Reading in the mist

Sangiacomo, Andrea; Hogenbirk, Hugo Dirk; Tanasescu, Raluca; Karaisl, Antonia; White, Nick

Published in:
Digital Scholarship in the Humanities

DOI:
10.1093/llc/fqac014

Publication date:
2022

Reading in the mist: high-quality optical character recognition based on freely available early modern digitized books

Andrea Sangiacomo , Hugo Hogenbirk, and Raluca Tanasescu, Antonia Karaisl and Nick White
Faculty of Philosophy, University of Groningen, Groningen, The Netherlands

Abstract

In this paper, we present a workflow for reworking digitized versions of early modern books, freely available in the public domain, in such a way that they will be capable of yielding high-quality optical character recognition (OCR) results suitable for computational text mining. Testing our method, we observed that anything above 90% OCR accuracy is sufficient for semantic analysis. In addition,the overall homogeneity in the OCR accuracy across the corpus proved to be more important than having perhaps only a few works with higher accuracy and the rest available in a lower quality. In terms of the OCR process, this paper illustrates how it was possible to reduce the processing time at maximum quality of a single book of average length (ca. 500 pages) from a minimum of 20 hrs to an average of about 3 hrs (though theoretically nearly infinitely reducible). This was achieved by replacing a step-by-step OCR process with a fully automated pipeline system run on an arbitrary number of servers, breaking up the full process of OCRing one book into minimal tasks that can be handled simultaneously by multiple servers.

Correspondence:
Andrea Sangiacomo, Faculty of Philosophy, University of Groningen, Oude Boteringestraat 52, Groningen 9717 GL, The Netherlands.
E-mail:
a.sangiacomo@rug.nl

Oh I see! Try this lens!
Just an open space—I see nothing in particular.
Well, now!
Pine trees, a lake, a summer sky.
That's better. And now?
A book.
Read a page for me.
I can't. My eyes are carried beyond the page.
Try this lens.
Depths of air.
Excellent! And now?
Light, just light, making everything below it a toy world.
Very well, we'll make the glasses accordingly.

E. L. Masters, *Spoon River Anthology*, Dippold The Optician

1 Open Computational Recycling

In less than two decades, the investigation of the history of how humans understood and conceptualized the world (history of philosophy, history of ideas, history of science, history of knowledge, intellectual history, and other related fields) is opening up to the use

https://doi.org/10.1093/llc/fqac014 Advance Access published on 6 April 2022

of computational methods and digital tools. The integration of computational approaches in historical disciplines has been recommended as both a key method to deal with long-term historical transformations and to increase the societal impact of historical research (Guldi and Armitage, 2014). The development of a computational history of science has been recognized as a promising perspective for future research in the field (Laubichler *et al.*, 2013). Attempts to use digital tools and computational approaches are also increasingly popular in the study of the history of early modern and modern philosophy and science (Valleriani, 2017; Sangiacomo, 2018; Sangiacomo and Beers, 2020), history of ideas (Betti and van den Berg, 2014, 2016), and intellectual history (De Bolla, 2013; Bourke, 2017). Moreover, facilitating the integration of digital approaches in more traditional fields such as the history of philosophy and science is also crucial for digital humanities in order to ensure a wider dissemination of new computational practices, to calibrate them against the research agendas of different disciplines (Hayles, 2012; Liu, 2013), and to solve disciplinary imbalances (Wang, 2018).

The study of past ideas often relies on the study of how these ideas were expressed in writing. Traditionally, this means accessing a library and starting to read books or other written materials. Digital humanities offer a vast array of tools and methods for semantic analysis of both small- and large-scale textual corpora (Jockers, 2013; Recchia *et al.*, 2017). However, implementing these methods requires corpora that are available in sufficiently high-quality digital transcriptions. Until now, the limited availability of high-quality transcriptions of early modern texts has been a major stumbling block for the implementation of computational approaches in the history of early modern philosophy and science in particular. The variety of fonts and layouts, combined with irregularities produced by the printing process itself, make the optical character recognition (OCR) of texts printed before 1800 difficult. This often leads to low-quality transcriptions, unreliable for digital mining. Extensive projects have been recently developed to address this issue. The 'IMPACT' (2008–12) and the 'eMOP' projects (2012–15) aimed to develop new approaches to increase OCR quality (Mandell *et al.*, 2017). The 'Text Creation Partnership' (TCP) established a large-scale campaign of human transcriptions that made available 48,300 titles from the 'EEBO' and 2,400 titles from the 'ECCO' collections. Meanwhile, various initiatives worldwide (among them, Google Books, Gallica, and archive.org) disseminate an increasing amount of early modern digitized texts online, although most often in quite poor OCR quality.

Today, scholars who would like to implement computational approaches for semantic analysis to study an early modern corpus seem to have three main options: (a) rely on human transcriptions of the corpus; (b) digitize *ex novo* the whole corpus in high resolution and then implement OCR techniques capable of yielding sufficiently high-quality results; or (c) rely on the already available low-quality OCR (when human transcriptions are not available). The first option might provide the best results, but it is also expensive and time-consuming, especially if the corpus one has to study has not been the object of a past transcription campaign. Hardly any individual scholar, or even a small team, could afford this investment. The second option can be equally expensive. Each library has different policies and fees for digitizing texts and it is likely that no single library collection possesses all the works that would ideally compose the corpus one wants to study. Moreover, this solution does not provide much guidance about how exactly to adjust the OCR technology in such a way that it can provide high-quality results, given the potential diversity among formats, fonts, and languages that can characterize a vast and diverse corpus. The third option is the easiest and cheapest to implement, but it is also the one that comes with greater risks of yielding questionable or meaningless results when used as the basis for computational semantic analysis, due to the excessive amounts of errors in the available OCR.

In this paper, we present a method that combines the strengths of these three approaches, while at the same time compensating for their relative weaknesses. We have designed a workflow for reworking already digitized early modern sources, freely available on the web, in such a way that they would be capable of yielding high-quality OCR results. We also tested this method by using a sample corpus gathered in this way for semantic analysis. The results from these tests show that anything above 90% OCR accuracy (in terms of accurate transcriptions of words per page in each work included in the corpus) is more than sufficient for

performative semantic analysis and the overall homogeneity in the OCR accuracy across the corpus is more important than having perhaps only a few works perfectly transcribed (100% accuracy) while the rest would be available in even just a slightly lower quality.

