Datasheet

PDF Alchemist

Overview

PDF Alchemist is a Windows and Linux software utility that converts PDF documents into HTML, XML, or EPUB files.

It is designed to allow a user to easily take content from a PDF document to:

• Display that content cleanly on a web site or a mobile device, without needing a lot of reformatting

• Store the content in export files or in a database for later processing and analysis

PDF Alchemist is provided as both a scriptable server tool and as a Software Development Kit.

It generates a single folder and set of export files for each PDF document processed and preserves the layout and format of the content, including text, graphics, and tables.

Images and font files embedded in the PDF source document are saved as separate files in the output.

Using PDF Alchemist

The product offers Optical Character Recognition (OCR) technology, allowing the software to scan images within PDF files and extract text found in those images.

If you have a set of PDF documents that you created by scanning hundreds or thousands of old printed pages, PDF Alchemist will be able to find the text in those PDF documents and export that text to an HTML file or XML file.

You can control how PDF Alchemist processes PDF documents.

If fidelity to the appearance of the source file matters to you, you can make sure that the product creates HTML files that will open in a browser and match what the original PDF looked like.

If getting text out of PDF documents is your priority instead, so that you can load it into a set of database tables for searching and analysis, PDF Alchemist can focus on collecting and storing content in a database file instead.

When converting pages, PDF Alchemist preserves font styles and layout, justification, indents, margins, lists, tables, and hyperlinks.

You can apply your own names to the export directories created for images and for font files extracted from a PDF document and create your own custom names for exported fonts and images.

PDF Alchemist can convert PDF form documents into HTML forms and convert PDF form actions into JavaScript code.

When exporting content from a PDF document to an EPUB file, PDF Alchemist can create a new section or chapter to correspond to every Table of Contents entry in the PDF source file.

You can also decide to divide the content exported from a PDF document into a separate HTML file at a specific number of pages.

For example, if you have a very long PDF document, you can create a series of HTML export files, each one 10 pages long.

You can also select a specific set or range of pages from a PDF source document to process and export.

With PDF Alchemist, you can apply borders to all of the tables exported from a PDF document or export all tables without borders. And if you want to only export tables, and disregard all of the other content found in a PDF source document, the product allows for that, too.

We provide a set of sample PDF documents with PDF Alchemist. You can process these files with the product to see how well PDF Alchemist manages exporting text and tables to HTML, and how well the OCR utility works.

Alchemist Features

- Build a table of contents in the output HTML file based on bookmarks in the source PDF file
- Find and export captions to images
- Choose the resolution of graphics images (DPI) that you want to export from a PDF source document
- Save or discard content from page headers and footers
- Export fonts found embedded in a PDF document to a separate fonts file
- Turn OCR processing on or off, and select the language to use
- Throw away style values and tags to produce cleaner text in the output file

Platforms Supported

- Windows 64-bit
- Linux 64-bit

For more information visit *www.datalogics.com* or subscribe to our blog at *blogs.datalogics.com*

Datalogics, Inc. | 101 N. Wacker Drive | Suite 1800 | Chicago, IL 60606 | USA