

# Readiris™ PRO 10

## USER'S GUIDE



Readiris Corporate

© 2005 I.R.I.S. All rights reserved

OCR technology by I.R.I.S.

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

ICR and bar code reading technology by I.R.I.S.

BCR and field analysis technology by I.R.I.S.

© 2005 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, graphics, business cards, bar codes and even handwritten text!

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 promptly 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-independent 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 117 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 Hebrew documents and 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, Hebrew and Asian documents - many untranscribable 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 timesteps as used in the source document are maintained across the recognition. The placement of columns, text blocks and graphics follows your original documents. Readiris can even include the background photo of a scanned page in the recognized document! 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.

Bar codes that occur on a scanned page can also be read, and the same goes for handwritten text - such text can be captured as long as you write well-spaced "block letters".

You can even recognize business cards with Readiris: scan your business cards, recognize them and convert them into an address database. Think of your last exhibition when you came back with an entire stack of business cards and it took your secretary two days to encode them!

The cards' data is extracted automatically from the image and the recognition result is assigned to specific database fields. Readiris extensively uses a knowledge database, thus acquiring the necessary intelligence to discriminate the first and last name, a city and its state, a telephone and a fax number etc. The resulting data can be sent directly to your contact management software such as Microsoft Outlook (Express) or any vCard compliant application.

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. Interval scanning allows you to scan multipage documents efficiently when your scanner is not equipped with a document feeder. (Readiris Corporate supports high-speed scanners and executes batch OCR on large image collections: blank pages can be used to segment scanned batches into separate documents, automatic bar code reading ensures the proper indexing of the "dematerialized" documents.)

## **TABLE OF CONTENTS**

---

Save Time, No More Retyping! .....	III
Table of Contents .....	V
Credits and Copyrights .....	VIII



## **Chapter 1: Installation**

System Requirements .....	1-1
Installing the Readiris Software .....	1-1
Uninstalling the Readiris Software .....	1-3
Readiris “uninstall” program .....	1-3
Windows (un)install wizard .....	1-4
Installing Software Options .....	1-5
Installed Files .....	1-7
Read Me file and documentation .....	1-7
Handprinting form .....	1-7
Scanner drivers .....	1-7
Register to Vote! .....	1-8
Getting Product Support .....	1-9
Getting in Touch with I.R.I.S. ....	1-10

## **Chapter 2: Guided Tour**

Starting the Software up .....	2-1
The First-Time Startup .....	2-2
Discovering the Readiris Interface .....	2-2
Getting Started with a First Tutorial .....	2-5
Zooming in on Images .....	2-10
One, Decomposing a Scanned Image .....	2-13
One and a Half, Sorting Windows .....	2-16
Two, Windowing a Scanned Image Manually .....	2-19
Three, Saving Windowing Templates .....	2-23
Readiris Takes You around the World .....	2-25
User Lexicons to “Boost” the Linguistics .....	2-30
Readiris Changes Languages As Needed .....	2-32
Reading Documents with Mixed Languages .....	2-34
Defining the Document Characteristics .....	2-36
Readiris Gets More Intelligent Each Time! .....	2-37
Learn .....	2-40
Don’t Learn .....	2-40
Delete .....	2-41
Undo .....	2-41
Finish .....	2-41
Abort .....	2-41
The Role of Font Dictionaries .....	2-41

Sending Results Directly to Your Application .....	2-44
Saving the Results in a Text File .....	2-48
Creating Portable Documents.....	2-51
... Or Reading Them .....	2-59
Recognizing Multiple Pages .....	2-62
Editing multipage documents .....	2-71
Starting a New Document .....	2-73
Recognizing Text Zones .....	2-74
Organizing the Text Output .....	2-76
Setting up Your Scanner .....	2-77
Let the Bad Color Not Be Seen .....	2-79
Different Devices, Different Resolution .....	2-82
Saving Default Settings .....	2-86
Saving Specific Settings .....	2-87
Scanning Documents .....	2-88
Adjusting the Scanned Images .....	2-91
Letting the OCR Wizard Work for You .....	2-96
Readiris Recreates Your Document Layout .....	2-97
Columns Please, Not Frames! .....	2-102
Text Formatting, Part 2 .....	2-105
Exporting Text Several Times .....	2-106
Saving Graphics Separately .....	2-107
Saving Colored Backgrounds .....	2-109
Taking Graphics to the Hilt .....	2-112
Reading Faxes and Deferred Recognition .....	2-114
Recognizing Document Batches .....	2-115
Establishing a Watched Folder .....	2-117
Organizing Batches .....	2-119
Recognizing Tables .....	2-122
Recognizing Handwritten Text .....	2-127
Reading Bars and Spaces .....	2-131
Reading Business Cards .....	2-132
Scanning Business Cards .....	2-133
It Takes a Business Card Reading Mode! .....	2-138
Recognizing Business Cards .....	2-141
Getting On-line Help .....	2-143



## CREDITS AND COPYRIGHTS

---

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

AutoFormat, Connectionist, the IBCR-II, 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](http://www.apache.org)).

Acrobat and Reader are (registered) trademarks of Adobe. 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.
- 64 MB RAM. 128 MB RAM is recommended to process greyscale and color images.
- 120 MB free disk space. 105 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.

Readiris Corporate requires a monitor with a 1,024 x 768 resolution.

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**

---

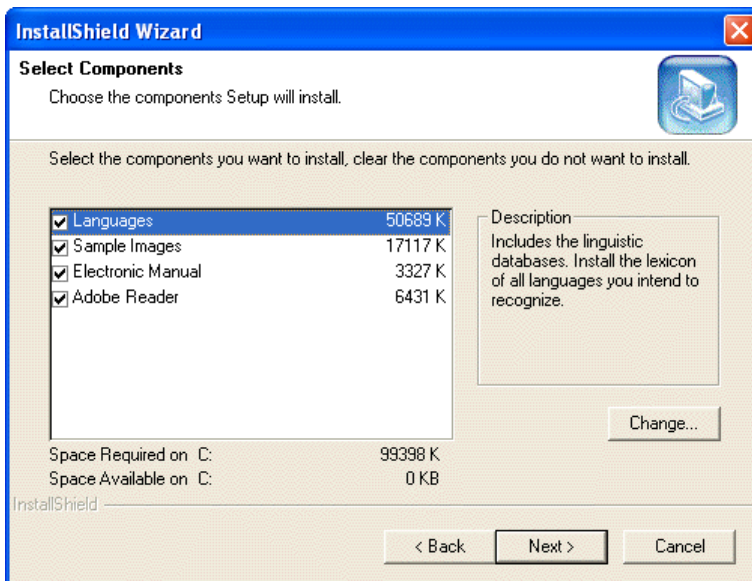
The Readiris software is delivered exclusively on an **autorunning CD-ROM**. To install, simply insert the CD-ROM in your CD-ROM drive and wait for the installation program to start running. Follow the on-screen instructions.



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.



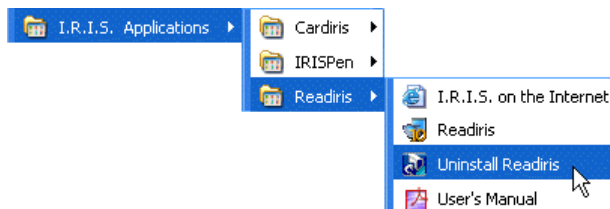
The same holds for a **shortcut** to Readiris on the Windows **desktop**. As a result, you are able to start Readiris directly from your desktop.

## 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.

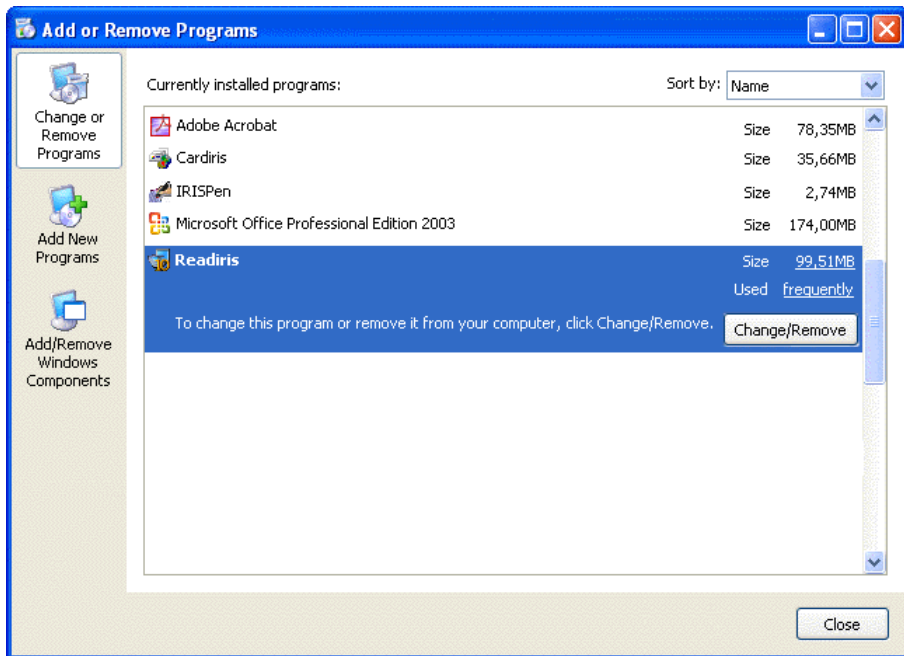




## Windows (un)install wizard

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

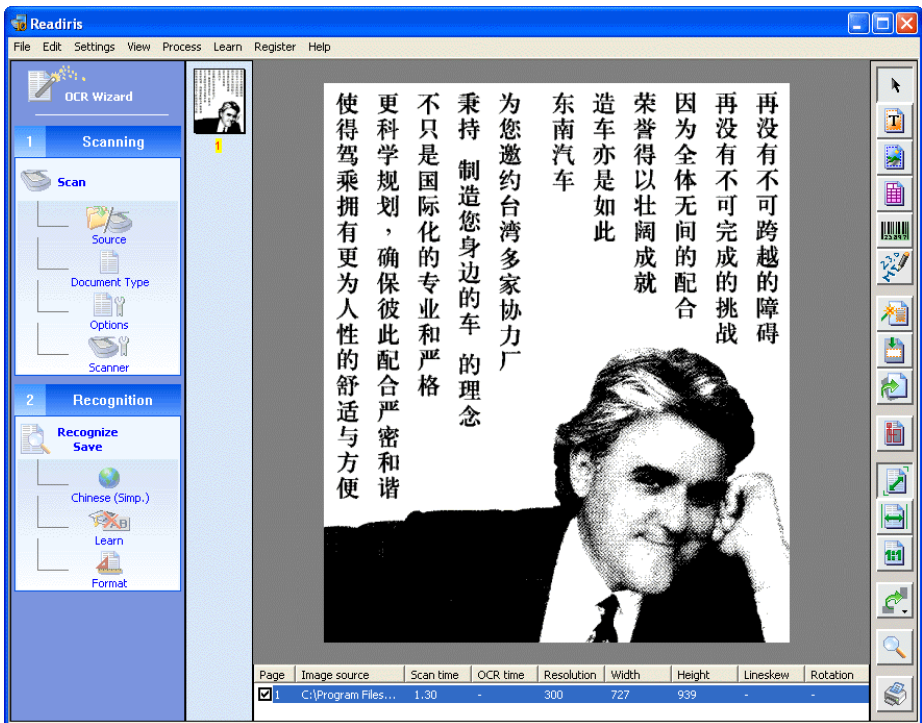
- Click "Settings" under the "Start" menu of Windows and go to the "Control Panel".
- Click the icon "Add/Remove Programs" under the control panel.



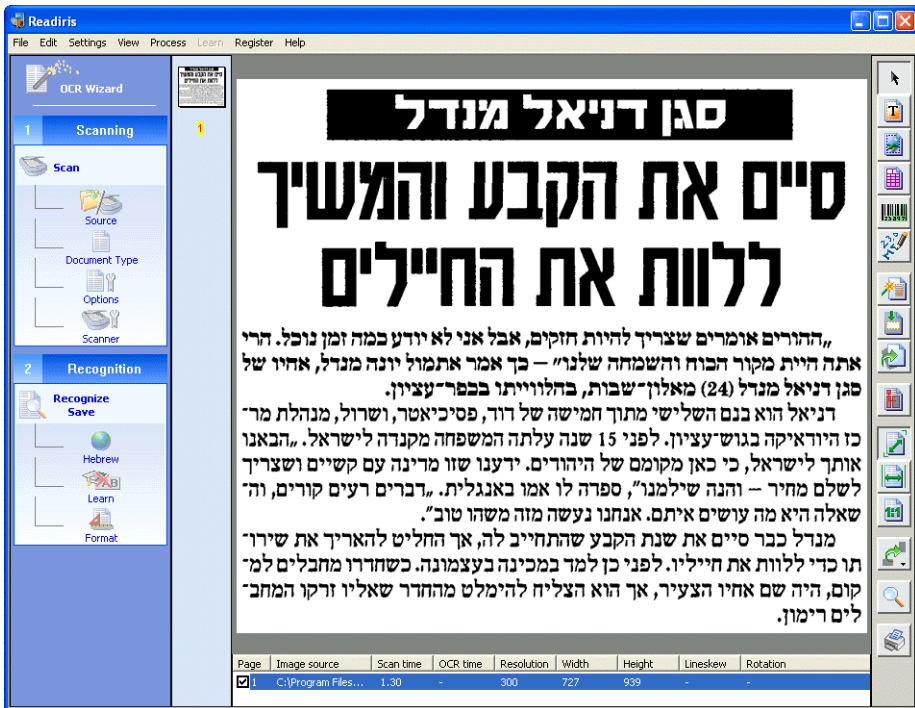
- Follow the on-screen instructions to remove the Readiris software.

## INSTALLING SOFTWARE OPTIONS

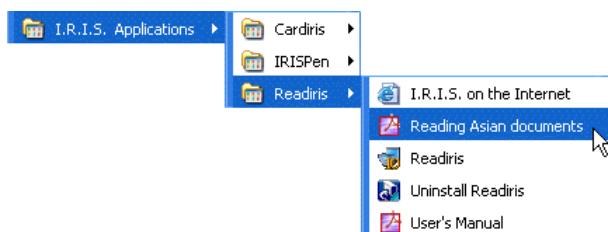
There are two software options available for the Readiris software: the “Asian OCR add-on” and the “Hebrew OCR add-on”. The “Asian OCR add-on” allows you to read Japanese, Traditional Chinese, Simplified Chinese and Korean.



The “Hebrew OCR add-on” predictably allows you to recognize Hebrew documents.



These software options are again delivered on an autorunning CD-ROM. By installing the “Asian OCR add-on”, specific documentation becomes available that discusses how you can recognize Asian documents.



## **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)

### **Handprinting form**

TEMPLATE.PDF Blank handprinting form for reprinting

TEMPLATE.DOC Blank handprinting form for editing

### **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 screenshot shows a web browser window titled "Readiris help". The address bar contains "Readiris help". The browser's navigation toolbar includes buttons for Hide, Back, Forward, Home, Print, and Options. Below the toolbar are tabs for Contents, Index, and Search. The left sidebar displays a tree view of the help content, with "Register your Readiris license" selected. The main content area is titled "Register Your Readiris License" and contains the following text:

### Register Your Readiris License

**Why you should register**

- Registering allows us to keep you informed of future **product developments** and **related I.R.I.S. products**.
- Registering entitles you to free **product support** and **special offers**.
- Depending on the software bundle, you'll receive the **softkey** in return as may be needed to continue using Readiris after one month.

**How to...?**

**Mail**

Send in your **registration card**.



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.

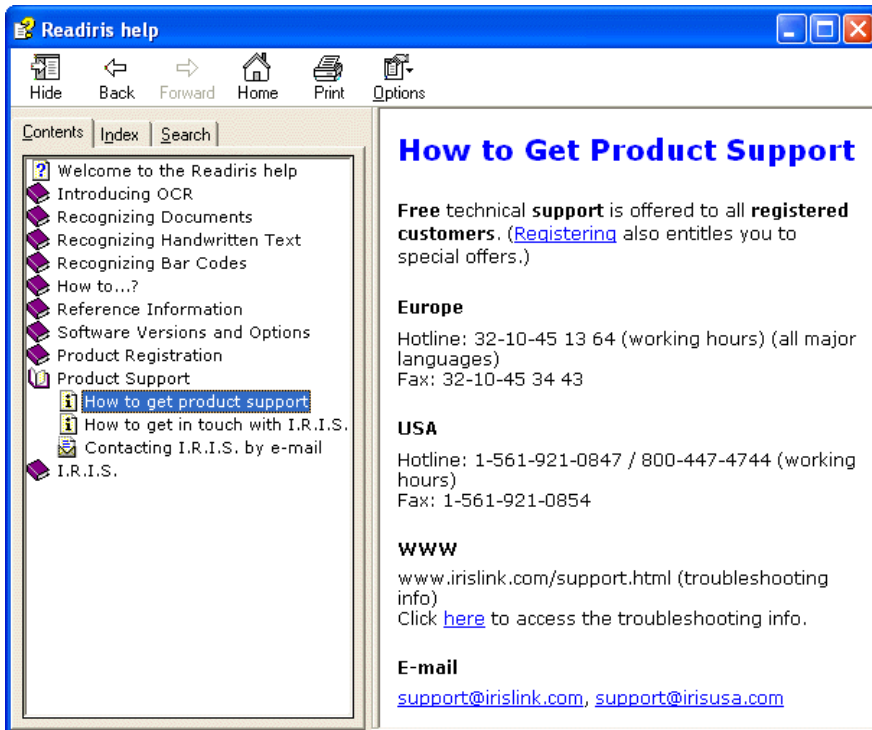
## GETTING PRODUCT SUPPORT

---

The command "Product Support" under the "Help" menu of Readiris details how you can get technical support. Please describe the phenomenon you experi-



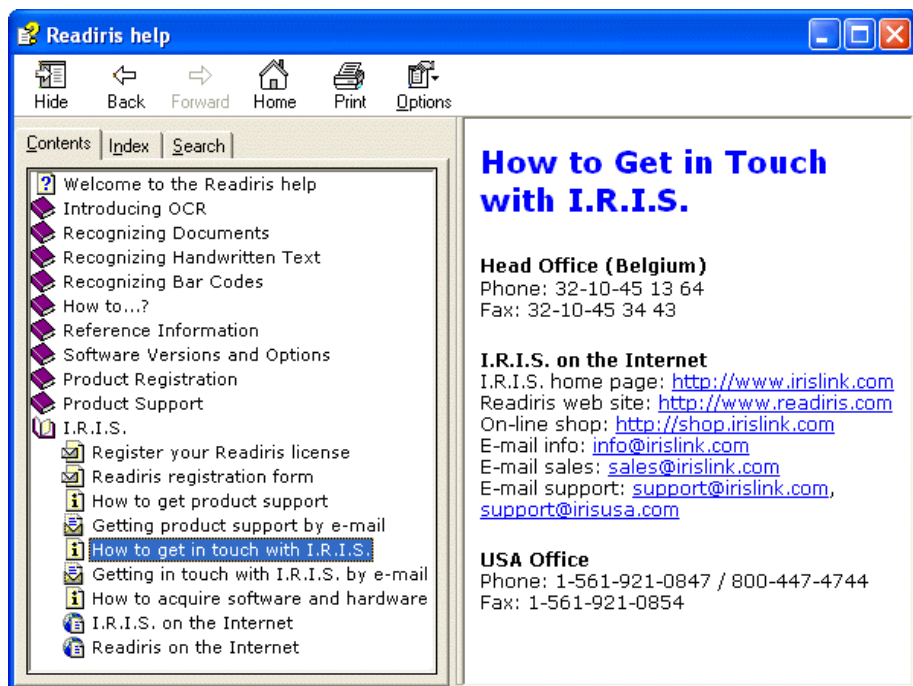
ence clearly and include all relevant data concerning Readiris, your scanner and your computer system.



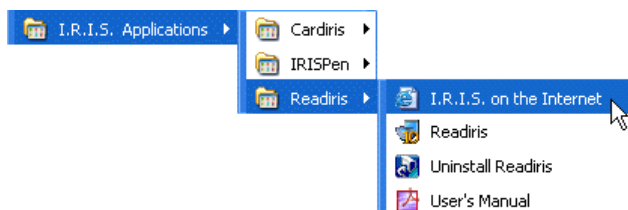
## GETTING IN TOUCH WITH I.R.I.S.

---

You can also contact I.R.I.S. to learn more about other solutions from its product range; 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.



# Chapter 2

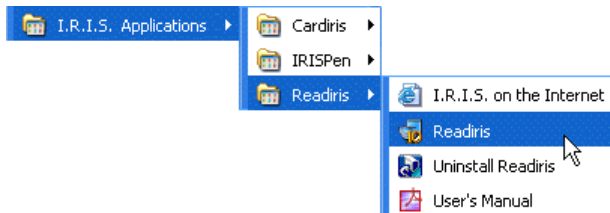
## GUIDED TOUR

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.



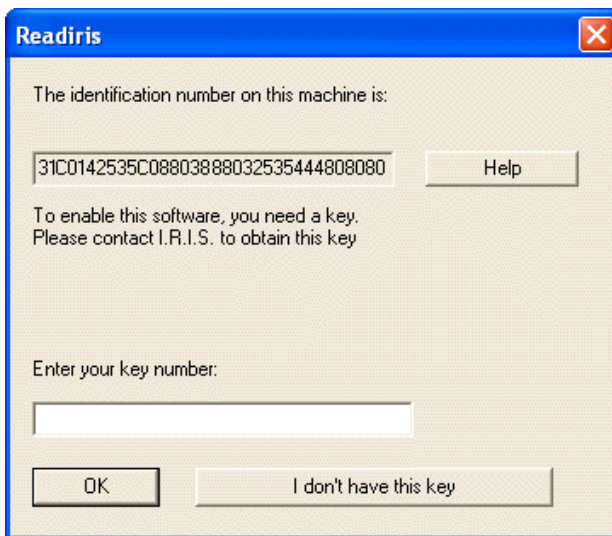
## THE FIRST-TIME STARTUP

---

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.

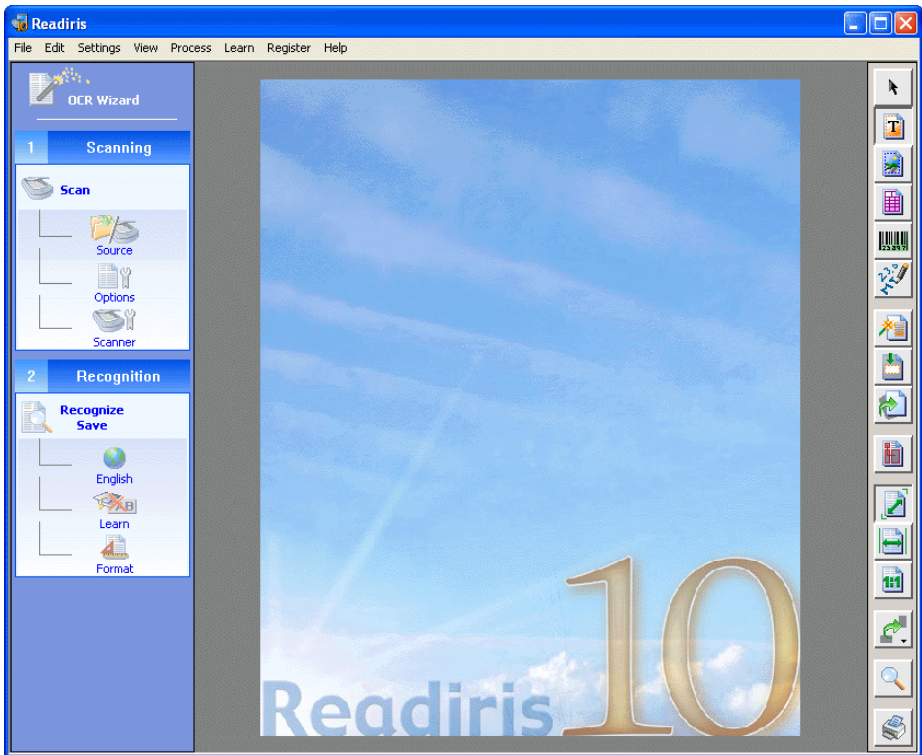


## DISCOVERING THE READIRIS INTERFACE

---

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.

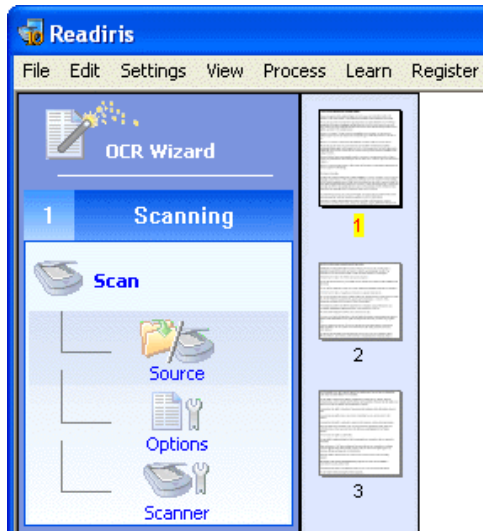
To learn which command corresponds to a certain button, hold your mouse pointer over it for a while: a **tooltip** will tell you what the button does.



