Optical Character Recognition

IMPACT Best Practice Guide

IMPACT project
Niall Anderson, British Library
Gunter Muhlberger, UIBK - University Innsbruck, Digitisation and Digital Preservation
Apostolos Antonacopoulos, PRIMA Research Lab, within the Informatics Research Institute (IRIS) at University of Salford

Table of Contents

A Best Practice Guide to OCR

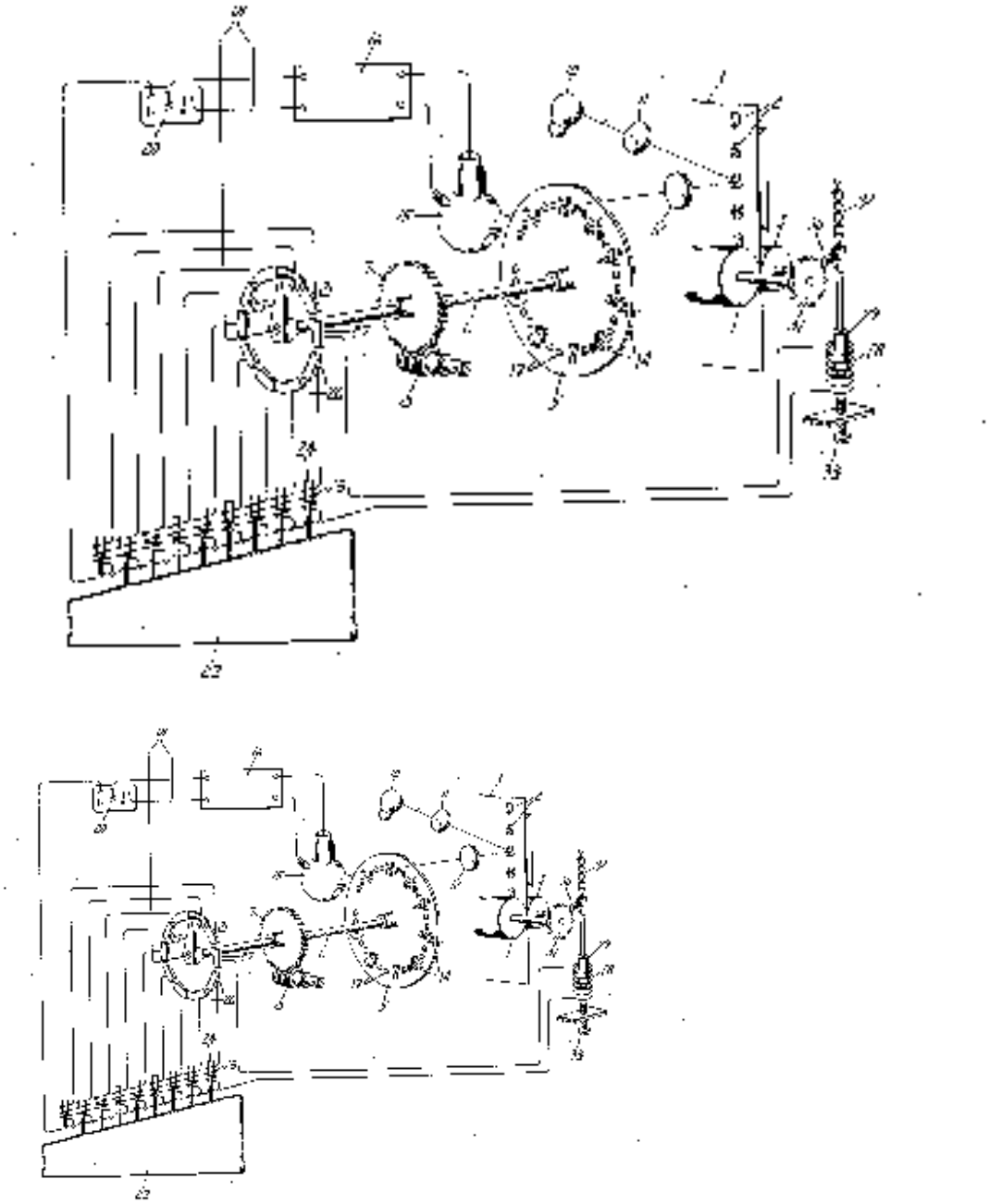
Background and developments to date

Analysing documents as if they were images is a theory and practice dating from at least the 1870s, when machines were invented that could automatically translate written telegram into morse code. But document analysis that specifically aims to convert paper-based textual symbols to a machine understandable format dates to Gustav Tauschek's "Reading Machine" of 1929 (granted US Patent 2,026,329 in 1935) and Paul W. Handel's "Statistical Machine" of 1931 (US Patent 1,915,993).

Both of these patents describe machines that use circular discs with template symbols cut out of them to allow light to shine through.

The image to be recognised was held in front of the disc, with the light shining behind the image through the template, towards an electrical photo sensor. The disc was rotated through each of its template holes

until the light no longer reached the photo sensor: this would indicate that the character on the page and the template hole were an exact match; thus identifying the character. The mechanism of Handel's machine can be seen below:

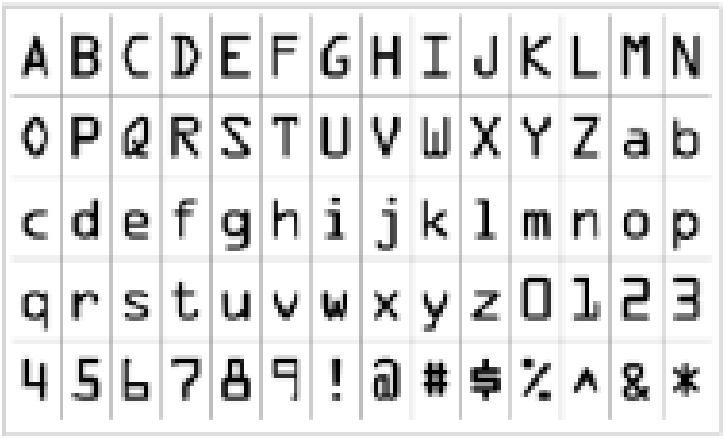


Source: US Patent Office

The next major steps in OCR came from the field of wartime cryptology, when the US Armed Forces Security Agency realised that codes could more easily be analysed and their structure unpicked if they were fed into a computer. This necessitated the translation of printed messages into machine language. The result was 1951's GISMO machine (U.S. Patent 2,663,758), which could analyse a line of text and convert

its shape into machine readable code. This signalled a move away from the strict character recognition technology of earlier devices towards image recognition. GISMO's creator, David H. Shephard, would go on to found the first major OCR company, the Intelligent Machines Research Corporation (IMR).

Throughout the 1950s and 60s, OCR developed broadly in line with the development of office computing. The OCR machines of the era were used solely to read and process contemporary material that had been produced specifically for those machines, so any text or representative shape that did not conform to a few unique, computer-generated fonts was not recognised or read. The first OCR-specific font was designed by RCA in 1965 for the Readers Digest and was called OCR-A:



Source: Radio Corporation of America

Readers Digest used an RCA-designed optical reader and computer to automatically process the OCR-A serial numbers on advertising and competition coupons returned to the magazine. RCA later expanded its character recognition operations to include automated ticket validation for TWA airlines. These early electronic optical readers could process up to 1,500 identical documents per minute, automatically rejecting those they could not process for human inspection.

