THE ROUTLEDGE COMPANION TO LIBRARIES, ARCHIVES, AND THE DIGITAL HUMANITIES

Edited by Isabel Galina Russell and Glen Layne-Worthey

First published 2025

ISBN: 9781032356259 (hbk)
ISBN: 9781032356280 (pbk)
ISBN: 9781003327738 (ebk)

8

GETTING BACK IN THE FLOW

An Outline for a Semi-Automated Digitisation Workflow to Improve the Quality of Digital Collections

Mirjam Cuper

DOI: 10.4324/9781003327738-11

The funder for this chapter is Koninklijke Bibliotheek.

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

8

GETTING BACK IN THE FLOW

An Outline for a Semi-Automated Digitisation Workflow to Improve the Quality of Digital Collections

Mirjam Cuper

KB, the National Library of the Netherlands

Digital resources and (digital) humanities research

Since the start of the digital era, more and more data is available in digital form. A lot of this data is 'born digital', such as social media posts, websites, and e-books. But that is not the only material that is digitally available. Heritage institutions and archives are also participating in this digital era, by digitising their heritage collections and making them available online for the public. Many heritage institutions use large-scale or mass digitisation projects to achieve this. This chapter aims to provide guidelines for institutions that wish to improve their support to the research community.

To clearly define what we mean by mass digitisation or large-scale digitisation processes, we have adapted Gooding's definition.[1] However, we do not distinguish between mass digitisation and large-scale digitisation projects. Furthermore, we have shortened Gooding's definition to the following three criteria:

- There is a set of heritage materials from national libraries or archives that needs to be digitised.
- There is a searchable interface through which the digitised material is made available for discovery and viewing.
- The material is digitised with the use of time-saving methods for digitisation and metadata creation (scanning, creating machine-readable texts through Optical Character Recognition (OCR) and automated layout recognition).

Digital heritage collections created by these large-scale digitisation processes are popular among humanities researchers. Thanks to the online accessibility of collections, research can be executed without having to travel to the archives where the original material is stored. This leads to a higher accessibility of these collections for researchers from all over the world. A lot of the digitised material is also transformed, either manually or automatically, to

DOI: 10.4324/9781003327738-11

computer-readable texts. These computer-readable texts create a whole new range of possibilities for computational research and methods to analyse the data, such as natural language processing and other machine-learning techniques.

A benefit of mass digitisation is the amount of material that is made available. However, the quality of the mass-digitised material varies greatly and is sometimes very low. An even larger problem is that most of the time, the quality of the digitised material is unknown. Due to this variability in quality and lack of information about the quality, problems can arise for researchers who want to perform research on this material. These problems can arise at various stages of the research.

Status quo: current common large-scale digitisation workflows

Large-scale digitisation workflows tend to have a similar structure. The workflow starts with the physical source material being digitised. The source material is scanned to create a digital copy of the material in the form of a digital image. This image is usually accompanied by either manually or machine-generated metadata. The next step is to enhance the digital image with a computer-readable version of the material's text content (if any). This is done using OCR software, which first uses layout recognition to divide the content of the image into text blocks, such as paragraphs, and then uses OCR to retrieve the digitised text. When this is done, the digitised material (often both image and computer-readable text) is published online. These kinds of large-scale projects are ideal for digitising lots of material relatively quickly. The process is highly automated, which reduces costs and time. However, within most large-scale digitisation projects, there is little to no quality control on the segmentation and quality of the digitised texts. Some organisations take samples to perform quality checks, but these are usually not extensive.

Generally, knowledge about digitisation processes is diffuse, and expertise regarding digitisation approaches, improvements, and challenges is not always exchanged between cultural heritage institutions. New technologies and research results do not always reach these institutions, or they do not know how to implement them. This leads to organisations developing solutions for problems that may have already been tackled by others. For example, when a researcher re-OCRs or manually improves the quality of digitised texts needed for research, this improved material is often not reingested into the institution's collection.

Case study: mass digitisation in the National Library of the Netherlands

At the KB, the National Library of the Netherlands, large-scale digitisation is used to create the digital heritage collections that are accessible via the online platform Delpher[2] and through an Application Programming Interface (API). The KB provides full-text Dutch-language digitised historical newspapers, books, journals, and copy sheets for radio news broadcasts.

The KB uses the same workflow as described above. Scanning the image, retrieving layout recognition, and OCR are outsourced. When the digitised material is returned from the outsourced producer, some automatic controls are performed. Batches of material are checked for, among others, structure, completeness, file formats, XML validation, photographic scan quality, page numbering, and correctness of metadata. Apart from these checks, random manual samples are taken from these batches to determine the quality of the digitised texts. The outcomes of these

quality controls are generalised to the complete batch and then checked against a predetermined cut-off point. When the quality of the batch is below this cut-off point, the whole batch is sent back to the producer with the request to re-digitise the material. Otherwise, the digital content is published on both Delpher and the API.

Every item receives an individual persistent identifier. Searching on both Delpher and the API is possible through metadata and full-text search (with the use of the OCR-ed text). The digitised material also contains layout information, which is used to locate search terms on the scanned image. On Delpher, these terms are highlighted on the image when viewed through the document viewer.