Before presenting our OCR method, it might be helpful to briefly mention the sources we used for this study. Our testing corpus has been inventoried by implementing a mixed method that combines scholarly bio-bibliographical dictionaries and the automated scraping of the *WorldCat* (Sangiacomo *et al.*, 2021). This process provided a final corpus consisting of 498 OCRed titles (out of a 'wish list' of 788 titles), published between 1600 and 1800, in three different geographical areas (Britain, France, and the Dutch Republic; hereafter referred to as British, French, and Dutch (sub)corpora), in three different languages (English, French, and Latin). This corpus was deemed relevant for the study of early modern natural philosophy and it was also taken as the corpus used to test the method described in this paper.

The preliminary step of our research consisted in looking for the titles on our ideal list that were already available online in the public domain or licensed permissively. As main providers for these works, we consulted Google Books, Bibliothèque Nationale de France (BNF), Munich Digitization Centre (MDZ) at the Bavarian State Library, Early English Books Online (EEBO), and, occasionally, e-rara.ch and archive.org. In general, higher quality scans are available at public libraries such as the Bavarian State Library or BNF; however, Google Books had a far greater number of the desired books available and the quality of their scans proved entirely sufficient for a high-quality OCR – provided we downloaded the original images directly from the interface, rather than obtain them in PDF format, where the quality was significantly lower. As such, over 70% of the initial list of relevant books could be obtained online; over 70% of those books collected were drawn from Google Books. Overall, Google Books proved not only sufficient in quality but also easy to integrate in the workflow and then yielded the majority of the books used.

As to the other sources, there were some slight biases in availability depending on the language and provenance of the original. For example, while hardly any of the books originally published in England or the Netherlands were available on BNF, some 13% of the books used for the French corpus came from BNF. Conversely, EEBO and TCP provided a significant share of the English corpus (ca. 9%), while no books from these sources were used for the Dutch or French corpora. It must be added, however, that since Google Books was the easiest to integrate into the workflow, preference was given to Google Books over BNF or EEBO or other comparable sources, whenever all were available. Thus, the quoted percentage of books used from each source does not reflect precisely their availability by source.

On a side note, EEBO and the related TCP project proved a double-edged sword: while the transcriptions provided by TCP were man-made and, naturally, of maximum accuracy, digitized books available on EEBO were often digital images from old microfilms and of generally very low quality.

In terms of processing time, the initial turnaround was about a week for five to ten books, from download to final delivery. After streamlining the process, as described in the following section, the output was increased to approximately fifty books a week.

2 OCR Method and Workflow

The following argument sketches the evolution of our OCR pipeline, a system developed to handle a large number of OCR processes at an optimum quality in as little time as possible. Processing the books included in our 'wish list' required the bespoke development of software in addition to existing programs. More poignantly, these tools not only had to be developed but embedded in a workflow that would allow for processing a large quantity of books with a quick turnaround, eliminating as much manual intervention from the process as possible while maintaining the best possible quality. The only third-party software we use is the open-source software Tesseract as our OCR engine, the latest version of which uses neural networks for training and recognition purposes (https://github.com/tesseract-ocr), and Amazon Web Services (AWS) cloud servers.

The standard OCR process of any digitized book normally runs through three steps: pre-processing, OCR, and post-processing. Pre-processing binarizes the digitized pages, that is, turns them into black and white, while removing undesirable speckles or discolorations or notes in the margins.

Binarization can happen at different thresholds with more or less aggressiveness and the choice has a significant impact on the legibility and suitability of the text in the image for OCR. For our purposes, we have developed a software that uses Sauvola and Pietikäinen's (2000) Adaptive Binarization algorithm, sped up with Shafait's (2008) integral image pre-calculation technique. This calculates an appropriate binarization threshold for each pixel on a page, using the context of the pixels around it to determine whether each should be black or white. This allows it to gracefully handle pages which have varying degrees of illumination, stains, or other issues, across a page. The following Fig. 1 shows the original image of a page and the same page binarized at various thresholds.

After pre-processing is completed, the OCR step takes the pre-processed pages and performs the actual OCR on the prepared images. Once this is done, the output is analyzed for quality in the post-processing step, determining whether the book has to be reprocessed or is fit for delivery, and the output is turned into searchable PDFs and concatenated text files. The quality of the output is significantly determined by two factors: the OCR model used and the pre-processing steps taken.

OCR engines are general tools that use a 'model' to recognize text in a specific language and script. While it is often possible to get very good results using a general-purpose model, which are generally provided with an OCR engine, more specialist material like historical, early modern printed works can benefit enormously from creating bespoke models. The outline of how this is done is by providing the engine with a pool of representative 'ground truth' training data, which are many samples of text line images with corresponding correct transcriptions. Our specific strategy for training bespoke OCR models has been described in detail in a different article (Hawk *et al.*, 2019). Briefly said, though, the larger and more diverse the ground truth pool within the boundaries of language and script type, the more effective the model for new,

