

# User Guide

Version 1.1

# IRISDocument 9.6 for IRIS Powerscan

11/29/2012  
I.R.I.S. Products & Technologies  
Dgi



## Table of Contents

Legal Notices .....	5
<b>Chapter 1 Introducing IRISDocument .....</b>	<b>7</b>
<b>Chapter 2 Getting started .....</b>	<b>9</b>
Running IRISDocument for IRIS Powerscan .....	9
<b>Chapter 3 Image enhancement .....</b>	<b>11</b>
<b>Chapter 4 Character Recognition.....</b>	<b>15</b>
Language.....	16
Secondary languages .....	16
Character pitch.....	17
Font type .....	17
Page range.....	18
Recognition.....	19
<b>Chapter 5 Image Compression.....</b>	<b>21</b>
General Image Compression.....	21
JPEG 2000 Compression .....	22
Compression .....	23
Optimization .....	24
<b>Chapter 6 XML Indexing .....</b>	<b>25</b>
Other OCR statistics .....	26
<b>Chapter 7 Document Names.....</b>	<b>27</b>

---

<b>Chapter 8 Output Formats .....</b>	<b>31</b>
Supported output formats .....	31
PDF .....	31
PDF Document types .....	31
PDF Options .....	35
Password-protected PDF .....	37
Digitally signed PDF .....	38
PDF/A .....	39
PDF - iHQC .....	41
XPS .....	43
XPS Document types .....	43
XPS Options .....	45
XPS - iHQC .....	46
Text-based output formats .....	47
Word, WordML, RTF and OpenDocument Text .....	47
Layout and other options .....	47
Other output formats .....	52
SpreadsheetML .....	52
(Unicode) Text .....	54
HTML .....	55
XML .....	55
Image files .....	56

<b>Chapter 9 Export Features .....</b>	<b>59</b>
<b>Index .....</b>	<b>61</b>

## **Legal Notices**

*IRISDocumentServer9.6\_dgi\_241012\_01*

### **Copyrights**

Copyrights © 2002 - 2012 I.R.I.S. All Rights Reserved.

I.R.I.S. owns the copyrights to the IRISDocument software, to the online help system and to this publication.

The information contained in this document is the property of I.R.I.S. Its content is subject to change without notice and does not represent a commitment on the part of I.R.I.S. The software described in this document is furnished under a license agreement which states the terms of use of this product. The software may be used or copied only in accordance with the terms of that agreement. No part of this publication may be reproduced, transmitted, stored in a retrieval system, or translated into another language without the prior written consent of I.R.I.S.

This User Guide utilizes fictitious names for purposes of demonstration; references to actual persons, companies, or organizations are strictly coincidental.

### **Trademarks**

The I.R.I.S. logo, IRISDocument, IRIS Powerscan and IRISDocument Server are trademarks of Image Recognition Integrated Systems S.A.

OCR ("Optical Character Recognition") technology, MICR ("Magnetic Ink Character Recognition") and barcode reading technology by I.R.I.S.

AutoFormat, ClearView, Connectionist, Linguistic and WID technology by I.R.I.S.

XML parser developed by Apache. This product includes software developed by the Apache Software Foundation.

All other products mentioned in this User Guide are trademarks or registered trademarks of their respective owners.

### **Patents**

iHQC™ patent-protected. US Patent No. 8,068,684.

# CHAPTER 1

## INTRODUCING IRISDOCUMENT

IRISDocument for IRIS Powerscan is a fully-integrated version of IRISDocument in IRIS Powerscan.

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

To process image files, IRISDocument uses I.R.I.S.' proprietary OCR technology (Optical Character Recognition), which supports as many as 137 languages. All American and European languages are supported, including the Central-European, Baltic and Cyrillic languages as well as Greek and Turkish.

A wide range of output formats are available: IRISDocument converts image files into text-searchable PDF and XPS files and into both text-searchable and editable Text, RTF, Word (.docx), OpenDocument Text, HTML, XML, WordML and SpreadsheetML files.

The PDF files you generate can be password-protected and digitally signed. Also the PDF/A format, the standard format for long-term archiving, is supported.

IRISDocument can also generate hyper-compressed PDF and XPS documents. By means of iHQC, which stands for intelligent High-Quality Compression, PDF and XPS documents can be compressed without loss of image quality. Three levels of iHQC are available for PDF documents, and one level for XPS documents. **iHQC Level I - Good Quality** is available for both PDF and XPS documents in

the standard version of IRISDocument Server. To make full use of the iHQC technology, an iHQC add-on is required.

Besides text recognition, IRISDocument offers powerful barcode recognition for document structuring purposes. Barcodes in image files (or in scanned documents) can be used as separators to indicate where new documents begin.

### **The IRISDocument Export features**

The documents that are processed can be exported to other applications by means of the **Export** feature.



# CHAPTER 2

## GETTING STARTED

### RUNNING IRISDOCUMENT FOR IRISPOWERSCAN

---

To start IRISDocument for IRIS Powerscan:

- From the Windows **Start** menu, point to **All Programs**, **IRIS Powerscan**, and then click **IRIS Powerscan**. Or click the shortcut on your desktop.
- **Open** the project in which you want to determine the **IRISDocument for IRIS Powerscan** settings.
- Click **Setup** on the **File** menu.
- Click the **Processing** icon and select **IRISDocument** in the list of connectors.
- Click **Setup** to determine the **IRISDocument for IRIS Powerscan** settings.

The settings are discussed in the sections below.