Mass-digitised collections and research problems

Kemman et al.[3] combined several studies about research phases into one representation of a research process within textual collections. They distinguished four main phases—discovery, selection, analysis, and dissemination—and several sub-phases. A schematic representation is shown in Figure 8.1. We will use this process as a guide to highlight some of the challenges that can arise when researchers want to use digitised heritage collections.

When searching through digitised heritage collections, there is always a chance of bias due to the fact that not all sources are digitised. In this chapter, we will take this bias for granted and will focus solely on problems that can occur with badly digitised material.

The *discovery phase* is generally the first moment a researcher interacts with a digitised collection. Once a research idea has been formed, data must be collected to perform the research. When a suitable collection is found, the researcher will start exploring and searching through the collection to gather the desired material.

Most heritage institutions that use mass digitisation have a search interface with which researchers can explore sources based on search queries. These queries can be based on metadata, full-text search, or both. It is evident that the results from a query are directly related to the quality of the metadata and digitised text. The lower the quality of the digitised text, the lower the chance that it can be found through these queries.[4] This introduces a bias in the search results: only material of high enough quality can be found and is accessible. Bazzo et al.[5] showed that problems with information retrieval can already occur at a word error rate of 5%.

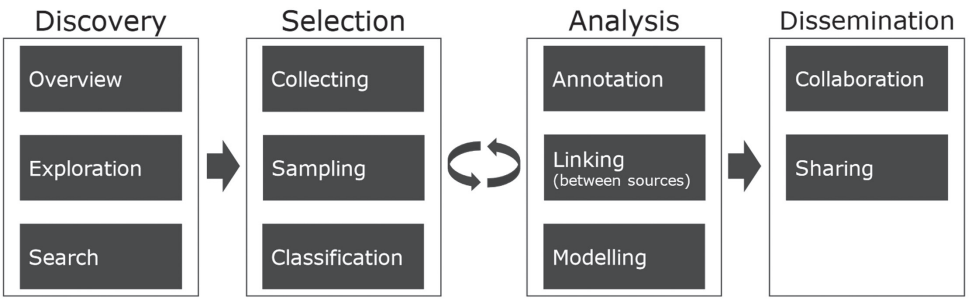

Figure 8.1 Schematic representation of the research process with textual collections.[a]

a Ibid.

Example 1: the influence of OCR errors on search results

To demonstrate the problems that can arise with full-text search on material with OCR problems, we did some small experiments.

Amsterdam versus Amfterdam

In the 17th century, it was common practice to use a 'long s' instead of a 'normal s' in certain word positions. However, a common mistake in OCR software is that this 'long s' is interpreted as an 'f'. When a researcher who is interested in news about Amsterdam in the 17th century would only search on 'Amsterdam', they will miss 47% of relevant articles, as shown in Figure 8.2.

Low quality, high impact

Digitised Dutch 17th-century newspapers often contain a lot of OCR errors. Therefore, a part of the corpus was manually corrected (see example 2 below, on the importance of persistent identifiers). We tested the difference in search results between the original OCR and the corrected texts. As shown in Figure 8.3, low-quality OCR can have a high negative impact on search results.

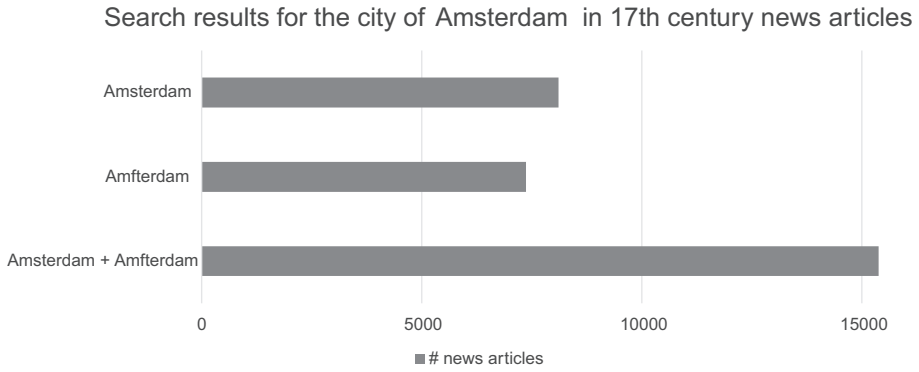

Figure 8.2 Appearance of articles about Amsterdam with different search queries.

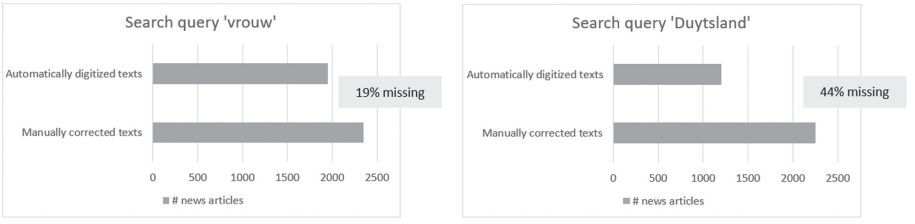

Figure 8.3 Difference in search results for original texts with OCR errors and with manually corrected texts.

In the *selection* phase, the researcher collects the materials relevant to their research. This can be done either manually through the search interface or automatically through an Application Programming Interface (API). The selection phase usually also depends on metadata and full-text data searches and has therefore potentially the same problems as the discovery phase.

Keeping the drawbacks of large-scale digitisation in mind, it is likely that Digital Humanities researchers planning to use computational methods would prefer to select materials based on the quality of digitisation. However, most institutions do not offer an indication of quality, which means that researchers are forced to either include all the data or measure the quality of the selected material themselves.

