Deciding whether Optical Character Recognition is feasible

Simon Tanner

King's Digital Consultancy Services

KDCS
Digital Consultancy

Contents

What is OCR **3**
Different uses for OCR **3**
Key issues for whether to use OCR **4**
Accuracy **5**
How to assess a resource for appropriateness for OCR **6**
Scanning methods possible **6**
Nature of original paper **6**
Nature of printing **7**
Nature of document **9**
Nature of output requirements **10**
Conclusions **11**

Deciding whether Optical Character Recognition is feasible

This document will introduce character recognition techniques. The document will illuminate the key factors in designing a text capture method. This document will also provide advice to digitisation projects on how to approach a text capture project and how the various mechanisms of scanning and text capture fit together.

What is OCR

Text capture is a process to convert analogue text based resources into digitally recognisable text resources. These digital text resources can be represented in many ways such as searchable text in indexes to identify documents or page images, or as full text resources. An essential first stage in any text capture process from analogue to digital will be to create a scanned image of the page side. This will provide the base for all other processes. The next stage may then be to use a technology known as Optical Character Recognition to convert the text content into a machine readable format.

Optical Character Recognition (OCR) is a type of document image analysis where a scanned digital image that contains either machine printed or handwritten script is input into an OCR software engine and translating it into an editable machine readable digital text format (like ASCII text).

OCR works by first pre-processing the digital page image into its smallest component parts with layout analysis to find text blocks, sentence/line blocks, word blocks and character blocks. Other features such as lines, graphics, photographs etc are recognised and discarded. The character blocks are then further broken down into components parts, pattern recognized and compared to the OCR engines large dictionary of characters from various fonts and languages. Once a likely match is made then this is recorded and a set of characters in the word block are recognized until all likely characters have been found for the word block. The word is then compared to the OCR engine's large dictionary of complete words that exist for that language.

These factors of characters and words recognised are the key to OCR accuracy – by combining them the OCR engine can deliver much higher levels of accuracy. Modern OCR engines extend this accuracy through more sophisticated pre-processing of source digital images and better algorithms for fuzzy matching, sounds-like matching and grammatical measurements to more accurately establish word accuracy.

Different uses for OCR

There are many uses for the output from an OCR engine and these are not limited to a full text representation online that exactly reproduces the original. Because OCR can, in many circumstances, deliver character recognition accuracy that is below what a good copy typist would achieve it is often assumed it has little validity as a process for many historical documents. However, as long as the process is fitted to the information requirement then OCR can have a place even when the accuracy is relatively low (see Accuracy below for more details).

Potential uses include:

- **Indexing** – the OCR text is output into a pure text file that is then imported to a search engine. The text is used as the basis for full text

searching of the information resource. However, the user never sees the OCR'd text – they are delivered a page image from the scanned document instead. This allows for the OCR accuracy to be quite poor whilst still delivering the document to the user and providing searching capability.

However, this mode of searching just identifies the document not necessarily the word or page on which it appears – in other terms it just indexes that those words appear in a specific item.

An example of this is the BUFVC British Universities Newsreels Database[1].

- **Full text retrieval** – in this mode the OCR text is created as above but further work is done in the delivery system to allow for true full text retrieval. The search results are displayed with hit highlighting within the page image displayed. This is a valuable addition to the indexing option from the perspective of the user.

  An example of this is the Forced Migration Online Digital Library[2].

- **Full text representation** – in this option the OCR'd text is shown to the end user as a representation of the original document. In this case the OCR must be very accurate indeed or the user will lose confidence in the information resource. All sorts of formatting issues in terms of the look and feel of the original are inherent within this option and it is rarely used with mark-up (see below) of some kind. The key factor is the accuracy and this leads to most projects having to check and correct OCR text to ensure the accuracy is suitable for publication with obvious time and cost implications.

- **Full text representation with xml mark-up** - in this option the OCR output is presented to the end user with layout, structure or metadata added via the XML mark-up. In the majority of cases where OCR'd text is to be delivered there will be at least a minimal amount of mark-up done to represent structure or layout. Currently this process normally requires the highest amount of human intervention out of all the options listed here as OCR correction is very likely with additional mark-up of the content in some way.

