USER'S GUIDE

Readiris Pro

© 2004 I.R.I.S. All rights reserved
OCR technology by I.R.I.S.
Connectionist, AutoFormat and Linguistic technology by I.R.I.S.

© 2004 I.R.I.S. All rights reserved

Save Time, No More Retyping!

Congratulations on acquiring Readiris. This software package will undoubtedly be of great help in recapturing your texts, tables and graphics.

As efficient as computers are, you have to key in your information first. If you have ever retyped a 15 page report or a large table of figures, you know how tedious and time-consuming it can be. Use this state-of-the-art OCR package to automatically enter text in your applications and you'll acquire an unprecedented level of efficiency and comfort!

Scan a printed or typed document, indicate the zones of interest - or have the system detect them for you - and execute the character recognition. Documents composed of many pages are processed from start to finish in a single effort. A few mouse clicks beat long hours of work as Readiris converts your paper documents into editable computer files: it's up to 40 times faster than manual retyping.

The wizard guides you through the OCR process comfortably: answer a few simple questions and you'll obtain quick and easy results with Readiris. You can send the reading results directly to your wordprocessor and spreadsheet. To recognize faxes and convert PDF documents, you can drag the image files from the Windows Explorer to the Readiris application window. Or right-click on an image to send it prompty to Readiris.

Readiris recognizes tabular data and recreates them as worksheets or as table objects inside your wordprocessor; your numeric data are immediately ready for further processing.

Based on the Connectionist technology from I.R.I.S., Readiris represents the best OCR has to offer. Font-independant feature extraction is complemented by self-learning techniques derived from a proprietary neural network. The system can learn new characters through context analysis: linguistic knowledge about syllables and words improves the OCR performance.

Readiris supports up to 107 languages: all American and European languages are supported, including the Central-European languages, the Baltic languages, Greek and the Cyrillic ("Russian") languages. (Optionally, you can read four

Asian languages - Japanese, Simplified and Traditional Chinese and Korean.) Readiris even copes with mixed alphabets: the software detects "Western" words that pop up in Greek, Cyrillic and Asian documents - many untranscrible proper names, brand names etc. are written using the Western symbols.

Readiris uses linguistics *during* the recognition phase, not after it. As a direct result, Readiris recognizes documents of all kinds with top accuracy, including low-quality documents, faxes and dot matrix printouts. It copes beautifully with badly scanned and copied documents containing too light or dark font shapes. Joined characters ("ligatures") are resolved and fragmented forms, such as dot matrix symbols, are recomposed.

User verification in pop-up style not only flags doubtful characters but also increases the system's precision. All solutions confirmed by the user are memorized, increasing speed and confidence as you go along. Using Readiris means rendering it more intelligent each time! This powerful learning tool allows you to train Readiris on special characters such as mathematic symbols and dingbats but also to handle distorted fonts as you will find in real documents.

To increase your productivity further, Readiris not only recognizes your texts, but can *format* them for you as well! Make use of "autoformatting" and Readiris recreates a facsimile copy of the scanned document: the word, paragraph and page formatting of the original document are retained.

Similar typefaces are used, the point sizes and typestyles as used in the source document are maintained across the recognition. The placement of columns, text blocks and graphics follows your original documents. And as Readiris supports greyscale and color scanning effortlessly, you can recapture any graphics - be they lineart, black-and-white photos or color illustrations. When a document contains tables, Readiris reorganizes them in real cells and recreates the cell borders of the original tables.

In other words, Readiris allows you to archive a true copy of your documents, be it editable and compact text files instead of scanned images! Various levels of formatting are available, the choice is up to the user.

Readiris supports a wide range of popular scanners: numerous flatbed scanners, sheetfed scanners, "all-in-one" devices or "MFPs" ("multifunctional peripherals") and digital cameras can be used. Readiris also supports the Twain scanning standard and some scanning platforms.

TABLE OF CONTENTS

Credits and Copyrights

The Readiris software is designed and developed by I.R.I.S. OCR, Connectionist, AutoFormat and Linguistic technology by I.R.I.S. I.R.I.S. detains the copyrights to the Readiris software, the OCR technology, the linguistic technology, the on-line help system and this manual.

AutoFormat, Cardiris, Connectionist, the I.R.I.S. Linguistic Technology, the I.R.I.S. logo and Readiris are trademarks of I.R.I.S.

XML parser developed by Apache. This product includes software developed by the Apache Software Foundation (www.apache.org).

Acrobat and Reader are (registered) trademarks of Adobe. AsianBridge is a trademark of TwinBridge. AsianSuite is a trademark of UnionWay. Excel, Windows and Word are registered trademarks of Microsoft. Intel is a registered trademark of Intel.

Chapter 1
INSTALLATION

This chapter discusses the system requirements and installation of the Readiris software.

SYSTEM REQUIREMENTS

This is the minimal system configuration required to use Readiris:

- a 486 based Intel PC or compatible. A Pentium based PC is recommended.
- 32 MB RAM. 64 MB RAM is recommended to process greyscale and color images.
- 110 MB free disk space. 95 MB of disk space suffices when you leave the sample files on the CD-ROM.
- the Windows XP, Windows ME, Windows 2000, Windows 98 or Windows NT 4.0 operating system.

Note that some **scanner drivers** may not work under the latest Windows version(s). Refer to the documentation supplied with your scanner to see which platforms are supported.

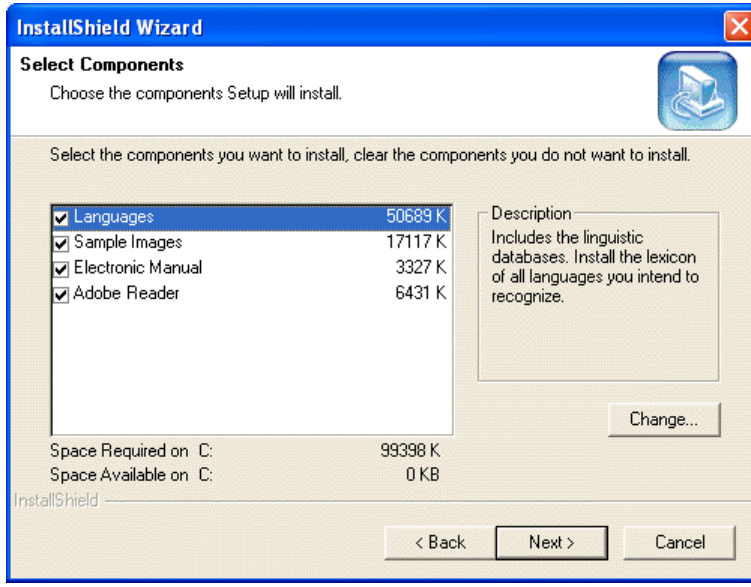
INSTALLING THE READIRIS SOFTWARE

The Readiris software is delivered exclusively on an **autorunning CD-ROM**. To install, simply insert the CD-ROM in your CD-ROM drive and wait for the installation program to start running. Follow the on-screen instructions.

Should the installation not begin to run when the CD-ROM is inserted in your CD-ROM drive, run the setup program MENU.EXE to install the software.

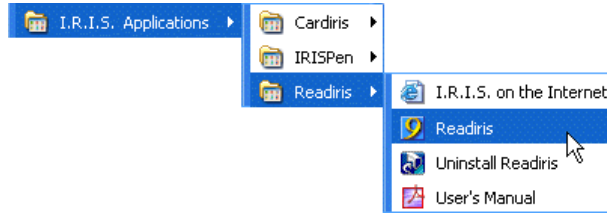
Users of Windows XP, Windows 2000 and Windows NT must ensure that they have the necessary **access rights** - contact the system administrator if necessary.

Some installation options are offered. Be sure to install the **linguistic databases** of all languages you intend to read. By default, all lexicons are installed. You are recommended to install the **sample images** which are used in the tutorials of this manual.



Similarly, install the Adobe Reader software required to access the software documentation, should this be necessary. The **electronic manual** is by default copied to your hard disk. You can also leave it on the CD-ROM.

The submenu "I.R.I.S. Applications - Readiris" under the "Programs" menu is created automatically by the installation program.



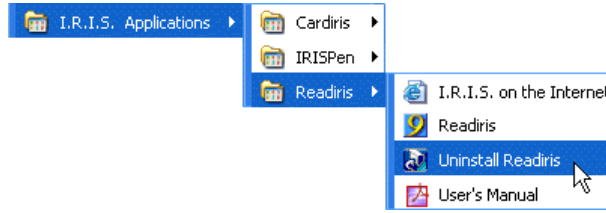
The same holds for a **shortcut** to Readiris on the Windows desktop. As a result, you are able to start Readiris directly from your desktop.

UNINSTALLING THE READIRIS SOFTWARE

There are only two correct ways of uninstalling Readiris: using the Readiris "uninstall" program and using the Windows (un)install wizard. You are strongly recommended *not* to uninstall Readiris or its software modules by manually erasing the program files.

Readiris "uninstall" program

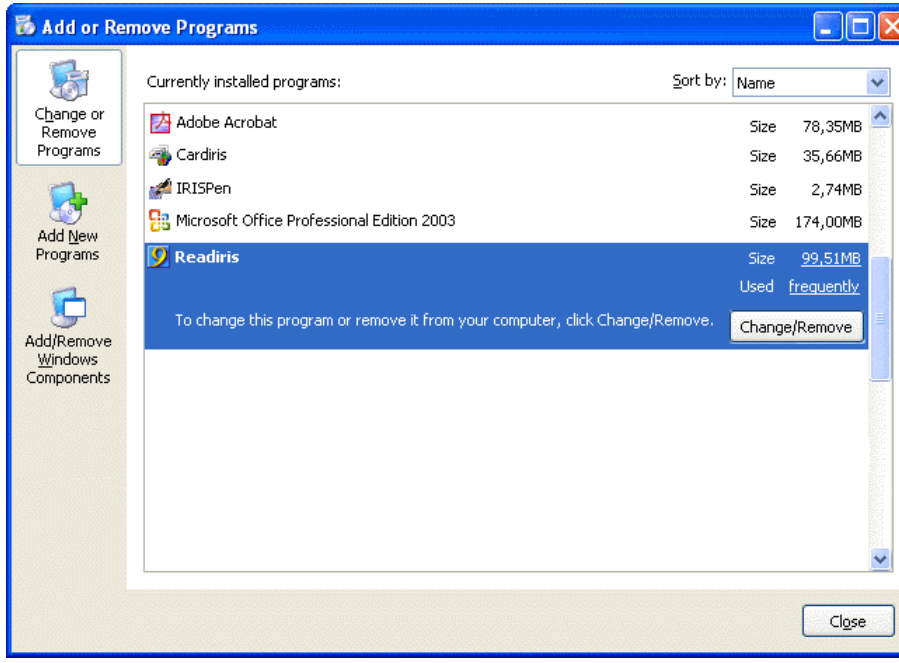
Select "Uninstall Readiris" under the submenu "I.R.I.S. Applications - Readiris" to start the Readiris "uninstall" program and follow the on-screen instructions.



Windows (un)install wizard

Execute the following steps to make use of the Windows (un)install wizard.

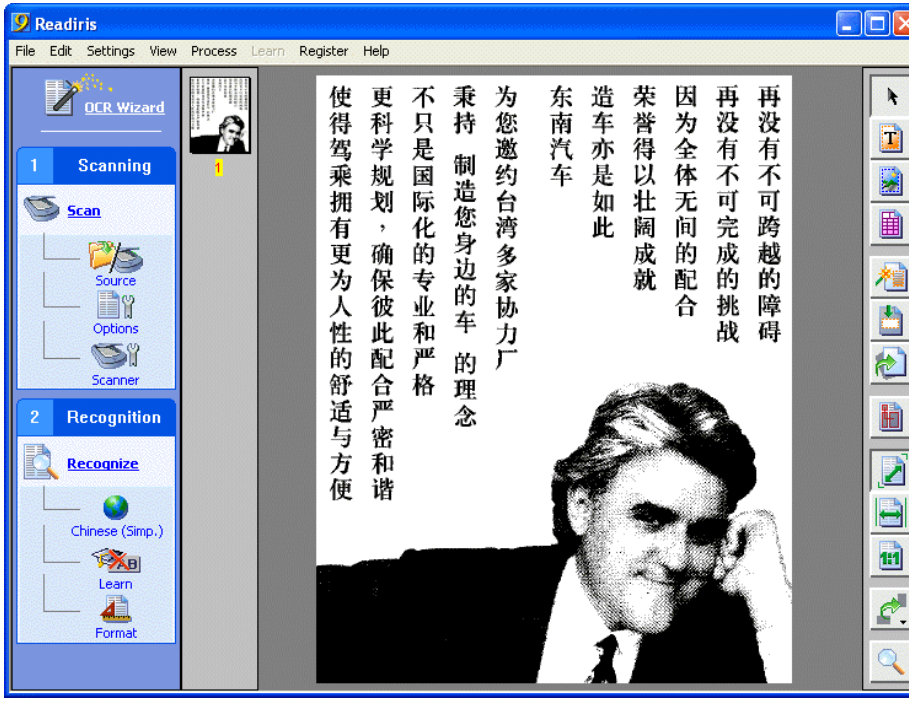
- ❑ Click "Settings" under the "Start" menu of Windows and go to the "Control Panel".
- ❑ Click the icon "Add/Remove Programs" under the control panel.



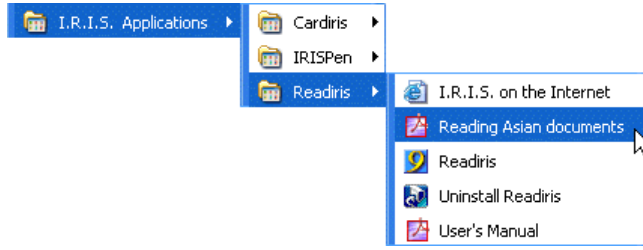
❑ Follow the on-screen instructions to remove the Readiris software.

INSTALLING SOFTWARE OPTIONS

There's a single software option available for the Readiris software: the "Asian OCR add-on". It allows you to read Japanese, Traditional Chinese, Simplified Chinese and Korean. This software is again delivered on an autorunning CD-ROM.



By installing this option, specific documentation becomes available that discusses how you can recognize Asian documents.

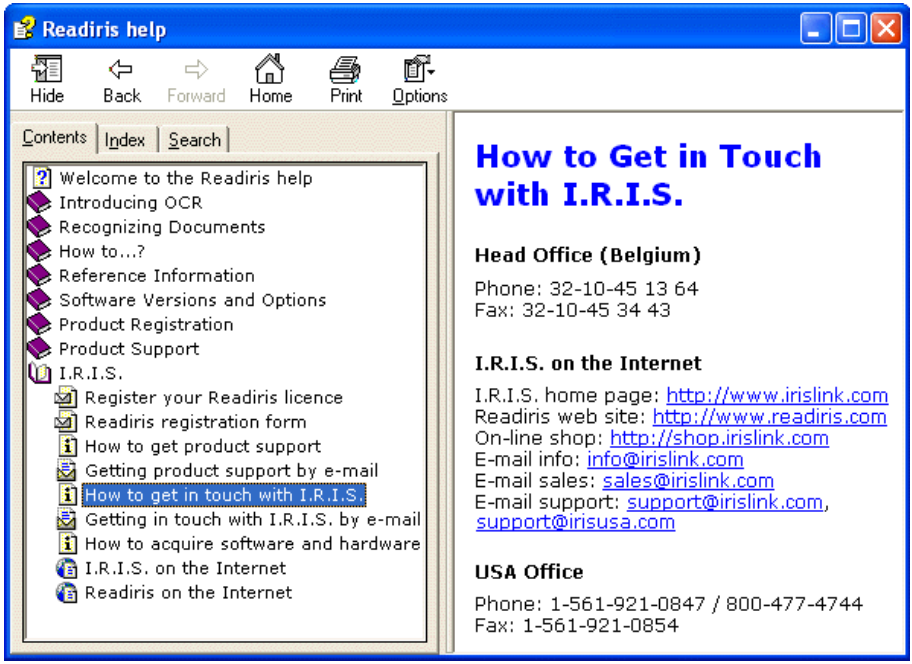


INSTALLING RELATED PRODUCTS

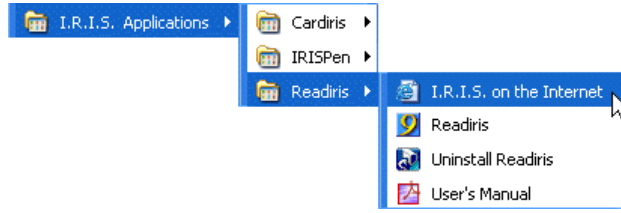
Depending on the software bundle you acquired, Readiris may be supplied with an evaluation version of the related product Cardiris, a **business card organizer**.

If this free software package is included on your Readiris CD-ROM, it is also installed using the autorunning CD-ROM and following the on-screen instructions.

Contact I.R.I.S. to learn more about complementary software; the command "Contact I.R.I.S." under the "Help" menu of Readiris details in which ways you can get in touch with I.R.I.S.



An application icon in the submenu "I.R.I.S. Applications - Readiris" under the "Programs" menu takes you directly to the I.R.I.S. **home page**. So does the Readiris startup screen and the command "I.R.I.S. on the Internet" under the "Help" menu of Readiris.



INSTALLED FILES

The installation program has created a folder where the Readiris files are located. Never try to uninstall Readiris or some of its modules by manually erasing the program files, use the Readiris "uninstall" program or the Windows (un)install wizard instead. See above.

Read Me file and documentation

README.HTM	"Read Me" file (in HTML format)
MANUAL.PDF	User's manual (in Adobe Acrobat format)

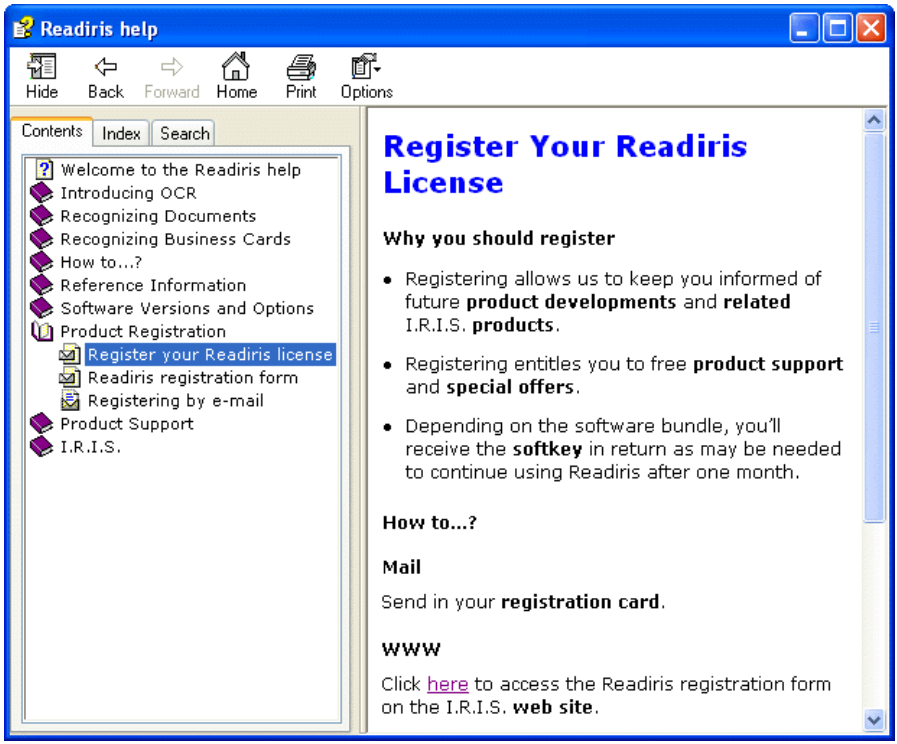
Scanner drivers

Don't hesitate to contact your scanner manufacturer or its representative should problems with scanner drivers continue. Most manufacturers allow you to download the latest versions of the scanners drivers from their web site.

REGISTER TO VOTE!

Don't forget to register your Readiris license! Doing so will allow us to keep you informed of future product developments and related I.R.I.S. products. The registration benefits, including free **product support** and **special offers**, are strictly limited to registered users.

You can register in many ways: by sending in your registration card or faxing its electronic counterpart, by calling I.R.I.S. during working hours and by filling out a registration form on the I.R.I.S. web site!

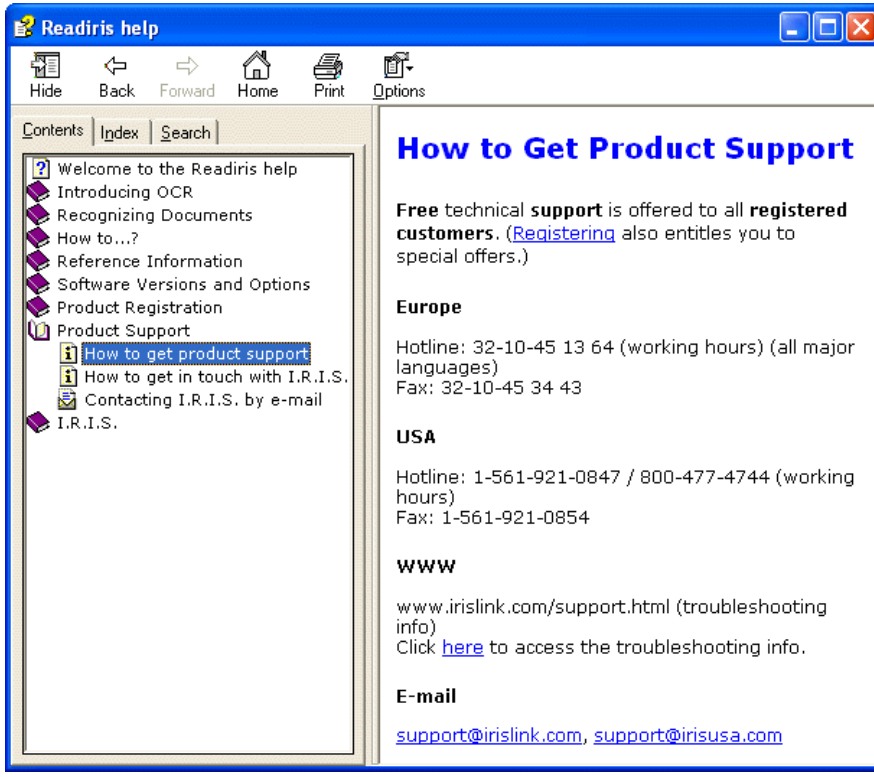


The Readiris **registration wizard** as you'll find under the menu "Register" of the Readiris software can guide you through the registration process comfortably.

Depending on the software version you acquired, you'll receive the **softkey** in return as may be needed to continue using the Readiris software after one month.

GETTING PRODUCT SUPPORT

The command "Product Support" under the "Help" menu of Readiris details how you can get technical support. Please describe the phenomenon you experience clearly and include all relevant data concerning Readiris, your scanner and your computer system.

