Optical Character Recognition (OCR) Service

IMAGE PDF FILES   SEARCHABLE PDF FILES
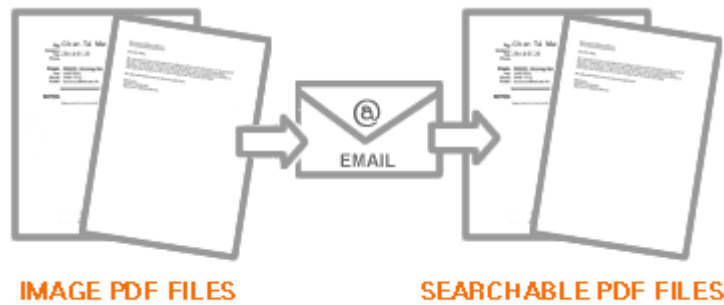
Many departments are equipped with multifunction printers (MFP) which could be used as a network scanner. Staff can use these MFPs to convert any paper document easily to PDF format for filing and for sharing. However, these PDF files contain only an image of the paper document which is not searchable (the content or the text is not recognized and indexed by search engine). Sometimes you may want to make it a searchable PDF with two layers: a layer of the image plus another layer of the text recognized from the image (i.e. PDF IMAGE+TEXT), say for indexing. One way is to use the software Acrobat to do an optical character recognition (OCR) to convert the document into a searchable PDF file.

To further automate the OCR process, ITS has recently setup an ABBYY OCR server to do the job for staff members. It is very easy to use. User simply email an image-based PDF document scanned by MFP (size of the PDF file must be less than 20MB) to the OCR server ocr@ied.edu.hk and the server will do the OCR and return a searchable PDF to the user via email shortly.

The OCR service supports not only PDF file, but also jpg, tiff and png file format.

The OCR server could recognize over 90% of English and Chinese (both Traditional and Simplified), but it is not 100% accurate.   Moreover, the accuracy of recognition depends on the quality of the original paper document/image.

The service allows users to convert paper document to searchable PDF and save it in our Document Management Systems (DMS). As searchable PDF will be indexed by the DMS, it allows users to search the documents in DMS by keywords found in the document.

PDF generated directly from Microsoft Word, Excel or PowerPoint is already searchable and there is no need to go through this OCR process to make it searchable. You could tell if a PDF is searchable by selecting the text in the PDF file. If you could select, copy and paste the text of the document to another application e.g. notepad or Microsoft Word, then the PDF file is searchable.

In case you have huge batch of files for conversion, special service can be arranged by ITS. Please contact Ms. Bonnie Shing at ext.6524 for details.