## CHAPTER 3

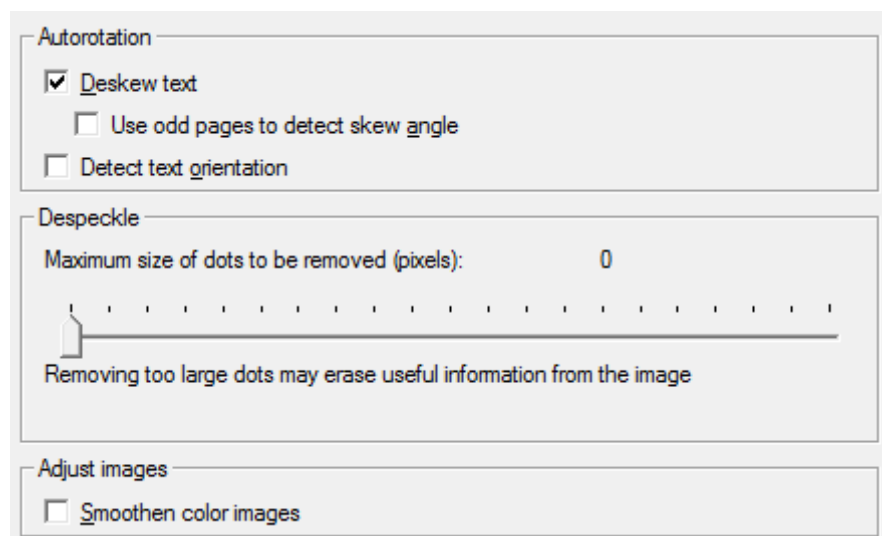
# IMAGE ENHANCEMENT

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.



The screenshot shows a dialog box with three sections: 'Autorotation', 'Despeckle', and 'Adjust images'. The 'Autorotation' section has three checkboxes: 'Deskew text' (checked), 'Use odd pages to detect skew angle' (unchecked), and 'Detect text orientation' (unchecked). The 'Despeckle' section has a slider for 'Maximum size of dots to be removed (pixels)' set to 0, with a warning below it: 'Removing too large dots may erase useful information from the image'. The 'Adjust images' section has one checkbox: 'Smoothen color images' (unchecked).

### 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.

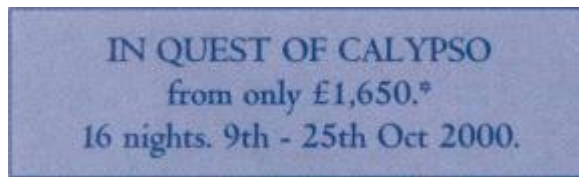
## **Despeckle**

Despeckling images makes them both crisper and smaller in size.

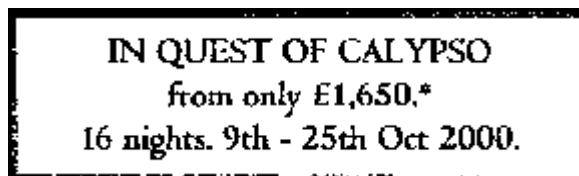
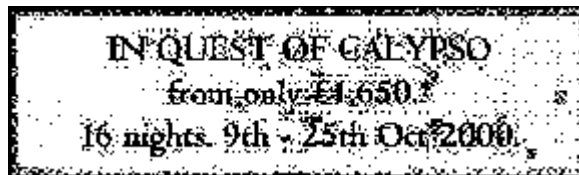
Move the slider to determine the **maximum size of the dots you want to remove** from black-and-white images.

## **Adjust images**

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



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



# CHAPTER 4

## 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 required character recognition options:

The screenshot shows a dialog box with several sections for configuring character recognition. The 'Language' section includes a dropdown menu set to 'English' and a list of secondary languages with checkboxes: Afaan Oromo, Afrikaans, Albanian, Arabic, and Asturian. The 'Character pitch' section has radio buttons for 'Automatic' (selected), 'Fixed', and 'Proportional'. The 'Font type' section has radio buttons for 'Automatic' (selected) and 'Dot matrix'. The 'Page range' section has radio buttons for 'All pages' (selected) and 'No pages', with a text box containing '1' and the label 'first page (s)'. The 'User lexicon' section features a text input field and a 'Browse...' button. The 'Recognition' section contains a slider bar with 'Speed' on the left and 'Accuracy' on the right, with a slider handle positioned approximately in the middle.

## Language

In order to recognize 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.

IRISDocument supports up to 137 languages. IRISDocument can optionally recognize four Asian languages (Traditional and Simplified Chinese, Japanese and Korean), Arabic and Farsi, and Hebrew.

Note that the character recognition can also be limited to numeric digits.

## Secondary languages

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

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

Select the required secondary languages in the list.

Note that if you select multiple secondary languages they must be of the same language group. Languages that do not belong to the same group will be disabled automatically.

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.

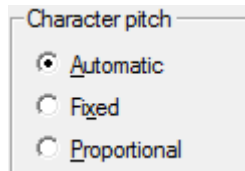
**The character recognition can be boosted by means of user lexicons:**

Create a .txt file containing the words you want IRISDocument to recognize. E.g. with Windows Notepad.

Click the **Browse** button and open the lexicon in IRISDocument.

## 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.

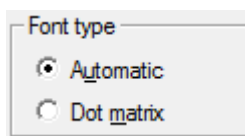


Select **Automatic** in order for IRISDocument to detect the character pitch automatically.

Note: the character pitch option does not apply to Asian or Arabic documents.