During the *analysis phase*, computational methods can be used for analysing large batches of material. Therefore, it is important that the included material is of high enough quality to support those methods. When the selected material contains a multitude of errors, this can negatively influence the outcomes of the algorithms and machine-learning models. A common quote among data and machine-learning specialists is 'garbage in, garbage out'. When training a model on material that is badly digitised, the result is likely to be a biased and unreliable model.

Research has been conducted to determine the influence of OCR quality on various computational tasks, such as Natural Language Processing, leading to a recommended OCR quality of at least 80%, and preferably 90% or higher.[6] In addition to OCR quality, the quality of the layout recognition is also important for researchers, particularly in the linguistic domain.[7]

The last phase is the *dissemination phase*. Here, the researcher shares the results of their work with others via publications or collaborations. The previously mentioned problems can lead to uncertainty about the research results: is the measured effect real or an effect caused by retrieval or quality bias?[8]

There is increasing awareness of the need to reuse data and models developed in research. One of the results of this awareness was the creation of the FAIR principles, which aim to support the reuse of scholarly data and tools. FAIR stands for: Findable, Accessible, Interoperable, and Reproducible.[9] As more institutions and publishers encourage researchers to publish as FAIR as possible,[10] it is important that the data management of institutions allows them to do so.

An important part of the FAIR principles is creating research that is reproducible. Therefore, the data used should be findable by and accessible to others after the research is published. To accomplish this, the FAIR principles emphasise the use of unique and persistent identifiers. In the realm of digitised collections, every item should get its own persistent identifier. This is especially important for research from which the original data cannot be shared due to copyright issues. With the persistent identifier, other researchers can ensure that they receive the same data directly from the originating institution.

Once in a while, there are initiatives to improve (parts of) digitised collections. When these improvements are being reintegrated into the digital collections, a choice should be made about adding a new persistent identifier along with these improvements. This is not an obvious choice as there are discussions on whether a persistent identifier belongs purely to the scan, or also to the processed digitised material. However, if these improvements are published without a new identifier or a form of version control, older research may be seemingly invalidated, as illustrated in example 2. The idea of what a persistent identifier is

can differ from institution to institution. Transparency about the approach in cases of data improvement can overcome incorrect assumptions.

Example 2: The importance of persistent identifiers or version control for research

In 2014, the Dutch Language Institute (INT), the National Library of the Netherlands (KB), and the Meertens Institute started a collaboration to manually correct around 120,000 newspaper articles published between 1618 and 1700. More than 200 volunteers participated. The improved texts were added to Delpher and the API, replacing the original, low-quality, OCR input.

This initiative was a great way to improve a poor part of the collection, which was hard to improve automatically. However, the original OCR was replaced with the new texts without changing the identifier or providing a form of version control. By keeping the same identifier and replacing the texts, the original OCR-ed texts are not retraceable for researchers.

This can cause confusion when, for example, a researcher reported about the low quality of the digitised texts. Research based on these old texts is not reproducible, since the new texts bear the same identifier as the old one. Figure 8.4 shows an example of different outcomes based on the original and improved texts with the same identifier. It displays two word clouds, showing the 20 most common words among news articles of the 'Oprechtse Haerlemse Courant' (a Dutch newspaper) from 1672. The left word cloud is based on the original OCR, the right word cloud is based on the manually corrected texts.

Figure 8.4 Example of word clouds based on the original OCR (left) and the manually corrected texts (right).

In an ideal world, every written item that ever existed would be digitised at a high-quality level. However, in the real world, there are various reasons why this is not possible. Firstly, not all historical material is preserved, a lot of material has been lost or destroyed. Secondly, some preserved material is so damaged or fragile that it is not possible to manually read the texts, let alone digitise them.

Even for material that is well-suited for digitisation, or has already been digitised, it is a challenge to ensure a high quality for the digitised material. If an optimal digitisation improvement workflow were to exist today, it would still take a lot of time to check and

improve the quality of all the digitised material currently available. In 2022, the KB alone has over 130 million digitised items.

Therefore, considering the researchers' need, the first and most important steps would be to provide them with transparency about the digitisation process and how improvements are implemented, transparency about the quality of the digitised materials, and the use of persistent identifiers or version control.

Getting back in the flow: theoretical outline of an optimised digitisation flow

We now expand on the modular, adaptive plug-in workflow as proposed by Cuper and D'Huys[11] and how this contributes to Digital Humanities research. The goal of this workflow is to achieve the best possible way to digitise heritage with the highest possible quality.

A schematic overview of the workflow is shown in Figure 8.5. The workflow is based on the processes that are used in current digitisation workflows and extended with various modules to provide transparency about quality and to improve the digitised material.

The main value of this optimised digitisation workflow lies in its adaptive, iterative approach. Whenever there is a (primary) solution for one of the modules, it can be inserted into the workflow. Modules that are inserted in the workflow can be upgraded to newer versions when available. This way, modules can be implemented, even when not every solution is available or perfect. To become feasible for large-scale digitisation projects, the workflow should ultimately be as automated as possible, only using human input when necessary.

