

White Paper



ABBYY® Historic OCR

The use of Gothic OCR in processing historical documents

Michael Fuchs, ABBYY Europe (München)

Historical documents in a digital world

Long ago, computers, the Internet and storing information became the norm. Libraries, as the archetypal preservers of printed documents, can no longer hide from the global whirl of digitization. As a result, work began several years ago on setting up digital libraries, such as the German Digital Library and the “europeana” in response to the need to have information available everywhere and accessible at any time.

The processes by which documents that were created digitally (digital-born files) are prepared for electronic access and Internet research differ fundamentally from the preservation, searching and accessing of paper-based texts. Having these types of texts integrated into our digital lives that has made Optical Character Recognition (OCR) such an important technology over the past few years. Modern-day printed documents can be digitalised much more easily, quickly, affordably and reliably than being typed out by hand thanks to today's highly automated OCR software.

Universities, libraries and archives wishing to digitise their historical document collections face an even greater challenge, because reliably capturing and recognising historical documents is no simple task. Blackletter fonts such as Textualis, Rotunda or Gothic, which were used for hundreds of years and are an important part of European letterpress tradition and therefore also significant for our European cultural heritage, have proven troublesome for automated text recognition software. In addition to the quality of paper used in historical texts, there was no standardised typeface when printing these older documents, which is why letters differ greatly from one document to another. Language and spelling have also changed immeasurably over the centuries, meaning that modern-day dictionaries often cannot be used for automatic correction.

The following section describes the elements that play a part in capturing historical documents using state-of-the-art OCR technology.

How optical character recognition works and why old documents present a great challenge

Optical Character Recognition (OCR) is a highly complex series of mathematical and linguistic processes. The most important steps required to optimally digitise (historical) documents are as follows.

Step 1 – From image capturing to image optimisation

The first – and one the most important – prerequisites for reliably capturing historical documents using OCR technology is creating scans of the best-possible quality. Digital pixel images can be created from an original document (single pages or a book) and also from microfilms. It is often important to use, special scanners that are also capable of digitalising difficult, analogous material in high quality. Certain minimum requirements must be met so that the OCR process has sufficient raw data at its disposal: A document page with normal-sized text (ca. 12 point) must have a resolution of at least 300 dpi and preferably be in greyscale or colour. Simple bi-tonal scanning in black and white may result in pasting over of important document information, resulting in poor text recognition results. A simple bitmap scan in black and white, for instance, does not allow applying the image-optimization options described below.

Another pre-processing step needs to be conducted after the documents have been digitised and saved as electronic image files. For example, this could mean adjusting the resolution to at least 300 dpi for images whose resolution is too low, separating double pages, rotating pages by 90, 180 or 270 degrees (so that all the files are facing in the same direction) and cropping of images.

During image preparation, typical scan errors are corrected as much as possible by the automatic de-skewing of document pages, the straightening of text rows and a controlled automatic removal of dust and background noise. Nevertheless, particular care must be paid to these touch-ups, so that during the image optimization process, small marks such as serifs on letters and punctuation marks are not removed.

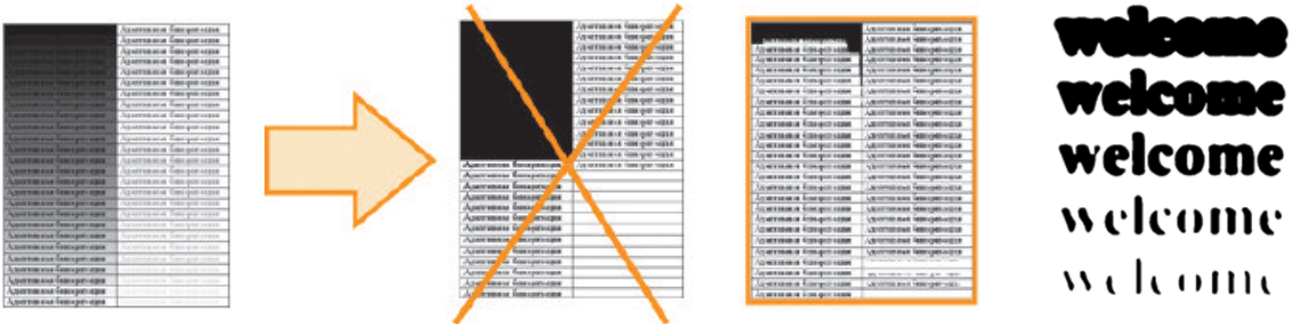
Most of the actual OCR processing always takes place at the level of a binary image – a black and white, bi-tonal raster image. To create this, ABBYY's OCR technology uses an intelligent background filtering with adaptive binarisation. This involves analysing the image using various algorithms and then gradually transferring it into a black and white image so that the text in the background is separated and the individual letters

are kept intact and are not too bold. As the figure below clearly shows, because of incorrect binarization valuable textual information can be lost during digitization, especially when dealing with documents with too little contrast or uneven text background.

unds. This is because the binarisation threshold has a direct influence on the quality of individual letters, thereby having an immediate effect on the overall quality of the OCR result.