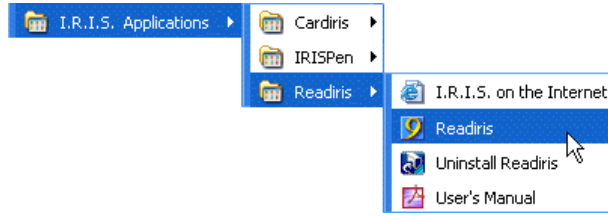

Readiris help
Hide
Back
Forward
Home
Print
Options
Contents
Index
Search
Welcome to the Readiris help
Introducing OCR
Recognizing Documents
How to...?
Reference Information
Software Versions and Options
Product Registration
Product Support
How to get product support
How to get in touch with I.R.I.S.
Contacting I.R.I.S. by e-mail
I.R.I.S.
How to Get Product Support
Free technical support is offered to all registered customers. (Registering also entitles you to special offers.)
Europe
Hotline: 32-10-45 13 64 (working hours) (all major languages)
Fax: 32-10-45 34 43
USA
Hotline: 1-561-921-0847 / 800-477-4744 (working hours)
Fax: 1-561-921-0854
WWW
www.irislink.com/support.html (troubleshooting info)
Click here to access the troubleshooting info.
E-mail
support@irislink.com, support@irisusa.com

Chapter 2
GUIDED TOUR

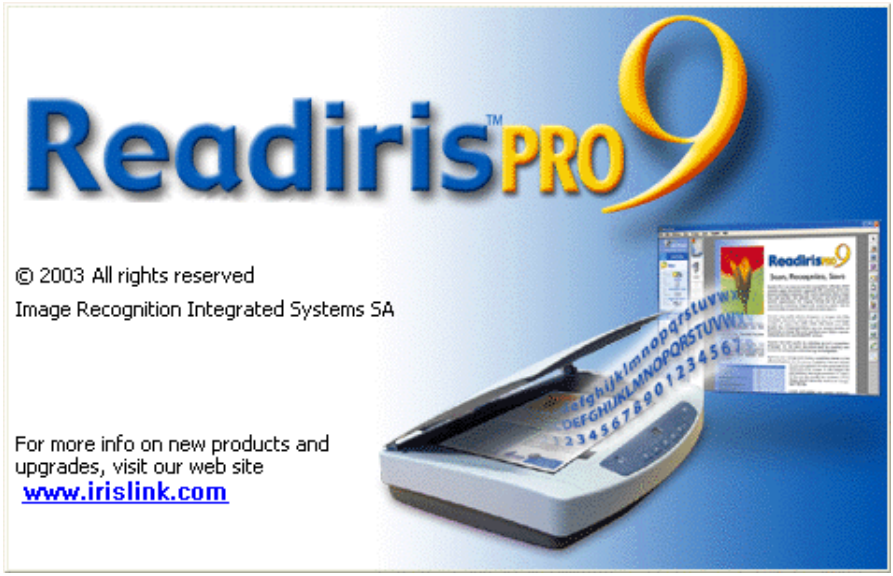
Readiris is a state-of-the-art OCR package equipped with numerous advanced features. We will discuss all major features in this chapter and add many tips and hints concerning the use of Readiris.

STARTING THE SOFTWARE UP

Click on the Readiris application in the submenu "I.R.I.S. Applications - Readiris", or click on the shortcut to the Readiris application on your desktop.



The Readiris startup screen and application window are displayed. The startup screen displays the version and copyrights of the Readiris software. It also gives direct access to I.R.I.S.'s **home page** - simply click on the URL to visit the I.R.I.S. web site. Clicking the mouse anywhere else makes this screen disappear.



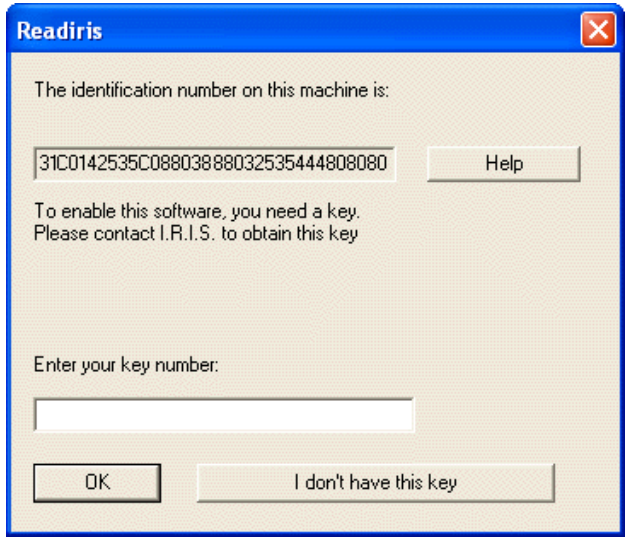
The next window concerns the OCR wizard; click "Cancel" for the time being.

The First-Time Startup

Depending on the software bundle you acquired, the first startup may be special: you may be prompted to register your licence.

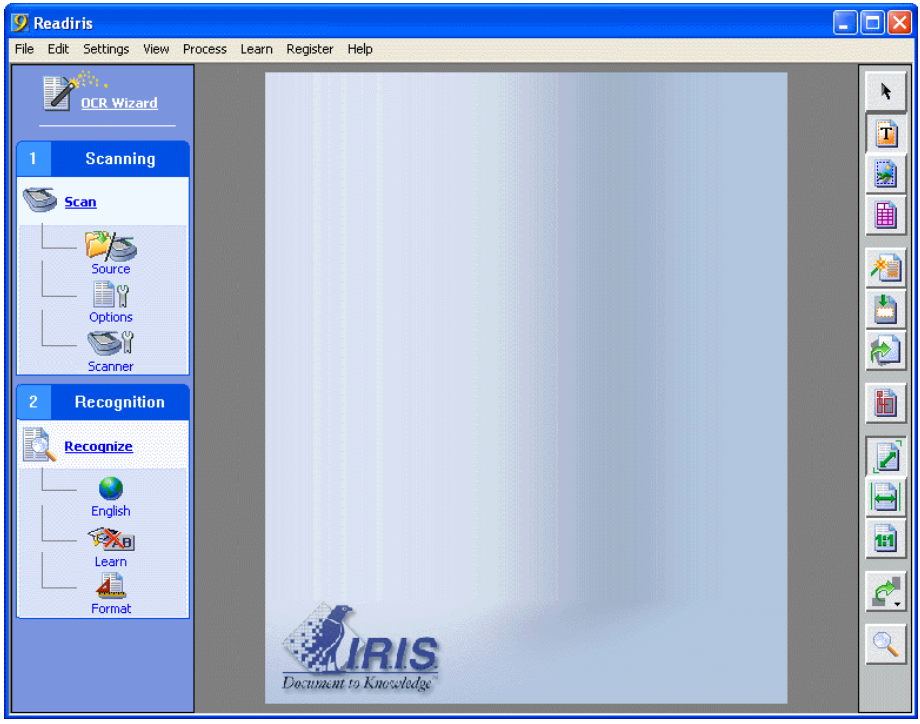
If this is the case, the use of Readiris is limited to 30 days, and by registering, you receive a free **softkey** from I.R.I.S. to continue using the software after the first month.

It takes your **identification number** to generate the softkey; be sure that this number is available or mentioned when you register your licence.



DISCOVERING THE READIRIS INTERFACE

The Readiris application window not only contains **command menus** but also two button bars that give quick access to all frequent commands. Initially, some command menus are dimmed: they concern the preview. As long as no image is opened, they are unavailable.

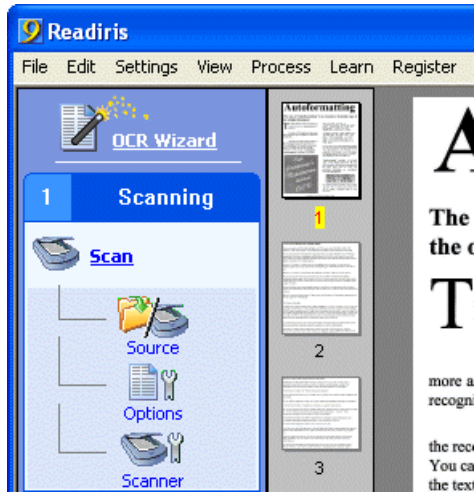


The same goes for the **image toolbar** on the right side of the application window: it contains all commands you need during the image preview. The **main toolbar** on the left gives quick access to all frequent general commands.

To learn which command corresponds to a certain button, hold your mouse pointer over it for a while: a **tooltip** will tell you what the button does.

The window pane or **image zone** is where the scanned images are displayed. You can drop image files onto the image zone (and on the Readiris icon) to recognize them.

As soon as pages gets processed, an additional toolbar, the **page toolbar**, is added on the left side: it represents the various pages of the document and gives access to the page commands using the right-click (the "Context" menu).



Getting Started with a First Tutorial

The best way to become familiar with the operation of Readiris is undoubtedly by using it. A number of **prescanned images** is provided with the software; they allow you to get started even when there is no scanner connected to your computer. Let's turn to these now.

The "Source" button on the main toolbar determines whether you are going to use a scanner or a prescanned image as image source.

Color, greyscale and black-and-white images are supported on an equal basis. Readiris allows you to open Adobe Acrobat PDF documents, JPEG images, Paintbrush (PCX) images, DCX fax images (a multipage version of the Paintbrush format), PNG images, TIFF images (uncompressed, LZW, PackBits, Group 3 and Group 4 compressed), multipage TIFF images and Windows bitmaps (BMP).

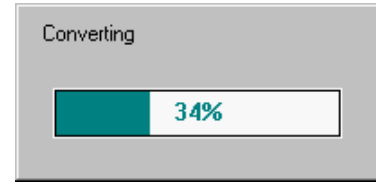
This capability is particularly useful to convert your **faxes** into editable text files.

As you are going to open a prescanned image, you should select "Image Files", and not the scanner, as image source with the "Source" button.

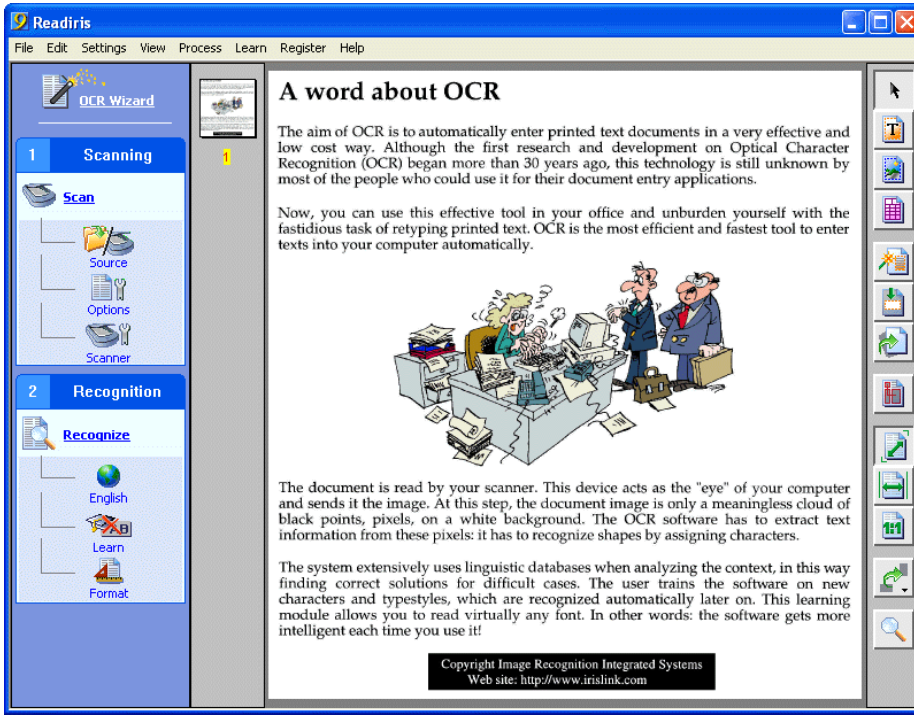
Next, click the "Open" button. (When you select the disk as image source, the "Scan" button is replaced by the "Open" button and the corresponding "Scan" command under the "Process" menu is replaced by the "Open" command.)

You could also select the command "Open" from the "File" menu and open a prescanned image directly - this works even if your scanner operates as current image source.

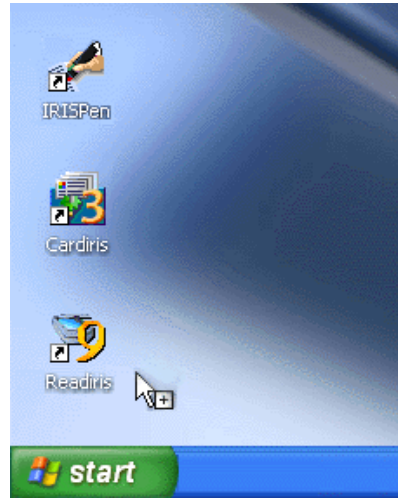
You are invited to select an image file. Select the file ENGLISH.JPG in the Readiris folder. As this sample file is a color image, it is not only read from disk: a "binarized", black-and-white version is created for the OCR process.



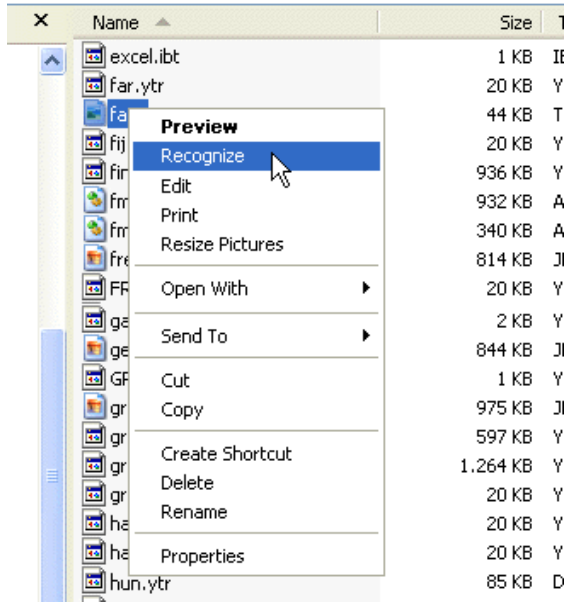
Finally, the image is displayed in the image zone. The page toolbar indicates that a single page is loaded into Readiris.



A third way of opening prescanned images is the use of **“drag and drop”**: drag images from the Windows Explorer onto the Readiris image zone or on the Readiris icon and they are promptly opened.



You can even open images from within the Windows Explorer: **right-click** an image file and select the command "Recognize" from the "Context" menu. (This command only appears when the file's file type is supported.)



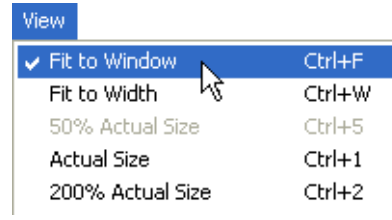
That does not mean the OCR is promptly executed: to give the user full flexibility, Readiris is simply started up and the image is opened.

The image toolbar on the right side of the Readiris application window contains all commands you need during the image preview: tools to indicate the zones of interest, to rotate the image, zoom in and out etc.

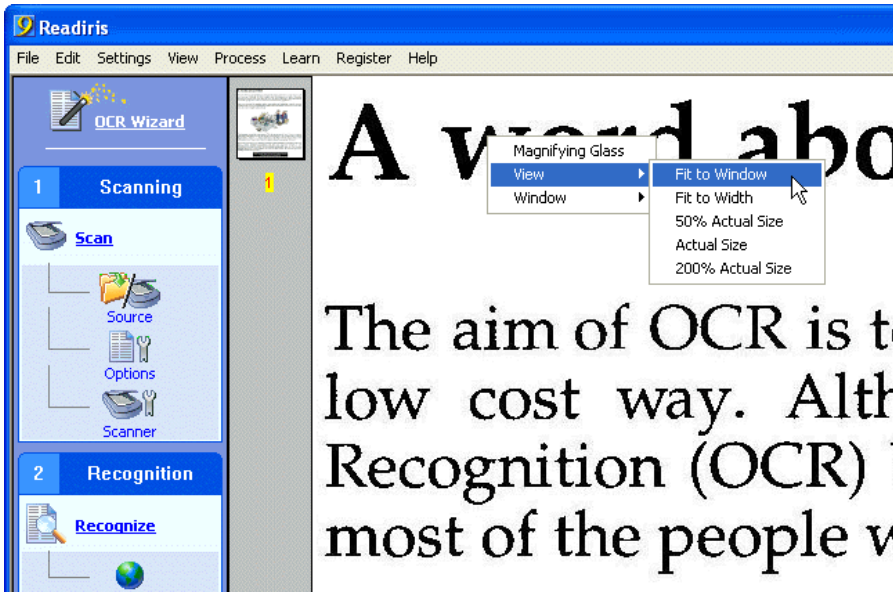
Zooming in on Images

Readiris has several commands that allow you to **zoom** in on a scanned image, for instance to verify the scanning quality.

The image toolbar contains buttons that allow you to zoom in at real size, to fit the image to the page width and to fit the entire image in the preview window. The "View" menu contains the same commands and adds two extra zoom levels: you can display the image at 50% and 200% of its actual size. At actual size, a screen pixel corresponds to an image pixel. (Shortcuts are available for all zoom levels!)

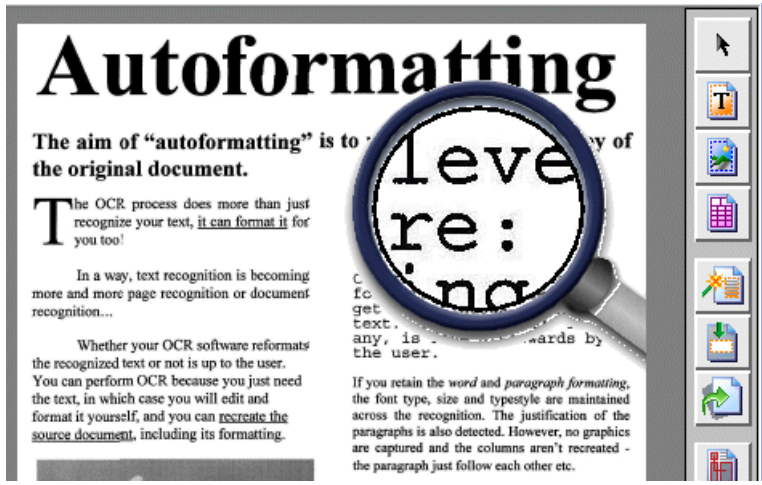


Also notice that the zoom levels are available on the right-click. Click with the right mouse button to invoke the "Context" menu and select the appropriate zoom level.



Furthermore, you can *double*-click the right mouse button over a region of the scanned image to zoom in at real size immediately. Repeat the operation to zoom out again.

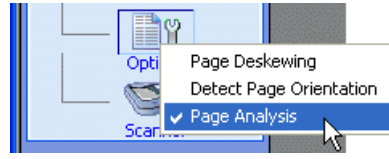
Finally, you can use the **magnifying glass** to zoom in on details of the scanned document. The magnifying glass is also available on the "Context" menu when you right-click the mouse over the image.



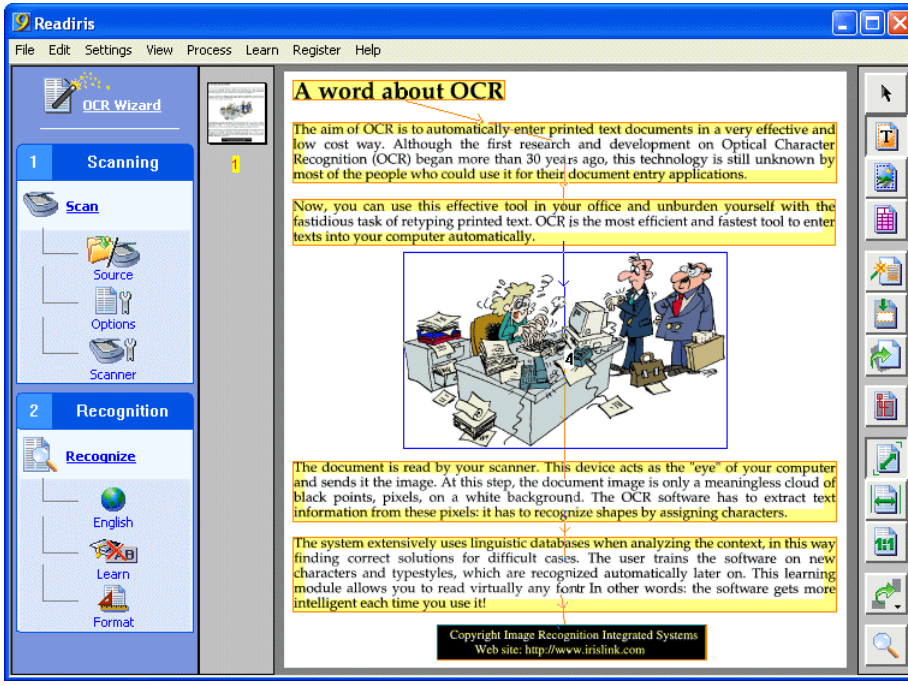
ONE, DECOMPOSING A SCANNED IMAGE

Now that the image is scanned, you have to indicate which parts you want to convert into editable text by drawing frames, so-called "windows", around the zones of interest.

Actually, Readiris will do this for you automatically when the option "Page Analysis" is enabled under the "Options" button on the main toolbar (or under the "Settings" menu).



Automatic page decomposition is particularly useful when **columnized texts** and documents with a complex page layout, possibly including graphics and tables, are recognized.

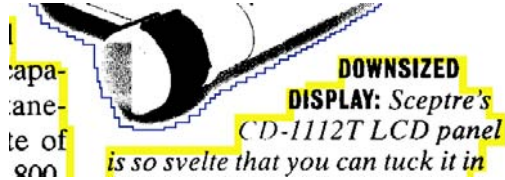


Page decomposition uses three **window types**: text, graphic and table windows. Readiris discriminates text blocks, tables and graphic zones containing photos, illustrations etc. on the page. (Saving graphics and recognizing tables will be discussed at great length below.)

A **color code** indicates the window type: text zones have a yellow border, graphic windows have a blue border and tables a purple border.

The number of windows is indicated at all times in the tooltips of the "Text Window", "Graphic Window" and "Table Window" tools.

Page analyisis is fast, skew-tolerant and highly accurate: it traces complex, "irregular" shapes.



The page analysis will even detect zones where you get **white text on a black background**. Recognizing such inserts is no problem: while the preview displays the scanned document correctly on-screen, Readiris "inverts" the image when the need arises to recognize such text blocks! (You can have your scanner generate *fully* inverted images to process pages with white text on a black background. See below.)

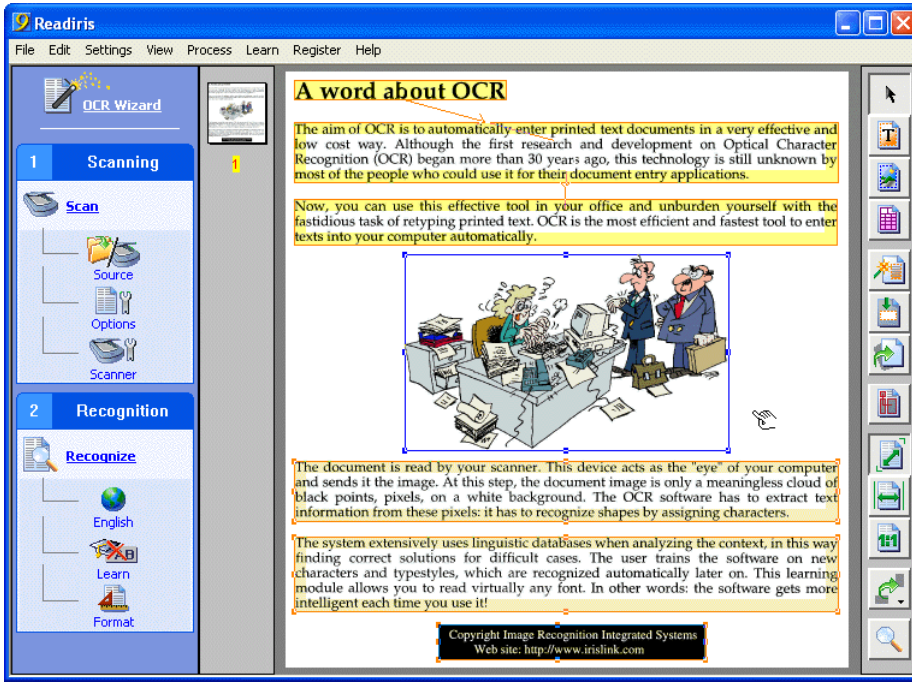
ONE AND A HALF, SORTING WINDOWS

Readiris not only detects the various blocks, but also *sorts* them: the zones are sorted top-down, left to right by default to cope with columnized documents.

