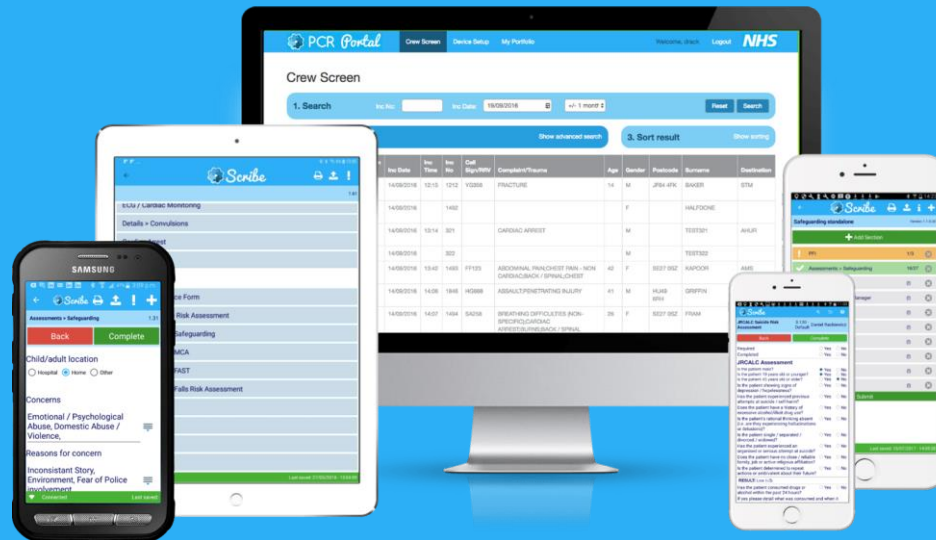


# Ceri Jones, 27th April 2021



## DOC-WORKS

*Understanding OCR*

# Introduction

# Aims


<b>Audience describes themselves as having...</b>	<b>What we hope you will gain from this discussion:</b>
... no idea what OCR is!	Be able to hold discussions where you can be clear on what OCR is and how it does or does not fit into your business
... some knowledge of OCR but not sure if it has relevance to their business	Start thinking and planning for changes within your business that might involve OCR or Digitisation
... experience with using it but did not find it worked out as planned	Know what has changed and what remains the same in the OCR technology space; hopefully giving some food for thought for how it might go to plan this time round!
... a specific challenge they think might be solved by using OCR today!	Have a starting point for the kind of internal discussions and planning required to achieve your objectives and the sorts of people who you should reach out to.

# Doc-works and OCR

- Founded in 2004 to be a *technology workshop* for **paper** documents, providing a route to digital information
- Over the last 17 years we've seen paper still growing year on year but less and less is **primary**
- Shift of focus to **pre-digitisation never-paper**

# Digitisation and OCR

## digitization

/dɪdʒɪtaɪ'zeɪʃ(ə)n/ 

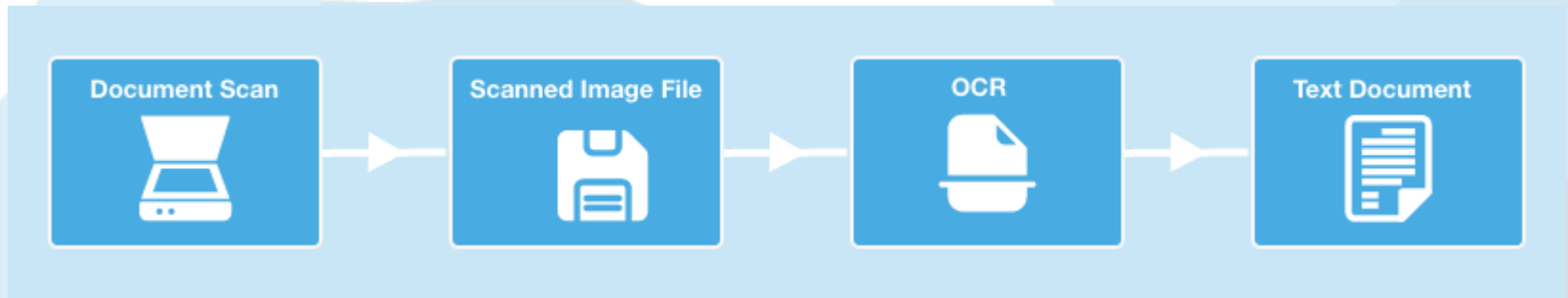
*noun*

1. the conversion of text, pictures, or sound into a digital form that can be processed by a computer:

Powered by [Oxford Dictionaries](#)

