Best practice for digitising small-scale Digital Humanities projects

Peggy Bockwinkel[1] and Dîlan Cakir[2]

[1] University of Stuttgart, Digital Humanities, Herdweg 51, 70174 Stuttgart, Germany
[2] University of Stuttgart, Stuttgart Research Centre for Text Studies, Azenbergstr. 12, 70174 Stuttgart, Germany

Abstract. Digital Humanities (DH) are growing rapidly; the necessary infrastructure is being built up gradually and slowly. For smaller DH projects, e. g. for testing methods, as a preliminary work for submitting applications or for use in teaching, a corpus often has to be digitised. These small-scale projects make an important contribution to safeguarding and making available cultural heritage, as they make it possible to machine read those resources that are of little or no interest to large projects because they are too special or too limited in scope. They close the gap between large scanning projects of archives, libraries or in connection with research projects and projects that move beyond the canonised paths. Yet, these small projects can fail in this first step of digitisation, because it is often a hurdle for (Digital) Humanists at universities to get the desired texts digitised: either because the digitisation infrastructure in libraries/archives is not available (yet) or it is paid service. Also, researchers are often no digitising experts and a suitable infrastructure at university is missing.

In order to promote small DH projects for teaching purposes, a digitising infrastructure was set up at the *University of Stuttgart* as part of a teaching project. It should enable teachers to digitise smaller corpora autonomously.

This article presents a study that was carried out as part of this teaching project. It suggests how to implement best practices and on which aspects of the digitisation workflow need to be given special attention.

The target group of this article are (Digital) Humanists who want to digitise a smaller corpus. Even with no expertise in scanning and OCR and no possibility to outsource the digitisation of the project, they still would like to obtain the best possible machine-readable files.

Keywords: digitisation, OCR, Gothic type, best practice, DH teaching

1 Introduction

1.1 The starting point: How to boost DH teaching?

The aim of our study was to establish best practice scenarios for DH projects in university teaching at the *University of Stuttgart* for which smaller corpora have to be scanned.[1] The hurdle to start a DH project should be kept as low as possible: In order to support the projects as best as possible, the necessary equipment (laptop and software) was made available for the teachers/researchers of the *Faculty of Philosophy and History* at the *University of Stuttgart*.

We wanted to create the most extensive possible synergies in terms of digitisation (scanning and OCR processes) – and with that, we wanted to be as sustainable as possible. Therefore all manuals and documentation of the OCR study presented in this paper are published on *github*.[2] For our study we considered the guidelines for digitisation provided by the DFG (German Research Foundation) [1], which are comparable to the FADGI guidelines for English [2].

1.2 Why another paper on digitising?

There are many studies on scan and OCR processes and their improvement. We will name three of them to make clear the added value of our study:

The first study is an analysis of OCR accuracy of historical newspaper digitisation [3]. It was introduced in a short paper by the National Library of Australia and published in 2009 and gives valuable hints how to improve the OCR. The analysis is made on basis of a large scale digitisation programme, which serves different needs compared to small scale programmes. For example: The scanned files were processed by contractors, which means, that the data volume of the files was adapted to the limited capacity of the contractor. Interestingly, it turned out in the end that the best solution for this programme was a manual correction of OCR mistakes by public users.

The second paper is an extensive study on improvements of the quality of mass OCR for "Old Prints" in German [4]. The study itself is – like the National Library study – very structured and reflected. However, it is not transferable to smaller projects, which means the resources and the equipment of large digitisation projects are very different and therefore not comparable with those of our study/target group. For example: The technical details refer to scanners/cameras that are designed for large scale digitisation. It is also difficult for newcomers to OCR to assess which suggestions (for improvement) could be used for smaller projects. For OCR professionals, this publication is an immense enrichment - for newcomers, it is rather overwhelming.

We are not the first to realise that large scale digitisation is difficult to apply to smaller projects. There is one study we would like to mention. It is on OCR tools and a proposal for a workflow for small-scale historical archives [5]. However, the technical

[1] The implementation of the study took place as part of the teaching project 'The Digital Archive' at the University of Stuttgart. The teaching project was intended to familiarise students of the humanities with the digital component in their first bachelor semesters. The project was carried out from 2013-2016 and financed by the MWK (Ministry of Science, Research and the Arts, Baden-Württemberg), see also www.uni-stuttgart.de/dda.

[2] https://github.com/Bockwinkel/OCR

challenges are rather high: The recommended OCR software is open source, which we are very much in favour. However, it is a command-line tool. We regard the hurdle for our target group to familiarise themselves with the command line and use tools that work exclusively via the command-line as very high.

Our conclusion is that these studies are useful for digitising experts, but not for novices like our target group (who have no interest in becoming OCR experts). Therefore, the aim of our study was to find a way to get the best OCR with the least (financial) effort on hard-/ software and manpower. Yet usability and sustainability should still be as high as possible. In detail we wanted to sort out how to get the best OCR results with an average book scanner or a normal photocopier with scan function. This is also a part of our sustainability strategy: To use the equipment, that is already there, so that no extra room and budget is needed.