Evidently, you can modify the **sort order**. To do so, click the "Sort" button on the image toolbar. The mouse cursor becomes a pointing hand as soon as the "sort mode" is enabled.

Click on the windows you want to include. Windows you do *not* click on are simply ignored, excluded from recognition. It's easy to see which windows are

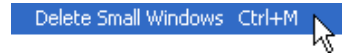
selected and which aren't: the selected windows have their full color, non-selected windows have a lighter color tone.



Page analysis is enabled by default. To force Readiris to decompose the current page - because you disabled page analysis by accident, because you erased some windows erroneously and want to redo the page analysis etc. -, you can simply click the button "Analyze Page" on the image toolbar.

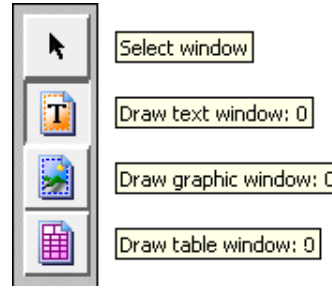
Select the document language *before* executing the page analysis when you are dealing with Asian documents. Specific routines are used for these languages: the interline spacing of Asian documents is in most cases bigger than in Western documents, the text is made up of small icons ("ideograms") that could easily be seen as graphic zones in Western documents and the text may run from top to bottom, from right to left. And if you forgot to select the proper language, select it afterwards. Readiris re-executes the page analysis automatically!

Some documents have many "stray" dots on the page, may generate a black page border around the actual image etc. To erase all small windows - it's assumed they don't contain any text - and re-sort the remaining zones, you can click the command "Delete Small Windows" under the "Edit" menu.



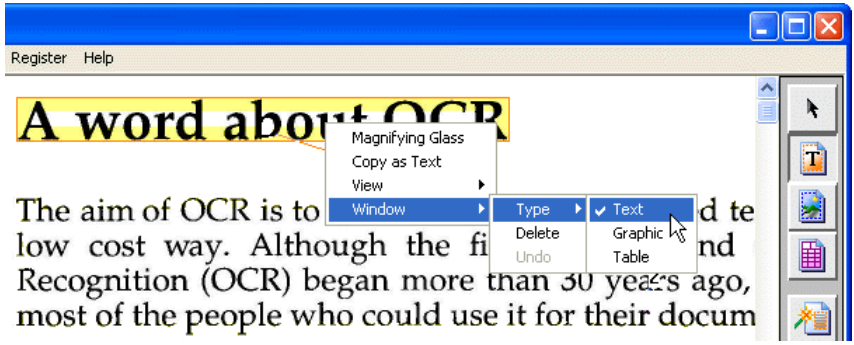
TWO, WINDOWING A SCANNED IMAGE MANUALLY

Page analysis is the automatic way of windowing a scanned page. Alternatively, you can zone an image manually with the **windowing tools** of Readiris.



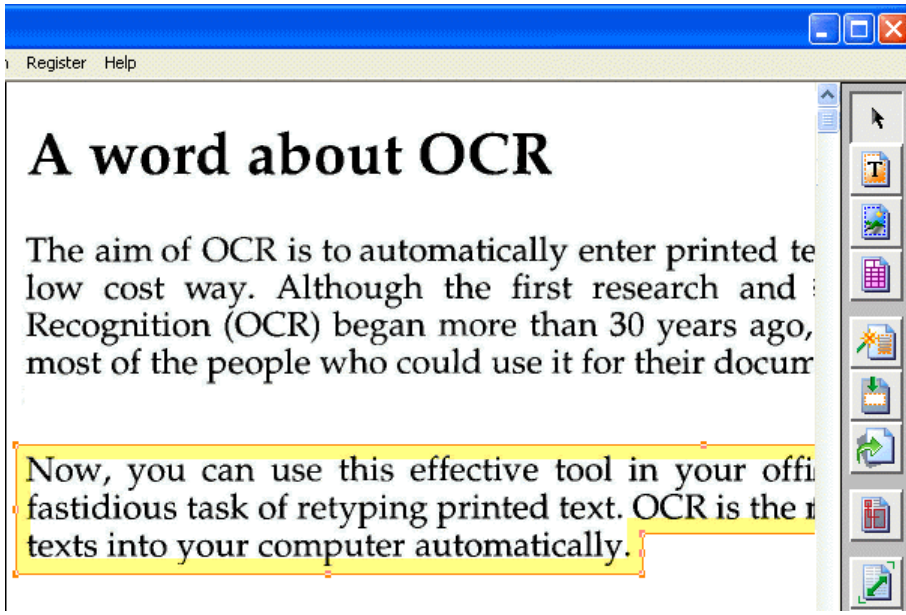
To **draw** a rectangle around a zone of interest, select the corresponding tool in the image toolbar, click the cursor in the upper left corner of the window, stretch the window by moving the mouse to the lower right corner and click again. (Sides smaller than 1 mm are not allowed, they wouldn't even contain a single character anyway.)

Not to worry should you have selected the wrong zone type: you can quickly change the type by right-clicking the mouse over a window and selecting the command "Window - Type" from the "Context" menu.



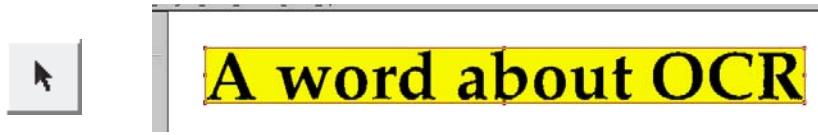
The windows are automatically sorted in the order of creation: arrows indicate the sort order.

You can also frame “irregular” text blocks by drawing **polygonal windows** around them. Non-rectangular windows are created by merging rectangular zones: as soon as two rectangles (of the same type) intersect, they become a single window automatically! In a way, you’re building a house by adding one room after the other... (Creating polygonal table windows doesn’t make any sense.)



Furthermore, manual windowing can be combined with window sorting: you can draw new windows even when the "sort mode" is enabled. You then use sorting to include a number of detected windows and manually create some other windows where the page analysis didn't yield the appropriate results. As soon as you start creating windows in the "sort mode", all zones you didn't select are promptly erased!

To modify, move and delete windows, you need to **select** them first. To do so, select the "Window Selection" or "arrow" tool in the image toolbar and click inside a window. Rectangular markers now appear at each corner and in the middle of the window sides.

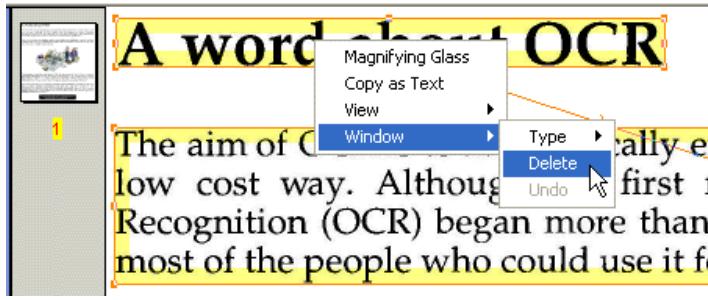


To **unselect** windows, click the mouse button elsewhere. To select **additional windows**, hold down the Shift key while clicking on these extra windows. To select a window and the **included windows** (of another type), hold down the Ctrl key while clicking on the main window.

So much for selecting windows. To **modify** a window, select it, put your mouse cursor over a marker and drag the side to change the window size.

To **move** a window, simply select it and drag it to another location.

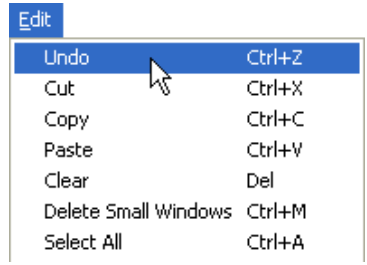
To **delete** windows, select them, right-click them and select the command "Window - Delete" from the "Context" menu. Doing so deletes all selected windows as well as the window under your mouse cursor.



Alternatively, you can select zones and choose the "Cut" or "Clear" command from the "Edit" menu. The "Cut" command cuts the window(s) to an internal buffer, "Clear" erases the window(s) irretrievably. When you paste zones, they are inserted in their original position, and you have to drag them to their new location.

In fact, *all* familiar commands from the "Edit" menu apply to the windows: you can delete, cut, copy and paste them! The "Undo" command also applies: if you

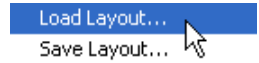
have unfortunately deleted, moved, resized etc. some windows, "Undo" will cancel the last operation.



Also note that shortcuts are available for all commands! Let's give an example: to erase all existing windows, you can choose the command "Select All" or its shortcut Ctrl+A and click the command "Clear" or its shortcut Delete. You are now ready to recreate the necessary layout. To restore the previous layout, you can choose "Undo" or the shortcut Ctrl+Z.

THREE, SAVING WINDOWING TEMPLATES

The resulting windowing layouts can be saved as **zoning templates** for future use with the command "Save Layout" under the "File" menu and loaded into memory with the command "Load Layout".

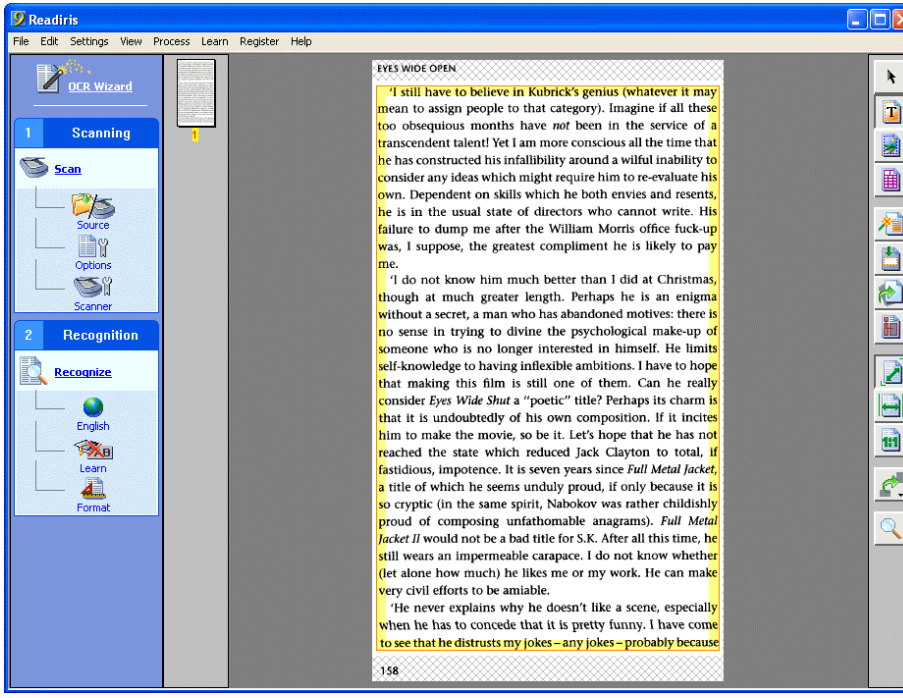


If you have to recognize documents with a similar layout, for instance a 50 page report where the header and footer should be excluded for obvious reasons, a single template can be applied to zone all 50 pages.

When you load a template into memory, page analysis is disabled automatically. The zoning template remains active until you re-enable page analysis on the main toolbar.

Actually, there's a nice alternative for zoning templates: the preview tool "Ignore Exterior Zone" limits the page decomposition to the "cropped" portion of the image.

Select this tool and frame the portion of the image you want to process. When you're dealing with a multipage document, you can exclude the same outer zone from page analysis on every page. (Re-execute the page analysis to cancel the image "cropping", or change the zones manually.)

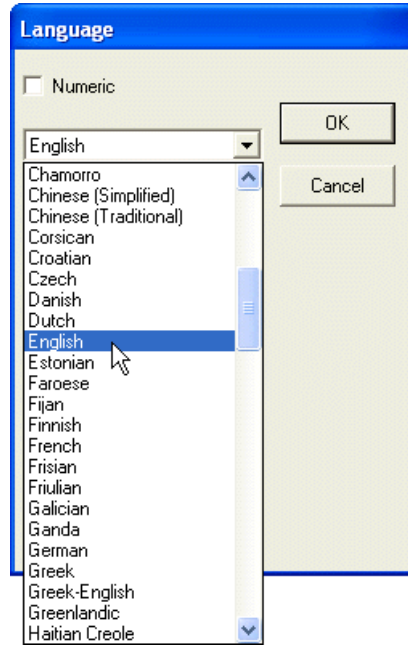


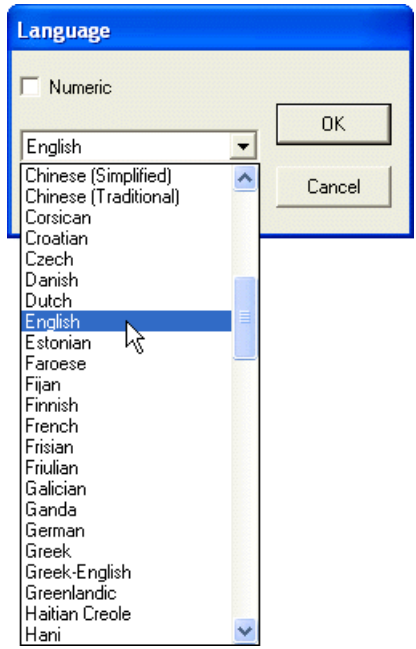
Readiris Takes You around the World

Assuming that the windows are correctly defined, you are now almost ready to execute the character recognition. We say "almost", because we haven't verified the language and document settings yet.

The language setting can be found on the main toolbar.

Click the "Language" button to modify the document language.





You can press a letter key to move to it directly: if English is currently selected, and you want to select Occitan, you can click the "O" key on your keyboard to go directly to the Occitan language. When several languages have the same initial, press the letter several times to go through the options. Let's give an example: Readiris reads English and Estonian. By pressing "E" once, you select English, by pressing "E" a second time, you select Estonian, and by pressing "E" a third time, you're back on English. (To go to *another* letter, say T, press BackSpace before you enter the "T" character.)

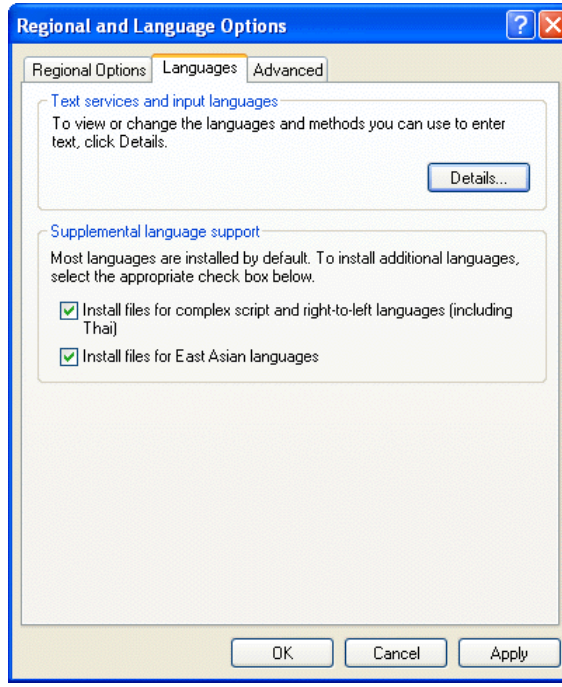
Readiris is far from limited to English: up to 107 **languages** are supported! All American and European languages are supported, including the Central-European languages, Greek, Turkish, the Cyrillic ("Russian") and the Baltic languages.

Optionally, you can read **Asian documents**: the extra module "Asian OCR add-on" offers recognition of Japanese, Simplified Chinese, Traditional Chinese and Korean. (Simplified Chinese is used on China's mainland and in Singapore, where Traditional Chinese is used by Hong Kong, Taiwan, Macau and the overseas Chinese communities.)

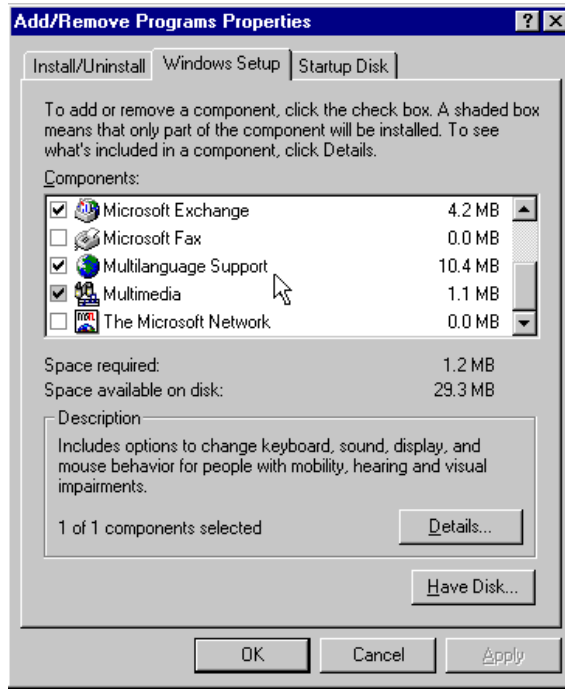
Also note that the British and American - or should we say "international"? - variants of the English language are distinguished.

It takes the appropriate Windows configuration to display Central-European, Greek, Turkish, Cyrillic and Baltic characters. You may have to install the **Windows multilanguage support** before your Windows system is able to cope with these languages.

On a Windows XP, 2000 and Windows NT 4.0 operating system, select the icon "Regional Settings (and Languages)" under the "Control Panel".



On a Windows ME and 98 operating system, select the icon "Add/Remove Programs" under the "Control Panel" to find out if the module “Multilanguage Support” is installed on your PC.



To view and edit Asian documents, you can install an Asian version of the Windows operating system or run specialized "emulating" software (such as UnionWay AsianSuite or TwinBridge AsianBridge) on a Western version of Windows to correctly represent the ideograms of these Asian languages. Finally, you can use Word 2003, Word 2002 or Word 2000 to view and edit such documents: Office 2003 System, Office XP and 2000 were specifically designed to cope with documents in many different languages.

Refer to the Readiris **"Read Me" file** for more information on this subject.

Selecting the proper document language is imperative. Based on the selection of a language, the software knows which **symbol set** to recognize. Multi-linguistic support ensures that "exotic" characters such as ç, ß, ñ, γ and ø are recognized correctly.

Secondly, the software extensively uses **linguistic databases** to validate its results. Suppose that you have to read the word "president" where an ink stain makes the "r" look like an "f". Looking things up in the English lexicon, Readiris will detect autonomously that the word "president" is being read and that it doesn't make any sense to recognize the symbol "f". This **"self-learning" technique** is of course highly dependent on the linguistic context.

Linguistics offer useful help to solve **ambiguous cases** such as an "O" which might be mistaken for a '0'. Another typical example is the letter "l" and number '1' which have an identical form in many fonts - think of texts produced on old typewriters! The linguistic context helps to determine whether you are dealing with "l" or '1'.

The illustration below shows various shapes of '1' and "l". The shapes on the first line are unambiguous, the shapes on the second line are ambiguous, but linguistics can solve them. When the context does not suffice, the user intervenes.

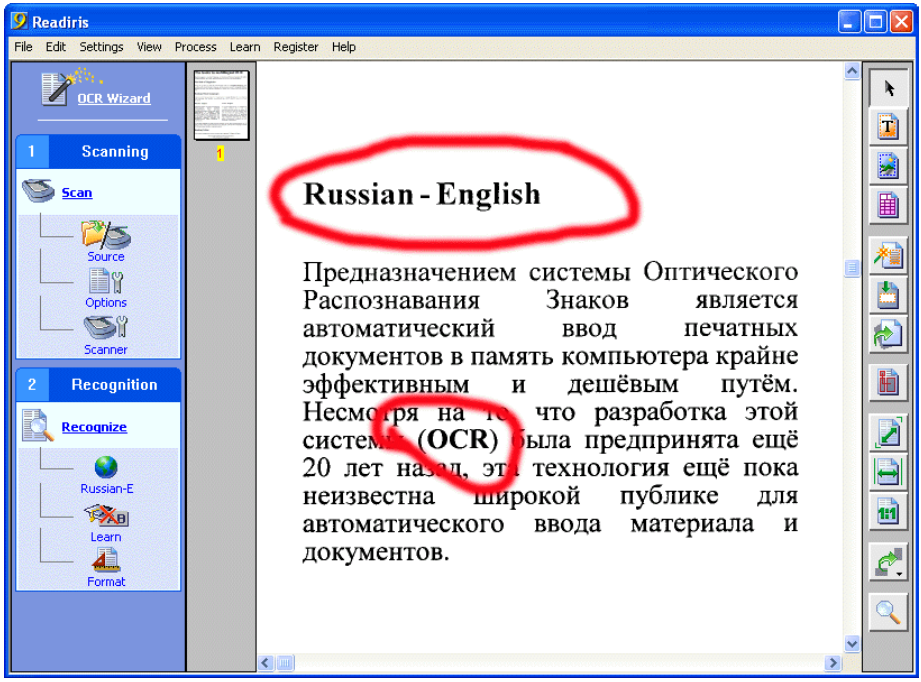
Readiris Changes Languages As Needed

But the buck doesn't stop here: Readiris can switch languages in the middle of a sentence without any help from the user! When Western words pop up in Greek, Cyrillic or Asian documents - many untranscrible proper names, brand names etc. are written using the familiar Western symbols -, Readiris can switch

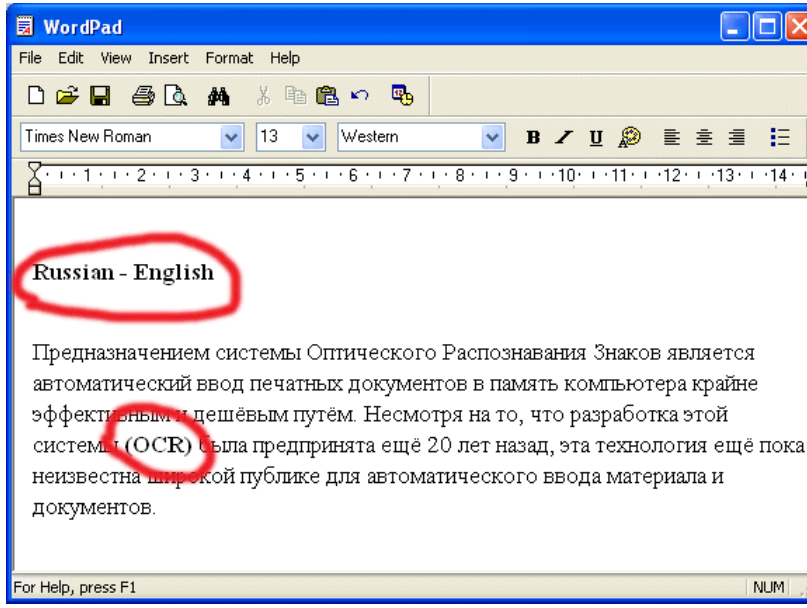
to the correct alphabet automatically. In other words, you can activate a **mixed alphabet** of Greek, Cyrillic or Asian and Western characters.