  Many examples of digital text resources with XML mark-up may be found through the Text Encoding Initiative website[3]. The projects listed there also demonstrate the variety in levels of mark-up that are possible making it possible to vary activity to match the projects intellectual requirements and economic constraints.

Key issues for whether to use OCR

There are several key issues to consider in deciding whether to use OCR at all or choosing between different possible appropriate uses for the text output. The main factors to consider are a combination of accuracy, efficiency and the value gained from the process. If the accuracy is below 98% then considerations of the cost in terms of time and effort to proof read and correct the resource would have to be accounted for if a full text representation is to be made. For instance, see the EEBO production

[1] http://www.bufvc.ac.uk/databases/newsreels/index.html
[2] http://www.forcedmigration.org/
[3] http://www.tei-c.org/Applications/index.html

description for how the accuracy issue changed their potential approaches[4]. If the OCR engine is not capable to delivering the required accuracy then rekeying the text may become viable, but only if the intellectual value to be gained from having the rekeyed text matches the projects goals and budgets. Otherwise, OCR for indexing and retrieval may be the most viable option.

Accuracy

The majority of OCR software suppliers define accuracy in terms of a percentage figure based on the number of correct characters per volume of characters converted. This is very likely to be a misleading figure as it is normally based upon the OCR engine attempting to convert a perfect laser printed text of the modernity and quality of, for instance, the printed version of this document. So, if told that even the better OCR software could get 1 in 10,000 characters wrong and that it will then likely get more than one or two characters wrong in this document would this seem quite so impressive?

It is more useful to know how accurate the OCR engine will be on pre-1950's printed texts of very varying quality in terms of print and paper quality. In this context, it is highly unlikely that we will get 99.99% accuracy and we could assume that even the very best quality printed pre-1950's resources will give no more than 98% (and most would be considerably less than that). In these scenarios the accuracy measure given by the software suppliers is not very useful in deciding whether OCR is appropriate to the original printed resource.

Regarding accuracy as a measurement of the amount of likely activity required to enable the text output to meet the defined requirements would be more useful. In this context we might look at the number of words that are incorrect rather than number of characters.

For example: a page of 500 words with 2,500 characters. If the OCR engine gives a result of 98% accuracy this equals 50 characters incorrect. However, looked at in word terms this could convert to 50 words incorrect (one character per word) and thus in word accuracy terms would equal 90% accuracy. If 25 words are inaccurate (2 characters on average per word) then this gives 95% in word accuracy terms. If 10 words were inaccurate (average of 5 characters per word) then the word accuracy is 98%. In terms of effort and usefulness the word accuracy matters more than the character accuracy – we can see the possibility of 5 times the effort to correct to 100% across the word accuracy range shown in this simple example. It is essential to remember that correcting OCR or text output is relatively expensive in terms of time and effort requiring both correction and proof reading activities – so it best to seek ways to avoid this additional activity if possible.

The other consideration might be the usefulness of the text for indexing and retrieval purposes. If it is possible to achieve 90% character accuracy and still get 90% word accuracy, then most search engines utilising fuzzy logic would get in excess of 98% retrieval rate for straightforward prose text. In these cases the OCR accuracy may be of less interest than the potential retrieval rate for the resource (especially as the user will never see the OCR'd text to notice it isn't perfect). In most prose circumstances significant words and names are repeated which improves even more the chances of retrieval and can enable high retrieval rates for OCR accuracies measuring lower than 90%.

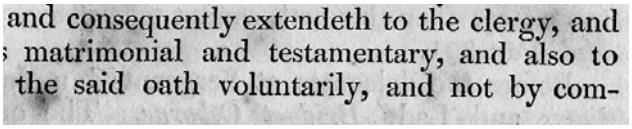
[4] http://www.lib.umich.edu/eebo/proj_des/pd_production.html

How to assess a resource for appropriateness for OCR

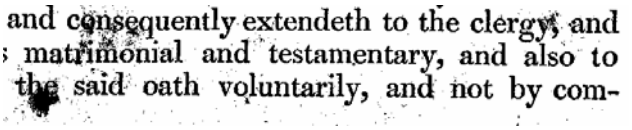
There are a number of key factors to consider when observing a printed resource and assessing whether it will produce the text resource accuracy desired through OCR technologies. Some of the main ones are listed below

