and will likely be helpful to validate the tools.

Other information that could be retrieved from these ordinances encompasses place names, dates, topics (categorisations), person names. In other words, performing Named Entity Recognition (NER) would be ideal for making the texts more searchable too. Once named entities have been recognised, one could visualise them on maps and timelines. These normative texts could be incorporated into Time Machine Projects[23] - which tend to leave the normative rules on the side.

Digital Historians - e.g. the Data for History Consortium[24] - are looking for ways of making geo-historical data interoperable in the semantic web. They suggest that this could be done through OntoME[25] which is designed for any object-oriented structured data model (based on CIDOC-CRM[26]), to make it easy to build, manage and align an ontology. Such an OntoME/Ontology for ordinances would be applicable on normative texts throughout Europe, providing the MPIeR-categorisation would be followed. Such an ontology would help solve language issues that occur in the current dataset, where French and Latin texts can be found in the Dutch collection (or vice versa). The OntoMe would help to structure the data in a machine-readable way, allowing to circumvent such challenges.

Early modern ordinances have long been left unattended, or at least research has not reached its full capacity. Datafication of these sources will bring more possibilities and will, in the longer run, enable us to see how cross-border influence (the entangled history) worked. One of the next steps is creating a Linked Data system to combine the available data. The map in Figure 9 shows - in light green - the research conducted at the MPIeR[27]; in dark green are other initiatives to inventorise the normative texts.[28] The multi-lingual SKOS will allow searches through different languages and, thus, across borders and through centuries. Hence, when this data does get connected and complemented with additional data, the possibilities to study the administrative norms of early modern (Western) Europe will almost become limitless.

[22] For example the rule-based, semi-automatic tool Layout Analysis tool Larex: https://github.com/OCR4all/LAREX

[23] See https://www.timemachine.eu/

[24] See http://dataforhistory.org/

[25] See http://ontome.dataforhistory.org/

[26] See http://www.cidoc-crm.org/

[27] See https://www.rg.mpg.de/forschungsprojekt/repertorium-der-policeyordnungen?c=2124983

[28] See e.g.: https://justitie.belgium.be/nl/informatie/bibliotheek/koninklijke_commissie_uitgave_belgische_oude_wetten_en_verordeningen; https://www.huygens.knaw.nl/projecten/resoluties-staten-generaal-1576-1796-de-oerbronnen-van-de-parlementaire-democratie/ [12-02-2020]; https://historischcentrumoverijssel.nl/digitalisering-historische-statenresoluties/ [12-02-2020]; https://www.huygens.knaw.nl/projecten/game-of-thrones/.

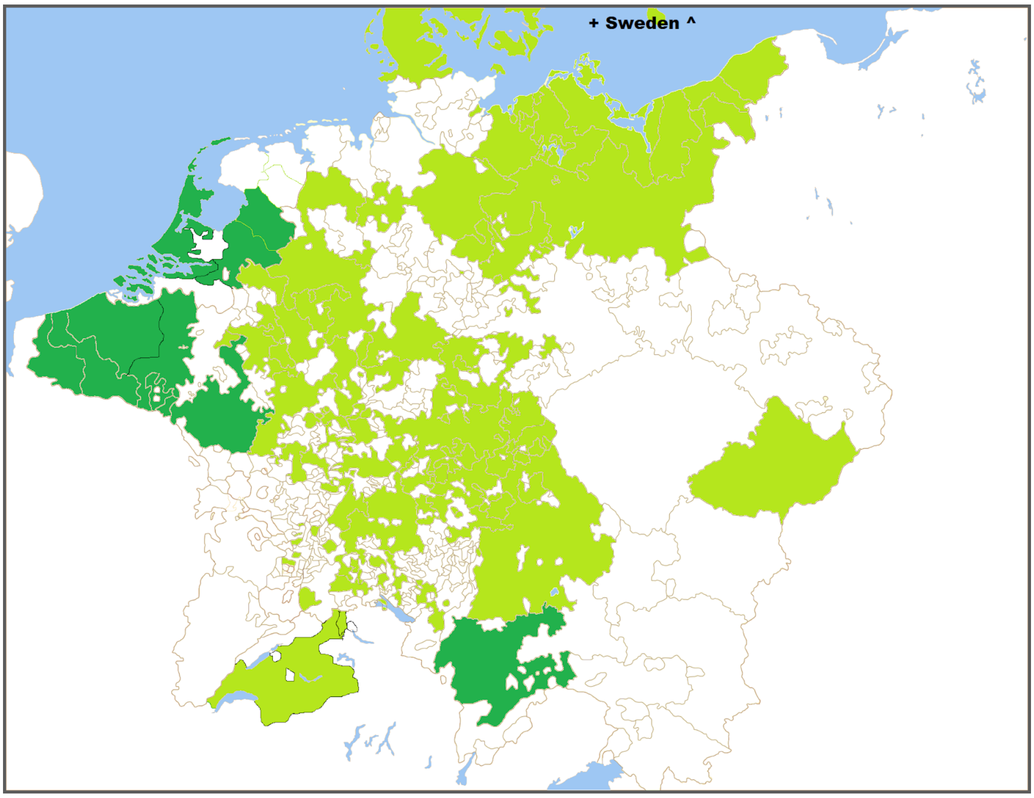


Figure 9: MPIeR Repertorium and other initiatives (dark green), period 1500-1800. Map: Blank map of the Holy Roman Empire in 1648, `https://commons.wikimedia.org/wiki/File:Holy_Roman_Empire_1648_blank.png`

6 Code and Data Availability

- The dataset used within Entangled Histories can be found at `https://lab.kb.nl/dataset/entangled-histories-ordinances-low-countries` and archived in Zenodo: `https://doi.org/10.5281/zenodo.3567844`.
- The code for the rule-based segmentation written by Sara Veldhoen is hosted at GitHub: `https://github.com/KBNLresearch/EntangledHistories`.
- The provisional SKOS used for the categorisation can be found at `https://doi.org/10.5281/zenodo.3564586`.

