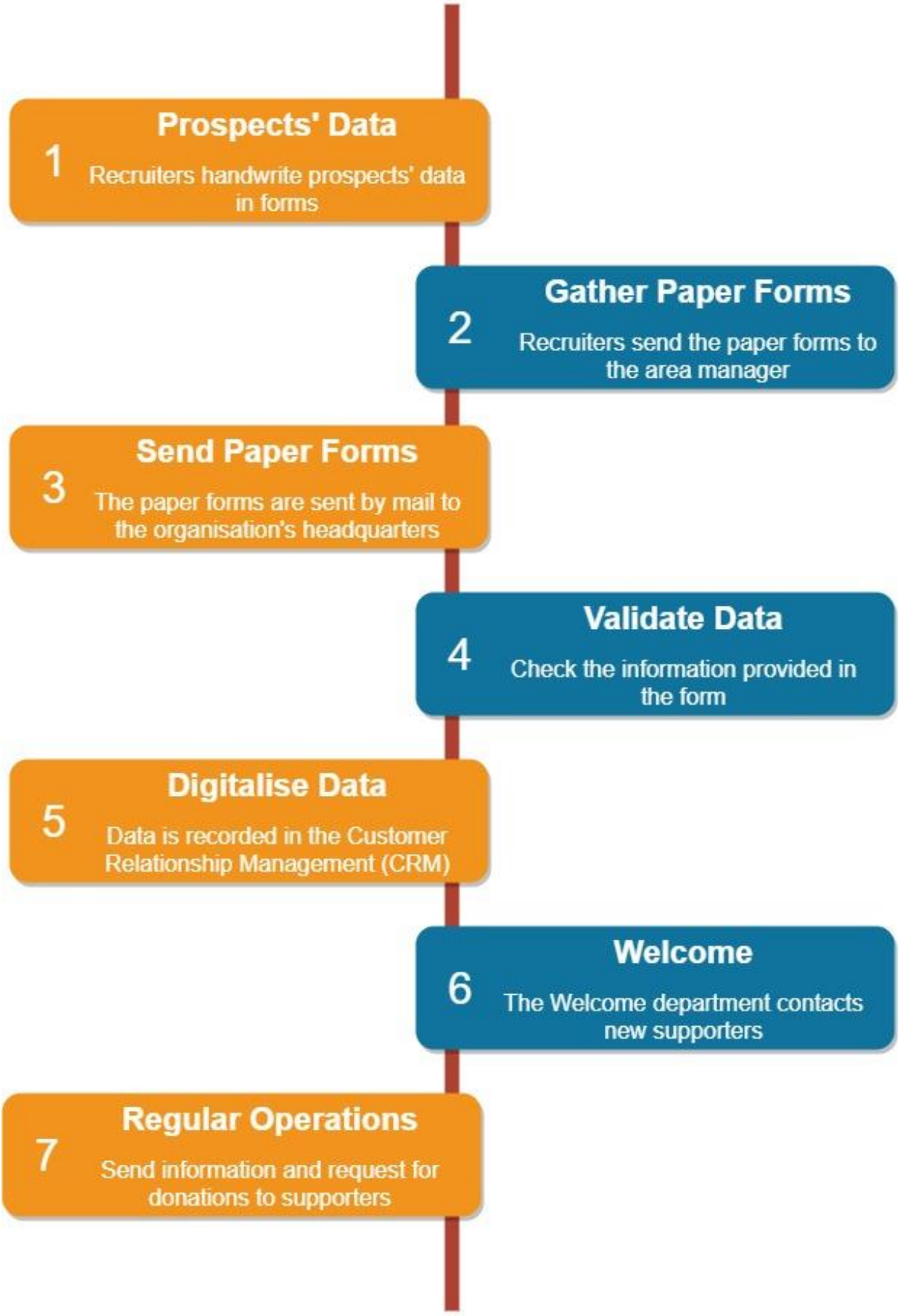
From Text to Digital

Introduction

I collaborate with an organisation which employs many recruiters to contact prospects face to face. Recruiters handwrite prospects' information in a form.

The area managers send the handwritten forms to this organisation's headquarters, and other employees read the questionnaire and manually record the information in the Customer Relationship Management (CRM) system for further use.


1
Prospects' Data
Recruiters handwrite prospects' data in forms
2
Gather Paper Forms
Recruiters send the paper forms to the area manager
3
Send Paper Forms
The paper forms are sent by mail to the organisation's headquarters
4
Validate Data
Check the information provided in the form
5
Digitalise Data
Data is recorded in the Customer Relationship Management (CRM)
6
Welcome
The Welcome department contacts new supporters
7
Regular Operations
Send information and request for donations to supporters

It is of paramount importance for this organisation to have the prospects' information digitalised as soon as possible. However, it currently takes a week since they get the information until the data is recorded in the CRM.