- OCR, Optical Character Recognition is an **example** of digitisation
- It specifically handles (scanned) images of handwritten or typed documents, extracting the data from them

# Data Extraction and OCR



- Where a document is in paper form there needs to be a “capture” event, usually through scanning\*
- Data Extraction occurs once the text of the document has been OCR'd

\* Scanning through simple multi-functional-devices (MFDs) all the way up to huge IBM scanners

# Where should the CFO be involved?

- A CFO should

## **At a minimum:**

Know the requirements of OCR\* within the finance function (e.g. implementing processes and systems around Accounts Payable)

## **Ideally:**

Work with the CTO to establish an holistic Digital Strategy, aiming to keep documents digital throughout

## **Gold Standard**

Further to the above, establish financial controls over the implementation of the digital strategy, avoid death by a thousand Apps and ensure business benefits are achieved

\* By OCR, please also include Data Extraction and more broadly keeping information digital throughout the business including the supply-chain

# Scanning & Indexing



# Outsourcing Scanning

- Businesses\* specialise in picking up, boxing, transporting to their facilities and scanning (then indexing) the documents collected (many offer secure shredding as well).
- Initially they tend to scan historic documents (“Back Scanning”) but usually offer a service to onwardly collect and scan new documents based on an agreed collection date.

# In-Sourcing Scanning

- Businesses also specialise in selling / leasing equipment and then providing the software and consultancy required for you to do your own scanning; in very specific cases this is a viable option\*
- For small volumes, businesses can use small inexpensive devices to scan the occasional paper invoice for example.

\* For example where your business is inevitably handling large volumes on inbound paper

# OCR and Validation

- OCR even with the most advanced technology will rarely return with 100% **confidence**, for that reason the data returned from the OCR engine will often be reviewed by human operators, a process known as **Validation**
- This human element is often missed from a cursory glance and is both important in terms of accuracy and adds a significant cost

# Indexing and Classification

- The purpose of data extraction is to either
  - Index a document so its scanned image can be found again (via an EDMS\*)
  - Push the data to another system (e.g. Accounts Payable, CRM etc)
  - Make available for reporting purposes (e.g. how many questionnaires said strongly agree?)
  - Classify the document so that it can be processed appropriately (e.g. sent to the right department)

\* Electronic Document Management System, sometimes simply “Document Management System”

# Purchase Invoice Processing

# Supplier based Templates

In many traditional implementations of OCR in an Accounts Payable setting, the consultants will set up a template (normally by supplier) of the types of invoices and where certain data can be found on the “page”; they will leave you with the tools and skills to keep these up to date and the “teach” the system new templates as they come along.

Pros	Cons
Should get a good read from supplier invoices which you have properly templated	Templates can take quite a bit of work and the skills required may not be already in your AP team - consultants are expensive
Tried and tested technology	Potential to miss out on newer technologies like <a href="https://aluma.io">aluma.io</a> which use a centralised repository of learned “templates”
Staff gain template management skills and reduce reliance on supplier	When staff leave they often take those skills with them

# Using associated data

- Keeping to the theme of Purchase Invoices, an increasingly attractive option is to send invoices along with an associated data file (many online accounts packages favoured by SMEs do this now)
- Often this comes in to the accounts payable department as an Excel or CSV file on the same email as the invoice itself
- A well implemented Accounts Payable system will ingest these automatically and not just load the PDF into the workflow but also load its associated data file (removing the need to extract the data at all)

# Invoice Processing - Line Level

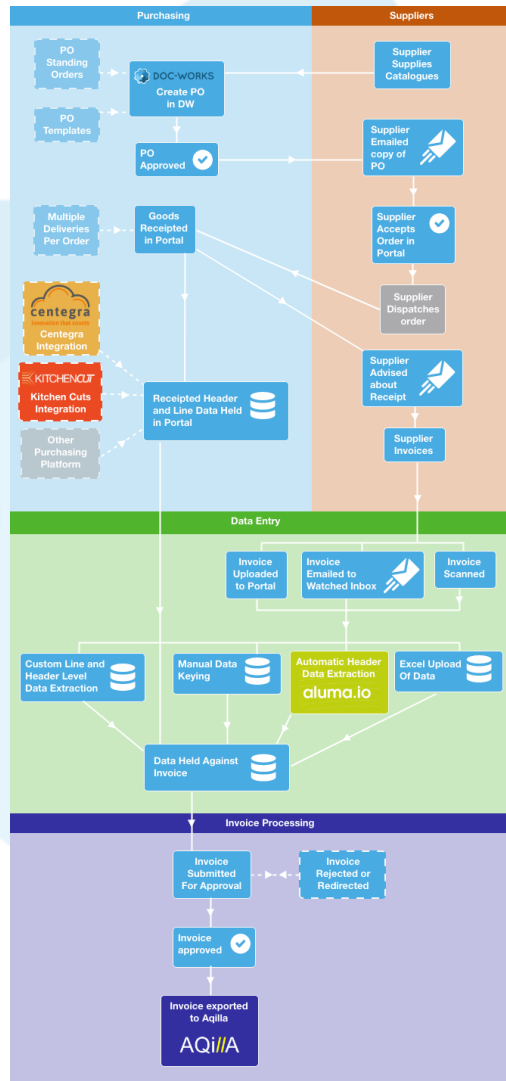
- Line Level data extraction is seen by many to be unachievable, here is why
  - Page breaks
  - Wild differences in format, even from one supplier or even page to page
  - Requirement to process some lines, ignore some, aggregate others
- What can be done?
  - Don't do it unless you need to, "we extract at line level" does not make sense if only 5% of all invoices really need line level extraction
  - Use supporting information, POs where available or associated substantiating Excel/CSV etc where not