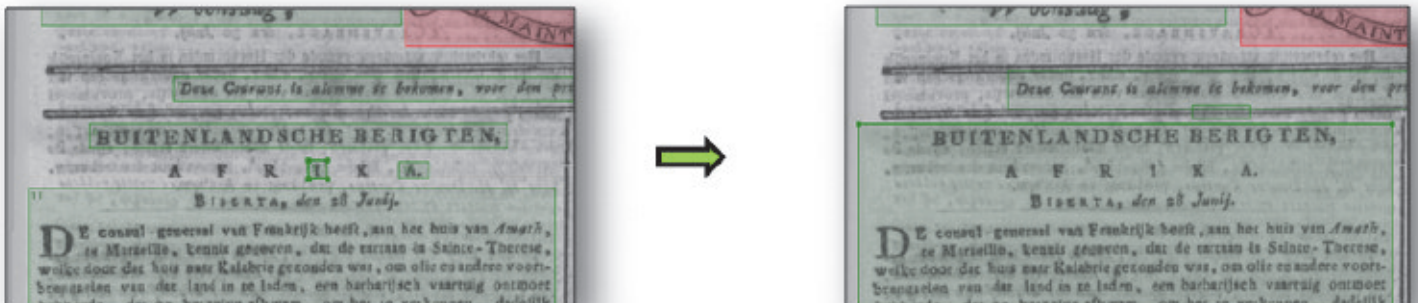
Valuable textual information for the OCR engine can be lost during digitisation due to incorrect binarisation.



Step 2 – Document analysis

A great deal of mathematics now comes into play for the layout analysis of the historical document, which is saved as an image file. Today's documents are usually very clearly and obviously structured and therefore rarely pose a problem for the automatic layout recognition software. Historical documents and newspapers, on the other hand, are usually structured more creatively and often have no standardised layouts. Identification of individual document elements, therefore, presents a great

challenge to OCR technology. In this step, the software must precisely capture the components that make up the image file. The OCR software then defines exactly from which areas, whether it is images, text, tables or similar. It is therefore important to use a correspondingly optimised analysis technology to decrypt historical documents. Only text areas that have been correctly and completely defined can produce a satisfactory final result in the OCR process.



If text blocks in a layout are recognised incorrectly, information will also be missing from the text recognition results. This figure shows the layout analysis improvements on old newspapers that ABBYY achieved during its IMPACT project.

Step 3 – Character recognition

After defining the various recognition areas, the actual recognition analysis begins, which painstakingly identifies each row and individually detects each letter and character. Recognition is based on classifying. ABBYY OCR technology uses various algorithms to recognise letters as correctly as possible and to verify the results, which are initially only suppositions, against character patterns saved in the system.

- A "Raster classifier" compares the pixel of each letter with the saved master characters
- A "Contour classifier" assesses the letter breaks using existing patterns
- A "Structure classifier" reduces the letters to simple vector frameworks, which are checked for their similarity to the saved master characters
- A "Specialist classifier" differentiates similar characters from one another, such as B from D or D from O