## Font type

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



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

**Far out in the uncharted back**

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.

Note: the Font type option does not apply to Asian, Arabic or Hebrew documents.

## **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 **IRISDOCUMENT.HTML**.

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

# CHAPTER 5

## IMAGE COMPRESSION

### GENERAL IMAGE COMPRESSION

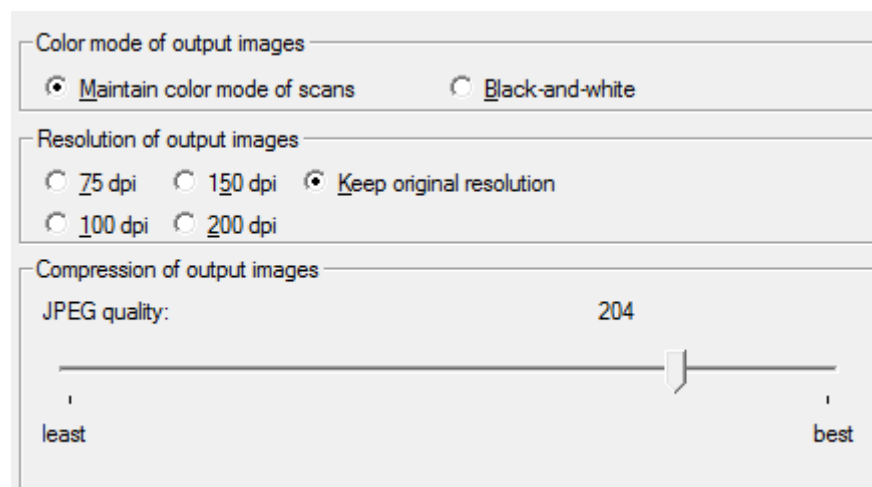
---

IRISDocument 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:

Note: take your storage limitations into account when selecting these 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. When generating color-grayscale images, reducing the resolution is not recommended.

## Compression of output images

Use the slider to determine the compression factor of JPEG images.

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

## JPEG 2000 COMPRESSION

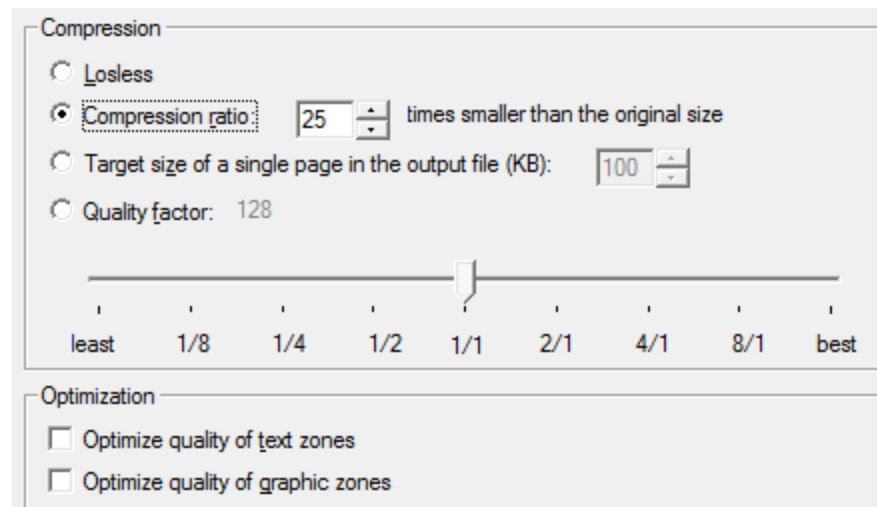
---

Next to the general image compression options, IRISDocument 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 options of your choice:



## 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 6

# XML INDEXING

After recognition, IRISDocument 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.

The screenshot shows a configuration panel for XML indexing, divided into three sections:

- Indexing:** Contains three checkboxes:  Enable XML indexing,  Include recognized text, and  Include recognized bar codes.
- File location:** Contains two radio buttons:  Batch folder and  Output folder.
- File name:** Contains two radio buttons:  index.xml (with a text input field containing 'index.xml') and  Use batch name.

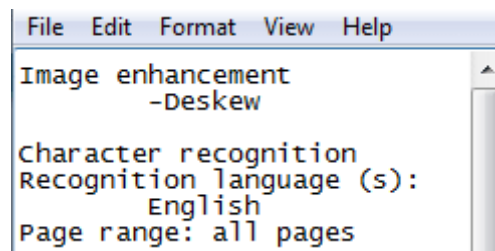
Note: do not confuse the generation of an XML index file with the generation of XML output.

The information in the XML file is used to export your processed documents to other applications. For more information see the section **Export features**.

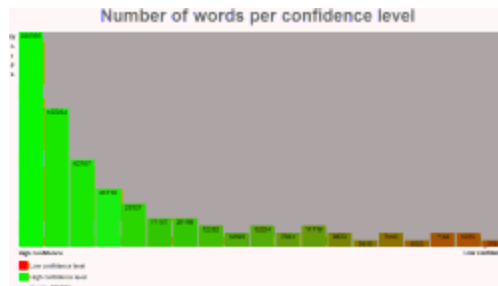
## Other OCR statistics

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

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



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 7

## DOCUMENT NAMES

The documents generated by IRISDocument are named automatically by default: IRISDocument 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:
  - The **Automatic** naming option is selected by default. The documents that are processed are given an 8-digit serial number, starting with 00000000, 00000001, ...

Any prefix can be added in front of this sequential name.