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).





Secondly, the **document panel** is then displayed below the scanned image. It displays **statistics** and information on all scanned pages - the image source and **resolution**, the scanning and recognition time etc. (The document panel comes with tooltips too...)

Page	Image source	Scan time	OCR time	Resolution	Width	Height	Lineskew	Rotation
<input checked="" type="checkbox"/> 1	C:\Program Files\Readiris\multipage.tif	2.74	-	300	2000	2388	-	-
<input checked="" type="checkbox"/> 2	C:\Program Files\Readiris\multipage.tif	2.08	-	300	2000	1888	-	-
<input checked="" type="checkbox"/> 3	C:\Program Files\Readiris\multipage.tif	2.05	-	300	1912	2004	-	-

## 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.

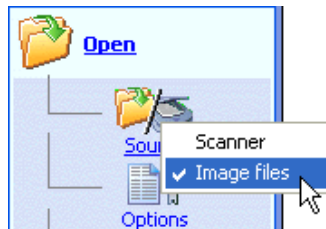


The "Source" button on the main toolbar determines whether you are going to use a scanner or a prescanned image as image source.

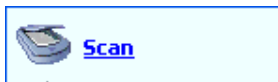
Color, greyscale and black-and-white images are supported on an equal basis. Readiris allows you to open Adobe Acrobat PDF documents, DCX fax images (a multipage version of the Paintbrush format), DjVu images (\*.djv, \*.djvu), JPEG images, JPEG 2000 images (\*.j2c, \*.jp2), PNG images, TIFF images (uncompressed, LZW, PackBits, Group 3, Group 4 and JPEG compressed), multipage TIFF images, Windows bitmaps (\*.bmp) and ZSoft Paintbrush images (\*.pcx).

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

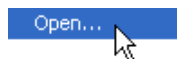
As you are going to open a prescanned image, you should select "Image Files", and not the scanner, as image source with the "Source" button.



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 and document panel indicate that a single page is loaded into Readiris.



Readiris

File Edit Settings View Process Learn Register Help

OCR Wizard

1 Scanning

Scan

Source

Options

Scanner

2 Recognition

Recognize

Save

English

Learn

Format

## Give the Brain a Break

*A device to avoid being sucked under by the information maelstrom*

IF THERE'S ONE PART of this magazine I TRY HARD NOT to read it's the section we call *Numbers*, near the front. My colleague who prepares it (I dare to write this because she's on vacation) has the knack of culling statistics which some mental nicotine makes you suck in: the proportion of Dutch couples who live in what used to be called sin (32%); the amount paid at auction for an Eric Clapton guitar (\$497,500); the estimated number of condos used in the world annually (9 billion); or how high the pile would be if Bill Gates stacked his wealth in \$1 bills (25,000 km).

The trouble with this fun is that it leaves stat clots in your brain. If you can't do some plumbing to restore thinking's normal flow, they accumulate like cholesterol up there in the grey matter's byways, causing serious mind-blocks. After three glasses of red you declare over dinner that Bill Gates has used nine billion condoms, Eric Clapton had a 25,000-km long guitar, and 32% of Dutch couples earn \$497,500 a year. The only cure for this condition of factoid overload—a sort of neurotic plague which is set to be the main cause of mental illness in the 21st century—is to learn to sift. Survivors will be those with the capacity to distinguish Internet surfing from info-surflet.

The first step to becoming a sworn sifter is easier than giving up smoking. It only requires a little virtual brain surgery to find a warm spot in one of your lobes and implant there an imaginary device smaller than the delete key on a computer. Psychotherapeutic, but as efficient as dopamine or serotonin, its medical acronym might be *DNK*, which is not a kinky variant of *DNK* but stands for Don't Need to Know.

After undergoing a voluntary *DNK* implant, I now shiver with pleasure on opening my morning newspapers as factoid after factoid that I don't need on my poor old soft disk is burned from memory's gates. An example: scientists in Texas say each year the Earth and the Moon move 3.82cm further apart, meaning that since Neil Armstrong was taking giant steps there 30 years ago, we are 114cm more distant. Standing at the back door looking up at the Big Cheese, does my soul benefit from this knowledge? No. Slam *DNK*!

There are forests of facts which can not be so easily shed, as with some of the far-from-trivial statistics that jolt in our *Numbers* spot. A recent example is that 29 million people in the world live in some form of slavery. That needs to be stored and pondered. As does, say, a recent finding by the London School of Economics that four million British children live in poverty, relative though that is. This knowledge can't be *DNK*ed if we are to remain half human, but precisely because there is such a welter of factoids the importance of big numbers like these dilutes into a milky way of mental numbness.

It's not just numbers, it's also words. Don't-Need-to-Know beginners can, for example, zap any report to do with rats. Not the sewer sort but the laboratory kind, the ones which almost daily lead to variants of: "Tweaking gene 859G and injecting fibroblastokininase greatly reduces hair loss in male rats over 50. Scientists at Brillolab caution that many years of research will be required before..."

Sport offers superb *DNK* material. You can whitewash all that coaches say before a match and most of what they say after it, along with what the players mutter about why they lost/won. It's the event, stupid! If you've watched Lance Armstrong through each stage of the Tour de France, do you really need to learn it wasn't easy on his legs?

Politics is the promised land of *DNK*. Meaning you blank all promises and pre-election bumps, all state-of-the-nation monologues. One needs to take in only how the votes fall, all denials (on the where-there's-smoke basis), and party funding figures. Where politics bleeds into economics, there is also room to turn a deaf ear. *DNK* all interest rate forecasts, stockgaging and merger talk because it's about as accurate as tarot, and anyway you never hear it in time to benefit from it.

Criticism can be a joy, but keep the *DNK* button ready for those experts who can't resist giving the whole plot away, explaining what the painting really means, or who majored in semantics. It's easy to build such personalized *DNK* lists. My own long one also includes all advertisements (especially those offering "personalized" things), and anything to do with any royal marriage, anywhere. Also flamed out are all numbers about flying saucers, mobile phone sales, desirable cholesterol levels, winners of the U.S. Masters, and guesstimates of how much Bill Gates has in the bank, or how it might stack up.

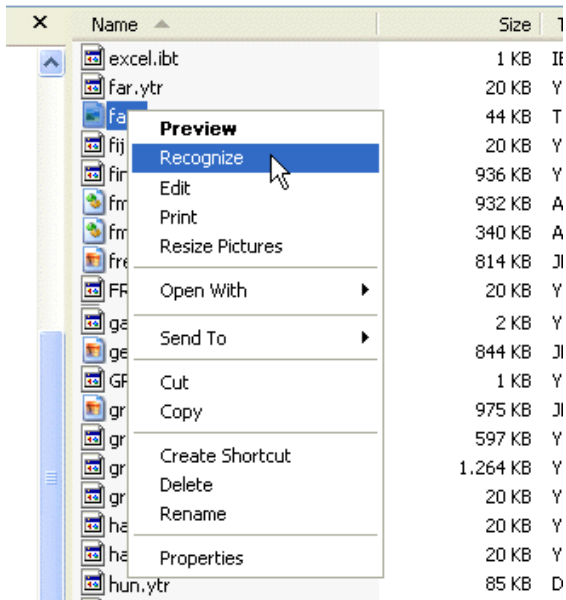
The *DNK* club is free and there are no membership numbers. Our catholic motto might sound familiar—if you haven't yet *DNK*ed out all slogans - It's No drain, no brain.

Page	Image source	Scan time	OCR time	Resolution	Width	Height	Lineskew	Rotation
1	C:\Program Files...	1:30	-	300	727	939	-	-

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.

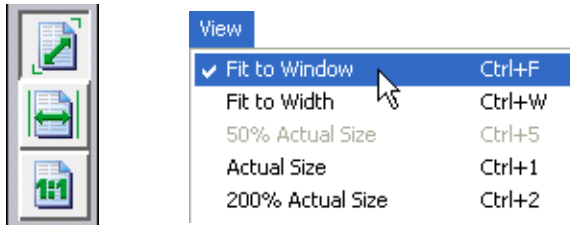
The image toolbar on the right side of the Readiris application window contains all commands you need during the image preview: tools to indicate the zones of interest, to rotate the image, zoom in and out etc.

## ZOOMING IN ON IMAGES

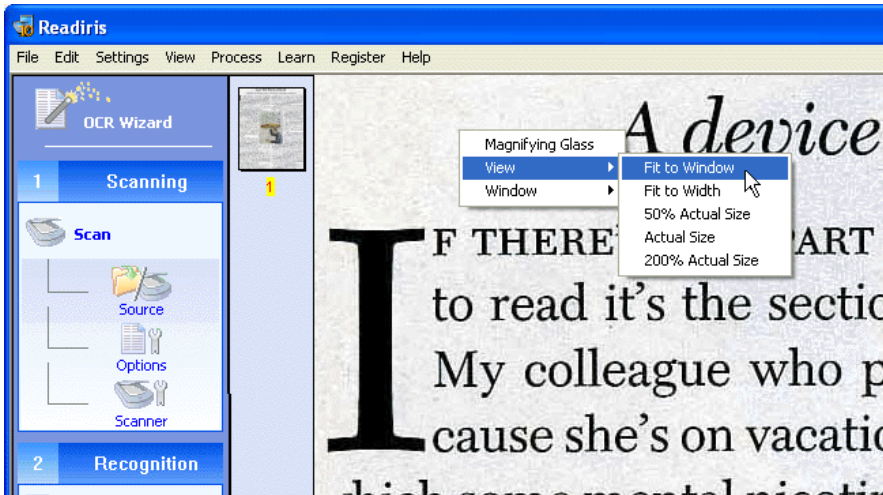
---

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.



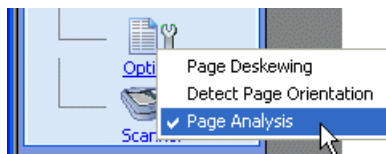




## ONE, DECOMPOSING A SCANNED IMAGE

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.

The screenshot shows the Readiris software interface. The main window displays a scanned document with the following content:

### Give the Brain a Break

2 device to avoid being sucked under by the information maelstrom

3 I make a new plan or time adjustment. I try not to read it's the section we call Numbers, near the front. My colleague who prepares it (I dare to write this because she's on vacation) has the knack of calling statistics which some mental arithmetic makes you suck in: the proportion of Dutch couples who live in what used to be called sin (32% of the amount paid at auction for an Eric Clapton guitar \$497,500); the estimated number of condos used in the world annually (9 billion); or how high the pile would be if Bill Gates stacked his wealth in \$1 bills (25,000 km).

The trouble with this fun is that it leaves fat clots in your brain. If you can't do some plumbing to restore thinking's normal flow, they accumulate like cholesterol up there in the grey matter's byways, causing serious mind-blocks. After three glasses of red you declare over dinner that Bill Gates has used nine billion condoms. Eric Clapton had a 25,000-km-long guitar, and 32% of Dutch couples earn \$497,500 a year. The only cure for this condition of factual overload—a sort of neuronal plague which is set to be the main cause of mental illness in the 21st century—is to learn to sit. Survivors will be those with the capacity to distinguish Internet surfing from info-surfing.

The first step to becoming a survivor is easier than giving up smoking. It only requires a little virtual brain surgery to find a warm spot in one of your lobes and implant there an imaginary device smaller than the delete key on a computer. Psychotherapeutic, but as efficient as dopamine or serotonin, its medical acronym might be **DNX**, which is not a kinky variant of sex but stands for Don't Need to Know.

After undergoing a voluntary **DNX** implant, I now shiver with pleasure on opening my morning newspapers as factual after fact that I don't need on my poor old soft disk is burned into memory's gates. An example: scientists in Texas say each year the Earth and the Moon move 3.82cm further apart, meaning that since Neil Armstrong was taking giant steps there 30 years ago, we are 114cm more distant. Standing at the back door looking up at the Big Cheese, **down my soul befriends** from this knowledge? No. Slurs **DNX** it.

There are forests of facts which can not be so easily shed, as with some of the far-from-trivial statistics that jolt in our numbers spot. A recent example is that 20 million people in the world live in some form of slavery. That tends to be stored and remembered. As does, say, a recent finding by the London School of Economics that four million British children live in poverty, relative though that is. This knowledge can't be tossed if we are to remain half-human, but precisely because there is such a wealth of facts that the importance of big numbers like these dilutes into a milky way of mental numbers.

It's not just numbers, it's also words. Don't Need-to-Know businessmen can, for example, nap any report to do with rats. Not the sweeter sort but the laboratory kind, the ones which almost daily lead to variants of "Tweaking gene 856G and injecting rhobotinidomazine greatly reduces their loss in male rats over 50. Scientists at Brillolab caution that many years of research will be required before..."

Sport offers superb **DNX** material. You can whitenoise all that coaches say before a match and most of what they say after it, along with what the players matter about why they lost/lost to the event, stupid! If you've watched Lance Armstrong through such stage of the Tour de France, do you really need to learn it wasn't **DNX** on his leg?

Politics is the promised land of sex. Meaning you blank all promises and pre-election bumps, all state-of-the-nation monologues. One needs to take in only how the votes fall, all details (on the where-there's-smoke basis), and party funding figures. Where politics bleeds into economics, there is also **DNX** on a computer. **DNX** also seems to burn a hole in your stockpiling and merger talk because it's about as accurate as that, and anyway you never hear it in time to benefit from it.

Criticism can be a joy, but keep the **DNX** button ready for those experts who can't resist giving the whole plot away, explaining what the painting really means, or who majored in semantics. It's easy to build such personalized **DNX** lists. My own list one also includes all advertisements (especially those offering "personalized" things), and anything to do with any marriage, anywhere. Also flamed out are all numbers about ring saucers, mobile phone sales, desirable cholesterol levels, winners of the U.S. Masters, and guestimates of how much Bill Gates has in the bank, or how it might stack up.

The **DNX** club is free and there are no membership numbers. Our catholic motto might sound familiar—if you haven't set raised out all slogans. It's No **DNX**, no **DNX**.

The interface also features a table at the bottom with the following data:

Page	Image source	Scan time	OCR time	Resolution	Width	Height	Lineskew	Rotation
<input checked="" type="checkbox"/>	C:\Program ...	7.92	-	300	2074	2602	-	-

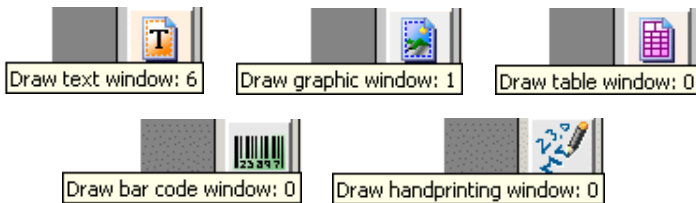
Page decomposition uses three **window types**: text, graphic and table windows. Readiris discriminates text blocks, tables and graphic zones containing

photos, illustrations etc. on the page. (Saving graphics and recognizing tables will be discussed at great length below.)

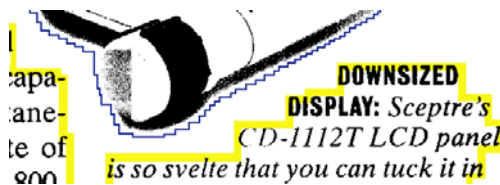
Two extra zone types are always drawn manually: bar code zones and handprinting zones. (More about bar code reading and the recognition of hand-written “block letters” later.)

A **color code** indicates the window type: text zones have an orange border, graphic windows have a purple border and tables a pink border. Bar code zones are green and handprinted zones are blue.

The number of windows is indicated at all times in the tooltips of the window tools.



Page analysis 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

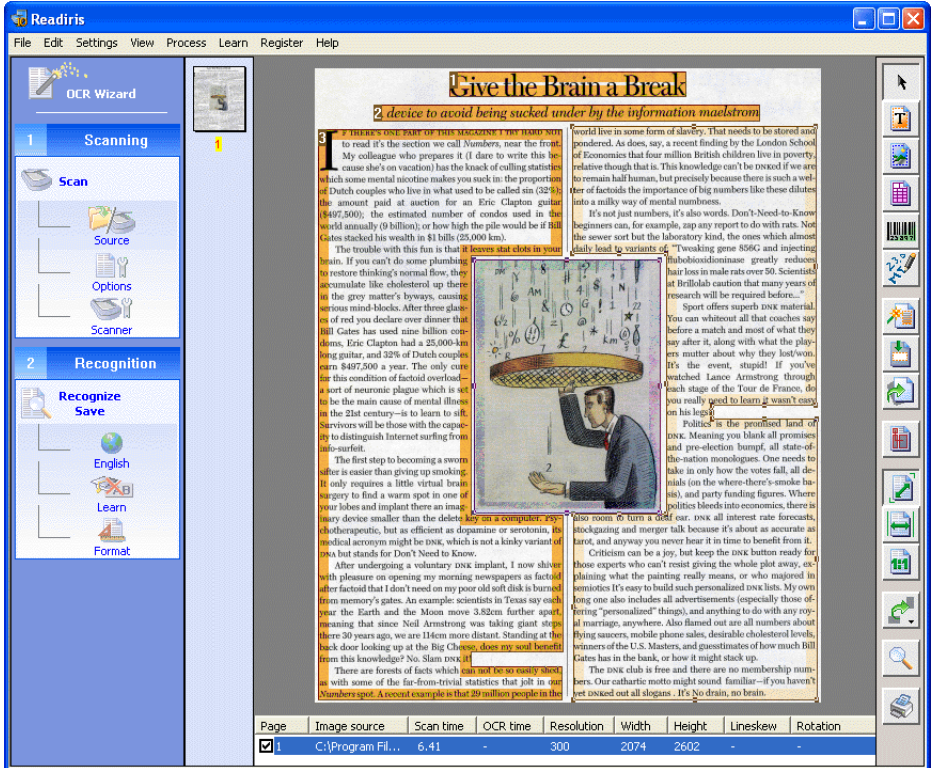
---

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 zones are selected and which aren't: the selected windows have their full color, non-selected windows have a lighter color tone and have no number.



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 and Hebrew 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. In Hebrew documents, the text runs 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.



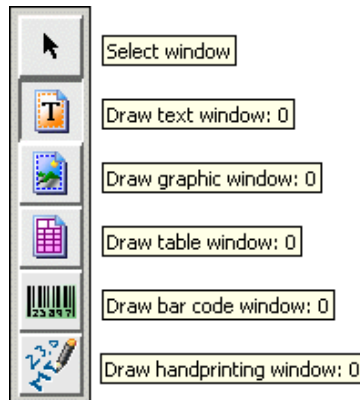
December 12, 1999  
Dear Friends,  
Twenty-five years ago this weekend, thousands of people showed up to the polls in my hometown. It was a special election, called by the local Board of Education.  
Imagine a whole town coming together to decide your fate like that. Of course, those weren't the exact words they were. (Should Michael Moore be removed from the Board of Education?), but that was pretty much how it felt. As I sat there all day in the new polling place (a school gymnasium), watching the people stream in to cast their ballot, I wondered why I had given up going away to college, just so I could be and by no change votes things in the school system. I sat there in those bleachers overlooking the voting booths and figured I was just nuts.  
Two years prior to this "judgment day," I had become the youngest elected official in the country -- and the first 18-year old ever elected to public office. A constitutional amendment had just passed, lowering the voting age from 21 to 18. It was Vietnam's me, and I guess the adults figured if we could go get drafted (one for others for no damn reason), then they figured they should let us vote and drink. Some trade-off.  
So, as I mentioned to you last year in one of those letters, I was a slightly school off teenager (this was before the era of school shootings and rituals). I decided the best way to fight the high school principal was to become his boss -- well, well, well a student. So, I ran for the school board on the platform of fixing him. Remarkably, I won. Nine months later, the principal was gone.

Another similar routine is automatic: the detection of zones on the page borders. When this routine is disabled, the page analysis ignores any zones that touch the page borders. When your scanner generates black borders around the actual image, page analysis tends to find zones where there’s only “noise”. Graphic zones on the page borders are left untouched: photos often touch the page borders, background graphics in most cases cover the entire page etc.



## 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.



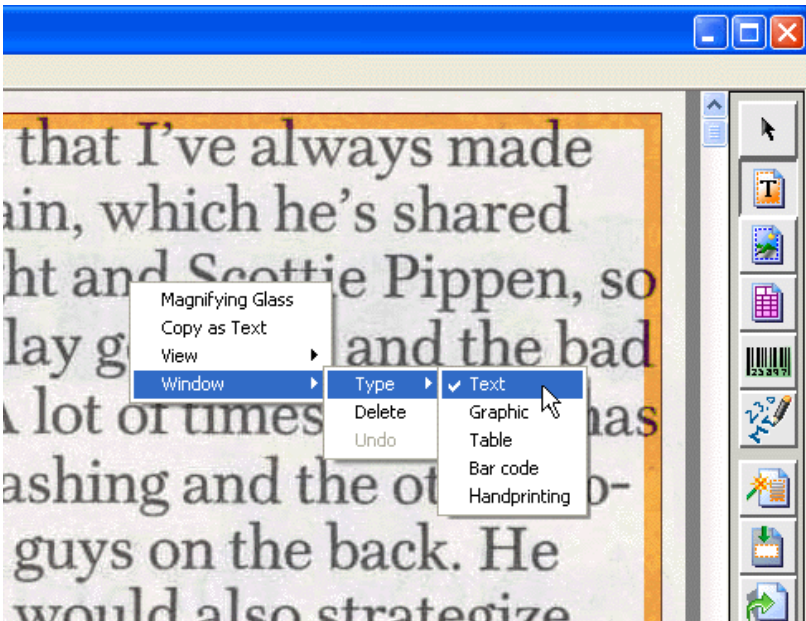
(As indicated earlier, bar code and handprinting windows are always drawn manually by the user: the page analysis does not detect them for you!)

To **draw** a rectangle around a zone of interest, select the corresponding tool in the image toolbar and drag the cursor from the upper left corner to the lower



right corner of the window. (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.

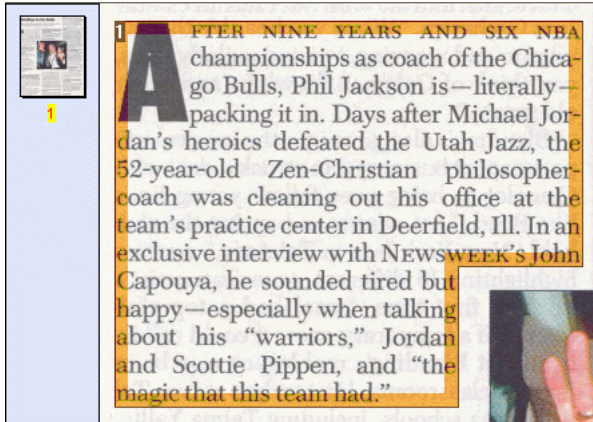


The windows are automatically sorted in the order of creation: numbers indicate the sort order.

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 and bar code 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.

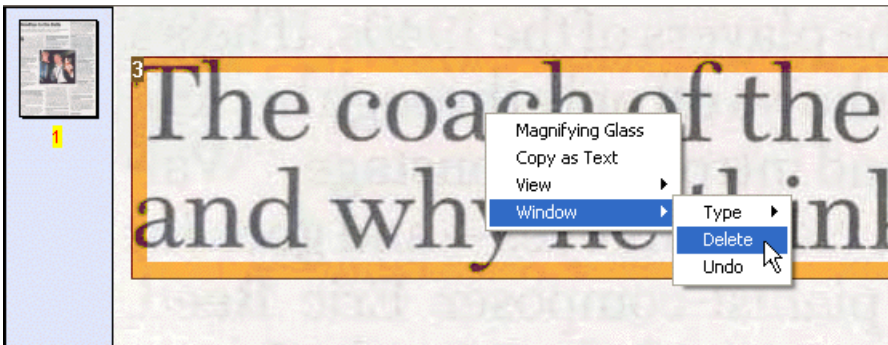


To select a window and the **included windows** (of another type), hold down the Ctrl key while clicking on the main window.

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.

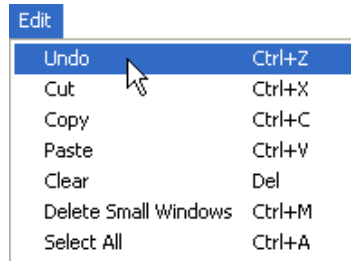
To **move** a window, simply select it and drag it to another location.

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.

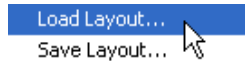
In fact, *all* familiar commands from the "Edit" menu apply to the windows: you can delete, cut, copy and paste them! The "Undo" command also applies: if you have unfortunately deleted, moved, resized etc. some windows, "Undo" will cancel the last operation.



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.

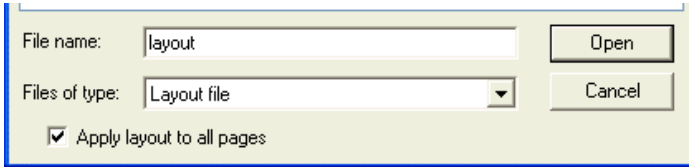
### THREE, SAVING WINDOWING TEMPLATES

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.