The aim of the paper is to present the results of the study and highlight what needs to be considered in order to achieve the best possible results during the digitising process. The paper might be useful for researchers and teachers who are no OCR experts, yet have to digitise small amounts of texts and want to learn which points they need to pay attention to.

2 Method

2.1 Initial situation for researchers

Humanities researchers are confronted with texts of two predominant type faces: Antiqua and Gothic type. Antiqua has been used for German texts since the first half of the 20th century. For texts from the time before, one is confronted with Gothic type, which has the reputation of being difficult to convert automatically into machine-readable texts. We wanted to find that out for ourselves and decided to use texts in Gothic type for our project.[3] For these texts there is another method that is very often used and preferred to the automated conversion, but which is also very time-consuming and therefore cost-intensive: the double keying method. Two different persons manually type the text, so that at the end two versions of the text are available as a file. Both files are compared with each other in order to be able to find all possible errors. We compared the automated conversion and the double keying to determine which method is the most effective for our situation.

[3] Automatically converted texts can also be checked for errors quickly, if you use software that uses different conversion methods. This is explained in more detail in a project similar to the one we introduce here and which was presented at the DHd2016 in Leipzig [6].

2.2 Choice of hardware and software

After talks with digitisation experts, we decided in favour of the *Abbyy Fine Reader* software.[4] The freely available open source software *OCRopus* or *Tesseract* has been set aside because a user interface is not or hardly available there, which was a disadvantage since our potential users from the humanities are accustomed to mature GUIs.

Two different scanners were available: a book scanner, in which the book is opened and the camera takes pictures from above, and a copier with scanning function, in which the book is placed on a glass plate and scanned.[5] Since each camera is different, systematic tests were carried out in advance to compare the quality of the cameras.[6] The parameters used were: dpi (200, 300, 400, and 600) and the colour setting, i. e. black and white, greyscale or colour.[7] With the book scanner, only the colour setting can be adjusted at 300 dpi each.

2.3 Questions and implementations

There were mainly two questions we wanted to answer: 1. Which scanner produces the lowest number of errors with which settings, i. e. with which parameters? The workflow for answering the first question looks like this:

Step 1: Systematic testing of the cameras with ‘sun picture’ and millimetre paper

Step 2: Scanning one page, Gothic type, with all the different settings on both scanners.

Step 3: Training *Abbyy* with single letters and the common German letter combinations ‘st’ and ‘tz’.

Step 4: Error analysis

2. Which method is more efficient: double keying or automated OCR conversion? To answer the second question, eight pages of a Gothic type text were scanned and the conversion was documented in order to compare the two conversion methods.

[4] We installed *ABBYY FineReader 12 Professional* (no volume license) on a notebook to be able to give it away to the researchers who want to use it.

[5] We used the scanners that were easily accessible for us: the book scanner at the university library and the institute copier.

[6] The 'sun picture’ and millimetre paper are a simple and inexpensive way to test the camera quality and to find out how well the camera is calibrated, see lecture by archivist Klaus Wendel (www.archium.org) on scanning and pre-scanning work during summer semester 2015. A website with detailed description of the lecture ‘Computer science for historians’ is found at http://www.uni-stuttgart.de/dda/lehre/InformatikHistoriker.html , last accessed 2017/11/01.

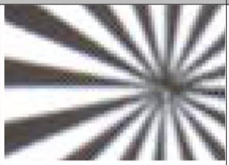
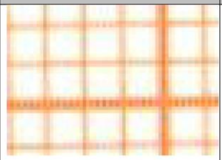
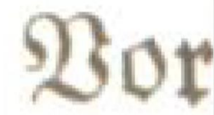
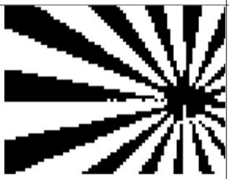
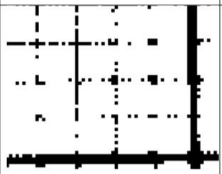
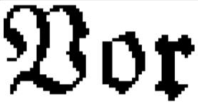
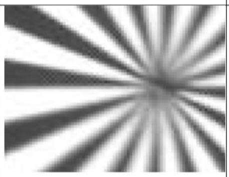
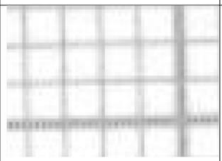
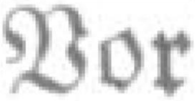
[7] ‘Colour’ is divided into full colour (256 colours) and automated colour.

3 Results

3.1 Results of 1st - 3rd step (first question)

After scans of the ‘sun picture’ and the millimetre paper were made, the results including the number of errors were compared on a double page (cf. table 1). You can clearly see the differences between black/white images and colour, or greyscale: The b/w images are much more pixelated despite the same dpi number. Not surprisingly, the number of errors on a double page scanned in b/w is also significantly higher than with the other two colour settings. The best result with seven errors was delivered by the colour scan. The greyscale scan has only two more errors.

