10.20378/irb-58260

The feasibility of machine-learning based workflows for editing serial historical sources. First results of the Schenkenschans customs registers project.

Werner Scheltjens, University of Bamberg

Preprint, forthcoming in: Ulrike Henny-Krahmer (ed.), *Machine learning and data mining for digital scholarly editions*, Norderstedt 2023.
Note: This paper is the peer-reviewed and accepted version of a paper presented at a conference on „Machine learning and data mining for digital scholarly editions“ (Rostock, 9-10 June 2022). See : https://www.i-d-e.de/aktivitaeten/veranstaltungen/machine-learning-and-data-mining-for-digital-scholarly-editions/

ABSTRACT

This paper discusses preliminary results of a project that aims to facilitate the study of logistics patterns in German-Dutch transport and trade on the Rhine in the early modern period by means of a digital edition of the customs registers of the Schenkenschans (SSZ). Based on an ongoing pilot study with a sample of the SSZ registers, the paper discusses the use of machine-learning based tools for HTR as starting point for creating a digital scholarly edition of serial historical sources. Building on a conceptualizing of the customs registrations in the SSZ as economic movement data, the paper briefly outlines how the scholarly edition supports data mining and documentation. Based on rigorous timing of the different steps in the proposed workflow, the pilot study assesses the feasibility of ML-based tools for HTR as a first step in the production of a digital scholarly edition of the SSZ for data mining and documentary purposes.

AUTHOR BIO

Werner Scheltjens (1978) obtained his PhD in history at the University of Groningen (2009) and his Habilitation (with venia for social and economic history and East European history) at the University of Leipzig in 2020. Since January 2021, he is professor of digital history at the University of Bamberg, Germany. His research focuses on the application of digital historical methods to conduct research on preindustrial economic history, maritime history, and historical metrology.

Introduction

Except for a few projects with high visibility, such as Slave Voyages, the Danish Sound Toll Registers Online, the combined VOC projects and the Prize Papers, recent discussions about the advancement of digital methods in the historical humanities seem to pay little attention to sources concerning preindustrial trade and transport (see e.g. Decker 2019; Romein et.al. 2020). However, sources such as merchant account books, manuals and notebooks, commodity price lists, ships' logbooks, port books or customs registers, cater to the interests of historians studying the preindustrial economy in a similar way as medieval and early modern municipal and monastic account books do (Vogeler 2015, 307). But in contrast to the latter (e.g. Burghartz 2015; Burghartz, Calvi & Vogeler 2017), the application of state-of-the-art methods of the digital humanities to serial historical sources of preindustrial trade and transport is still in its infancy. Digital scholarly editing, in particular, is an uncommon starting point for working with serial historical sources.

This is not surprising. For starters, with the exception of the flagship projects mentioned above, sources of preindustrial trade and transport are usually more fragmented and smaller. Perhaps they are also harder to relate to than city and monastery account registers. More importantly, however, working with serial sources of preindustrial trade and transport has its origins in the different research tradition of the historical social sciences, which favours structured representations of content over the typographical features of the source and typically reproduces content in an abridged form. Often parts of the source are aggregated or even summarized for economic reasons (e.g. to fit into one printed volume). Elements that do not (seem to) carry semantic meaning are omitted and, as a rule, the contents of the source are reorganized in order to fit into a predefined, often tabular structure (Vogeler 2015, 308-310).

Nevertheless, it should not be overlooked that printed source publications of preindustrial trade and transport form a long scholarly tradition in their own right. Famous examples include the *Rijksgeschiedkundige Publicaties* (RGP) in the Netherlands (e.g. Unger 1939; Niermeyer & Smit 1968-1997; Lindblad 1995); source publications related to the study of the Hanse (e.g. Jenks 2012; *London Customs Accounts*), Hamburg (Schneider, Krawehl & Denzel 2001), overland trade (Straube 2015) and riverine trade (e.g. Scholz-Babisch 1971; Rauscher 2015); and sources published in the context of the Annales School (e.g. Chaunu 1955-1957). Obviously, this vast and international tradition comprises some printed scholarly editions as well, such as the account books of Hildebrand Veckinchusen (Lesnikov 1973), but most of the above-mentioned scholarly works were not published with the goal of applying philological editing principles to historical sources.

The research tradition of the historical social sciences informs the methodology for publishing sources of preindustrial trade and transport in electronic form. Relational data models continue to prevail (cf. Gil 2021). Famous projects that have adopted a relational data model include the *Dutch Asiatic Shipping in the 17th and 18th centuries* (DAS); *Sound Toll Registers Online* (STRO; Gøbel 2010; Scheltjens & Veluwenkamp 2012; Veluwenkamp, Scheltjens & Van der Woude 2021); the *SlaveVoyages* project and the *Trans-Atlantic Slave Trade Database* (Eltis 2012), *Navigocorpus* (Marzagalli, Dedieu, Pourchasse & Scheltjens 2012); *PORTIC - PORTs, and Information and Communication Sciences and Technology. Querying and Visualizing eighteenth-century shipping and trade dynamics in the digital era* (PORTIC); *TOFLIT18 - Transformations of the French Economy through the Lens of International Trade, 1716-1821* (TOFLIT18); *Der Donauhandel. Quellen zur Österreichischen Wirtschaftsgeschichte des 17. und 18. Jahrhunderts* (Donauhandel); and most recently *AveTransRisk* (Piccinno & Iodice 2021), the *London Customs Accounts*, or *Portfolio, the Exchequer Portbooks project* (Dunn 2020).[1] Although these electronic source publications typically do not present themselves as digital scholarly editions, they do share with the latter the critical interaction with textual sources and the goal of creating opportunities for exploring and interpreting the contents of the source in novel ways.

Research questions

Inspired by the breakthrough of digital approaches in the historical humanities, Georg Vogeler and others have repeatedly called for the application of methods of digital scholarly editing to sources of economic history, especially of medieval accounts and registers (e.g. Vogeler 2015,

[1] For links to the websites of the projects (names in italics), see the references.

2016, 2019). However, even today, digital scholarly editing of economic history sources is still not a very common practice compared to the construction of relational databases and spreadsheets (e.g. Isenmann 2015; Kypta et.al. 2019, 467-468). For several reasons, the barriers of entry to digital scholarly editing for researchers with a background in the historical social sciences still seem to be quite high (Vogeler 2015, 318-320). This paper will not embark on a general discussion of what is needed to convince researchers of the benefit of a digital scholarly edition compared to the more familiar, database-oriented methods of the historical social sciences. Rather it focuses on questions related to the applicability and feasibility of machine-learning-based tools for editing serial historical sources, like 'how do I get started?', 'what kind of preparations are necessary?', 'what kind of results can I expect, and what can I do with them?', or 'how long does it take to process one page?'.