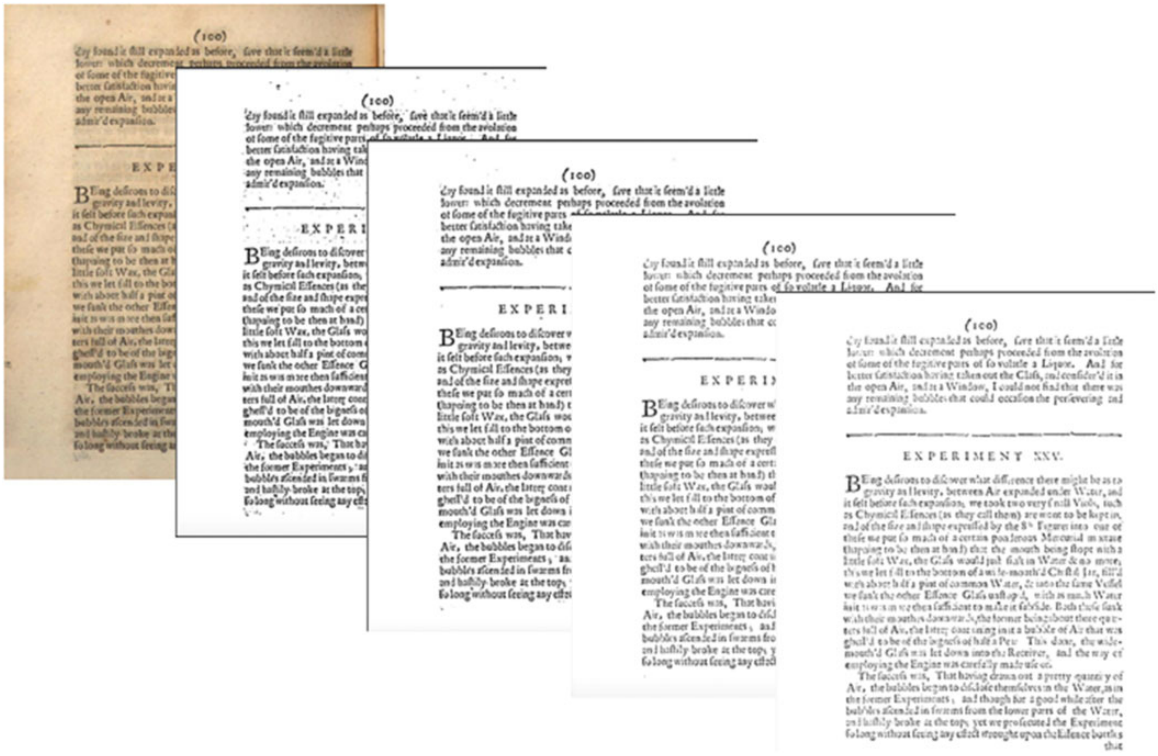

Fig. 1 Examples of different binarization thresholds

similar texts. In this context, the first model trained that achieved sufficient accuracy with unseen texts contained ca. 2,100 image-text pairs from fifteen different books. Throughout the project, we re-trained the model with additional ground truth from the works processed; our latest model was trained on ca. 5,600 image-text pairs from a significantly higher spread of books. Training one Tesseract model up to 100,000 iterations would take 2 to 3 days.

However, the current discussion will focus on the changes made to pre-processing, workflow, and infrastructure to demonstrate the gains made in quality and reduction of processing time. Overall, it was possible to develop a process ensuring maximum quality while reducing the processing time of a single book of average length (ca. 500 pages) from a minimum 20 hr to an average of about 3 hr (though, theoretically nearly infinitely reducible). This was achieved by replacing a step-by-step OCR process with a fully automated pipeline system run on an arbitrary number of servers, breaking up the full process of OCRing one book into minimal tasks that can be handled simultaneously by multiple servers.

The measures of success for this part of the project are both the average time needed to process one book and the consistency and quality of the final output. While the measurement of processing time is quite straightforward, the assessment of quality in the realm of OCR is only ever approximate, never definite, unless a transcription of 100% accuracy already exists. To test the general accuracy a newly trained model can achieve, we use ground truth specifically created and set aside for this purpose. For the majority of the corpus, however, where no such ground truth exists, we use the confidence score provided by Tesseract (exemplified in Fig. 2) on its output as an indicator of the quality of our output, combined with spot-checking. Said confidence score measures the probability with which the OCR engine deems an output to be correct. While that does not equate to word or character accuracy, it does give a qualitative (more than quantitative and absolute) assessment of how the output compares. From our experience, consistent scoring above 75% confidence corresponds with acceptable output of ca. 90–95% accuracy. The eventual accuracy of the whole pipeline amounts to 90–95% which has been checked by sampling the texts.

We harness this output by creating graphs for each book that track the average word confidence of each page. While this does not highlight single mistakes on a page and does not yield the reason for bad output, those graphs quickly indicate on which pages the OCR process encountered problems (e.g., because the page was blank, contained images or non-Roman characters, etc.).

Since the quality of the output depends on a number of factors, not all of which are dependent on pre-processing or OCR models, it is difficult to define the average improvement in quality achieved through our process. However, we developed a workflow that, as far as pre-processing is concerned, would ensure the optimum quality as a matter of process. The biggest gain, therefore, is the removal of any doubt from the process whether a different binarization threshold would be preferable.

3 Iteration of workflow

Our initial process required manual intervention for each step (pre-processing, OCR, and post-processing) using one relatively powerful server. A book was submitted for pre-processing, then OCRed, then analyzed with the help of a confidence graph as described above. If the result was satisfactory, the OCR transcript could be delivered; if the quality was lacking, the book would be re-processed with a different binarization threshold. Diagram 1 below illustrates this initial workflow.

Since we were working with only one server at the time, capacity was limited, and a maximum of three to four books could be processed simultaneously without overburdening the system. Pre-processing could take around 1–2 hr for a 500-page book; OCR might take approximately 4 hrs, since each page takes an average of 30 seconds to OCR on a standard server. The analysis was partly manual, in itself time-consuming (ca. 2 hrs) and fairly speculative. If the quality of the result was not sufficient, the book had to be resubmitted with a different pre-processing threshold—which was a frequent occurrence—, the whole process would be repeated and the total processing time doubled from 8 to 16 hrs, not counting the time lag and manual interventions.