These and other electronic character readers were remarkably similar to the earliest Tauschek and Handel machines, being comprised of a disc with fixed slits that reflected light passed through and a moving disc with character templates cut out to verify recognition. The main difference was that the reflected image was broken into discrete black and white signals, which were processed in a photo-multiplier tube and converted into binary digits for computer recognition.

The next generation of OCR equipment used a cathode ray tube, a thin beam of light and photo-multipliers to effect a technique called "curve following". As the name suggests, curve following meant using light to trace the contours of a particular character, rather than simply dividing an image into black and white components and interpreting the black matter as text. This allowed for the recognition of dark inks other than black and also for greater adaptability with regard to typeface recognition.

In 1974, Ray Kurzweil started the Kurzweil Computer Products, Inc. and developed the first omni-font optical character recognition system. Kurzweil's original designs were for a reading machine for the blind, which would allow people to understand written text by having a computer read it to them out loud. This device required the invention of two experimental devices: the flatbed scanner and the text-to-speech synthesiser. The Kurzweil Reading Machine was the size of a tabletop and made its public debut at a press conference in the offices of the American National Federation of the Blind in 1976. The first production model was bought by the musician Stevie Wonder.

Commercial versions of the Kurzweil machine followed in 1978 and Kurzweil would sell his company to Rank Xerox in 1980.

The next major step in the history of OCR came with the development of optical array scanning in the early 1980s. By allowing for the scanning and storage of images in any format and to a chosen degree of optical complexity, array scanners allowed for the capture of complete documents to a high standard for the first time, and thus allowed OCR technology to take on material from earlier printing ages. Modern scanners and OCR software still takes it cue from these image processing devices.

How OCR works

Before an OCR system can begin to recognise the symbols present in an input document image, several important processing stages must be carried out to convert the input into a usable form:

- **Digitising the Input -**if the document to be recognised exists as a physical paper printout, the first step that must be carried out involves converting it to a representation that the recognition system can manipulate. Since the majority of modern recognition systems are implemented on computer hardware and software, each page is typically converted to a compressed or uncompressed digital image format.

  This conversion from printed page to digital image often involves specialised hardware like an optical scanner that attempts to determine the colour value at evenly spaced points on the page. The scanning resolution will determine how many of these points will be inspected per unit of page length. Typically this is specified in dots or pixels per inch: thus a document scanned at a resolution of 300ppi will have been sampled at 300 evenly spaced points for each inch of each page.

  While the optical scanner is the traditional means by which paper images become digitised, sometimes digital cameras are used to capture and digitise document images. Camera capture has the advantage of being able to capture documents (like thick books or product packaging) that might prove difficult or impossible using a flatbed scanner

- **Binarisation -**for character recognition purposes, one generally does not need a full colour representation of the image, and so pages are often scanned or converted to a greyscale or bitonal colour depth. In greyscale, each pixel represents one of 256 shades of grey, while in a bitonal image each pixel is assigned one of two values representing black or white. While both of these methods will allow an image to be stored in a smaller digital space (and thus allow more images to be stored overall), the images themselves can suffer from information loss due to the approximation of colour values in the image.

  Working with bitonal images is generally held to be the most efficient practice. Converting a colour or greyscale image to bitonal format is referred to as binarisation. Approaches to binarisation typically fall into one of two categories. “Global” methods treat each pixel independently, converting each to black or white based on a single threshold value. If a pixel’s colour intensity is higher than the global threshold it is assigned one value, otherwise it is assigned the opposite value. By contrast, “local” methods make use of the colour information in nearby pixels to determine an appropriate threshold for a particular pixel.

- **Smoothing Out Textured Backgrounds -**a related problem facing character recognition systems is the separation of text from a textured or decorated background. Textured backgrounds can interfere with accurate binarisation of input document images by obscuring legitimate characters or being read as characters themselves, and they can make segmentation and recognition of characters much more difficult.

  One of the more common approaches is to make use of mathematical morphological operators. After the image has been binarised, a small structuring element (a group of pixels with specified intensities) is created and swept across the image. At each step the pixels in the structuring element and their

corresponding image pixels are compared, and depending on the result of this comparison as well as the operation being performed, the image pixel underneath the centre of the structuring element is updated.

- **Noise Removal -**during the scanning process, differences between the digital image and the original input can occur. Hardware or software defects, dust particles on the scanning surface and improper scanner use can all change pixel values from those that were expected. Such unwanted marks and differing pixel values constitute noise that can potentially skew character recognition accuracy.

  In addition, certain marks or anomalies present in the original document before being scanned (in particular dark page borders) constitute unwanted information that can increase the time spent processing an image and the space needed to store it.

  There are two broad categories of digital noise. “Additive” noise occurs when an unwanted visual effect has a greater colour depth than the median colour threshold in the overall document, and is thus represented in the binary image as a black character. Conversely, “substractive” (or dropout) noise occurs when a character’s colour depth is less than the median colour threshold of the overall document, leading to legitimate characters being whited out. In both cases, noise can be retrospectively removed by tweaking the image thresholds or using a smoothing filter (which averages out neighbouring pixel intensities), but these are imprecise techniques. This should underline that for good character recognition a good scan is a paramount.

- **Page Deskewing -**a common problem that occurs when a flatbed scanner is employed to digitise paper documents is that the paper is often placed so that it does not lie exactly perpendicular with the scanner head. This can lead to the image being skewed. Depending on the severity of the effect, skewed images can lead to difficulties when attempts are made to segment the input image into columns, lines, words or individual character regions.

  Most scanning software now contains a deskew feature as part of its post-capture processing suite. There are a variety of different methods for both detecting and correcting skew, but most work by identifying upward-pointing straight lines within a character and measuring the degree to which they all point in the same direction: this gives a good indication of the degree of skew. Deskew software will then treat all characters within a skewed section as though they were part of a bounding rectangle and rotate the rectangle to the angle where the skew is minimal.

- **Page Dewarping -**non-linear warping is often observed in document images of bound volumes when captured by camera or flatbed scanning or due to environmental conditions (e.g. humidity that results in page shrinking). Text in such cases is strongly distorted and this not only reduces the document readability but also affects the performance of subsequent processing such as document layout analysis and optical character recognition.

  Over the last decade, many different techniques have been proposed for page dewarping. According to whether auxiliary hardware or information is required, the proposed techniques can be classified into two main categories based on 3D document shape reconstruction and 2D document image processing. Techniques of the former category involve image capture with special setup (stereo-cameras, structured light sources, laser camera), prior metric knowledge or they rely on restricted models to obtain the 3D shape of the page. The main drawback of these techniques is that they require the use of calibrated digital cameras. On the other hand, techniques in the latter category, without any dependence on auxiliary hardware or information, use only 2D information from document images. These techniques can also handle cases of arbitrary warping such as page shrinking due to humidity.

- **Layout Analysis -**once a page has been digitised, denoised, deskewed and dewarped, the final preprocessing task involves identifying the regions of text to be extracted from the page. Like most of the other phases of document analysis, there are many ways to attempt to tackle this problem, and doing so leads naturally into the follow-up tasks of isolating and segmenting the lines, words, and individual character images for the purposes of recognising them.

