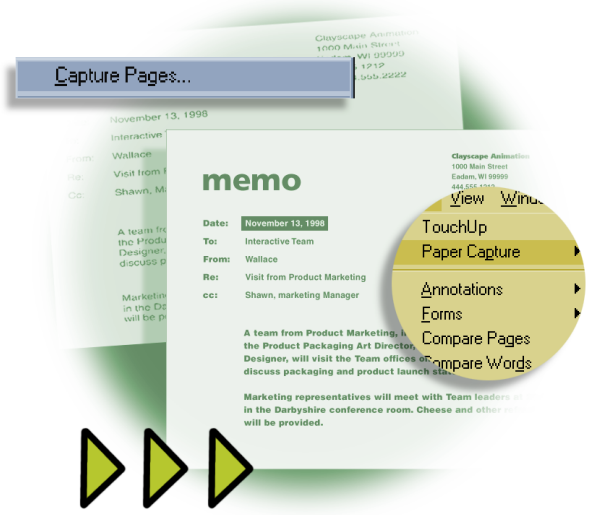
Lesson 10

Creating PDF Documents from Paper and the Web



Acrobat lets you create editable and searchable PDF documents by converting or "capturing" scanned documents and Web pages. You can use the resulting PDF documents for a variety of archival, presentation, and distribution needs.

In this lesson, you'll learn how to do the following:

- Capture a PDF Image file.
- Show and correct Capture suspects.
- Convert a Web page to PDF using Web Capture.

This lesson will take about 35 minutes to complete.

If needed, remove the previous lesson folder from your hard drive, and copy the Lesson10 folder onto it.

Capturing a fax image file

You can use the Import feature to convert image files, such as TIFF images or scanned paper documents, to PDF Image Only pages. In PDF Image Only format, all elements on a page can only be edited as bitmap images; text characters cannot be searched or edited. If your imported document contains text, you may want to convert the document to PDF Normal format so that the text can be edited and searched in Acrobat. You use the Capture feature to convert documents to PDF Normal format.

We've provided a fax document, scanned and saved as a TIFF image file, for you to import and capture.

Scanning text you plan to capture

- *For normal text, set up the scanner to create black-and-white (or 1-bit) images.*
- *Black-and-white images and text must be scanned at 200 to 600 dpi. Color images and text must be scanned at 200 to 400 dpi.*

Note: *Pages scanned in 24-bit color, 300 dpi, at 8.5-by-11 inches are very large files (24 MB); your system must have at least twice that amount of virtual memory available to be able to scan. If you're scanning in color, check that you have at least 50 MB of space available on your hard drive before beginning the scanning process.*

- *For color or grayscale pages with large type, consider scanning at 200 dpi for faster processing.*
- *For most pages, scanning at 300 dpi produces the best captures. However, if a page has many unrecognized words or very small text (9 points or below), try scanning at a higher resolution (up to 600 dpi). Scan in black and white whenever possible.*
- *Do not use dithering or halftone scanner settings. These settings can improve the appearance of photographic images, but they make it difficult to recognize text.*

• *For text printed on colored paper, try increasing the brightness and contrast by about 10%. If your scanner has color-filtering capability, consider using a filter or lamp that drops out the background color.*

• *If your scanner has a manual brightness control, adjust it so that characters are clean and well formed. If characters are touching because they are too thick, use a higher (brighter) setting. If characters are separated because the characters are too thin, use a lower (darker) setting.*

Note: *The Capture Pages command is designed primarily for black-and-white text, but it can be adjusted to work with color text if there is a high contrast and a minimum of background color or graphics. Capture Pages handles text that is rotated by as much as 7°. For complex color OCR work, see the Adobe Web site (www.adobe.com) for information on the full Acrobat Capture product.*

–From the online Adobe Acrobat User Guide, Chapter 4

Characters that are too thin, well-formed characters, and characters that are too thick

Importing the fax

1 Start Acrobat.

2 Choose File > Import > Image.

3 Open the fax document:

• In Windows, select Fax.tif in the Lesson10 folder, located inside the Lessons folder within the AA4_CIB folder on your hard drive, and click Open.