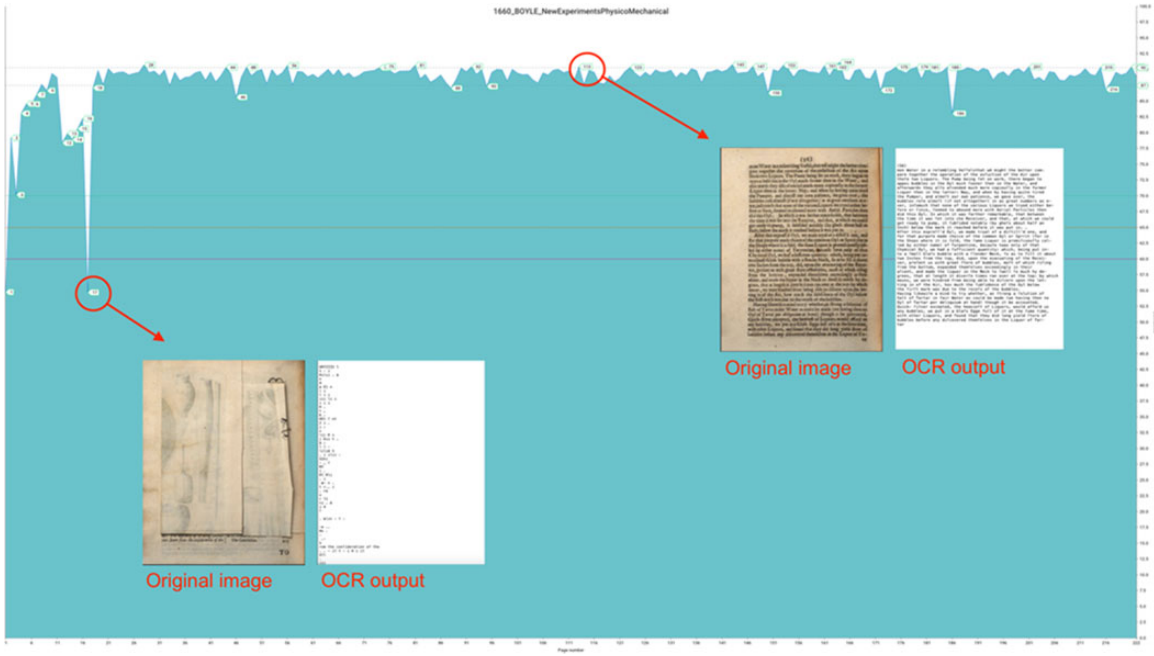

Fig. 2 Average confidence score for each page in a processed book

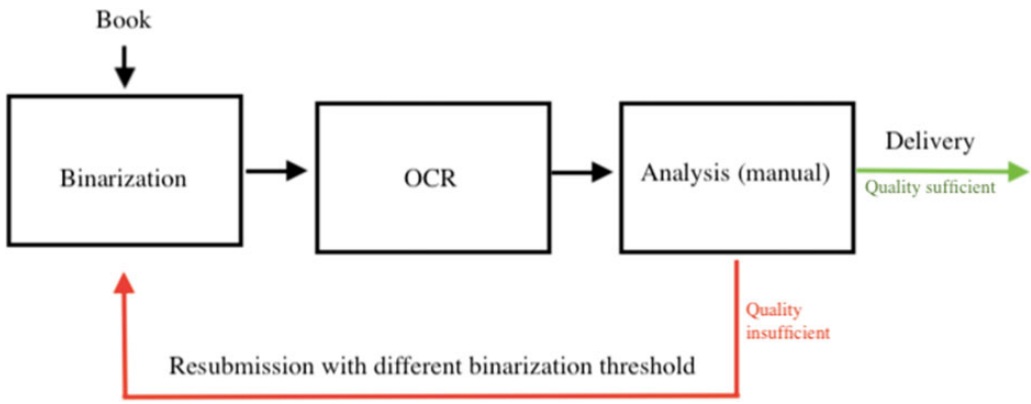

Diagram 1. Initial workflow

The central issue in this process is the identification of the correct binarization threshold. Firstly, it is hard to choose the right threshold for a book in advance and, secondly, different pages in one and the same book might profit from different thresholds. Resubmitting a book to the pre-processing step with a different binarization threshold could solve only the first issue, but not the second. Moreover, this resubmission step was the most time-consuming element in the whole workflow.

As a solution, a new approach to pre-processing was introduced whereby each page in a book is processed at multiple thresholds (four, in our case). Each version for each page is submitted to the OCR process. Once processed, the page with the highest average word confidence for each page is chosen for

each output. This improved workflow is illustrated in Diagram 2.

This new workflow removed the issue of resubmitting and ensured that for each page the optimum pre-processing thresholds were chosen, or rather, the pre-processed version of a page that scored the highest average word confidence. While confidence is not the same as accuracy, it does give a reliable assessment in relative terms and manual validation has shown that among differently pre-processed pages, the one with the highest average word confidence score reliably corresponded with the best output. In that sense, the problems outlined above (choosing the right threshold from the outset and catering to different conditions across the book) were both solved. Although the qualitative difference between various thresholds might not be more than a fraction of percentage points in accuracy in most cases, the difference can have a severe impact on the OCR quality in some isolated cases of pages or whole books. Streamlining this multi-threshold pre-processing, therefore, not only prevents bad processing of fringe cases from the start; more poignantly, even where the difference in quality is not big, the need for manual checking is eliminated because each page in the book is processed at the best possible quality by definition.

This improvement of the process increased the quality and eliminated much manual intervention from the process. On the downside, however, this new method significantly increased the amount of time needed to process a book since each page was now reproduced in four differently binarized versions and each submitted to the OCR process. Thus, the OCR processing time was also quadrupled. Since each page takes a minimum of 30 s to OCR, processing a book of 500 pages would mean 2,000 times 30 s for OCR alone (i.e., 1,000 min or 16.7 hrs per book). This means that even if the spectre of resubmission and thereby doubling of processing time was removed, the time needed to process a book from the outset already took that amount of time as a minimum, even with manual intervention removed. To put the time constraint in context, processing some 100 books would take some 1,670 hrs (i.e., almost 70 days in processing time alone, not including post-processing of each book and delivery), which made this no more than a compromise..

To eliminate this time constraint, the next step was to preserve the improved quality and workflow but to change the technical infrastructure in order to make it possible for this higher volume of processing to happen in a minimal amount of time. Up to this point, we had only been using a single server for the entire workflow. The next adjustment, thus, changed the process in a way to make it possible for multiple servers to do the processing simultaneously.