Starting from the observation that it is hard to answer such basic questions with regard to the feasibility of applying (partly) automated procedures during the process of editing serial historical sources of preindustrial trade and transport, a pilot study was set up. Its goal was to gain insight into the application of ML-based tools for handwritten text recognition (HTR). The pilot study was constructed around the customs registers of the Schenkenschans (*Ger.* Schenkenschanz Zollregister, further abbreviated as SSZ), a Dutch source of preindustrial transport and trade that captures ships passing the German-Dutch border on the river Rhine between 1630 and 1810. The pilot study anticipates the use of the SSZ for the analysis of preindustrial logistics patterns. Based on a careful monitoring of the time spent on the different processing steps for HTR, the paper aims to gain insight into the following issues: Under which conditions does the partly automated workflow for HTR support a more efficient use of scarce resources? What does such a workflow look like? How does the application of a ML-based workflow for HTR affect the later stages of editing serial historical sources such as the SSZ?

The Schenkenschans customs registers

The research questions formulated above outline the methodological context of a project that started October 2021 at the University of Bamberg (Scheltjens & Rolker 2021). The project aims to facilitate the study of logistics patterns in German-Dutch transport and trade on the River Rhine in the early modern period by means of the '*digitale Erschließung*' of the SSZ customs registers. As Patrick Sahle has pointed out, the untranslatable verb '*erschließen*', or the noun '*Erschließung*', encompasses "(…) any activity that increases the amount of information concerning a specific object and thus enhances its accessibility and usability" (Sahle 2016: 23). One of the ways to 'open up' or 'uncover' the SSZ for historical research is that of the digital scholarly edition. This is the path pursued by the pilot study, which aims to answer if and how a digital edition of the SSZ may be created relying on tools for HTR. The pilot study aims to find out whether a digital edition is at all feasible for the SSZ and what role (semi-)automated procedures may play during the process. Therefore, an important part of the pilot study focuses on time management and the allocation of scarce resources to the different stages of handwritten text recognition.

Preserved at the *Utrechts Archief* in the Netherlands, the handwritten registers cover the years 1630-1810 with gaps (see below). They provide information written in the Dutch language about ships and timber rafts moving upstream and/or downstream on the Rhine and passing by the Schenkenschans customs station, located close to the Dutch-German border. Except for a number of smaller scholarly contributions dealing with timber rafts on the Rhine in the seventeenth and eighteenth centuries (Van Prooije 1990) and some statistics compiled by Verheul (1994) and later used by Van Zanden and van Riel (1994 [2021], 67) and Combrink

(2021, 25), the source has hardly been used in historiography. Even the most basic information about the registers, such as the number of register entries, is unavailable in the literature. Although the lack of orientation with regard to the source and its contents is challenging, the absence of directions from historiography also provides an opportunity. The SSZ offer an excellent playground for testing and experimenting with available digital tools and research methods.

Prior to designing the workflow, a pre-processing step was devoted to exploring the source. Work with the SSZ started with a collection of about 5,000 double-page scans that were kindly put at the disposal of the project by the Dutch historian Leendert van Prooije, who maintains contact with the *Utrechts Archief* that hosts the source. In a first step, the double-page images were separated into recto and verso images to make further processing easier. Empty pages, i.e. pages from the original customs books that - if at all - merely contain a few lines to indicate columns, were omitted. These empty pages occur at the end of almost every annual register. Moreover, in several registers, the backs (verso) of pages were left blank. These blank pages were omitted as well, since they do not contain any data that requires processing. Thanks to this procedure, the total number of pages to be processed could be reduced to 6,329.

Basic information was gathered about the number of pages in each register, the distribution of different hands and layouts across the document collection, and the data items in each register. The SSZ collection contains registers for 126 years within the 1630-1810 period. For four longer periods, no registers are preserved. The first gap in the SSZ collection is from May 1631 to April 1651; the second from January 1663 to April 1672; the third from May 1722 to April 1737 and the fourth from May 1742 to April 1749. Minor gaps occur between January and April 1680; May 1702 and April 1703; May 1712 and April 1713; May 1720 and April 1721 and May 1757 and April 1758.

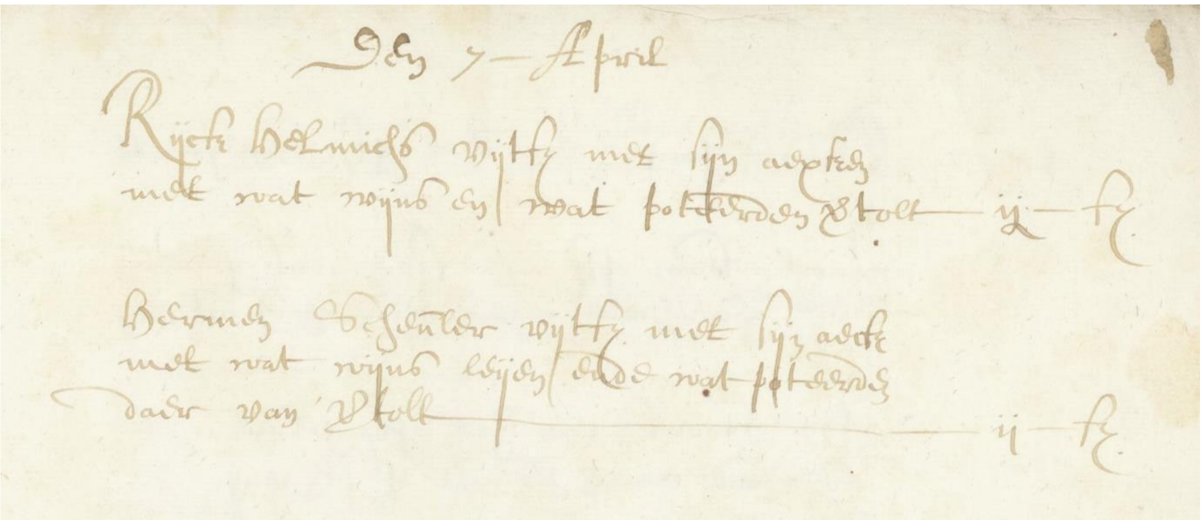
The registers were kept in 19 different handwriting styles that are unevenly distributed in the collection. Some hands occur only once on very few pages, other hands persist for several years in a row. This is particularly true for the periods 1683/84 to 1708/09 (937 scans) and 1763/64 to 1795/96 (2,457 scans) (see table 1 for details). Together, these two hands are responsible for more than half of the entire SSZ collection. In contrast, the handwriting style changed almost every year in the early registers of the second half of the seventeenth century. Some hands are known by name, but often the writer did not sign off the register at the end of the year and therefore cannot be identified directly from the source.