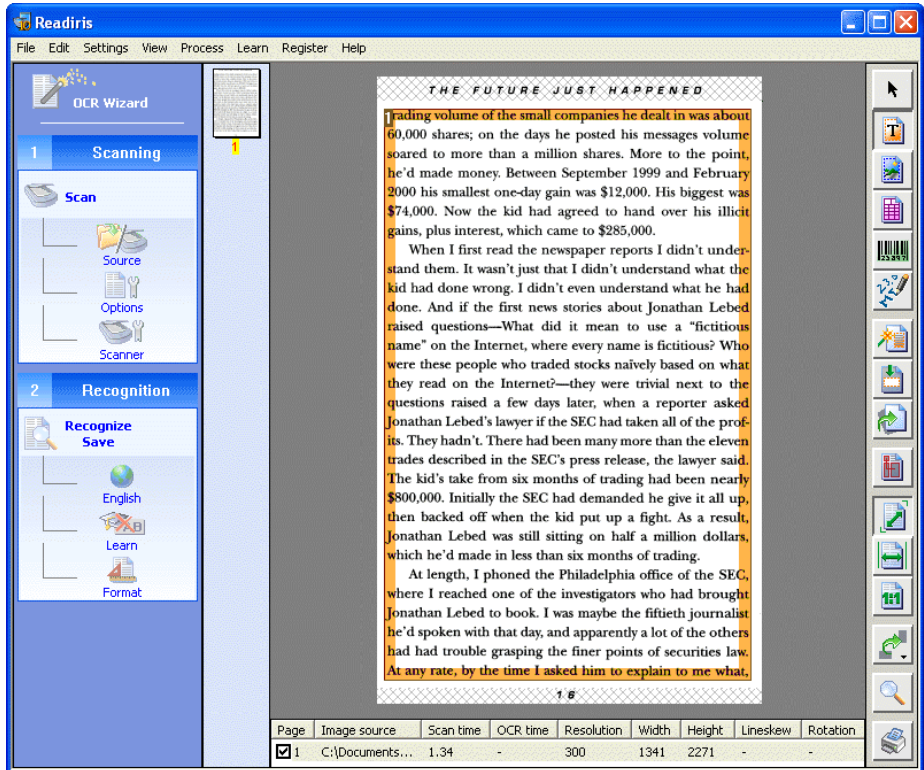
When you load a template into memory, page analysis is disabled automatically. The zoning template remains active until you re-enable page analysis on the main toolbar. When you're loading a layout into memory, you can enable the option "Apply Layout to All Pages" to apply it promptly to all pages of the current document.



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

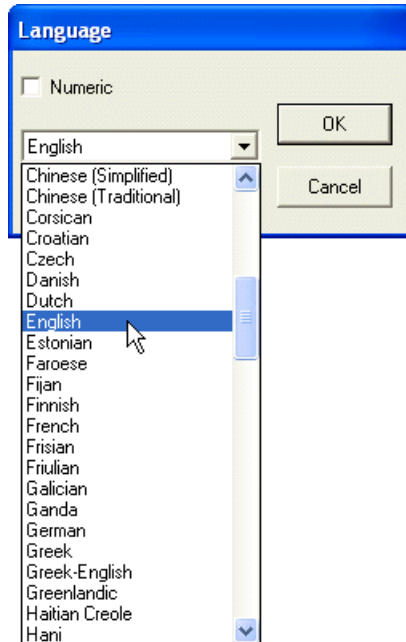
Assuming that the windows are correctly defined, you are now almost ready to execute the character recognition. We say “almost”, because we haven’t verified the language and document settings yet.



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.)

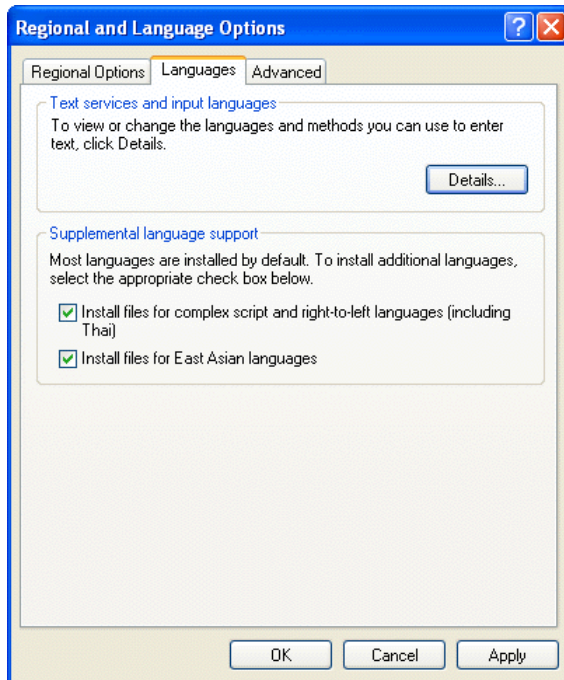
Readiris is far from limited to English: up to 117 **languages** are supported! All American and European languages are supported, including the Central-European languages, Greek, Turkish, the Cyrillic ("Russian") and the Baltic languages.

Optionally, you can read **Hebrew** and **Asian documents**: the extra module "Hebrew OCR add-on" predictably offers recognition of Hebrew documents, the software option "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. The same goes for Spanish and Mexican.

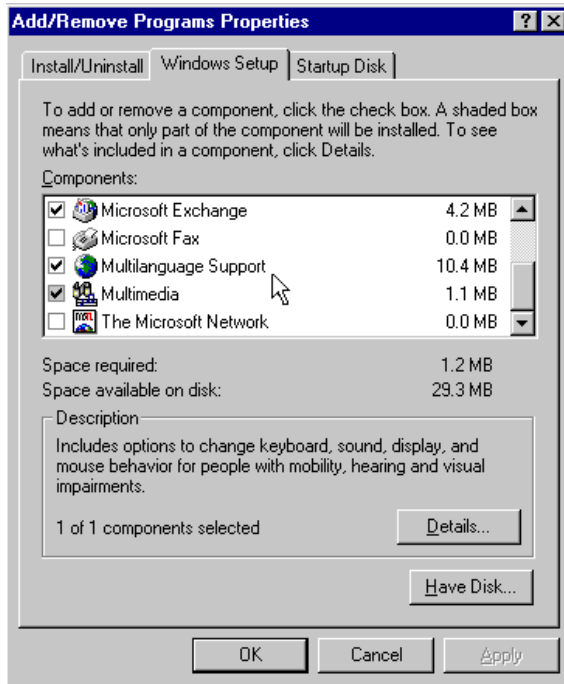
It takes the appropriate Windows configuration to display Central-European, Greek, Turkish, Cyrillic and Baltic characters. You may have to install the **Windows multilanguage support** before your Windows system is able to cope with these languages.

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 and Hebrew documents, you can install an Asian and Hebrew version of the Windows operating system or 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-lin-



guistic 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.

193 1950s. 1hr  
Well, Rossellini

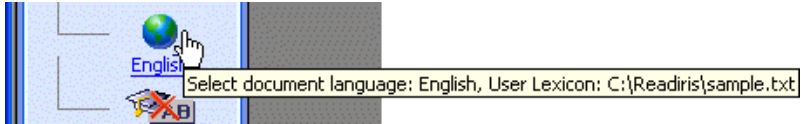
## USER LEXICONS TO “BOOST” THE LINGUISTICS

---

You can take the linguistic “feedback” one step further by customizing it: as powerful as the standard lexicons may be, users of Readiris Corporate can “boost” the **OCR accuracy** further by loading user lexicons with the command "User Lexicon" under the "Settings" menu.

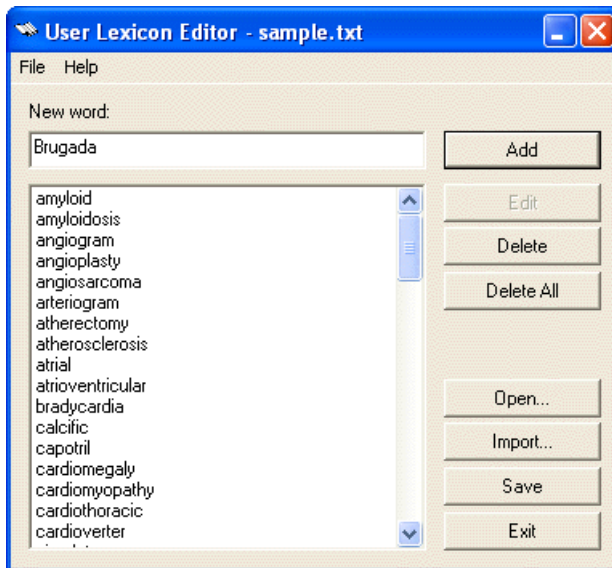


The tooltip of the "Language" button indicates which user lexicon is currently active.



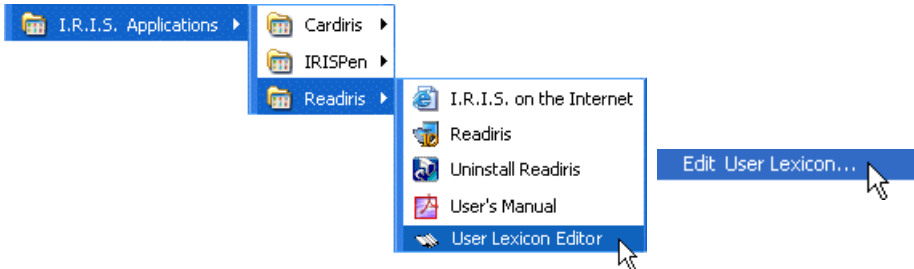
User lexicons are word lists containing any term that does not occur in the “basic”, general purpose lexicons. Think for instance of technical, scientific, legal or other **company-specific terms**!

Readiris is accompanied by the User Lexicon Editor, a utility designed to create and maintain such **user lexicons**. This tool is very user-friendly; consult its on-line help should you have any doubts about its operation.





You'll find this editor in the submenu "I.R.I.S. Applications - Readiris" and under the "Settings" menu of Readiris.



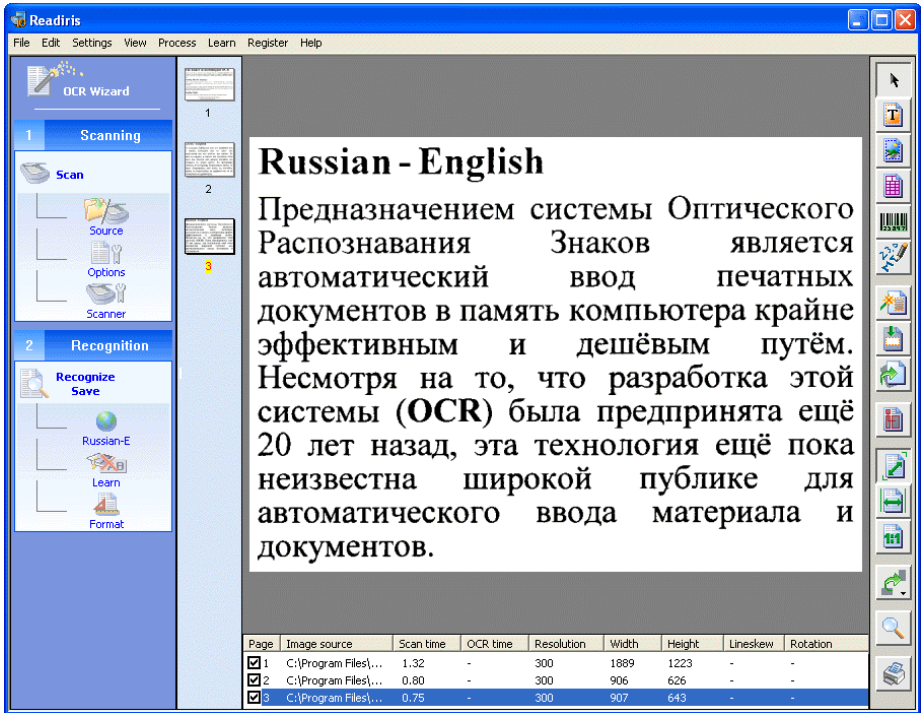
## READIRIS CHANGES LANGUAGES AS NEEDED

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, Hebrew or Asian documents - many untranscribable 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, Hebrew or Asian and Western characters.

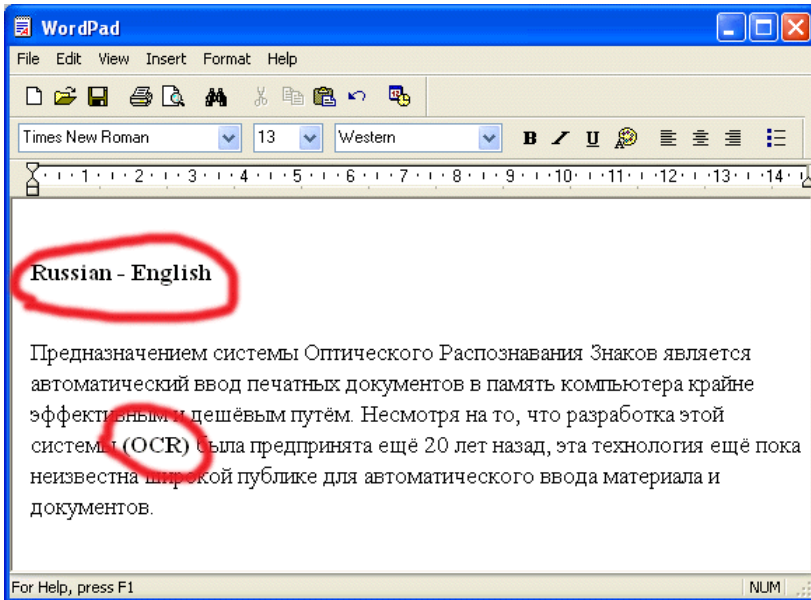
Be sure to select "Greek-English" or the appropriate Cyrillic language setting - for instance "Byelorussian-English". In other words: don't try to just select "Greek" or "Byelorussian" as document language and hope that the Western symbols will come out fine!



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



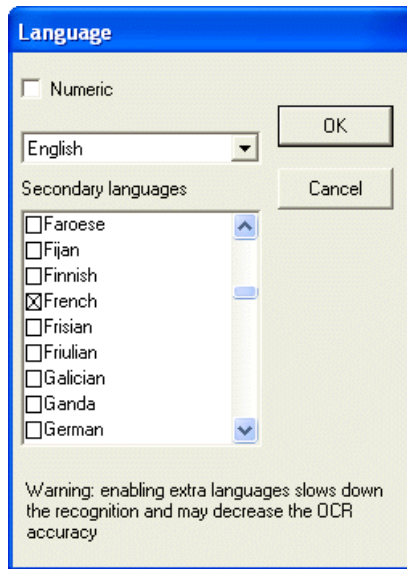
The end result looks like this when opened with the wordprocessor - you may have to select a Cyrillic **font** to display the Russian text correctly.



## READING DOCUMENTS WITH MIXED LANGUAGES

---

Readiris Corporate is more powerful when it comes to reading documents that cover several languages: with that version, you can select a primary language and (up to 4) **secondary languages** (of the same language group).

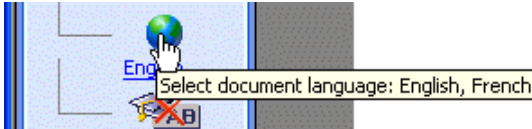


To read a document where the French translation occurs alongside an English text, you'd select English as main language and French as secondary language or vice versa. Mind you, you don't just extend the character set to include the accentuated characters such as ç, é and ù. Both linguistic databases provide linguistic “feedback” to the OCR process, and this allows Readiris to “auto-detect” which language is used where in your document!

You can select up to 5 languages simultaneously. Only languages that belong to the same language group can be combined: languages that are covered by another Windows code page than the “main” language can *not* be activated simultaneously. But first and foremost, don't select languages that don't apply: the bigger the character set, the slower the recognition and the bigger the risk for OCR errors!



Should you need to learn quickly which languages are currently selected, the tooltip of the "Language" button enumerates them...



## 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.

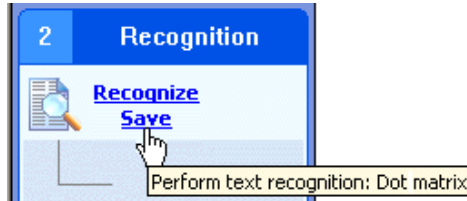
The setting "Automatic" means that Readiris will detect the font mode automatically. Let Readiris "auto-detect" the font mode in all cases - unless you are sure only dot matrix documents are being read! (Obviously, "Automatic" is the default value.)



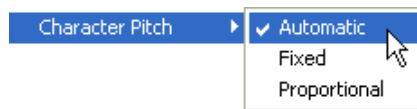
The font type is indicated in the tooltip of the "Recognize-Save" button: when no message is added to the tooltip, the "auto-detection" of the printing quality



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 “i” or “j”. Virtually all books, magazines and newspapers are printed in proportional pitch.

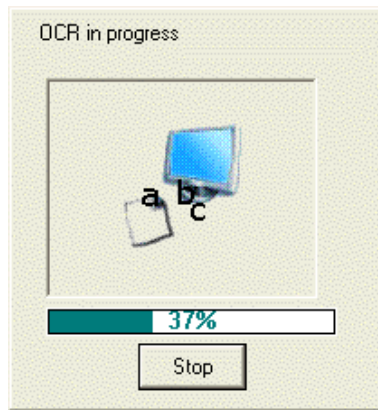
The simplest solution is to leave this option at all times on the default value "Automatic", which means that Readiris will detect the character pitch automatically.

## **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-Save" 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.

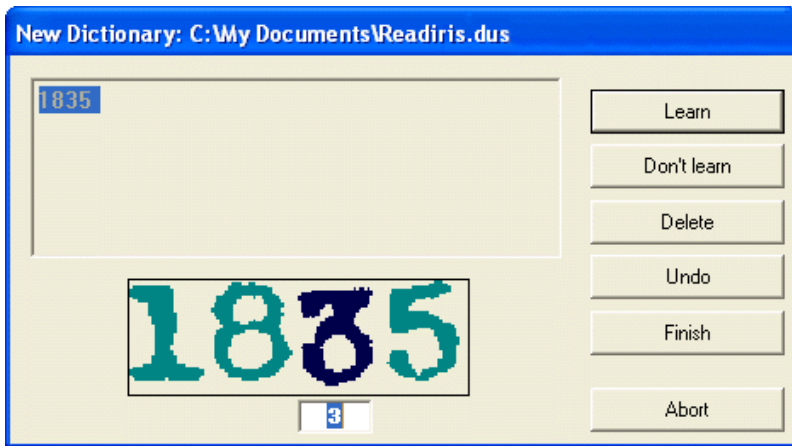
(Interactive learning does not apply to Asian documents: learning does not make sense for these languages which use thousands of different symbols - and you'd have to be able to enter the ideograms, not an easy task when using a Western keyboard!)

**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 " $\pi$ " 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 doubtful 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

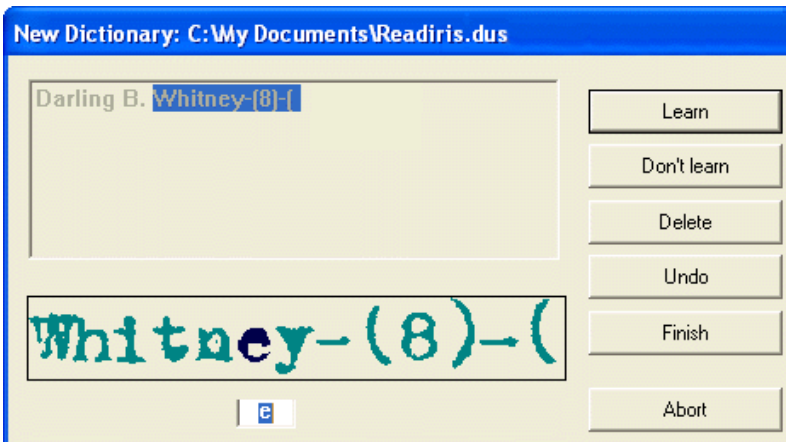
You agree with the proposed solution or correct it. The program saves this doubtful character in the font dictionary as “sure”, final. Future recognition will no longer require your intervention, the shape is considered learnt once and for all.

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 detailed 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**

---

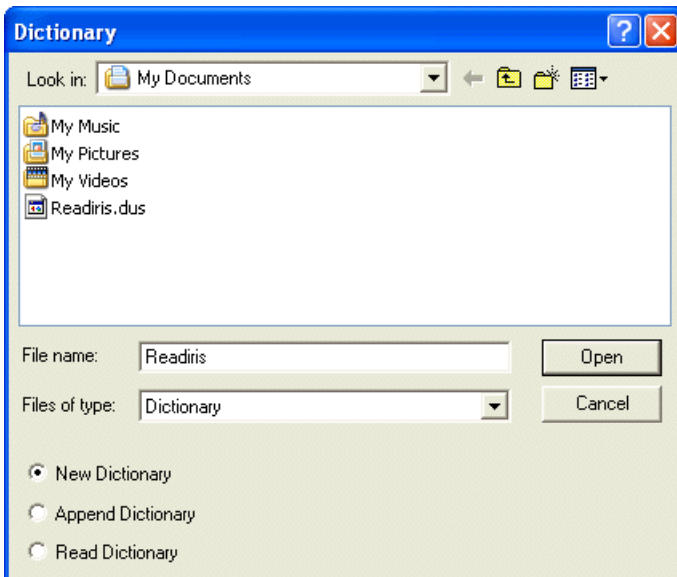
The results of each training session are temporarily held in the computer's memory but can and should be stored in files called "dictionaries" for future use.

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



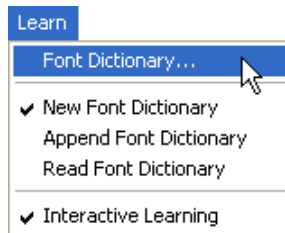
The font dictionaries should be loaded into memory when you want to recognize similar documents in order to make use of the extra intelligence they contain; in this way, Readiris takes into account the intelligence stored in these font libraries. You could say that Readiris gets more intelligence each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

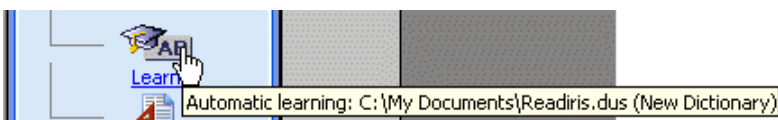
The append mode indicates that the training results will be saved in an *existing* dictionary: the recognition makes use of the extra intelligence already contained in the dictionary, and you add new font shapes to it. In simple terms, this option allows you to build up a font dictionary in several steps.

(When you enter a file name 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.

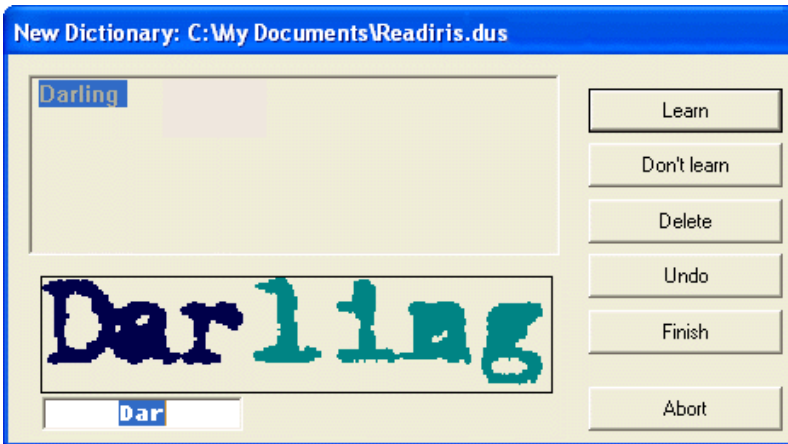
Select the new mode when a single page is recognized. To recognize many pages of the same type - pages with the same fonts and printing quality - select the new mode for the first page, the append mode for a few pages more and the read mode for the rest of the document(s).

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.

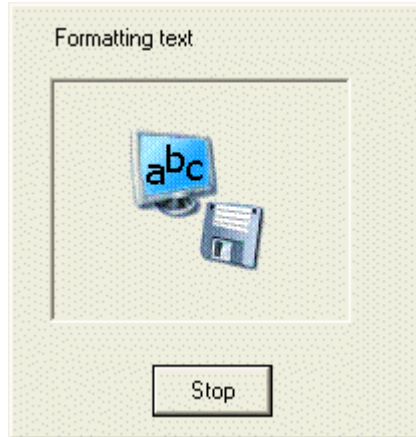


## SENDING RESULTS DIRECTLY TO YOUR APPLICATION

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.)