In order to employ several servers but avoid having them duplicate tasks, a pipeline was built that managed pre-processing and OCR of each book through a queue system. Concretely, a book submitted to the pipeline would be entered into the 'pre-processing queue.' Any active server made available would be alerted to a task in the queue; the book would then be binarized at four different thresholds and each binarized page would be added to the next queue in the system, the 'OCR queue.' Once the entire book is pre-processed, the task is deleted from the 'pre-processing queue.' Meanwhile, the pages entered in the

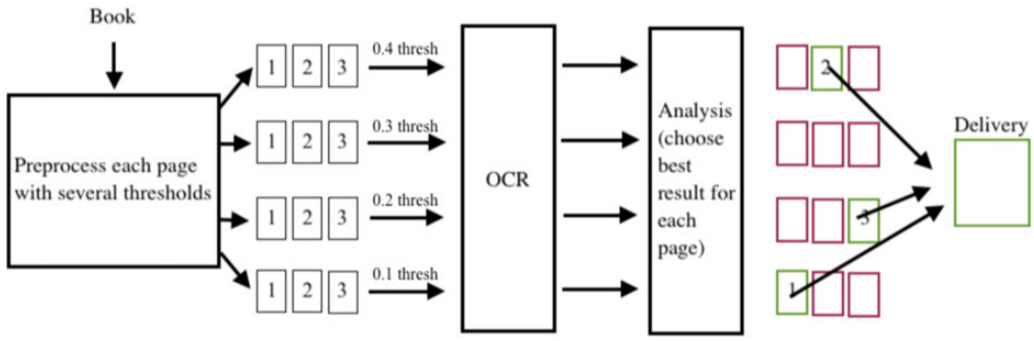

Diagram 2. Improved workflow with multiple binarization thresholds

'OCR queue' are picked up and processed by available servers. Each page processed is deleted from the queue; once the entire book is processed, it is passed on as a whole to the 'Analyze queue' stage where all pages are evaluated in a graph and the best versions for each page are compiled to make up the full processed book, which is available for delivery once the process analysis is complete. Figure 3 illustrates this queue system.

While it seems inordinately untidy to enter single pages rather than the entire book as tasks in the 'OCR queue,' this step made it possible to reduce the processing time to an absolute minimum. Splitting up a book into one task per page, rather than one task per book, makes it possible for any number of servers to work on the same book simultaneously. The time needed to OCR all the pages can therefore be reduced almost infinitely, even if each page is OCRed four times (once for each of the four different binarization thresholds). Moreover, since the task as a whole is still the same, having more servers sharing the same task is not actually any more expensive in terms of total processing power. Moreover, a technically speaking unlimited number of books can be processed simultaneously without slowing down the process.

To illustrate the amount of time saved by this step, a 500-page book, binarized at four thresholds, would (as per the above) require approximately 1,000 minutes processing time, that is, ca. 16.7 hrs. Divided among six servers running simultaneously, however, the OCR processing time can be cut down to less than 3 hrs. Technically speaking, the number of servers set to the task is not limited. Theoretically, therefore, it is possible to use seventeen servers to do the same amount of processing at the same cost in less than an hour.

Our server infrastructure used to support this workflow is quite straightforward. We created a disk image (or 'snapshot' in AWS terminology) containing our core pipeline software on top of a basic Debian cloud image and configured to update and start automatically using *systemd*. We could then create new server instances as we needed them with one command, which would immediately check the work queues and start processing. To ensure the servers didn't stay up for longer than needed, we included an inactivity timer into the pipeline which automatically shut down the server after a short period if there had been no work to do.

The pipeline was designed to send a regular 'heartbeat' message to the queue system while processing a job, so that if it is disrupted for any reason, the job is soon made available for other servers to complete instead. This makes the pipeline very robust to failures

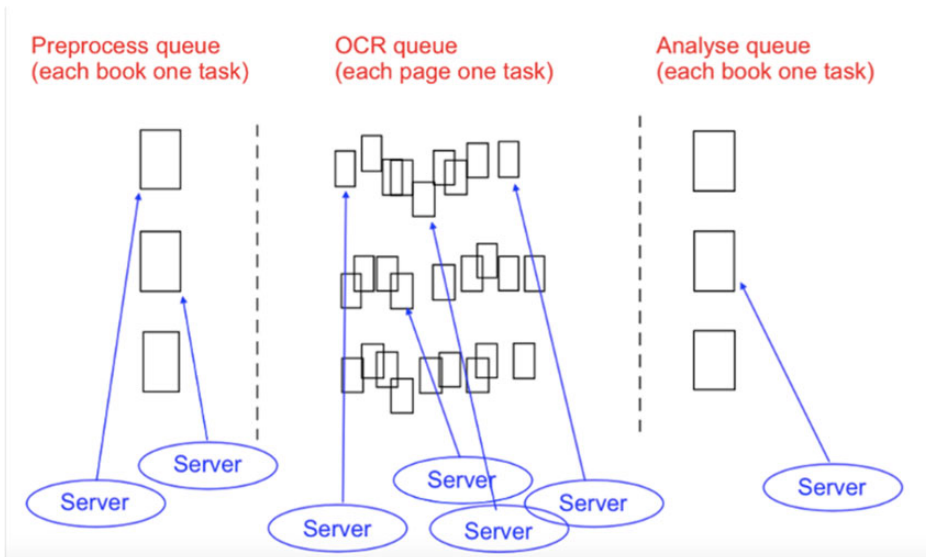

Fig. 3 The queue system

or unexpected outages, which has meant that we have been able to take advantage of the unexpensive 'Spot Instance' servers that AWS offers, which can be cancelled without warning at any time.

In summary, the crux to developing a robust and time-sensitive OCR pipeline relied to a great extent not just on the excellence of the software inputs (OCR models and pre-processing software) but also on the development of an infrastructure that makes the time-sensitive processing of a large volume of books possible. This is important not only for delivering the output on time, but also for yielding plenty of capacity and space to allow for testing and iteration. In that sense, the goal of our pre-processing adjustment was to ensure the 'best possible' quality regarding different binarization thresholds; the infrastructure, then, was built in response, to make this condition a feasible option.

While we have largely solved the issues to do with pre-processing, other factors remain that can have an impact on OCR quality: low-quality scans, script in non-Latin alphabets, mathematical formulae beyond the OCR model's scope, and images or decorative borders on the pages. Still, with the exception of these cases, our current OCR models combined with the pre-processing system described normally achieved an average word accuracy between 90% and 95% for books printed in the seventeenth and eighteenth centuries. On average, this quality tends to be a bit higher for books printed in the eighteenth century, albeit the percentage point difference compared with older books would not exceed 1% or 2%. Although the basis of our trained models is built on Tesseract's standard model for modern English texts and a slight bias can be expected toward historic English texts, retraining with an appropriate amount of historic French and Latin ground truth brought the accuracy of books written in these languages up to the same bracket.