Scanning methods possible

Bit-depth is the number one factor that can improve OCR accuracy once a base level of 300 dpi resolution is achieved. If the image can be represented as greyscale (8-bit) or better, then this is more likely to improve the OCR accuracy than almost any other scanning mechanism. So if given the choice of increasing resolution or increasing bit-depth (i.e. from black and white to greyscale) then go for the greyscale. Remember that all OCR engines will struggle to recognise anything well if the resolution is below 300 dpi and that this is the absolute minimum baseline for scanning.


and consequently extendeth to the clergy, and
matrimonial and testamentary, and also to
the said oath voluntarily, and not by com-

Greyscale is clearly readable

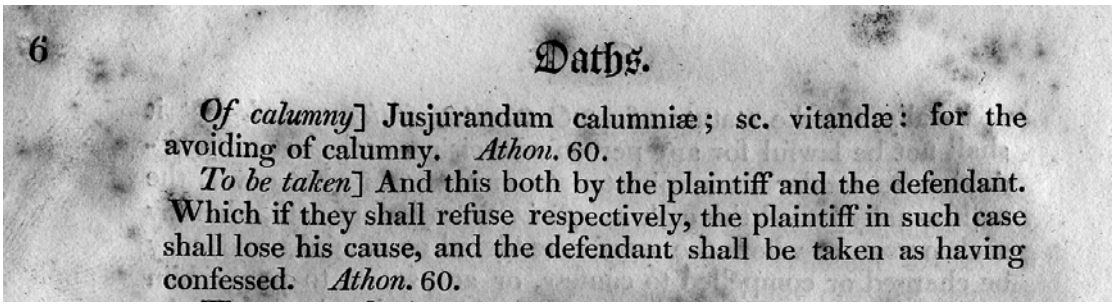

and consequently extendeth to the clergy, and
matrimonial and testamentary, and also to
the said oath voluntarily, and not by com-

Bitonal emphasises foxing and obscures characters in words (*consequently*, *clergy*, *matrimonial* and *the)* that would be captured accurately from the greyscale image.

Thus, the lower the standard of scanned image then the worse the OCR accuracy is likely to be. There are other factors related to the nature of the original that will affect the standard of the scanned image and these should be accounted for as well.

Nature of original paper

The original paper on which text appears is critical to the scanned image standard and is often the core reason why greyscale (8-bit) improves the OCR accuracy. If the paper is dirty, torn, foxed, otherwise marked, has creases or show through then this will all mar the cleanliness of the text characters on the page. If the OCR engine cannot discriminate between the character and the paper background noise then it will be more likely to misrepresent the character. Greyscale images as opposed to B&W give the OCR software a better chance of discriminating between text and noise and thus improve the accuracy.


6 Oaths.

Of calumny] Jusjurandum calumniæ; sc. vitandæ: for the avoiding of calumny. *Athon.* 60.

To be taken] And this both by the plaintiff and the defendant. Which if they shall refuse respectively, the plaintiff in such case shall lose his cause, and the defendant shall be taken as having confessed. *Athon.* 60.

This example shows foxing and also text show through – both of which will mar the OCR accuracy compared to text against a clean white background (like this report).

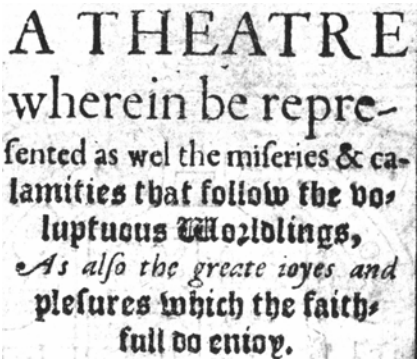
KDCS Digital Consultancy

Nature of printing

The nature of the printed text in the original may make a significant difference to OCR accuracy. Obviously if the text is printed poorly or if it was typed and characters are broken, faded or have indistinct edges then this will affect the ability of an OCR engine to recognise patterns and differentiate between similar shaped characters. So the clarity of the printing is a factor to consider. Print clarity may also have been improved by some fonts over others and also by the use of larger point sizes. Character sizes of below 6 points in the original will limit the accuracy likely to be achieved, although increasing resolution in the scanned image to 600 dpi and using greyscale may improve matters.