- The naming option **Use indexing fields** only applies to IRISDocument Server for IRIS Powerscan. In order to use an index field as document name, you must first create index fields in IRIS Powerscan.
- When you select **Use Name of first image**, all your output documents will start with the name of the first input file in the batch or watched folder that was processed by IRISDocument. E.g. SampleDocument, SampelDocument0, SampleDocument01, ...
- When you are processing multiple image folders each containing image files that belong together, select the



naming option **Use name of image folder**. Your output documents will be named ImageFolder0, ImageFolder1 for instance.

- You can also use the content of a specific barcode in your document as document name. Select the option **Use content of barcode** and indicate which barcode in the document must be used as document name.
- Or you can use the content of the barcodes on the first page as document name. Select the option **Use content of first-page barcodes** and indicate at which barcode IRISDocument must start.
- The **first sentence of the recognized text** can also be used as naming option.
- When generating PDF and PDF-iHQC output files at the same time, or XPS and XPS-iHQC output files at the same time, a **suffix** is added to the document name by default: e.g. document.ihqc.pdf. This way a clear distinction is made between regular PDF and XPS documents and their iHQC-compressed counterparts. This suffix can be changed to your liking.

Note: when you are only generating PDF-iHQC output files, and no 'regular' PDF files, no suffix is added to the files. The same goes for XPS-iHQC files.

Document names

Automatic  
Prefix:

Use indexing field

Use name of first image

Use name of image folder

Use content of bar code

Use content of first page bar codes  
Start at barcode:

Use first sentence of recognized text

Optional iHQC file name suffix:

# CHAPTER 8

## OUTPUT FORMATS

### SUPPORTED OUTPUT FORMATS

---

IRISDocument supports a wide range of output formats: PDF, PDF-iHQC, PDF/A, PDF/A-iHQC, Word (.docx), 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 IRISDocument to generate and determine their **layout** and other **options**.

Note that all output formats can be enabled simultaneously.

### PDF

---

#### PDF Document types

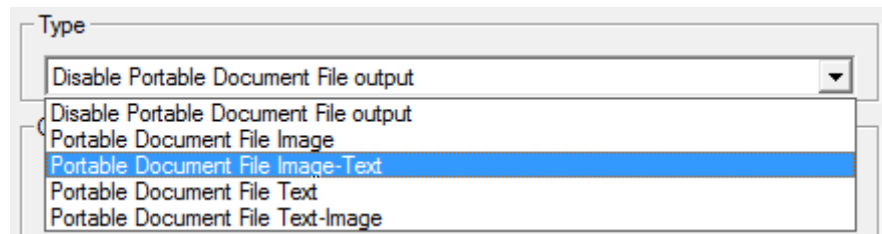
IRISDocument generates four types of PDF files: **Text**, **Text-Image**, **Image-Text** and **Image**.

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

IRISDocument 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**, **PDF Image-Text**, **PDF Text** or **PDF Text-Image**.



### PDF Image

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

With IRISDocument 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

IRISDocument 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

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

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

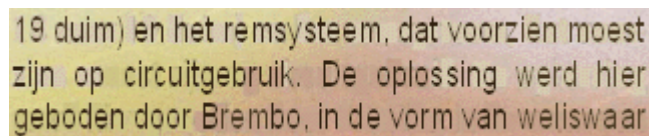
**Tip:** use **text-only** PDF files to save disk space.

## PDF Text-Image

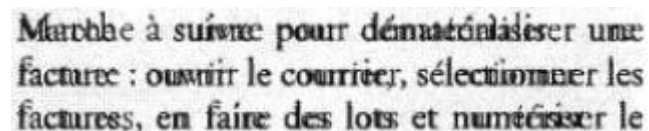
IRISDocument recognizes text and creates **searchable** PDF documents that contain the page image and the recognized text.

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:



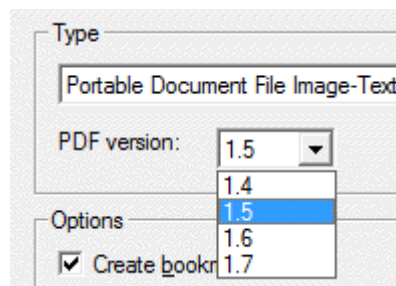
19 duim) en het remsysteem, dat voorzien moest zijn op circuitgebruik. De oplossing werd hier geboden door Brembo, in de vorm van weliswaar



Marche à suivre pour dématérialiser une facture : ouvrir le courrier, sélectionner les factures, en faire des lots et numériser le

- Select which **version** of PDF document you want to generate: 1.4, 1.5, 1.6 or 1.7.

IRISDocument by default generates **PDF 1.5** documents.



## Notes

- If you want to generate **PDF/A compliant files**, select **PDF version 1.4**.

Note that PDF 1.4 documents are not compatible with certain options:

- PDF version 1.4 is compatible with iHQC Level I and II, but not with Level III or custom compression.
- PDF version 1.4 is not compatible with JPEG 2000 compression.
- PDF version 1.4 is not compatible with Wavelet compression (which applies special compression to graphics).
- In case you are using Adobe Acrobat to view PDF files:
  - It takes Adobe Acrobat 5.0 and higher to open PDF 1.4 documents.
  - It takes Adobe Acrobat 6.0 and higher to open PDF 1.5 documents.
  - It takes Adobe Acrobat 7.0 and higher to open PDF 1.6 documents.
  - It takes Adobe Acrobat 8.0 and higher to open PDF 1.7 documents.

