IRISPdf™ 6 for IRISPowerscan™

User's Guide

Table of Contents

Copyrights

IRISPdf 6 for IRISPowerscan-dgi-080916-01

Trademarks

CHAPTER 1
INTRODUCING IRISPDF FOR IRISPOWERSCAN

IRISPdf for IRISPowerscan is a fully-integrated version of IRISPdf in IRISPowerscan.

IRISPdf turns IRISPowerscan into a production OCR/ICR solution to scan, structure, sort, index and convert volumes of scanned documents into highly compressed electronic data.

IRISPdf supports up to 137 languages and uses optimized OCR technology (Optical Character Recognition) to recognize image files and process them into a wide range of output formats: text-searchable PDF and PDF/A files, and both searchable and editable Text, RTF, Word, OpenDocument Text, HTML, XML, WordML and SpreadsheetML files.

Next to that, IRISPdf can apply intelligent high-quality compression (iHQC) to PDF files.

IRISPdf also generates enhanced and compressed image files of different types: TIFF, multipage TIFF, JPEG, JPEG 2000 and Windows bitmap. IRISPdf also produces image-based PDF documents.

The processed documents can be exported to other programs by means of the export feature.

IRISPdf add-ons

In order to optimize the functionality of IRISPdf, several software add-ons are available.

IRISPdf can optionally be equipped with the following add-ons:

- **Asian OCR add-on**, to process 4 Asian languages (Traditional and Simplified Chinese, Japanese and Korean)
- **Hebrew OCR add-on**, to process Hebrew documents
- **Arabic OCR add-on**, to process Arabic documents

Installing IRISPdf add-ons

The IRISPdf add-ons must be directly activated in IRISPowerscan by means of a software key provided by I.R.I.S.

To acquire the software key:

- On the IRISPowerscan **Help** menu, click **Add-ons**, then select the add-on to be activated.
- Send the hardware key by e-mail to support.pro@irislink.com.
- Introduce the software key you receive by returned mail in the software key field.

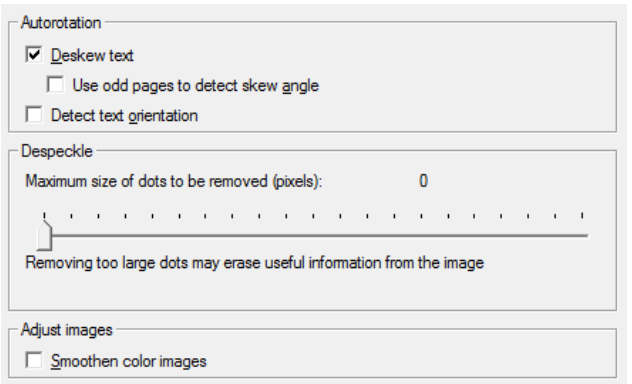
CHAPTER 2
IMAGE ENHANCEMENT

The image enhancement feature optimizes the OCR accuracy and image quality, and reduces the file size.

To access the image enhancement options:

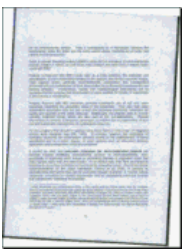
- Open the **Processing** section and click the **Image Enhancement** tab
- Set the image enhancement settings

  Note: do not select options that do not apply, however, they only slow down the recognition process.



Autorotation

- The **Deskew text** option automatically straightens pages scanned at an angle.



Deskewing improves the quality of scans and reduces the file size.

- Enable the option **Use odd pages to detect skew angle** to make the text deskewing faster.

  This option is designed for front-rear scanning. Only the front side is used to detect if the text is skewed.

- Enable the option **Detect text orientation** to rotate pages automatically when they have been scanned at a 90°, 180° or 270° angle.

  This option is useful when you're scanning documents with both portrait and landscape oriented pages.

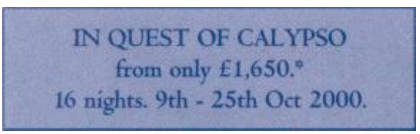
Despeckle

Despeckling images makes them both crisper and smaller in size.

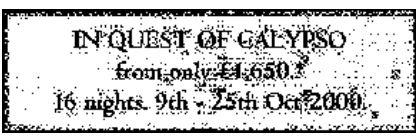
Select the maximum size of the dots you want to remove from black-and-white images with the **slide toolbar**.

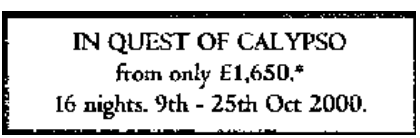
Adjust images

Select the option **Smoothen color images** to render grayscale and color images more homogeneous.



Smoothening is sometimes the only way to separate **text** from a **colored background**.