As transparency about the digitisation process is important, all digitisation and improvement steps should be documented. This documentation should not only include information about the used software and undergone improvements but should also contain information about the protocols regarding persistent identifiers and version control. Using this

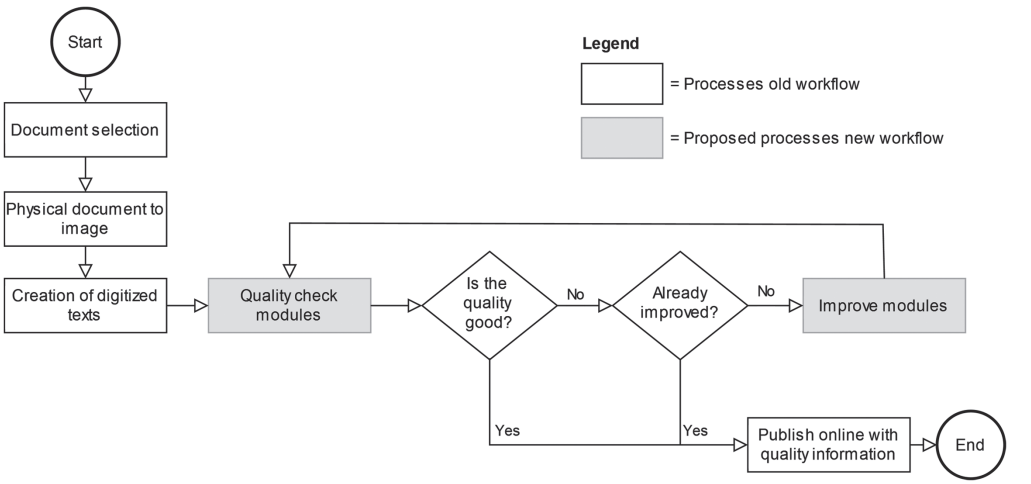

Figure 8.5 Simplified example of the modules in the proposed workflow, adapted from Cuper and D'Huys.[a]

a Ibid. Rights to the material are controlled by a third party; used by permission.

documentation, researchers should be able to revert to a previous version of the digitised material. Furthermore, the documentation should contain information about the quality. The documentation should be both human and machine-readable and freely accessible, so it can easily be used in the selection phase of research.

An important aspect to ensure that the workflow is successful is collaboration with other institutions and researchers. Currently, knowledge about digitisation best practices is diffuse and too rarely shared. This is a shame because it leads to reinventing the wheel, which is a waste of time and resources. Sharing ideas and solutions accelerates improvements in the workflow. The workflow is not only suitable for new material but can also be used to provide transparency about existing digital collections and to support their improvement. The workflow contains quality checks on an item level which can be used for targeted, case-by-case improvements.

But how will the Digital Humanities benefit from this workflow? As mentioned in the previous section, one of the problems is the lack of transparency about the process of digitisation and the quality of the digitised material. To solve this, the proposed workflow incorporates documentation to record all the steps that the material has undergone and to provide an indication of quality.

There are several benefits when institutions provide an indication of the quality of the individual items. Firstly, researchers can use this indication while pre-selecting material for their research, thereby providing an option to reduce bias due to OCR errors. Secondly, a uniform quality indication leads to easier comparisons between research on material from the same institution. Thirdly, researchers can document the quality-based selection criteria in their publication, leading to increased reproducibility. Last but not least, researchers can use this indication instead of running their own quality checks. This not only saves time for the researcher but is also a more sustainable solution as fewer computational resources are needed.

Improved OCR leads to better findability of material through both the search interface and API. This reduces bias in search results. When the availability of high-quality material increases, more reliable data can be used to train and test machine-learning models, which improves their robustness.

To publish research according to the FAIR principles, researchers should make sure that their data is findable and that their research is reproducible. By using new persistent identifiers or version control after every improvement, older versions of material can always be found, thereby enabling researchers to adhere to these principles.

Explaining the modules

It is time to take a closer look at the rationale behind the modules and to provide guidelines for implementing them. The modules can be divided into two parts: *quality check* modules and *quality improvement* modules.

When material is digitised, there are various steps to get from the source material to the machine-readable format. The way in which a former step is performed can directly influence the next steps. There are various elements for which quality should be measured:

- the metadata;
- the content of the image;
- the recognised layout;
- the digitised text.

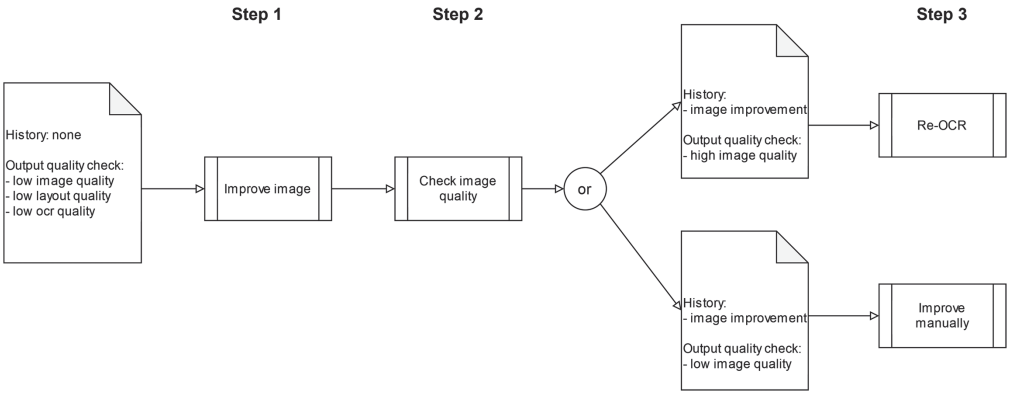

Figure 8.6 Simplified example of targeted improvements based on quality checks.