Be sure to select "Greek-English" or the appropriate Cyrillic language setting - for instance "Byelorussian-English". In other words: don't try to just select "Greek" or "Byelorussian" as document language and hope that the Western symbols will come out fine!

Here's an example where a Russian text contains some English words - open the image file ALPHABETS.TIF if you want to try it for yourself!



The end result looks like this when opened with the wordprocessor - you may have to select a Cyrillic **font** to display the Russian text correctly.



To **mix other languages**, simply select the language with the most extended character set. If you have a document where the, say, French translation is placed alongside an English text, you have to select French as language to ensure that the accentuated characters such as ç, é and ù get recognized correctly.

DEFINING THE DOCUMENT CHARACTERISTICS

Now that the language is set, we'll turn to the other document characteristics. You can fine-tune the recognition by specifying some document features: the font type and character pitch. (These commands do not apply to Asian documents.) Let's clarify what this means.

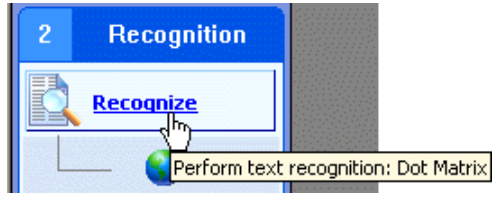
Let's start with the command "Font Type" under the "Settings" menu. The font modes separate "normal" documents from **dot matrix** printed documents. "Draft" or "9 pin" dot matrix symbols are made up of isolated, separate dots, and highly specialized recognition routines are used to recognize them.

ape-descended life

"Letter quality" dot matrix printing, also called "25 pin" or "NLQ" dot matrix, requires the "normal" setting, as do the **printing qualities** typeset, typewritten, laser printed and inkjet printed.

The setting "Automatic" means that Readiris will detect the font mode automatically. Let Readiris "auto-detect" the font mode in all cases - unless you are sure only dot matrix documents are being read! (Obviously, "Automatic" is the default value.)

The font type is indicated in the tooltip of the "Recognize" button: when no message is added to the tooltip, the "auto-detection" of the printing quality applies, when the message "Dot Matrix" shows up in the tooltip, the dot matrix reading mode is enabled.



The **character pitch** can be set with the command "Character Pitch" under the "Settings" menu.

With *fixed* or "monospaced" fonts, all symbols of the font have the same width. `An "i" takes up as much horizontal space on a line as a "w", as is the case in this sentence.` Think of documents produced using a typewriter, where the carriage moves a fixed distance for each typed symbol.

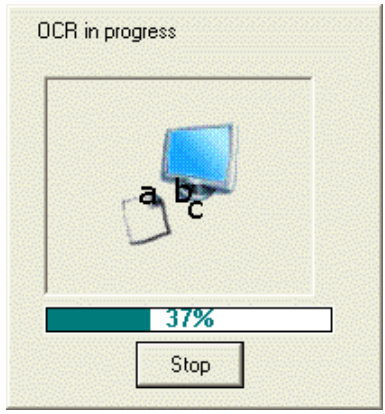
A *proportional* pitch means that the width of a character depends on its shape. Symbols like "m" and "w" are wider, take more horizontal space on a line than the "thin" characters "l" or "j". Virtually all books, magazines and newspapers are printed in proportional pitch.

The simplest solution is to leave this option at all times on the default value "Automatic", which means that Readiris will detect the character pitch automatically.

READIRIS GETS MORE INTELLIGENT EACH TIME!

When the document language is selected and document characteristics are set, enable the interactive learning and click the "Recognize" button.

The OCR progress is indicated on-screen. You can click the "Stop" button to abort the text recognition.



At the end of the recognition, Readiris enters the interactive learning phase when the learning is enabled with the "Learn" button on the main toolbar.

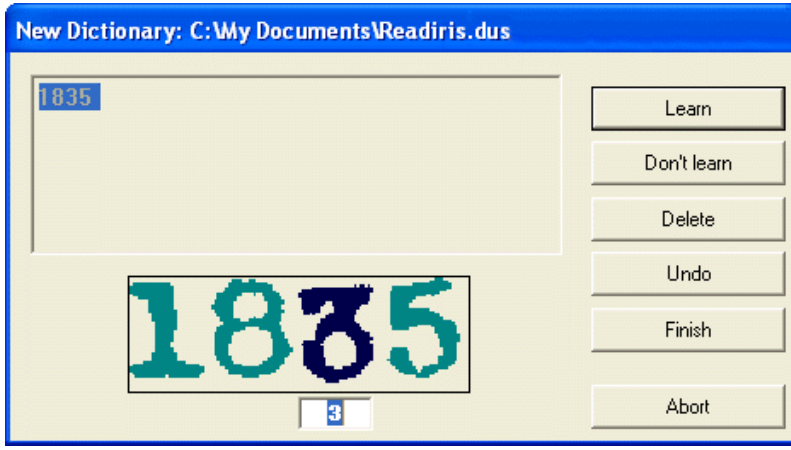
(Interactive learning does not apply to Asian documents: learning does not make sense for these languages which use thousands of different symbols - and you'd have to be able to enter the ideograms, not an easy task when using a Western keyboard!)

Font training can substantially enhance the accuracy of the recognition system. When the user tries to read distorted, defaced forms as are found in real documents or stylized font shapes which Readiris does not recognize optimally, training can overcome this temporary "failure".

User learning is also used to train the system on **special symbols** which Readiris is unable to recognize, such as mathematical and scientific symbols and dingbats. Some examples: Readiris can be trained to recognize the "π" symbol as "pi" or the dingbat "☎" as "Tel". (However, the list of recognized symbols cannot be extended with the symbols "π"and "☎"!)

The recognized text is displayed progressively and the system stops on doubtful characters, or - if you are dealing with touching characters ("ligatures") - on doubtful character strings. They are always presented in their context, the doubt-

ful characters are highlighted. Unrecognized characters are represented by a tilde (the "~" symbol).



The first thing you should do is verify if you activated the correct font dictionary and dictionary mode - these are always indicated in the title of the learning window. If that is not the case, click the "Abort" button - the document image is redisplayed with the zoning as was created -, enable the right font dictionary or dictionary mode and run the OCR again. (The operation of font dictionaries will be discussed shortly.)

If necessary, enter a character (or character string) for the incorrect or unknown shape and click one of the following buttons.

Learn

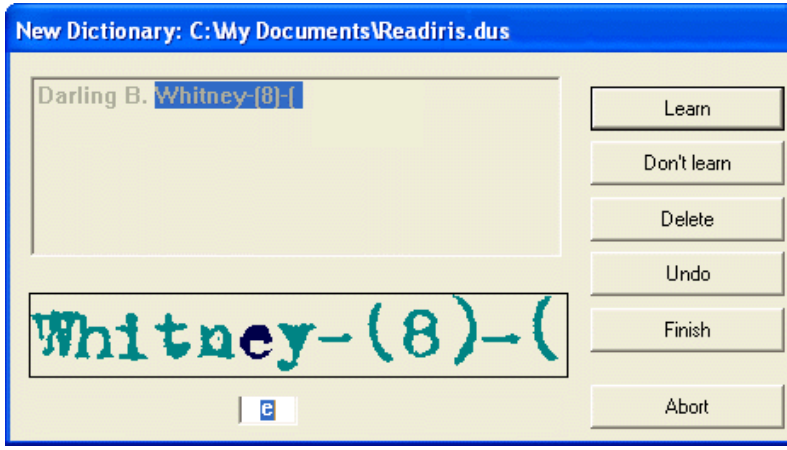
You agree with the proposed solution or correct it. The program saves this doubtful character in the font dictionary as "sure", final. Future recognition will no longer require your intervention, the shape is considered learnt once and for all.

In the example above, the system stops on a soiled character, and we click "Learn" to accept a shape which cannot be confused with other characters.

Don’t Learn

You agree with the proposed solution or correct it. The difference with the "Learn" button is that the learnt symbol gets the status “unsure” in the dictionary. For future recognition, the system will propose the “learnt” solution but still require a confirmation.

This button is used for symbols which might be confused with others: a defaced "e" which might be mistaken for a "c", a damaged "t" which closely resembles an "r" etc.



The "e" above is seriously damaged - in fact it is close to the "e" symbol -, and you should click "Don’t Learn" so as not to confuse the two symbols.

Delete

The displayed form is eliminated from the output. This button is used to ignore “noise” on the documents - spots, coffee stains etc. - which might get recognized

as points, commas and what have you -, and to erase any other unwanted symbol.

Undo

You go back to correct mistakes. You can undo the 32 last decisions.

Finish

The learning process is aborted but the OCR continues in automatic mode. All decisions by the system thereafter are accepted without user validation.

Click this button when you see that the recognition is highly accurate and does not require detailled proofreading.

Abort

Don't confuse "Finish" with the "Abort" button: with "Abort", no output is generated and you start all over, with "Finish", the text is created, it just isn't proofread in detail!

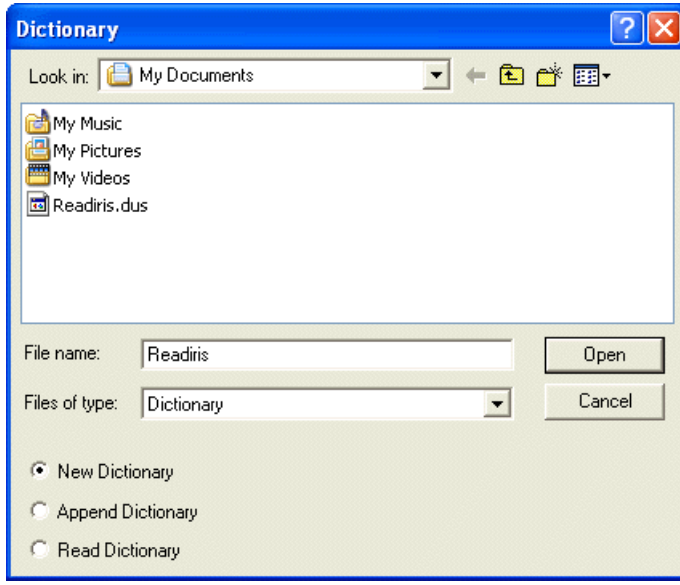
THE ROLE OF FONT DICTIONARIES

The results of each training session are temporarily held in the computer's memory but can and should be stored in files called "dictionaries" for future use.

(Don't confuse font dictionaries with (user) lexicons! Font dictionaries contain character shapes learnt during the interactive OCR phase, lexicons are linguistic databases that assist the recognition.)

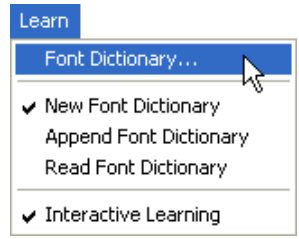
The font dictionaries should be loaded into memory when you want to recognize similar documents in order to make use of the extra intelligence they contain; in this way, Readiris takes into account the intelligence stored in these font libraries. You could say that Readiris gets more intelligence each time you use it!

How does this work? The operation of font dictionaries is controlled by the "Learn" menu: you have to select a dictionary with the command "Font Dictionary" and determine its mode of operation.



Font **dictionaries** are limited to 500 shapes, and you are recommended to create separate dictionaries for specific applications, for instance per type of document. Dictionaries have the default extension *.DUS. Training no longer has effect when the dictionary is full: the results of the learning are no longer held in memory or written to a dictionary.

You can set the dictionary mode inside the command "Font Dictionary" or directly under the "Learn" menu. Three dictionary modes are available: new, append and read.



By selecting "New Font Dictionary", you indicate that the training results will be saved in a *new* dictionary. (If you select an existing dictionary, its contents will be erased.)

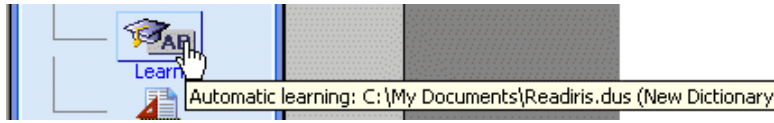
The append mode indicates that the training results will be saved in an *existing* dictionary: the recognition makes use of the extra intelligence already contained in the dictionary, and you add new font shapes to it. In simple terms, this option allows you to build up a font dictionary in several steps.

(When you enter a filename for a new dictionary and activate the "append" mode, an empty font dictionary is created and you complete it.)

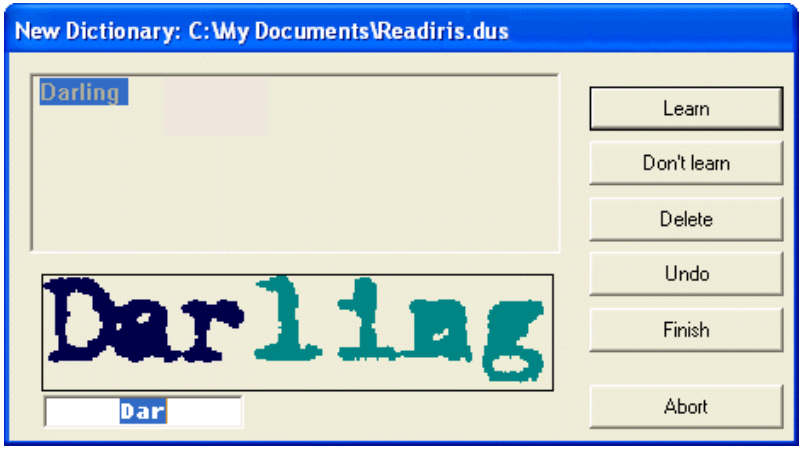
With the last option, "Read Font Dictionary", the dictionary functions in read-only mode: you make use of the dictionary *without* adding new font shapes to it.

Select the new mode when a single page is recognized. To recognize many pages of the same type - pages with the same fonts and printing quality - select the new mode for the first page, the append mode for a few pages more and the read mode for the rest of the document(s).

Know that the tooltip of the "Learn" button indicates at all times which font dictionary is currently active and in which mode that dictionary operates.



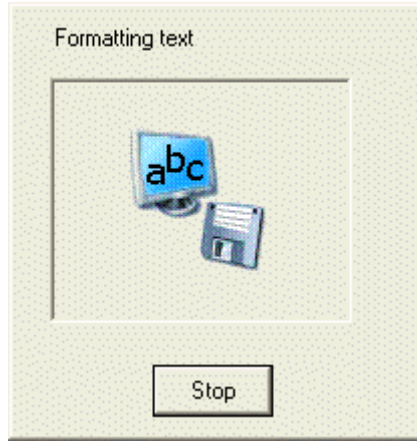
When you enter the interactive learning, the dictionary and its operating mode are indicated in the window title; you should click the "Abort" button and start over in case they are wrong.



Sending the Result Directly to Your Application

The interactive training concludes the character recognition. As Microsoft Word operates as output target by default, your wordprocessor is started up automatically at the end of the recognition (if necessary) and the recognized text is inserted.

You may get a progress bar on-screen as the recognized document gets formatted. (Whether this progress bar appears on-screen or not depends on the size of the document and the complexity of the formatting to be performed.)

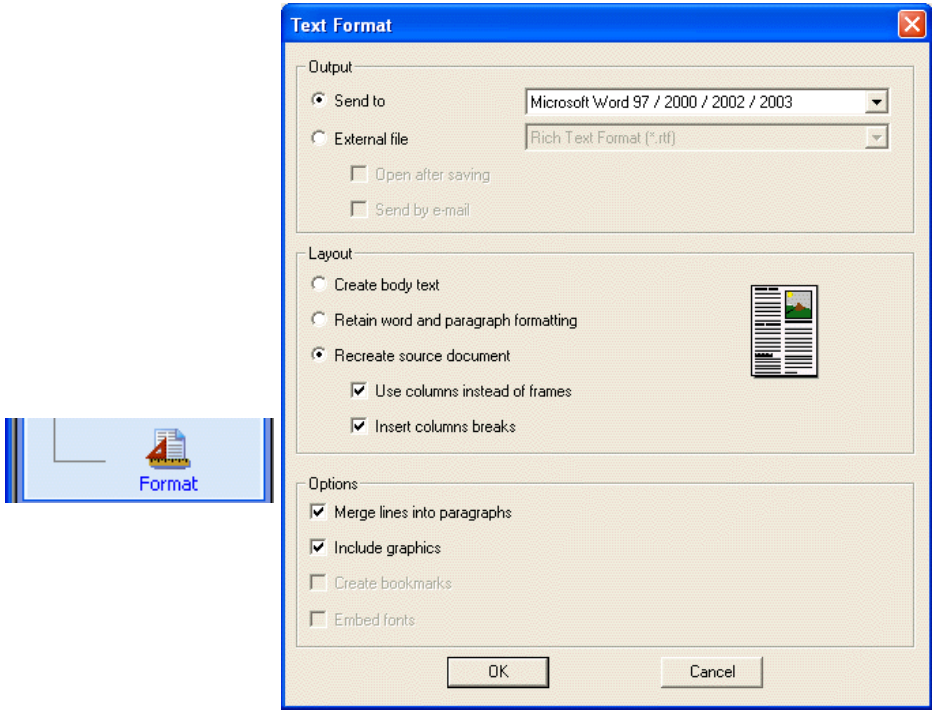


The scanned image is displayed again with the zoning as created to be available for further processing, it stays there until you scan another page.

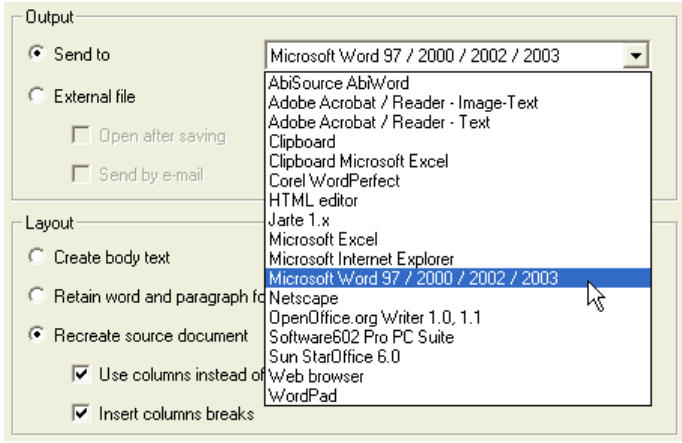
You have indeed converted a paper document into an editable computer file, be it 40 times faster than manual retyping! Go ahead and compare it with the image you have inside your Readiris window.

Actually, Readiris offers three different methods when it comes to saving the OCR result: sending the recognized document directly to a target application, saving the result in an external file and copying the result to the Windows clipboard.

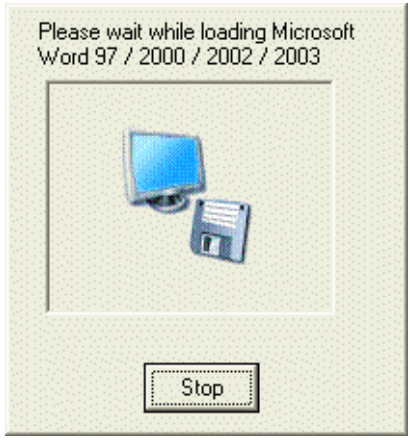
The **output target** is selected using the "Format" button on the main toolbar (or the command "Text Format" under the "Settings" menu).



The "Send to" feature offers a direct OCR link between your scanner and your Windows applications: you **send** the scanned documents directly to your wordprocessor, spreadsheet or web browser, to Adobe Reader etc.!



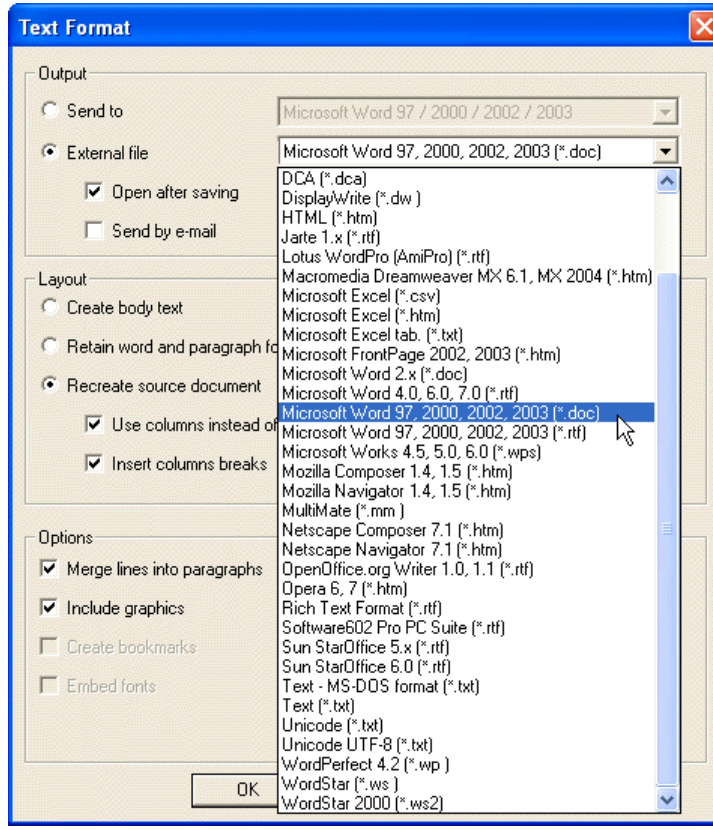
At the end of the recognition, the target application is started up and the recognized document is opened inside a new text file or worksheet.



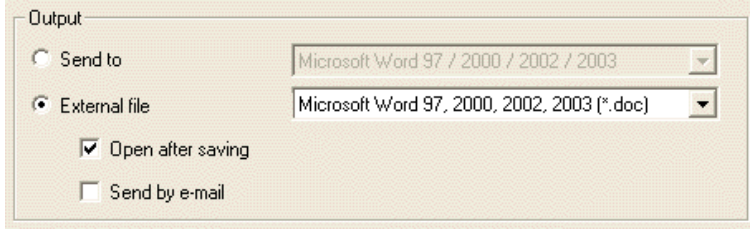
Don't forget that the option "Send to" also allows you to copy the recognized text to the Windows **clipboard**, so there is no strict need to export the result... or save it to an external file!