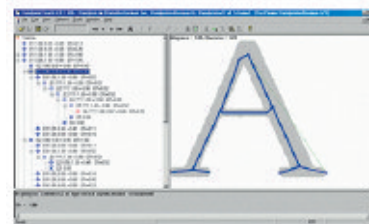
Raster classifier



Contour classifier



Structure classifier



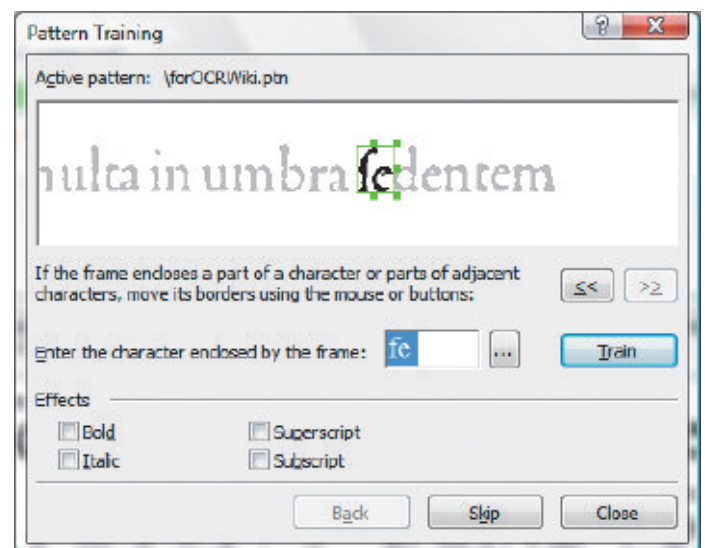
Feature differentiating classifier



ABBYY uses a number of special expert algorithms to recognise characters and achieve the best recognition results.

Recognising individual special characters

OCR technology uses typical patterns that are saved in the software (see above) to recognise standard characters. Many OCR solutions can also be gradually trained during the analysis process to see specific special characters in a document or type of document that are not included in the software's own stored patterns. This achieves very good results, particularly in recognising unusual characters or special ornate symbols. Individual patterns of this kind are always used in combination with the pre-saved classifiers, which increase the recognition rate, even for difficult historical texts, by providing more analytical information. From a technical standpoint, it must be noted that in order to train the system for characters, the resolution of processed images and template documents must be the same. It is important that the character height in pixels same, only then the pattern training can deliver improvements. The training only represents a further assessment feature for hard-to-recognise characters, one cannot expect "miracles" from this process.



Software can be trained to recognise special unknown characters and increase the recognition rate.

Language definition

In an additional step, the individual characters are put together to form whole words. For this process, it is important that the OCR software knows exactly in which language the document is written. Only by selecting this preference can the word results be checked for their relevance. The language definition provides a great deal of additional and useful information for analyzing the scanned text template. If an English document is processed, for example, the umlaut option, which is irrelevant for English, is omitted from the analytical process. ABBYY OCR technology also has the ability to identify documents that combine several languages.

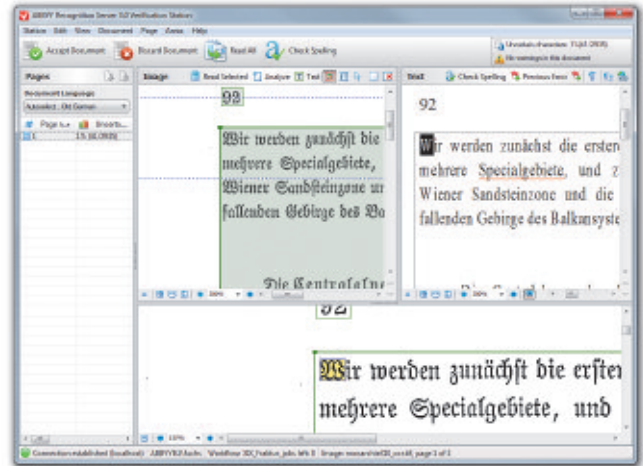
Dictionaries

To optimise text recognition at the word level ABBYY OCR products make use of the latest findings in linguistic technologies and also use morphological dictionaries as "analysis helpers" in the background. This is important so that even when the software only vaguely recognises characters, such as in a very poor-quality text document, the best possible decision can still be made. If an "ü", for example, cannot be recognised properly due to poor scanning quality, the software uses the information it has on hand to capture the text correctly. With the additional information "This is a German text" and a comparison with the dictionaries stored in the software ("Munich" is not in that dictionary, but "München" with a "ü" is), the text can be reliably recognised despite the document's poor quality.

What works well in more modern texts cannot easily be applied to the recognition of historical texts. As illustrated here, however, the standards are lacking for spelling, to which OCR technology could originate itself if electronic dictionaries from earlier times were available. However, in order to enable reliable recognition even for difficult historical documents, it is advisable to create new specific word lists or even entire dictionaries, especially for large volumes and special topic areas. With an OCR development tool kit like ABBYY FineReader Engine, the operations for setting up such a special dictionaries can be automated with the available existing interfaces (APIs).

Step 4 – Optional testing and post-correction by the user

After performing the previously described processing steps, there is the possibility for users or operators to intervene in the processing to improve the overall results. Manual post-correction can be useful, especially with layout analysis (definition of text, image or table blocks). A manual check can also be used to correct vaguely-recognised characters and words.



The Verification Station in Recognition Server 3.0 allows a recognised text to be checked before exporting it. Developers and users of FineReader Engine (SDK) also have Visual Components at their disposal.