• In Mac OS, select Fax.tif in the Lesson10 folder, located inside the Lessons folder within the AA4_CIB folder on your hard drive, and click Add. Then click Done.

The fax image file is imported as a new PDF Image Only document.

4 Select the touchup text tool (𝕋), and click in the fax text.

Notice that you cannot edit the text in the document. The TIFF file has been imported as a PDF Image Only file; that is, all elements in the document, including the text, behave as bitmap pictures.

5 Choose File > Save As, rename the file **Fax1.pdf**, and save it in the Lesson10 folder.

You'll convert the file to PDF Normal format using the Capture Pages command. A PDF Normal document contains editable text that can be altered, scaled, and reformatted.

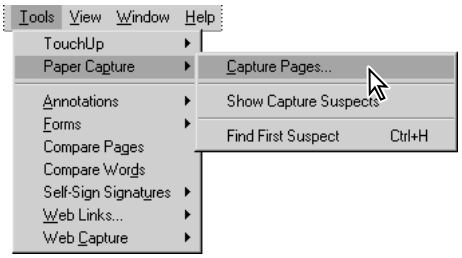
Capturing the fax image

1 Choose Tools > Paper Capture > Capture Pages.

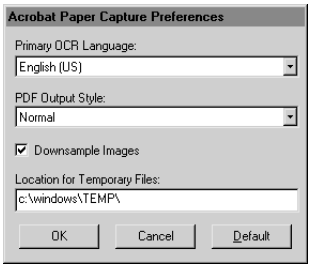
You can restrict the capture to certain pages of the document and specify other capture settings.

2 Select Current Page to capture the page currently displayed on-screen.

3 Click Preferences. In the Preferences dialog box, choose English (US) for the language and Normal for the style.



Choose Capture Pages command.



Set preferences.

4 Click OK to exit the Preferences dialog box, and click OK again to capture the fax document.

5 Select the touchup text tool (𝕋), and click in the document. Notice that you can now edit all horizontal text. The vertically oriented text along the side of the page is still treated as a bitmap image.

About PDF file types

You can use Acrobat with a scanner to create a PDF file from a paper document. The resulting file is a PDF Image Only file—that is, a bitmap picture of the pages that can be viewed in Acrobat but not searched.

If you want to be able to search, correct, and copy the text in an Image Only file, you must "capture" the pages in the file to convert the file to PDF Normal. When you capture pages, Acrobat applies optical character recognition (OCR) and font and page recognition to the text images and converts them to scalable text. You can also convert a file to PDF Original Image with Hidden Text when you capture pages. This type of file has a picture of the pages in the foreground, with the scalable captured text behind it.

PDF Normal files are generally the smallest files, making them ideal for online distribution. PDF Original Image with Hidden Text files are recommended when you need to have searchable text but must keep the original scanned image of a page for legal or archival purposes.

On an Asian-language system, or on a nonnative system with the Asian languages installed, you can scan (but not capture) documents with Asian text.

–From the online Adobe Acrobat User Guide, Chapter 4

Cropping the file

Now you'll crop the captured file to a standard page size.

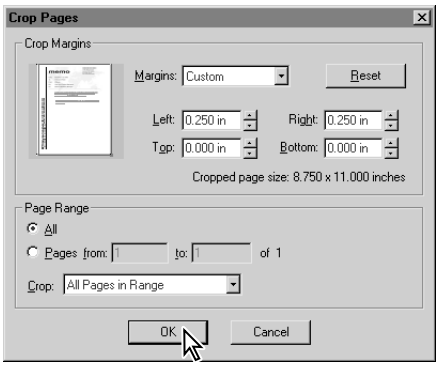
Note: *You must capture a file before cropping or rotating it. If you capture a file after you crop or rotate it, you'll lose the changes you have made to the file.*

1 Click the Fit in Window button ( ).

Look at the page status bar. Notice that the page size is 9-by-11 inches. You'll crop the page to standard letter size, 8.5-by-11 inches.