# Using Purchase Orders

- Purchase Order systems are not just a means of controlling purchasing, they are also a means of vastly improving the Accounts Payable process
  - Once a budget holder has pre-approved a purchase order, they should not be asked to approve it again once the invoice arrives and is matched to that PO
  - If the purchase order is also “receipted” - marked as having the goods or services fully and satisfactorily delivered - then any invoice for that order need not be returned to the purchasing manager to confirm it can be paid
  - Lastly, the invoice itself can be immediately pre-loaded with both header and line level data based on the data held on the Purchase Order, avoiding data entry or OCR/Validation

# Complex AP implementations



To the left we have an illustration of a complex Accounts Payable implementation, points to note:

- The OCR portion of this is the single olive green box in the light green section (and even so this box also represents determining where from the OCR'd text to look for the fields we need)
- The top left section is all about Purchase Orders and Receipts (in this example 3 different purchase ordering systems were used as well as supporting invoices which would not have POs at all)
- The top right section is the functionality available to Suppliers, here they are encouraged to upload their invoices and match them to one of the approved Purchase Orders as well as keep any pricing matrices up to date
- The light green section (apart from containing the requirement for OCR) also shows that that associated data can be loaded via a watched email inbox (as in the previous slide)
- The light purple section at the bottom, explores the approval process based on the rules provided and the eventual seamless upload to the accounting system (in this case Aqilla)

# OCR in other settings

# When paper just is the way!

- Some business HAVE to deal with paper, it can even be core to their activities
  - One example is where patient records are still required because either electronic solutions are still in progress or they were not available at the time and paper was a fallback
  - Documents are either returned to a central location or scanned at the department in which they were created
  - Once the scanned image is available the system OCRs it - in this case using specialised OCR which is designed for handwriting
  - The data extracted from OCR is validated on just a handful of fields so that it can be matched to the underlying patient systems (and enriched), some of these fields determine whether other fields need validating (secondary validation)

# OCR for classification

- As mentioned before OCR can be very useful for classifying documents, i.e. what kind of document is this
  - One example is where a customer had 10s of thousands of documents scanned given just a document date, they were asked to remove anything older than 7 years if type X and 10 years if type Y... the trouble is they could not tell
  - They turned to OCR (specifically aluma.io) which quickly provisionally classified the documents and we provided the customer with samples of the provisionally classified documents, went through a few iterations of “teaching” and then successfully classified well over 95% of the documents
  - The last 5% we presented back in a manual image classifier screen (image on the right, a panel on the left to quickly decide on its classification)

# OMR - Optical Mark Recognition

- Especially where handwriting is concerned the wisest choice to track important fields as **MARKS** instead of **WORDS**
  - One example is where exam papers have multiple choice answers - here we look for a mark in box A, B, C, D or E
  - A lot of this is based on good design, don't have fields which request the person filling in the form to write Y or N in a Yes No option, give them Y and N boxes to colour in
  - Widely used now on paper are 1 and 2 dimensional barcodes, these are amazing for safely keeping data in a machine readable format, even if subsequently printed and scanned
    - The most common formats of 2d barcodes are QR codes (that you scan with a phone to link to a website and mailmarks at the top of letters)
    - 1d barcodes are older and you will have seen them on products you buy at shops for decades and are used for scanning through the tills



# Beyond OCR

# Machine Readable PDFs

- Where documents arrive from suppliers as PDFs, these are often machine readable (open them and see if you can search for a word or even double click on a word and see if it **highlights**)
  - Machine readable PDFs do not need to be OCR'd - they not only already contain nice readable text but you can 100% rely on that text
  - You do still need extract the data you need and that is not as simple as it sounds, for example a supplier rarely puts their name on an invoice, they tend to use logos which are essentially useless to the computer
  - Just because the information is in readable text does not mean to say we can immediately work out where the invoice number (it might be that this appears without even a label at the top right of the page) and line level (see previous slide) can still be a major challenge.