Saving the Results in a Text File

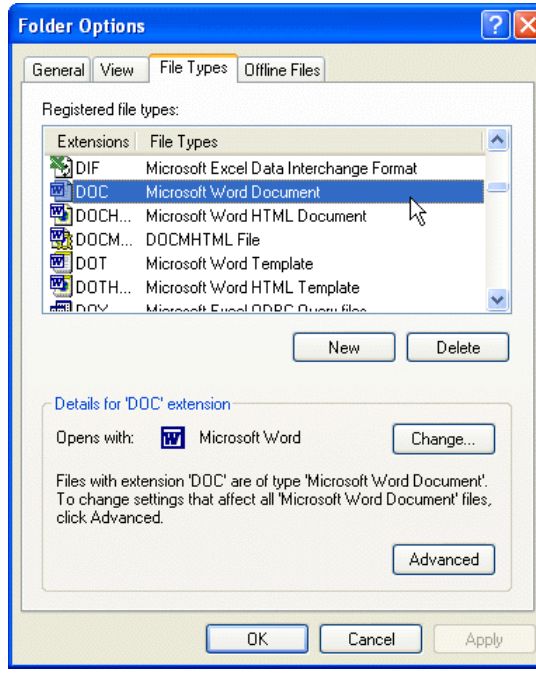
You can indeed write the OCR result to an "external" file. Here again, Readiris supports a wide range of file formats incorporating all popular wordprocessors, spreadsheets, web applications etc.



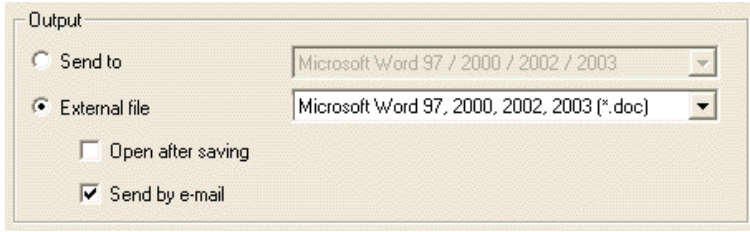
The option "Open after Saving" is largely similar to the “send” feature: you open the recognized document once it’s saved.

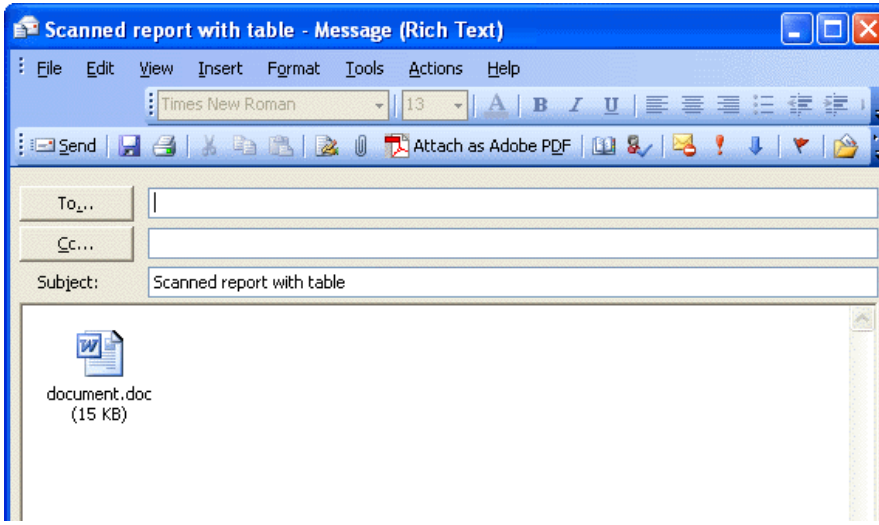


However, the method used to address the target application is different. This time, the **Windows file types** determine which application will be started up. It's as if you double-clicked the output file in the Windows Explorer... (With the option "Send to", Readiris addresses specific target applications directly.)



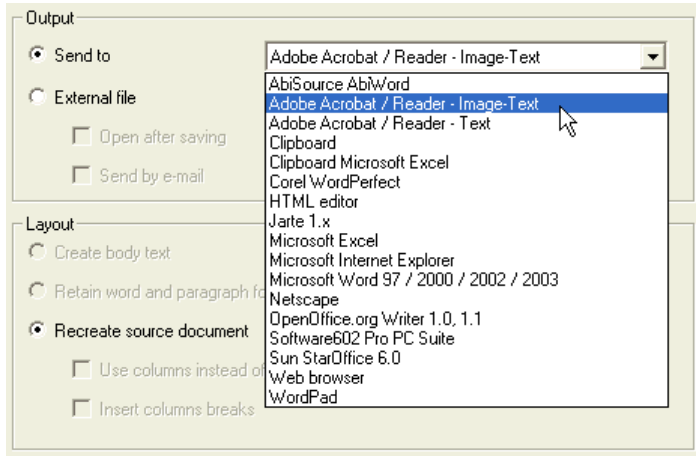
The option "Send by E-mail" creates a new **mail** message and inserts the recognized document as mail attachment. Do you know a faster way of distributing a paper document quickly...?

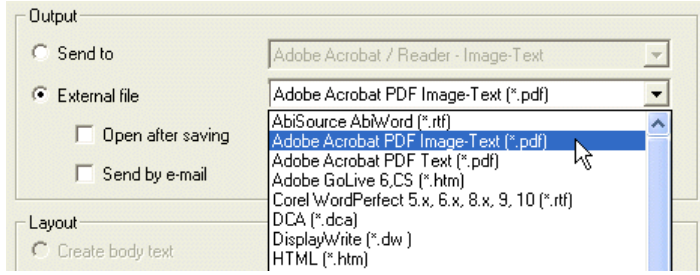




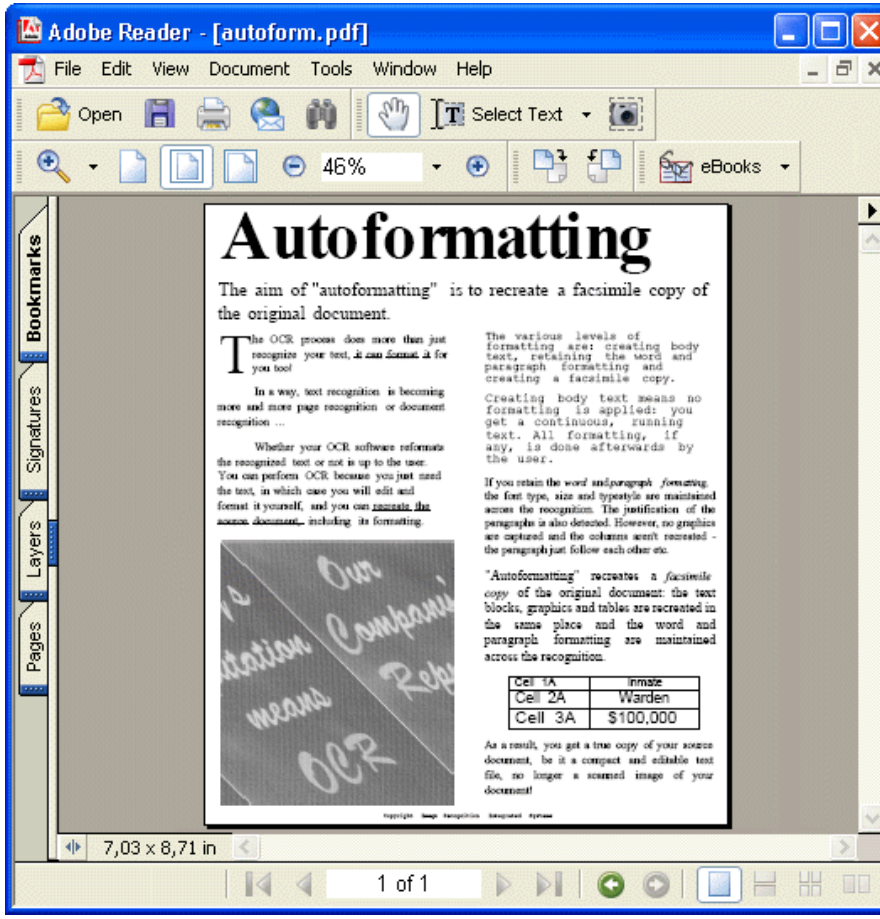
CREATING PORTABLE DOCUMENTS...

We'll go deeper into one format: **Adobe Acrobat PDF**. Readiris allows you to create PDF documents of two types - PDF Text and PDF Image-Text.





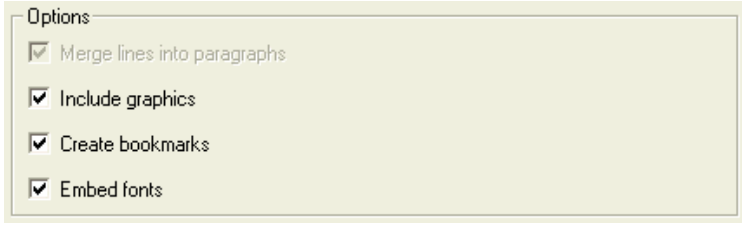
What's the difference between the two? When you select the format "PDF Text", Readiris creates a PDF file that contains the text result. (Graphics may occur but only when graphic zones occur on the page - photographs, artwork etc.) In other words: the page image is *not* contained in the single-layered PDF file!



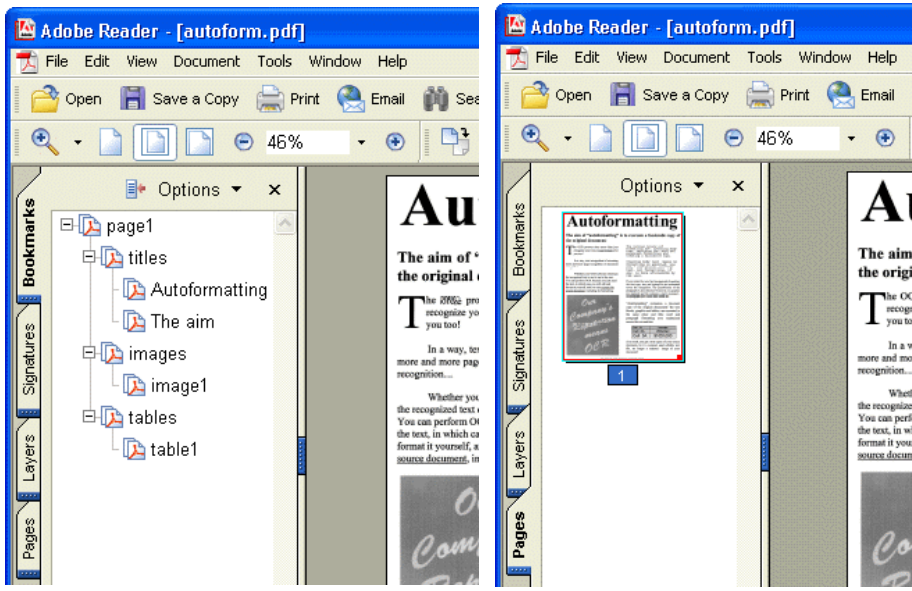
The format "PDF Image-Text" yields different results: Readiris creates a searchable PDF file that contains the recognized text *and* the page image. The

page image is contained above the text in a two-layered PDF file. Use the "Search" tool of Adobe Reader and this becomes quickly obvious!

Click the "Format" button to discover two options that concern the Acrobat PDF format: "Create Bookmarks" and "Embed Fonts".



The option "Create Bookmarks" sees to it that a **bookmark** is created for each document element - the graphics as well as the text blocks and tables. For the text zones, Readiris applies an intelligent algorithm to come up with a title, a "summary" per zone; the tables and graphics are simply numbered. (Another navigational element of PDF documents, page **thumbnails**, can be created dynamically by your Adobe Reader software!)



The option "Embed Fonts" embeds the fonts in the PDF files. Embedding fonts prevents font substitution when readers view and print the recognized document. It ensures that readers - whatever their computer configuration may be - see the text in its original fonts. However, embedding fonts increases the file size of the recognized documents (somewhat)!

... OR READING THEM

Let's look the other way for a moment. As Readiris offers full support of the Adobe Acrobat PDF format, you won't just generate PDF files, you can also *read* them!

“Repurposing” PDF documents may be a major application of Readiris. There are several reason why this is the case. First of all, it’s a way of converting images into text: open image-based PDF documents, execute the recognition and save the OCR result to a text document (in any supported text format). Text files are editable, image files are not.

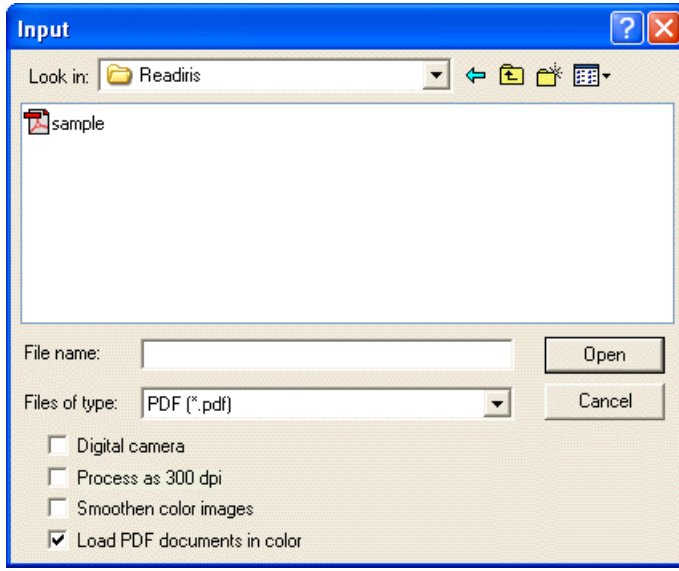
Second case: you can convert image-based PDF files to text-based PDF documents. You then execute the recognition on “image-only” PDF files and save the OCR results... as text-based PDF documents! Text-based PDF files are searchable and editable, “image-only” PDF files are not.

Finally, converting PDF files is a way of “unlocking” PDF content. You can recognize “read-only” PDF documents, where the text is normally inaccessible. With unprotected PDF files, the content can be retrieved (copied and saved to a text file), with “read-only” files, the content cannot be extracted. These documents can only be viewed and printed!

An important nuance: Readiris does not open password-protected PDF documents, even if all other PDF security barriers are broken down by Readiris!

Proceed as usual: load PDF files into memory as you open prescanned images - faxes, snapshots made with your digital camera etc. Click the "Stop" button or press Escape to interrupt the loading process between two pages.

There’s a specific option that concerns PDF files. You can open them as color and as black-and-white documents. This option is offered because rasterizing color documents is much slower!



RECOGNIZING MULTIPLE PAGES

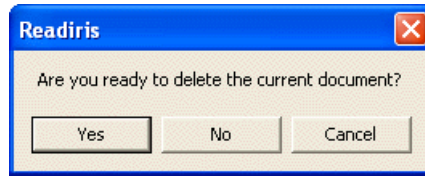
After the OCR, the scanned image is redisplayed with the zoning as created to be available for further processing.

You can now open the recognized text with your wordprocessor or text editor, import it into your desktop publishing software or any other text-based application. Go ahead and compare it with the image you have inside your Readiris window.

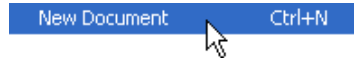
But how do you save the text of additional pages? Or in other words: how do you process documents consisting of multiple pages? It's actually very simple: go on recognizing pages and save the results to the same file! (Make sure that file isn't currently open, because that will prevent you from writing to it!) Secondly,

don't forget to put the font dictionary in the append mode so that you can continue the font training comfortably.

As soon as you scan pages (or open image files) inside a document, you have to decide whether you want to start a new document or complete the current document.

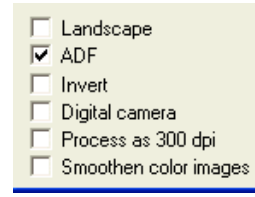


Answer "no" to add pages to the current document, answer "yes" to create a new document. This answer has the same effect as the command "New Document" under the "File" menu.



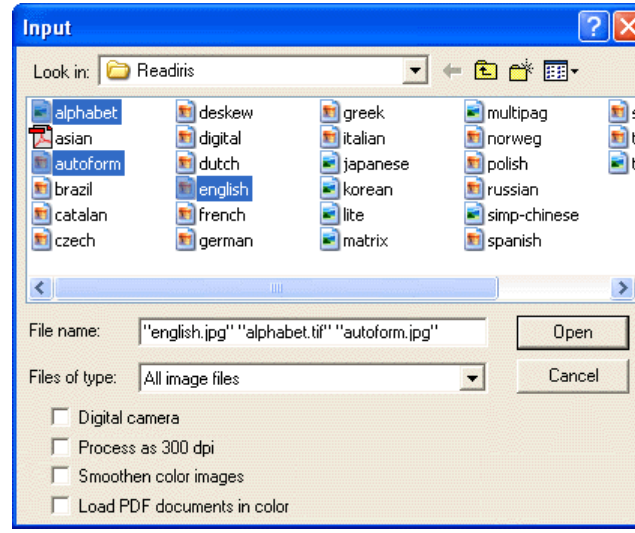
But there's a more efficient way of recognizing several pages than scanning and OCRing them one after the other: processing **multipage documents** directly!

To scan a document composed of several pages in one operation, enable the document feeder of your scanner with the option "ADF" under the "Scanner" button.

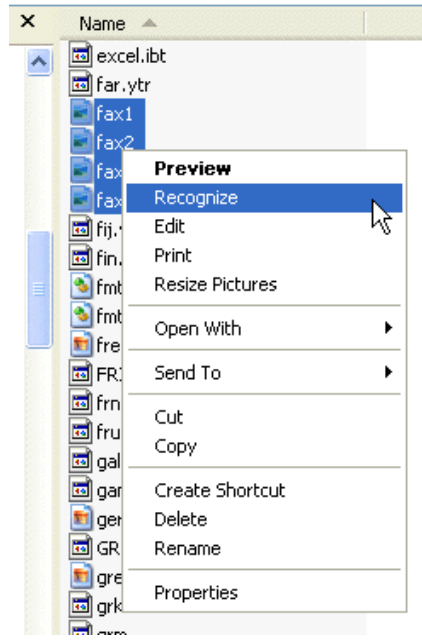


Place the pages of your document in the automatic document feeder and start the scanning: all pages are scanned until the document feeder is empty.

You can also *open* multiple prescanned images. To load several images, select the first image and hold down the Ctrl key as you select additional images. To load a continuous range of images, select the first image and hold down the Shift key as you select the last image.



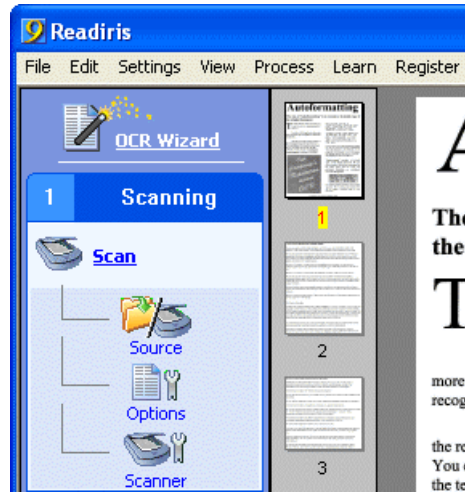
The same effect can be obtained comfortably from within the Windows Explorer: select several image files, right-click and select the command "Recognize" from the "Context" menu. You can repeat this operation: all images you send to Readiris append the current document until you click the command "New Document".



You can even *drag* several prescanned images from the Windows Explorer onto the Readiris window! The same argument holds: all images you drag onto the Readiris window are added to the current document until you click the command "New Document".

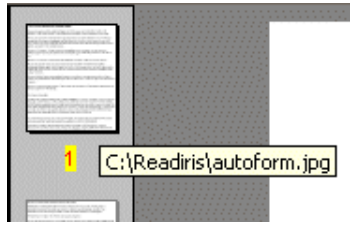
Readiris sorts the images automatically - image 001.tif precedes 002.tif precedes 003.tif etc.

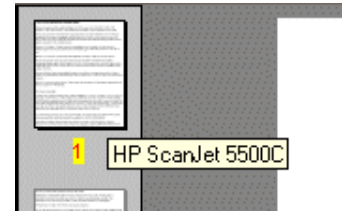
The **page toolbar** on the left side - it is displayed as soon as pages get processed - represents the various pages of the document and gives access to the page commands (using the right-click).



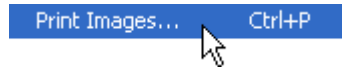
The current page is highlighted in the page toolbar and mentioned in the Readiris title bar.

The page toolbar comes with a tooltip: hold your mouse pointer over a page thumbnail to learn which image was loaded into the memory. (If a multipage image was opened, there's obviously just one file for all the images.) When you are *scanning* multipage documents, the tooltip simply mentions the scanner model.

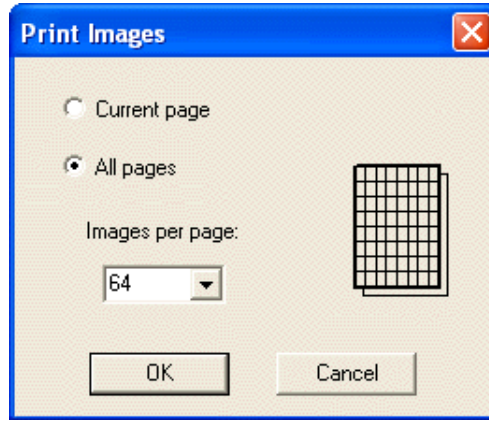




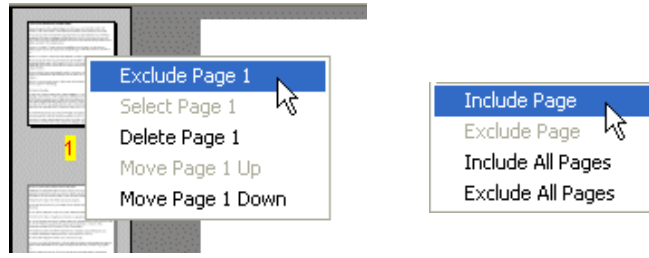
You can quickly **print** the scanned **images** with the command "Print Images" under the "File" should you need an overview of your document.



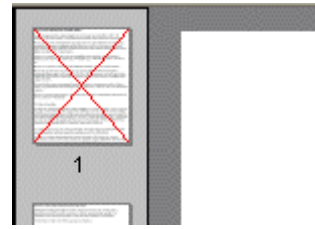
You can print the current page or all pages. Select the number of pages or thumbnails you want printed on a page.



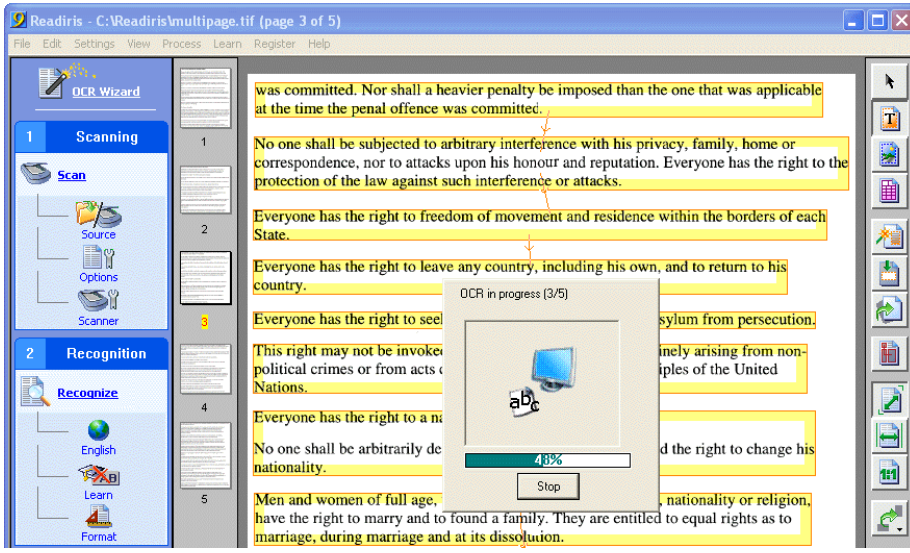
But you don't have to print all pages either: the page toolbar (and the "Edit" menu) allow you to exclude pages (temporarily). Right-click a page thumbnail and select the command "Exclude Page" on the "Context" menu or display a page and select the command "Exclude Page" from the "Edit" menu to exclude it from the printing (and recognition) process. Select the command "Include Page" to include it again. For greater flexibility, the "Edit" menu offers equivalent commands that apply to *all* pages.