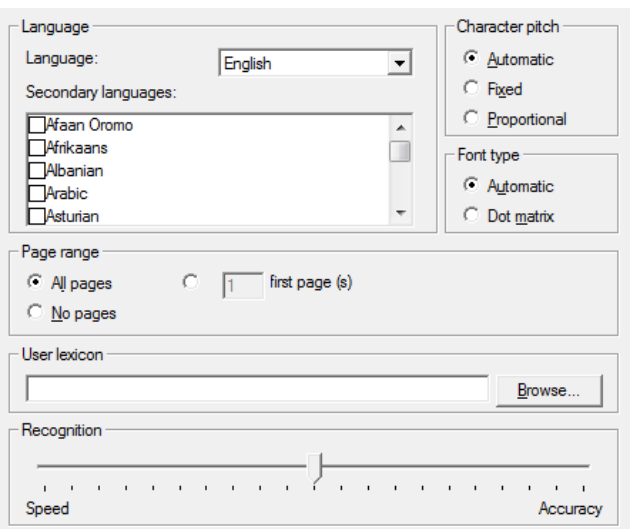


CHAPTER 3
CHARACTER RECOGNITION

The accuracy of the OCR process depends on many factors, such as the selected language, the document characteristics etc.

To access the character recognition options:

- Open the **Processing** section and click the **Character Recognition** tab
- Select the desired character recognition options:



Language

In order to recognize scanned documents, the document language must be specified. Based on the language selection, the software knows which symbol sets to recognize.

Select the language of your choice in the **Language** drop-down list.

> IRISPdf supports up to 137 languages. IRISPdf can optionally recognize four Asian languages (Traditional and Simplified Chinese, Japanese and Korean), Arabic and Hebrew.
>
> Note that the character recognition can also be limited to numeric digits.

Secondary languages

Next to the primary language, IRISpdf allows you to select up to 4 secondary languages.

This way IRISPdf uses mixed character sets, enabling it to recognize Western words that pop up in Greek, Cyrillic and optionally Asian, Arabic or Hebrew documents.

Check the boxes of the desired secondary languages in the list.

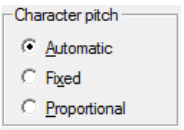
> Do not select languages that do not apply: the bigger the character set, the slower the recognition and the higher the risk of OCR errors.

The character recognition can be boosted by means of user lexicons:

> Create a .txt file containing the words you want IRISPdf to recognize. E.g. with Windows Notepad.
>
> Click the **Browse** button and open the lexicon in IRISPdf.

Character pitch

The character pitch is the **number of characters per inch** in a typeface.



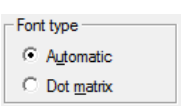
Select **fixed pitch** if all characters of the typeface have the same width. This is often the case in old typewriter documents.

Select **proportional pitch** when the characters of the typeface have a different width. Virtually all fonts you find in newspapers, magazines and books are proportional.

Select **Automatic** in order for IRISPdf to detect the character pitch automatically.

Font type

IRISPdf distinguishes between "regular" and dot matrix printed documents.



Dot matrix symbols (of the type 9 pin) are made up of isolated, separate dots.

Special segmentation and recognition techniques are used to recognize such documents.

Select **Dot matrix** to recognize so-called "draft" or "9 pin" dot matrix printed documents and **Automatic** to recognize "25 pin" or "NLQ" (Near Letter Quality) dot matrix, or other "normal" printing.

Page range

The character recognition can be applied to all pages, no pages and a certain number of pages.

The third option allows you to mix **text-based** and **image-based** pages in a single PDF file.

To create a mixed PDF file:

- Open the **Processing** menu and click the **Character Recognition** tab.
- In the **Page range** section, select the option **X first page(s)**
- The number of pages you indicated will be recognized. The pages following that number will only be scanned.

  This option increases the speed of the OCR process by avoiding the recognition of irrelevant pages and reduces the file size of the output.

Recognition

The recognition slide toolbar allows you to select the right trade-off between **OCR speed** and **OCR accuracy**.

Fast recognition can be used for documents with high-quality images while **Accurate recognition** should be preferred when the image quality is lower.

The confidence level of the OCR process can be checked in the log file **IRISPDF.HTML**.

This trade-off between speed and accuracy is available for the Latin, Cyrillic and Greek alphabets.

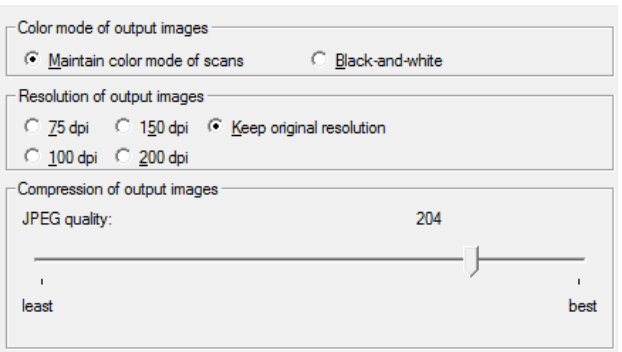
CHAPTER 4
IMAGE COMPRESSION

GENERAL IMAGE COMPRESSION

IRISPdf allows you to generate compact images. The images in scanned documents can be compressed and their color mode and output resolution changed by means of extensive **image compression options**.

To access these options:

- Open the **Image Compression** section and click the **General** tab
- Select the appropriate image compression options



Color mode of output images

You can either choose to maintain the color mode of scans or save scans in black-and-white.