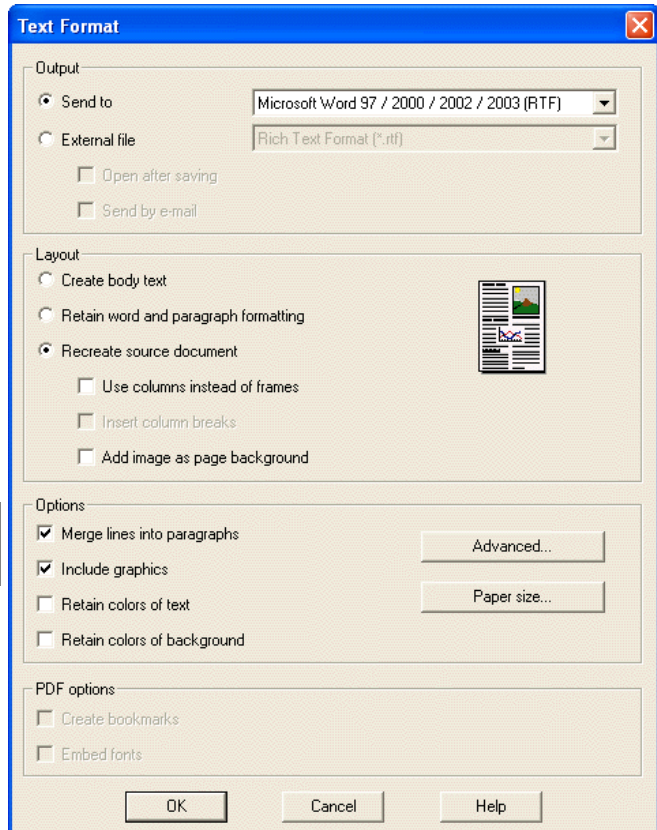


The scanned image is displayed again with the zoning as created to be available for further processing, it stays there until you scan another page.

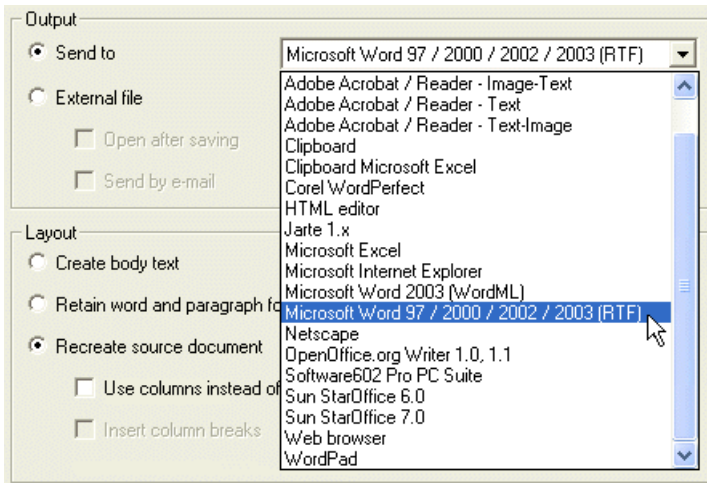
You have indeed converted a paper document into an editable computer file, be it 40 times faster than manual retyping! Go ahead and compare it with the image you have inside your Readiris window.

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.

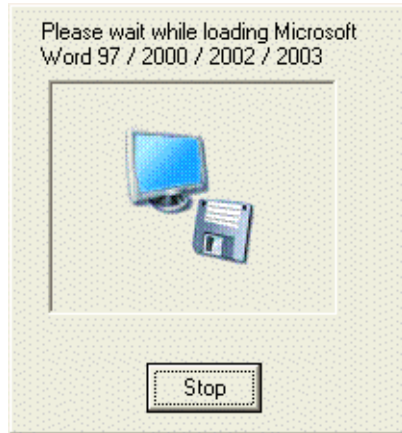
The **output target** is selected using the "Format" button on the main toolbar (or the command "Text Format" under the "Settings" menu).



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.

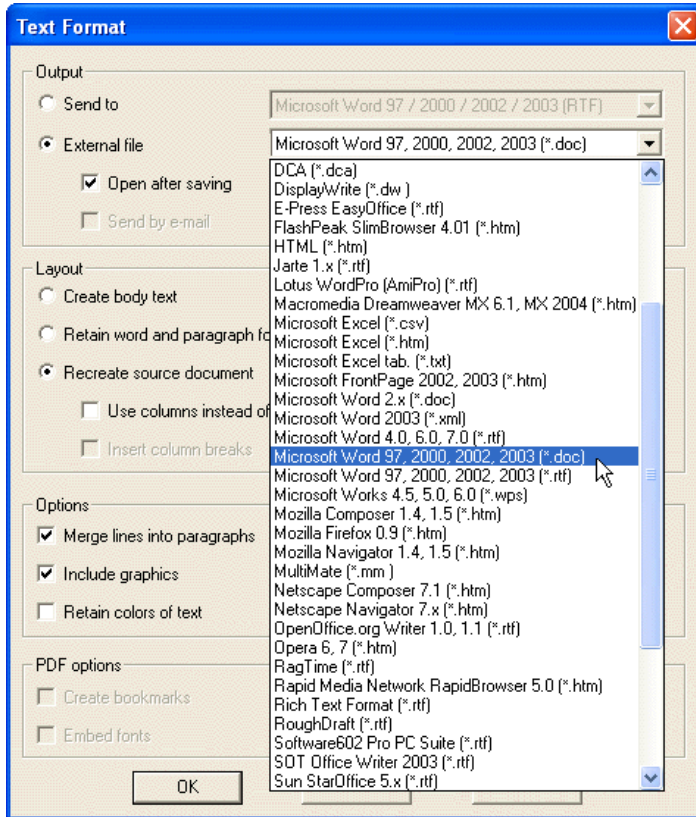


Don't forget that the option "Send to" also allows you to copy the recognized text to the Windows **clipboard**, so there is no strict need to export the result... or save it to an external file!

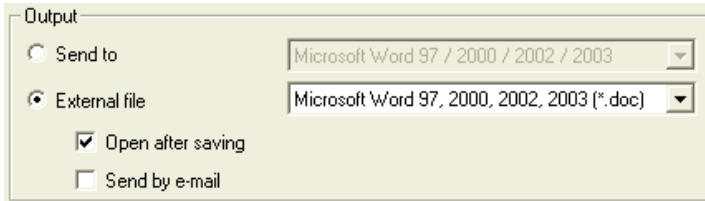
## **SAVING THE RESULTS IN A TEXT FILE**

---

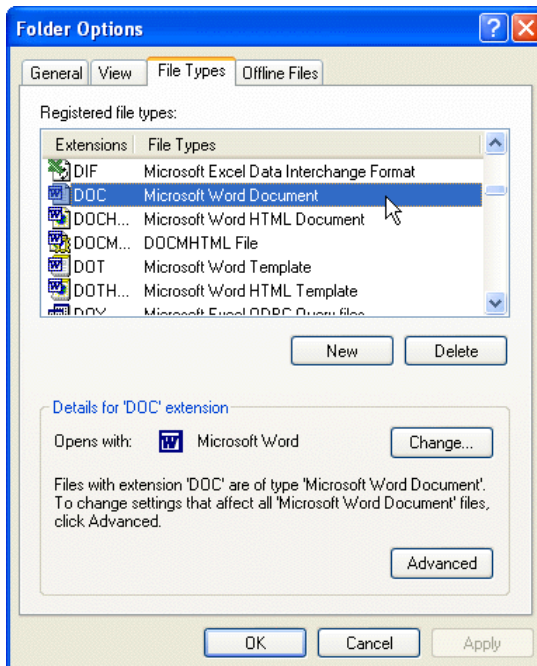
You can indeed write the OCR result to an "external" file. Here again, Readiris supports a wide range of file formats incorporating all popular wordprocessors, spreadsheets, web applications etc. (Amongst others, Readiris supports WordML, the new text format of Microsoft Office 2003!)



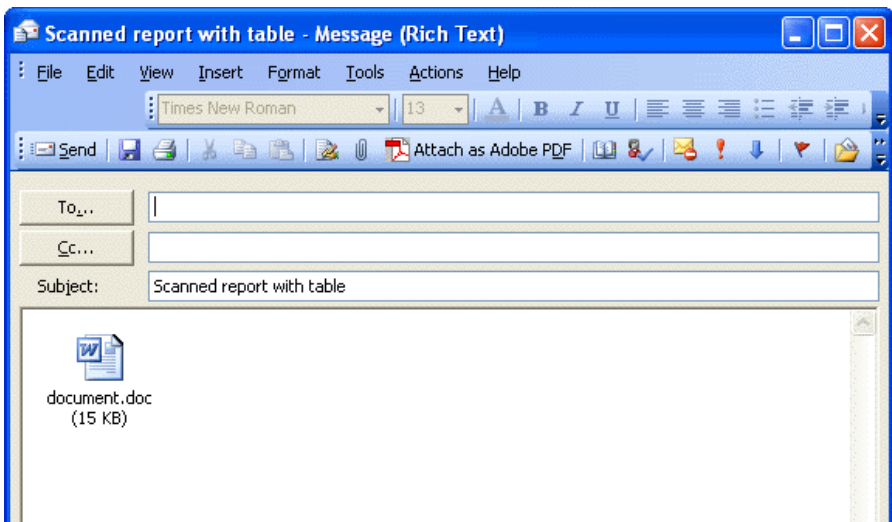
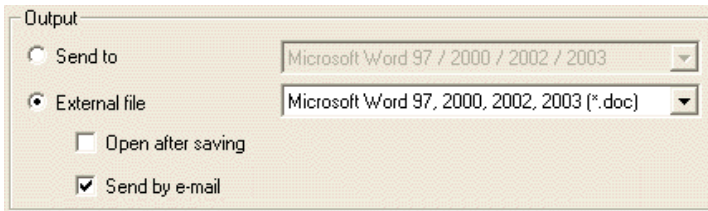
The option "Open after Saving" is largely similar to the "send" feature: you open the recognized document once it's saved.



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.)

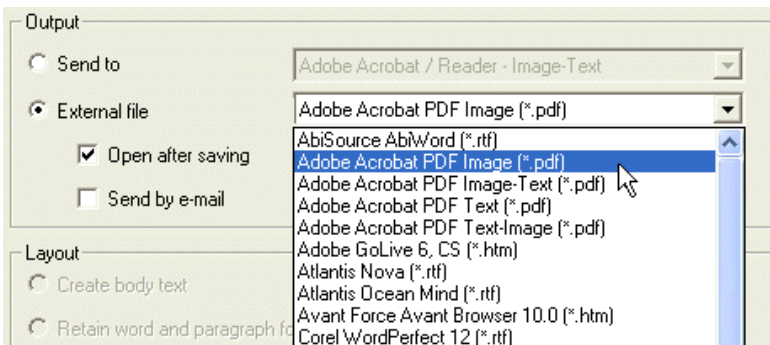
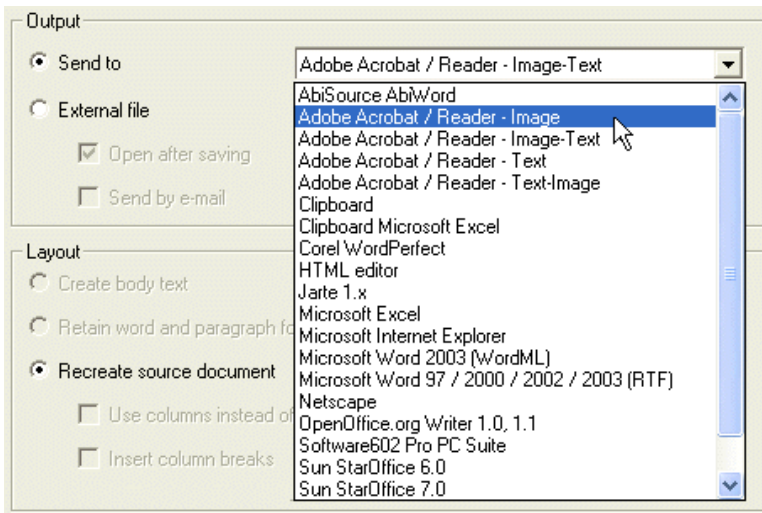


The option "Send by E-mail" creates a new **mail** message and inserts the recognized document as mail attachment. Do you know a faster way of distributing a paper document quickly...?



## CREATING PORTABLE DOCUMENTS...

We'll go deeper into one format: **Adobe Acrobat PDF**. Readiris allows you to create text- and image-based PDF documents.



What's the difference between these formats? 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, art-



work etc.) In other words: the page image is *not* contained in the single-layered PDF file! The format "PDF Image" is also single-layered, but it obviously only contains the scanned image, no OCR results!



Adobe Reader - [autoform.pdf]

File Edit View Document Tools Window Help

Open Save Print Mail Select Text

46%

eBooks

# Autoformatting

The aim of "autoformatting" is to recreate a facsimile copy of the original document.

The OCR process does more than just recognize your text, it can format it for you too!

In a way, text recognition is becoming more and more page recognition or document recognition ...

Whether your OCR software reformats the recognized text or not is up to the user. You can perform OCR because you just need the text, in which case you will edit and format it yourself, and you can ~~format the source document~~, including its formatting.

The various levels of formatting are: creating body text, retaining the word and paragraph formatting and creating a facsimile copy.

Creating body text means no formatting is applied: you get a continuous, running text. All formatting, if any, is done afterwards by the user.

If you retain the word and paragraph formatting, the font type, size and typeface are maintained across the recognition. The justification of the paragraphs is also detected. However, no graphics are captured and the columns aren't recreated - the paragraphs just follow each other etc.

"Autoformatting" recreates a facsimile copy of the original document: the text blocks, graphics and tables are recreated in the same place and the word and paragraph formatting are maintained across the recognition.

Cell 1A	inmate
Cell 2A	Warden
Cell 3A	\$100,000

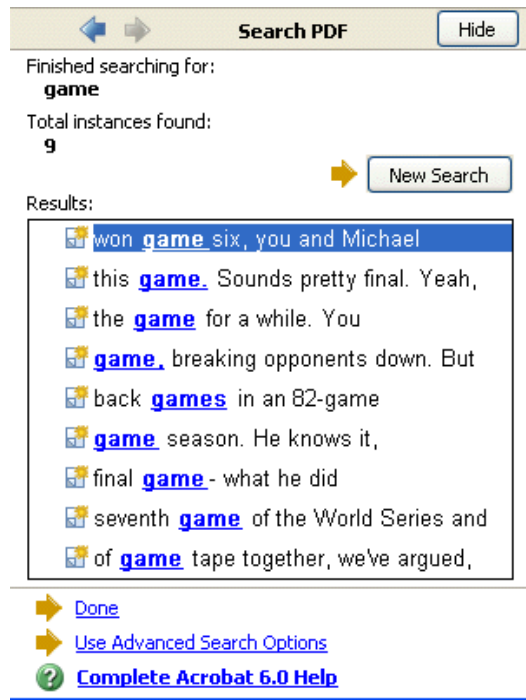
As a result, you get a true copy of your source document, be it a compact and editable text file, no longer a scanned image of your document!

7,03 x 8,71 in

1 of 1

The formats "PDF Text-Image" and "PDF Image-Text" yield different results: Readiris creates a searchable PDF file that contains the recognized text *and* the

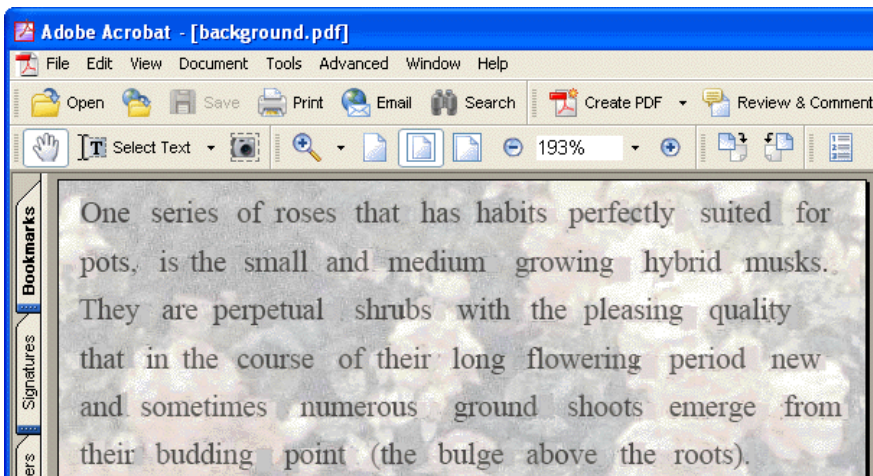
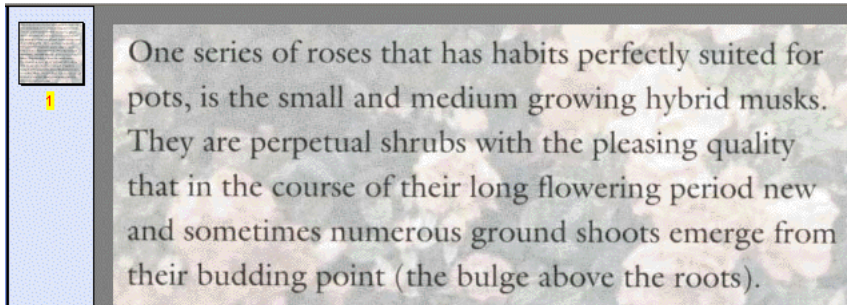
page image. With “text-image” PDF files, the text is placed *above* the page image in the two-layered PDF file; with “image-text” PDF files, the text is contained *under* the page image. Use the "Search" tool of Adobe Reader and this becomes quickly obvious!



PDF files of the type “text-image” are actually pretty sophisticated: the pixels of the recognized text are erased to create a legible document! Displaying recognized text in, say, black on top of black character bitmaps would give you text with a heavy shadow... You can recognize the sample image



BACKGROUND.JPG if you want to give it a try. (Readiris Corporate offers the same functionality for other text formats as well...)



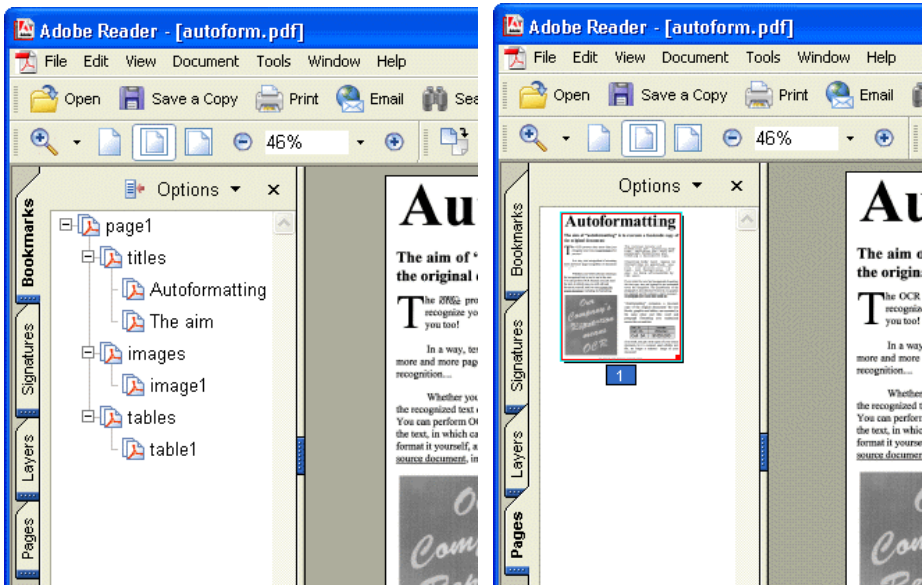
All text-based PDF files encode web site URLs as visible links: click them to visit the mentioned web site!



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

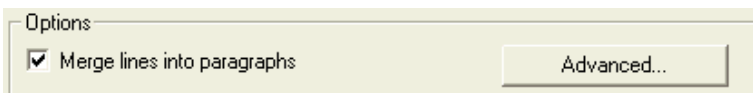


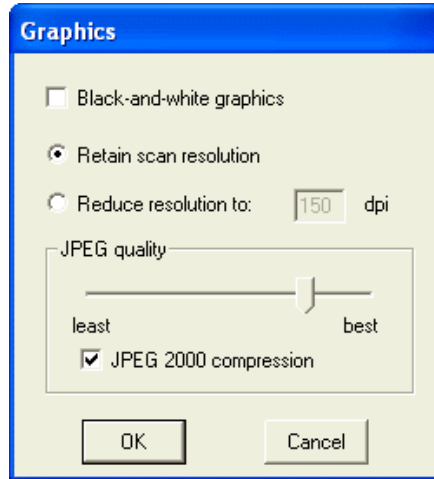
The option "Create Bookmarks" sees to it that a **bookmark** is created for each document element - the graphics as well as the text blocks and tables. For the text zones, Readiris applies an intelligent algorithm to come up with a title, a "summary" per zone; the tables and graphics are simply numbered. (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)!

A further option of PDF files is "hidden" under the advanced graphic options in the "Format" dialog. When you generate PDF files, you can select the compression method for the graphics inside PDF files - JPEG or JPEG 2000. (JPEG 2000 is the newest, more compact version of the JPEG standard.)





## ... 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 reasons 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.



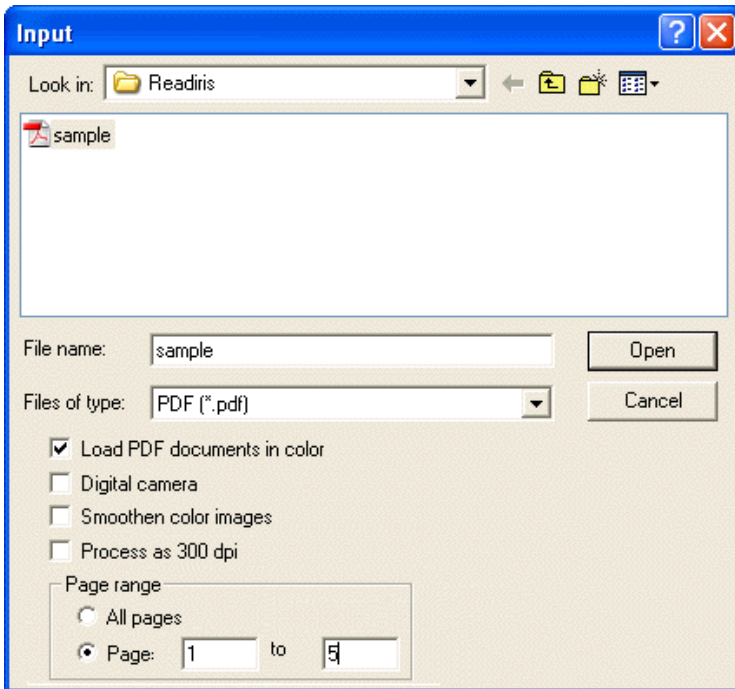
Finally, converting PDF files is a way of “unlocking” PDF content. You can recognize “read-only” PDF documents, where the text is normally inaccessible. With unprotected PDF files, the content can be retrieved (copied and saved to a text file), with “read-only” files, the content cannot be extracted.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

Two important nuances must be noted: Readiris does not open password-protected PDF documents, even if all other PDF security barriers are broken down by Readiris! (To be specific: “master passwords” that set the permissions of PDF documents don’t bother Readiris, “user passwords” required to open a PDF document do.) Secondly, Readiris does not convert PDF documents that contain JPEG 2000 compressed images.

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!





Secondly, you may want to indicate which pages you want to convert. If your objective is, say, to capture just a chapter of a lengthy PDF publication, it doesn't make any sense to load the entire book into Readiris... Indicating the proper **page range** can save you lots of time! (This also holds for multipage TIFF images.)



## RECOGNIZING MULTIPLE PAGES

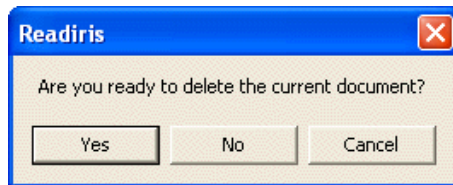
---

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.

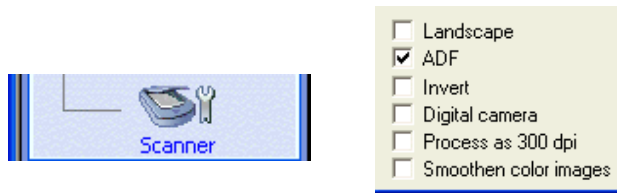


Answer "no" to add pages to the current document, answer "yes" to create a new document. This answer has the same effect as the command "New Document" under the "File" menu.



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

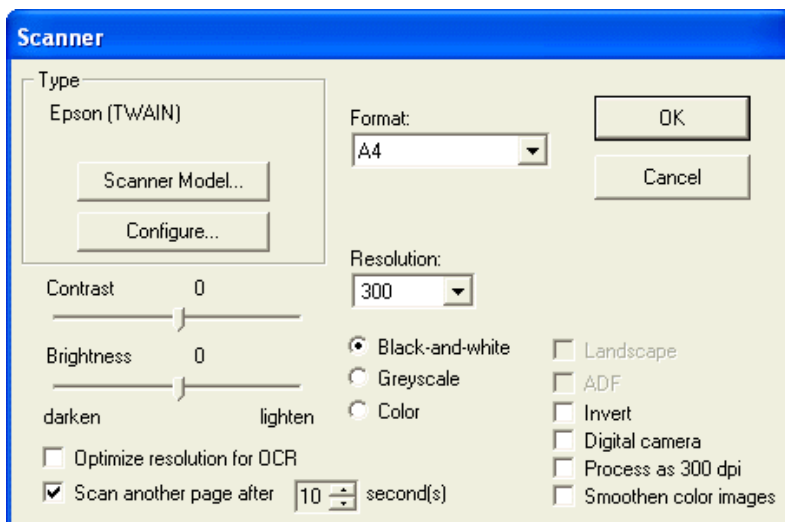
To scan a document composed of several pages in one operation, enable the document feeder of your scanner with the option "ADF" under the "Scanner" button.



