Creating Searchable PDFs From Scanned Documents - A Guide

Version 1.0.1 January 2006

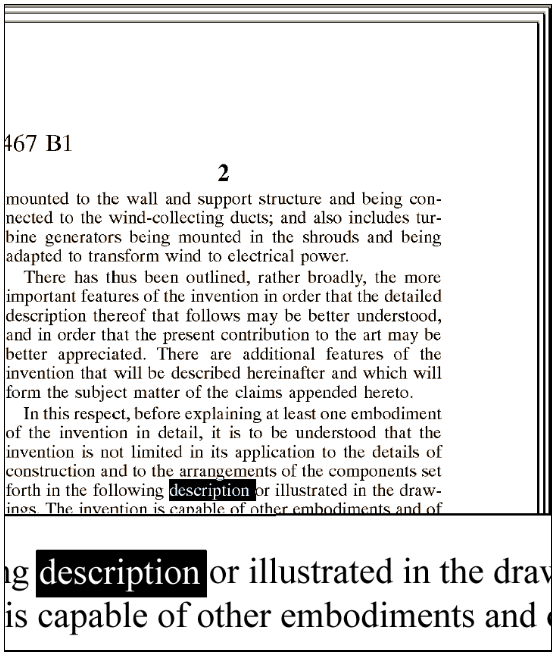
http://www.aquaforest.com/

1 Searchable PDF Explained

This brief document aims to provide guidance for the creation of searchable PDF files from scanned documents, whether standard TIFF Files or Image-Only PDF files.

1.1 What is a Searchable PDF?

A searchable PDF file is a PDF file that includes text that can be searched upon using the standard Adobe Reader “search” functionality. In addition, the text can be selected and copied from the PDF. Generally, PDF files created from Microsoft Office Word and other documents are by their nature searchable as the source document contains text which is replicated in the PDF, but when creating a PDF from a scanned document and OCR process needs to be applied to recognize the characters within the image.



467 B1

2

mounted to the wall and support structure and being connected to the wind-collecting ducts; and also includes turbine generators being mounted in the shrouds and being adapted to transform wind to electrical power.

There has thus been outlined, rather broadly, the more important features of the invention in order that the detailed description thereof that follows may be better understood, and in order that the present contribution to the art may be better appreciated. There are additional features of the invention that will be described hereinafter and which will form the subject matter of the claims appended hereto.

In this respect, before explaining at least one embodiment of the invention in detail, it is to be understood that the invention is not limited in its application to the details of construction and to the arrangements of the components set forth in the following description or illustrated in the drawings. The invention is capable of other embodiments and of

g description or illustrated in the drav
is capable of other embodiments and

1.2 Inside a Searchable PDF

In the context of Document Imaging, a searchable PDF will typically contain both the original scanned image plus a separate text layer produced from an OCR process. The text layer is defined in the PDF file as invisible, but can still be selected and seached upon. PDF files are able to store images using most of the native compression schemes used in TIFF files, so for example Group 4 TIFF files do not usually require any format conversion.

2 OCR Accuracy

A number of factors affect the accuracy of the text produced by the OCR process – 100% accuracy is certain possible under good conditions but each of the following issues, and OCR processing options will have an impact.

2.1 Original Image Quality

Although some pre-processing options such as despeckle and deskew can help in some cases, the visual quality of the original scan is of paramount importance.

2.2 Image DPI and Format

The image resolution should be at least 150 DPI for OCR processing, and preferably 300 DPI for optimal results, although for good quality scans 200 DPI is often sufficient. Non-lossy formats (TIFF Group 4, LZW etc) are preferred over lossy formats such as JPEG compression.

2.3 Despeckle

This pre-processing option removes isolated “dots” within the image which can cause recognition problems, and makes the result image “cleaner”.

2.4 Deskew

This option can improve OCR results by straightening crooked pages.

2.5 Auto-Rotate

OCR processing usually recognizes text written top-to-bottom, left-to-right, so pages that are orientated any other way (usually landscape pages) need to be re-oriented to enable recognition.

2.6 Language Settings

The language setting determines the set of characters that will be recognized, and the dictionary that will be used as a guide.

3 The Conversion Process

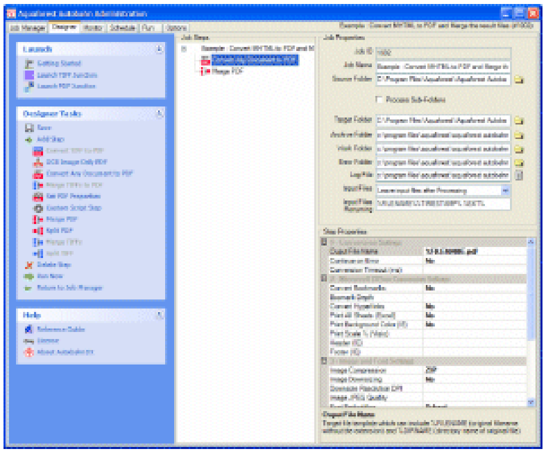
3.1 Conversion with TIFF Junction

Aquaforest's TIFF Junction is able to convert large volumes of TIFF and Image PDF files to searchable PDF with a high degree of accuracy. When converting from a TIFF file, the process is fairly straightforward; each page image is run through the OCR process according to the options set, and the text layer and image are used to construct the PDF file.

When converting from Image PDF Files, an additional stage is required which creates a TIFF file from the PDF document. By default this is done by rasterizing each PDF page to a bitmap and then converting to TIFF. This ensures a complete representation of each page is made, and is suitable for documents that actually have more than just a single image on each page (for example a Bates number as text) but can be slower than the "image extraction" method which directly extracts the images from each page.

3.2 Managing and Scheduling Jobs

Many conversion jobs can benefit from functionality such as Watched Folders, Scheduled Jobs, Windows Service and .Net API. To add these capabilities to TIFF Junction, Autobahn DX is available which includes TIFF Junction as one of it's components.



4 Hardware and Performance

4.1 CPU Power

The OCR process is highly CPU intensive and will benefit from being given as much CPU power as possible. As a guide about 1,000 pages per hour can be processed on a 2.5GHz processor, although this will vary according to the source document and OCR options chosen.

4.2 Exploiting Multiple CPUs

To take advantage of multiple CPUs, multiple conversion jobs should be run concurrently. This can most conveniently be done by using the Job Management facilities of Autobahn DX.

4.3 Memory

Memory can be a limiting factor when creating the final PDF, in the case of very large documents. A rule of thumb would be to have 1GB – 1.5 GB of memory per processor.

5 Further Information

Please contact info@aquaforest.com