> When you have chosen to maintain the color mode of scans while generating PDF documents, color and grayscale graphics will be saved in the JPEG format by default.
>
> Bitonal images (black-and-white) are saved in the TIFF format with Group 4 compression.

Resolution of output images

The resolution used to scan images does not necessarily have to be the **output resolution** of the images. You can store images in resolutions of **75**, **100**, **150** and **200** dpi or **keep their original resolution**.

> Note, however, that reducing the resolution of grayscale and color images is a processor-heavy task.

Compression of output images

Use the slide toolbar to determine the compression factor of JPEG images.

Note that the settings determined under the **General tab** apply to *all* graphics generated by IRISPdf.

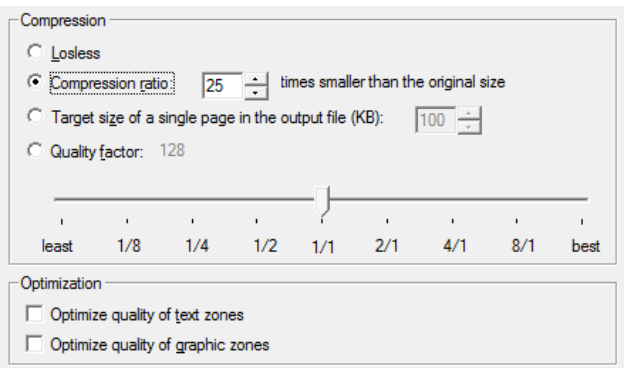
JPEG 2000 COMPRESSION

Next to the general image compression options, IRISPdf allows you to apply **JPEG 2000 compression** to color and grayscale images.

Note that JPEG 2000 compression does not apply to iHQC documents as iHQC is 15 times more efficient.

To access the JPEG 2000 compression options:

- Open the **Image Compression** section and click the **JPEG 2000** tab
- Select the appropriate compression options



Compression

The file size of scanned images can be influenced in several ways:

- Select **lossless** compression for optimal results
- The **Compression ratio** allows you to determine how many times you want the scanned images to be smaller than their original
- You can also determine the desired target size of a single page in the output file

  Indicate the file size for a single page from 1 KB to 10,240 KB. The default value is 100 KB.

- Select a **Quality factor** to determine the **degree of loss** allowed during the compression process

  Move the slide toolbar to select a value from 0 to 256: 0 guarantees the highest image quality, 256 the best compression. The default value is 128.

Optimization

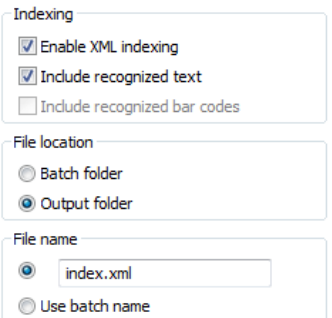
- **Optimize quality of text zones** maintains a high quality for text and table zones, reducing the quality of graphic zones in the output images.
- **Optimize quality of graphic** zones has the opposite effect.

  Note that **character recognition** must be enabled for these options to be available, otherwise the system can't detect which zones contain text and which areas contain graphics.

CHAPTER 5
XML INDEXING

After recognition, IRISPdf by default generates an **XML index file**, containing detailed information on the scanned documents, including the recognized text.

To access the XML indexing options, open the **Batch Output** section and click the **XML Indexing** tab.

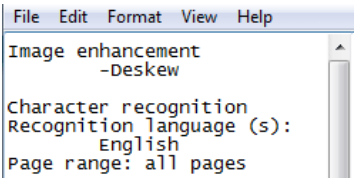


Note: do not confuse the generation of an XML index file with the generation of XML output.

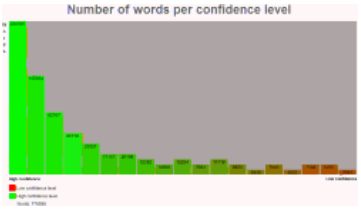
Other OCR statistics

Next to an XML index file, IRISPdf automatically generates a log file and an OCR confidence file after document processing. These files are located in the output folder of IRISPowerscan.

The **log file** lists all OCR parameters determined in IRISPdf.



The **OCR confidence file** allows you to monitor the confidence of the OCR process by means of two charts providing word-based and character-based statistics.



Should the confidence level not be satisfactory, ask yourself the right questions. E.g. Have the settings been determined correctly? Do the scanned documents have a sufficiently high resolution? Has the correct language been selected?

CHAPTER 6
DOCUMENT NAMES

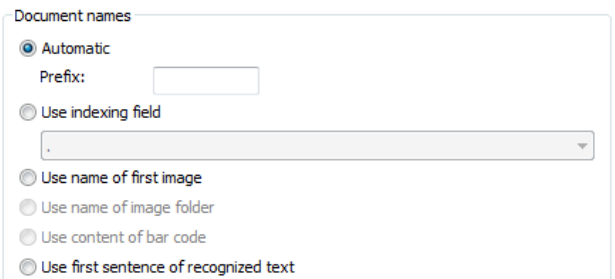
The documents generated by IRISPdf are named automatically by default: IRISPdf uses 8-digit names, starting from 00000000.