Table 1. The results of the scans with the book scanner brought together in one chart; the best result with seven errors is highlighted in green.

colour settings	dpi	„sun picture“	millimetre paper	detail of gothic types	number of mistakes in one double page, gothic types
colour	300				7
black/ white	300				17
grey scale	300				9

3.2 Results of 4th step

In the 4th step, the errors were documented and analysed. *Abbyy* was trained for 15 to 30 minutes per double page. Individual letters and the letter sequences ‘st’ and ‘tz’ were trained, as they are a common letter combination in German. Table 2 lists the errors already mentioned in Table 1: seven errors with a double page in colour and nine errors with a double page in greyscale. It is noticeable that the error 'st' instead of 'si' was most common, i. e. whenever 'si' appeared in the text, 'st' was detected instead.[8] This error can be reduced by training *Abbyy* not only on the letters 'st', but also on 'si'. Assuming that the error 'st' instead of 'si' would be reduced to zero, the greyscale scan would show the best result with only four errors per double page. The colour scan, on the other hand, still shows six errors. The preferred setting therefore is greyscale instead of colour. This analysis reveals that a different setting is better than the first test (table 1) suggested.[9]

Table 2. The results of the scans with the book scanner brought together in one chart with a closer look at the kind of mistakes - the best result turns out to be grey scale. It is highlighted in green. '→' stands for 'instead of'.

		e →c	u →n	st →si	st →il	ss →st	st →ss	st →ff	b →d	m →n .	l →!	s →j	s →f	b →h	t →f	i →l	c →:	e →o	e →h	L →T
colour	300			1							1	1	1	2	1					
black/ white	300	1		6						3						1	1	1	1	1
grey scale	300	1		5	1	1			1											

3.3 Method comparison *Abbyy* vs. Double Keying (second question)

The result is quite clear, in terms of both workload and errors: The automated conversion with *Abbyy* took 1 hour; 25 errors had to be corrected during post-processing. The double keying took a total of 6 hours and 89 errors had to be corrected. Of course, the errors are cleaned up at the end of the process, so that the files have almost zero errors after processing. The final result is therefore the same for both methods, except that the time required for double keying is six times longer than for automated conversion. Therefore, it is preferable to use automated conversion in any case.

4 Conclusion

It was shown that the choice of camera/scanner and the selected settings are important. It is always worthwhile to carry out test scans in advance and train with *Abbyy* on a trial basis. This is the only way to find out which settings are best suited and which letter combinations need to be trained to improve the OCR. In addition, a modest quality of the books that should be digitised is very important: water stains, domed or damaged paper, etc. leads to more errors in the machine-readable text.[10]

[8] Holley (2009) also mentions this bi-gram mistake and suggests a confusion matrix. For details, see [3].

[9] Considering also the results of the copier, the best setting is full colour, 400 dpi of the copier. It produces only five mistakes of the 'st' instead of 'si' mistake and is therefore the preferred choice.

[10] To test the limits of the OCR software, we decided to convert a scan in a poor quality (the quality of the book itself as well as the scan). On 1 ½ pages we had over 700 errors and a workload of four hours, which is a result that is not acceptable. Nevertheless, it shows the importance of a modest source and a scan in an appropriate quality.

The analysis of the results has shown that not only the **number** of errors is important, but also the **type** of error. Frequent mistakes can often be avoided by a new and targeted training, which means the selection of the scan settings might be adjusted after a closer examination of the type of errors. In our case, we were able to reduce the number of errors from seven to four per double page by additionally analysing the type of error. Only through the additional analysis it became clear that a different scanner setting with an additional training provides the best results.

All results and best practices are available on *github*[11] in order to spare smaller projects the extensive familiarisation with the topic of scanning and OCR. The material can be used to trace a thoroughly tested workflow in which individual components can also be exchanged, e. g. *Abbyy* with a freely available open source software.

References

1. Deutsche Forschungsgemeinschaft. (2016). *DFG-Praxisregeln "Digitalisierung"*. Retrieved from http://www.dfg.de/formulare/12_151/
2. Federal Agencies Digital Guidelines Initiative. (2016). *Technical guidelines for digitizing cultural heritage materials. Creation of raster image files.* Retrieved from http://www.digitizationguidelines.gov/guidelines/
3. Holley, R. (2009). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine 15* (3/4).
4. Stollwerk, Ch. (2016): Machbarkeitsstudie zu Einsatzmöglichkeiten von OCR-Software im Bereich „Alter Drucke" zur Vorbereitung einer vollständigen Digitalisierung deutscher Druckerzeugnisse zwischen 1500 und 1930. *DARIAH-DE working papers.*
5. Blanke, T.; Bryant, M.; Hedges, M (2012): Ocropodium: open source OCR for small-scale historical archives. *Journal of Information Science 38 (I)*, 76-86.
6. Boenig, M., Würzner, K.-M., Binder, A. (2016). Über den Mehrwert der Vernetzung von OCR-Verfahren zur Erfassung von Texten des 17. Jahrhunderts. In *DHd 2016, Modellierung, Vernetzung, Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma,* (p. 103-108). Retrieved from http://dhd2016.de/boa.pdf

[11] https://github.com/Bockwinkel/OCR