Although we see the importance of metadata, it has its own challenges and requires its own expertise. Therefore, we decided to keep metadata out of scope for the remainder of this chapter.

Apart from the metadata, the various elements have a sequential order in terms of influence. When the quality of the content of the image is low, this likely influences the outcome of the layout recognition and the quality of the OCR. Sometimes, the image quality is good enough, but the layout recognition is still low. This can happen, for example, with periodicals or newspapers, where the layout is variable, with changes in font size and text boxes of different sizes. Even when the separate words are recognised correctly, this can still lead to low-quality OCR-ed text, for example, when the texts of separate articles are combined into one due to layout recognition errors. With this in mind, targeted improvements based on the outcomes of the various quality checks would be preferable. An example of this is shown in Figure 8.6.

However, at the moment, there is not a good automated solution for all the quality modules. This emphasises the need for a plug-in workflow, where institutions can start with a smaller selection of models. This way, libraries and archives don't have to wait until all the quality checks are done and automatic improvements are in place, but they can already start improving their transparency and digitised material based on the knowledge of what they *can* use at this moment.

Quality check modules

For targeted improvement, knowledge about the quality of various aspects of the digitised content is essential. Therefore, we propose various quality check modules.

Image quality check: There are various reasons for the content of an image to be of low quality. For example, skewed or bent pages, ink bleed-through, or a damaged page due to mould, stains, or rips. A lot of algorithms are already available to detect skewness and noise in documents.[12] Furthermore, it could be interesting to experiment with a machine-learning model that can distinguish images with high-quality OCR output from images where manual intervention is probably needed.

Layout recognition quality: A current review showed that there are not yet standards or systematic methods to measure the quality of layout recognition.[13] Experimenting with 'intersection over union' methods to detect overlapping segments can be conducted to see if they provide useful information about layout quality.

Optical character recognition quality: Various approaches have been developed to detect errors in (digitised) texts without comparing them to correct versions of these texts. These approaches can be roughly split into two kinds: isolated-word approaches and context-dependent approaches. With the isolated-word approach, the word is examined apart from its context, whereas the context-dependent approach takes the context into account.

This difference can be illustrated by an example. Imagine that the OCR output returns the sentence: 'the neighbour mows the glass'. An isolated word approach looks only at the word 'glass', which is an existing word in English. So it would mark it as a non-error. The context-based approach however uses other words from the sentence, such as 'mows'. As this is not a common context in which the word 'glass' is used, it detects the word as an error.

One commonly used isolated-word approach to detect errors is the lexical approach. Here, every word in the text is checked against a lexicon to see if this word exists in the given language. Another example of an isolated-word method is garbage detection. With this method, each word is checked against certain rules, such as whether a word contains non-letter characters, or the same character repeatedly in a row, to see if the word contains 'garbage' or not.[14] For context-dependent approaches, various methods are used. Schaefer and Neudecker use word-embeddings to detect OCR errors, while other studies use n-grams.[15]

In addition to these examples, many more approaches are available. An extensive overview of commonly used methods has been documented by Nguyen.[16] As all approaches have their drawbacks in terms of accuracy, initiatives have arisen in which multiple methods are combined to provide a more accurate indication of quality.[17]

Quality improvement modules

Once the quality of the digitised material is known, it can be divided into material that is good to publish online as it is, and material that could benefit from improvement. There are several ways to try and improve this low-quality material at various levels.

Improving the image: There are multiple studies where an attempt is made to enhance an image before (re)-OCRing it. A lot of these techniques are commonly used in image processing, such as binarisation, skew-correction, noise removal, and contrast adjustment. These methods are also applicable to historical documents, such as shown in various studies. Yahya et al.[18] performed a review of three enhancement methods on historical pages with a damaged background. Other studies reported effective methods to restore ink bleed-through.[19]

Re-scan source material: In some cases, it can be beneficial to re-scan the source material. When scans were made years ago, stored at low resolution or on low-quality storage, it can be useful to re-scan the source material with the newest techniques to obtain better results. The same applies to damaged or degraded source material for which a higher quality version is available.

Due to the mass digitisation processes, sometimes images are skewed or pages are folded or bent, where with a little bit more time the material can be scanned without these issues. In such cases, targeted re-scanning can also be worth the effort.

Re-OCR with different software: Over the past years, various (open source) OCR tools have been developed, and existing packages are constantly updated. As some of these tools claim to be optimised for historical documents, it can be worthwhile to experiment with whether certain tools may perform better for material with specific characteristics. It may also be worthwhile to test other techniques on printed material, such as handwritten text recognition (HTR), as shown by Romein et al.[20] Also, comparing and combining the most common suggestions from various OCR outputs can sometimes increase the quality. Furthermore, some (parts of) material, such as advertisements, tables, and music scores are extremely hard to digitise correctly with current OCR software, so it could be beneficial to invest time and resources in training tools specifically for that purpose.

Layout recognition: Just as with OCR, it may be beneficial to try different layout recognition software on the material to see if this increases the quality of the layout recognition. Comparing and combining the outcome of various algorithms may also improve quality.