To access the document naming options:

- Open the **Document Output** section and click the tab **Document Names**
- Select the appropriate document naming options

  Although IRISPdf names documents automatically by default, there are several other naming possibilities.

  Note that you must first create index fields in IRISPowerscan before you can use the name of the index fields as document name. Refer to the IRISPowerscan documentation to learn how to do so.



CHAPTER 7
SUPPORTED OUTPUT FORMATS

IRISPdf supports a wide range of output formats: PDF, PDF-iHQC, PDF/A, PDF/A-iHQC, Word, RTF, WordML, OpenDocument Text, XML, HTML, Text, SpreadsheetML and several types of image files.

Note that all output formats are disabled by default, however.

To generate output files:

- Open the **Document Output** section and click on the tabs of the desired output formats.
- Select the output formats you want IRISPdf to generate and determine their **layout** and other **options**.

  Note that all output formats can be enabled simultaneously.

PDF DOCUMENT TYPES

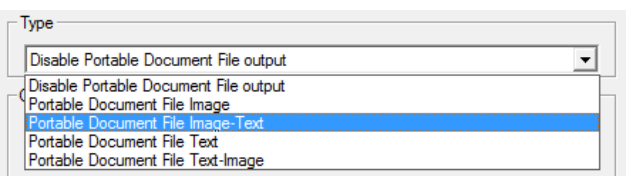
IRISPdf generates four types of PDF files: **Text**, **Text-Image**, **Image-Text** and **Image**.

IRISPdf also generates both **password-protected** and **digitally signed** PDF output and offers **PDF/A** output for long-time preservation.

IRISPdf can also apply **iHQC compression** to reduce the file size of PDF output to a minimum.

To generate PDF output:

- Open the **Document output** section and click the **PDF** tab
- Select the desired PDF type in the **Type** drop-down list



PDF Image

This format generates **image-only** PDF documents, it does not execute OCR.

> With IRISPdf it is also possible to mix text-based and image-based pages in a single PDF file. See the **Character Recognition** section.

PDF Image-Text

IRISPdf recognizes text and creates **searchable** PDF files that contain the page image and the recognized text.

The page image is placed on top of the text.

> With this format you can search words inside documents and view their true image as it was scanned.

Tip: use the graphics options in the **Image Compression** section to determine the color mode, resolution and JPEG quality of the graphics stored inside PDF files.

Tip: use the image enhancement options in the **Processing section** to improve the image quality and reduce the file size of **PDF Image** and **Image-Text** files.

Note that iHQC compression is available for **PDF Image** and **Image-Text**.

PDF Text

IRISPdf recognizes text and creates **searchable** PDF files.

The page image is not contained in these single-layered PDF files.

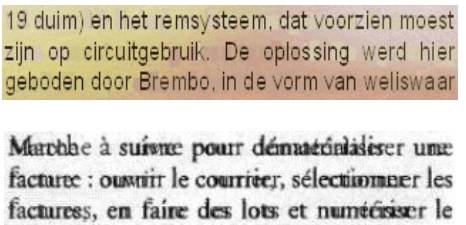
> Use **text-only** PDF files to save disk space.

PDF Text-Image

IRISPdf recognizes text and creates **searchable** PDF documents that contain the page image and the recognized text.

The page image is contained beneath the text.

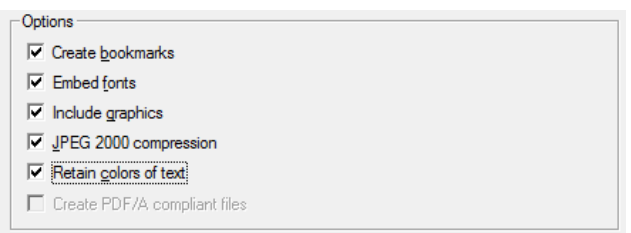
The pixels of the recognized text are erased to create a legible document. Otherwise, the text would have a heavy shadow as illustrated below:



PDF OPTIONS

Depending on the PDF type you have chosen, several options are available.

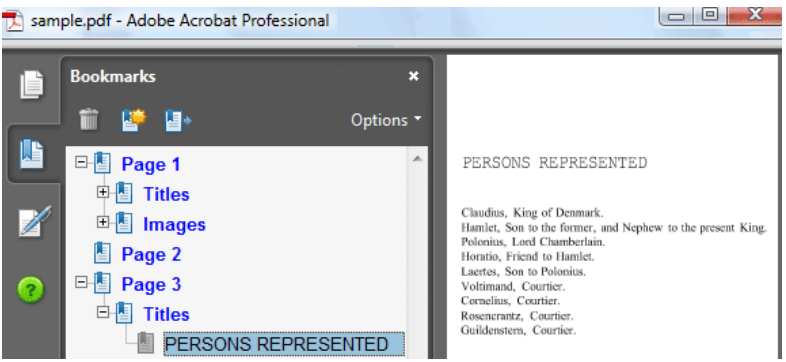
IRISPdf allows you to create bookmarks, embed fonts, include graphics, JPEG 2000 compress images, retain colors of text and create PDF/A compliant files.