The thumbnails of excluded pages are stricken out. Mind you, printing the current page always works, even if it is "disabled" for the time being!



Load the sample image MULTIPAGE.TIF and start the recognition. The various pages are displayed one after the other; the Readiris title bar indicates the page number.

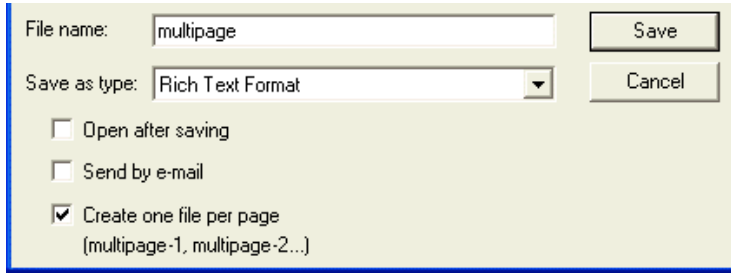


If the interactive learning is enabled, you go through the recognition and learning phases page by page. The dictionary mode "New" is used for the first page and the mode "Append" for the successive pages.

When you click the "Finish" button, all decisions by the system thereafter are accepted without user validation. In other words, the interactive learning is aborted for *all* pages; the OCR for this document continues in automatic mode.

The recognition result of multipage documents is saved in a single output file. (When the recognition result is sent to a target application, multiple pages get created inside a single document.)

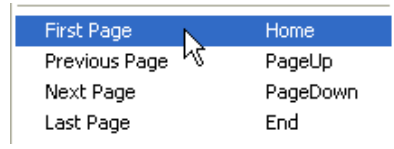
At least, that's the case when the option "Create One Document per Page" is disabled when you save the recognized document. This option sees to it that each page of a multipage document is saved in a separate file. If the user gives the file name text.doc, the files will be called text-1.doc, text-2.doc etc.



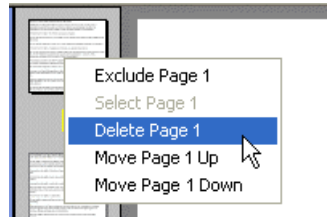
EDITING MULTIPAGE DOCUMENTS

The user can edit multipage documents, mainly to correct scanning errors: he can delete pages from the document and move pages to other locations in the document.

The navigation first. To *go to a page*, click on its icon in the page toolbar or hold your cursor over its thumbnail, invoke the "Context" menu by right-clicking and use the command "Select Page". To go to the previous page, you can use the shortcut PageUp, to go to the next page, press PageDn. Press Home to go to the first page, press End to go to the last page. Or use the corresponding commands under the "View" menu.

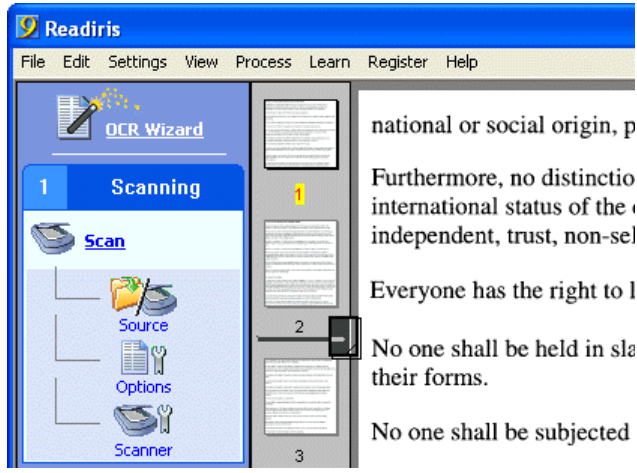


Let's edit the document now. To *delete a page* from the document, hold your cursor over its thumbnail, right-click it and use the command "Delete Page". And we remind that you can temporarily exclude pages, not delete them, from the recognition (and image printing) process: the page toolbar (and the "Edit" menu) offer the necessary commands.



To *move a page up* in the document, use the command "Move Page Up", and to *move a page down*, use the command "Move Page Down".

To *move a page* to a totally different location in the document, drag its icon to that new location.



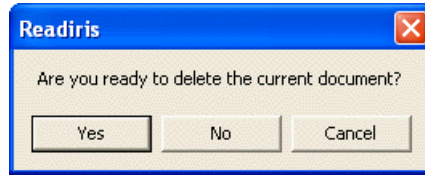
Starting a New Document

You can use the command "New Document" under the "File" menu to close the current document.

This command "cleans the slate". Any document loaded into memory - containing a single page or multiple pages - is erased. You are now ready to create a new document.

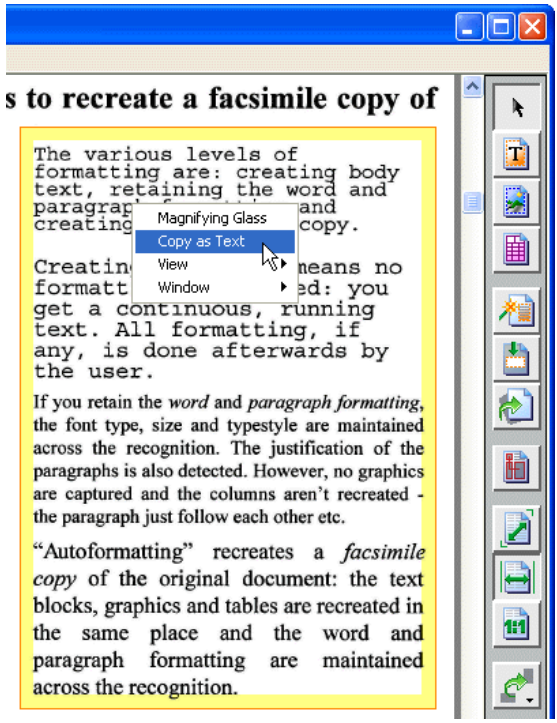
But you can also create a new document from within the current document. As long as the OCR was not executed, the system assumes that you want to add pages to the current document. You can for instance scan all the pages in the scanner's autofeeder, fill the feeder again and start over. All pages scanned will compose a single document. Or you could scan a number of pages and add some image files, say, faxes. These pages again form a single document, all you have to do is change the image source in between with the "Source" button.

When the OCR *was* already executed and you re-initiate the scanning (or the loading of images), you are prompted to start a new document or complete the current document.



RECOGNIZING TEXT ZONES

We now know how to recognize pages and how to process multipage documents. But can we recognize less than a page with equal comfort? We can! Right-click your mouse and select the command "Copy as Text" from the "Context" menu: the text window under the mouse gets recognized and sent to the clipboard.



The current system settings - language, font type etc. - apply. The OCR result is placed on the clipboard as "running", unformatted text.

Organizing the Text Output

Saving or exporting the text means more than selecting an output method or defining a filename for the output file. You also select a file format and determine

the appearance of the recognized text. In short, you have to decide where you want to take the text before you launch the execution.

Some options of the "Format" button allow you to influence the look of the text output.

The **text flow** of the output document is directly influenced by the option "Merge Lines into Paragraphs".

Keep this option enabled to have Readiris detect the paragraphs: Readiris will then apply the normal **wordwrap** typical of wordprocessors, otherwise, a carriage return is added after each line and hyphenated words remain so! Paragraph detection is enabled by default.

Let's give an example to clear things up. When the first three lines of a column are "The new presi-", "dent waved from the balcony." and "His wife had joined him.", the paragraph detection gives you the following result: "The new **president** waved from the balcony**. His** wife had joined him." The hyphenated parts of the word "president" were "reglued" and a space was added at the end of the first sentence, thus creating naturally flowing text.

Had paragraph detection *not* been enabled, the original layout would have been retained, with a carriage return added at the end of each line.

This option is *not* available when the PDF format is selected: Adobe Acrobat PDF files always store text line by line!

(The "Format" button contains some formatting options we haven't discussed yet - this will be done shortly.)

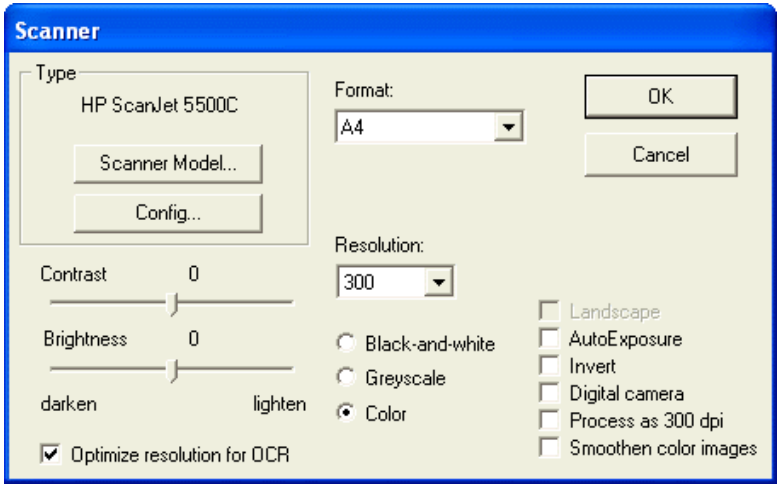
SETTING UP YOUR SCANNER

Let's set our scanner up now. It is assumed that the scanner hardware and necessary drivers are installed correctly.

If your Readiris software licence was bundled with a scanner or digital camera, this step probably is unnecessary as your hardware may already be set up under Readiris.

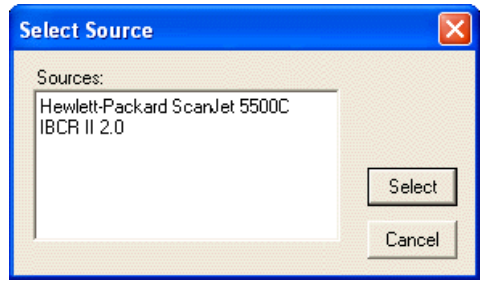
Click the "Scanner" button on the main toolbar.

Click the button "Scanner Model" to determine your **scanner model**.



When you select the option "<Image>" as “scanner”, prescanned images function as image source at all times - you won’t have even to select the disk as image source with the "Source" button on the main toolbar.

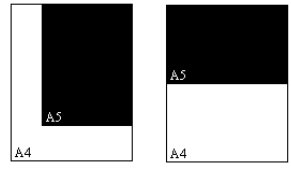
The "Configure" button is only available when you scanner allows it. It gives access to some advanced scanning parameters; with Twain scanners, clicking the "Configure" button allows you to select the Twain source. (You can also use the command "Select Source" under the "File" menu.)



Once the scanner is selected, the same window may allow you to set the scanning resolution, the page format and orientation, brightness and contrast and may allow you to indicate whether you are going to use the scanner's document feeder. With Twain compliant scanners, all scanning parameters are often set inside the Twain interface.

Set the **brightness**, and, if available, the **contrast**.

By enabling the option "Landscape", you indicate that the selected page orientation is wide ("landscape") instead of tall ("portrait"). The page orientation actually applies to reduced page formats: on an A4 flatbed scanner, you can scan, say, A5 pages (half that big) in portrait or landscape format, but you can obviously only scan the full A4 surface in one direction!



The option "Invert" allows you to generate **"inverted" images** in the black-and-white scanning mode - you can activate this option to process full pages with white text on a black background.

Bring Color to Your Text Scans!

Readiris supports black-and-white, greyscale and color images on an equal basis, so you are free to choose the **color mode** that best suits your needs. To include lineart graphics in the recognized documents, scan in black-and-white, to include black-and-white photos, scan in greyscales, to include color pictures, scan in color.

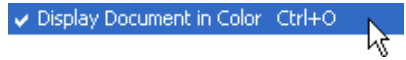
But why would you reduce the bit depth of the images during the scan? It goes without saying that greyscale and color images are slower to acquire and require more RAM memory than "bilevel" images.

Scanning in greyscale and color isn't just useful to save the graphics with sufficient quality, in some instances, it's also useful or necessary to obtain good OCR results! When text is printed on a color background, scanning in color may create the tone differences that are lacking in black-and-white images. When there is only limited contrast between the text and the background, the background can create "noise" that renders the recognition difficult or impossible!

Think for instance of black text printed on a dark background: when scanning such a document in black-and-white, you may not be able to "drop" the background color without losing the text information as well, as much as you may try to adjust the scanner brightness...

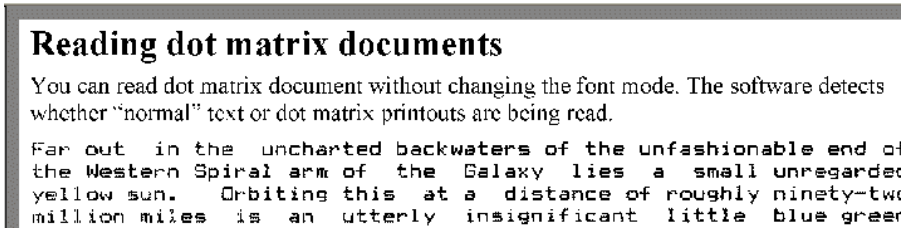
Readiris creates a black-and-white version for every greyscale and color image. Thanks to its intelligent routines, even tough cases get solved - here's how a "difficult" image gets binarized!

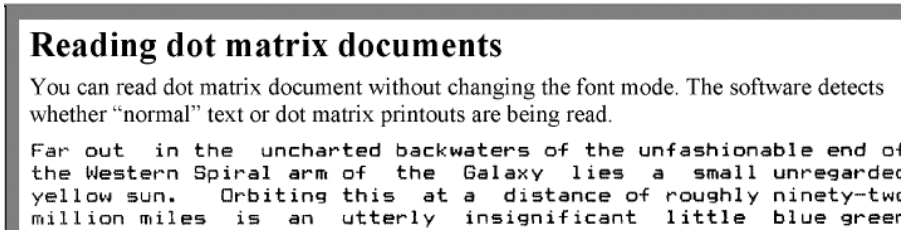
To view a scanned image in black-and-white, disable the option "Display Document in Color" under the "View" menu.



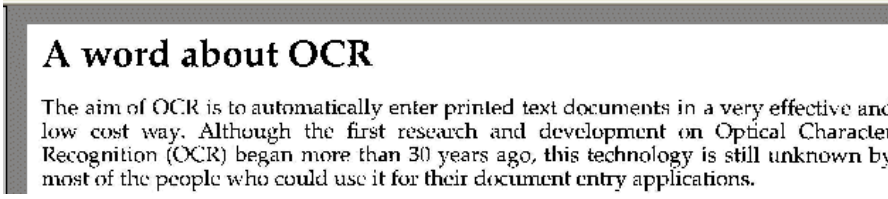
When this option is enabled, you won't see any black-and-white images on your computer screen - even when you're actually scanning bilevel images! That's because the option "High-Quality Display" under the "View" menu optimizes the images for an optimal on-screen legibility.

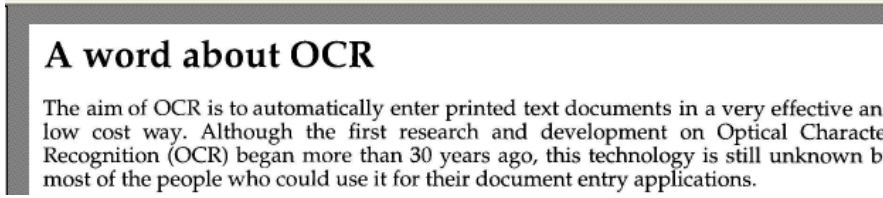
This specialized high-resolution display technique converts black-and-white images into greyscale images.





Greyscale and color images are softened, smoothened.





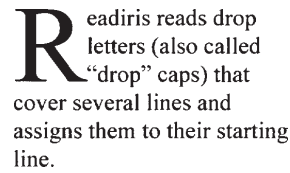
As a result, there's no need to zoom in, even on laptops with an LCD screen or desktop computers with a low-end 14" screen. High-quality display is enabled by default, but may be superfluous on high-resolution computer screens.

DIFFERENT DEVICES, DIFFERENT RESOLUTION

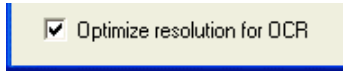
Whatever your scanning mode may be, use a scanning **resolution** of 300 dpi for normal applications. Use a higher resolution of 400 dpi for small print (below 10 point) and when the document is very degraded.

Readiris reads **point sizes** of 6 to 72 point (0.08" to 1 or 0.21 to 2.54 cm).

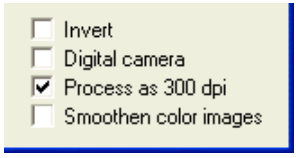
Readiris also recognizes "**drop letters**", large caps that cover several lines. (These can of course be no bigger than 72 point!)

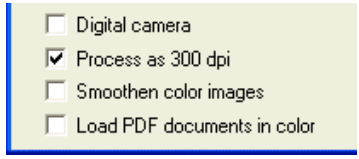


As optimal OCR requires a resolution between 300 and 400 dpi, Readiris warns you when you're submitting images with a resolution lower than 200 dpi or higher than 800 dpi. However, Readiris can correct scans with too much detail for you! Enable the option "Optimize Resolution for OCR" in the scan settings to do so. Whenever the image resolution of your scans exceeds 600 dpi, the resolution is reduced for the OCR process.

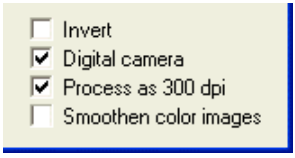


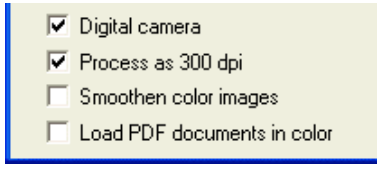
There are other ways of avoiding this warning: you may be reading **faxes** - which have a resolution of 100 or 200 dpi -, when you're creating images with a digital camera - where the resolution is unknown - and when you're opening images where the file header contains an incorrect resolution. To process such images hassle-free, enable the option "Process as 300 dpi". This setting applies to both direct scanning and the opening of prescanned images.





When your images are acquired by a **digital camera** instead of a scanner, it is mandatory that you enable a special option (that also applies to scans and prescanned images).

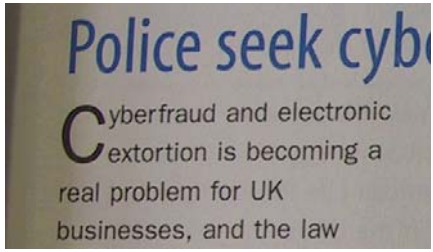


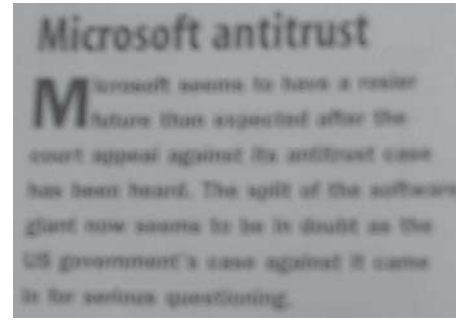


By doing this, you enhance the image before it gets recognized. There are specific challenges to be met when it comes to digital cameras: they produce low-resolution images - even when you hold the camera very close over your document - and the image resolution is in any case unkown.

There are some "finer points" to be aware of when it comes to successfully recognizing images captured with a digital camera.

First of all, select the highest possible image resolution. Create for instance 2,048 x 1,536 size images when 1,024 x 768 and 640 x 480 images are also supported. Secondly, enable the "macro" mode of your camera to take closeups - which is always the case when you photograph documents. (This mode was designed to capture flowers, insects etc.) Otherwise, the images are unsharp and illegible.

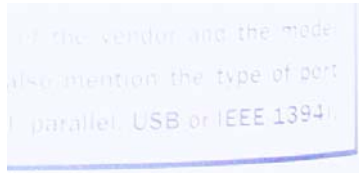




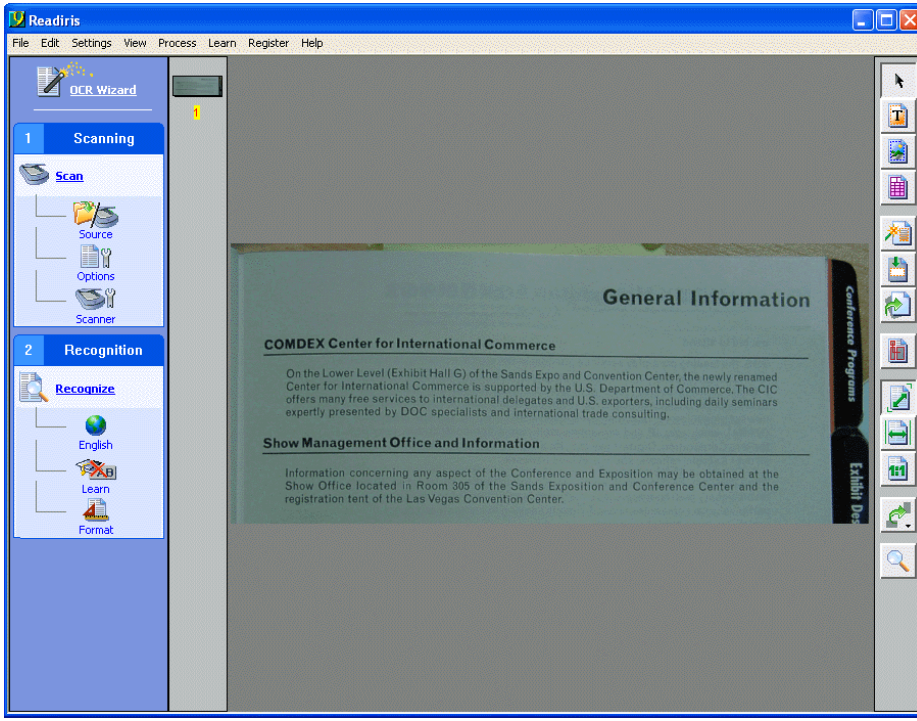
Limit yourself to no or small compression: important compression reduces the sharpness of the captured text. Zoom manually to crop your document - some cameras are bundled with photo stitching software, but don't bother using it for document capture.

Hold the camera directly above the document to avoid capturing the document at an angle. However, avoid shadows cast on the document by the camera or your hand! Produce stable images. Consider mounting your camera on a tripod when necessary.

Disable the flash when you're filming glossy paper, otherwise the image may be too light. Generally speaking, adapt the brightness and contrast to the environment - day light, lamp light, neon light etc. (Some cameras can be calibrated by filming a white document.)

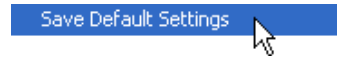


To give it a try, open the image DIGITAL.JPG in the Readiris folder and execute the recognition.



SAVING DEFAULT SETTINGS

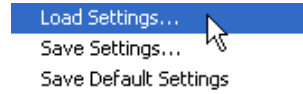
Set all scanning parameters correctly and click the command "Save Default Settings" under the "File" menu to save the current settings as default settings for future use.



Settings files contain more than the scanner **settings**: they also determine whether you are going to use interactive learning, which language the documents have, which output mode is used - for instance send text to WordPad - etc. In short, *all* operational settings of Readiris are stored in the settings files.