## PDF Options

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

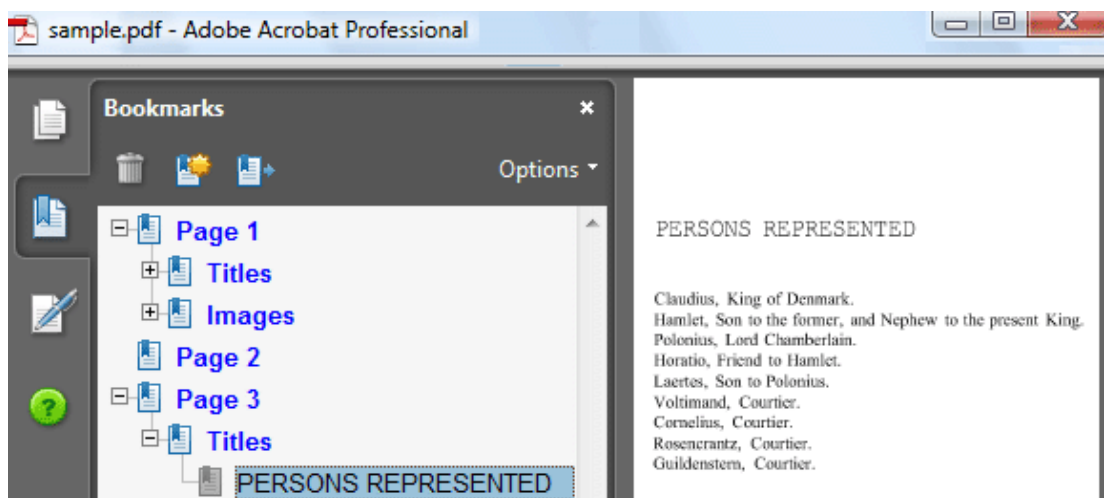
IRISDocument allows you to create bookmarks, embed fonts, include graphics, JPEG 2000 compress images, retain colors of text and create PDF/A compliant files.



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, IRISDocument **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



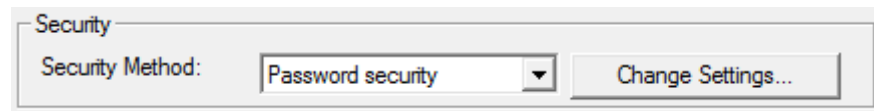
be selected when you have chosen PDF Text.

## Password-protected PDF

Next to regular PDF output, IRISDocument offers **password-protected PDF** output.

### To apply password-protection:

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

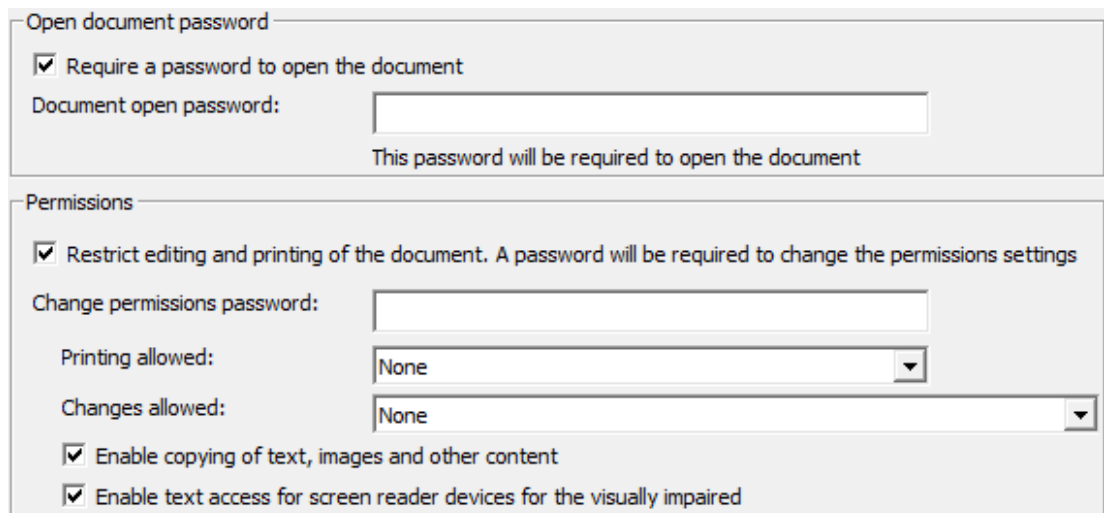


- Select **Password security** in the **Security Method** drop-down list.

The **Change Settings** button becomes available.

- Click it to change the password security settings.

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

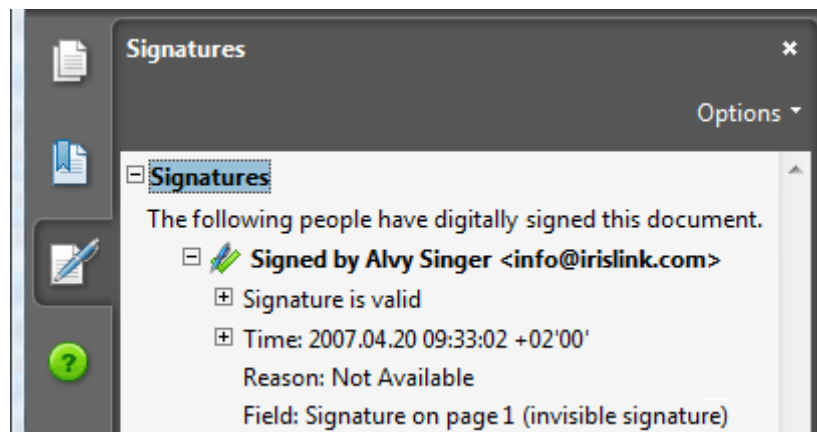


## Digitally signed PDF

Next to regular and password-protected PDF output, IRISDocument 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.