To access the PDF options, open the **Document Output** section and click the **PDF** tab.

Create bookmarks

The option **Create bookmarks** creates bookmarks for each text block, graphic and table in Adobe Acrobat PDF files.



Embed fonts

Select the option **Embed fonts** to embed the fonts in Adobe Acrobat PDF files.

> Embedding fonts prevents font substitution and ensures that readers, regardless of their computer configuration, see the text in its original fonts.
>
> Embedding fonts increases the file size of recognized documents somewhat.

Include graphics

The option **Include graphics** includes the graphics in PDF **Text** documents.

> This option is enabled by default for the PDF types **Image**, **Image-text** and **Text-Image** and cannot be deselected.

> Including graphics is essential to create a true copy of source documents.

JPEG 2000 compression

By default, IRISPdf **JPEG 2000 compresses** grayscale and color images in PDF documents.

These settings apply to all graphics inside PDF files.

> Note that JPEG 2000 compression is not available for PDF/A and PDF-iHQC output.

Retaining colors of text

The option **Retain colors of text** maintains the original colors of the text across the recognition.

This option is **always** enabled for **PDF Text-Image** output and **can** be selected when you have chosen PDF Text.

PASSWORD-PROTECTED PDF

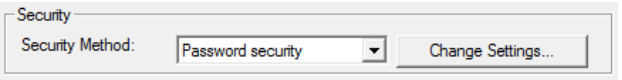
Next to regular PDF output, IRISPdf offers **password-protected PDF** and **PDF-iHQC** output.

Passwords may contain up to 255 characters.

> **Warning**: there is no way to recover passwords from a document.

To apply password-protection:

- Open the **Document Output** section and click the **PDF** tab

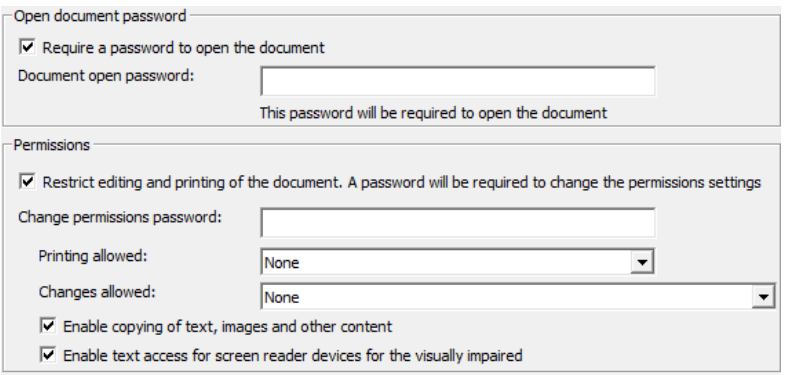


- Select **Password security** in the **Security Method** drop-down list.

  The **Change Settings** button becomes available.

- Click it to change the password security settings.

  These settings are the standard protection features offered by Adobe Acrobat.



DIGITALLY SIGNED PDF

Next to regular and password-protected PDF output, IRISPdf offers digitally signed **PDF**, **PDF/A**, **PDF-iHQC** and **PDF/A-iHQC** output.

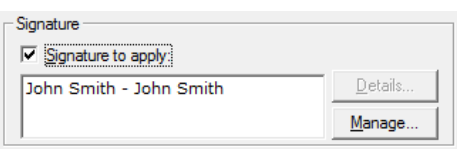
Digital signatures authenticate the identity of the document author, certify a document and help prevent unwanted changes. They are very hard to forge as they contain encrypted information unique to the signer.

The author signature is invisible: it appears in the **Signatures** tab of Adobe Acrobat and Adobe Reader. To ensure legibility of all scanned information, IRISPdf does not place a signature on the pages of recognized documents.

Warning: it is up to the user to create a self-signed digital ID or to obtain a certificate from a third-party signature handler. Refer to the manual of Adobe Acrobat for specific instructions.

To apply a digital signature:

- Open the **Document Output** section and click the **PDF** tab
- Check the box **Signature to apply** to apply a digital signature



- Click the signature you wish to apply

  The **Details** button will become available.

- Click the **Details** button to view all available information on the current signature

- Click the **Manage** button to manage any digital signature installed on your PC

  You can edit, remove, import and export the digital certificates.

PDF/A

Next to "regular" PDF documents, IRISPdf offers **PDF/A** and **PDF/A-iHQC** output.

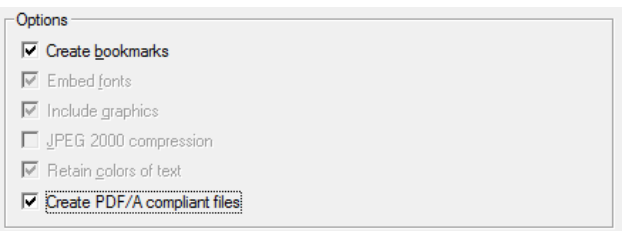
PDF/A files are used for long-term archiving and contain only what is strictly needed for opening and viewing files during their expected lifetime.