While large, single column paragraphs of body copy may be fairly trivial to pick out of a document image, this process becomes increasingly more complex as one tries to extract the contents of tables, figure captions, text appearing within a graph or halftone image, etc. In multi-column documents and documents containing figure captions, determining the correct reading order of the text is not easy. Additionally, historical machine printed documents have some notable characteristics which make the segmentation problem difficult and very challenging. These include the existence of non-constant spaces between text lines, words and characters as well as the existence of various font sizes, marginal text, ornamental characters and graphical illustrations.

The process of identifying the entire textual region and reading order is often given the term zoning. There are many different ways of doing this, but all methods boil down to partitioning a page into its constituent parts and assigning each one an identified region type (such as a table) and within those regions a character type and orientation.

In recent years, a process known as *logical layout analysis* has been used to differentiate between regions of the same broad type (for instance distinguishing a work's title whether it appears in the body text, an abstract, or a footnote). Though this fine-grained distinction is most useful for other tasks like information retrieval and document classification, it still provides additional information that can aid in the character recognition process. For instance, knowing in advance that a text region has been identified as a page number allows one to significantly restrict the set of output symbols (to numbers or roman numerals), and thus reduce the likelihood of a misclassification.

- **Optical Character Recognition -** once a page has been suitably digitised, cleaned up, and had its textual regions located, it is ready to be segmented so that the individual symbols of interest can be extracted and subsequently recognised. Again, there are a number of different methods for extracting this data, but most OCR engines begin by identifying a character and running that character's properties through an internal database of symbols until it finds a match. Once it has positively identified a character, it moves on to the next within the relevant textual region.

  More complex OCR engines attempt to limit the number of characters they need to process by using probabilistic analysis of the incidence of letters within a particular language. In English, for instance, the letter q is almost invariably followed by the letter u, so a probabilistic search engine will begin from the assumption that the two letters are contiguous.

Best Practice in the Use of OCR

As noted above, there is no single correct method for producing or displaying machine-readable text. In considering which approach may be right for your project, you should consider the following factors:

- **Use should reflect project goals -** as noted above, if your project needs text files that will be read by a search engine to support full text searching, then text files produced by an OCR application may be sufficient. Depending on the source material, however, you may find that even the best OCR output will require some manual monitoring or correction in order to produce satisfactory search results. Making this decision will involve careful analysis of the level of search accuracy you require and the practical implications that will go towards achieving various levels of accuracy. Is it necessary, for example, for someone to find every occurrence of a search term? Are some sections of the text more critical for searching than others? The answer to these and comparable questions will help to determine the best approach.

  Similarly, you may decide that you wish to support features beyond full text searching and these may require a higher level of machine readability. In order to broaden search options, for instance, you may wish to apply SGML, XML or other tagging to support the focused searching of key elements or sections

of the text. Finally, you will want to decide if you will display the OCR text files to users. Should your project include display, you will want to consider what this might mean for the level of accuracy required for OCR output. You will want to determine what level of tolerance the user may have — or may not have — for errors in the OCR text files. An understanding of the quality expectations of the users and how the displayed files may be used will be helpful in this analysis.

- **Understand the characteristics of your source material -** as noted above, the quality of the paper and the characteristics of the printed source material including language, font, layout, and graphical elements, can impact on the overall quality of the OCR output. Understanding the full range of features present in the source will allow you to determine the most efficient way — or even whether — to employ OCR. If, for instance, your source material contains a large number of full-page graphics or includes a large number of special characters or scientific or mathematical symbols, you may wish to develop an automated means for filtering these pages for special OCR or other treatment.

  OCR engines heavily rely on lexical data. OCR engines usually come with predefined “dictionaries” for a set of mostly modern languages. Obtaining appropriate lexical data for historical language or special terminology is often a problem.

- **Develop quality control measures -** with a clear understanding of the source material characteristics, it will be possible to establish text accuracy targets that will successfully support the mission of the project. But in order to give these targets meaning you will need to establish a quality control program to ensure that established targets are met. Quality control measures may vary from a complete review of all phases of text production to a sampling approach, in which only a subset of pages is selected for review. If a complete review is cost-prohibitive, as it likely will be for mass digitisation projects, it may be important to employ a sampling standard such as ISO 2895-1[1]. It is also important to have procedures in place regarding any data errors that are found. Will these errors be corrected by in-house staff? Or, if the work was produced by a third party, will the re-work be done by the vendor? How will quality standards be communicated to the staff involved? Whatever the approach to a quality control program, it will be important to consider carefully the staffing, budget, and space implications of the work.

- **Understand the impact of scale -** it is very important to realise that the solution that is appropriate for a short-term project of 10,000 pages may not be appropriate for a longer-term project of 10 million pages. Even for short-term projects, it may be that the approach which works smoothly in the start-up phase begins to break down as the project ramps up in scale. It is important to be sensitive to the impact of scale and to take this into account in project schedules and budgets as changes in scale can have a dramatic impact on both.

- **Location of OCR processing (outsourcing or in-house) -** Production of OCR can be, and frequently is, contracted to a third-party vendor; however, outsourcing may not be appropriate for all projects. In considering whether to outsource or to retain this process in-house it is important to look at a variety of questions. You will want to ask which approach makes the most effective use of hardware, software and staff. Which approach provides the greatest value—as distinct from offering the lowest cost? How important is it to retain local control over all aspects of the process? What impact will each approach have on key production schedules? What skills can an outside vendor offer, and what benefit is lost to the project by not having these skill sets locally?

  Addressing these and similar questions will help to develop an understanding of the comparative advantages that each approach may offer. Outsourcing is discussed in depth in the IMPACT Best Practice Guide to Outsourcing.

- **Project Duration -** the time required to run OCR software can vary greatly depending on the characteristics of the source material, the specific software, the number of engines, the hardware running the process, and the quality control measures developed for the project. Even within a single printed

[1] *Large Scale Digitisation Initiatives*; 2009; Conteh, A: http://www.slideshare.net/JISCDigi/aly Retrieved 13.03.2011

work, significant deviation can occur as the content varies. Running a representative test sample of the source material can be helpful, but it will not necessarily reveal the complexities of ongoing production processes. Given the speed of technical innovation for hardware and software, it is important to take into account the expected duration of the project. It may be that the hardware and software applications that were the best choice at the beginning of a project become dated and require reassessment as the project moves forward. And even if the original hardware and software applications continue to function, the project may gain important cost benefits from upgrading. Consequently, a careful, periodically recurring evaluation of new developments and technologies is recommended for projects of all but the shortest duration.

- **Costs -** while it may be relatively easy to project costs based on an initial run of sample data, it is difficult to anticipate how actual production costs could differ from estimates made from pilot data. Typically, production costs are higher than expected, especially in the early days of a project. Machines may not perform at the speed advertised or as tested; drives may fail; software upgrades may be needed. As the scale of production increases, additional purchases may be needed to reach the desired production level and to maintain server response time within an acceptable range. Similarly, as project scale increases more staff time is required to tend to systems administration and developing a technical infrastructure suited to the operation. It is helpful to anticipate the unexpected in these areas and to build into project budgets, if possible, some ability to respond to these unforeseen needs as they arise.

Considering each of these factors may help in identifying the best approach to producing machine-readable text for a particular project. However, each of these elements must be examined in the context of the overall mission of the project. It is only in this context that the best method that offers the fullest level of support for the goals of the project within the constraints of time and budget can be identified.

For an interactive guide to costing a digitisation project from beginning to end, consult the IMPACT Cost Estimator [https://www.surfgroepen.nl/sites/impactproject/oc/GA%20Annex%201%20Description%20of%20Work/OC2/OC2.1/IMPACT_Storage-Estimator_BSB_version3-2.xls].