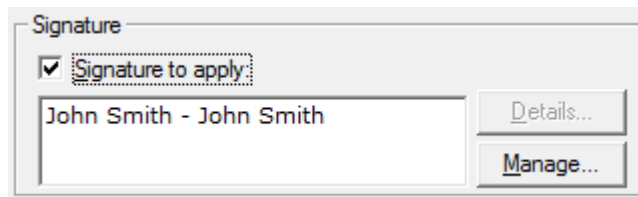
The author signature is invisible: it appears in the **Signatures** tab of Adobe Acrobat and Adobe Reader. To ensure legibility of all scanned information, IRISDocument does not place a signature on the pages of recognized documents.



**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, IRISDocument 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 IRISDocument 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, IRISDocument 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, IRISDocument now offers PDF/A files **without embedded fonts**. As the text is placed beneath the image, no font embedding is necessary. This way, IRISDocument 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 files are not compatible with iHQC Level III or with custom compression.

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

## PDF - iHQC

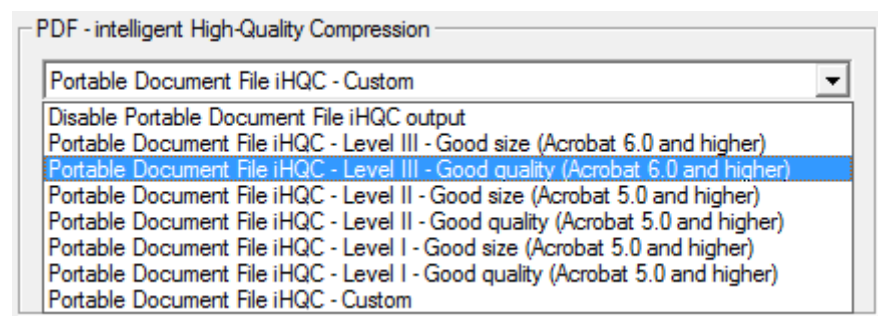
Next to four types of regular PDF files, IRISDocument also offers two types of PDF-iHQC output: PDF Image-text and PDF Image.

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.

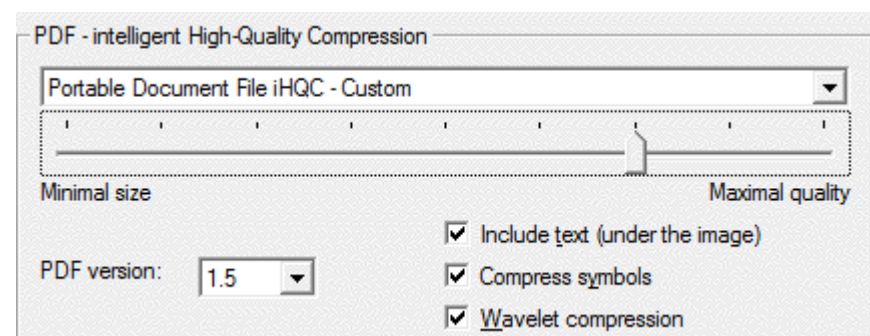
Note that at least Adobe Acrobat 8 is recommended for viewing PDF-iHQC documents.

### 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.



- Select the **PDF version** you want to generate: 1.4, 1.5, 1.6 or 1.7.

## Notes

- If you want to generate PDF/A compliant files, select **PDF version 1.4**.

Note that PDF 1.4 documents are not compatible with certain options:

- PDF version 1.4 is compatible with iHQC Level I and II, but not with Level III or custom compression.
- PDF version 1.4 is not compatible with JPEG 2000 compression.
- PDF version 1.4 is not compatible with Wavelet compression (which applies special compression to graphics).
- In case you are using Adobe Acrobat to view PDF files:
  - It takes Adobe Acrobat 5.0 and higher to open PDF 1.4 documents.
  - It takes Adobe Acrobat 6.0 and higher to open PDF 1.5 documents.
  - It takes Adobe Acrobat 7.0 and higher to open PDF 1.6 documents.
  - It takes Adobe Acrobat 8.0 and higher to open PDF 1.7 documents.

## Options

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

## XPS

---

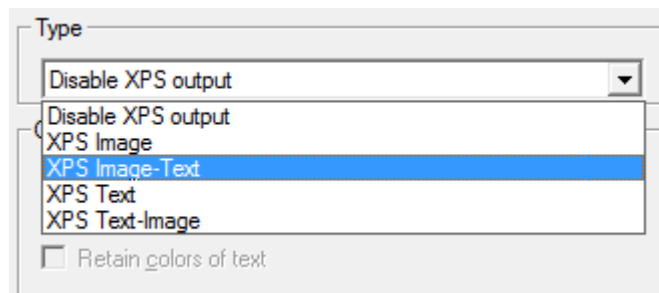
### XPS Document types

IRISDocument generates four types of XPS files: **Text**, **Text-Image**, **Image-Text** and **Image**.

IRISDocument can also apply iHQC compression (intelligent High-Quality Compression) to maximally reduce the file size of the XPS output.