4 Playing with Imperfect Accuracy

The OCR quality obtained by implementing the method just described is very high, but it does not provide 100% accuracy. While one might wish for perfection, this degree of accuracy (or inaccuracy left, if one wants to see the flip side of it) proved to be more than sufficient for the purposes of computational text mining. This point is important since it takes some pressure out of the need for working necessarily with flawless OCR results, which might be difficult when a project has to rely on already digitized texts or just with a diverse corpus of early modern books covering a few centuries. Albeit not flawless from a human point of view, our tests and investigations revealed two important points on this front: not only is 100% word accuracy not strictly needed in order to yield meaningful statistical results, but, more importantly, homogeneity in the overall word accuracy level across the corpus has a direct impact on statistical and semantic analysis.

In order to illustrate these points, we offer two considerations: (a) we run some tests in which we manually corrupted the same texts at a fixed ratio and checked how this was reflected in the semantic analysis performed on that text; and (b) we illustrate how using the OCR results produced with the above-described method produced insightful results and the 'OCR noise' was somehow easily detected and handled through semi-automated cleaning (by removing, for instance, stop words and other recurrent 'noise' that could be spotted by the human eye). By contrast, introducing just one work in a fully accurate transcription within our corpus (hence, making its overall OCR accuracy slightly less homogenous) resulted in apparent biases in favor of that work.

(a) The impact of OCR inaccuracy highly depends on the type of research task the OCRed materials are used for. For some of such tasks, results have been obtained of the impact of the OCR inaccuracy and the required minimal accuracy before meaningful results can be found. In general, anything above 85–90% accuracy is consistently found to be sufficient, although lower accuracy can sometimes be sufficient as well (Hill and Hengchen, 2019; Van Strien *et al.*, 2020). However, not all research tasks are accounted for in the literature.

In our case, we conducted internal tests for the task of identifying the correct top-n collocates for a particular wordtype in a corpus. The impact of OCR (in)accuracy on this task has not been tested previously, although the study of collocation is usually very relevant for semantic analysis (Brezina *et al.*, 2015; De Bolla *et al.*, 2019). For present purposes, we define 'correctness' as the overlap with the extracted

collocates that one would find in a 100% accurate corpus. This is investigated for four different cases: extracting fifteen, thirty, forty-five, and sixty collocates for a word from a text. Notice that this does not entail a qualitative investigation of the collocates found since this would also take into account the problems a collocate analysis might incur irrespective of OCR inaccuracies. We used a reverse method, namely, an algorithm that simulates OCR inaccuracies by iteratively corrupting digitally native texts (Wikipedia entries). The collocates derived from the original text are compared with the texts output by the OCR simulation at differing levels of inaccuracy. This is done both for a method that makes use of a naïve inaccuracy inducer (randomly switching *n*% of all characters to another character) and one that incorporates some of the biases found in actual OCR (in this case, we consider the very common mistake of reading the long *s* as *f*). Results are provided here below in Figs 4 and 5.

The results are very promising, showing how unbiased inaccuracy (completely random mistakes) can still lead to high average levels of success from 70% accurate texts upward. For texts that simulate also some of the more often encountered particular problems of OCR, somewhat higher levels of accuracy are required, still allowing for significantly good results from 72% upward. From 80% upward, almost perfect results are generated, irrespective of the size of the group of collocates that is extracted from the text.

These results need qualification. For one, the specific research task defined significantly impacts the required OCR accuracy (see below for such a case). Secondly, introducing biases particular to the OCR (instead of a fully naïve scrambling) shows slightly worse results. This might indicate that a more accurate rendering of the particularities of OCR inaccuracy might lead to worse results. However, the reduction in accuracy is very small. In addition, a more accurate rendering of OCR inaccuracies would make the results specific to the OCR machines that are modeled in this rendering. The above results model no particular OCR machine, making the results more broadly applicable.

In addition, the expectation is that one could further increase the robustness of OCR-inaccuracy in research practice. The removal of words that are found to interfere considerably upon the inspection of the collocates or the *ad hoc* replacement of certain word types that are known not to occur in the corpus, but are the result of incorrect interpretation of certain characters by the OCR (e.g., 'caufa' has no meaning in our Latin corpus, but the term 'causa' does) can

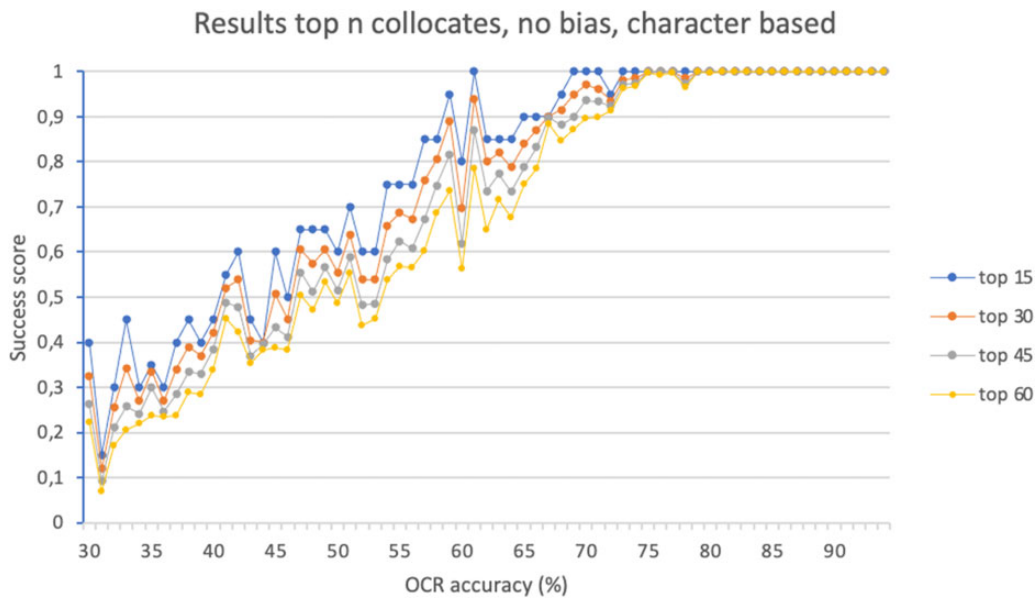

Fig. 4 Accuracy for the top *n* collocates on different levels of accuracy, with no biases