Hand	Scans	Period
1	43	1630-1631
2	69	1651-1653
3	51	1653-1655
4	152	1655-1660
5	38	1655-1656
6	148	1662; 1672-1676
7	15	1676-1677
8	102	1677-1681
9	56	1681-1683
10	937	1683-1709
11	133	1709-1712
12	341	1713-1722
13	453	1737-1738; 1740-1742; 1749-1751; 1754-1756; 1758-1760

14	439	1738-1740; 1751-1754; 1756-1757; 1760-1762
15	51	1762-1763
16	2457	1763-1796
17	456	1796-1803
18	294	1803-1808
19	94	1808-1810

Table 1: Number of scans per hand in the SSZ customs registers, with indication of period and name of the writer (if available).

The contents and layout of the registers usually changed when the writer and thus the handwriting changed, but sometimes a writer could also change his writing style and registration practices. As a result, the collection comprises 29 different layouts written in 19 hands. Among them, the first register (1630/31) is an outlier compared to the rest of the collection (see figure 1).

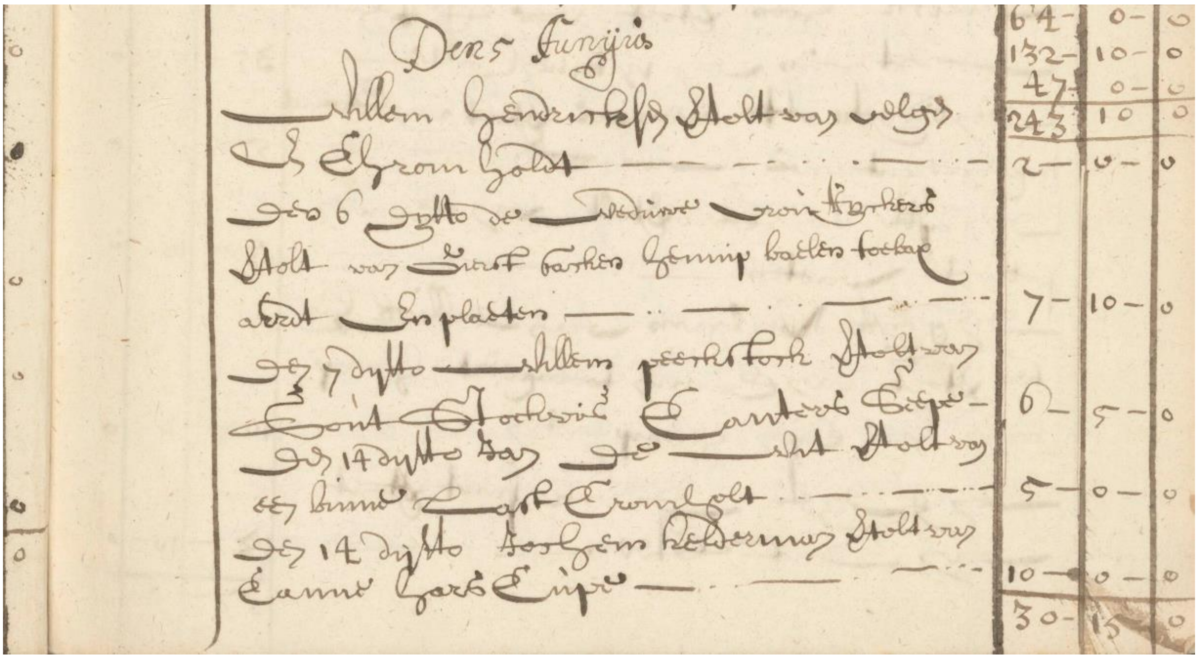


Den 7 April
Rijck Helmichs withe met sijn aecken
met wat wijns en wat potterijen tolt — ij — st

Berendt Schouten withe met sijn aecke
met wat wijns [illegible] ende wat potterijs
daer van tolt — ij — st

Figure 1: Sample from the 1630/31 Schenkenschans customs register.

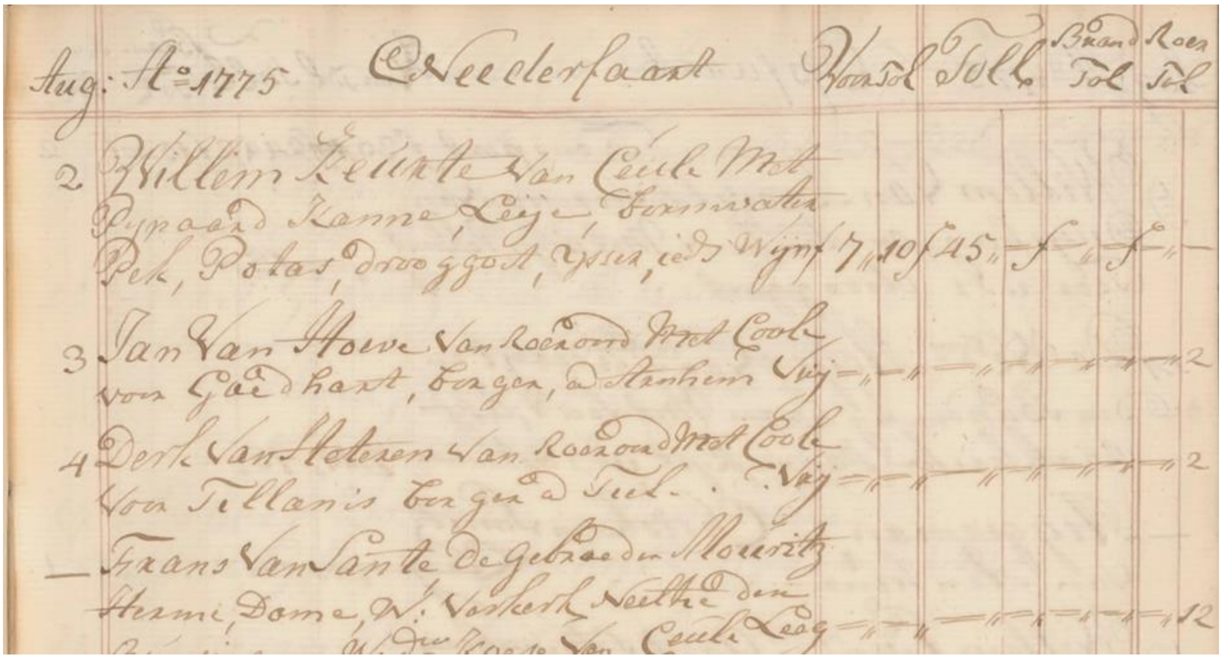
Not only does it predate the next register in the collection by 20 years, it also has an entirely different layout. In contrast with the rest of the collection, the 1630/31 register does not have a tabular structure. The tabular structure of the later registers could either be explicit, when vertical lines were visible for columns, or implicit, when the page layout follows invisible column divisions. Separated by the date of passage, each entry mentions the name of the skipper, the direction of the passage (up- or downstream), the type of ship, cargo items carried on board and duties paid.