They have thought of using tablets with 4G connection to speed up the process. However, for fear that this solution implies a high risk of equipment loss or damage, and it is expensive, they want to consider other possibilities.

In this post, I propose a solution for this organisation to choose and minimise the time for the information to be available by improving steps 4 and 5 in the diagram above. I also explain some areas that need to be discussed internally to find the best solution for every organisation.

The final solution is meant to accelerate and optimise the work of the human teams responsible for data validation and digitalisation. It will reduce the amount of repetitive work and allow the team members to dedicate more time to business-critical tasks where their intelligence and skills are more valuable. This solution is not meant to replace those teams nor their members.

As the organisation becomes savvier in data capture automation, this solution will evolve and be optimised to increase efficiency.

Scenario

Form Description

The forms contain the next main data blocks:

1. Recruiter information,
2. Supporter's personal information,
3. Donation,
4. Payment method and details,
5. Signature, and
6. Free text space for additional information.

Additionally, the lateral margins are in different colours. Each colour represents the geographical location of the organisation's office responsible for the recruiter.

Data Validation

Before the data is digitalised and recorded in the CRM, fact-checkers must validate some information:

1. Bank accounts or credit card numbers are valid,
2. The supporter is over 23 years old, and
3. The form is signed.

If the form is not valid, it is not digitalised. The fact-checker contacts the recruiter. Then the recruiter must contact the prospect and try to get the correct or missing data.

Textual Data Capture

Data Capture consists of any method of collecting information and then changing it into a form that can be read and used by a computer[1].

There are several methods to capture data from unstructured documents[2]. I describe in this section the ones which are relevant to this scenario.

OCR (Optical Character Recognition) is a technology which converts scanned or photographed images of machine-printed characters into electronic information for processing. For example, OCR is used for invoice capture.

ICR (Intelligent Character Recognition) is the computer translation of hand-printed and written characters. For example, data that is hand-written on forms. Then the forms are scanned, and the image of the captured data is then analysed and translated by sophisticated ICR software to electronic information.

Most ICR platforms are self-learning systems that allow users to add new characters or representations to the recognition database. This procedure increases accuracy as the system is used. Given variances in handwriting styles, capture rates can vary dramatically. However, you can expect a capture rate of 95+% if you are capturing from a structured form.

OCR is often used to mean textual data capture from hand-written documents as well as machine-printed ones. In this post, I will use the term OCR in this generic way.

Finally, **bar code recognition** allows increasing the effectiveness of the data capture process by adding relevant information about the form per se. For example, we can use bar codes in this organisation to replace the colour code in the forms which represent the office and automate the recognition of this information.

Barcodes contain information which has been generated using a barcode font to provide a symbolic and encoded representation of a number, text, or a series of numbers and text which can be decoded by computers, optical scanning devices and apps on smartphones.

There are many different barcode types[3]. The best barcode for this organisation is GS1-128, which is very reliable . It has a check code that prevents misreading.

---

[1] Cambridge Dictionary – Data Capture.

[2] ProcessFlows – Methods of data capture

[3] GS1 Barcodes

Amendment November 23rd, 2019: *My friend Juan Antonio Martínez Rodríguez-Bermejo has extensive experience in barcode implementations. He has reviewed this design, correct my recommendation on the barcode for this project, and recommended GS1-128. Thanks, Juan Antonio.*

Barcodes are often used to separate batches of paper documents for image scanning. In this scenario, barcodes attached to separator pages contain only general information. Adding barcode pages adds a negligible amount to the document preparation time, but it saves time further down the line as users can load batches of documents into the scanner at the same time.

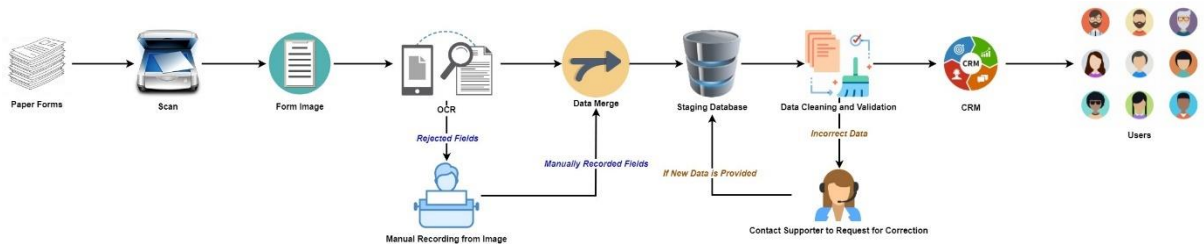
If we want to use the images in our automated document management process, barcodes can facilitate the automatic indexing of our scanned documents into an electronic document repository.

