Biotech Batch Record Digitization: Challenges and Role of AI – Part 3

The following two images show a sample result of image processing techniques applied which resulted in the improvement of Tesseract OCR performance.

A sample of a handwritten section, before and after image processing

Tesseract OCR's performance was measured across all the handwritten sections in the entire 5 batch records resulting in 15 PDFs of varying sizes. The average performance of the OCR on the entire documents set was observed to be 34.35% before training, while after training the Tesseract OCR results improved to 47.1%. Using OCRs which could utmost provide maximum 50% accuracy on handwritten texts cannot be used for developing an automated system.

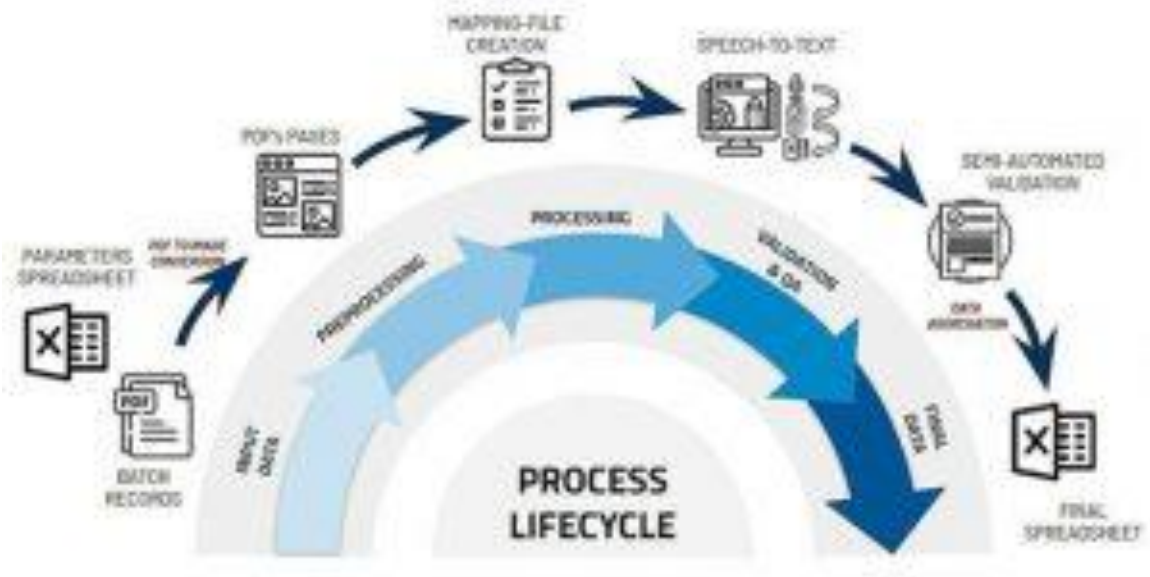
The OCR technology has yet to reach its full maturity. There are multiple instances of the document and handwritten data variations observed in our daily activities. No doubt that the OCR technology has seen tremendous improvements in the past 5 years, but it is yet to become independent of variations like different handwritings, font sizes, shades of ink, image noise and associated garbage values, bordered and borderless tables, etc. It is due to these limitations of the OCR technology, that a completely or partially automated solution cannot be developed.