#### To generate XPS output:

- Open the **Document Output** section and click the **XPS** tab.
- Select the appropriate XPS file format in the **Type** drop-down list:



### XPS Image

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

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

## **XPS Image-Text**

IRISDocument recognizes text and creates searchable XPS 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 consult 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 XPS 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 these XPS types.

## **XPS Text**

IRISDocument recognizes text and creates **searchable** XPS files.

The page image is not contained in the XPS files.

Use **text-only** XPS files to save disk space.

## **XPS Text-Image**

IRISDocument recognizes text and creates **searchable** XPS 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:



19 duim) en het remsysteem, dat voorzien moest zijn op circuitgebruik. De oplossing werd hier geboden door Brembo, in de vorm van weliswaar

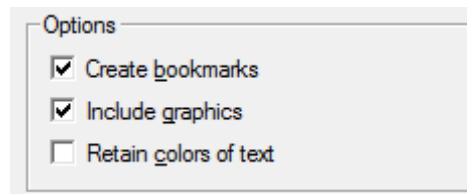
Marche à suivre pour dématérialiser une facture : ouvrir le courrier, sélectionner les factures, en faire des lots et numériser le

## XPS Options

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

If necessary, refer to the section **XPS Document types** to learn how to activate XPS output.

IRISDocument allows you to create bookmarks, include graphics and retain colors of text.



### Creating bookmarks

The option **Create bookmarks** creates bookmarks for each text block, graphic and table in Microsoft XPS files.

### Including graphics

The option **Include graphics** includes graphics in **XPS Text** documents.

This option is always enabled for the other XPS types; it is essential to create a true copy of source documents.

## Retaining colors of text

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

The option is enabled by default for **XPS Text-Image** files and can be enabled for **XPS Text** files.

## XPS - iHQC

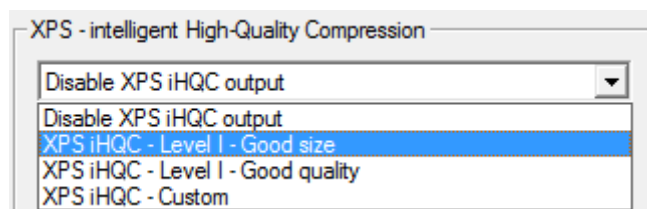
Next to four types of XPS files, IRISDocument also offers two types of XPS-iHQC output: XPS Image-text and XPS Image.

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

**iHQC Level I Good Quality** is provided for free with IRISDocument. To make full use of this compression technology, the **iHQC add-on** is required.

### To generate XPS-iHQC output:

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



## **TEXT-BASED OUTPUT FORMATS**

---

### **Word, WordML, RTF and OpenDocument Text**

IRISDocument offers several types of text-based output formats: it generates versatile **Word (.docx)**, **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 IRISDocument to generate and determine their **layout** and other **options**.

**WordML** is supported by Microsoft Word 2007 and 2003.

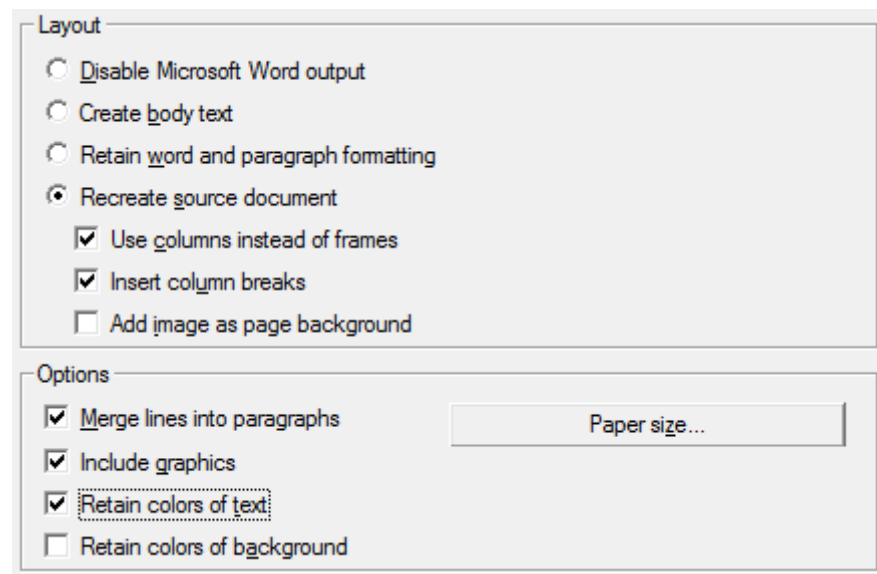
**OpenDocument Text** is an XML-based open format supported by several recent word processors. An open source plug-in is required for Microsoft Word to support this format.

### **Layout and other options**

Numerous layout options are available for the text-based output formats **Word (.docx)**, **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 IRISDocument.
- **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.

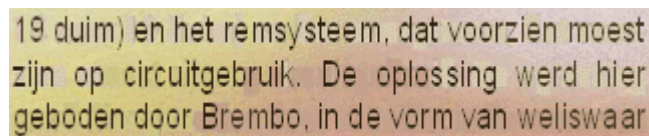
Any text you edit, add or remove, remains inside its column; no text ever flows automatically across a column break.

**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.

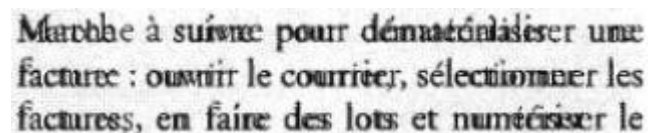
Note that this option is not available for **HTML** output.