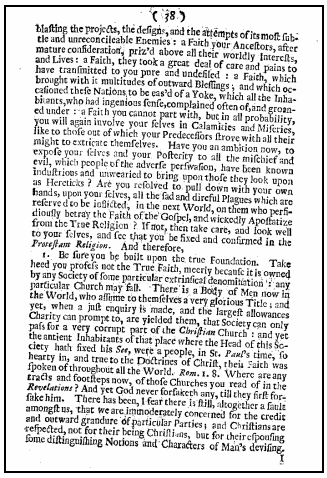
The other features of the printing that will affect quality are uniformity, text alignment and arrangement on the page, graphics and picture content.

- **Uniformity** - The more uniform the text and print quality the more likely that the OCR engine will perform in a predictable manner. Even if the text is poor then as long as it is uniform the OCR engine can be trained to better performance. However, if the text is not uniform or shows wide variation then the OCR engine may struggle and training will not make any significant overall improvement in accuracy.


A THEATRE
wherein be repre-
ſented as wel the miſeries & ca-
lamities that follow the vo-
luptuous Worldlings,
As alſo the greate ioyes and
pleſures which the faith-
full do enioy.

This variation in font and print quality will make OCR more difficult to automate or train.

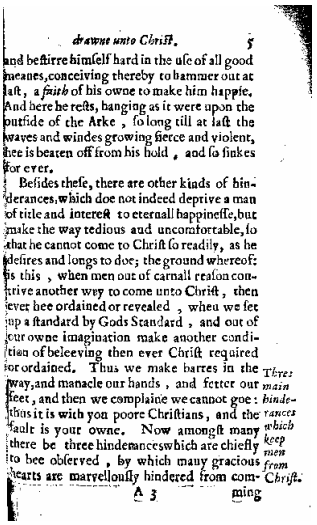
- **Text alignment** – How well is the text aligned with the page? Common text alignment issues are: if the text is at a slant to the page; the lines of text are wavering up and down across the page; dipping at the side of the page where the binding has pulled the page downwards or where columns of text are very close together or out of synch. All of these issues affect the OCR engines ability to analyse the layout of the text and divide the page into component words and characters. Therefore the more variation from everything being aligned at right angles to each other the more work the OCR engine has to do and the higher the likelihood of error.


(38)

blafting the projects, the defigns, and the attempts of its moſt ſubtle and unreconcileable Enemies: a Faith your Anceſtors, after mature conſideration, priz'd above all their worldly Intereſts, and Lives: a Faith, they took a great deal of care and pains to brought with it multitudes of outward Bleſſings; and which occaſioned theſe Nations to be eas'd of a Yoke, which all the Inhabitants, who had ingenious ſenſe, complained often of, and groan-ed under: a Faith you cannot part with, but in all probability, you will again involve your ſelves in Calamities and Miſeries, like to thoſe out of which your Predeceſſors ſtrove with all their might to extricate themſelves. Have you an ambition now, to expoſe your ſelves and your Poſterity to all the miſchief and evil, which people of the adverſe perſwaſion, have been known induſtrious and unwearied to bring upon thoſe they look upon as Heretikes? Are you reſolved to pull down with your own hands, upon your ſelves, all the ſad and direful Plagues which are reſerv'd to be inflicted, in the next World, on them who perfidiouſly betray the Faith of the Goſpel, and wickedly Apoſtatize from the True Religion? If not, then take care, and look well to your ſelves, and ſee that you be fixed and confirmed in the *Proteſtant Religion*. And therefore,

1. Be ſure you be built upon the true Foundation. Take heed you profeſs not the True Faith, meerly becauſe it is owned by any Society of ſome particular extrinſical denomination: any particular Church may fail. There is a Body of Men now in the World, who aſſume to themſelves a very glorious Title; and yet, when a juſt enquiry is made, and the largeſt allowances Charity can prompt to, are yielded them, that Society can only paſs for a very corrupt part of the *Chriſtian* Church: and yet the antient Inhabitants of that place where the Head of this Society hath fixed his See, were a people, in St. *Paul's* time, ſo hearty in, and true to the Doctrines of Chriſt, their Faith was ſpoken of throughout all the World. *Rom.* 1. 8. Where are any tracts and footſteps now, of thoſe Churches you read of in the *Revelations*? And yet God never forſook any, till they firſt forſake him. There has been, I fear there is ſtill, altogether a fault amongſt us, that we are immoderately concerned for the credit and outward grandure of particular Parties; and Chriſtians are reſpected, not for their being Chriſtians, but for their eſpouſing ſome diſtinguiſhing Notions and Characters of Man's deviſing.