	64	0	0
Den 5 Junijus	132	10	0
	47	0	0
Willem Hendrickss Stoltwaag volgs	243	10	0
[illegible] Ebrom holdt	2	0	0
Den 6 dytto de weduwe van Herbert			
Stoltwaag Gerrit Barkes hennip batten tockap			
aerdt en platen	7	10	0
Den 7 dytto Willem Peerkstock Stoltwaag			
[illegible] Cantes Eycke	6	5	0
Den 14 dytto Gaag de [illegible] Stoltwaag			
een binnen Last Cromholt	5	0	0
Den 14 dytto Jorghem Kelderman Stoltwaag			
Caneel hart Eyke	10	0	0
	30	15	0

Figure 2: Sample from the 1656 Schenkenschans customs register.

Starting in the 1650s, the registers gradually obtained a more or less standardized character. For example, the excerpt from the 1656 register (see figure 2) exemplifies the clear structure of the register. The left margin is empty; the main column contains the customs entries headed by a date at the beginning of the month; and three columns on the right contain information about duties paid. The upper and lower right corners of this snippet contain intermediate totals; at the top for May 1656, at the bottom the sum of a few entries in June 1656. In the late seventeenth century, the registrations became even more succinct and, for the most part, were limited to just one line. However, the key data items remained the same: date of passage, name of skipper, cargoes carried and customs paid. During the first half of the eighteenth century, information about the port of origin, destination, domicile, and ship type started to complement the key data items. If at all, before that time, this information was only occasionally included in the registers. It is only in the 1760s, however, that the additional data items were captured systematically in the registers and that more detailed registration procedures replaced previous practices. The excerpt from a 1776 register (see figure 3) shows that, by that time, the use of the header to capture information about dates, direction and the different taxes due had become the rule.



Aug: Aº 1775 Neederwaart Konsol Toll Brand Tol Roer Tol

2 Willem Keunte van Ceulen Met
Pyrmont Kanne, Leije, Formwater
Pek, Potas, droog goet, ysser, en Wyn f 7 10 f 45 f f

3 Jan van Hoeve van Roerord Met Cool
voor Goedhart, bergen, en Arnhem Vrij 2

4 Derk van Heteren van Roerord Met Cool
voor Tillanis bergen en Tiel. Vrij 2

— Frans van Sante de Gebroeders Mouritz
Hermi, Dome, W. Verkerk, Neethe den
Ceulen Leeg 12

Figure 3: Sample from the 1775/76 Schenkenschans customs register.

After 1750, the registers also contain more details about the cargoes carried by the ships (e.g. their quantities and – sometimes – recipients) and the timber rafts passing through customs (e.g. their dimensions) than in earlier times.

Customs entries as economic movement data

The SSZ are well suited for the study of preindustrial logistics patterns. Alongside the cliometric study of transportation, which focuses on such topics as the speed of transport and the choice of transport routes (Bogart 2019), and the history of commerce and trade, which deals with merchant networks, trade agreements, commercial institutions and commercial practices (e.g. Häberlein & Jeggle 2010) as well as with the quantitative analysis of trade flows (e.g. Daudin & Charles 2015; Charles et al. 2022), the study of logistics patterns offers a third way of using sources of preindustrial trade and transport. Whereas the former branches of economic history predominantly rely on either macro level statistics or micro level case studies to achieve their goals, logistics patterns analysis requires a different approach. By looking for regularities, repetition, flexibility and outliers in the movement of ships passing through customs stations, arriving at or departing from ports during long periods of several years, decades and even centuries, logistics pattern analysis aims to identify the mechanisms and strategies at work in the transport services sector. Relying on the systematic processing of large amounts of more or less homogenous data about the movement of ships and their masters as well as additional information about the ship itself, the domicile of the master, preferred sailing routes and commodities carried, logistics pattern analysis distinguishes between occasional and specialist transport services providers and identifies different types of specialization. For example, it is possible to identify groups of masters sharing similar behaviour as well as clusters of domiciles with a certain level of specialization on one or another commodity and/or route. Moreover, in the course of the process, traces of the lives of many ‘anonymous’ people can be reconstructed, which adds a social history dimension to the analysis (cf. Gleba & Petersen 2015, 9).

In order to use the SSZ as a source for analysing logistics patterns in German-Dutch preindustrial transport and trade on the Rhine its entries should be understood as economic movement data (Scheltjens & Dopfer 2012). Each entry in the SSZ captures the movement of a ship, with its master and cargo items, at a certain time and in a certain direction. This movement can be followed up by other ship movements in the same and opposite directions, during the same year, but also in (non-)consecutive years. From the perspective of a single master, a movement in the register for one year can be related to other movements during the same year and/or other movements in different years. From the perspective of many masters sharing certain characteristics (e.g. the same domicile; carrying the same good), movements in one year can be related to other movements during the same or different years. Groups of masters sharing a number of characteristics for a particular period formed clusters that changed over time. Gaining insight into the emergence, composition, rise and decline of clusters identified in the source is a major task in the analysis of logistics patterns, which is inspired by recent advances in the study of movement patterns (Dodge 2021; Dodge, Weibel & Lautenschütz 2008).

Prior to mining the entries in the SSZ as economic movement data, all key data items in the registers (date, name of master, direction, commodities, port of origin, destination, domicile, ship type) should be annotated. Another requirement is that, in addition to the key data items, annotation distinguishes between ships going in different directions (upstream / downstream), on the one hand, and ships and timber rafts passing through customs, on the other hand. One last requirement for mining the entries in the SSZ as economic movement data is that linking and ranking economic movements within one register and across several registers should be possible. Therefore, at the very least, orthographic differences in the names of masters have to be addressed, and probably the names data in the registers will have to be linked to external sources in order to facilitate the identification process of individual masters. When these conditions are met and the semantic tagging of all entries is finished, the process of data mining the SSZ in search for logistics patterns can start.