7 Acknowledgements

We would like to thank a number of people for their help and feedback throughout the project, which formed the basis of this paper: Martijn Kleppe, Lotte Wilms, Steven Claeyssens, Ronald Nijssen, Jeroen Vandommele, Rosemarie van der Veen-Oei, Mirjam Cuper and Ewald Pauwels; Julie M. Birkholz is thanked for proof-reading the article.

Annemieke would like to thank the KB National Library of the Netherlands for the hospitality and help with the project; in particular the team-members Sara Veldhoen and Michel de Gruijter. At Ghent University Annemieke would like to thank: René Vermeir, Dirk Heirbaut, Georges Martyn Thomas Donald Jacobs, Kevin Dekoster, and Klaas van Gelder as well as the members of the Institute for Early Modern History and the GhentCDH Research Lab for their continuous support.

8 Funding acknowledgement statement

This publication has been made possible through the KB National Library of the Netherlands Researcher-in-Residence-project. Each year - since 2014 - two researchers are appointed for six months to do research within the KB. The researcher can benefit from the knowledge and experience of our KB programmers and collection specialists, while the KB gathers valuable information to improve its services to digital humanities researchers.

This publication also benefited from the knowledge gained during the NWO Rubicon Grant 446-16-019 that Annemieke Romein received (2016-3) for her project *Law and Order: Low Countries? The relations between provincial and urban institutions in state-building in the Low Countries 1579-1701*.

References

Clinchant, S., H. Déjean, J. Meunier, E. M. Lang, and F. Kleber
2018. Comparing machine learning approaches for table recognition in historical register books. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, Pp. 133–138.

Der Weduwen, A.
2018. *Selling the republican ideal : state communication in the Dutch Golden Age*. PhD thesis, University of St Andrews. Not published.

Hoftijzer, P. G.
2015. *Europäische Geschichte Online : EGO : The Dutch Republic, Centre of the European Book Trade in the 17th Century*. Mainz: Leibniz-Inst. f. Europ. Geschichte.

Kleppe, M., S. Veldhoen, M. van der Waal-Gentenaar, B. den Oudsten, and D. Haagsma
2019. Exploration possibilities Automated Generation of Metadata. https://doi.org/10.5281/zenodo.3375192.

Koolen, M. and R. Hoekstra
2019. Reusing Existing Structures for Access to Large Historical Corpora. Presentation at DHBenelux 2019, Liège. http://2019.dhbenelux.org/wp-content/uploads/sites/13/2019/08/DH_Benelux_2019_paper_10.pdf.

Kotkas, T.
2014. *Royal police ordinances in early modern Sweden : the emergence of voluntaristic understanding of law*, The Northern world, 1569-1462 ; vol. 64. Leiden [etc.]: Brill.

Muehlberger, G., L. Seaward, M. Terras, S. Ares Oliveira, V. Bosch, M. Bryan, S. Colutto, H. Déjean, M. Diem, S. Fiel, B. Gatos, A. Greinoecker, T. Grüning, G. Hackl, V. Haukkovaara, G. Heyer, L. Hirvonen, T. Hodel, M. Jokinen, P. Kahle, M. Kallio, F. Kaplan, F. Kleber, R. Labahn, E. M. Lang, S. Laube, G. Leifert, G. Louloudis, R. McNicholl, J. Meunier, J. Michael, E. Mühlbauer, N. Philipp, I. Pratikakis, J. Puigcerver Pérez, H. Putz, G. Retsinas, V. Romero, R. Sablatnig, J. Sánchez, P. Schofield, G. Sfikas, C. Sieber, N. Stamatopoulos, T. Strauß, T. Terbul, A. Toselli, B. Ulreich, M. Villegas, E. Vidal, J. Walcher, Weidemann Max, H. Wurster, and K. Zagoris
2019. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5):954–976.

Prasad, A., H. Déjean, and J. Meunier
2019. Versatile layout understanding via conjugate graph. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Pp. 287–294.

Quirós, L.
2018. Multi-task handwritten document layout analysis. *CoRR*, abs/1806.08852.

Romein, C. A.
2019. De computer de wet laten herschrijven...?! Presentation KB Weetfabriek 23 September 2019, see `https://www.youtube.com/watch?v=XZzL5j_sjkw`. Archived on Zenodo: `https://doi.org/10.5281/zenodo.3562881`.

Stolleis, M., K. Härter, L. Schilling, and M.-P.-I. f. E. Rechtsgeschichte
1996. *Policey im Europa der frühen Neuzeit*, Ius commune (Klostermann).: Sonderhefte. V. Klostermann.

Suominen, O.
2019. Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly*, 29(1):1–25.

Tafti, A. P., A. Baghaie, M. Assefi, H. R. Arabnia, Z. Yu, and P. Peissig
2016. OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym. In *Advances in Visual Computing*, G. Bebis, R. Boyle, B. Parvin, D. Koracin, F. Porikli, S. Skaff, A. Entezari, J. Min, D. Iwai, A. Sadagic, C. Scheidegger, and T. Isenberg, eds., Pp. 735–746, Cham. Springer International Publishing.

Tennis, J. T. and S. A. Sutton
2008. Extending the simple knowledge organization system for concept management in vocabulary development applications. *Journal of the American Society for Information Science and Technology*, 59(1):25–37.

Řehůřek, R. and P. Sojka
2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Pp. 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.