Skewed image


drawne unto Chriſt. 5

and beſtirre himſelf hard in the uſe of all good meanes, conceiving thereby to hammer out at laſt, a *faith* of his owne to make him happie. And here he reſts, hanging as it were upon the outſide of the Arke, ſo long till at laſt the waves and windes growing fierce and violent, thee is beaten off from his hold, and ſo ſinkes for ever.

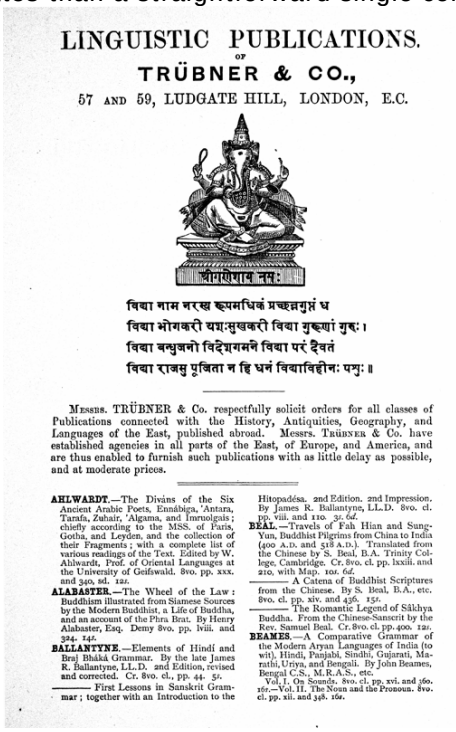
Beſides theſe, there are other kinds of hinderances, which doe not indeed deprive a man of title and intereſt to eternall happineſſe, but make the way tedious and uncomfortable, ſo that he cannot come to Chriſt ſo readily, as he deſires and longs to doe; the ground whereof: is this, when men out of carnall reaſon contrive another way to come unto Chriſt, then ever hee ordained or revealed, when we ſet up a ſtandard by Gods Standard, and out of our owne imagination make another condition of beleeving then ever Chriſt required or ordained. Thus we make barres in the way, and manacle our hands, and fetter our feet, and then we complaine we cannot goe: this is with you poore Chriſtians, and the fault is your owne. Now amongſt many there be three hinderances which are chiefly to bee obſerved, by which many gracious hearts are marvellouſly hindered from coming

A 3

Three main hinderances which keep men from Chriſt.

Skewed image plus binding obscuring and distorting text

- **Complexity of alignment** – Should the text be presented in multiple columns (as in newspapers for instance) or be arranged around pictures or graphics then this may complicate the document analysis and also lead to higher error rates than a straightforward single column alignment.


LINGUISTIC PUBLICATIONS.
OF
TRÜBNER & CO.,
57 AND 59, LUDGATE HILL, LONDON, E.C.

विद्या नाम नरस्य रूपमधिकं प्रच्छन्नगुप्तं धन
विद्या भोगकरी यशःसुखकरी विद्या गुरूणां गुरुः।
विद्या बन्धुजनो विदेशगमने विद्या परं दैवतं
विद्या राजसु पूजिता न हि धनं विद्याविहीनः पशुः॥

MESSRS. TRÜBNER & Co. respectfully solicit orders for all classes of Publications connected with the History, Antiquities, Geography, and Languages of the East, published abroad. Messrs. TRÜBNER & Co. have established agencies in all parts of the East, of Europe, and America, and are thus enabled to furnish such publications with as little delay as possible, and at moderate prices.

AHLWARDT.—The Divâns of the Six Ancient Arabic Poets, Ennâbiga, 'Antara, Tarafa, Zuhair, 'Alqama, and Imrulqais; chiefly according to the MSS. of Paris, Gotha, and Leyden, and the collection of their Fragments; with a complete list of various readings of the Text. Edited by W. Ahlwardt, Prof. of Oriental Languages at the University of Greifswald. 8vo. pp. xxx. and 340. 8d. 12s.