Automatic post-processing: (Semi-)automated post-processing methods have been researched extensively. As with OCR quality methods, these approaches can be divided into isolated word approaches and context-dependent approaches. Some examples of isolated word approaches are merging OCR outputs, lexical approaches, and topic-based language models. Context-based approaches are usually based on machine-learning techniques. As the number of methods is too long to extensively cover in this chapter, we refer to Nguyen,[21] who has described most of these methods and techniques in depth.

Manual post-processing: In some cases, manual corrections are unavoidable, for example, when the source material is badly damaged or when ink has bled. In such cases, automated processes will likely not provide the desired results. Human input can then be used to improve the digitised material as much as possible. This can be done for both layout recognition and the text itself. Manual improvements can be done by the institution's employees, but to process a lot of material, the input from volunteers is recommended. Various crowdsourcing approaches have already been carried out to improve the quality of collections. Two examples are the crowdsourcing project led by the Meertens Institute to improve the quality of digitised 17th-century newspapers (see example 2 for a description) and the Trove project. In 2008, the National Library of Australia presented an interface called Trove, in which users could indicate erroneous digitised texts and correct them. After six months, 2 million lines of text in 100,000 articles were corrected by around 1300 volunteers. As a precaution, a roll back option was implemented so text could be restored in case of vandalism of text. However, after six months, no signs of vandalism were detected.[22] Up to this day, Trove is still online and the manual correction options are still available.[23] Both examples show that users of digitised collections want to engage and participate in improving their quality.

Implementation module: As researchers work with digitised collections, the quality of their selected materials can sometimes be too low for the desired research methods. When this happens, some researchers decide to improve it, either manually or automatically. An example is the research described by Romein et al.,[24] in which the incomprehensible digitised texts of ordinances were improved. However, most of the time when these kinds

of improvements are done, they do not end up in the heritage institutions' collections as improved versions. One of the reasons is that in a lot of institutions, protocols are missing for reingesting material from other sources. It is a shame that institutions often miss such opportunities, in which improvements in digitised material are so readily available. To overcome this problem and to avoid repeated efforts, institutions could create protocols and guidelines for researchers on how to store and share the data in such a way that it can easily be ingested in the collections.

How to get there? An iterative approach

Unfortunately, there is no ready-made solution available that can be implemented in our proposed workflow and that adheres to the above-mentioned requirements for research. However, libraries and archives do not have to wait until a complete solution is created. They can start by implementing small elements and modules and improving these through an iterative process. For example, we introduced methods to give a comprehensive overview of the OCR quality of digitised texts. These can be used to distinguish between texts that are suitable for improvement and those which are not. In the latter situation, the existing solutions for post-processing can be used, even though these have imperfections. Manually correcting these texts is another alternative.

Case study of the KB

The KB has started a project to improve the transparency and quality of its digitised collections. This project is executed as a close collaboration between the research department, the digitisation department, collection experts, and content providers.

The first goal was to find measures to provide an indication of OCR quality without having a correct version of the text to compare it with. Therefore, various experiments were performed on candidate measures to check their usefulness and reliability.[25] Furthermore, the research department performs and supports ongoing research from both internal and external researchers about the quality and improvement of digitised collections.

As most heritage institutions run into the same issues, collaboration is recommended to share knowledge and experience of digitisation projects, thereby focusing on exchanging ideas and reutilising solutions. It is also strongly advised to openly publish newly created solutions for modules and to share the code. This enables other institutions to use this information, saving time and resources, and consequently also increasing sustainability. Furthermore, collaboration leads to the acceleration of improvements.

Additionally, it is encouraged to reach out to Digital Humanities researchers and data scientists to increase knowledge and technical expertise. Many (partial) solutions can potentially be implemented. There is no need to duplicate work, cope with potential shortages in technical skills, and lose scarce time when the work has already been done. Cooperating with researchers also enables institutions to reingest material that researchers have improved.

Communication with researchers who use the collections is essential; this is the key to targeted improvements of the collections. This chapter mentioned known issues in Digital Humanities research; however, it is still important to contact specific research groups and collect their wishes and priorities. With different types of research, researchers' needs vary.

Many automated approaches require training data sets. These can be manually corrected versions of texts, scans that are labelled as readable or unreadable, manually corrected segmentation, and more. These training sets are used to train machine-learning models to perform such tasks automatically. Building such training corpora can be done while implementing and testing various modules. But these kinds of activities are also highly suitable for crowdsourcing tasks, where volunteers help to build the desired sets.

As machine-learning is rapidly evolving, it is good practice to stay informed about new research and techniques. It is highly likely some challenges do not have solutions yet but will soon. It is also advised to explore beyond the field of digital heritage and Digital Humanities. Techniques commonly used in other domains, like image processing, speech recognition, and automatic translations, may be transferable solutions for challenges in digitising heritage.

Although there is a substantial initial time investment due to the manual work involved in creating and updating the various modules, it should be possible to add more automated steps in time. This will increasingly automate the workflow, which would ultimately result in a semi-automated workflow that is rarely manually interfered with. The eventual goal is to create a transparent, digitised collection with high-quality content.

Improve your collections, improve Digital Humanities research!

How to get there? Quick start:

- Don't wait, just start now!
- Spread the word and collaborate.
- Make use of what is already there.
- Don't be afraid to use crowdsourcing.
- Implement, test, improve, and repeat.
- Look beyond your own borders.