Avoiding problems in OCR

As the foregoing will have made clear, one of the chief contributing factors to the success of the OCR process is the quality of the initial scan. Three broad factors external to the quality of the OCR software itself can have potential negative effects on high OCR accuracy. These are:

- The underlying qualities of the source material
- The initial image capturing process (including the technology used to create the image)
- Image enhancing techniques employed to improve legibility/readability

This section deals with these common barriers to high OCR accuracy and how to avoid them. Note that negative factors resulting from source material may not be correctable, though they may guide selection of material for OCR.

Negative factors resulting from the source material

Yellowed paper.

Background colour may change across document due to age of material, causing fuzziness at edges of characters. This may have a minor effect on OCR accuracy.

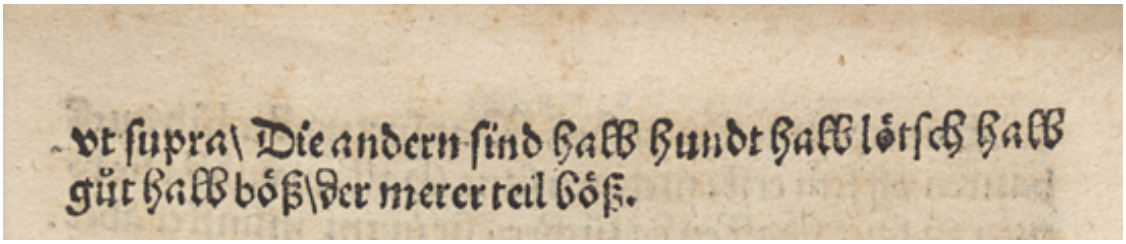


Image taken from the collection of the **Bayerische Staatsbibliothek,** as part of the IMPACT random dataset. Reprinted by permission.

Warped paper

The combination of humidity and age means that the pages of many books are warped rather than flat. As a result, the text warps too. This will have a relatively high impact on OCR accuracy, especially if found in combination with bad printing

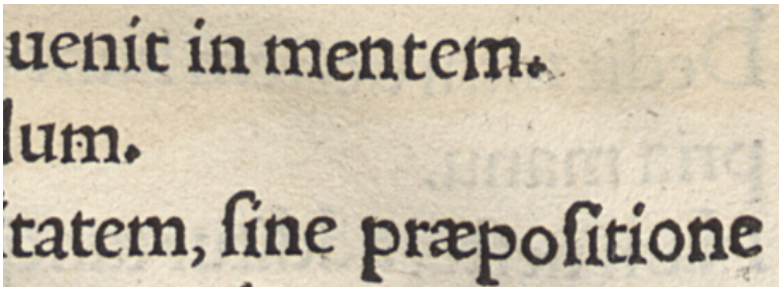


Image taken from the collection of the **Bayerische Staatsbibliothek,** as part of the IMPACT random dataset. Reprinted by permission.

Ink transfer

Ink transfer occurs when two leaves of a book have been pressed together before the ink dries. This can have very negative effects on OCR accuracy.

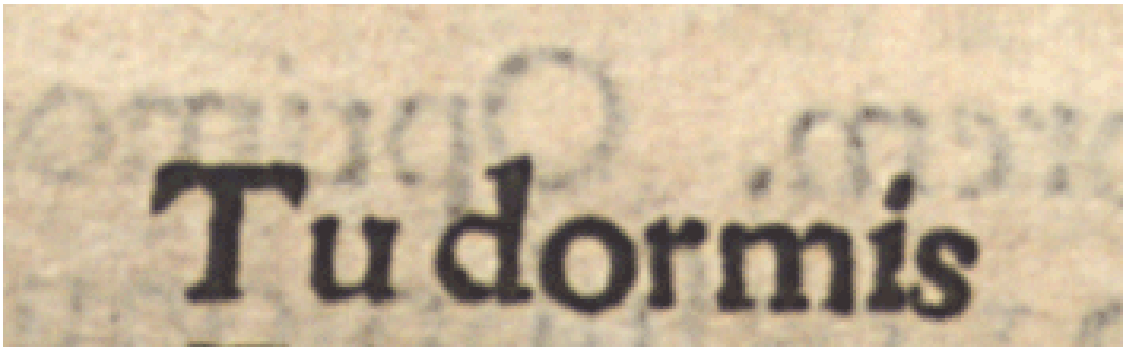


Image taken from the collection of the **Bayerische Staatsbibliothek,** as part of the IMPACT random dataset. Reprinted by permission.

Show through of ink

In documents printed on very thin paper, the ink from one side of the page will often show through to the other. This can have severe negative effects, comparable in all ways to bleed through

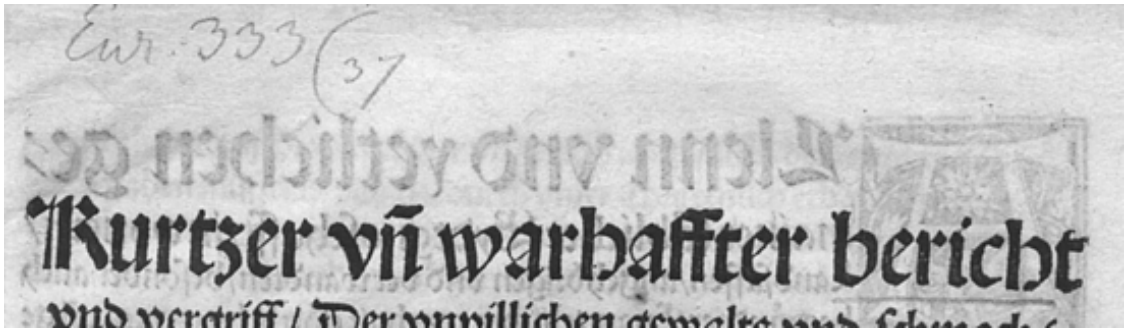
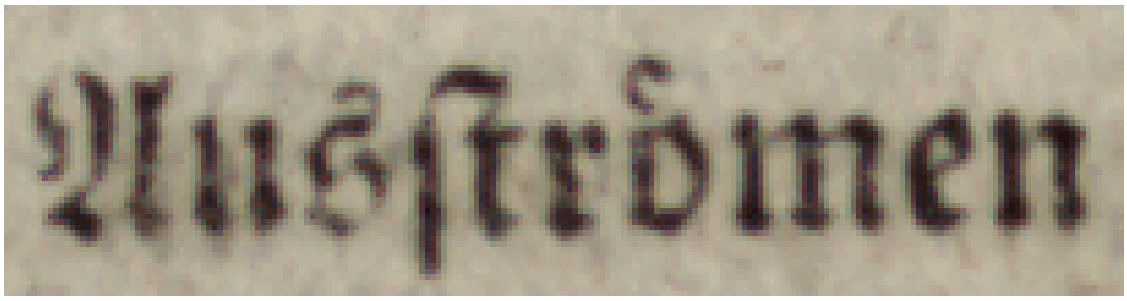


Image taken from the collection of the **Bayerische Staatsbibliothek,** as part of the IMPACT random dataset. Reprinted by permission.