# M/L, NLP and AI

- OCR is often mentioned alongside tricky little acronyms like:
  - AI or Artificial Intelligence is a broad term used to describe machines behaving with human-like intelligence
  - M/L or Machine Learning is really an *application* of AI; here we ask the machine to achieve ever improving results by letting it try its task (for example OCR) and then let it know if that was successful, the machine then can alter its approach and if that gets better results it will keep going along that track
  - NLP or Natural Language Processing is a specific branch of AI which can be used to examine text taken from OCR (just in the same way as when you say: Hey Siri / Alexa / Google) and tries to handle the input as if it was being “understood”

# Keeping things Digital *throughout*

*As amazing as OCR is, the requirement for it is generally an indicator that our modern plans to pass information around seamlessly have come to a shuddering halt*

Whether it is because of a supplier printing out an invoice and popping it in the post, or even a member of your own staff sending a PDF by email onto the next department instead of progressing it through your own systems... in the end something has fallen off the Digital Highway

Work with your departments (and your customers and suppliers) to establish a continuous flow of data, this will often require process re-engineering as well as software engineering

# Going Paperless / Less Paper

*The goal of modern business is to remove paper (or more accurately non-Digital information)*

Look at all the places within the organisation where information enters or leaves a department non-Digitally; list and prioritise these for making them digital from the start. Ask:

- What are the volumes?
- Does the non-digital format impact the next department (whether it is in your business or a supplier/customer)?
- Is the information required for any reports?
- Is information linked to any regulatory/compliance concerns?
- What is the effort required to make this digital?
- **Am I happy for this to remain as paper?**

# preDigitisation

*Digitisation is now commonly used to describe pre-Digitisation: documents being digital from the start*

What are the methods you can use to achieve that?

- Use existing software where possible, it might be that you have not used some functionality available to you or that you are not adequately licensed, so a user is choosing to output to Excel/PDF etc to pass information to another department (speak to your supplier but understand their advice is based on their interests)
- Use data capture forms to replace paper or Excel forms that are being completed, ensure you use a centralised approach to all departments and make sure it is full integrated and flexible
- With suppliers and customers ensure you are aware of data interchange options, APIs are now commonly available

# Wrap-up

# Myth Busting

<b>Myth</b>	<b>Fact</b>
OCR is a magic wand that will turn your paper (or non-digital) records into clean, accurate data rich documents	OCR is an essential tool for extracting information from paper or scanned non-text images
To handle a build up of paper documents you need to buy scanners and get scanning	In many cases paper need only be stored for a period then shredded, not everything needs to be scanned and third parties can help you with this.
You need to OCR documents that arrive by email, for example PDFs of Purchase Invoices.	Most suppliers send PDF invoices which are already machine readable - i.e. the text can be extracted without OCR
I need to index everything on my documents in order to get the benefit of OCR and data extraction	A smart approach is to extract just the information required to match to another data source that you already have - for example Payroll Number would give you Employee Name, Address etc via lookup!

# Do's and Don'ts

Do's	Don'ts
<p>Speak to suppliers across a range of technology and services; many will be happy to give you advice before they start charging you consultancy!</p>	<p>Be led by a single supplier - even from a large reputable firm; it may well be that they come and fix a problem you do not have: <b>to a hammer everything looks like a nail</b></p>
<p>Consider OCR/Digitisation holistically; discuss your overall digital strategy both throughout all your own departments and beyond into the supply chain.</p>	<p>Work in isolation, solve a single instance of where information is “trapped” in a form that you cannot get at (e.g. paper)</p>
<p>Change your processes (and attempt to change those of your suppliers and customers) to reduce the requirement for complex data extraction using methods like OCR.</p>	<p>Print out document that you received as PDFs or Word/Excel etc! Don't add to the burden of OCR with avoidable print outs.  <i>Be aware of where paper is <b>primary</b> or <b>secondary</b></i></p>
<p>Look to enrich information from systems instead of OCR / indexing all required fields</p>	<p>Extract information you don't need.</p>

# Current Climate

- With a large number of people working from home, this has pushed forward the Digital Strategies of many businesses (especially in the public sector)

*This creates a pressure on the private sector to catch up in order to remain technically relevant*





Q&A