A ‘traditional’ way of proceeding with the SSZ would be to develop a relational data model that facilitates capturing the key data items and relations of each entry in the registers in a manual process. Whereas the size of the source may not be entirely problematic in that regard, the variations in the structure and contents of the registers do make this a cumbersome and unrewarding task. Not only would adherence to ‘traditional’ data modelling methods limit the possibilities for mining the source as a collection of economic movement data, it would also restrict the potential scholarly impact of the project. At best, another tailor-made solution to working with serial historical sources would be produced.

Therefore, rather than sticking to the predominant database-oriented methodology for source publications in the historical social sciences, we suggest comprehensive handwritten text recognition of the entire SSZ using the ML-based tool *Transkribus* (Muehlberger et.al. 2019; Hodel et.al. 2021) as an alternative way of processing the registers. This approach will allow producing a digital edition of the SSZ in accordance with the *de facto* standards of the Text Encoding Initiative (www.tei-c.org), while also facilitating the creation of structured and semantically enriched datasets for logistics pattern analysis. As a result, rather than being a research outcome in its own right, the digital scholarly edition will serve as comprehensive documentation and point of reference, control and perhaps replication of the logistics pattern analysis that is to be conducted based on the extracted source data.

An ML-based workflow for processing the SSZ registers

The remainder of the paper describes the main phases of the ML-based HTR workflow. Due to the ongoing status of the pilot study, the results presented below are preliminary and limited to the first 13 registers in the source, which cover the years 1630-1631 and 1651 to 1662. Nevertheless, it is hoped that some first insights regarding the potential use of a digital edition of the SSZ for data mining may be derived from their description.

Layout analysis and segmentation

A key task of the pilot study is to examine and test the implementation of ML-based solutions for handwritten text recognition such as *Transkribus* (Muehlberger et.al. 2019; Hodel et.al. 2021). During the first phase, the focus lied on pre-processing the SSZ customs registers, i.e. on the analysis of the registers' layout and their subsequent segmentation. To tackle this task, 13 registers for the 1630-1662 period were uploaded to the program. This subset contains 369 pages with seven different layout and handwriting styles, meaning that both changed almost every year. After having experimented with automatic identification of text regions and baselines as well as with training layout analysis using *P2PaLA*, both of these preliminary processing steps were not pursued further in the pilot study. The results produced automatically were unsatisfactory. Neither text regions nor baselines were recognized adequately; automatic segmentation was largely unsuccessful due to the limited availability of training data. The possibility to draw tables manually was declined as well. Although the registers do have a more or less regular structure and pages are often divided into a few columns, the Table-feature in *Transkribus* appeared to be too rigid and slow to process the page structure efficiently.

The solution so far has been to identify text regions manually, making a distinction between the opening and closer parts of the register, on the one hand, and the core of the register that contains the customs entries, on the other hand. The page opener contains information about the starting date of a particular page, sometimes the direction of the registered ships and the different taxes to be paid. The page closer, if present, contains intermediate sums as well as monthly totals. In contrast with classical text editorial uses of the structural elements header and footer, they can occur more than once on the same page. This is the case, for example, when one page contains entries for the end of one month and entries at the start of the next month (see also figure 2). Often, after a month's ending, intermediate sums were given in the registers and the new month was announced with another header. To avoid confusion, the structural tags <register_opener>, <register_closer> and <register_entries> were created to tag these elements on each page alongside the identification of text regions. The result of this process is a separation of <register_opener> and <register_closer> information from the customs entries.

The main part of the page - the customs entries - was not processed further at this stage. No attempt was made to identify each entry in a separate text region. This would be very time-consuming and prone to errors. Rather the identification of baselines and therefore of customs entries has been guided by the fact that the customs entries are essentially text strings stretching over one or a few lines. These strings have a more or less regular format and contain several recurring elements, i.e. the key data items discussed earlier. As a rule, they start with a date (or part of it) and end with the amount of customs paid. Therefore, it is envisaged to rely on a script for splitting the main part of each page into single customs entries. In order to facilitate splitting the entries as text strings, the baselines that identify the text to be read within the text regions were set to capture the entire width of all text (=words, numbers and dashes) on one line, across column borders. Baselines have been identified automatically within the text regions. In many cases, this worked fine, but manual corrections, such as merging baselines that constitute one line of text, have been necessary to make sure that all text had been captured adequately. This step in the pre-processing phase has also been used to check the reading order of each line of text and each text region.

Layout analysis and segmentation process were carried out for all 13 registers in the selection. First, text regions for opener, closer and entries were added manually. Then, the *CITlab Advanced Layout Analysis Method* was used to add baselines to the text regions. In a third step, baselines were corrected and made to capture the entire line width when necessary. Finally, the reading order of text regions and baselines was checked on each page. The average time spent on conducting layout analysis and segmentation varied from 1.4 minutes per page for the 1652/53 register to 5 minutes per page for the 1630/31 register. The latter register was considerably more time-intensive to process due to its different page composition and layout compared to the other registers (see figure 1).

Handwritten Text Recognition

Following layout analysis and segmentation of the subset of the SSZ, handwritten text recognition (HTR) constituted the next major phase in the ML-based workflow. HTR was pursued in a pragmatic way. First, several trained base model that are provided by *Transkribus* were tested and compared. If initial HTR results with the existing base model were acceptable, we used it to train a tailor-made model. Different hands were combined in one model for as long as the same base model delivered acceptable initial results. If the initial results were poor and no base model could be found, it was decided that a new tailor-made HTR model should be trained from scratch.

Based on this process, *Republic_7* (HTR ID: 23967) turned out to be the preferred base model for most registers in the pilot selection of the SSZ. Although the initial Character Error Rate (CER) and the Word Error Rate (WER) were high (see table 2), HTR with this model could still be used as the starting point for training a new model for 10 out of 13 registers. In three cases, no acceptable base model could be found. In order to train an SSZ model based on the selected base model, 10 pages of ground truth were produced for each layout/handwriting style. After exclusion of three registers from HTR model training building on an existing recognition model, 10 SSZ registers with five different layout and handwriting styles remained. Hence, 50 pages were manually corrected for ground truth production. After having trained the model, the error rates of the base model were compared with those of the trained model for a number of registers (see table 2). The comparison shows that the quality of the HTR improved significantly when the trained model was used. In all cases, the use of the trained model led to a reduction of the CER for selected registers to average values below three percent (see table 2). Another round of correction of the transcriptions processed with the trained SSZ model finalized the second major phase in the ML-based workflow.