ALABASTER.—The Wheel of the Law: Buddhism illustrated from Siamese Sources by the Modern Buddhist, a Life of Buddha, and an account of the Phra Bat. By Henry Alabaster, Esq. Demy 8vo. pp. lviii. and 324. 14s.

BALLANTYNE.—Elements of Hindí and Braj Bhákhá Grammar. By the late James R. Ballantyne, LL.D. 2nd Edition, revised and corrected. Cr. 8vo. cl., pp. 44. 5s.

——— First Lessons in Sanskrit Grammar; together with an Introduction to the Hitopadésa. 2nd Edition. 2nd Impression. By James R. Ballantyne, LL.D. 8vo. cl. pp. viii. and 110. 3s. 6d.

BEAL.—Travels of Fah Hian and Sung-Yun, Buddhist Pilgrims from China to India (400 A.D. and 518 A.D.). Translated from the Chinese by S. Beal, B.A. Trinity College, Cambridge. Cr. 8vo. cl. pp. lxxiii. and 210, with Map. 10s. 6d.

——— A Catena of Buddhist Scriptures from the Chinese. By S. Beal, B.A., etc. 8vo. cl. pp. xiv. and 436. 15s.

——— The Romantic Legend of Sâkhya Buddha. From the Chinese-Sanscrit by the Rev. Samuel Beal. Cr. 8vo. cl. pp. 400. 12s.

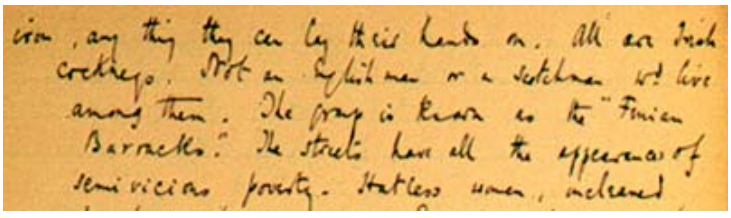
BEAMES.—A Comparative Grammar of the Modern Aryan Languages of India (to wit), Hindi, Panjabi, Sindhi, Gujarati, Marathi, Uriya, and Bengali. By John Beames, Bengal C.S., M.R.A.S., etc.
Vol. I. On Sounds. 8vo. cl. pp. xvi. and 360. 16s.—Vol. II. The Noun and the Pronoun. 8vo. cl. pp. xii. and 348. 16s.

- **Lines, graphics and pictures** – The presence of lines or graphic elements including photographic pictures can lead to OCR errors without human intervention. Lines (both horizontal and vertical) within the text can become noise that is recognised as vagrant punctuation. Graphics and photographs may confuse layout analysis or may even contain text that is not required for the output (car registration in picture for instance) and thus deliver a seemingly spurious result.

In this image the inclusion of picture content, the complexity of the text's interaction with the music and the horizontal lines all conspire to make this image very difficult to OCR with any accuracy. By far the most cost effective and accurate way to capture this text would be to rekey it.

- **Handwriting** – Handwriting may be character recognised only under very controlled circumstances and not usually as an automated activity. Usually handwriting recognition is limited to handwriting in Forms designed specifically for the purpose (e.g. application forms, tax forms).

  If approaching a printed text document with handwritten notes or marginalia then the handwritten is unlikely to be accurately recognised and thus adds to the overall noise and inaccuracies in text output.



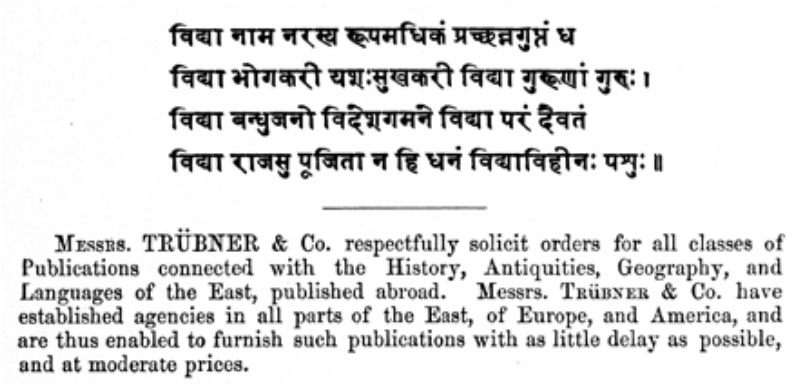
This example is highly unlikely to OCR with any accuracy and is not cost effective to attempt.