The PDF/A files generated by IRISPdf are ISO standard (ISO 19005-1:2005) and PDF/A-1b compliant.

To generate PDF/A output:

- Open the **Document Output** section and click the **PDF** tab
- Select the PDF file format of your choice in the **Type** drop-down list
- Clear the **JPEG 2000 compression** option

  The option **Create PDF/A compliant files** will become available.



- Select that option to create PDF/A compliant files.

  Important: When producing **PDF Text** files, IRISPdf embeds all fonts automatically in PDF/A output to ensure that documents can be opened and viewed as created in the future.

  When producing **PDF Image-text** files, however, IRISPdf now offers PDF/A files **without embedded fonts**. As the text is placed beneath the image, no font embedding is necessary. This way, IRISPdf produces more compact PDF/A output while the document text is still searchable and copyable.

  Notes:

  To avoid data loss, PDF/A files cannot be password-protected.

  PDF/A compliant files do not support JPEG 2000 compression.

  PDF/A compliant output is currently not available for Asian languages.

  It takes Adobe Acrobat 5.0 (or Adobe Acrobat Reader 5.0) or higher to generate and open PDF/A files.

PDF - iHQC

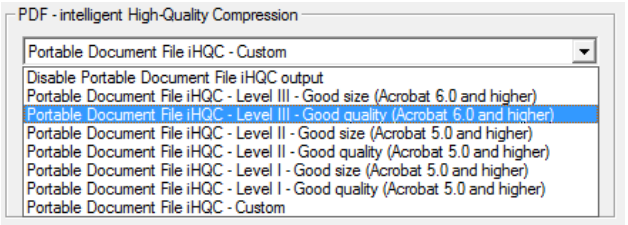
Next to four types of regular **PDF files**, IRISPdf also offers two types of PDF-iHQC output: **PDF Image-text** and **PDF Image**.

iHQC stands for **intelligent High-Quality Compression**, I.R.I.S.' proprietary, efficient compression technology. iHQC is to images what MP3 is to music and what DivX is to movies.

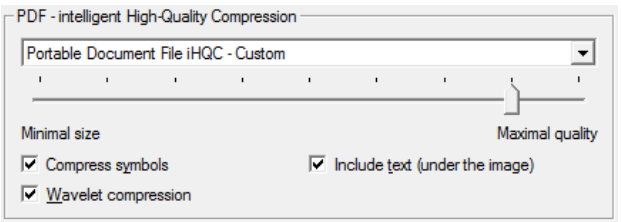
Note that Adobe Acrobat 8 is recommended for viewing PDF-iHQC documents.

To generate PDF - iHQC output:

- Open the **Document Output** section and click the **PDF-iHQC** tab.
- Select the **compression level** you want to apply.



You can also **customize** the compression level, by selecting the **custom** option.



Compress symbols applies specialized compression to text whereas **Wavelet compression** applies specialized compression to graphics.

Options

IRISPdf offers the same options for PDF-iHQC and regular PDF output. Refer to the section **PDF options**.

Note that PDF/A compliant files cannot be generated in combination with Level III compressed files.

TEXT-BASED OUTPUT FORMATS

WORD, WORDML, RTF AND OPENDOCUMENT TEXT

IRISPdf offers several types of text-based output formats: it generates versatile **Word**, **WordML, RTF** and **OpenDocument Text** output.

To generate text-based output files:

- Open the **Document Output** section and click on the tabs of the desired output formats
- Select the output formats you want IRISPdf to generate and determine their **layout** and other **options**

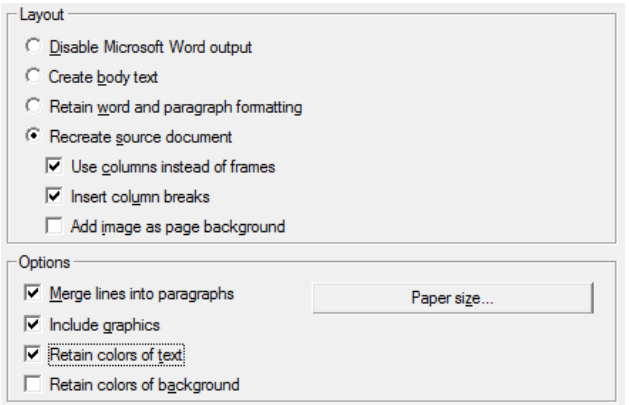
> **WordML** is supported by Microsoft Word 2007 and 2003.
>
> **OpenDocument Text** is an XML-based open format supported by several recent word processors. An open source plug-in is required for Microsoft Word to support this format.

LAYOUT AND OTHER OPTIONS

Numerous layout options are available for the text-based output formats **Word**, **WordML**, **RTF** and **OpenDocument Text**.

> Note that many of the options described below also apply to **HTML** output files.

Open the **Document Output** section and click the tabs of the desired output formats to access the options.