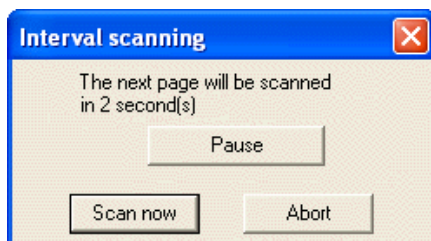
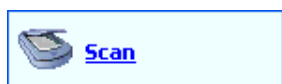
Place the pages of your document in the automatic document feeder and start the scanning: all pages are scanned until the document feeder is empty.

When your flatbed scanner is *not* equipped with a document feeder, **interval scanning** allows you to scan multipage documents efficiently. The scanner automatically scans another page after a user-selected number of seconds; the interval allows you to replace the page you put on your scanner's flatbed.

Indicate the interval you need to place another page on the scanner's flatbed in the scanner settings; click the "Scanner" button and define an appropriate value for the option "Scan Another Page after x Second(s)".

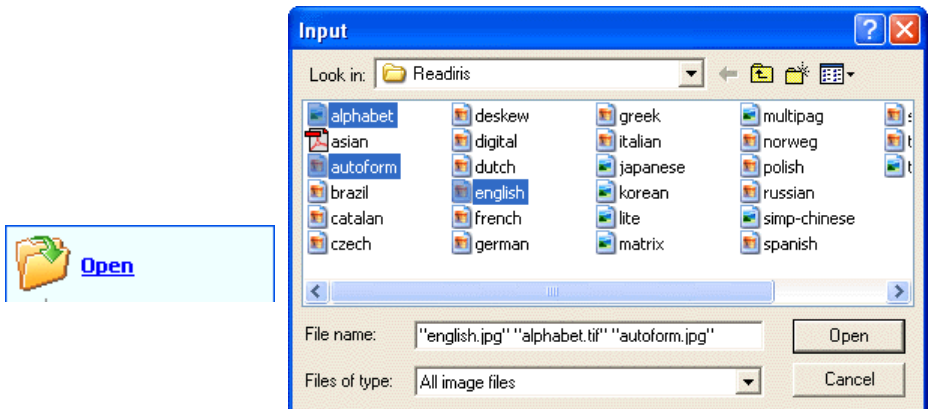


Click the "Scan" button to initiate the scanning. Click "Abort" in the interval scanning dialog to end the automatic scanning.

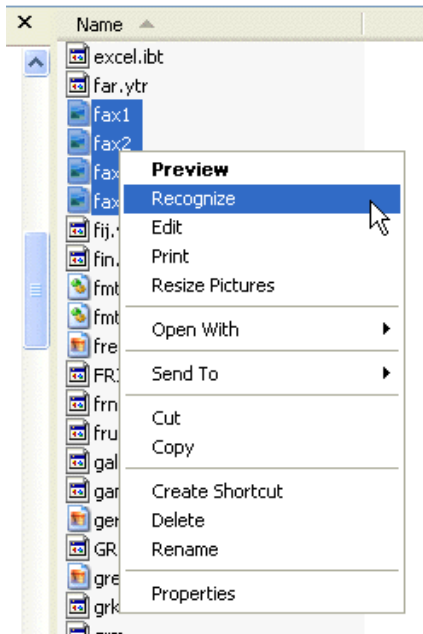


Readiris offers full flexibility: you can bypass the scanning interval to scan instantly and freeze it to take a phone call! Click "Pause" in the automatic scanning dialog to freeze the scanning interval; click "Resume" when you're ready to continue. Or click "Scan Now" in the interval scanning dialog to scan immediately: the interval is cut short!

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.



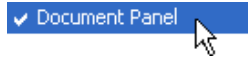
The same effect can be obtained comfortably from within the Windows Explorer: select several image files, right-click and select the command "Recognize" from the "Context" menu. You can repeat this operation: all images you send to Readiris append the current document until you click the command "New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

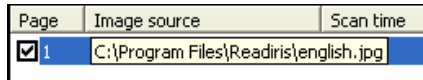
The **document panel** displays **statistics** and information on all scanned pages - the image source and resolution, the scanning and recognition time etc. The document panel can be hidden and displayed with the corresponding option under the "View" menu.



You also learn which image was loaded into memory. If a multipage image was opened, there's obviously just one file for all the images. When you are *scanning*, the document panel simply mentions the scanner model.

Page	Image source	Scan time	OCR time	Resolution	Width	Height	Lineskew	Rotation
<input checked="" type="checkbox"/> 1	C:\Program Files\Readiris\multipage.tif	2.74	-	300	2000	2388	-	-
<input checked="" type="checkbox"/> 2	C:\Program Files\Readiris\multipage.tif	2.08	-	300	2000	1888	-	-
<input checked="" type="checkbox"/> 3	C:\Program Files\Readiris\multipage.tif	2.05	-	300	1912	2004	-	-

Drag the column resizing cursor to change the size of a column. (You cannot change the *order* of the columns.) Or hold your mouse cursor over a column when it is too short to display the data: a tooltip will display the full data!



You can display the same information for all pages with the command "Info" under the "File" menu and you can display that information per page by holding your mouse cursor over a page thumbnail in the **page toolbar** on the left side. This toolbar is displayed as soon as pages get processed. It gives access to the page commands (using the right-click).

Include page yes  
 Image source C:\Program Files\Readiris\english.jpg  
 Scan time 8.52 seconds  
 OCR time -  
 Resolution 300 dpi  
 Width 2245 pixels  
 Height 2841 pixels  
 Lineskew -  
 Rotation -

- Exclude Page
- Select Page
- Delete Page
- Move Page Up
- Move Page Down



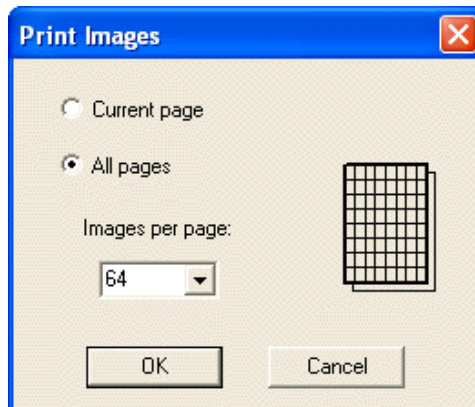
The current page is highlighted in the document panel and page toolbar and mentioned in the Readiris title bar.

To delete a page, select it in the document panel and press the Delete key! (Or select it in the page toolbar, right-click it and select the command "Delete Page" from the "Context" menu.)

You can quickly **print** the scanned **images** with the "Print" button on the image toolbar (or with the command "Print Images" under the "File" menu) 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 document panel (and the corresponding commands in the "Edit" menu and the contextual page commands



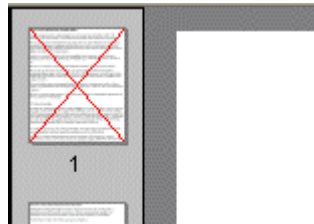
of the page toolbar) allow you to exclude pages (temporarily). Simply click the page number in the document panel to exclude it from the printing (and recognition) process. Click a second time to include it again. For greater flexibility, the "Edit" menu offers equivalent commands that apply to *all* pages.

Page	Image source	Scan time
<input type="checkbox"/> 1	C:\Program Files\Readiris...	0.94
<input checked="" type="checkbox"/> 2	C:\Program Files\Readiris...	0.71
<input checked="" type="checkbox"/> 3	C:\Program Files\Readiris...	0.73

Include Page
Exclude Page
Include All Pages
Exclude 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.



Readiris

File Edit Settings View Process Learn Register Help

OCR Wizard

1 Scanning

Scan

Source

Options

Scanner

2 Recognition

Recognize

Save

English

Learn

Format

1

2

3

5

1 everyone has the right to freedom of thought, conscience and religion; this right includes freedom to change his religion or belief, and freedom, either alone or in community with others and in public or private, to manifest his religion or belief in teaching, practice, worship and observance.

2 everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.

3 everyone has the right to freedom of peaceful assembly and association.

4 no one may be compelled to

5 everyone has the right to take freely chosen representatives.

6 everyone has the right to equ

7 the will of the people shall be expressed in periodic and general suffrage and shall be held by s

Everyone, as a member of soc through national effort and int and resources of each State, of dignity and the free developm

8 everyone has the right to work, to free choice of employment, to just and favourable conditions of work and to protection against unemployment.

9 everyone, without any discrimination, has the right to equal pay for equal work.

10 everyone who works has the right to just and favourable remuneration ensuring for himself and his family an existence worthy of human dignity, and supplemented, if necessary, by other means of social protection.

Everyone has the right to form and to join trade unions for the protection of his interests.

OCR in progress (4/5)

42%

Stop

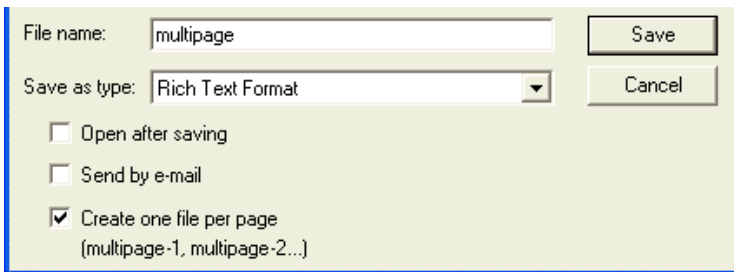
Page	Image source	Scan time	OCR time	Resolution	Width	Height	Lineskew	
<input checked="" type="checkbox"/>	2	C:\Program Files\...	1.39	1.62	300	2000	1888	-
<input checked="" type="checkbox"/>	3	C:\Program Files\...	1.35	2.10	300	1912	2004	-
<input checked="" type="checkbox"/>	4	C:\Program Files\...	1.49	-	300	2004	2000	-

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 File 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. (This option is not available when you send the OCR results to a target application, only when you create an external file.)



File name:

Save as type:

Open after saving

Send by e-mail

Create one file per page  
(multipage-1, multipage-2...)

## EDITING MULTIPAGE DOCUMENTS

---

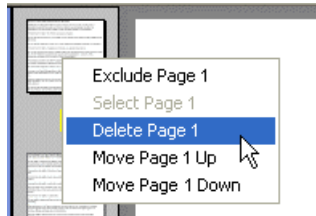
The user can edit multipage documents, mainly to correct scanning errors: he can delete pages from the document and move pages to other locations in the document.

The navigation first. To *go to a page*, click it in the document panel or 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.



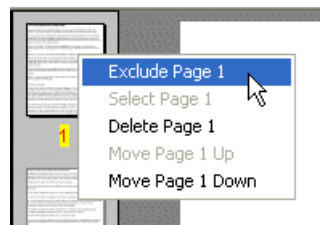
First Page	Home
Previous Page	PageUp
Next Page	PageDown
Last Page	End

Let's edit the document now. To *delete a page*, select it in the document panel and press the Delete key. Or hold your cursor over its thumbnail, right-click it and select the command "Delete Page" from the "Context" menu.



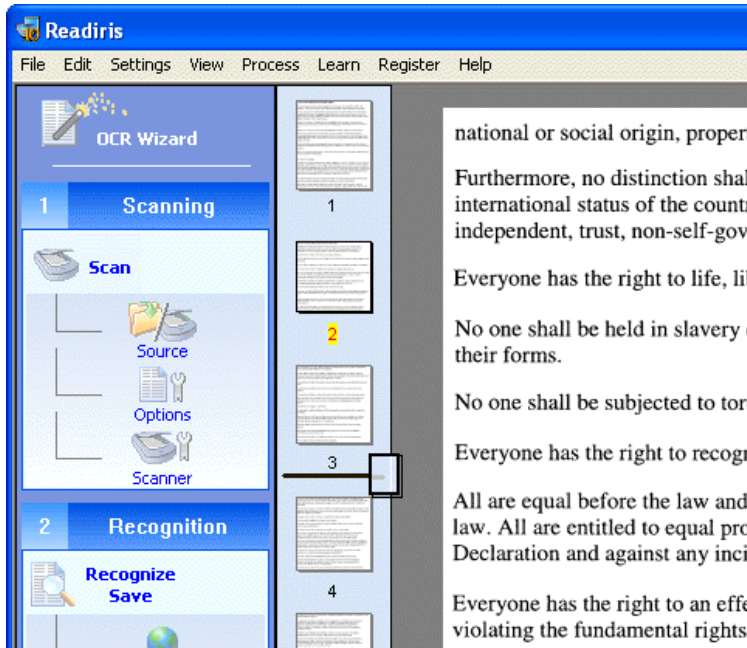
And we remind that you can temporarily exclude pages, not delete them, from the recognition (and image printing) process: the document panel, the page toolbar and the "Edit" menu offer the necessary commands.

Page	Source	Scan time
<input type="checkbox"/> 1	C:\Docume...	5.68
<input checked="" type="checkbox"/> 2	C:\Docume...	9.45



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.

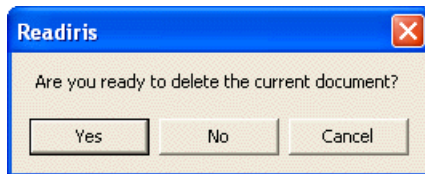


This command "cleans the slate". Any document loaded into memory - containing a single page or multiple pages - is erased. You are now ready to create a new document.



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.

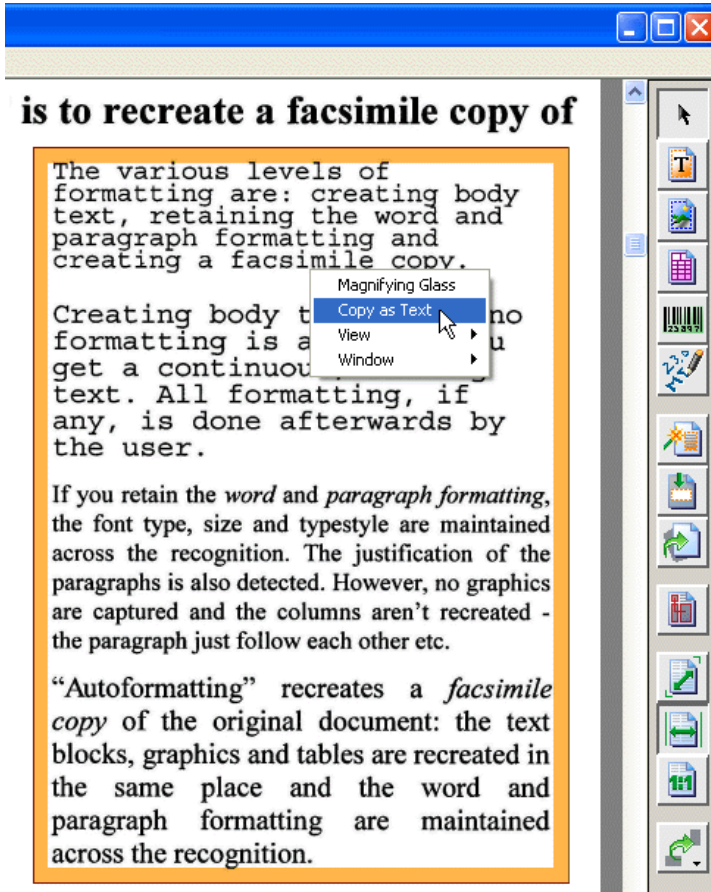
When the OCR *was* already executed and you re-initiate the scanning (or the loading of images), you are prompted to start a new document or complete the current document.



## 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 file name 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

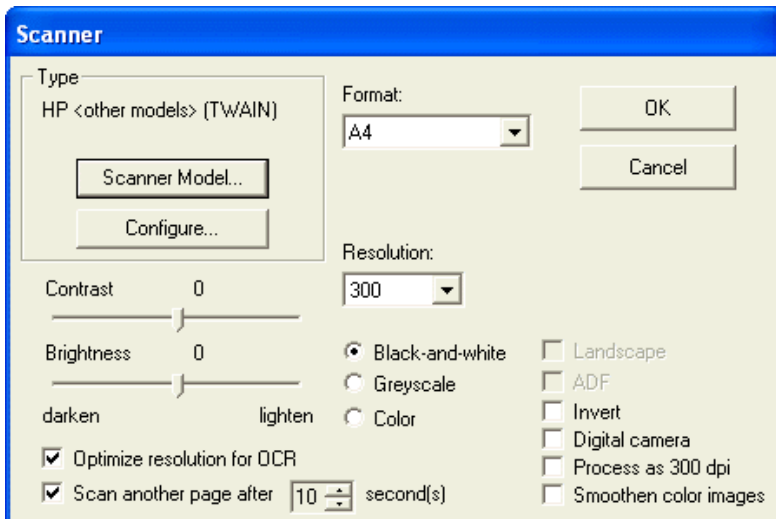
Let's set our scanner up now. It is assumed that the scanner hardware and necessary drivers are installed correctly.

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.



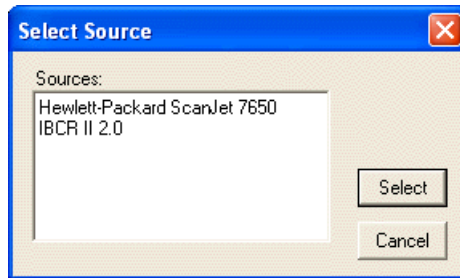
Click the button "Scanner Model" to determine your **scanner model**.





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.

Interval scanning we already discussed. It is highly efficient way of scanning multipage documents when your flatbed scanner is not equipped with a document feeder.

## **LET THE BAD COLOR NOT BE SEEN**

---

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 back-



ground color without losing the text information as well, as much as you may try to adjust the scanner brightness...

**MASAYOSHI SON**, 42, president and CEO, is the master Net empire builder. His conglomerate holds stakes in 300 Internet companies in the U.S., Japan, Europe, and other Asian countries. Today, Softbank manages about \$4 billion in venture capital funds for global investments.

**YASUMITSU SHIGETA**, 35, has invested in more than 70 Web or mobile Net-based ventures in Japan and the U.S., including Tumbleweed Communications and Phone.com. Shigeta is also developing new businesses that take advantage of the growth of the Internet and mobile communications.

**MASAYOSHI SON**, 42, president and CEO, is the master Net empire builder. His conglomerate holds stakes in 300 Internet companies in the U.S., Japan, Europe, and other Asian countries. Today, Softbank manages about \$4 billion in venture capital funds for global investments.

**YASUMITSU SHIGETA**, 35, has invested in more than 70 Web or mobile Net-based ventures in Japan and the U.S., including Tumbleweed Communications and Phone.com. Shigeta is also developing new businesses that take advantage of the growth of the Internet and mobile communications.

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!

**MASAYOSHI SON**, 42, president and CEO, is the master Net empire builder. His conglomerate holds stakes in 300 Internet companies in the U.S., Japan, Europe, and other Asian countries. Today, Softbank manages about \$4 billion in venture capital funds for global investments.

**YASUMITSU SHIGETA**, 35, has invested in more than 70 Web or mobile Net-based ventures in Japan and the U.S., including Tumbleweed Communications and Phone.com. Shigeta is also developing new businesses that take advantage of the growth of the Internet and mobile communications.

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

Display Document in Color Ctrl+O



Actually, you won't see any black-and-white images on your computer screen - even when you're indeed scanning bilevel images! That's because Readiris optimizes the images for an optimal on-screen legibility. I.R.I.S.' specialized high-resolution display technique converts black-and-white images into greyscale images.

### **Reading dot matrix documents**

You can read dot matrix document without changing the font mode. The software detects whether "normal" text or dot matrix printouts are being read.

Far out in the uncharted backwaters of the unfashionable end of the Western Spiral arm of the Galaxy lies a small unregarded yellow sun. Orbiting this at a distance of roughly ninety-two million miles is an utterly insignificant little blue green

### **Reading dot matrix documents**

You can read dot matrix document without changing the font mode. The software detects whether "normal" text or dot matrix printouts are being read.

Far out in the uncharted backwaters of the unfashionable end of the Western Spiral arm of the Galaxy lies a small unregarded yellow sun. Orbiting this at a distance of roughly ninety-two million miles is an utterly insignificant little blue green

Greyscale and color images are softened, smoothed.

### **A word about OCR**

The aim of OCR is to automatically enter printed text documents in a very effective and low cost way. Although the first research and development on Optical Character Recognition (OCR) began more than 30 years ago, this technology is still unknown by most of the people who could use it for their document entry applications.



## A word about OCR

The aim of OCR is to automatically enter printed text documents in a very effective and low cost way. Although the first research and development on Optical Character Recognition (OCR) began more than 30 years ago, this technology is still unknown by most of the people who could use it for their document entry applications.

As a result, there's no need to zoom in, even on laptops with an LCD screen or desktop computers with a low-end 15" screen.

Zoom in at real size (or higher) to see the "raw" image as it was scanned!

## 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).

6 point

# 72 point

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

**R**eadiris reads drop letters (also called "drop" caps) that cover several lines and assigns them to their starting line.

**L**e Festival de Wallonie est dans sa phase ultime et nale avec l'ouverture de ses dernières sections.  
Cette semaine, c'est la branche

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. Amongst others, the image resolution of such images is marked in red in the document panel.

Page	Image source	Scan time	OCR time	Resolution	Width	Height
<input checked="" type="checkbox"/> 1	C:\Documents and...	1,24	-	96	400	447

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.

Optimize resolution for OCR  
 Scan another page after  second(s)

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.

Invert  
 Digital camera  
 Process as 300 dpi  
 Smoothen color images

Load PDF documents in color  
 Digital camera  
 Smoothen color images  
 Process as 300 dpi



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 unknown.

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.

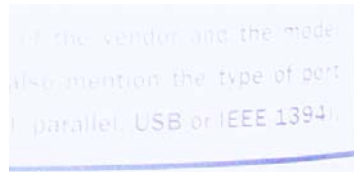




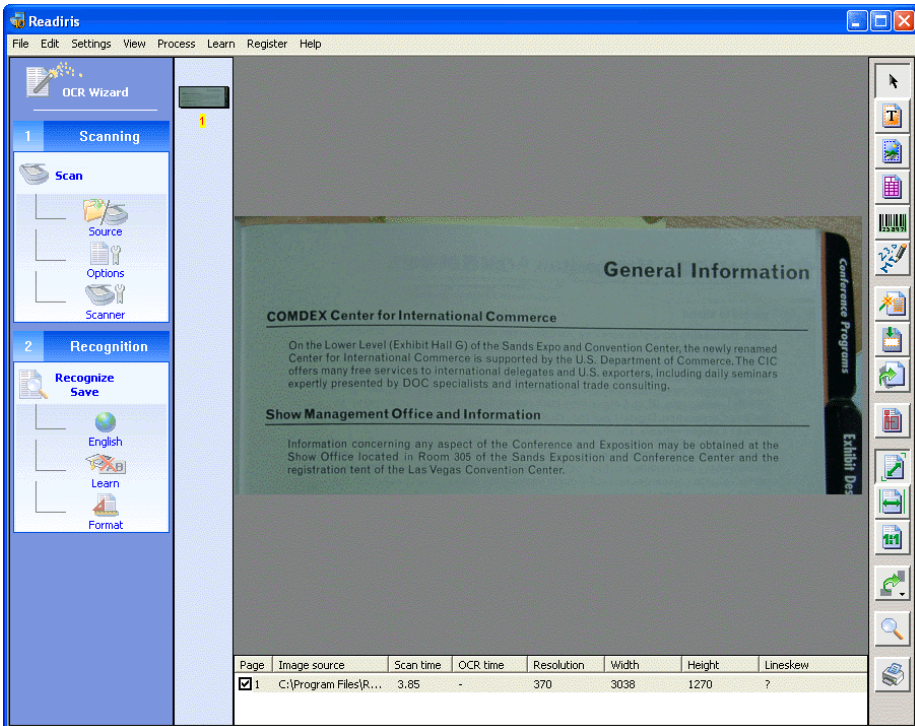
Limit yourself to no or small compression: important compression reduces the sharpness of the captured text. Zoom manually to crop your document - some cameras are bundled with photo stitching software, but don't bother using it for document capture.

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.

Save Default Settings

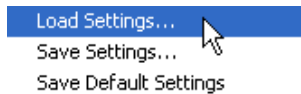


Settings files contain more than the scanner **settings**: they also determine whether you are going to use interactive learning, which language the documents have, which output mode is used - for instance send text to WordPad - etc. In short, *all* operational settings of Readiris are stored in the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

## **SAVING SPECIFIC SETTINGS**

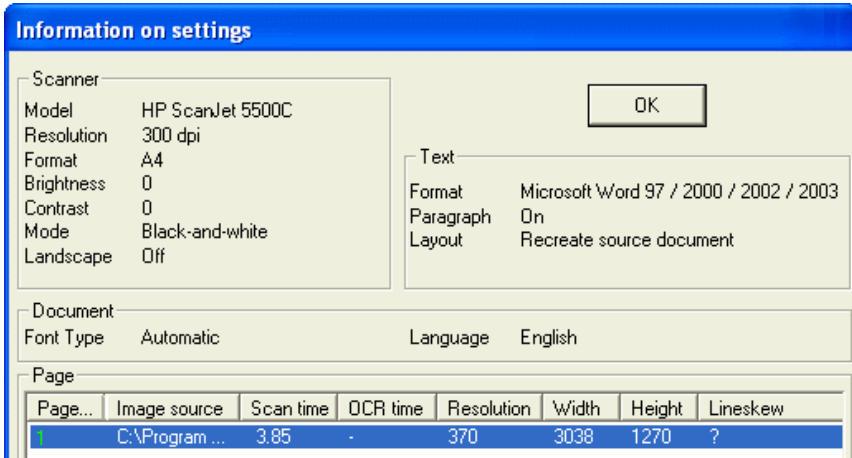
---

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.