Register	Base Model (Republic_7)		Own model (SSZ_v1)	
	CER	WER	CER	WER
1630-1631	10,56	35,42	2,65	10,78
1653-1654	26,64	65,34	2,95	9,55
1662	21,18	47,63	1,86	5,46

Table 2: CER and WER using base Model Republic_7. Results for selected SSZ registers

Timing of ground truth production and manual correction of the transcriptions generated using the trained model revealed that ground truth production took on average between 9.5 and 17.5 minutes per page depending on the hand itself, the density of the handwriting, and thus the number of entries to be reviewed. Manual correction benefited from a significantly reduced number of errors, which results in average processing times per page between 4.5 and 7.5 minutes. In general, the time invested for producing ground truth, even based on an existing recognition model, is not much lower than the time needed for manually transcribing these pages. Significant gains can be expected only when the trained model is used for processing the remaining pages of a particular hand or set of hands.

Feasibility assessment

The discussion of the two main phases of the ML-based workflow for HTR indicates that the proposed workflow comprising (1) partly automated layout analysis and segmentation, (2) preliminary HTR relying on an existing base model, and (3) the subsequent optimization of HTR with a limited amount of training data can yield significant improvements in the quality of the automatic transcription. The CER went from a range of 10.56% to 26.64% to less than 3% for the registers transcribed so far. As was shown in the previous section, a transcription of SSZ registers that can be used in a digital scholarly edition is just one round of manual corrections away. Table 3 summarizes the results of monitoring the average time spent on processing one page of the SSZ (see table 3).

			Time spent (average in minutes per page)			
Register	Pages	LA	LA & Seg	GT	Manual correction	Entire register
1630-1631	41	I	5	9.5	7.5	14.5
1651-1652	40	II	2	No fitting base model found		
1652-1653	28	II	1.4	No fitting base model found		
1653-1654	25	III	3.2	16	7.2	13.9
1654-1655	26	III	2.1	n/a	Ongoing	
1655	23	IV	3.3	n/a	Ongoing	
1655-1656	38	VII	2.9	No fitting base model found		
1656	21	IV	4.9	14	8.7	15.7
1657	19	IV	3.7	n/a	5.5	9.2
1658	18	IV	3.4	n/a	6.0	9.4
1659	35	V	3.8	n/a	6.5	10.3
1660	31	V	4.3	11.5	6.0	11.8
1662	24	VI	3.6	17.5	4.3	13.4

*Table 3: Monitoring results for layout analysis and segmenation (LA & Seg), ground truth production (GT) and post-correction of pages transcribed with own model, averages in minutes per page. Rounding may result in minor differences. "n/a" means that no ground truth production was necessary for this register. The formula for calculation the average time per page for the entire register is ((Pages * LA&Seg) + (10 * GT) + ((Pages – 10) * manual correction))/Pages.*

The monitoring results indicate that – on average – processing one page of the SSZ registers from scratch takes roughly between 10 and 15 minutes. The result of such ML-based workflow is a segmented page and a corrected transcription of that page. The parameters specified in the workflow and monitoring results presented above can be used to estimate the total time necessary for processing the remainder of the SSZ collection using the same ML-based workflow (see table 4). In table 4, the estimates differentiate between layout analysis and segmentation, ground truth production and manual correction of transcriptions with a trained model. The 13 registers (369 pages) of the pilot study for the basis for these estimates, but are not included in them. To arrive at an optimistic estimate, we assumed that 10 pages of ground truth will be produced for each remaining hand in the SSZ collection, i.e. 120 pages of ground truth; for the pessimistic estimate, we assumed 10 pages of ground truth for each remaining layout in the collection, or 220 pages of ground truth.

	Optimistic	Pessimistic
Basic figures about the collection		
Total size of collection	5,960	
Size of ground truth	120	220
Size of post-correction	5,840	5,740
Timing parameters (average per page)		
LA & segmentation	1.4	5.0
Ground truth production	9.5	17.5
Post-Correction	4.3	8.7
Estimates (in minutes)		
LA & segmentation	8,940	29,800
Ground truth production	1,140	3,850
Post-Correction	25,112	49,938
TOTAL (rounded, in hours)	586	1,393

Table 4: Optimistic and pessimistic estimates for processing the entire SSZ collection.

The estimates are still rather rough. They do not take into account the possible efficiency gains of training a segmentation model, nor do they consider the option that less pages of ground truth might also be sufficient. Moreover, the estimates take the minimal and maximum monitoring results for each step in the ML-based workflow for HTR, and do not consider other combinations of these values. On top of that, first tests already indicate that starting from existing HTR models is likely to become even easier for later registers (especially for the second half of the eighteenth century, when the registers were kept by a professional clerk with a clear handwriting, as can be seen in figure 3). Starting from existing HTR models for initial transcription and ground truth production has been and will continue to be beneficial for the entire workflow of the project. The limitations of the estimates notwithstanding, the pilot study clearly indicates that, assisted by a ML-based workflow, the production of a corrected transcription with minimal segmentation of the entire SSZ collection is feasible. It would require roughly between 15 and 35 weeks of full time work devoted to the project. When starting from scratch, the time spent on developing the pilot study as well as the time needed to process the few registers for which no suitable base model was found, should be added to this total.

This pilot study results indicate that, with reasonable effort, the full transcriptions of the SSZ can be made available for the study of preindustrial logistics patterns. Since *Transkribus* allows for semantic annotation within the programme as well as export of the transcription results in a variety of formats (e.g. TEI-XML, plain text, PAGE-XML), different strategies for analysing the SSZ as a collection of economic movement data are possible. The choice of file format and annotation strategy lies outside of the scope of this paper. What is clear, however, is that the data produced with the ML-based workflow for HTR fulfil the basic requirements for further analysis as economic movement data. Key data items as well as register entries as a whole can be annotated relatively easily, probably in a semi-automatic way, using available tools for and relying on recent experiences with NER (e.g. Ehrmann et al. 2021). This kind of annotation will facilitate mining the registers in search for regular, irregular, repetitive or flexible economic movements at the micro level of the individual shipmaster, for linkages and ranking orders within one and across multiple registers, as well as for the identification of clusters of movements and their characteristics.

Closing remarks

In this paper, the use of HTR as starting point for a digital scholarly edition of the SSZ was tested by means of a pilot study. The preliminary results of the pilot study have clearly shown that HTR can be applied successfully to serial historical sources of preindustrial trade and transport. Work on the pilot study has shown that the proposed workflow can be divided usefully into 'simple' preparatory tasks related to the layout of the source, on the one hand, and more profound 'ground truth' production based on existing HTR models and manual correction of pages transcribed with a trained HTR model, on the other hand. Detailed monitoring of the time spent on each of the steps in the ML-based workflow for HTR contributed greatly to assessing the feasibility of the proposed workflow and made it possible to estimate the total number of working hours needed to prepare the groundwork for a digital scholarly edition of the entire SSZ collection.