Layout

- **Create body text** avoids text formatting by IRISPdf
- **Retain word and paragraph formatting** takes an intermediate position between body text and autoformatting

  The font type, size and type style are maintained across the recognition.

  The tabs and the alignment of each block are recreated.

  The text blocks and columns aren't recreated; the paragraphs just follow each other.

  The tables are recaptured correctly.

  These two options are also available for **SpreadsheetML** output.

- **Recreate source document** recreates a facsimile copy of the original document

You get a true copy of your source document, no longer a scanned image.

- **Use columns instead of frames** determines *how* the autoformatting will be done: the text blocks, tables and graphics can be stored in frames or flowing columns (if any)

  Columnized texts are easier to edit than documents containing several frames: the text flows naturally from one column to the next.

  Note: when the system is unable to detect columns in the source document, this formatting mode uses frames as a fallback position.

  Note that this option is not available for **HTML** output.

- **Insert column breaks** determines whether you insert hard column breaks at the end of each column

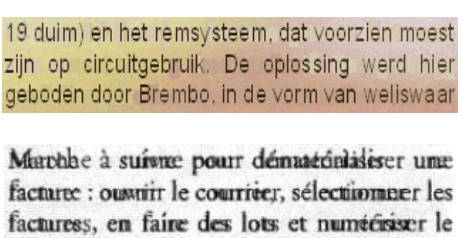
  Any text you edit, add or remove, remains inside its column; no text ever flows automatically across a column break.

  Tip: disable this option when you have columnized body text. You'll ensure the natural flow of the text from one column to the next.

  Note that this option is not available for **HTML** output.

- The option **Add image as page background** places the scanned image as page background beneath the recognized text

  The pixels of the recognized text are erased to create a legible document. Otherwise, the text would have a heavy shadow as illustrated below:



This option increases the file size of the output files substantially, however.

Note: this option is not available for **WordML** files.

The format **PDF Text-Image** provides the same result for PDF files.

The option **Retain colors of background** provides a less drastic, more compact alternative, as illustrated above.

Note that IRISPdf detects any web page URLs and e-mail addresses in scanned documents and recreates them as hyperlinks in the output.

Options

- **Merge lines into paragraphs** enables automatic paragraph detection

  IRISPdf wordwraps the recognized text until a new paragraph starts and reglues hyphenated words at the end of a line.

- **Include graphics** includes the graphics in autoformatted files

  This is essential to create a true copy of a document.

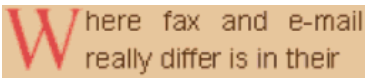
  Use the graphics options of the **Image Compression** section to determine the color mode and resolution of the graphics stored inside the output files.

  Use the **image enhancement** options of the **Processing** section to improve the image quality and reduce the file size.

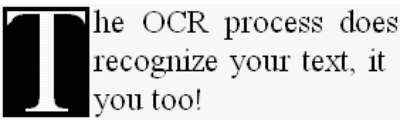
- **Retain colors of text** maintains the original colors of the text across the recognition

- **Retain colors of background** maintains the spot colors of the page background across the recognition

A uniform background color - if there is one in the source document - is created per paragraph in the output file.



This option also recreates inverted drop letters.



The option **Add image as page background** offers a more drastic, less compact alternative, as illustrated above.

Retaining the colors of the background implies that the colors of the text are maintained simultaneously.

When you recognize tables and save the document as a SpreadsheetML worksheet, this option maintains the background color of each cell.

	A	B	C
1	Performance optical media		
2	CD-ROM Digital Versatile Disk	Average access time (msec)	CPU utilization (%)
4	CD-ROM 24x speed	80	58.2
5	CD-ROM 32x speed	60	72.1
6	DVD	58	78.9
7	Tested on 333 MHz Pentium II with 64 MB RAM and 4 GB HD		

Preferred paper sizes

When you are exporting **Word, WordML, RTF** or **OpenDocument Text** documents, you can select preferred paper sizes.

IRISPdf will go through the active paper sizes in the indicated order and uses the first paper size that is sufficiently large to hold the scanned document.

OTHER OUTPUT FORMATS

SPREADSHEETML

IRISPdf offers versatile SpreadsheetML output. This format is supported by Microsoft Excel 2007, 2003 and 2002.

As documents often contain more than only tables, it is useful to activate SpreadsheetML as a "secondary" format alongside (an)other format(s). It is only used for those pages that contain tables, for all other pages the SpreadsheetML output format is disabled.

To generate SpreadsheetML Output:

- Open the **Document Output** section and click the **SpreadsheetML** tab
- Select the **Layout** and other **options** of your choice:

Layout

The layout options **Create body text** and **Retain word and paragraph formatting** are available, just as in text-based output formats.

Options

- The option **Merge lines into paragraphs** enables automatic paragraph detection

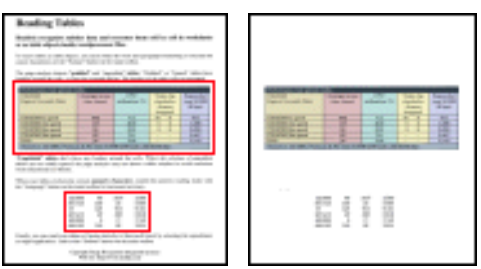
IRISPdf wordwraps the recognized text until a new paragraph starts and reglues hyphenated words at the end of a line.