This command also displays the information you find on the document panel for all pages.

## 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  $0.5^\circ$ ) 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.

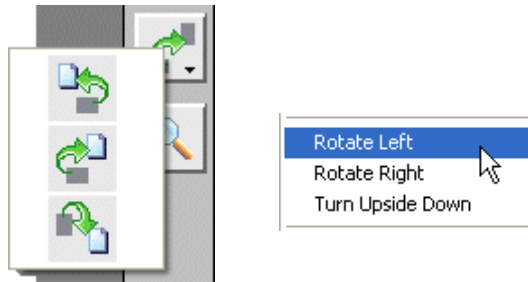


If you forgot to enable this option, use the "Deskew Page" button on the image toolbar (or the command "Deskew Page" under the "Process" menu) to "straighten" pages which were scanned at an angle.

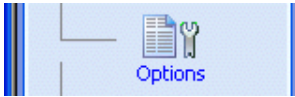


The deskewing takes a few seconds: the image is analyzed to detect the skew angle - if any -, the color or greyscale image *and* its black-and-white version are deskewed and the page analysis gets re-executed.

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.



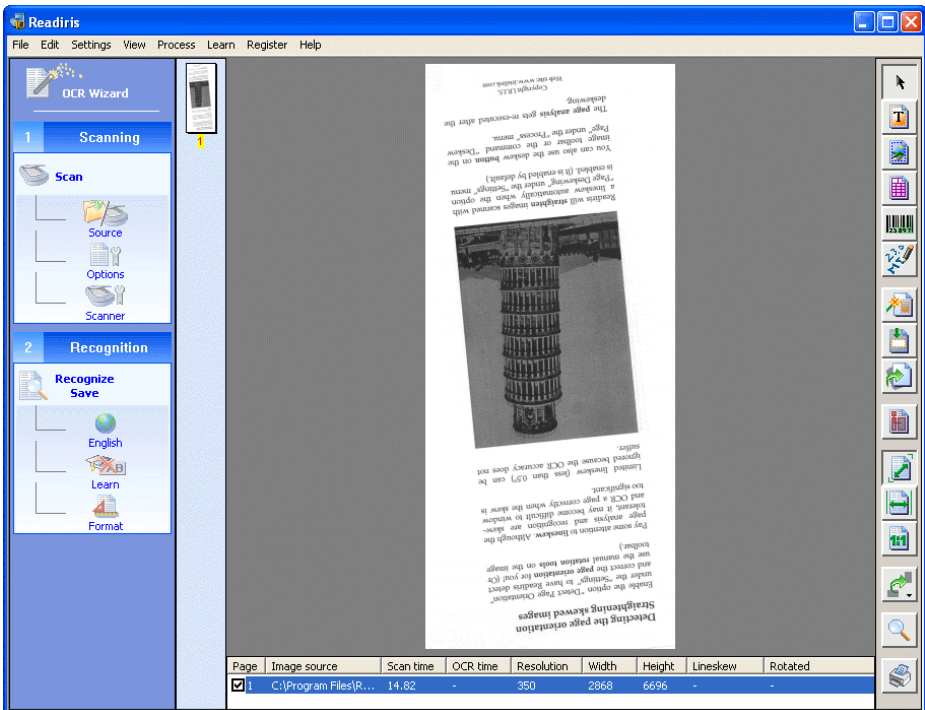
However, Readiris can correct badly oriented pages for you. Enable the option "Detect Page Orientation" under the "Options" button (or under the "Settings" menu) and Readiris will correct the page orientation where needed.



Detect Page Orientation



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.



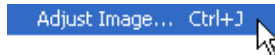
Also note that the document panel indicates which skew angle was corrected and which rotation was executed!

Width	Height	Lineskew	Rotation
173	114	5.69°	turned upside down

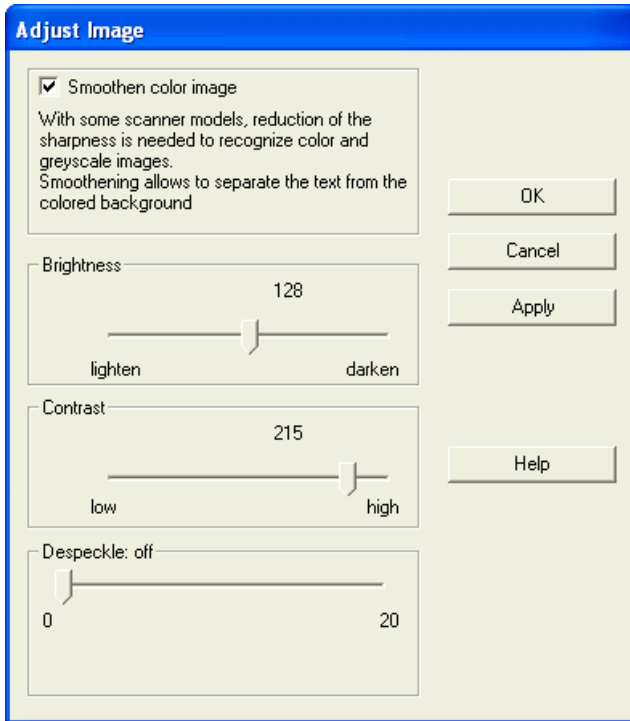
## **ADJUSTING THE SCANNED IMAGES**

---

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.

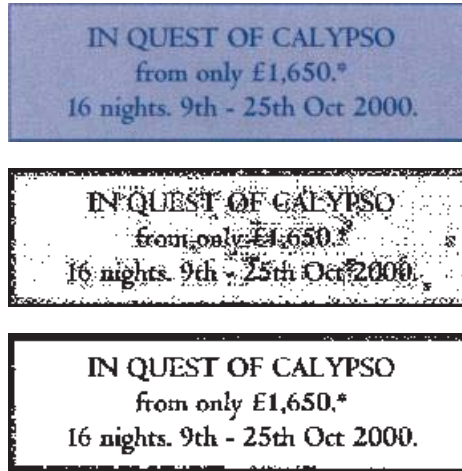


The option "Smoothen Color Image" renders greyscale and color images more homogeneous by “flattening”, smoothing out relative differences in intensity. As a result, a stronger contrast is created between the foreground - the text - and the background - a color, artwork etc.

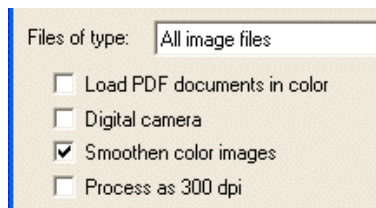
This **preprocessing** feature may seem highly technical and difficult to understand, but it certainly has its role to play: with some scanner models, this reduction of the sharpness is needed to recognize color and greyscale images.



Smoothing 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 smoothing can also be enabled when you load prescanned images into memory!



The **brightness** now. This setting determines the overall brightness of the image: any darkening or lightening of the image applies to all pixels. The objective is to rid yourself of the page background. We'll give two examples. In the first example, every zone of the image is dark. We therefore lighten the image to eliminate the page background. The foreground - the text - remains sufficiently dark to get detected by the binarization. Example 2: the image is so light even the



foreground text doesn't show up in the binarized image! We darken the image so as to make the text legible.

**Verenigde Staten,  
een antwoord te vi  
maar met name or**

wyjścia każdego  
brawia, że nasze



wyjścia każdego  
brawia, że nasze

**Verenigde Staten,  
een antwoord te vi  
maar met name or**

wyjścia każdego  
brawia, że nasze

The **contrast** determines the local contrast between the darker and lighter zones of the image. (The text is usually darker than the background - the reverse is true when you're dealing with inverted text.) The objective is to make the character shapes stand out nicely against their (colored) background. Here's an example where we need to increase because the default setting yields broken characters.

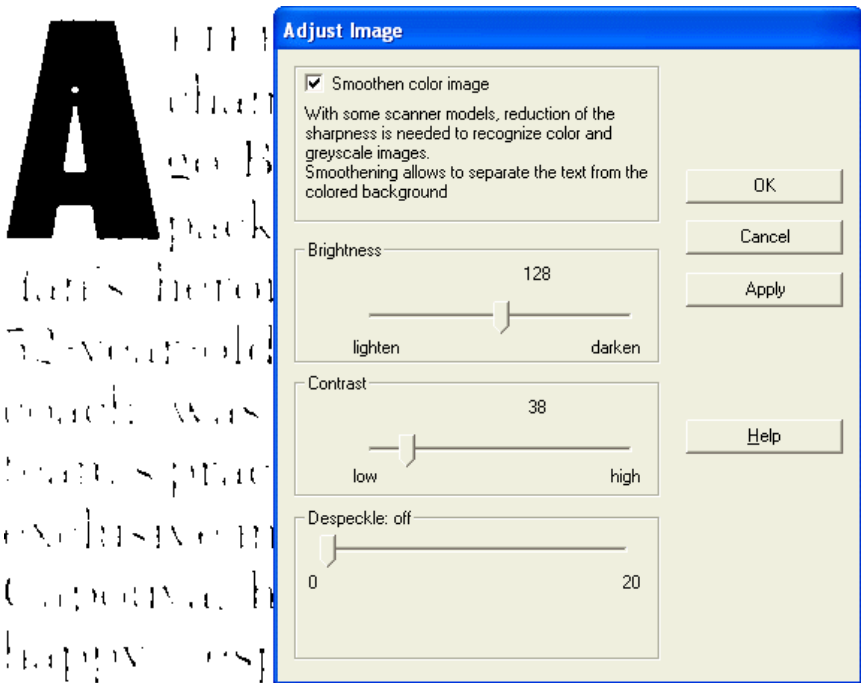
**A Look at International  
Planning the Future .....**

**A Look at International  
Planning the Future .....**

**A Look at International  
Planning the Future .....**

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!



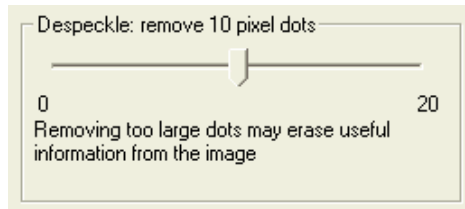
These options concern color and greyscale images, the last one, "Despeckle", exclusively concerns black-and-white images. "Despeckling" means that the "parasite pixels" (also called "salt and pepper noise") will be removed from black-and-white images.



# The Olympic Games

## The Olympic Games

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.!



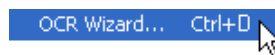
The best way of optimizing the images for the OCR process is this: place the adjustment window where it doesn't prevent you from judging the image adjustment you execute. Adapt the parameters - clicking "Apply" each time - until the image is crisp and clear.

## 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

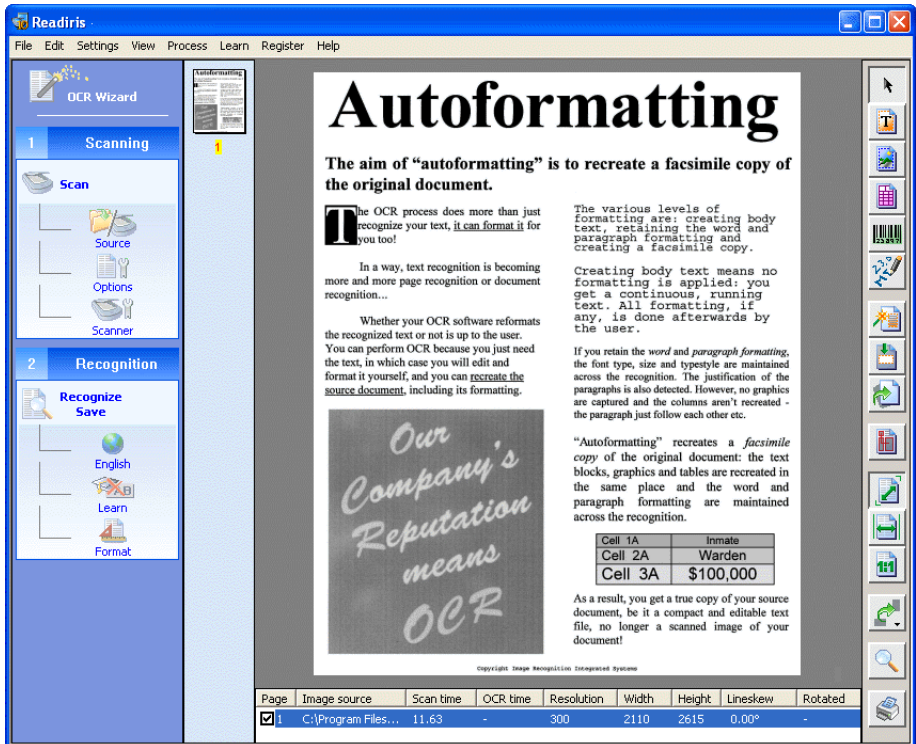


text block are recreated. So are the bulleted and numbered lists. Any e-mail addresses and URLs of web pages get detected and recreated as hyperlinks in the output. The placement of columns, text blocks and graphics follows your original document.

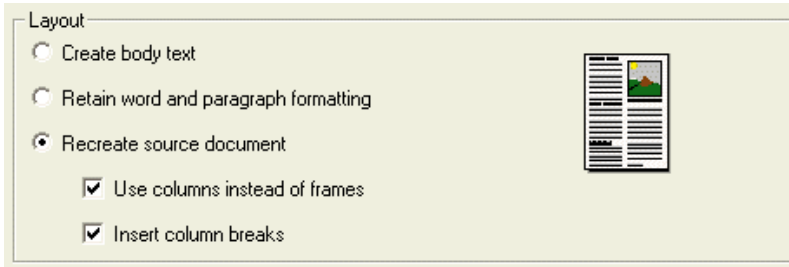
In other words, Readiris allows you to archive a true copy of your documents, be it a editable and compact text file instead of a scanned image!

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 AUTOFORMAT.JPG which is found in your Readiris folder.



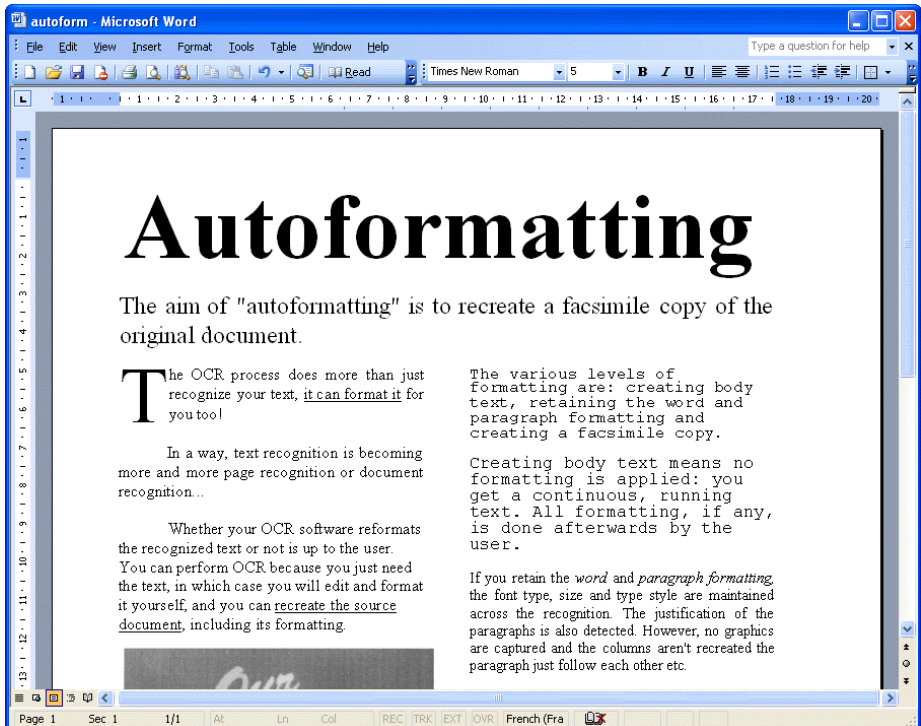
Click the "Format" button on the main toolbar and choose to send the OCR result to Microsoft Word or select the RTF (Rich Text Format) or Word (\*.doc) format. Secondly, select "Recreate Source Document" as layout option. (The option "Merge Lines into Paragraphs" is enabled by default to apply wordwrap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



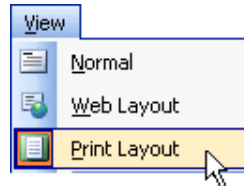
Whether layout reconstruction is available depends on the selected output mode. Some “poor” formats generating “plain” text such as Text (ANSI), MS-DOS Text (ASCII) etc. do *not* support advanced formatting codes and therefore cannot offer autoformatting. The Adobe Acrobat PDF format on the other hand 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.





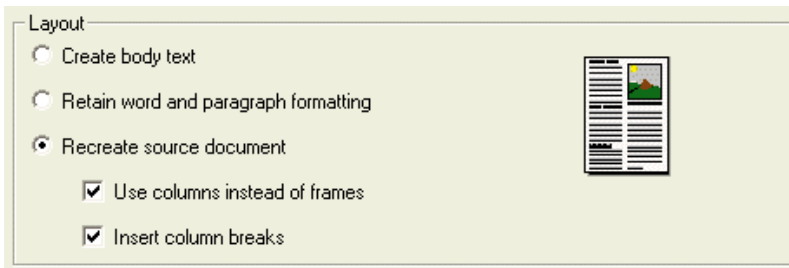
To see the effect correctly, you need to enable the "WYSIWIG" mode of your wordprocessor, mostly called "page layout" mode. However, if you send the recognized document directly to Microsoft Word, the page or print layout view is activated automatically!



In short, Readiris not only recognizes your texts, but can format them for you as well. OCR isn't just text recognition anymore, it has become **document recognition** as well!

## COLUMNS PLEASE, NOT FRAMES!

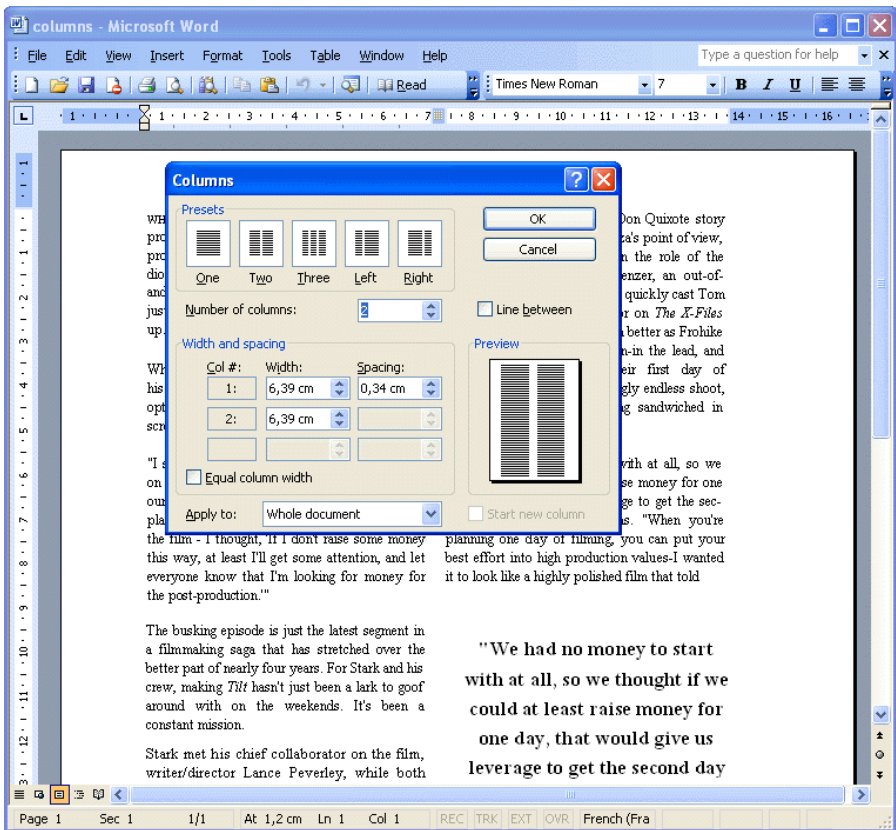
The formatting option "Use Columns instead of Frames" determines *how* the "autoformatting" gets done: the text blocks, tables and graphics can either be stored in frames or in editable **columns**.



"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.



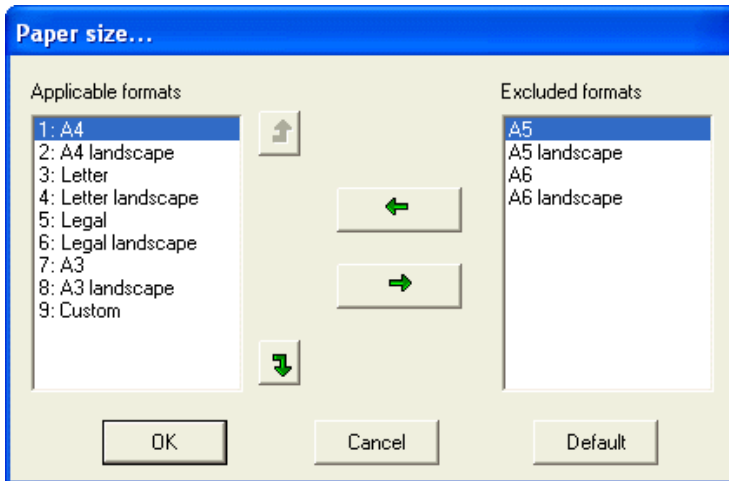
With column breaks, any text you edit, add or remove remains inside its column; no text ever flows automatically across a column break. All text that follows a column break is moved to the top of the next column!

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.

There’s one aspect where you may decide *not* to recreate the source document: the page size of your output documents. What do we mean here? Let’s give some examples: you’re scanning A4 pages but you create Letter output because that format is easier to print, whereas A4 requires a manual feed. Or you may be an attorney; you scan Letter documents that you save in the Legal format.

That’s why Readiris allows you to define preferred paper sizes for the output documents. Click the button "Paper Size" in the "Text Format" dialog.



Select the applicable and excluded paper sizes: the applicable paper sizes can be used to format the recognized documents, the excluded formats won't ever be used. Sort the applicable paper sizes: Readiris goes through the various page sizes in the indicated order and utilizes the first paper size that is sufficiently large to hold the scanned document. The button "Default" re-applies the default settings. (These take your Windows settings into account!)

Know that this option does not apply to HTML files - a text format designed for the Internet that doesn't have any page formats! Nor does it apply to PDF files, which apply a custom fit to recreate the source document accurately.

## **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-Save" 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.

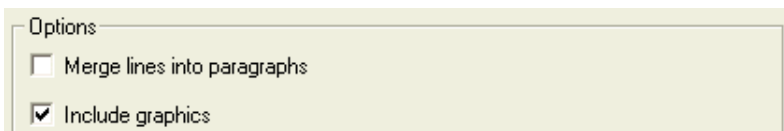


The same goes for any other element you change: when you add a page to your OCR job, only that page will be recognized. When you create a new text zone on any page, only that zone will be recognized before the results get exported.

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 document; 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.





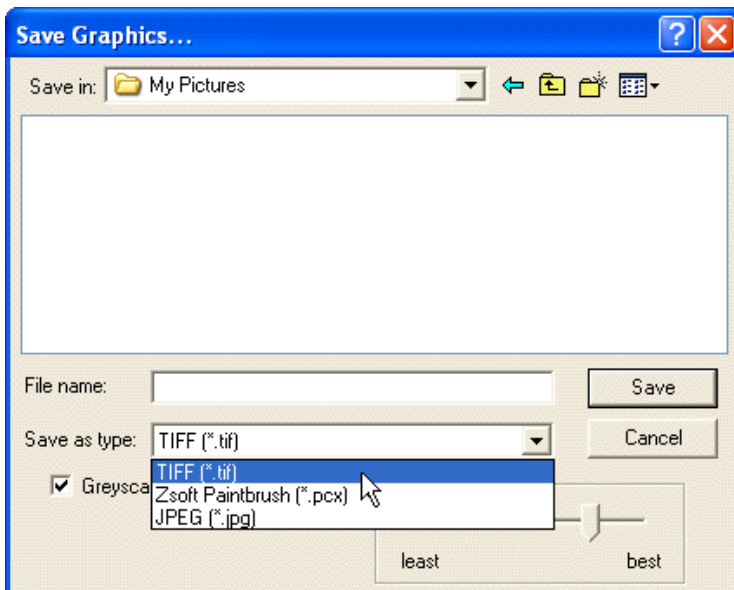
Next, choose the command "Save Graphics" under the "File" menu.

Save Graphics...