References

Utrechts Archief, Bestand Kapitel St. Marien, Inv. Nr. 1666.

Transkribus: www.readcoop.eu
P2PaLA: https://github.com/lquirosd/P2PaLA and
https://readcoop.eu/de/transkribus/docu/p2pala/
CITlab Advanced Layout Analysis Method: https://readcoop.eu/transkribus/docu/layout-analysis-help/

1. *AveTransRisk. Average-Transaction Costs and Risk Management during the First Globalization (Sixteenth-Eighteenth Centuries)*. https://humanities-research.exeter.ac.uk/avetransrisk/
2. Bogart, Dan. 2019. "Clio on Speed. A Survey of Economic History Research on Transport." In *Handbook of Cliometrics. Second edition*, edited by Claude Diebolt and Mark Haupert, 1453-1478. Cham: Springer Nature Switzerland.
3. Burghartz, Susanna, Ed. 2015. *Jahrrechnungen der Stadt Basel 1535–1611 – digital*. Basel / Graz. http://gams.uni-graz.at/context:srbas
4. Burghartz, Susanna, Sonia Calvi, and Georg Vogeler, Eds. 2017. *Urfehdebücher der Stadt Basel – digitale Edition*. Basel / Graz. http://gams.uni-graz.at/context:ufbas

5. Chaunu, Pierre. 1955-1957. *Séville et l'Atlantique, 1504-1650. Structures et conjoncture de l'Atlantique espagnol et hispano-américain (1504-1650).* 8 vols. Paris: Éditions de l'IHEAL. http://books.openedition.org/iheal/5670
6. Combrink, Tamira. 2021. "Slave-based coffee in the eighteenth century and the role of the Dutch in global commodity chains." *Slavery & Abolition* 42, Issue 1: 15-42. https://doi.org/10.1080/0144039X.2020.1860465
7. Daudin, Guillaume and Loïc Charles, Eds. 2015. *Eighteenth-Century International Trade Statistics. Sources and Methods*. (Special Issue of the Revue de l'OFCE 140). Paris: OFCE. https://www.ofce.sciences-po.fr/pdf/revue/140/revue-140.pdf
8. Charles, Loïc, Guillaume Daudin, Paul Girard and Guillaume Plique (2022) Exploring the transformation of French trade in the long eighteenth century (1713–1823): The TOFLIT18 project, Historical Methods: A Journal of Quantitative and Interdisciplinary History, DOI: 10.1080/01615440.2022.2032522
9. *Der Donauhandel. Quellen zur Österreichischen Wirtschaftsgeschichte des 17. und 18. Jahrhunderts*. https://www.univie.ac.at/donauhandel/
10. Dodge, Somayeh. 2021. "A data science framework for movement." Geographical Analysis 53, Issue 1: 1-20. https://doi.org/10.1111/gean.12212
11. Dodge, Somayeh, Robert Weibel and Anna-Katharina Lautenschütz. 2008. "Towards a taxonomy of movement patterns." *Information Visualization* 7, no. 3-4: 240-252. https://doi.org/10.1057/PALGRAVE.IVS.9500182
12. Dunn, Oliver. 2020. "A sea of troubles? Journey times and coastal shipping routes in seventeenth-century England and Wales." *The Journal of Transport History* 41, no. 2 (August): 184-207. https://doi.org/10.1177/0022526619886061
13. *DAS - Dutch Asiatic Shipping in the 17th and 18th centuries.* http://resources.huygens.knaw.nl/das
14. Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named Entity Recognition and Classification on Historical Documents: A Survey. *ArXiv:2109.11406 [Cs]*. http://arxiv.org/abs/2109.11406
15. Eltis, David. 2012. "Some implications from the Transatlantic Slave Trade for Maritime databases." *International Journal of Maritime History* 24, no. 1: 257-264. https://doi.org/10.1177/084387141202400113
16. Gil, Tiago Luís. 2021. *How to make a database in historical studies*. Cham: Springer Nature Switzerland.
17. Gleba, Gudrun and Niels Petersen. 2015. "Einleitung." In *Wirtschafts- und Rechnungsbücher des Mittelalters und der Frühen Neuzeit. Formen und Methoden der Rechnungslegung – Städte, Klöster und Kaufleute*, edited by Gudrun Gleba and Niels Petersen, 7-12. Göttingen: Universitätsverlag. https://doi.org/10.17875/gup2015-825
18. Gøbel, Erik. 2010. "The Sound Toll Registers Online Project, 1497–1857." *International Journal of Maritime History* 22, no. 2: 305–324. https://doi.org/10.1177/084387141002200213
19. Häberlein, Mark and Christoph Jeggle, Eds. 2010. *Praktiken des Handels. Geschäfte und soziale Beziehungen europäischer Kaufleute in Mittelalter und früher Neuzeit*. Konstanz: UVK Verl.-Ges.
20. Hodel, Tobias, David Schoch, Christa Schneider und Jake Purcell. 2021. "General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example." *Journal of Open Humanities Data* 7: 13–23. https://doi.org/10.5334/johd.46
21. Isenmann, Mechtild. 2015. "Das ‚Handlungs- und Bilanzbuch' Paulus I. Behaims (1519-1568). Finanzgeschäfte und Klientel eines Nürnberger Financiers. Ein Werkstattbericht." *Annales Mercaturae. Jahrbuch für internationale Handelsgeschichte / Yearbook for the History of International Trade* 1: 37-60.
22. Jenks, Stuart. 2012. *Das Danziger Pfundzollbuch von 1409 und 1411*. Köln / Weimar / Wien: Böhlau Verlag.