- The option **Retain colors of text** maintains the original colors of the text across the recognition
- The option **Retain colors of background** recreates the background color of each cell

	A	B	C
1	Performance optical media		
2	CD-ROM	Average access	CPU
3	Digital Versatile Disk	time (msec)	utilization (%)
4	CD-ROM 24x speed	80	58.2
5	CD-ROM 32x speed	60	72.1
6	DVD	58	78.9
7	Tested on 333 MHz Pentium II with 64 MB RAM and 4 GB HD		

- The option **Ignore all text outside the tables** saves the tables and ignores all other recognition results

All data inside the tables is recaptured; any data outside the table(s) is not.



You can limit the recognition to a numeric character set. Only the digits 0 to 9 will be recognized.

- The option **Convert figures into numbers** encodes the recognized figures as numbers

As a result, you can execute arithmetical operations on those cells. The text cells (in any table) remain text.

Excel exclusively executes mathematical operations on data that is encoded as numbers.

Create one worksheet per

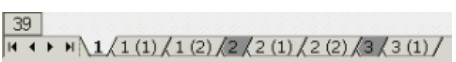
- The option **Create one worksheet per Page** sees to it that one worksheet is created per scanned page

  If a page contains tables and text, all will be placed on the same worksheet.

  Note that only the figures inside the tables are encoded as numbers. When the option **Convert data to numbers** is enabled, text inside and outside the table remains text.

- The option **Create one worksheet per Table** places each table in a separate worksheet and includes the recognized text (outside the tables) in yet another worksheet

  If the recognized document contains ***several pages***, you'll see that structure repeated per page.



(UNICODE) TEXT

IRISPdf offers versatile unicode **Text** output.

To generate Text output:

- Open the **Document Output** section and click the **Text** tab
- Select the file **type** of your choice
- Use the option **Unicode Text** to generate Unicode text output

The advantage of Unicode is that you can encode any language - and view and edit the result with the proper word processor (Word 2007, 2003, 2000).

- Use the option **Unicode UTF-8** to generate Unicode UTF-8 output

  Unicode UTF-8 is a web-based text format.

Option

Merge lines into paragraphs enables automatic paragraph detection.

IRISPdf wordwraps the recognized text until a new paragraph starts and reglues hyphenated words at the end of a line.

HTML

IRISPdf offers versatile HTML output, a highly popular text format supported by all web browsers.

To generate HTML output:

- Open the **Document Output** section and click the **HTML** tab
- Select the appropriate **layout** and other **options**

These options are highly similar to the ones available for text-based output files. Refer to the section **Layout and other options**.

XML

IRISPdf offers versatile **XML** output.

Do not confuse XML output with **XML indexing**.

To generate XML output:

- Open the **Document Output** section and click the **XML/WordML** tab.
- Select the file **type** of your choice:
  - **Compact XML** creates the smallest XML documents

    The text is legible to the human eye as it is stored line by line, block by block

    Any application capable of parsing XML files (e.g. Internet Explorer) can be used to study the OCR results.

    Any XML parser can be used to edit and parse the XML documents.
  - **Detailed XML** adds much detail to the recognized text

    The text is *not* legible to the human eye because the XML document contains detailed formatting information (type styles, position of each character on the page etc.). The text is stored character by character, word by word.

    It takes an XML parser to make sense of the XML output.

IMAGE FILES

Alongside several text formats, IRISPdf offers image output.

Images can be exported as BMP, JPEG, JPEG 2000 and TIFF files.

To generate image output:

- Open the **Document Output** section and click the **Image files** tab.
- Select the appropriate **file format** in the drop-down lists to generate **bitonal** and/or **color-grayscale** images:
    - The following graphic formats are supported for **bitonal images**: **TIFF** and **multipage TIFF** (both with Group 4 compression) and **Windows bitmaps**.
    - The following graphic formats are supported for **color-grayscale images**: **JPEG**, **JPEG 2000**, **TIFF** and **multipage TIFF** (both with JPEG and JPEG 2000 compression) and **Windows bitmaps**.

Tip: use the image enhancement options in the **Processing section** to improve the image quality.

Warning: Windows bitmaps do not offer any compression. A single A4 color page may take some 25 MB disk space on your hard disk.

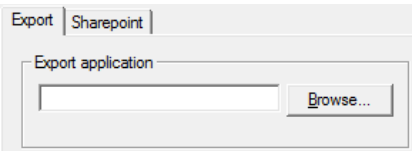
Note that you can also generate **image-only** PDF and PDF-iHQC files.

Also note that scanned images can be saved in black-and-white and color-grayscale mode simultaneously.

CHAPTER 8
EXPORT APPLICATION

By means of the export application, processed documents can be exported to other programs automatically after OCR. Simply indicate to IRISPdf which program to run.

Open the **Export** section and click the **Browse** button to search for an appropriate application.



INDEX