Notes

1 Paul Gooding, *Historic Newspapers in the Digital Age* (London and New York: Routledge, 2017), 4–5.
2 Koninklijke Bibliotheek, "Delpher," www.delpher.nl/
3 Max Kemman et al., "User Needs for a Text Suite for Advanced Digital Research," (Dialogic, 2022).
4 Myriam Traub et al., "Impact of Crowdsourcing OCR Improvements," *JCDL '18: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (2018).
5 Guilherme Bazzo et al., "Assessing the Impact of OCR Errors in Information Retrieval," *Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science (Springer)*, no. 12036 (2020).
6 Daniel van Strien et al., "Assessing the Impact of OCR Quality on Downstream NLP Tasks," *International Conference on Agents and Artificial Intelligence (ICAART 2020)* (2020).
7 Clemens Neudecker et al., "A Survey of OCR Evaluation Tools and Metrics," *HIP '21: The 6th International Workshop on Historical Document Imaging and Processing* (2021).

8 Myriam Traub, Jacco Van Ossenbruggen, and Lynda Hardman, "Impact Analysis of OCR Quality on Research Tasks in Digital Archives," *19th International Conference on Theory and Practice of Digital Libraries* (2015).
9 Mark D. Wilkinson et al., "The Fair Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data*, no. 3 (2016).
10 *Sorbonne Declaration on Research Data Rights.*
11 Mirjam Cuper and Sarah D'Huys, "Optimizing the Digitization Workflow of Heritage Institutions to Increase Quality of Digitized Texts," *Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries (TPDL 2022)*. http://dx.doi.org/10.1007/978-3-031-16802-4_51
12 Arwa AL-Khatatneh, Sakinah Ali Pitchay, and Musab Al-qudah, "A Review of Skew Detection Techniques for Document," *17th UKSIM-AMSS International Conference on Modelling and Simulation* (2015); Ajay Kumar Boyat and Brijendra Kumar Joshi, "A Review Paper: Noise Models in Digital Image Processing," *Signal & Image Processing: An International Journal* (2015).
13 Neudecker et al., "A Survey of OCR Evaluation Tools and Metrics."
14 Kazem Taghva et al., "Automatic Removal of "Garbage Strings" in OCR Text: An Implementation," (2001); Richard Wudtke, Christoph Ringlstetter, and Klaus Schulz, "Recognizing Garbage in OCR Output on Historical Documents," *Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data* (2011).
15 S. El Atawy and A. Abd ElGhany, "Automatic Spelling Correction Based on N-Gram Model," *International Journal of Computer Applications* 182, no. 11 (2018); Robin Schaefer and Clemens Neudecker, "A Two-Step Approach for Automatic OCR Post-Correction," *Proceedings of LaTeCH-CLfL* (2020).
16 Thi Tuyet Hai Nguyen et al., "Survey of Post-OCR Processing Approaches," *ACM Computing Surveys* (2022).
17 Mirjam Cuper, "Examining a Multi Layered Approach for Classification of OCR Quality without Ground Truth," *DH Benelux Journal* 4: The Humanities in a Digital World (2022); Pit Schneider and Yves Maurer, "Rerunning OCR: A Machine Learning Approach to Quality Assessment," *Journal of Data Mining and Digital Humanities* (2021).
18 Sitti Rachmawati Yahya et al., "Review on Image Enhancement Methods of Old Manuscript with the Damaged Background," *009 International Conference on Electrical Engineering and Informatics* (2009).
19 Drira Fadoua, Frank Le Bourgeois, and Hubert Emptoz, "Restoring Ink Bleed-through Degraded Document Images Using a Recursive Unsupervised Classification Technique," Bunke, H., Spitz, A.L. (eds) *Document Analysis Systems VII. DAS 2006. Lecture Notes in Computer Science (Springer)*, vol 3872 (2006); Anna Tonazzini, Emanuele Salerno, and Luigi Bedini, "Fast Correction of Bleed-through Distortion in Grayscale Documents by a Blind Source Separation Technique," *IJDAR*, no. 10 (2007).
20 C. Annemieke Romein, Sara Veldhoen, and Michel de Gruijter, "The Datafication of Early Modern Ordinances," *DH Benelux Journal* 2: Digital Humanities in Society (2020).
21 Nguyen et al., "Survey of Post-OCR Processing Approaches."
22 Rose Holley, "Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers" (National Library of Australia, 2009).
23 "Trove," https://trove.nla.gov.au/
24 Romein, Veldhoen, and de Gruijter, "The Datafication of Early Modern Ordinances."
25 Mirjam Cuper; C. van Dongen, T. Koster. (2023), "Unraveling Confidence: Examining Confidence Scores as Proxy for OCR Quality". In: Fink, G.A., Jain, R., Kise, K., Zanibbi, R. (eds) Document Analysis and Recognition – ICDAR 2023. ICDAR 2023. Lecture Notes in Computer Science, vol 14191. Springer, Cham. https://doi.org/10.1007/978-3-031-41734-4_7

References

AL-Khatatneh, Arwa, Sakinah Ali Pitchay, and Musab Al-qudah. "A Review of Skew Detection Techniques for Document." *17th UKSIM-AMSS International Conference on Modelling and Simulation* (2015).