Nature of document

Once considering all the physical characteristics of the printed page then it is important remember the content related issues for OCR. OCR engines work by having as many examples as possible of character shapes from as many different language and national printing traditions as possible. They are supported by extensive

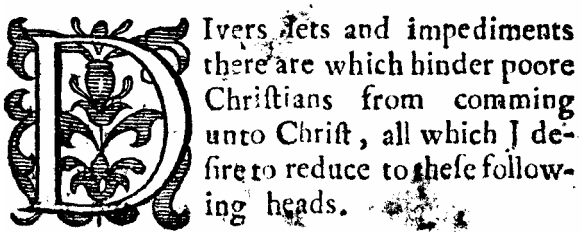
dictionaries and vocabularies to enable natural language matching of grouped characters against known words to statistically improve the accuracy of the OCR output.

Therefore OCR engines may be fooled by alphanumeric content (such as part number for an engine e.g. MK2-3) or by personal names and place names that are unusual as they may not be in the engines dictionary of acceptable words. Further to this issue symbols, long s, diacritics and mathematical notation may play havoc with an OCR engines accuracy levels. Text content from pre-1900 may contain words that are no longer in common usage and a number of languages have gone through significant changes. Thus words may not be found in the OCR engines dictionary.


विद्या नाम नरस्य रूपमधिकं प्रच्छन्नगुप्तं धन
विद्या भोगकरी यशःसुखकरी विद्या गुरूणां गुरुः।
विद्या बन्धुजनो विदेशगमने विद्या परं दैवतं
विद्या राजसु पूजिता न हि धनं विद्याविहीनः पशुः॥

MESSRS. TRÜBNER & Co. respectfully solicit orders for all classes of Publications connected with the History, Antiquities, Geography, and Languages of the East, published abroad. Messrs. TRÜBNER & Co. have established agencies in all parts of the East, of Europe, and America, and are thus enabled to furnish such publications with as little delay as possible, and at moderate prices.

In this example, the Asian text would not be recognised by an OCR engine unless it had the specific dictionary engaged and also was told what language it was. Even then this form of text is usually not amenable to OCR.


Ivers lets and impediments there are which hinder poore Chriſtians from comming unto Chriſt, all which I deſire to reduce to theſe following heads.

In this example, the long s character will cause difficulties for OCR engines (see *Christians, desire* etc). Plus the use of words with alternate spellings to those used in modern language (*comming, poore*) will also diminish the OCR engines capacity to recognise words.

Consideration of the content and whether this is supported by the OCR engines word and font dictionaries is very worthwhile. Word dictionaries can always be added to in an effort to improve accuracy (such as by adding surname lists or street name content).

Nature of output requirements

What is needed from the OCR output will define and constrain the effectiveness of the OCR process. Output of text to an ASCII text file with just a list of words is very straightforward and can be heavily automated. Output to a PDF file can also be automated. But once the output has to represent the layout and structure of the original page then this creates technical difficulties that are usually only resolved by human intervention and therefore be more costly in time and effort. XML mark-up covers such a range of possibilities (from just marking up the pagination to marking up the meaning of a single character) that this can be sometimes automated but is usually human mediated and thus has various cost factors associated with it.

Conclusions

Looking at original printed resources to assess whether OCR will be an effective process requires a combination of factors to be compared. The key issue will be the likely accuracy of the OCR engine in producing text output appropriate to the project requirements. If the accuracy is not in an acceptable range for the desired purpose then to improve this will require human intervention – particularly OCR engine training, correction of text and proofreading. Once human intervention of any significance is introduced then the time, effort and thus project cost increases very quickly.

The factors to compare are:

- Scanning – the standard of the page image
- The original – its nature and condition
- Nature of the printing
- Content
- Output requirements

Once these are considered in relation to the intellectual value desired from the OCR'd text then decisions regarding project budgets and the effort required may be made. It is worth remembering that there are always a number of ways to deliver text to the user and that OCR is just one of the tools along the path to achieve this and so if it is not deemed the best method there will be other pathways to the project goals.