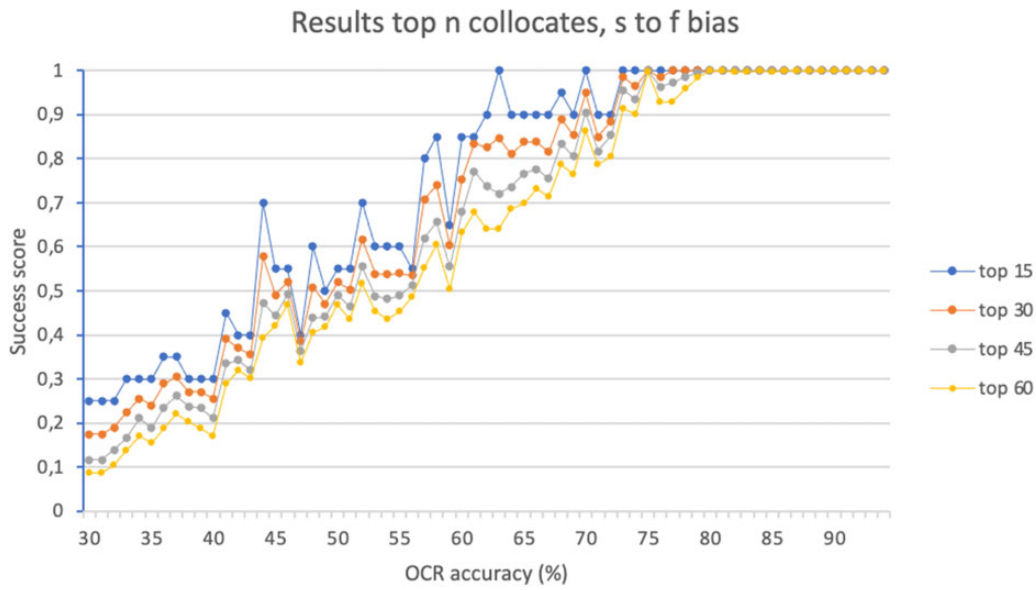

Fig. 5 Accuracy for the top *n* collocates on different levels of accuracy, with biases

Table 1. Comparative network positioning by degree and Eigenvector of Cavendish's twentieth century and re-OCRed modern editions cf. tf-idf vector correlation

Work (node)	Degree	Eigenvector centrality
Cavendish, *Observations* twentieth century	9.44	0.1
Cavendish, *Observations* (1666)	6.84	−0.01
Cavendish, *Grounds* twentieth century	9.18	0.1
Cavendish, *Grounds* (1668)	8.27	−0.01

significantly improve results with relatively little effort and little chance of making unknown mistakes.

(b) While 100% accuracy might not be necessary, working with an average homogeneous OCR accuracy across the corpus has a direct and significant impact on the results obtained by text mining. Recent literature shows that OCR accuracy definitely has an impact on downstream NLP tasks, from sentence segmentation and name entity recognition to word vectorization and topic models (Van Strien *et al.*, 2020). We would like to illustrate this point by discussing one particular case, in which we processed the English subcorpus digitized using the OCR method described above, which also included two texts derived from contemporary editions and hence flawless in terms of OCR accuracy (i.e. 100% versus the average 90–95% for the rest of the corpus).

The corpus was analyzed using td-idf vectors (Lavin, 2019) in order to identify a degree of similarity between the various works included based on textual features. We included the contemporary editions of Margaret Cavendish's works (*Observations* and *Grounds*) in the corpus formalized as a network, in which the nodes are the documents and the edges are the tf-idf correlation scores established between the nodes. Table 1 shows that the twentieth-century files ranked prominently in terms of degree (number of links) and Eigenvector centrality (overall node prominence). They are well connected and typically connect to other well connected works.

However, when we replaced the contemporary editions with the re-OCRed early modern editions of Cavendish's works, her place in the network changed radically. For instance, while the twentieth-century *Observations* file is ranked in the first half of the corpus in terms of number of links, the period file becomes the lowest ranked, having the lowest number of connections and one of the highest betweenness centralities. The latter fact may be a direct result of the OCR quality of the digitized version, more in line with the rest of the corpus.

A high betweenness centrality calculated on the grounds of tf-idf vector correlation scores typically shows

Table 2. Comparative network positioning by betweenness centrality and strongest correlation of Cavendish's twentieth century and re-OCRed modern editions, cf. tf-idf vector correlations

Work (node)	Tokens	Vectors	Betweenness centrality	Strongest correlation
Cavendish, *Observations* twentieth century	49,411	5,068	0.73	WILSON,
Cavendish, *Observations* (1666)	61,343	16,925	0.98	*The Principles of Philosophy*(1754)
Cavendish, *Grounds* twentieth century	27,789	3,597	0.66	
Cavendish, *Grounds* (1668)	31,167	6,106	0.68	

a peculiar type of writing, but one that sits on the shortest paths between other texts. We interpret this as a reflection of Cavendish's sophisticated profile as philosopher, poet, scientist, fiction writer, and playwright. The betweenness centrality scores presented in Table 2 suggest that the digitization of the twentieth century edition leveled out her writing by having an impact on the generated word vectors, making it more similar to the other works in the corpus, although the other files themselves were early modern editions. Both early modern editions contain a higher number of tokens and vectors, which resulted in higher betweenness centrality scores and, hence, in Cavendish appearing on more shortest paths between various other works in the analyzed corpus. One of the main reasons for this discrepancy is the elimination of Old English allographs, such as long s, in the more recent digitized versions. Nevertheless, the strongest correlations established by her books across the same corpus do not differ, as Cavendish remains strongly connected to Wilson's *Principles of Philosophy* irrespective of the OCR quality of the files.

We then suggest that OCR accuracy does have an influence on the way books in a corpus related to each other in some respects, but not to the extent to which it would level out particularly strong connections. These results corroborate existing findings according to which an 80% quality (and ideally 90%) OCR accuracy should not impact downstream NLP tasks like the ones we have used.