Bazzo, Guilherme, Gustavo Lorentz, Danny Vargas, and Viviane Moreira. "Assessing the Impact of OCR Errors in Information Retrieval." *Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science (Springer)*, (2020): 102–109.

Boyat, Ajay Kumar, and Brijendra Kumar Joshi. "A Review Paper: Noise Models in Digital Image Processing." *Signal & Image Processing: An International Journal* 2, no. 6 (2015): 63–75.

Cuper, Mirjam. "Examining a Multi Layered Approach for Classification of OCR Quality without Ground Truth." *DH Benelux Journal* 4: The Humanities in a Digital World (2022): 43–59.

Cuper, Mirjam, and Sarah D'Huys. "Optimizing the Digitization Workflow of Heritage Institutions to Increase Quality of Digitized Texts." *Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries (TPDL 2022)* (2022): 490–94. http://dx.doi.org/10.1007/978-3-031-16802-4_51

Cuper, Mirjam., van Dongen, Corine, Koster, Tineke. (2023). "Unraveling Confidence: Examining Confidence Scores as Proxy for OCR Quality". In: Fink, Gernot A., Jain, Rajiv, Kise, Koichi, Zanibbi, Richard (eds) *Document Analysis and Recognition – ICDAR 2023. ICDAR 2023. Lecture Notes in Computer Science*, vol 14191. Cham: Springer. https://doi.org/10.1007/978-3-031-41734-4_7

El Atawy, S., and A. Abd ElGhany. "Automatic Spelling Correction Based on N-Gram Model." *International Journal of Computer Applications* 182, no. 11 (2018): 5–9.

Fadoua, Drira, Frank Le Bourgeois, and Hubert Emptoz. (2006). "Restoring Ink Bleed-through Degraded Document Images Using a Recursive Unsupervised Classification Technique." In: Bunke, Horst, Spitz, A. Lawrence (eds) *Document Analysis Systems VII. DAS 2006. Lecture Notes in Computer Science*, vol. 3872, pp. 38–49. Cham: Springer.

Gooding, Paul. *Historic Newspapers in the Digital Age*. London and New York: Routledge, 2017.

Holley, Rose. "*Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers*." National Library of Australia, 2009.

Kemman, Max, Nick Jelicic, Guido de Moor, Marenne Massop, and Tommy van der Vorst. "*User Needs for a Text Suite for Advanced Digital Research*." Dialogic, 2022.

Koninklijke Bibliotheek. "Delpher." www.delpher.nl/.

Neudecker, Clemens, Konstantin Baierer, Mike Gerber, Clausner Christian, Antonacopoulos Apostolos, and Stefan Pletschacher. "A Survey of OCR Evaluation Tools and Metrics." *HIP '21: The 6th International Workshop on Historical Document Imaging and Processing* (2021): 13–18.

Nguyen, Thi Tuyet Hai, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. "Survey of Post-OCR Processing Approaches." *ACM Computing Surveys* 54, no. 6 (2022): 1–37.

Romein, C. Annemieke, Sara Veldhoen, and Michel de Gruijter. "The Datafication of Early Modern Ordinances." *DH Benelux Journal* 2: Digital Humanities in Society (2020).

Schaefer, Robin, and Clemens Neudecker. "A Two-Step Approach for Automatic OCR Post-Correction." *Proceedings of LaTeCH-CLfL* (2020): 52–57.

Schneider, Pit, and Yves Maurer. "Rerunning OCR: A Machine Learning Approach to Quality Assessment." *Journal of Data Mining and Digital Humanities* 2 (2021).

Sorbonne Declaration on Research Data Rights, January 27th 2020, https://sorbonnedatadeclaration.ent.upmc.fr/data-Sorbonne-declaration.pdf

Strien, Daniel van, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara Mcgillivray, and Giovanni Colavizza. "Assessing the Impact of OCR Quality on Downstream NLP Tasks." *International Conference on Agents and Artificial Intelligence (ICAART 2020)* (2020).

Taghva, Kazem, Tom Nartker, Allen Condit, and Julie Borsack. "Automatic Removal of "Garbage Strings" in OCR Text: An Implementation." (2001).

Tonazzini, Anna, Emanuele Salerno, and Luigi Bedini. "Fast Correction of Bleed-through Distortion in Grayscale Documents by a Blind Source Separation Technique." *IJDAR* 10 (2007): 17–25.

Traub, Myriam, Thaer Samar, Jacco van Ossenbruggen, and Lynda Hardman. "Impact of Crowdsourcing OCR Improvements." *JCDL '18: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (2018): 29–36.

Traub, Myriam, Jacco Van Ossenbruggen, and Lynda Hardman. "Impact Analysis of OCR Quality on Research Tasks in Digital Archives." *19th International Conference on Theory and Practice of Digital Libraries* (2015).

Wilkinson, Mark, Michel Dumontier, IJsbrand Aalbersberg, and et al. "The Fair Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data*, 3 (2016).

Wudtke, Richard, Christoph Ringlstetter, and Klaus Schulz. "Recognizing Garbage in OCR Output on Historical Documents." *Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data* (2011): 1–8.

Yahya, Sitti Rachmawati, Siti Norul Huda Sheikh Abdullah, Khairuddin Omar, Mohamad Shanudin Zakaria, and Choong-Yeun Liong. "Review on Image Enhancement Methods of Old Manuscript with the Damaged Background." *International Conference on Electrical Engineering and Informatics* (2009): 62–67.