23. Kypta, Ulla, Julia Bruch, and Tanja Skambraks, Eds. 2019. *Methods in Premodern Economic History. Case studies from the Holy Roman Empire, c.1300 - c.1600*. Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-030-14660-3
24. Lesnikov, M.P. *Die Handelsbücher des Hansischen Kaufmannes Veckinchusen*. Berlin, Akademie Verlag, 1973.
25. Lindblad, J. Thomas. 1995. *Dutch entries in the pound-toll registers of Elbing, 1585-1700*. The Hague: Instituut voor Nederlandse Geschiedenis.
26. *London Customs Accounts*. https://www.hansischergeschichtsverein.de/london-customs-accounts
27. Marzagalli, Silvia, Jean-Pierre Dedieu, Pierrick Pourchasse and Werner Scheltjens. 2012. “Navigocorpus: a brief overview of the potential of a database.” *International Journal of Maritime History* 24, no. 1: 331-359.
28. Muehlberger, Guenter et.al. 2019, "Transforming scholarship in the archives through handwritten text recognition. Transkribus as a case study." *Journal of Documentation* 75, 5: 954-976. https://doi.org/10.1108/JD-07-2018-0114
29. Niermeyer, J.F. and J.G. Smit. 1968-1997. *Bronnen voor de economische geschiedenis van het Beneden-Maasgebied 1104-1534*. 2 vols. ‘s-Gravenhage: Martinus Nijhoff. http://resources.huygens.knaw.nl/economiebenedenmaas
30. Piccinno, Luisa and Antonio Iodice. 2021. “Managing shipping risk. General average and marine insurance in early modern Genoa.” In Maritime Risk Management. Essays on the history of marine insurance, general average and sea loan, edited by Phillip Hellwege and Guido Rossi, 83-109. Berlin: Duncker & Humblot.
31. *Portfolio - Exchequer Portbook Project. Digitisation and Transcription of the English and Welsh Exchequer Portbooks held at the UK National Archives, 1565-1798 [TNA E 190 & 122]*. https://portfolio.winchester.ac.uk/
32. *PORTIC - PORTs, and Information and Communication Sciences and Technology. Querying and Visualizing eighteenth-century shipping and trade dynamics in the digital era*. https://anr.portic.fr/en/home/
33. Prize Papers Portal. https://portal.prizepapers.de/index/
34. Rauscher, Peter. 2015. “Die Aschacher Mautprotokolle als Quelle des Donauhandels (17./18. Jahrhundert). ” In *Wiegen - Zählen - Registrieren. Handelsgeschichtliche Massenquellen und die Erforschung mitteleuropäischer Märkte (13. - 18. Jahrhundert)*, edited by Peter Rauscher and Andrea Serles, 255-306. Innsbruck - Wien - Bozen: Studienverlag.
35. Sahle, Patrick. 2016. “What Is a Scholarly Digital Edition?” In *Digital Scholarly Editing: Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo, 19–40. Cambridge, UK: Open Book Publishers.
36. Scheltjens, Werner and Christof Rolker. 2021. *Digitale Erschließung einer seriellen Quelle für die niederländisch-deutsche Rheinschifffahrt in der Frühen Neuzeit: Pilotstudie über die Zollregister von der Schenkenschanz (1630-1810), von der automatischen Handschrifterkennung bis zur Online Datenbank*. https://fis.uni-bamberg.de/cris/project/pj00640
37. Scheltjens, Werner and Kurt Dopfer. 2012. “Unified rule approach and the semantic enrichment of Economic movement data.” In *Towards a Universal Ontology of Geographic Space*, edited by T. Podobnikar, M. Ceh, 229-247. IGI Global.
38. Scheltjens, Werner and Jan Willem Veluwenkamp. 2012. “Sound Toll Registers Online: introduction and first research examples.” *International Journal of Maritime History* 24, no. 1: 301-330. https://doi.org/10.1177/084387141202400115
39. Schneider, Jürgen, Otto-Ernst Krawehl and Markus A. Denzel, Eds. 2001. *Statistik des Hamburger seewärtigen Einfuhrhandels im 18. Jahrhundert. Nach den Admiralitäts- und Convoygeld-Einnahmebüchern*. St. Katharinen.
40. Scholz-Babisch, Marie. 1971. *Quellen zur Geschichte des klevischen Rheinzollwesens vom 11. bis 18. Jahrhundert*. 2 vols. Wiesbaden.
41. *SlaveVoyages - Explore the Origins and Forced Relocations of Enslaved Africans Across the Atlantic World*. https://www.slavevoyages.org/

42. *STRO - Sound Toll Registers Online*. www.soundtoll.nl
43. Stapel, Rombert. 2018. "Historical Atlas of the Low Countries (1350-1800)", https://hdl.handle.net/10622/PGFYTM, IISH Data Collection, V7
44. Straube, Manfred. 2015. *Geleitswesen und Warenverkehr im thüringisch-sächsischen Raum zu Beginn der Frühen Neuzeit*. Köln / Weimar / Wien: Böhlau Verlag.
45. *TOFLIT18 - Transformations of the French Economy through the Lens of International Trade, 1716-1821*. http://toflit18.medialab.sciences-po.fr/#/home
46. Unger, W.S. 1939. *De tol van Iersekeroord. Documenten en rekeningen 1321-1572*. 's-Gravenhage: Martinus Nijhoff. http://resources.huygens.knaw.nl/retroboeken/iersekeroord/
47. Van Prooije, Leendert A. 1990. "De invoer van Rijns hout per vlot 1650–1795." *Economisch- en sociaal-historisch jaarboek* 53: 30-79.
48. Van Zanden, Jan Luiten and Arthur Van Riel. 2021 [first edition 1994]. *The Strictures of Inheritance. The Dutch economy in the nineteenth century*. Princeton: Princeton University Press.
49. Veluwenkamp, Jan Willem, Werner Scheltjens and Siem Van der Woude. 2021. "Sound Toll Registers Online." *TSEG-The Low Countries Journal of Social and Economic History 18*, 1: 147-160. https://doi.org/10.18352/tseg.1203
50. Verheul, M.M. *Anderhalve eeuw Rijnvaart: een kwantitatieve studie naar de Rijnhandel met het Duitse achterland op basis van de tolregisters van schenkenschans 1650–1800* (unpublished thesis, University of Utrecht, 1994).
51. Vogeler, Georg. 2015. "Digitale Edition von Wirtschafts- und Rechnungsbüchern." In *Wirtschafts- und Rechnungsbücher des Mittelalters und der Frühen Neuzeit*, edited by Gudrun Gleba and Niels Petersen, 307–328. Göttingen: Universitätsverlag. https://doi.org/10.17875/gup2015-825.
52. Vogeler, Georg. 2016. "The Content of Accounts and Registers in their Digital Edition. XML/TEI, Spreadsheets, and Semantic Web Technologies." In *Konzeptionelle Überlegungen zur Edition von Rechnungen und Amtsbüchern des späten Mittelalters*, edited by Jürgen Sarnowsky, 13-41. Göttingen: University Press Göttingen.
53. Vogeler, Georg. 2019. "The ‚assertive edition‘. On the consequences of digital methods in scholarly editing for historians." *International Journal of Digital Humanities* 1: 309-322. https://doi.org/10.1007/s42803-019-00025-5