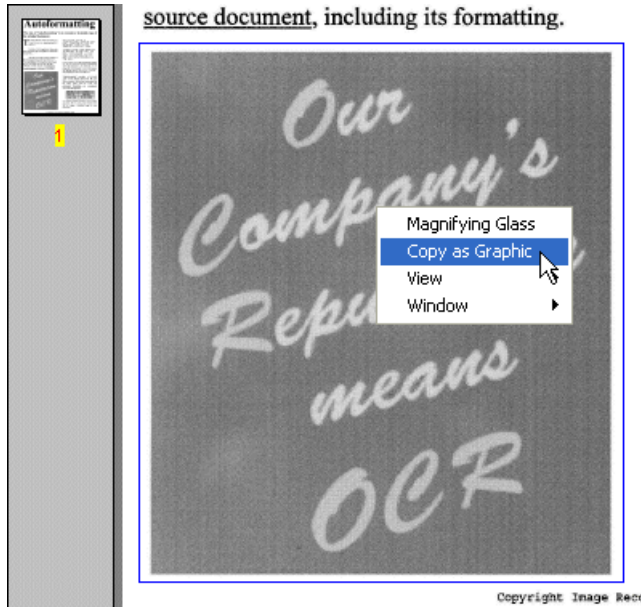
You are prompted to specify a file name. Determine which graphic file format you will use. Select a format that's supported by your paint or photo retouching software. The JPEG, TIFF and ZSoft Paintbrush (\*.pcx) formats are supported. Readiris Corporate also supports the compact format JPEG 2000 (\*.j2c)!

Enable the option "Greyscale/Color" to save the graphic as a color or greyscale graphic. When you save black-and-white graphics in the TIFF format, Group 4 compression is used. When you save greyscale and color graphics in the TIFF format, JPEG compression is used and you can adjust the JPEG quality.





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!

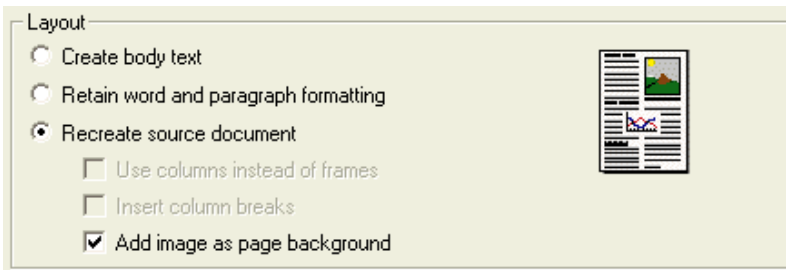


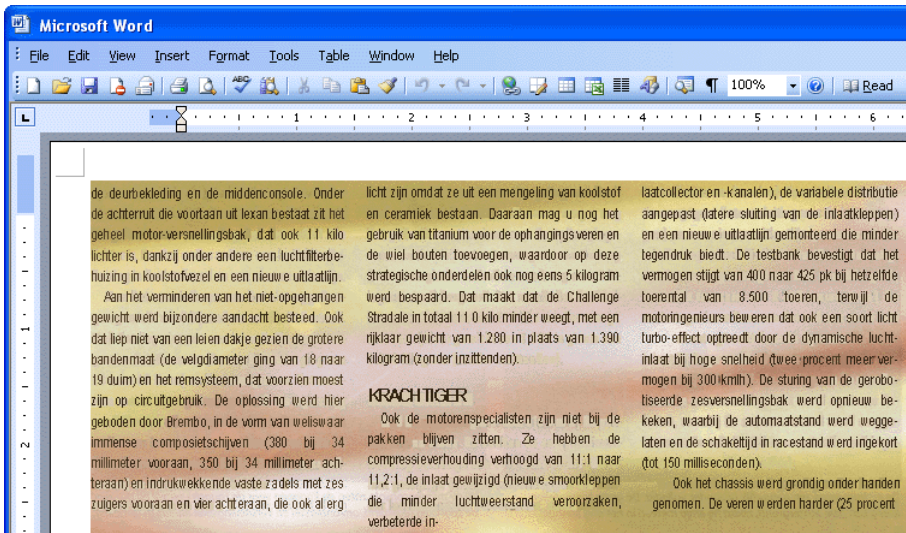
## **SAVING COLORED BACKGROUNDS**

With Readiris Corporate, you can also save the background inside the recognized documents! The option "Add Image As Page Background" places the scanned image as page background under the recognized text; this feature is supported for HTML, RTF, Word (\*.doc) and WordML (\*.xml) output. (When you generate PDF files, you can produce the same effect by selecting the format "PDF Text-Image".)

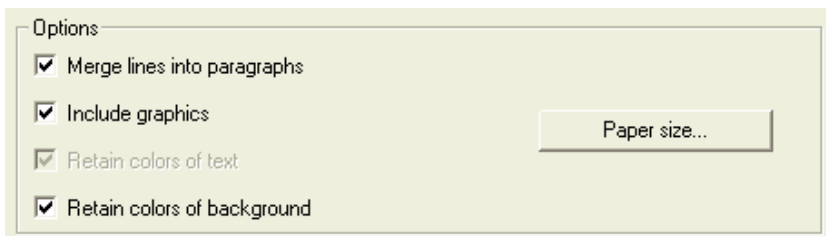


It goes without saying that this option increases the file size of your recognized documents substantially as this option adds the scanned image with all its detail to the document output in the background. *All its detail?* Not really: as also holds for PDF “text-image” output, the pixels of the recognized text are erased to create a legible document. Displaying recognized text in black on top of black character bitmaps would give you text with a heavy shadow... (The sample image BACKGROUND.JPG illustrates how it works.)





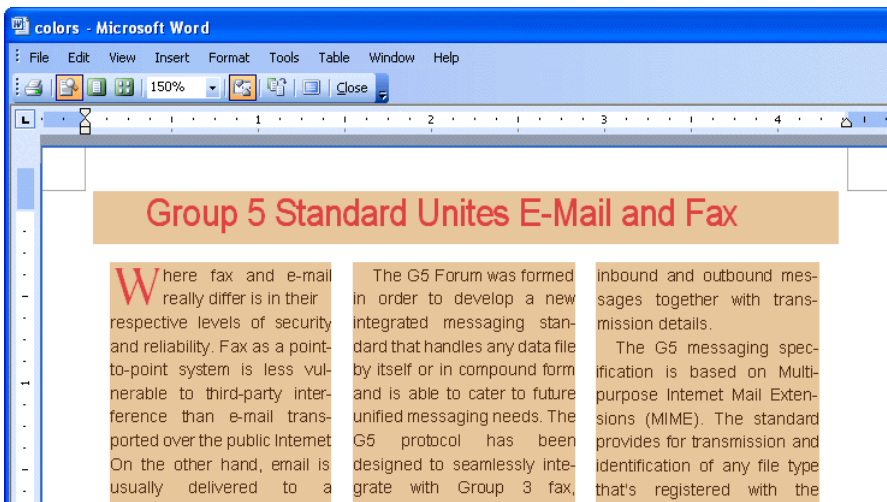
Readiris also offers a less drastic and more compact alternative: the option "Retain Colors of Background" maintains the spot colors on the page across the recognition. (This option implies that you retain the colors of the text.)



You'll get one uniform background color - if there is one in the source document - per paragraph. This again applies to HTML, Word and WordML documents. The details of, say, a full-page photograph in the background are *not* main-



tained this time, the spot color of a text frame is is. (Recognize the sample image COLORS.JPG to give it a try...)

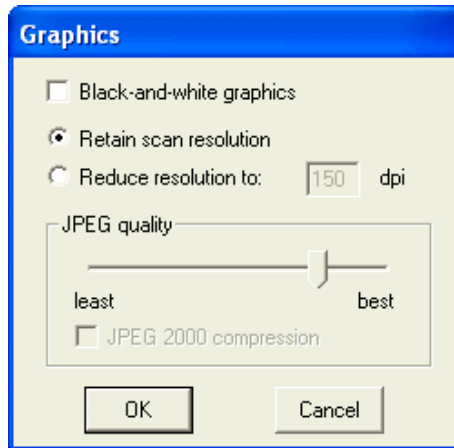


## TAKING GRAPHICS TO THE HILT

---

Readiris Corporate offers other advanced options for the graphics. You'll find these under the button "Advanced" of the "Text Format" command.

These settings also apply to *all* graphics - the graphic zones included in recognized documents and the page image you place above the text in an "image-text" Adobe Acrobat PDF file!



Determine the color mode: save your graphics in color-greyscale or as black-and-white images. Select the resolution of the images in the recognized documents: maintain the scan resolution or reduce it. (You cannot *increase* the resolution in this way.) When the recognized documents get posted on a web site as HTML files, you'll undoubtedly want to reduce the graphics to screen resolution. Hence, for HTML files, the graphics are by default reduced to 72 dpi. (You can fill out a higher value manually.) Finally, you can select the JPEG quality. (JPEG images are used to store color and greyscale graphics in PDF documents, Word and RTF documents etc.)

(The option "JPEG 2000 Compression" we already discussed: it applies JPEG 2000 compression to all graphics and images contained in PDF documents.)

These options allow you to influence the **file size** of your output documents substantially! Let's give an example: when you generate "image-text" PDF files with bilevel graphics, you store the images in Group 4 compressed TIFF files. Save the same scans as color images and you store JPEG files by default with a high (0.8) quality.

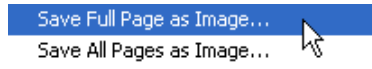


## READING FAXES AND DEFERRED RECOGNITION

---

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 be prompted to save the entire page as a PDF, JPEG, JPEG 2000 (\*.j2c), TIFF or ZSoft Paintbrush (\*.pcx) file.



Select the command "Save All Pages as Image" to save a multipage document. Two file formats are available here: PDF and 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, DCX fax images (a multipage version of the Paintbrush format), DjVu images (\*.djv, \*.djvu), JPEG images, JPEG 2000 images (\*.j2c, \*.jp2), PNG images, TIFF images (uncompressed, LZW, PackBits, Group 3, Group 4 and JPEG compressed), multipage TIFF images, Windows bitmaps (\*.bmp) and ZSoft Paintbrush images (\*.pcx).

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.

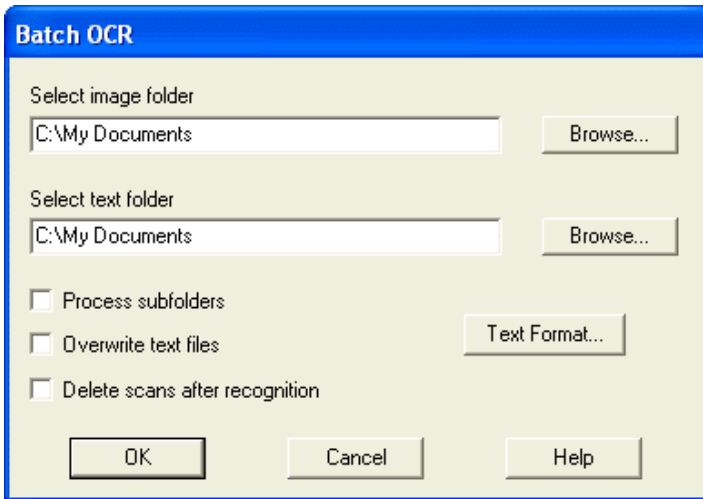
## **RECOGNIZING DOCUMENT BATCHES**

---

Readiris Corporate is much more powerful when it comes to recognizing prescanned images: you can recognize entire document batches automatically and you can establish a watched folder. Let's study this in detail!

**Batch OCR** executes the recognition on all prescanned images in a specific folder. You could for instance scan by day and read by night... Acquire all documents to be recognized; when you're done, run the batch OCR and leave your PC to it. When you return the next day, all your documents have been recognized!

The recognized documents get the same file name as the image files. The file extension obviously depends on the selected output format: image file 001.TIF gets converted into 001.DOC when Word is the selected output format.



Select the image and the text folder. The text folder can be different from the image folder - but doesn't have to! When the image folder is identical to the output folder, you'll find the text documents alongside the scans (unless you enable the option "Delete Scans after Recognition")!

The option "Process Subfolders" determines whether the subfolders of the image folder will be processed as well. Doing so makes sense when the prescanned documents are placed in specific subfolders. You may for instance have a folder named "2-18-2005" that indicates the date and subfolders "1", "2" etc. that contain the actual documents.

If enabled, this option triggers the processing of *all* subfolders; you cannot limit the OCR to (a) specific subfolder(s). When the text folder differs from the image folder, the subfolders get recreated, "mirrored" in the output folder!

Click "OK" to execute. No further effort is required to convert the documents: the recognition process is entirely automatic. (Interactive learning does not apply here.)



Furthermore, you are limited to the “external” text formats. The OCR results are saved on disk. Sending the reading results directly to a target application, sending them by e-mail or opening them automatically after the recognition simply doesn't make any sense when the OCR gets executed on an “unattended” PC. But all other OCR options - the current document language, font type etc. - apply: be sure to enable the proper options before you trigger the recognition!

Batch OCR comes with some advanced features. The option "Overwrite Text Files" determines whether the OCR process can replace previous recognition results. Disable this option when you add new image files to a folder that contains images that have been processed previously. (Otherwise your image files would be re-recognized when you process the folder a second time...)

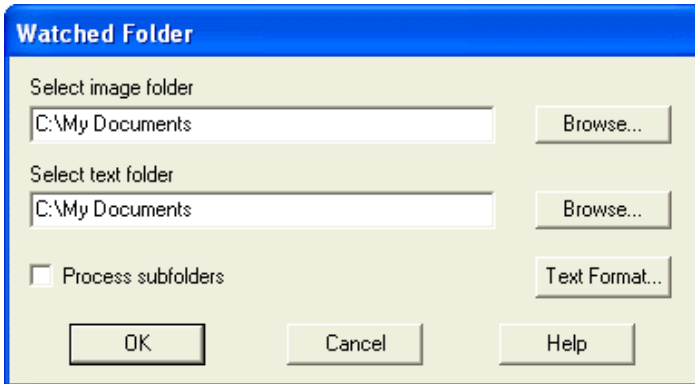
The option "Delete Images after Recognition" determines whether the image files will be deleted after the recognition. In other words, Readiris can “clean up” your image folder for you!

Enable this option when you exclusively store-archive the recognized documents, discarding the “temporary” image files. Disable this option when you save both the scans and the recognized documents. (We repeat that Readiris generates PDF documents of the type “image-text”: this output format saves the scanned image *and* the recognized document in a single file!)

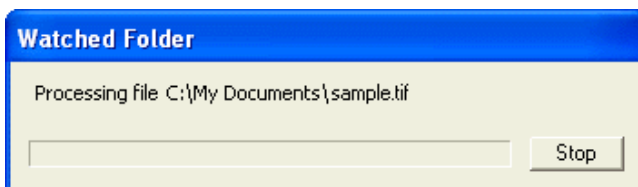
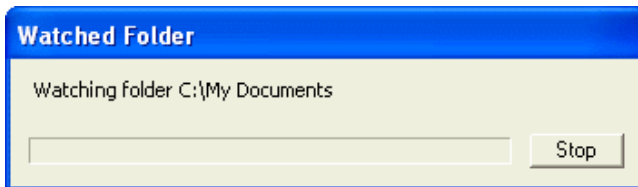
## **ESTABLISHING A WATCHED FOLDER**

---

The use of a “watched folder” is largely similar to the operation of batch OCR. The major difference is that, this time, the user does not trigger the recognition. On the contrary, Readiris systematically executes the recognition on any image file that gets dropped in a specific folder. You can leave the OCR software running day after day... Acquire new documents and they will be recognized promptly.



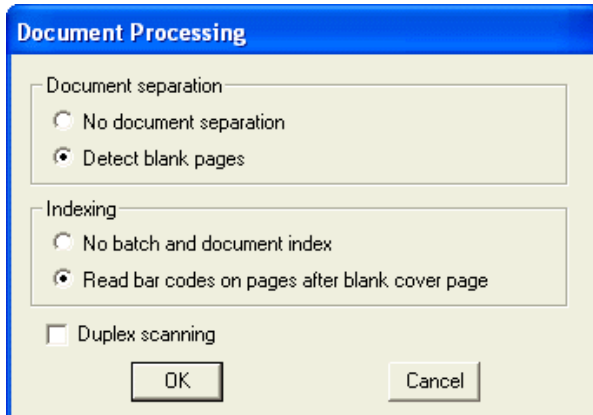
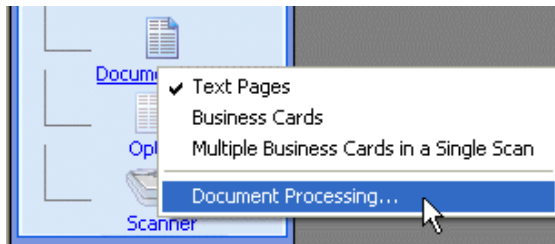
Click "OK" to activate the “supervision” of the **watched folder** and Readiris processes the images progressively as they get created-placed in the watched folder. Click "Stop" to interrupt the supervision.



As is the case with batch OCR, Readiris processes the images of all supported file formats. You cannot limit the OCR to a specific graphic format: Readiris processes the images of *all* supported file formats. (Any files with another file format are simply ignored.)

## ORGANIZING BATCHES

None of this means that the (pre)scanned pages you scan are by necessity processed blindly, in a straightforward way. On the contrary, Readiris Corporate comes with sophisticated routines that allow to process the scanned batches intelligently. Go to the command "Document Processing" on the "Document Type" button (or under the "Settings" menu) to discover how it all works.



Blank cover pages can be inserted between the pages to **separate** the **documents**. A "blank" page is a page with hardly any black pixels - black borders

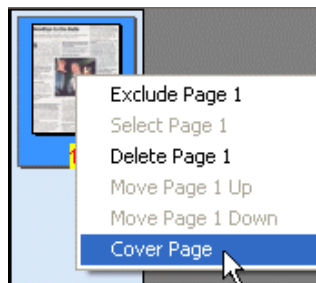


excluded. The page toolbar highlights cover pages, the document panel marks them.



Page	Cover page	Bar code	Image source	Scan time
<input checked="" type="checkbox"/> 1	yes		C:\Documents an...	0.48
<input checked="" type="checkbox"/> 2	-	97890546	C:\Documents an...	1.27

You can also segment a stream of scanned images manually: select any page - blank or not - in the page toolbar, right-click it and enable the option "Cover Page" from the "Context" menu. (Know that the contents of a cover page are lost: cover pages serve as separator but are *not* read!)



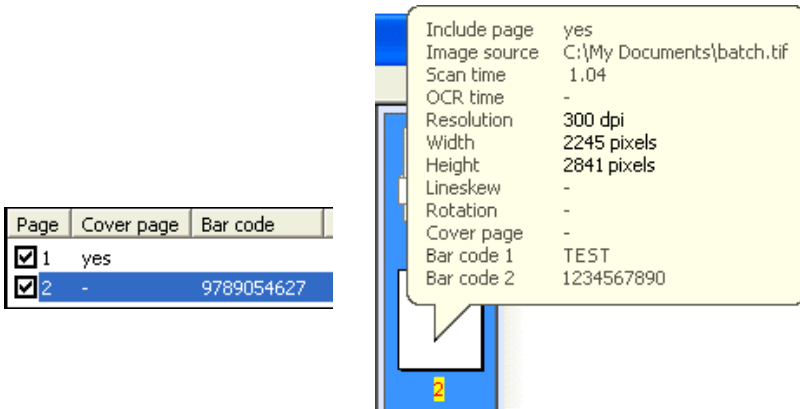
Breaking up the various documents in a stream of scanned pages is the first step. **Indexing** them accurately is the next step. Bar code reading is applied to this end. All bar codes that occur on a page that follows a blank cover page get read automatically. (To give it a try, recognize the sample image BATCH.TIF.)



The option "Duplex Scanning" indicates that you're scanning the front and rear sides of documents. The rear sides will obviously be ignored to detect a bar

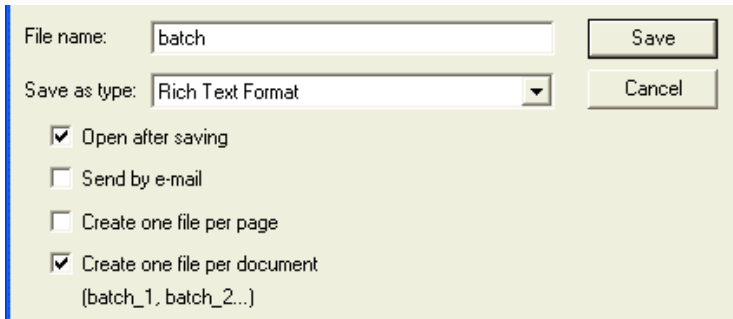
code on the next page: the next page becomes the next *front* image, not the next image!

The Readiris interface is helpful in other ways when you start processing document batches: the document panel and the tooltips on the page icons indicate the value of the recognized bar codes!



The bar code reading results are saved in the XML index, not in the recognized document!

When you process batches manually, the option "Create One File per Document" in the "Save" dialog saves separated documents, not complete batches. (This only works when you create external files, not when you send the output directly to a target application.)



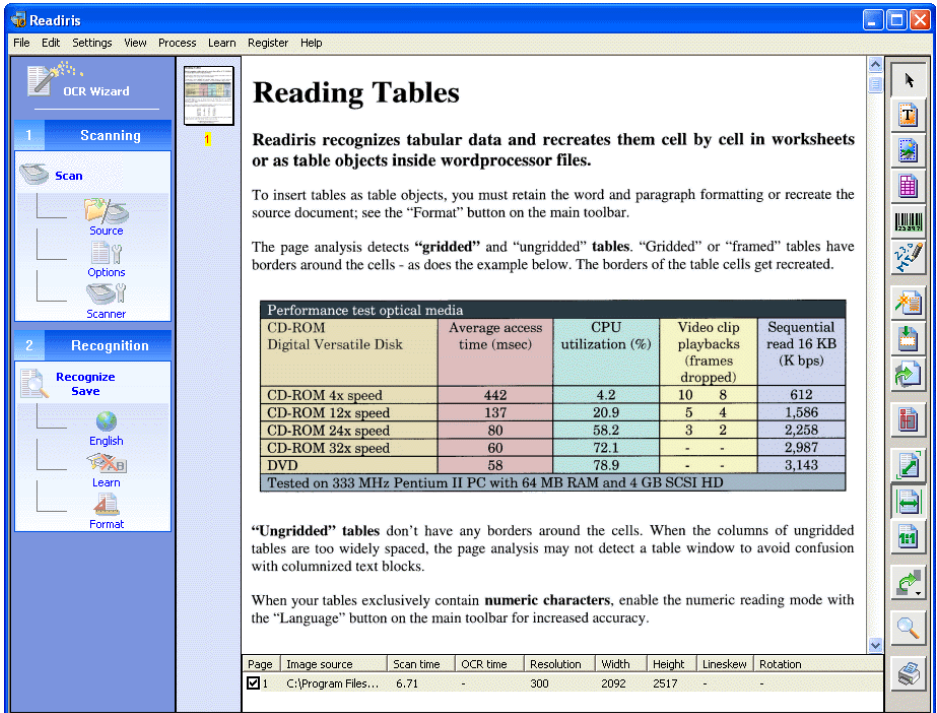
## RECOGNIZING TABLES

---

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 reconstruct 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.)



table - Microsoft Word

File Edit View Insert Format Tools Table Window Help

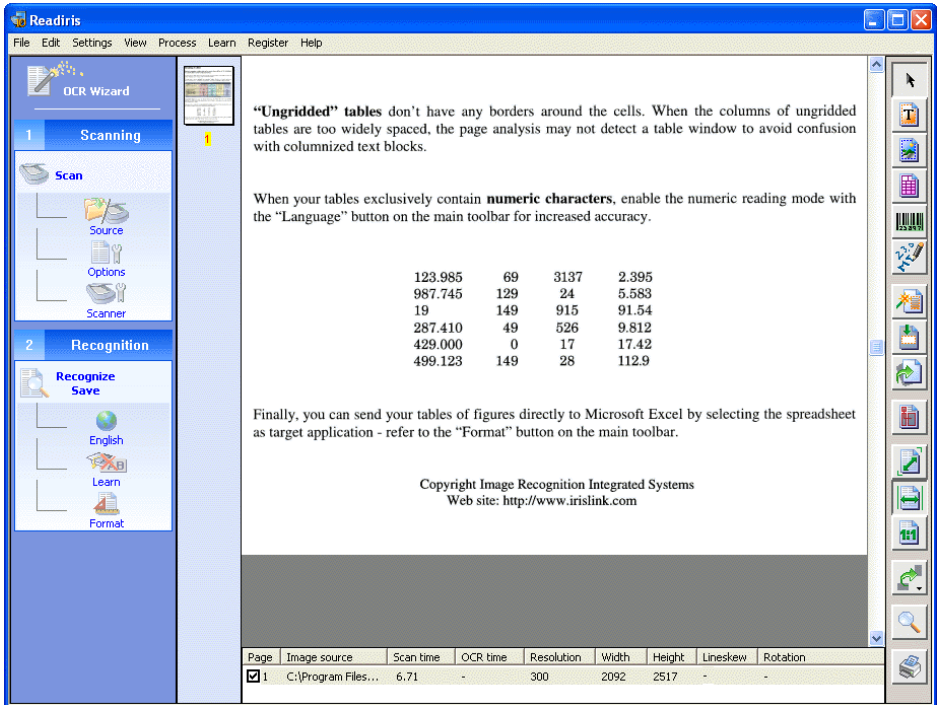
Type a question for help

Times New Roman 12 B I U

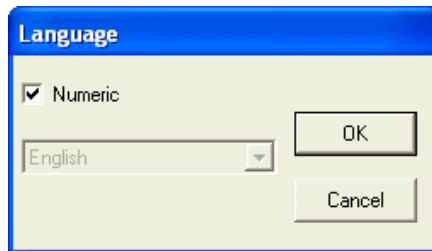
Performance test optical media					
CD-ROM Digital Versatile Disk	Average access time (msec)	CPU utilization (%)	Video clip playbacks (frames dropped)		Sequential read 16 KB (K bps)
CD-ROM 4x speed	442	4.2	10	8	612
CD-ROM 12x speed	137	20.9	5	4	1,586
CD-ROM 24x speed	80	58.2	3	2	2,258
CD-ROM 32x speed	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!