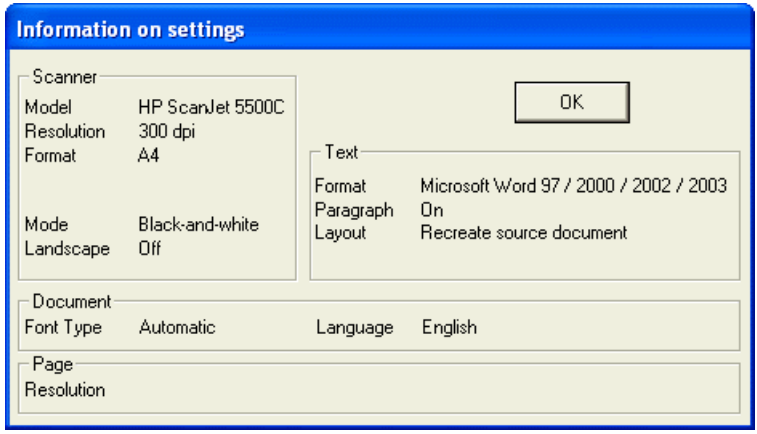
SAVING SPECIFIC SETTINGS

The default settings will obviously be used at each program startup, but you can save specific settings as well to avoid having to redefine the operational parameters. The commands "Save Settings" and "Load Settings" under the "File" menu take care of this.



Let's give an example: if you regularly have to OCR English documents with a specific layout, you are recommended to create a settings file for this type of document. You would then select "English" as the document language, load a specific zoning template to avoid having to reapply the same windowing each time, disable learning but activate a font dictionary in the "read" mode because the same typefaces are used systematically etc.

If you are unsure what the current settings are, you don't have to "plunge" into every menu and command to discover what they are. You can use the command "Info" from the "File" menu to get an overview.



SCANNING DOCUMENTS

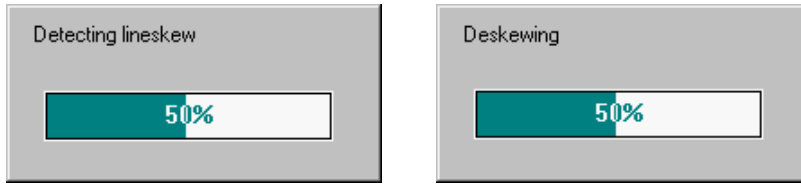
Now that our scanner is set up, we want to get started scanning documents. There are some elements you should be aware of.

First of all, pay some attention to lineskew. Although the page analysis and recognition are skew-tolerant, it may become difficult to window and OCR a page correctly when the skew is too significant. Limited lineskew (less than 0.5°) can be ignored because the OCR accuracy does not suffer.

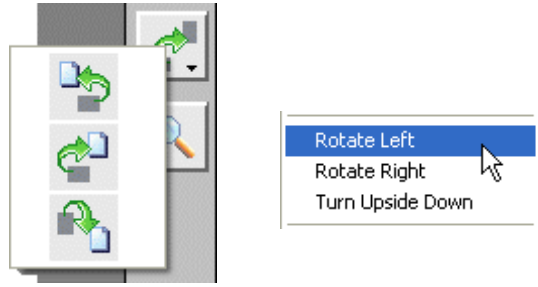
The option "Page Deskewing" under the "Options" button (and under the "Settings" menu) determines whether pages which were scanned at an angle will be **deskewed**, straightened automatically - limited lineskew gets ignored. This option is disabled by default.

If you forgot to enable this option, use the "Deskew Page" button on the image toolbar (or the command "Deskew Page" under the "Process" menu) to "straighten" pages which were scanned at an angle.

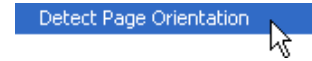
The deskewing takes a few seconds: the image is analyzed to detect the skew angle - if any -, the color or greyscale image *and* its black-and-white version are deskewed and the page analysis gets re-executed.



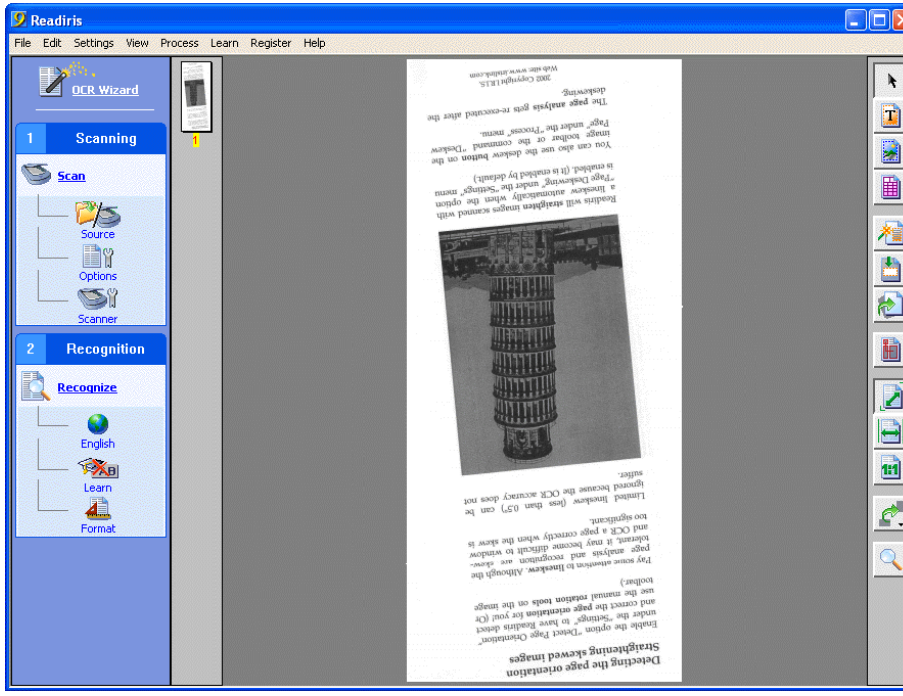
You may also need to adjust the page orientation. Use the **rotation** tools on the image toolbar. (Corresponding commands are found under the "View" menu.) Three rotation directions are available: to the left, to the right and upside down. Rotation also takes a few seconds as the image itself is updated, not just the display on-screen.



However, Readiris can correct badly oriented pages for you. Enable the option "Detect Page Orientation" under the "Options" button (or under the "Settings" menu) and Readiris will correct the page orientation where needed.

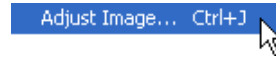


You can make good use of the image DESKEW.JPG in the Readiris folder if you want to try it. Enable the options "Page Deskewing" and "Detect Page Orientation" before you open the image and let Readiris restore the Tower of Pisa the way we like it.

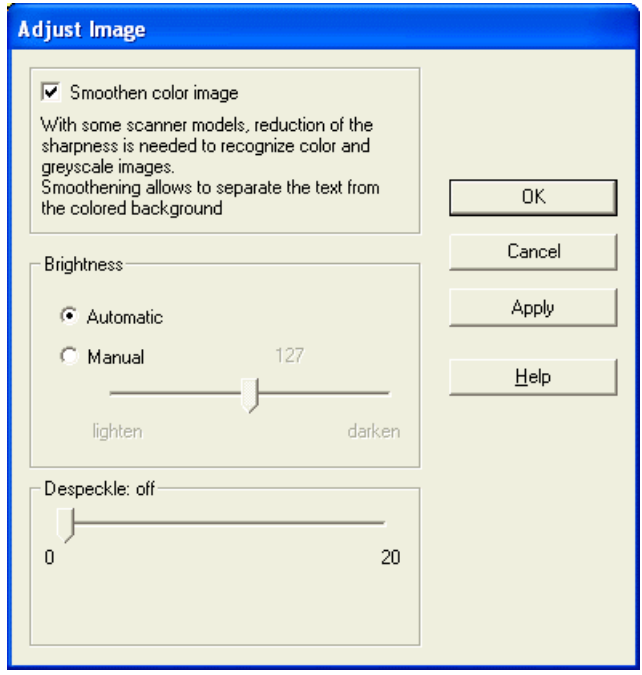


ADJUSTING THE SCANNED IMAGES

As was already indicated, powerful intelligent routines automatically convert color and greyscale images into black-and-white. Should this still be necessary, the user can optimize the image further for the consecutive OCR process. Select the command "Adjust Image" under the "Process" menu to do so.

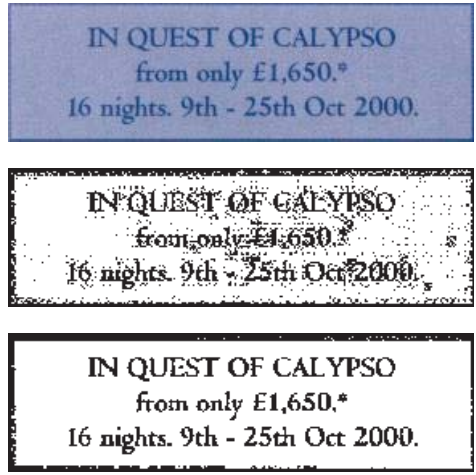


When you access this command, the black-and-white version is displayed automatically. (It's as if you disabled the option "Display Document in Color"!) There are some complicated concepts here, and we need to discuss them in detail.

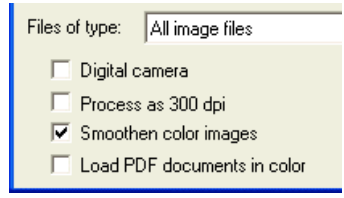


The option "Smoothen Color Image" renders greyscale and color images more homogeneous by “flattening”, smoothing out relative differences in intensity. As a result, a stronger contrast is created between the foreground - the text - and the background - a color, artwork etc.

This **preprocessing** feature may seem highly technical and difficult to understand, but it certainly has its role to play: with some scanner models, this reduction of the sharpness is needed to recognize color and greyscale images. Smoothening is sometimes the only way separate text from the colored background! Below is a sample image that is simply illegible without image smoothing.

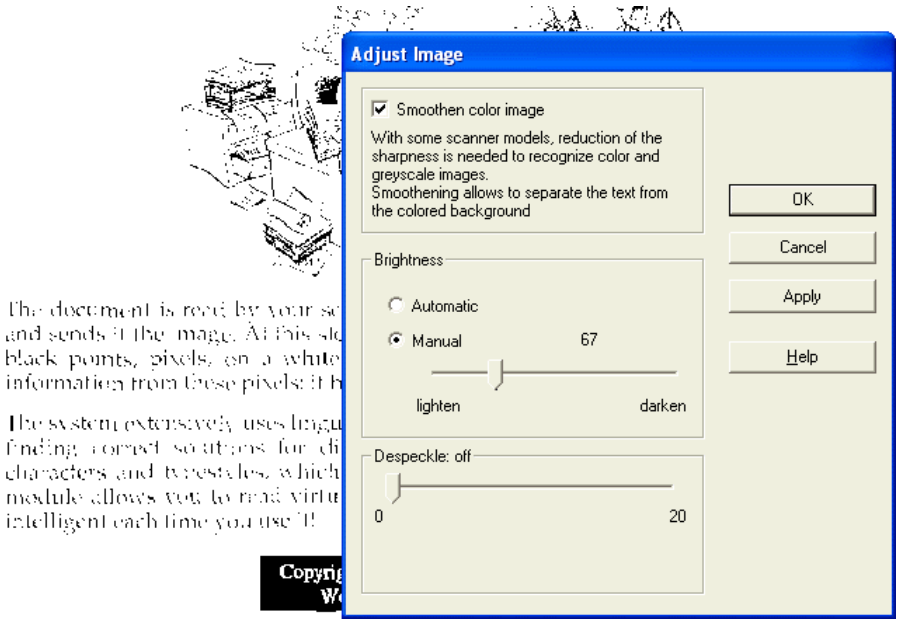


The image smoothening can also be enabled when you load prescanned images into memory!

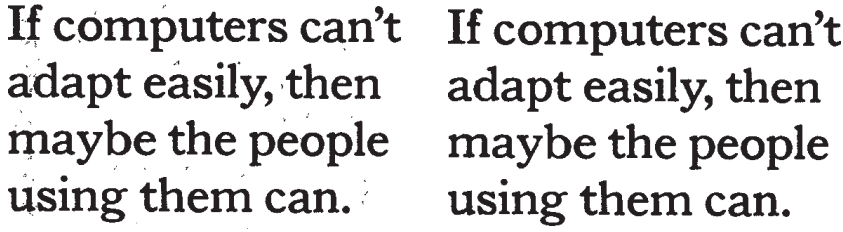


The brightness now. By "brightness", we actually mean the black-and-white threshold. The setting "Automatic" determines the bilevel threshold automatically. Apply a different threshold when necessary by darkening or lightening the black-and-white image: when you darken the image, more pixels become black in the black-and-white version, when you lighten the image, less pixels become black in the black-and-white version.

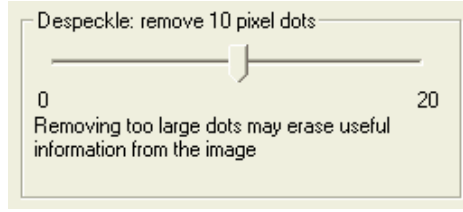
Note above all that no image adjustment is executed until you click the "Apply" button! By clicking "OK", you execute the adjustment *and* close the window. Here's an example where we lightened the black-and-white image dramatically - though admittedly not with OCR accuracy in mind!



The first two options concern color and greyscale images, the last one, "Despeckle", exclusively concerns black-and-white images. “Despeckling” means that the “parasite pixels” (also called “salt and pepper noise”) will be removed from black-and-white images.



Be sure that you don't erase spots that are too big, otherwise you might start erasing the dots on "i", portions of dot matrix letters etc.!

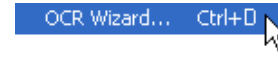


The best way of optimizing the images for the OCR process is this: place the adjustment window where it doesn't prevent you from judging the image adjustment you execute. Adapt the parameters - clicking "Apply" each time - until the image is crisp and clear.

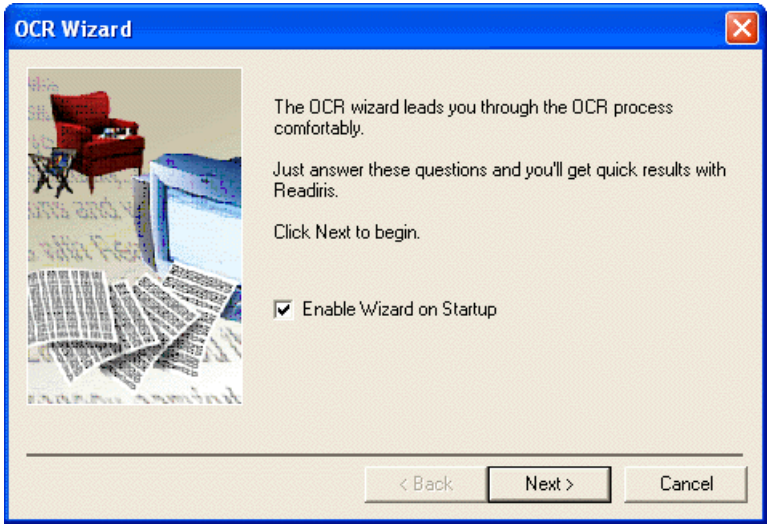
LETTING THE OCR WIZARD WORK FOR YOU

Let's get started capturing documents now. Instead of going through all the parameters, we'll use the **OCR wizard**, a very comfortable way of recognizing pages.

Click the "OCR Wizard" button on the main toolbar (or select the command "OCR Wizard" under the "Process" menu).



The wizard guides you through the OCR process comfortably: answer a few simple questions and you'll obtain quick and easy results with Readiris.



Actually, the OCR wizard starts running each time you start up Readiris; you can avoid this by disabling the option "Enable Wizard on Startup" in the first screen of the wizard (and with the equivalent option under the "Settings" menu).

Readiris Recreates Your Document Layout

The OCR wizard renders the recognition process highly automatic, but "automatic" OCR should *not* be confused with autoformatting! "Autoformatting" means that Readiris recreates a **facsimile copy** of the scanned document: the word, paragraph and page formatting of your original document are applied.

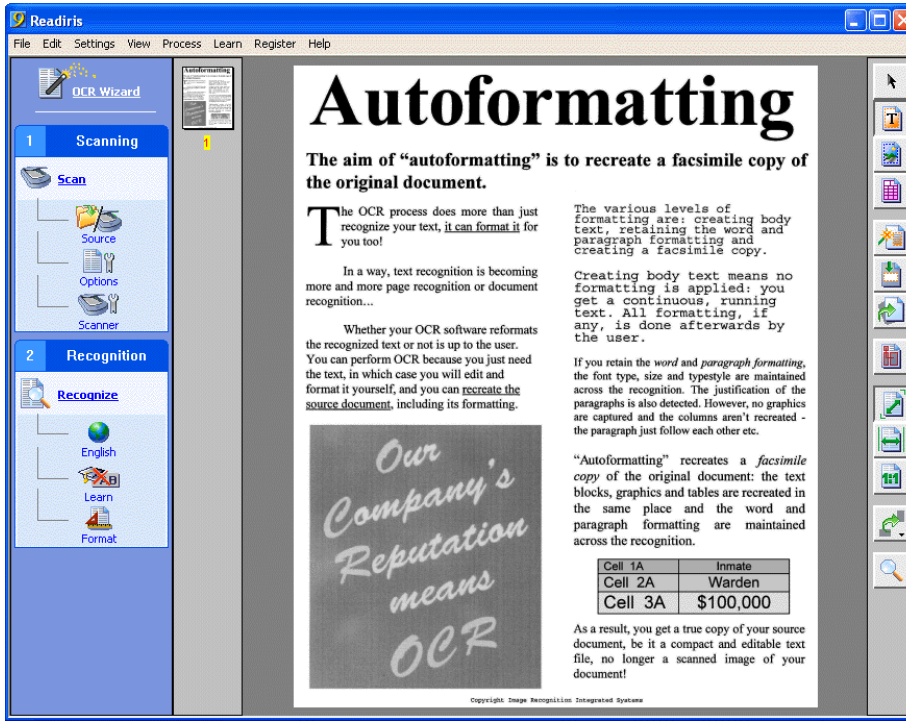
Similar typefaces (serif and sans serif, proportional and fixed, normal and condensed) are used as in the source document, the point sizes and typestyles (bold, italic, underlined, superscript and subscript) are maintained across the recognition. The tabs and the alignment (left, centered, right and justified) of each

text block are recreated. So are the bulleted and numbered lists. Any e-mail adresses and URLs of web pages get detected and recreated as hyperlinks in the output. The placement of columns, text blocks and graphics follows your original document.

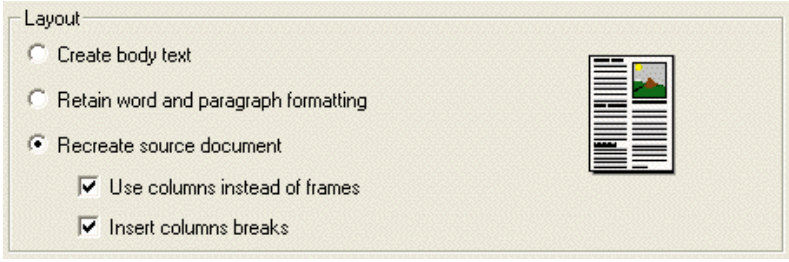
In other words, Readiris allows you to archive a true copy of your documents, be it a editable and compact text file instead of a scanned image!

All this implies that the sorting of windows only *partially* applies when "autoformatting" is used: you can include and exclude zones, but any re-ordering of zones is simply ignored!

Here's an example of how it works. To get acquainted with this feature, open the image AUTOFORM.JPG which is found in your Readiris folder.

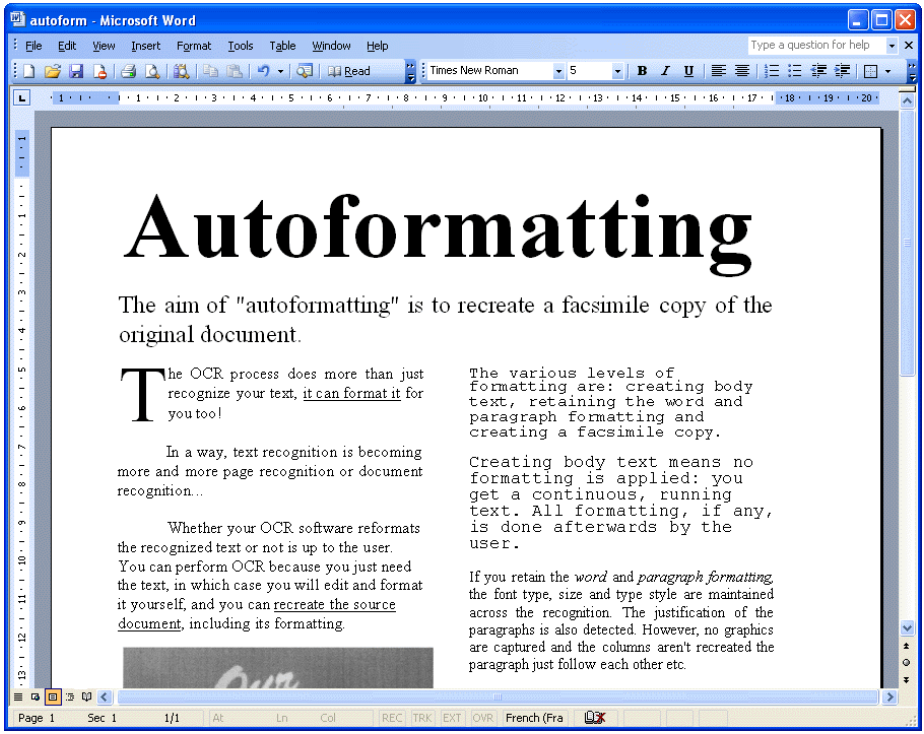


Click the "Format" button on the main toolbar and choose to send the OCR result to Microsoft Word or select the RTF (Rich Text Format) or Word (DOC) format. Secondly, select "Recreate Source Document" as layout option. (The option "Merge Lines into Paragraphs" is enabled by default to apply wordwrap within the paragraphs.)

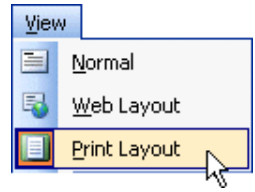


Whether layout reconstruction is available depends on the selected output mode. Some "poor" formats generating "plain" text such as Text (ANSI), MS-DOS Text (ASCII) etc. do *not* support advanced formatting codes and therefore cannot offer autoformatting. The Adobe Acrobat PDF format on the other hand was designed to copy the look of your documents: PDF documents by nature imply autoformatting.

When the recognized text is opened using a wordprocessor, the text looks like this without *any* intervention by the user.



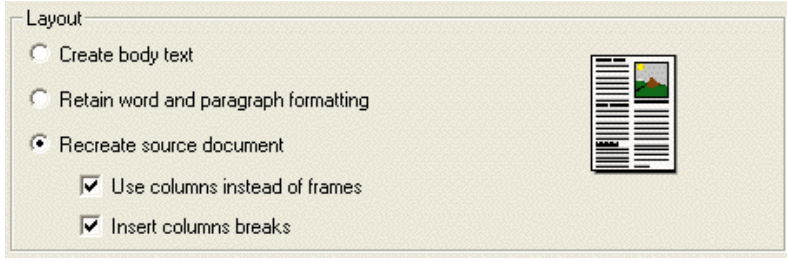
To see the effect correctly, you need to enable the “WYSIWIG” mode of your wordprocessor, mostly called “page layout” mode. However, if you send the recognized document directly to Microsoft Word, the page or print layout view is activated automatically!