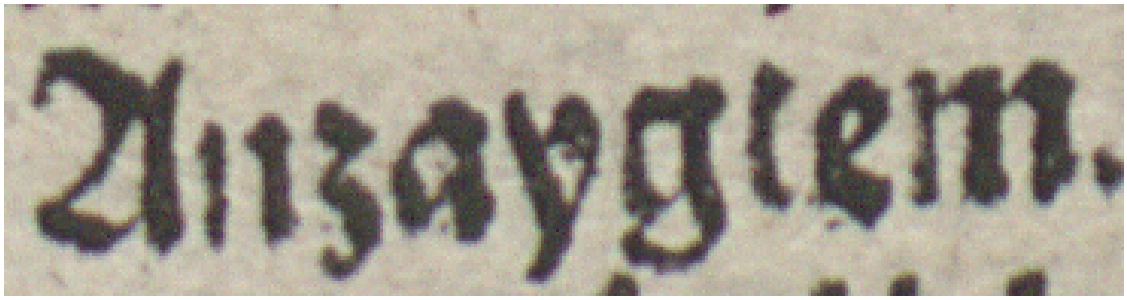
Bad printing

Blurred, fat, broken or faded characters. These can have highly negative effects, particularly where characters are blurred together or broken into more than one shape



Blurred

Broken and dotted



Fat

Images taken from the collection of the **Bayerische Staatsbibliothek,** as part of the IMPACT random dataset. Reprinted by permission.

More than one ink colour

Moderately severe effects can occur. OCR may not recognise complete characters or words

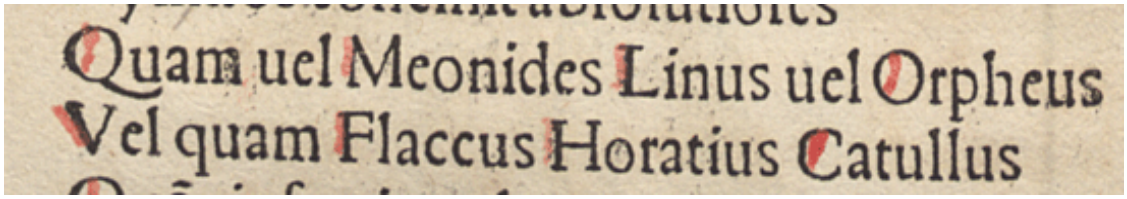


Image taken from the collection of the **Bayerische Staatsbibliothek,** as part of the IMPACT random dataset. Reprinted by permission.

Annotations

These can include notes and drawings by users; also library stamps and watermarks. These are very unlikely to be read correctly, but a deeper problem is that annotations can confuse the segmentation process of OCR engines (e.g. the means by which it identifies character/word/line blocks).

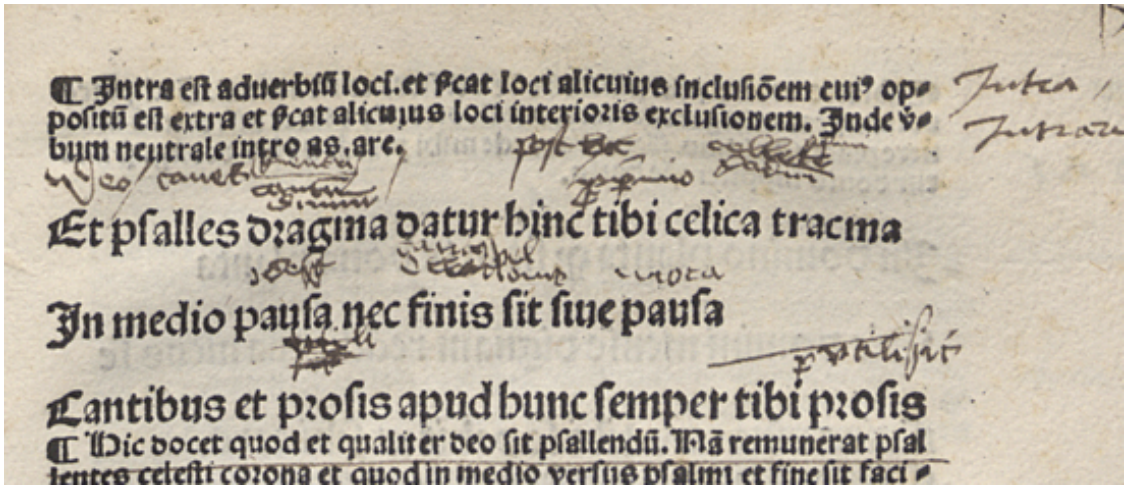


Image taken from the collection of the **Bayerische Staatsbibliothek,** as part of the IMPACT random dataset. Reprinted by permission.

Lack of lexicon data

This occurs when the OCR engine does not have access to relevant language data for the document. This leads to unrecognised words, and perhaps more damagingly to false substitutions of words

Negative factors resulting from the image capture process

In general, the best OCR results will be produced by creating images at a resolution of 400 pixels per inch (ppi) or above, in colour or grey-scale.[2] These standards will preserve the vast majority of the detail in the original item, where lower resolutions will result in progressively worse OCR recognition.

However, capturing images in such high quality has cost and storage implications, particularly if the master image file is to be archived indefinitely. Some institutions engaged in mass digitisation create an OCR output from the master file, and replacing the master image with a compressed access copy to conserve storage space.

These are the most common ways in which image capture can negatively affect OCR accuracy:

Narrow binding

May result in geometrical distortion. This may result from carelessness by the camera/scanner operator in correcting for the tightness of the binding, or from the inability of a particular scanner to deal with books that will not comfortably open beyond 60°. It can have very high negative consequences on OCR accuracy;

Image not cropped

Borders and unnecessary detail from facing page may appear in scan. It is particularly prevalent in scans from microfilm. While very common, it will generally have a low effect on OCR accuracy; most OCR engines can correct for these details;