Speech-to-Text

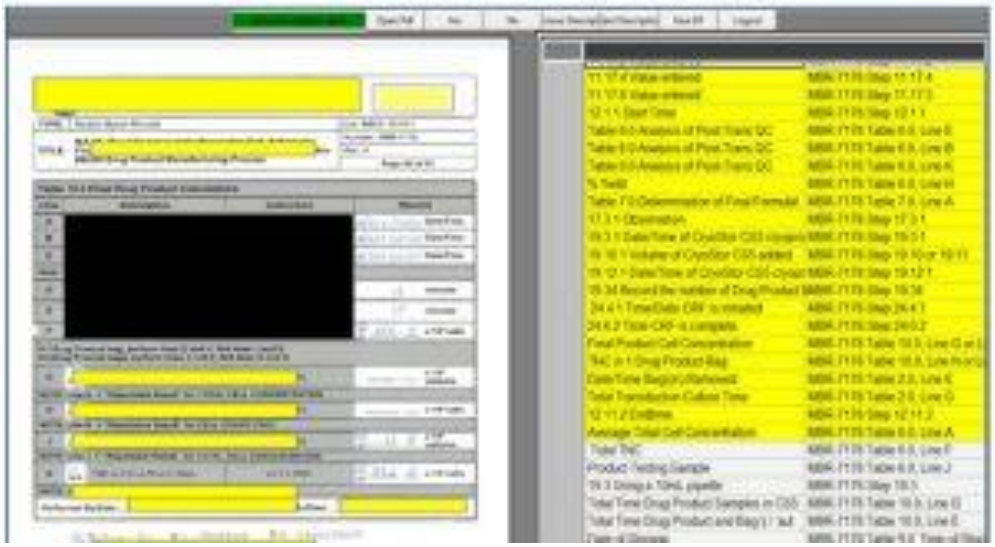
The speech-to-text conversion software is matured enough and can produce results with 90-95% accuracy in the achieved text outputs. The performance can be further improved up to 100% if the speech recording is made in a silent environment and the

pace of speech and accent is maintained. The Google Cloud Speech service was initially used but was found to be slowing down the process of data extraction, as it needs the audio to be first recorded and saved in a file, which is then sent to the cloud service and it returns a text output. The text output is provided to the data entry system, which takes care of the text being allocated to the correct parameter. The audio file needs to be recorded for each data point, which makes the entire operation of the data fill-up process very slow.

As an alternative, Speech-to-text software running locally can be used. The workflow is created to integrate the speech-to-text method into data extraction and data entry software. The following diagram denotes the workflow of the software for new batch-record processing.



The speech-to-text approach is integrated into software that can take care of the remaining workflow for data entry into a database format. The input to the workflow is the PDF and a list of parameters that need to be extracted from PDF. The pages in the PDF are converted into images for ease of processing. A mapping file is created which can map the parameters found in the PDFs to the associated page within the PDF. The pages of interest are brought to the user for each parameter so that the user can start reading the information on the page for the parameter. The speech-to-text software running in the background keeps listening to the voice and starts creating the equivalent text in a text file. Once all the parameters are recorded, the text file can be processed within the software to automatically fill-up the information for each parameter in a structured format. A validation step is also introduced within the software so as to allow the user to check the processed key-value pairs of information and correct the data if any errors are detected.



The results for time study are recorded over multiple iterations for all the five batch records. On an average, around 120 minutes were needed for each batch record, containing 3 PDFs each, which consisted of around 200 pages per batch and around 150 parameters to be extracted from those PDFs per batch. While an initial setup time of 5 hours was needed to create and validate the mapping file, which is a onetime process for each type of CMO and Drug combination. The manual data entry operation can process one or two batch records per day, resulting in 5 to 10 batch records per week. The speed achieved via the above solution is trice of that of the manual data entry process.
Considering the above setup, and two roles involved in this process, one for data entry using speech-to-text technology, while the other for validation, we could scale-up the software to process around 30 drug manufacturing batch-records for a single CMO and Drug combination in a week.

The speech-to-text system results in nearly 100% accuracy and the system can be scaled up to work much faster than manual data entry and data validation process.

Conclusion

After comparing the various types of OCR solutions, it can be concluded that OCR solutions are not effective against the variations in handwritten texts. The maximum of 50% accuracy was achieved in the trials performed on the handwritten information in the PDFs, while the Speech-to-text solution achieved nearly 100% accuracy of data extraction using the human-in-loop method for the data validation. Speech-to-text seems to be a much viable solution in the industry, as it can outperform the current manual data entry approaches. Also, a combination of OCR (for pre-processing) and speech recognition yields an average of 5X improvement in overall processing speeds with zero/near-zero error rates.