Important note: In general, no manual correction is applied in mass digitization projects as it would take far too long to do so on hundreds of thousands of pages. In practice, automated, programmable post-correction based on the XML export is used. Nevertheless, in critical cases the correction station's result can be directly inspected and verified.

Step 5 – Synthesis and export documents formats

In practice, the ability to generate different output formats with various options can be very important, as one universal format is not usually suitable for all downstream work and archiving steps. It is precisely for this reason that ABBYY technology supports various export formats. Whether it is plain text, Office, XML format or searchable PDF (/A) documents, users can export the scanned text documents into the digital format they require.¹ Currently historical documents can be exported either as full-text, searchable digital books or PDF documents. An XML export with additional information (for example character coordinates) enables the results to be embedded in digital libraries and made available to a wider group of users. The digitised historical documents can then be found not only by searching for manually generated metadata, such as the author's name or the book title, but also as a full text documents using specific keywords.

¹ For more information about available export formats, please go to www.abbyy.de.

Development of ABBYY's Gothic OCR

ABBYY has been working for many years in close cooperation with libraries, universities, research institutions and other technology partners so that advanced OCR technology can automatically and reliably detect Gothic writing – a type of blackletter font that was typically used in texts published between 1800 and 1938. As an OCR specialist, ABBYY has provided its many years of experience in recognising and digitally preparing historical documents to many national and international projects for digitalising libraries. With METAe and IMPACT, ABBYY was also involved in two major research projects for the European Union for enhancing Gothic OCR. To further improve the technological capabilities for automated detection of blackletter fonts, the ABBYY research team created special "classifiers" or alphabets for the detailed analysis of Gothic characters. This involved a significant development effort. For each Gothic character, an average of about 2,500 variations were filed, a whole new alphabet pattern was created and 31,000 pages collected from different historical sources and tested in detail. Only by processing a variety of sample texts could the ABBYY OCR engine acquire the necessary fine-tuning to be able to recognise automatically the peculiarities of the Gothic alphabet, such as ligatures or combined lettering.

METAe

The European Union research project METADATA ENGINE (METAe) worked from 2000 to 2003 on automated recognition of layout and structure in books and magazines as well as the development of an OCR for Gothic fonts. In a research community of 14 partners from Europe and the United States, including leading national libraries, the project also explored possible uses for the systematic digitization of books and magazines in libraries and archives.² With FineReader XIX, ABBYY developed

a special Omnifont OCR solution for Gothic fonts for the METAe project. This font was widely used in many European countries and, in the case of the German language; it was still being used in 1941 in about 80 per cent of all printed documents. The Omnifont OCR ABBYY FineReader XIX solution can recognise Gothic fonts without prior training. ABBYY created five historical dictionaries for this, which reflect the historic state of spelling in English, French, German, Italian and Spanish, and contains 50,000 to 100,000 historical roots of words covering over 90 per cent of words occurring in historical texts.³

IMPACT

In the IMProving ACcess to Text (IMPACT) project, ABBYY also played an important role and supports the European Union Commission's research project for preparing state-of-the-art OCR technology.⁴ With the IMPACT project, the European Union wants to provide a wide readership outside of the research field with access to historical texts and promote the digitization of European cultural heritage through the continuous enhancement and improvement of relevant technologies. In close cooperation with the IMPACT project team, ABBYY's research and development team recently enhanced both the technology for image processing applications as well as the technology for the analysis of document layouts. Taking into account numerous sample documents provided as part of the project by Europe's leading libraries, ABBYY was handling the adjustment of OCR core components for a variety of historical printed documents into various European languages. On this basis it has been possible to achieve important advances in technology for improving the quality of Gothic character recognition. In addition, as part of the IMPACT project ABBYY developed a special OCR XML output option for restoring logical document structures.

² For more information, see <http://www.frakturschrift.com/de/projects:metae>.

³ A software development kit for FineReader XIX, including the Gothic Classifier and historical dictionaries, is available from ABBYY.

⁴ For more information, see <http://www.frakturschrift.com/de/projects:impact>.

Why should OCR be applied to old documents?

OCR plays an important role in the digitization and decoding of our printed cultural heritage at the national, European and global level. Outside a small circle of researchers, most readers today are not capable of reading Gothic writing. Through the use of technologies such as Gothic OCR, which makes the reliable text recognition and comprehensive digitization of older documents increasingly possible, historical sources are becoming available to a much larger readership than ever before. Therefore, software like ABBYY's omnifont OCR, which can reliably decipher Gothic fonts without prior training, is of great importance, not only in the context of individual digitization projects for libraries, but also for general mass digitization of historical text documents.