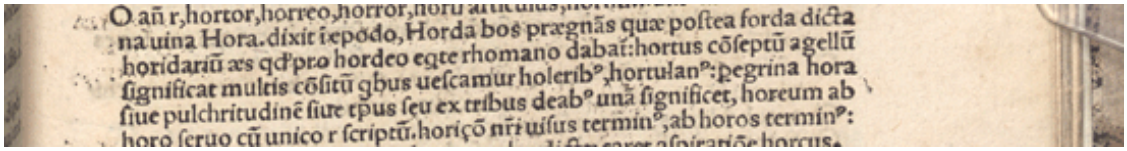


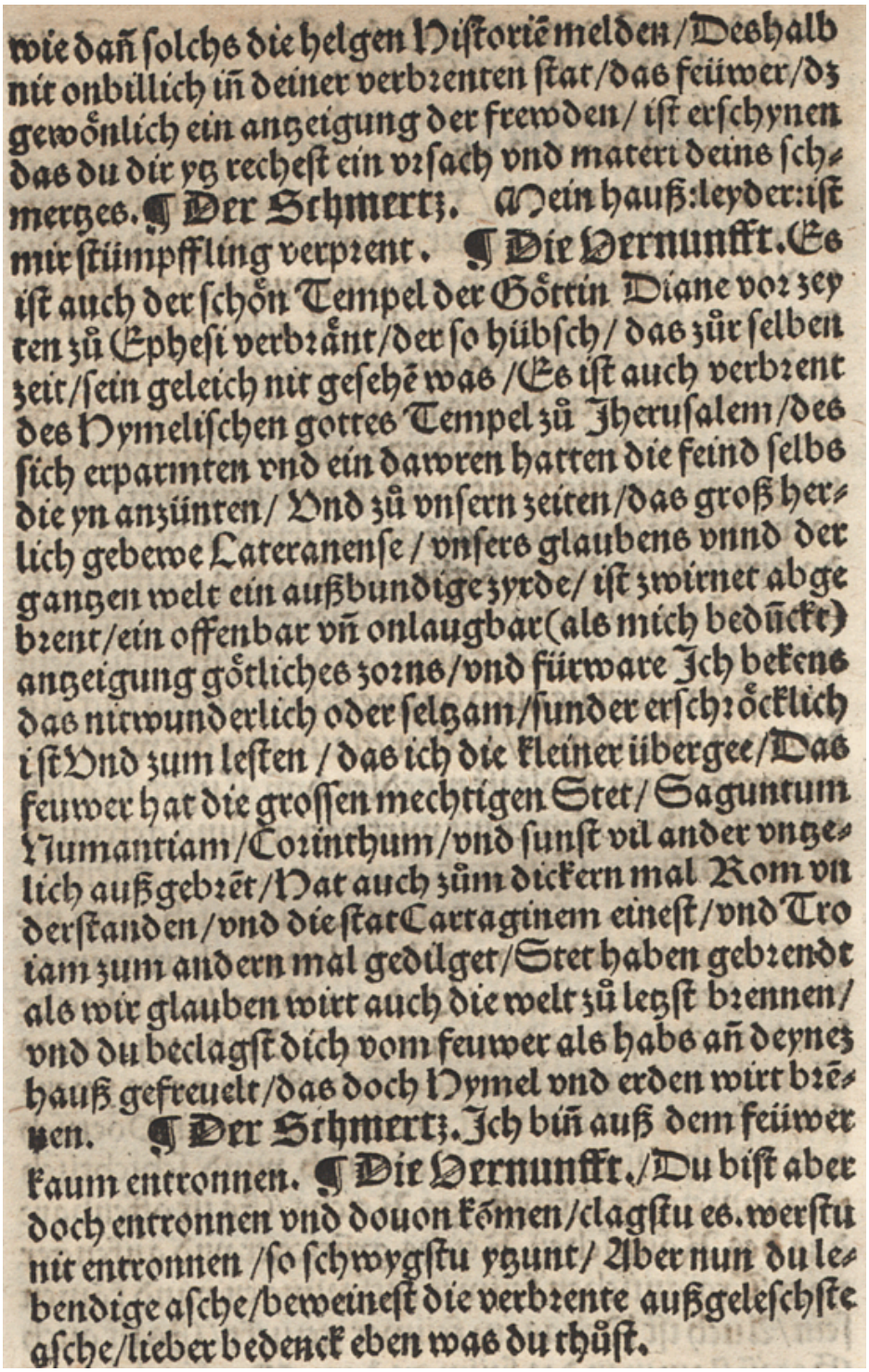
Image taken from the collection of the Bayerische Staatsbibliothek, as part of the IMPACT random dataset. Reprinted by permission.

Image skew

In mass digitisation conditions, the scanner will not be adjusted page-by-page for the relevant print space (i.e. just that part of the material that is needed). This can lead to the creation of skewed images. Because

[2] The superiority of greyscale over bitonal scanning has been called into question in a recent study: *Going Grey? Comparing the OCR Accuracy Levels of Bitonal and Grayscale Images;* Powell, T. and G. Paynter, National Library of New Zealand; D-Lib Magazine, Volume 15, Nos. 3&4, ISSN 1082-9873: http://www.dlib.org/dlib/march09/powell/03powell.html Accessed 13.03.2011

a common problem, OCR engines have been trained to correct for it, so effects on OCR accuracy are generally low. However, the severity of the effect on OCR depends naturally on the degree of skew.



wie dañ ſolchs die helgen Hiſtoriẽ melden/Deshalb
nit onbillich iñ deiner verbrenten ſtat/das feüwer/dz
gewönlich ein anzeigung der frewden/ iſt erſchynen
das du dir yetz recheſt ein vrſach vnd materi deins ſch-
mertzes. ¶ Der Schmertz. Mein hauß:leyder:iſt
mir ſtümpfflíng verprent. ¶ Die Vernunfft. Es
iſt auch der ſchön Tempel der Göttin Diane vor zey
ten zů Epheſi verbrãnt/der ſo hübſch/ das zůr ſelben
zeit/ſein geleich nit geſehẽ was /Es iſt auch verbrent
des Hymeliſchen gottes Tempel zů Jheruſalem/des
ſich erparmten vnd ein dawren hatten die feind ſelbs
die yn anzünten/ Vnd zů vnſern zeiten/das groß her-
lich gebewe Lateranenſe / vnſers glaubens vnnd der
gantzen welt ein außbundige zyrde/ iſt zwirnet abge
brent/ein offenbar vñ onlaugbar(als mich bedũckt)
anzeigung göttliches zorns/vnd fürware Jch bekens
das nit wunderlich oder ſeltzam/ſunder erſchröcklich
iſt Vnd zum leſten / das ich die kleiner übergee/Das
feuwer hat die groſſen mechtigen Stet/ Saguntum
Numantiam/Corinthum/vnd ſunſt vil ander vnge-
lich außgebrẽt/Hat auch zům dickern mal Rom vn
derſtanden/vnd die ſtat Cartaginem eineſt/vnd Tro
iam zum andern mal gedilget/Stet haben gebrendt
als wir glauben wirt auch die welt zů letzſt brennen/
vnd du beclagſt dich vom feuwer als habs añ deynez
hauß gefreuelt/das doch Hymel vnd erden wirt brẽ-
nen. ¶ Der Schmertz. Jch bin̄ auß dem feüwer
kaum entronnen. ¶ Die Vernunfft./Du biſt aber
doch entronnen vnd douon kömen/clagſtu es.werſtu
nit entronnen /ſo ſchwygſtu eyunt/ Aber nun du le-
bendige aſche/beweineſt die verbrente außgeleſchſte
aſche/lieber bedenck eben was du thůſt.

Image taken from the collection of the **Bayerische Staatsbibliothek,** as part of the IMPACT random dataset. Reprinted by permission.

Factors resulting from image enhancement techniques

Most industrial image capturing technology is bundled with a suite of image processing and enhancement software. Some of the above-mentioned factors that negatively affect OCR results can be “corrected” by this software, so that a skewed image can be reoriented, or a page automatically cropped so no unnecessary detail is included in the image file, etc. However, the results of these image enhancing tools are often difficult to objectively judge, because the software works only with its parent scanner. The effectiveness of the tools is therefore directly related to the initial effectiveness of the scan.

In addition to this, image enhancement tools of this type often feature a range of advance features that allow images to be altered post-capture, so page images can be automatically sharpened or have their contrast boosted, etc. But problems can occur when images are enhanced in a large batch: while the OCR accuracy from some images will be enhanced by the application of image sharpening, some could be negatively affected.

OCR engines also include as standard a number of image enhancement tools. The most common are:

- **Binarisation tools -** where a colour or grey-scale picture is converted into a pure black and white (binary) image. Making an image bi-tonal in this fashion helps the OCR engine to separate textual detail from background detail. It will also reduce the effects of bleed-through and show-through;
- **Deskew -** which corrects the angle of a page;
- **Dewarping -** which corrects for the curve of a page and the characters on it;
- **Border removal -** which removes the borders at the edge of a scan. This minimises the storage space required by a particular image and can also improve OCR – seeking as it does to ensure that no extraneous details are recognised as characters;
- **Page segmentation -** which can divide a page image into headers and text, individual paragraphs and columns and, in the case of newspapers, into individual articles.