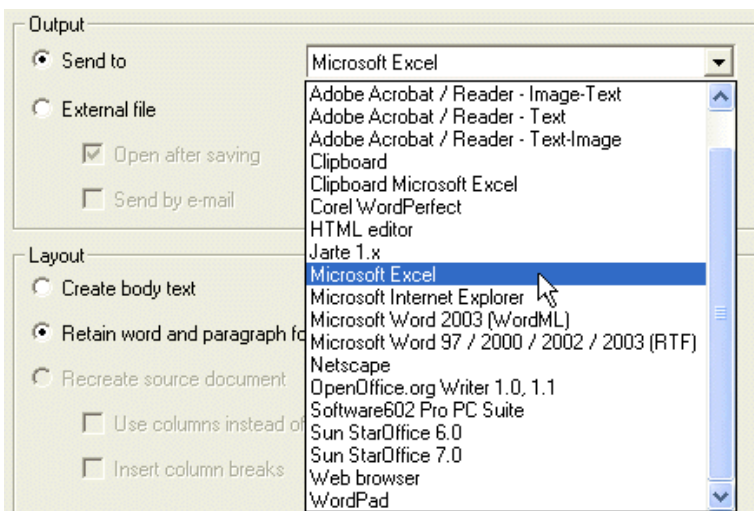


For optimal OCR accuracy, you should limit recognition to the **numeric symbols** with the "Language" button. (The numeric mode is not strictly numeric, it includes the symbols “0” to “9”, “+”, “\*”, “/”, “%”, “,” (comma), “.” (dot), “(”, “)”, “-”, “=”, “\$”, “£”, “¥” and the “€” symbol.)



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.

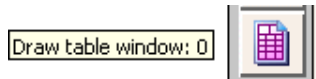
This time, we will send the OCR result directly to the spreadsheet Microsoft Excel, so we select Excel as target application under the "Format" button.



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.

	A	B	C	D	E
1	123.985	69	3137	2.395	
2	987.745	129	24	5.583	
3	19	149	915	91.54	
4	287.410	49	526	9.812	
5	429.000	0	17	17.42	
6	499.123	149	28	112.9	
7					

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.



## RECOGNIZING HANDWRITTEN TEXT

We've recognized scanned documents, tables, faxes, snapshots taken with a digital camera, we've saved graphics and we've converted PDF documents. Readiris adds yet another reading capability: the recognition of handwritten texts.

Actually, we should say handprinted text, not handwritten text! Hand *writing* is used to describe continuous, “cursive” handwritten text. The symbols within a



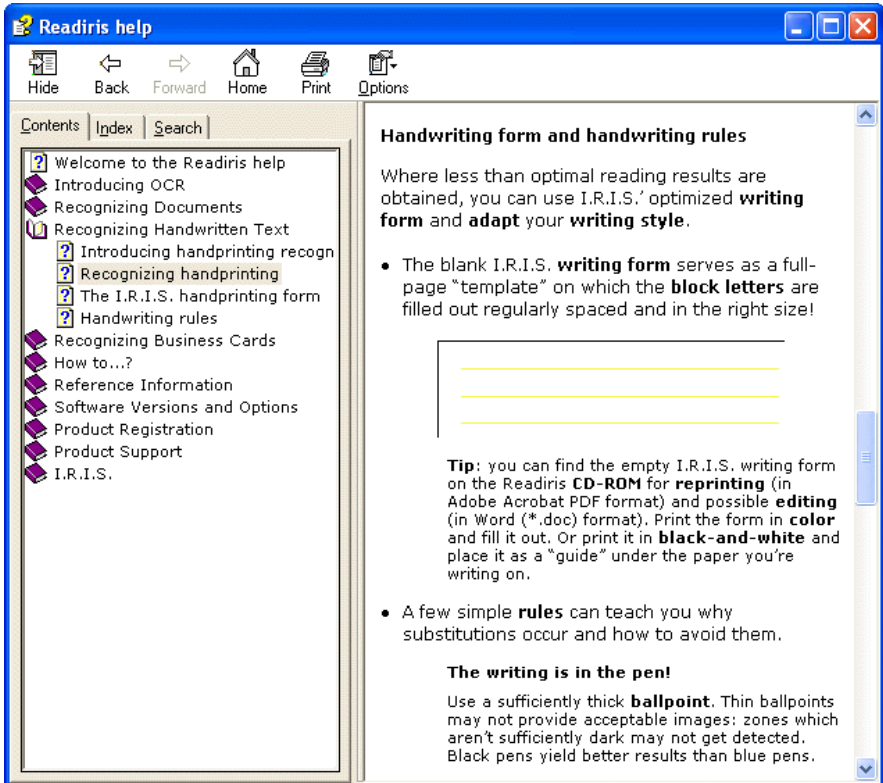
word or character string touch, it is impossible to say where one symbol ends and another starts. With *handprinting*, the “block letters” are separated and the recognition software has an easier job to isolate the individual characters.

*Kalicienskie*  
*001-0405369-28*

It takes highly specialized software - “ICR” or “Intelligent Character Recognition” software - to recognize handprinted symbols. I.R.I.S.’ powerful ICR technology is based on more than one million writing samples! Readiris supports all natural writing styles - American or European. No imposed style is required.

Handprinting recognition is limited to the numerals (0-9), the uppercase letters (A-Z) and the punctuation symbols “,” (comma), “.” (dot) and “-” (hyphen).

Where less than optimal reading results are obtained, you can adapt your writing style and use I.R.I.S.’ optimized writing form. Consult the on-line help of Readiris to discover the writing rules. A few simple tips can teach you why substitutions occur and how to avoid them. The blank I.R.I.S. writing form serves as a full-page “template” on which the block letters get filled out regularly spaced and in the right size! You can find the empty form on the Readiris CD-ROM for reprinting and editing.



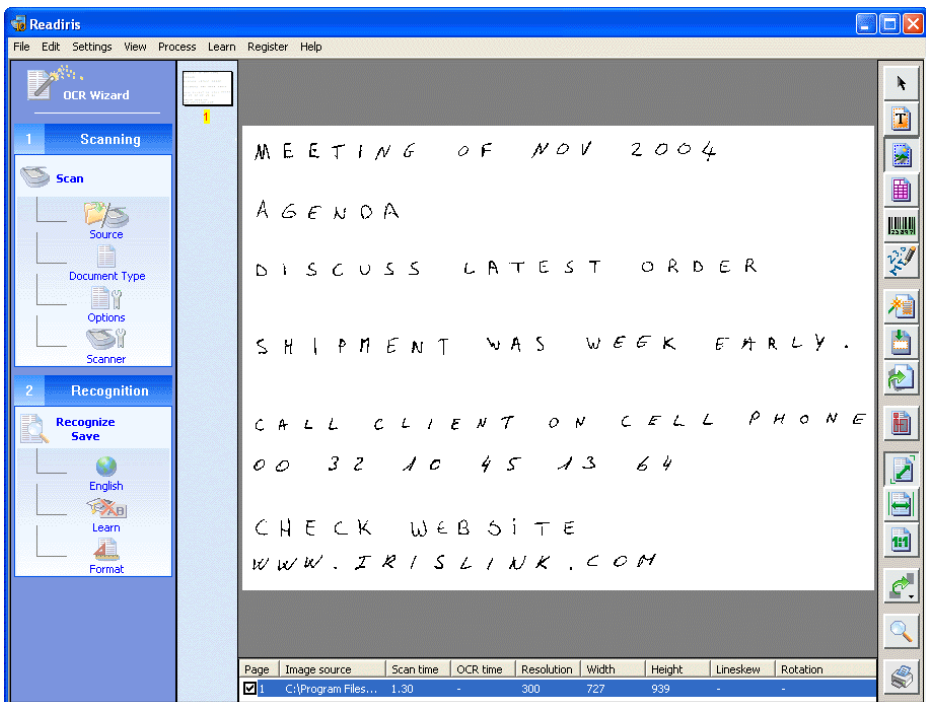
So now we know how we can take machine-legible handwritten notes during a meeting. How can we proceed to recognize these notes afterwards with Readiris? Draw a handprinting window around the handprinted text and execute the recognition. (You can give it a try with the sample image **HANDPRINTING.TIF!**)



Draw handprinting window: 0



The document characteristics - language, font type and character pitch - do not apply to handprinting. You're limited to a basic English - or should we say "Latin"? - character set of (uppercase) block letters. Nor does interactive learning apply: learning does not make sense in an environment where everybody has a particular handwriting style. (As indicated, the ICR technology is based on more than one million writing samples...)



## READING BARS AND SPACES

And Readiris reads bar codes too...! Bar codes that figure in scanned images can be read and included as recognized data inside the output documents.

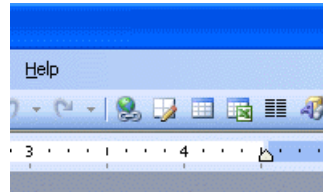
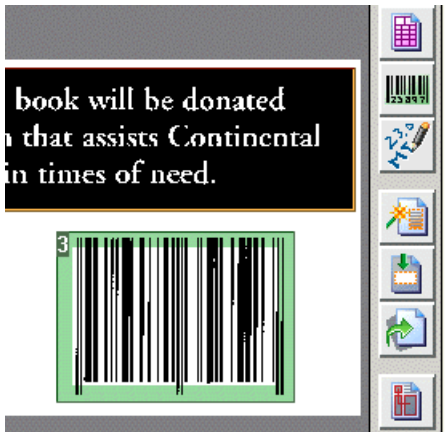
Bar codes are composed of parallel bars and spaces between them. Pre-defined combinations of bars and spaces represent specific characters. There are several bar code standards or "symbolologies". All widespread bar code symbolologies are supported.



Only laserprinted and inkjet printed bar codes have sufficient quality. Exclude matrix printed bar codes: they do not produce sufficient contrast and their resolution is mostly limited to 60 dpi! Readiris recognizes well contrasted bar codes best; black bars on a white background yield the best results. Most bar code types require a "quiet zone" around the actual bar code. Bar codes don't render partial results; a missing start or stop character or an incorrect check digit always lead to a misread, a zero result!



Draw a bar code window around each bar code - the page analysis does not detect them - and execute the recognition. The bar codes are read and included in the text output. You can also right-click a bar code zone and select the command "Copy As Data" from the "Context" menu; the bar code is read and sent to the clipboard... (The check characters of some bar code standards are verified but stripped from the reading results.) The sample image BARCODE.TIF illustrates how it works.



s book will be donated  
 :ion that assists Continental  
 ies in times of need.

978047135652

## READING BUSINESS CARDS

We've recognized scanned documents, tables, faxes, snapshots taken with a digital camera, we've saved graphics and we've converted PDF documents. Readiris Corporate adds yet another reading capability: the recognition of business cards ("BCR" - "Business Card Reading")!

With Readiris, you can scan your business cards, recognize them and convert them into an **address database**. In this context, OCR allows to encode calling cards without the time-consuming task of retyping them. Think of your last exhibition when you came back with an entire stack of business cards and it took your secretary two days to encode them!

The card's data is extracted automatically from the image and the recognized data is assigned to specific database fields. Readiris extensively uses a knowledge database, thus acquiring the necessary intelligence to discriminate the first and last name, a city and its state, a telephone and a fax number etc. Each



country has a different "style" of composing business cards; Americans compose an address differently than the French do etc.

This works for up to 28 **countries**: North and South American business cards and business cards from the European countries, including the Eastern-European nations. (Optionally, you can read Asian business cards from China, Japan, Korea and Taiwan.)

The resulting data is available for **export**. You can save your contacts in a structured text file - for instance as comma delimited data or in the vCard format - to import them in any address database - for instance Microsoft Access.

Alternatively, you can send the contacts directly to your contact management software Microsoft Outlook (Express) and to your **PDA** software Palm Desktop. Business card reading smoothly complements such applications as contact managers, databases or even wordprocessors whose mail merge function allows to print letters, envelopes and labels.

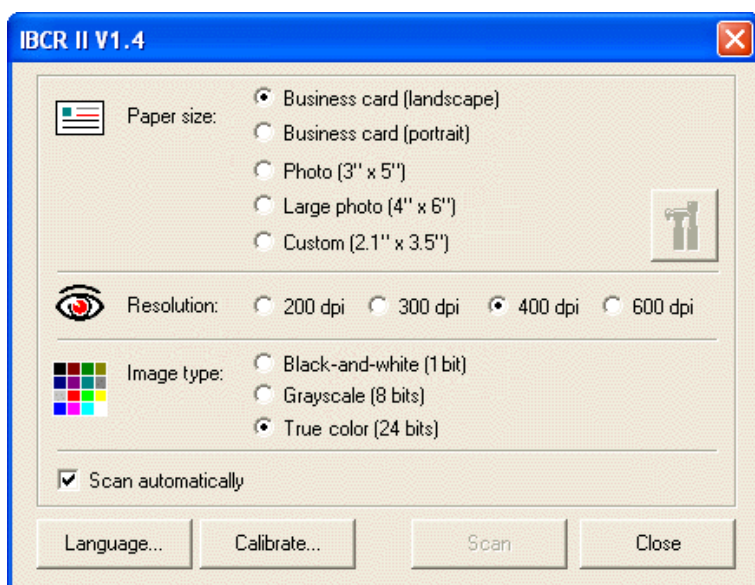
## **SCANNING BUSINESS CARDS**

---

How does business card reading work? For starters, the scanner must be set up appropriately.

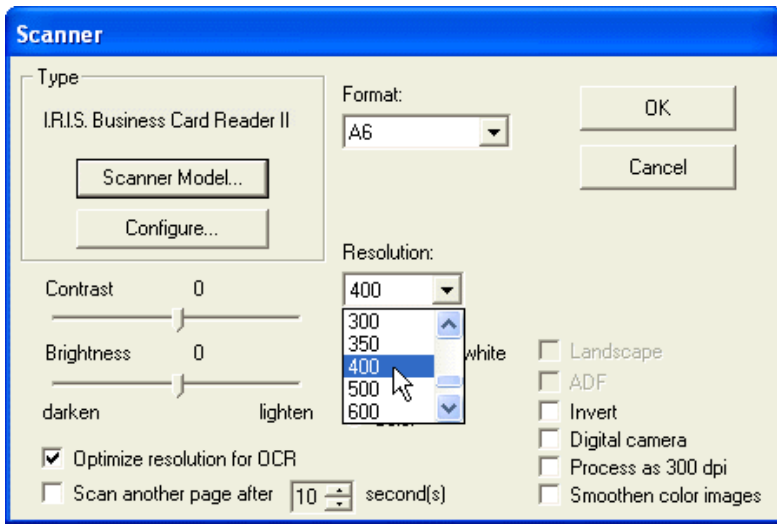
When business card reading is a major application - you're, say, scanning a great many business cards on a booth at a fair -, you can use I.R.I.S.'s dedicated card scanner **IBCR-II**. Dedicated **business card scanners** are optimized for scanning business cards and have many advantages over other scanner types: they hardly take up any space on your desk, swiftly cope with calling cards of varying paper and printing qualities and you never have to wonder about selecting the correct card format.

Use the Twain interface of the IBCR-II to enable the **automatic scanning mode**! Give the scan command once and enable the option "Scan Automatically" in the Twain user interface.



You can now insert one business card after the other: as soon as a business card is placed in the scanner, the scanning starts...

To recognize business cards successfully, we recommend you to select a scanning **resolution** of 400 dpi.



When you're using a **flatbed scanner**, you can scan several business cards simultaneously on the scanner's flatbed and have them segmented by the software. The background must be black if Readiris is to extract the various business cards. There's a very easy way of accomplishing that: scan your cards with the lid open!

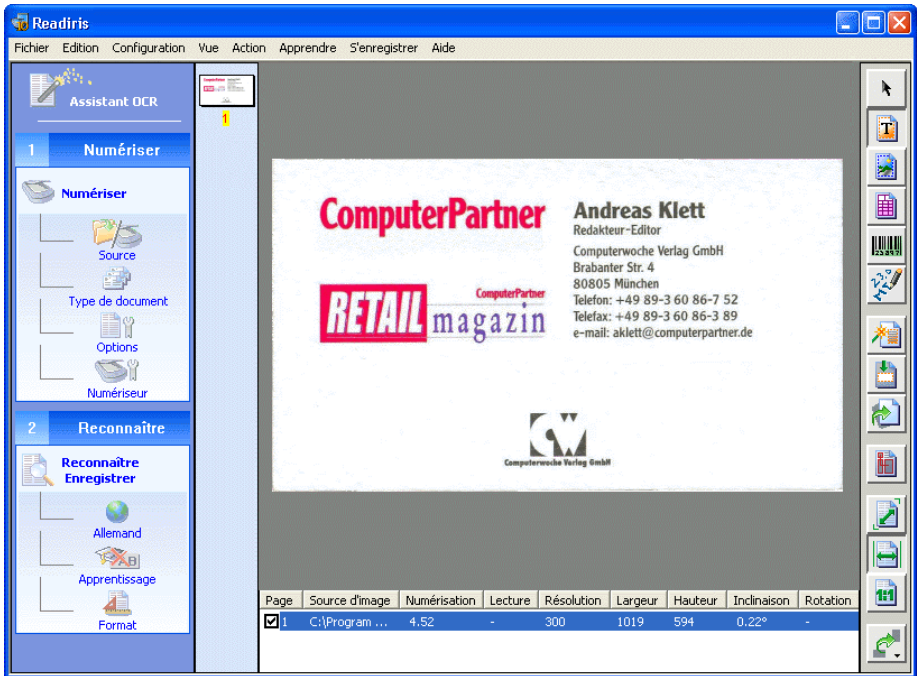


Select the option "Multiple Business Cards in a Single Scan" with the "Document Type" button on the main toolbar (or with the command "Document Type" under the "Settings" menu) and scan your business cards.

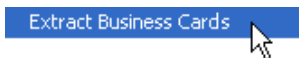


With flatbed scanners, it takes a click on the "Scan" button to initiate the image capture! (You can also select the command "Scan" from the "Process" menu to scan an image.)

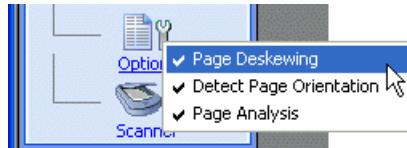
The complete page as you scanned it will never appear as such on the computer screen; only the segmented business cards will.



If you forgot to enable the image “splitting” with the multiple card mode, click the command "Extract Business Cards" under the "Process" menu to segment the large image into the actual card images, throwing away the superfluous black borders.



But whatever your scanner type may be, some options are enabled invisibly to optimize the reading process: page deskewing and the detection of the page orientation. Click the "Options" button in the main toolbar to discover them.



It is hardly conceivable that the business cards you put on a flatbed will (all) be perfectly straight, so let the software handle things for you... (Should you have disabled this option by accident, there's a "Deskew Page" button on the image toolbar to deskew the business cards afterwards, but things may quickly become tedious if you have to straighten every scanned business card manually!)

The same goes for the detection of the orientation: let the software handle this for you, otherwise you'll have to rotate business cards that were placed upside down or at a 90° angle on the scanner flatbed. (Use the rotation buttons on the image toolbar should you have disabled this option...)



## IT TAKES A BUSINESS CARD READING MODE!

The individual card images are now properly placed in the image window. Make sure that you enabled the **card reading mode** with the "Document Type" button (or with the equivalent command under the "Settings" menu).



Select the option "Business Cards" when you're scanning the business cards one by one, select "Multiple Business Cards in a Single Scan" to scan several business cards simultaneously on your scanner's flatbed. This option doesn't just concern the image splitting of your scans: it takes one of these two options to *recognize* business cards! Select "Text Pages" to disable the business card reading mode again.

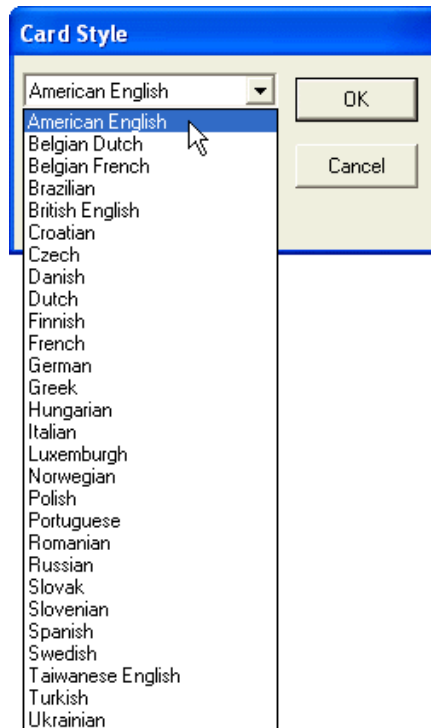
The button "Document Type" confirms that the business card reading mode is enabled.



Enabling this special mode is necessary because special OCR routines are then used that allow the software to assign the recognized data to specific database fields - company name, e-mail addresses and web sites etc. When Readiris recognizes the character string "John Williams", it "knows" that John is a first name and Williams a last name. In the same way so does the system recognize titles, such as "vice president", "engineer", cities such as Boise and Chicago, states such as Oregon and Maryland etc.

This works for up to 28 countries: North and South American business cards and business cards from the European countries, including the Eastern-European nations, are supported. (Reading Chinese, Japanese, Korean and Taiwanese business cards requires the optional module "Asian BCR add-on". Business cards from these Asian countries in English are supported by the "standard" Readiris Corporate software!)

As soon as the business card reading mode is enabled, the contents of the "Language" dropdown change! Always select the correct country.



By selecting the business card's "language" in the button bar, you not only indicate the language of the text to be recognized, you also indicate the general layout, style of the business card. Does it have an American or British look and feel to it? Each country has a different "style" of composing business cards, Americans compose an address differently than the French do, Dutch ZIP codes and telephone numbers have a different syntax than their British counterparts etc. (Some card styles correspond to several languages: Belgium and Canada



have two official languages, Switzerland has three! Readiris detects the selected language automatically...)



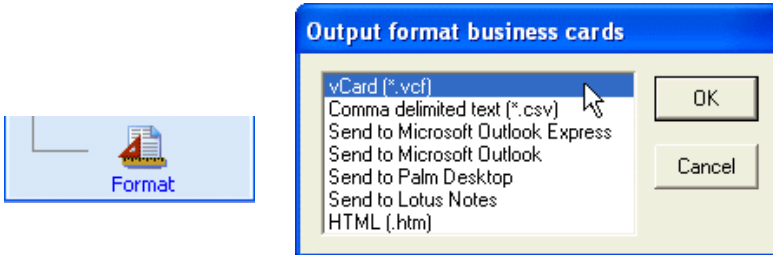
The field analysis comprises a third task: Readiris not only analyzes but also *formats* the recognized text. The system filters all irrelevant data from a business card, even if it plays an active role! If your business card mentions the information "Telephone: (508) 898-42 89", the resulting output will be "5088984289" in the telephone field. The word "telephone" is dropped, even if Readiris made good use of it to detect the location of the telephone number. The brackets, the hyphen and the spaces within the telephone number are also deleted.

## RECOGNIZING BUSINESS CARDS

Process your scans as usual by clicking the "Recognize-Save" button.



However, you are recommended to determine the output format before you do so. Click the "Format" button to this end.



A number of popular Personal Information Managers (“**PIMs**”) are supported directly: Microsoft Outlook (Express) and your **PDA** software Palm Desktop. To interface with other applications, use the “universal” formats vCard and comma delimited data. (vCard files are swiftly imported into any vCard compliant application: double-click a vCard file and the contacts are added to your Windows Address Book!)

The text result looks for instance like this when you send the results directly to Microsoft Outlook (Express).





**Readiris help**

Hide Back Forward Home Print Options

Contents Index Search

**?** Welcome to the Readiris help

- ◆ Introducing OCR
- ◆ Recognizing Documents
- ◆ Recognizing Handwritten Text
- ◆ Recognizing Bar Codes
- ◆ How to...?
- ◆ Reference Information
- ◆ Software Versions and Options
- ◆ Product Registration
- ◆ Product Support
- ◆ I.R.I.S.

**Welcome to Readiris™ Help...**

- Use on-line help to learn more about Readiris.
- Quickly find answers to questions.
- Connect to the I.R.I.S. web site for latest tips and product updates.

©2005 Copyright [I.R.I.S.](#) All rights reserved

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.