5 Conclusions

We would like to end our discussion by stressing the positive and hopeful perspective that our investigation reveals. For a long time, researchers interested in applying digital methods to written early modern materials were faced with almost insurmountable difficulties due to digitizing and OCRing the works they wished to study. The method we presented in this paper shows how even a small team can manage to re-use the wealth of already digitized materials available in the public domain and then re-work them in order to obtain a high OCR quality (90% or above). We also tested the robustness of our outputs for the sake of digital text mining and verified that, albeit not absolutely flawless, it is more than sufficient for investigating it with sophisticated digital and statistical methods and for yielding significant results. We also observed that our method tends to create a homogeneous level of OCR quality across the corpus that it delivers, which might be more important for the sake of obtaining meaningful results, than including a more diverse range of sources with greater differences in terms of OCR accuracy. This means that it might be worth investing the time of re-OCRing a whole corpus by using a single method like the one we describe, which yields homogenous OCR results, rather than simply incorporating heterogeneous OCR sources and perhaps sprinkling them with a few flawless transcriptions. While what is available today in terms of digitized materials and OCR quality tends to be heterogeneous and often below the threshold of sufficient accuracy, it is not necessary to re-do the whole digitization process. It is feasible to improve on what is already available and then work around the relatively minimal degree of inaccuracy and noise that is left.

Funding

This article is part of a project that has received funding from the European Research Council (ERC) under

the European Union s Horizon 2020 research and innovation pro- gramme [grant number 801653].

References

Betti, A. and van den Berg, H. (2014). Modelling the history of ideas. *British Journal for the History of Philosophy*, **22**(4): 812–35.

Betti, A. and van den Berg, H. (2016). Towards a computational history of ideas. In Wieneke, L., Jones, C., Düring, M., Armaselu, F. and Leboutte, R. (eds), *Proceedings of the Third Conference on Digital Humanities in Luxembourg with a Special Focus on Reading Historical Sources in the Digital Age*, Vol. 1681, CEUR Workshop Proceedings, CEUR-WS.org.

Bourke, E. (2017). Female involvement, membership, and centrality: a social network analysis of the Hartlib circle. *Literature Compass*, **14**(4): e12388. doi:10.1111/lic3.12388

De Bolla, P. (2013). *The Architecture of Concepts: The Historical Formation of Human Rights*. New York: Fordham University Press.

De Bolla, P., Jones, E., Nulty, P., Recchia, G., and Regan, J. (2019). Distributional concept analysis: a computational model for history of concepts. *Contributions to the History of Concepts*, **14**(1): 66–92.

Brezina, V., McEnery, T., and Wattam, S. (2015). Collocations in context: a new perspective on collocation networks. *International Journal of Corpus Linguistics*, **20**(2): 139–73.

Guldi, J. and Armitage, D. (2014). *The History Manifesto*. Cambridge: Cambridge University Press.

Hayles, N. K. (2012). *How We Think: Digital Media and Contemporary Technogenesis*. The University of Chicago Press, Chicago and London.

Hawk B., Karaisl A., and White, N. (2019). Modelling medieval hands: practical OCR for Caroline minuscule. *Digital Humanities Quarterly*, **13**(1). http://www.digitalhumanities.org/dhq/vol/13/1/000412/000412.html

Hill, M.J. and Hengchen, S. (2019) Quantifying the impact of dirty OCR on historical text analysis: Eighteenth century collections online as a case study. *Digital Scholarship in the Humanities*, **34**(4): 825–843. https://doi.org/10.1093/llc/fqz024

Jockers, M. (2013). *Macroanalysis: Digital Methods & Literary History*. Urbana, Chicago: University of Illinois Press.

Laubichler, M. D., Maienschein, J., and Renn, J. (2013). Computational perspectives in the history of science: to the memory of Peter Damerow. *Isis*, **104**(1): 119–30.

Lavin, M. J. (2019). *Analyzing Documents with TF-IDF. Programming Historian*. https://programminghistorian.org/en/lessons/analyzing-documents-with-tfidf#how-the-algorithm-works.

Liu, A. (2013). The meaning of the digital humanities. *PMLA*, **128**(2): 409–23. doi:10.1632/pmla.2013.128.2.409

Mandell, L., Neudecker, C., Antonaco, A. *et al.* (2017) Navigating the storm: IMPACT, eMOP, and agile steering standards. *Digital Scholarship in the Humanities*, **32**(1): 189–94.

Recchia, G., Jones, E., Nulty, P., Regan, J., and de Bolla, P. (2017). Tracing shifting conceptual vocabularies through time. In Ciancarini, P., Poggi, F., Horridge, M. ***et al.*** (eds), *Knowledge Engineering and Knowledge Management. EKAW 2016. Lecture Notes in Computer Science*. Vol. **10180**. Dordrecht: Springer.

Sangiacomo, A. (2018). Modelling the history of early modern natural philosophy: The fate of the art-nature distinction in the Dutch universities. *British Journal for the History of Philosophy*, **27**(1): 46–74. doi:10.1080/09608788.2018.1506313

Sangiacomo, A. and Beers, D. (2020). Divide et impera: modelling the relationship between canonical and non-canonical authors in the early modern natural philosophy network. *Hopos*, **10**: 365–413.

Sangiacomo, A., Tanasescu, R., Donker, S., and Hogenbirk, H. (2021). Expanding the corpus of early modern natural philosophy: initial results and a review of available sources. *Journal of Early Modern Studies*, **10**(1): 107–15.

Sauvola, J. and Pietikäinen, M. (2000) Adaptive document image binarization. *Pattern Recognition*, **33**(2): 225–36. https://doi.org/10.1016/S0031-3203(99)00055-2

Valleriani, M. (2017). The tracts on the sphere: Knowledge restructured over a network. In Valleriani M. (ed.), *The Structures of Practical Knowledge*. Dordrecht: Springer, pp. 421–73.

Van Strien, D., Beelen, K., Coll Ardanuy, M., Hosseini, K., McGillivray, B., and Colavizza, G. (2020). *Assessing the Impact of OCR Quality on Downstream NLP Tasks. Proceedings of the 12th International Conference on Agents and Artificial Intelligence*. Vol. 1: ARTIDIGH, Valletta, Malta, pp. 484–96.

Wang, Q. (2018). Distribution features and intellectual structures of digital humanities: a bibliometric analysis. *Journal of Documentation*, **74**(1): 223–46.