The image enhancement software in scanners and OCR engines can produce unwanted effects when not set up with enough care:

Bitonal output reduces readability

Scanners and OCR engines will automatically transform grey or colour images into binary images before OCR. If a document is particularly difficult or if the parameters for binarisation are not set with enough care, OCR accuracy can be severely reduced;

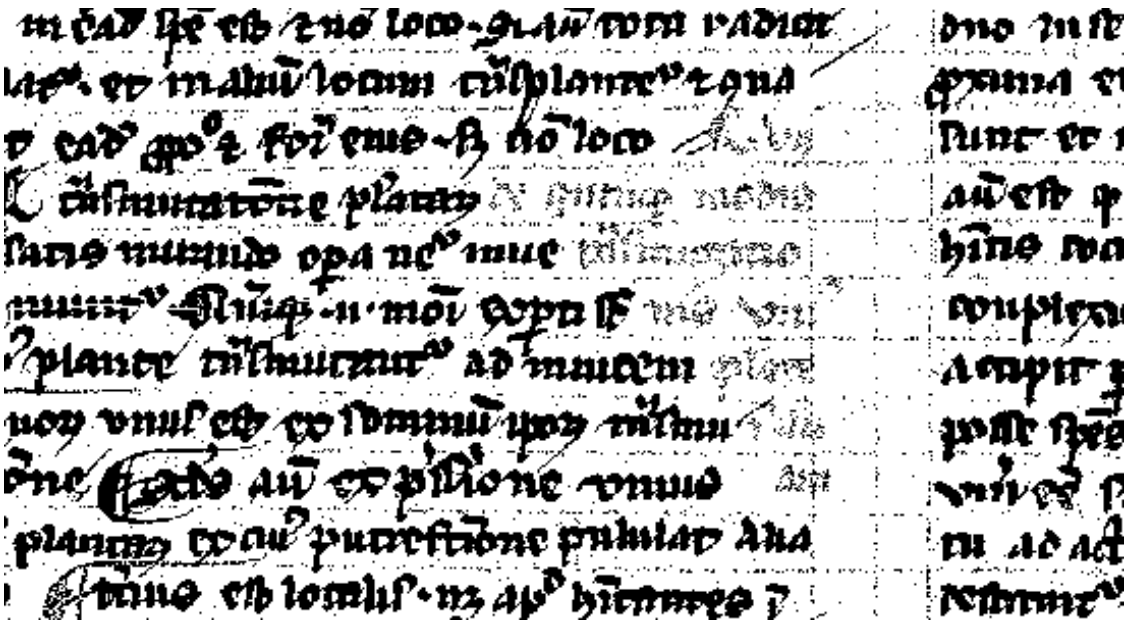


Image taken from the collection of the **Bayerische Staatsbibliothek,** as part of the IMPACT random dataset. Reprinted by permission.

Under-correction of processing software for poor quality scan

The image processing software in a scanner or OCR engine may not be sufficiently sophisticated to correct for an image that has been captured with insufficient regard to lighting, contrast, etc. This is particularly true of bitonal image creation and often results in broken text. This can have very high negative effects on OCR

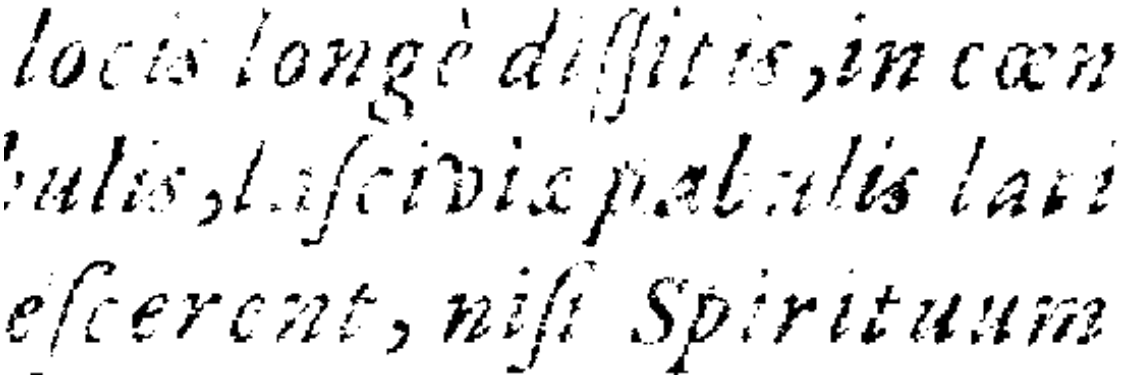


Image taken from the collection of the **Bayerische Staatsbibliothek,** as part of the IMPACT random dataset. Reprinted by permission.

Implementing OCR

Sample costs of OCR

In 2008 The European Library Plus project conducted a survey of its library members[3], asking them what they paid per OCR newspaper page. The costs exclude the prices of hardware and software:

Table of OCR page prices calculated by TELplus content-providing partners

Library	Outsourced/In-house	Price per page
Austrian National Library (ONB)	Outsourced	0.25 €
National Library of the Czech Republic (NLP)	Outsourced	0.026 € - 0.032 €
National Library of Spain (BNE)	Outsourced	0.04 € - 0.06 €
French National Library (BnF)	Outsourced	0.063 €
National Library of The Netherlands (KB)	Outsourced	0.15 €
The British Library (BL)	Outsourced	0.32 €
Martynas Mažvydas National Library of Lithuania (LNM)	Outsourced	0.61 €
National Széchényi Library of Hungary (OSZK)	In-house	0.05 € - 0.10 €
Slovak National Library (SNK)	In-house	0.10 € - 0.50 €
National Library of Estonia (RR)	In-house	0.20 € - 0.66 €
National and University Library of Slovenia (NUK)	In-house	0.30 € - 0.65 €

The prices listed in this report vary widely for a number of reasons. First, price per page goes down depending on the overall size of the digitisation project; second, if OCR output is provided in a single format it will be cheaper than OCR output in a number of formats (as happens at the KB); thirdly, if the basic material is difficult to OCR accurately it will cost more time and money to get an acceptable OCR output. (All of these findings are discussed in more detail in the original document.)

However, perhaps the most significant finding is that whether OCR is outsourced and in-house does not appear to be a particular factor in the overall cost. As noted, the survey discounts the cost of buying and maintaining software and hardware, but after the initial capital outlay there appears to be little material difference.

For an interactive guide to establishing likely costs from beginning to end of a digitisation workflow, consult the IMPACT Cost Estimator [https://www.surfgroepen.nl/sites/impactproject/oc/GA%20Annex%201%20Description%20of%20Work/OC2/OC2.2/Cost%20calculator/oc2_cost_calculator_v3.05.xls].

[3] *Survey of existing OCR practices and recommendations for more efficient work*; 2008; Korb, J.; http://www.theeuropeanlibrary.org/telplus/documents/TELplusD1%202_Final.pdf Retrieved 12.02.2010

Evaluation and Quality Assurance of OCR Results

The most effective and comprehensive way of measuring the accuracy of OCR results remains manual revision, but this is cost prohibitive at any level, and unworkable as a single method of evaluation in a mass digitisation context.