In short, Readiris not only recognizes your texts, but can format them for you as well. OCR isn't just text recognition anymore, it is becoming more and more **page** or document **recognition** as well!

COLUMNS PLEASE, NOT FRAMES!

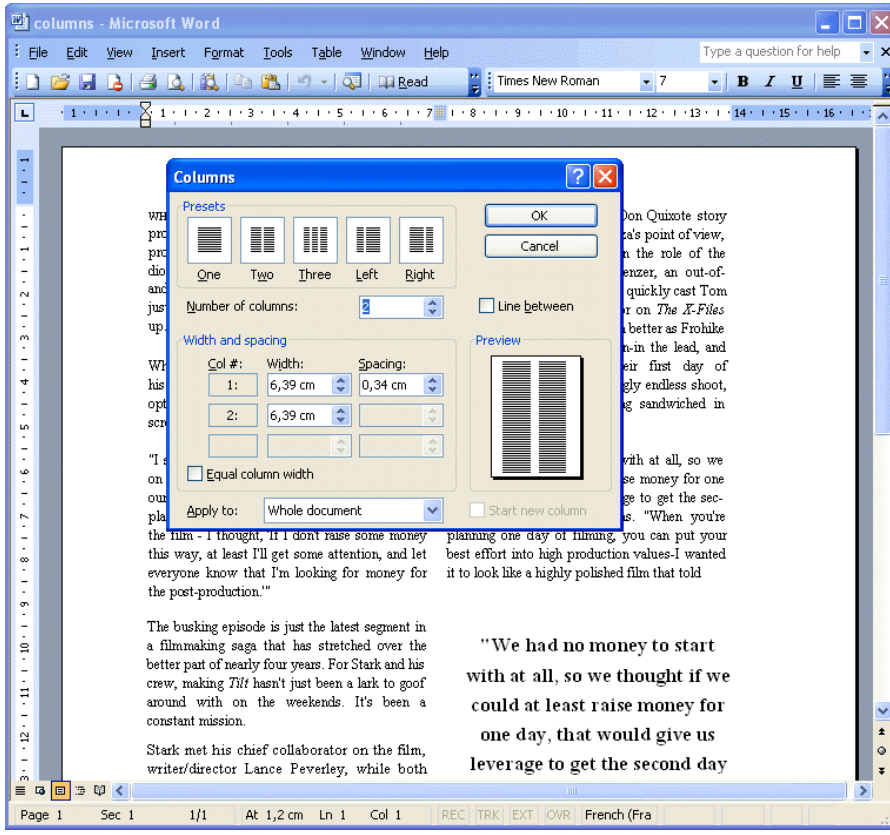
The formatting option "Use Columns instead of Frames" determines *how* the "autoformatting" gets done: the text blocks, tables and graphics can either be stored in frames or in editable **columns**.



"Frames" are separate containers for text used to position several blocks of text, graphics and tables on a page. With columns, the text flows naturally from one column to the next, and columnized texts are much easier to edit.

We now assume that real columns do occur on the scanned document: when the system is unable to detect columns in the source document, this formatting mode uses frames anyway as a "fallback" position!

You can make good use of the image COLUMNS.JPG in the Readiris folder if you want to try it.



The option "Insert Column Breaks" refines the recreation of columns: it determines whether you insert “hard” column breaks at the end of each column or not. With column breaks, any text you edit, add or remove remains inside its column;

no text ever flows automatically across a column break. All text that follows a column break is moved to the top of the next column!

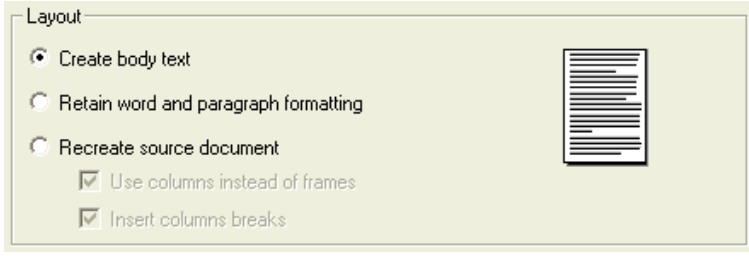
Enable this option when you want to maintain column breaks where these were detected in the recognized document - whatever text editing gets done after the OCR. In newspapers and magazines, the various columns on a page often correspond to different article "threads". Having text flow from one column to the next "on the sly", covertly may not be a good idea!

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

TEXT FORMATTING, PART 2

The other layout options are "Create Body Text" and "Retain Word and Paragraph Formatting".

As the icon on the right side illustrates, creating **body text** means you create a non-formatted, "running" text. The text will be captured, but its formatting is entirely ignored. Use this option when you just need to recapture a text but not its layout.



Body text is also what you get when you quickly recognize a text zone by right-clicking it and selecting the command "Copy as Text": when the recognition is done, you'll paste body text into your text application.

The option "Retain Word and Paragraph Formatting" represents the middle road: the **word formatting** - font type, point size and typestyle - is retained

across the recognition, and so is the **paragraph formatting** - the tabs and the alignment.

Don't confuse this formatting option with "full" autoformatting: this option just puts one paragraph after the other, it does not recreate columns or copy the relative position of the various zones.

Exporting Text Several Times

Actually, you can export the OCR results several times without repeating the recognition! Change the text format and the formatting options under the "Format" button and click the button "Recognize" again. No OCR is executed this time - unless you defined new windows or modified existing ones! Otherwise Readiris just reformats the OCR results and saves them in the new text format or sends them to the target application you've just selected.

The same goes for any other element you change: when you add a page to your OCR job, only that page will be recognized. When you create a new text zone on any page, only that zone will be recognized before the results get exported.

You could for instance recognize a 10 page document and save it in a Word file. Then you quickly scan the abstract found on the cover page and send it by e-mail to an impatient colleague to finally scan the appendix - a table - and save all results in an HTML file to be posted on your company's web site.

Saving Graphics Separately

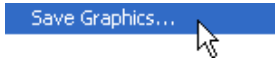
In our example, the graphic was included in the recognized text; whether this is the case depends on the formatting option "Include Graphics". Whether it is

possible to save graphics inside the text again depends on the output mode. "Poor" text formats such as Text (ANSI) etc. don't store graphics!

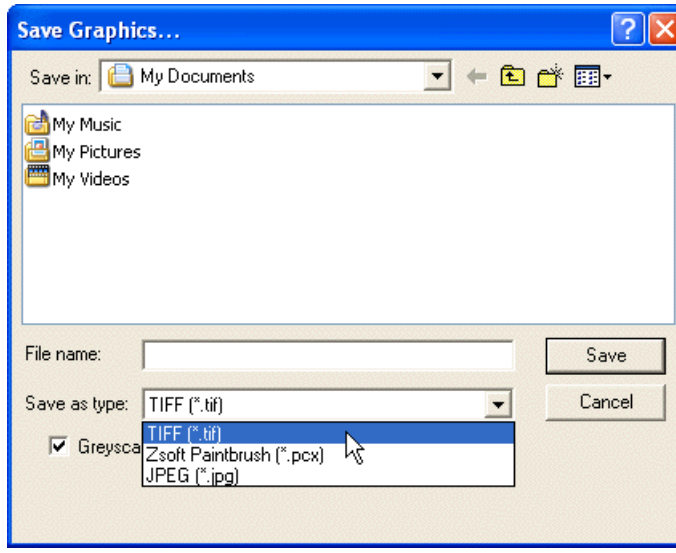
Still, with Readiris, you can save graphics without performing text recognition. As Readiris generates black-and-white, greyscale and color images, you can capture lineart graphics and photos.

How? Draw a graphic zone around the illustrations, cartoons etc. you need. Creating graphic windows manually is done in the same way as drawing text and table windows, simply select the "Graphic Window" tool now.

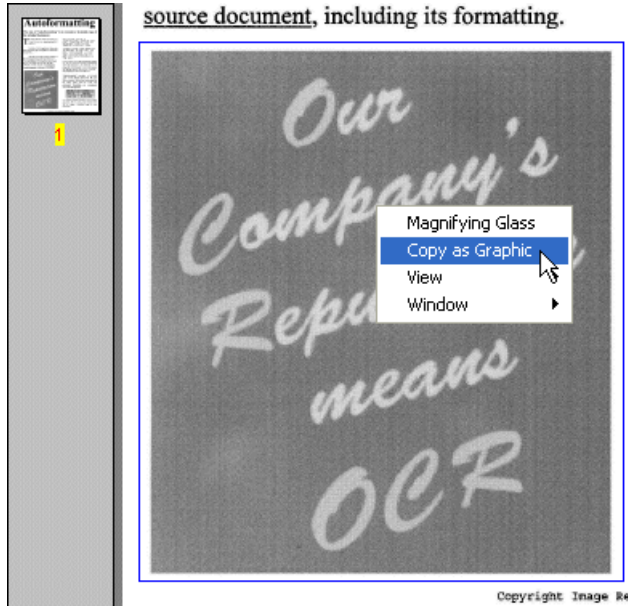
Next, choose the command "Save Graphics" under the "File" menu.



You are prompted to specify a filename. Determine which graphic file format you will use. Select a format that's supported by your paint or photo retouching software. The JPEG, TIFF and Paintbrush (PCX) formats are supported. Enable the option "Greyscale/Color" to save the graphic as a color or greyscale graphic.



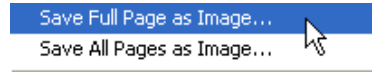
To send a graphic to the clipboard rather than save an image file, right-click your mouse over a graphic window and select the command "Copy as Graphic": the graphic zone under the mouse pointer is ready to be pasted!



READING FAXES AND DEFERRED RECOGNITION

Saving images as image files opens another possibility: you can save the *full* page and perform **deferred OCR** on it later on. That's what we did with the prescanned images of our tutorials.

Simply scan the document. Select the command "Save Full Page as Image" under the "File" menu to save a single page. You'll again be prompted to save the entire page as TIFF or Paintbrush (PCX) file.



Select the command "Save All Pages as Image" to save a multipage document. A single file format is available here: multipage TIFF.

You can now select the disk as image source and open the image file with the "Open" button (or with the corresponding command under the "Process" menu). (If you use the "Open" command under the "File" menu, you don't even have to update the image source.)

As color, greyscale and black-and-white images are supported on an equal basis, Readiris opens Adobe Acrobat PDF documents, JPEG images, Paintbrush (PCX) images, DCX fax images (a multipage version of the Paintbrush format), PNG images, TIFF images (uncompressed, LZW, PackBits, Group 3 and Group 4 compressed), multipage TIFF images and Windows bitmaps (BMP).

This capability is particularly useful to convert your **faxes** into editable text files! Readiris uses extra intelligence when it comes to reading faxes: the software detects the typical fax resolutions - 100 x 200 dpi ("normal quality"), 200 x 200 dpi ("fine quality") and 200 x 400 dpi ("superfine quality") - and "preprocesses" these images automatically to ensure optimal OCR results.

Nevertheless, it's still a good idea to ask your correspondents to send faxes with the "fine" quality - those faxes will yield better OCR results.

Don't forget that you can right-click on images in the Windows Explorer and select the command "Recognize" from the "Context" menu to open images! Alternatively, you can use "drag and drop": drop image files from the Windows Explorer onto the image zone or icon of Readiris and they are promptly opened.

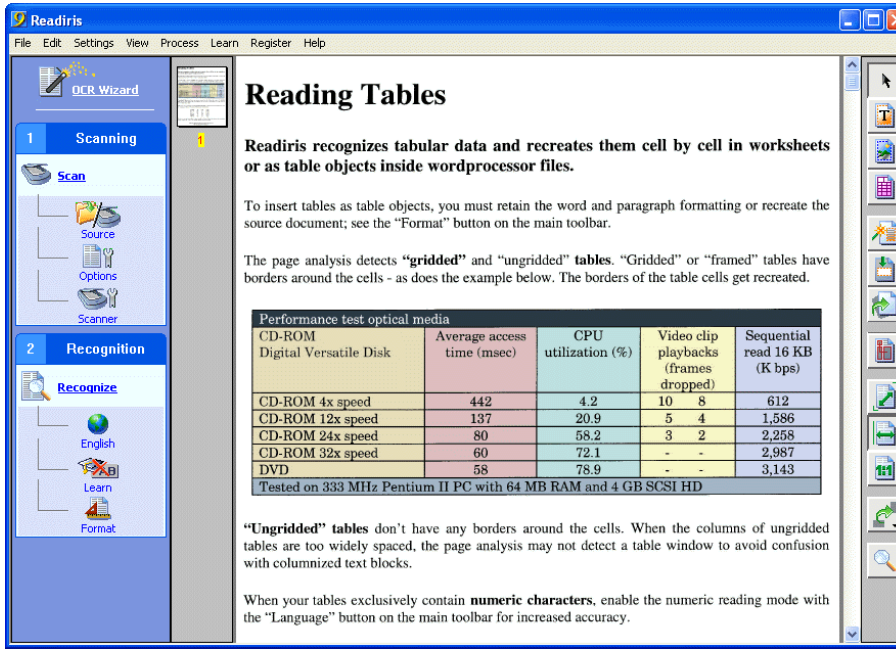
Recognizing Tables

So far, we've recognized texts and faxes and we've saved graphics. Let's process a table now. Take a table of figures and scan it, or open the sample image TABLES.JPG in your Readiris folder.

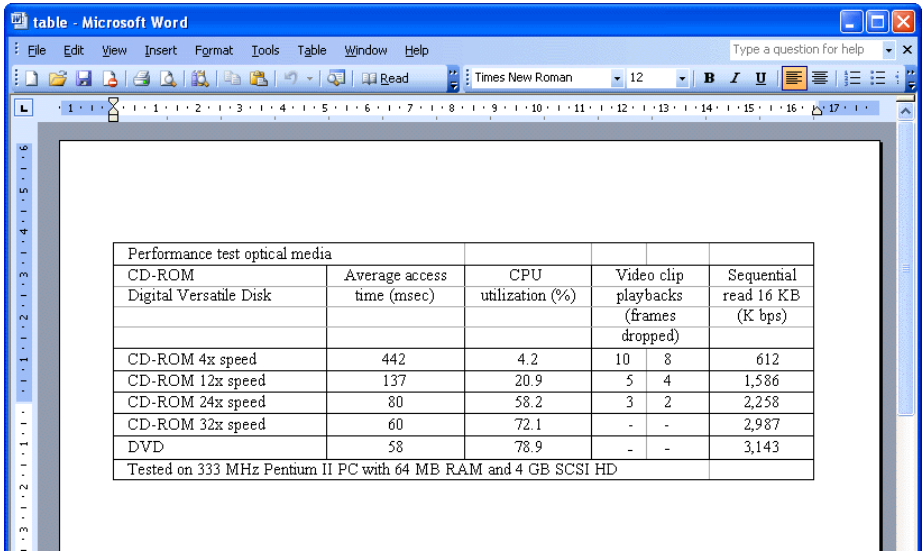
Actually, the image TABLES.JPG contains two tables, and that's no coincidence! The page analysis zones them as table windows, and Readiris will recon-

struct them for you by recreating the tables cell by cell in your spreadsheet or by inserting a table object inside your wordprocessor files.

Let's explore the different solutions, starting with the "gridded" or "framed" table - it has borders around the cells.

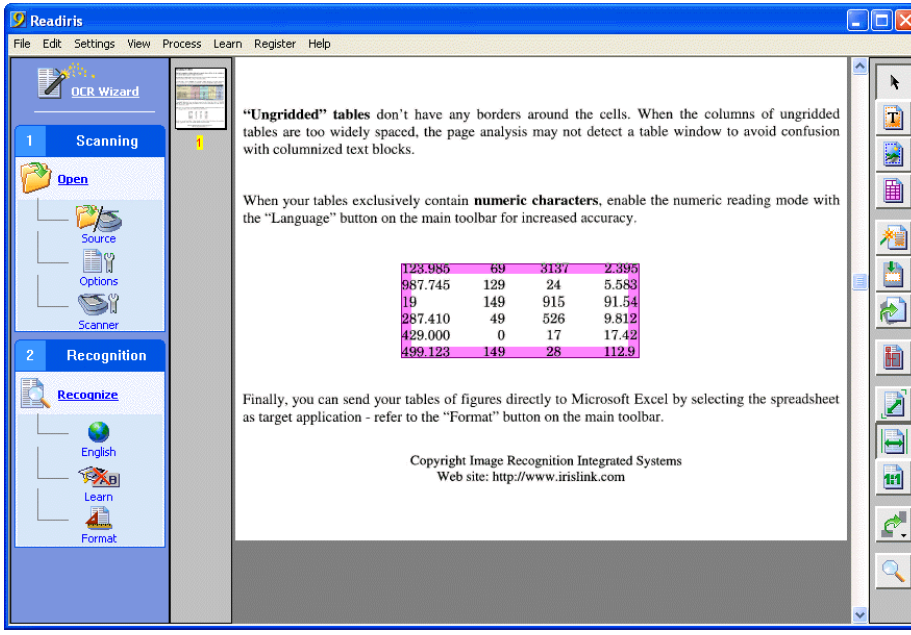


Run the recognition with the layout option "Retain Word and Paragraph Formatting" or "Recreate Source Document" enabled and the table gets recreated. Open your wordprocessor to have a look at the result: the cells and the borders were recreated by Readiris one by one! (You could obviously have included the text paragraphs in the text file as well.)



Performance test optical media					
CD-ROM Digital Versatile Disk	Average access time (msec)	CPU utilization (%)	Video clip playbacks (frames dropped)		Sequential read 16 KB (K bps)
CD-ROM 4x speed	442	4.2	10	8	612
CD-ROM 12x speed	137	20.9	5	4	1,586
CD-ROM 24x speed	80	58.2	3	2	2,258
CD-ROM 32x speed	60	72.1	-	-	2,987
DVD	58	78.9	-	-	3,143
Tested on 333 MHz Pentium II PC with 64 MB RAM and 4 GB SCSI HD					

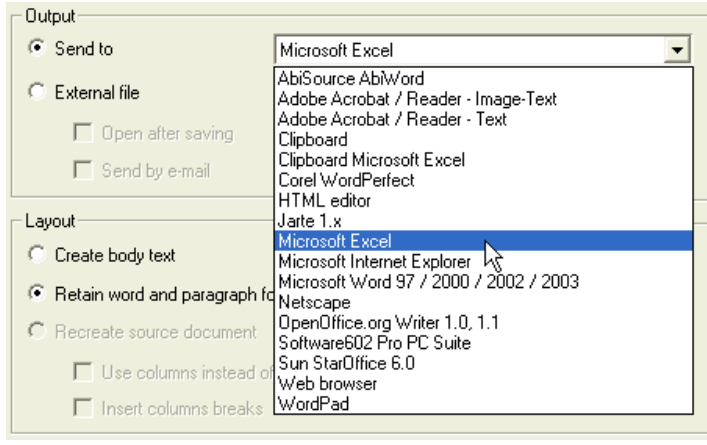
Now the “ungridded” example - it has no borders around the cells. Note that the page analysis nevertheless detects the table!



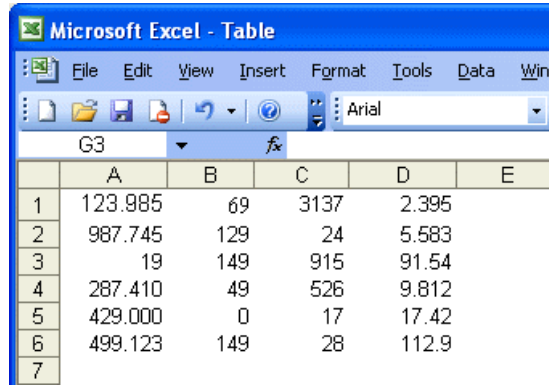
For optimal OCR accuracy, you should limit recognition to the **numeric symbols** with the "Language" button. (The numeric mode is not strictly numeric, it includes the symbols 0 to 9, +, *, /, %, , (comma), . (dot), (,), -, =, $, £, ¥ and the € symbol.)

As you can only do this when the table doesn't contain any alphabetic symbols - otherwise the text portions won't be recognized correctly - we can activate the numeric mode now but couldn't do it for the first table.

This time, we will send the OCR result directly to the spreadsheet Microsoft Excel, so we select Excel as target application under the "Format" button.



The spreadsheet is started up automatically and the result looks like this: the typical table structure with rows and columns is recreated, and you are immediately ready to process the data.

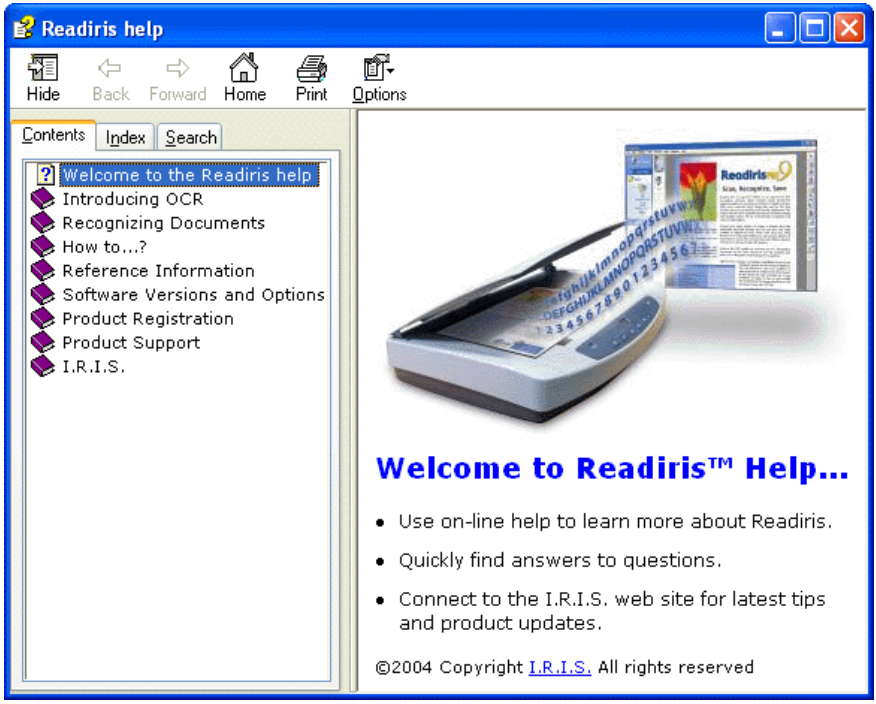


You may come across "ungridded" tables the page analysis does not detect as table zones because the columns are too widely spaced - Readiris tries to avoid confusion with columnized text blocks. To create a table window manually, click on the "Table Window" tool in the image toolbar and proceed as usual; the button's tooltip again indicates the number of table windows.

GETTING ON-LINE HELP

This concludes our overview of Readiris. Some last-minute information may not be included in this manual. We thus recommend you to consult the on-line help system for additional information on Readiris.

Go to the "Help" menu to do so. The command "Help Topics" and its shortcut key F1 allow you to navigate through the many help topics.



The other commands of the "Help" menu tell you how to get product support, how to contact I.R.I.S., give direct access to the I.R.I.S. home page etc.