As part of the barcode project planning, we must choose and match the barcode type with the device we are going to use to decode the barcode information – a computer, a hand-held scanner or with a smartphone or tablet.

We will also need to decide how many barcodes we need, how much and what type of information we want it to contain, and what we want to achieve with it.

Solution

The diagram below shows the Data Capture system and workflow I propose to this organisation.



Scanner

We need a scanner for this Data Capture system. A flat-bed scanner is cheaper, and it should give good results, though it is time-consuming to scan many pages. This organisation needs to do much scanning, so a sheet-fed scanner is going to be more efficient.

Additionally, some scanners include OCR software. They can also scan both sides of the paper at once. The scanner's OCR software usually runs on a PC.

OCR Software

If we don't want to purchase a scan with OCR software, either because of the price, the functionality or the architecture of our solution, there are several options. This article analyses different possibilities to help us make a decision.

On a separate note, many OCR applications are based on the Tesseract OCR engine. This engine was initially developed by Hewlett-Packard in England in the 1980s. HP made it open source in 2005, and Google now maintains the source code. We can also use Tesseract OCR engine for our solution.

Wikipedia warns that "Tesseract's output will be very poor quality if the input images are not preprocessed to suit it: Images (especially screenshots) must be scaled up such that the text x-height is at least 20 pixels, any rotation or skew must be corrected or no text will be recognized, low-frequency changes in brightness must be high-pass filtered, or Tesseract's binarization stage will destroy much of the page, and dark borders must be manually removed, or they will be misinterpreted as characters"[4].

Staging Database and Loading Data into the CRM

The fields in our form are well defined, and we need to validate the data before importing it to our CRM. For this reason, I propose to use a PostgreSQL database, an open-source relational database, as a Staging Database.

We will extract data from our Staging Database into a CSV file and loaded into Salesforce, the corporative CRM.

Define Retention Periods

Storage has a price, and data represent a profit. It is necessary to define before the project starts the retention periods for:

- Form images,
- Records in the database,
- Information in the CRM, and
- Backup strategy for all the above.

Likewise, we must define the procedures to purge data according to our retention periods.

Good Practices

The solution proposed in this post will be more efficient if the next good practices are followed. Increasing efficiency will reduce the number of manual, repetitive tasks team members will need to perform.

- Determine if the document or form can be updated to improve the capture/recognition process and method.
    - Structured forms. E.g., write information in the right box, do not write outside boxes, etc.

[4] Wikipedia - Tesseract (software)

    - Add/ remove/ modify fields.
    - Paper is thick enough, and data from one side is not visible on the other side.
- Investigation of the existing line of business systems, to determine what additional metadata can be extracted, and add it to the bar code.
- Information in the forms is highly sensitive. Digital Rights Management (DRM) is commonly used to encrypt PDF files with different encryption algorithms. Our provider must offer point-of-entry encryption (meaning the file is encrypted as it exits the scanning application) and that no unprotected temporary files remain.
- To improve usability and increase the accuracy of OCR and other recognition technologies. Image enhancement is required in the textual data capture solution. Typical image enhancement might include deskew[5], despeckle[6] and rotate functions. Brilliant capture should also include options to remove blank pages, remove separator sheets, autorotate, remove lines, and adaptive thresholding. Adaptive thresholding technology assists in cleaning "dirty" documents or documents that have a coloured background which interferes with the foreground data.
- Statistical sampling of automatic data captures to monitor accuracy.
- Define testing techniques appropriate for the teams involved.

Measure OCR Quality

Depending on the confidence the OCR has in a write-in field, it is trying to read, it either accepts the inferred result or rejects it. Rejected field snippets are sent to human keyers, who attempt to manually type the correct answer to the field snippet. The term "Reject Rate" is used to measure the fraction of fields input to the OCR that is not read automatically.

The fields typed manually are then merged with the accepted fields and stored in a database until the next batch to load them in the CRM. There are always errors in the final merged data fields, either from the OCR accepted fields, or from the human keyers. The errors in the accepted OCR fields are measured by a quantity called OCR "Error Rate".

Usually, the quality of OCR systems is measured by plotting OCR Error Rate versus Reject Rate.

Return of Investment

To know the cost per page of our solution and determine the Return of the Investment (ROI), we have to add up the scanner, OCR software and working hours price, and divide it by the number of pages.

---

[5] The process of straightening an image that has been scanned or photographed crookedly — that is an image that is slanting too far in one direction, or one that is misaligned. This process is done in the post-production stage using graphics software. Webopedia – image de-skew

[6] To remove speckles from. www.yourdictionary.com – despeckle

$$Cost\ per\ Page = \frac{[Scanner] + [OCR\ sofware] + [Working\ hours]}{Number\ of\ Pages}$$

Additionally, we must include the cost of storage and database licensing as part of the total cost of the project to evaluate the ROI of our solution.