One of the simplest alternatives has been to manually check an OCR engine's software log for a random batch of files. The software log is where the OCR engine documents its own (self-perceived) success rate, ranking character and word matches against the software's internal dictionary. Checking the log therefore enables the user only to assess how successful the software thinks it's been, and not how successful it's actually been.

Pursuing a similar methodology, but in more depth, are human eye comparisons between the text in digital image file and a full text representation of the OCR output from that text. This is obviously more time consuming than checking against the software log, but gives a much more accurate idea of the success rate of the OCR. Key to this method is ensuring that the sample images checked are representative of the entire corpus to be digitised, in type and quality; otherwise the results for the particular sample can be far from the overall average.

Useful as this test can be, it does not on its own take into account the "significance" of a misrecognised character or word. Certain OCR misrecognitions may have little or no effect on either the searchability or legibility of a text-based digital image, but will still drag down the overall OCR average. Simon Tanner, in his paper 'An Analysis of OCR Accuracy within The British Library Newspaper Digitisation Programme'[4], suggests that certain common words in a language (in English, these tend to be widely used articles, pronouns, prepositions, and prepositional pronouns) might usefully be discarded from analysis, giving a picture of the usefulness of the OCR output rather than its strict accuracy.

All of these methods can be used at a project's start-up phase – as a benchmarking exercise for both hardware and software – or throughout a project's life cycle. But because they are necessarily limited in scope due to any project's timescale and resourcing, they tend to wane in importance as a project progresses. A simple, statistical method for monitoring OCR success throughout a project is to include the software log's success rate in the OCR output file (ALTO files allow you to do this), or at least keep it separately. Looked at en masse, it will give an overview of where the OCR engine thinks it's succeeding and where it thinks it's failing. If there are wide discrepancies between one batch of files and another, the software log will allow the institution to prioritise those files where OCR accuracy is low, and to manage (and hopefully mitigate) those discrepancies.

Conclusion and further developments

As the foregoing makes clear, to maximise OCR success great care must be taken in the planning and execution of projects. The supply of good quality image inputs into the post-capture processing stages will reduce the complexity and cost of processing and reduce dependence on correction techniques subject to the inherent material characteristics. It is strongly recommended to allow for pilot processing of a representative selection of material through all stages of the project including OCR – this will allow you to make informed choices in the trade off between quality, volume, cost and speed.

But what of the future of OCR technology? Increasingly, digitising institutions and research partners are moving towards concepts of Adaptive OCR; that is to say, towards developing software engines that learn from user feedback and from their own software as they process material, adding new words to a custom dictionary, or recognising and anticipating a degree of skew or warping throughout a single volume. IMPACT is fundamentally concerned with developments of this sort.

[4] *An Analysis of OCR Accuracy within The British Library Newspaper Digitisation Programme*; 2008; Tanner, S; http://www.impact-project.eu/uploads/media/Simon_Tanner_Measuring_OCR_Accuracy.pdf

By the end of 2011, IMPACT will publically introduce novel hierarchical segmentation models that allow the discrete problems of text block, text line, word and character segmentation to be addressed separately, while at the same time allowing for interplay between the three levels. This segmentation scheme will include advanced methodologies that will push forward the state-of-the-art and efficiently handle documents with complex and dense layout.

In addition to this, IMPACT will develop the science of machine language recognition. Current OCR dictionaries tend only to deal with very common proper nouns in a language, and have little if any support for historical variation in spelling. Given that the bulk of text material being digitised in Europe comes from before the 20th century, this leaves OCR engines with a significant gap in linguistic knowledge. By 2011, the IMPACT Project aims to produce both dictionaries of historical variation and of Named Entities for Dutch, German and English[5]. It is envisaged that the Named Entity material can be enlarged upon by users through the IMPACT Project website.

In addition to this, IMPACT will lead in enhancement and enrichment strategies for already digitised materials.

Increasingly, organisations that are engaged in digitisation are looking for ways to both maximise user experience through full text representation, and to compensate for the shortfall in accuracy of their original OCR results. This developing field is known as “post-correction” and IMPACT aims to lead in its development.

There are two main types of post-correction in development:

Automatic post correction

The raw OCR file can be post processed and corrected with the help the specialised dictionaries treating historical language and named entities mentioned above and with the help of context-aware language models. This lexical data can also be used directly by the OCR engine.

In addition to this the IMPACT Project has built a language modelling module, which will work with the historical dictionaries to identify patterns of relationship between words and thereby increase the comprehensiveness of the overall dictionary[6].

Cooperative Correction

Another experimental field showing some encouraging results is cooperative correction, whereby users help in correcting digitised documents, The full text of a digital document is shared with a resource’s users, along with a toolkit that allows them to correct the original OCR results. The IMPACT Project is building a sophisticated “carpeting” tool for collaborative correction, whereby a great many characters and words with suspect OCR results can be corrected and filed. The more characters a user corrects, the more the tool’s ability to recognise characters grows – so users will be developing the OCR engine.

Three other largescale projects are experimenting with cooperative correction. The first is Distributed Proofreading[7], started under the aegis of Project Gutenberg, which aims to provide accurate OCR copies of out-of-copyright books through the collaboration of volunteers online. From DP’s website:

“During proofreading, volunteers are presented with a scanned page image and the corresponding OCR text on a single web page. This allows the text to be easily compared to the image, proofread, and sent

[5] *Historical Lexicon Building*; 2009; Depuydt, K.; http://www.impact-project.eu/uploads/media/Katrien_Depuydt_Historical_Lexicon_Building.pdf. From the proceedings of The First IMPACT Conference, Den Haag, April 6-7 2009

[6] *Language Technology*; 2009; Schulz, K.; http://www.impact-project.eu/uploads/media/Klaus_Schulz_Language_Technology.pdf. From the proceedings of The First IMPACT Conference, Den Haag, April 6-7 2009. Retrieved 13.03.2011

[7] Distributed Proofreaders: Site Concept; 2009: http://www.pgdp.net/c/ Retrieved 13.03.2011

back to the site. A second volunteer is then presented with the first volunteer's work and the same page image, verifies and corrects the work as necessary, and submits it back to the site. The book then similarly progresses through two formatting rounds using the same web interface.

“Once all the pages have completed these steps, a post-processor carefully assembles them into an e-book, optionally makes it available to interested parties for 'smooth reading', and submits it to the Project Gutenberg archive.”

The second is the National Library of Australia’s Australian Newspaper website, which rather than send copies of an image and its corresponding text to a reader, actually allows users to correct the text direct to the site. According to Rose Holley, manager of the site, some 2,000,000 lines of text (or 100,000 articles) were corrected by volunteers between the site’s launch in August 2008 and January 2009.[8]

The last is the National Library of Finland’s Digitalkoot project, in which the library and its partner Microtask seeks to index the library’s digital holdings by encouraging users to engage with the material through games and gamesplay.[9]

[8] *Many Hands Make Light Work*; 2009; Holley, R.; http://www.impact-project.eu/uploads/media/Rose_Holley_Many_Hands_Make_Light_Work.pdf. From the proceedings of The First IMPACT Project, Den Haag, 6-7 April 2009. Retrieved 13.03.2011

[9] Electrifying our cultural heritage; Digitalkoot; National Library of Finland; 2011; http://www.digitalkoot.fi/en/splash Retrieved 13.03.2011