- 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:



19 duim) en het remsysteem, dat voorzien moest zijn op circuitgebruik. De oplossing werd hier geboden door Brembo, in de vorm van weliswaar



Marche à suivre pour dématérialiser une facture : ouvrir le courrier, sélectionner les factures, en faire des lots et numériser le

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 IRISDocument 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.

IRISDocument 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.

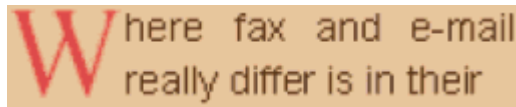
This is essential to create a true copy of a document.

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

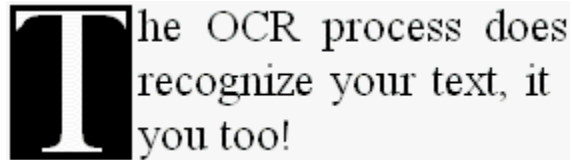
Use the **image enhancement** options of the **Processing** section to improve the image quality and reduce the file size.

- **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	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		

## Preferred paper sizes

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

IRISDocument 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

IRISDocument offers 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.

IRISDocument 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.

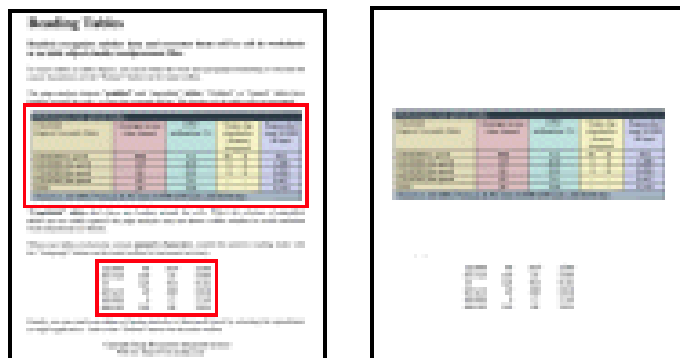


- The option **Retain colors of background** recreates the background color of each cell.

	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 are then 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

IRISDocument offers 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.

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

## **HTML**

IRISDocument offers HTML output.

### **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**

IRISDocument offers **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, IRISDocument 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 9

# EXPORT FEATURES

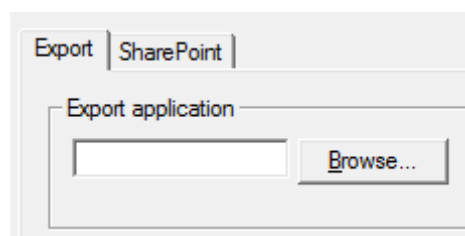
After you have processed your documents with IRISDocument, they can be exported automatically to other applications by means of the custom export feature.

The **custom export feature** connects IRISDocument to external applications that, for instance, open the processed documents in a viewer, import them in an imaging system, republish them on an intranet server, inform users by e-mail that new material is available, etc.

To export documents to other applications, the custom export feature makes use of the XML index that IRISDocument generates by default. As a result, you need to select custom export applications that are capable of parsing the XML index files generated by IRISDocument.

### To select an export application:

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



Any external application that is capable of parsing XML index files is supported. Select **Internet Explorer** (Iexplore.exe) for instance.

Internet Explorer will parse the XML index and will display the documents as soon as they have been processed by IRISDocument.

Make sure you did not clear the option **Enable XML indexing** on the **XML indexing** tab in the **Batch Output** section.

Note that you can develop **custom export applications** that parse the XML index and execute additional tasks on your documents.

- Click the **Run** command on the **File** menu to process and export your documents.

IRISDocument will process the documents and transfer the file name of the XML index to the external program on the command line.

# INDEX

<b>B</b>	
Bitmaps .....	58
Bitonal images .....	58
Black-and-white images .....	58
<b>C</b>	
Character pitch .....	19
Character recognition.	19
Color images .....	58
Confidence file .....	27
<b>D</b>	
Deskew .....	15
Despeckle .....	15
Detect text orientation	15
Digitally signed PDF	.39
Document naming .....	29
Document output .....	33
<b>E</b>	
Embedded fonts .....	41
Export application .....	61
<b>F</b>	
Font type .....	19
<b>G</b>	
General image compression .....	23
Grayscale images .....	58
<b>H</b>	
HTML .....	56
<b>I</b>	
Image compression ...	23, 24
Image enhancement...	15
Image files .....	58
Index file .....	27
Installation .....	9
<b>J</b>	
JPEG .....	58
JPEG 2000 .....	24



<b><i>L</i></b>		<b><i>R</i></b>	
Languages .....	19	RTF .....	48
Layout options .....	48	<b><i>S</i></b>	
Log file.....	27	Secondary languages..	19
<b><i>M</i></b>		Smoothen color images .....	15
Mixed character set....	19	Software installation ....	9
Multipage TIFF.....	58	SpreadsheetML .....	53
<b><i>O</i></b>		Statistics .....	27
OCR .....	19	Supported languages ..	19
OpenDocument Text..	48	Supported output formats .....	33
<b><i>P</i></b>		<b><i>T</i></b>	
Page range.....	19	Text .....	56
Password-protected PDF .....	38	Text-based output formats .....	48
PDF document types..	33	TIFF .....	58
PDF options .....	36	<b><i>W</i></b>	
PDF/A .....	41	Word.....	48
PDF-iHQC .....	42	WordML.....	48
Pitch .....	19	<b><i>X</i></b>	
Product support.....	11	XML.....	57