As OCR software becomes increasingly easier to use, old texts are being rediscovered and used for reprints. Users in the academic field benefit particularly from digitisation. By capturing text digitally, old documents can now be searched through, allowing for the more efficient research of the documents. The conversion of old texts into modern digital formats such as XML with specific meta-information (e.g. information about the original layout of the document), searchable PDFs and e-books, using OCR software thereby expands the possibilities of distributing historical documents considerably, without having to resort to using valuable and possibly fragile paper originals.

Quality of historical OCR-based text recognition

Although many projects are dedicated to the continuous development of technologies for Gothic recognition and significant advances have been made in recent years, questions are still raised regarding recognition quality. For example, "when is 'good' really 'good enough'?" Even if the subject has now almost become a philosophical one, some quality-related aspects are shown below. The overall accuracy that is achievable, generally depends on the many different parameters as briefly outlined below.

- Quality of the original paper
- Quality of the scan
- Correct scan parameters
- Quality of the image processing
- Quality of the document analysis to accurately identify all text and image areas in the document
- Accurate reconstruction of the layout
- Retention of the reading sequence
- Optimised character recognition for antiqua fonts
- Use of an OCR solution specialised in Gothic script
- Availability of suitable dictionaries
- Options for manual and automated (re) correction

Of course, in practice there is always the desire for nearly 100 per cent accurate character recognition. As described above, the particularities of historical text documents means that this cannot always be achieved automatically. The very high-quality standards that can be achieved with the OCR application for modern documents require considerably more time when dealing with historical material. Therefore it is much more costly, both with regard to the preparatory processes, implementation of the project itself and the manual or automatic post-processing of the OCR results.

If users of digital libraries would have the choice between "only" ten very high quality books or 1000 digitally captured historical documents, many of them would probably prefer the

latter. In practice, a decision must always be made very carefully just how much work can or should be invested for which application purpose. Another important topic to mention is that historical and modern word spelling have to be considered when stating a quality discussion.

As historical documents are naturally tied to their historical context, which means both written in an old-fashioned language and using historical spelling, which in some cases differs drastically from current spellings, the use of intelligent search technologies in the development of these documents is necessary in every case, e.g.

- old spellings (in German): Theile, Mittheilung, reduzirt;
- new spellings (in German): Teile, Mitteilung, reduziert.

Search technologies that are used to find words in "historic" spelling even when the modern spelling variant is typed in, is usually also capable of finding words with OCR errors, because it uses fuzzy-search algorithms.

The additional advantage of having historical documents in a modern font is obvious. The uninformed reader can then also read and use Gothic fonts more easily.

For readers who use digital libraries and are interested in historical documents, it is usually an interested and educated reader who can understand the content correctly despite possible inaccuracies in recognition and interpretation.

Conclusion

Projects with Gothic OCR are always relatively complex and, especially for the automated detection of historical documents, no one project will be the same as another. Differences in the quality of templates, layouts, fonts or missing dictionaries place many obstacles in the software's path. Nevertheless, improvements achieved in processing historical documents mean that today's OCR software can also be applied to image collections of historical documents that are already scanned. Required computing power and also any potential licensing costs for using OCR solutions have long since proven to no longer be the decisive factor in not giving old library texts the chance for a "second" digital life.⁵

Today, there are already new possibilities being proposed for the further development and correction of digital documents. The increased spread of open systems, known as crowd sourcing, for example, whereby volunteers read through texts digitised books and correct them, is a highly attractive option. This would mean that old texts and sources do not lose their importance for understanding the past and the present.

ABBYY®

ABBYY Europe GmbH

Elsenheimerstrasse 49
80687 Munich, Germany
Tel: +49 89 511 159 0
Fax: +49 89 511 159 59
sales_eu@abby.com
www.ABBYY.com
www.ABBYY.de

Bureau France
4, rue Leroux
94100 Saint-Maur des Fossés
France
sales_france@abby.com
www.france.ABBYY.com

ABBYY UK Ltd.

Abbey House, Grenville Place
Bracknell RG12 1BP, United Kingdom
Tel: +44 1344 392 610
Fax: +44 1344 392 611
sales_UK@abby.com
www.ABBYY.com

⁵ For more information about Gothic OCR, please go to www.frakturschrift.de.

Image Sources & Copyrights 1st page bottom left: ©Lin Kristensen, New Jersey, USA http://commons.wikimedia.org/wiki/File:Old_book_-_Timeless_Books.jpg This file is licensed under the Creative Commons Attribution 2.0 Generic All other images, ABBYY