2 Choose Document > Crop Pages.

3 In the Crop Pages dialog box, enter **0.25** for both Left and Right, and click OK. If needed, click OK again to the confirmation message to crop the page.



4 Choose File > Save to save the Fax1.pdf file.

Correcting suspects

The Capture Pages command converts a bitmap text image into its equivalent text characters. If Acrobat suspects that it has not recognized a word correctly, it displays the bitmap image for the word in the document and hides its best guess for the word behind the bitmap. You can view these *suspect* words in the captured document.

Showing suspects

Choose Tools > Paper Capture > Show Capture Suspects.

The suspect words appear highlighted in the document. Next, you'll examine each suspect and correct or accept Acrobat's best guess for the word.

Correcting suspects

1 Choose Tools > Paper Capture > Find First Suspect.

The original bitmap word appears enlarged in the Capture Suspect window, and Acrobat's best guess for the word appears highlighted in the document. You can correct a suspect word by typing in the desired characters, or you can accept Acrobat's best guess for the word.

2 If needed, click Accept & Next until you arrive at the suspected word "office." If needed, zoom in on the document page to view Acrobat's best guess for the suspect.

Since Acrobat's guess is obviously wrong, you'll type in the correct word.

3 Type **offices** on the keyboard. Notice that the word is updated in the document as you type. Then click Accept (Windows) or Accept & Next (Mac OS) to convert the image text to the word you just typed.

4 Choose File > Save As, make sure that Optimized is selected, and save Fax1.pdf in the Lesson10 folder. Click Yes (Windows) or Replace (Mac OS) to confirm replacing the file. The Save As command lets you save a smaller, optimized version of your finished file.

5 Choose File > Close to close the file.

Converting a Web page to PDF

You can use Acrobat to download pages from the World Wide Web and convert them to PDF. You can define a page layout, set display options for fonts and other visual elements, and create bookmarks for Web pages that you convert to PDF.

Because captured Web pages are in PDF, you can easily save, distribute, and print them for shared or future use. Acrobat gives you the power to convert remote, minimally formatted files into local, fully formatted PDF documents that you can access at any time.

Configuring your Internet or proxy settings

Before you use Web Capture, you must configure your Internet or proxy settings for access to the Word Wide Web.

1 Choose File > Preferences > Internet Settings.

2 Do one of the following:

• In Windows, click the Connection tab in the Internet Properties dialog box, and provide the necessary information for your setup. Your system administrator or ISP will give you the information you need.

• In Mac OS, select Use an HTTP Proxy Server, and then enter your proxy URL and port number in the text boxes.

In Windows, if you do not configure your Internet settings using the Internet Settings preferences, Internet Explorer must be installed and the Internet Properties dialog box configured to allow access to the World Wide Web. In particular, the Proxy Server box on the Connection tab must have a valid proxy address if you are accessing the Web through a firewall in an enterprise environment. Once Internet Explorer has been installed and configured, you may use any browser as your default browser. If your version of Internet Explorer does not have an Internet Properties dialog box, you must upgrade to a current version of Internet Explorer (available from the Microsoft Web site).

Setting options for converting Web pages

You set options for capturing Web pages before you download the pages. Here, you'll set options for the structure and appearance of your captured pages.

1 Choose File > Open Web Page.

Note: *If the Open Web Page command does not appear under the File menu, choose File > Preferences > Web Capture, and deselect Consolidate Menu Items in Top-level Menu. When this option is selected, all commands pertaining to Web capture appear under a separate Web menu.*

2 Click Conversion Settings.

3 In the Conversion Settings dialog box, click the General tab.

4 Under General Settings for Generated PDF, select the following options:

• Create Bookmarks to New Content to create a structured bookmark for each downloaded Web page, using the page's HTML title tag as the bookmark name. Structured bookmarks help you organize and navigate your captured pages.

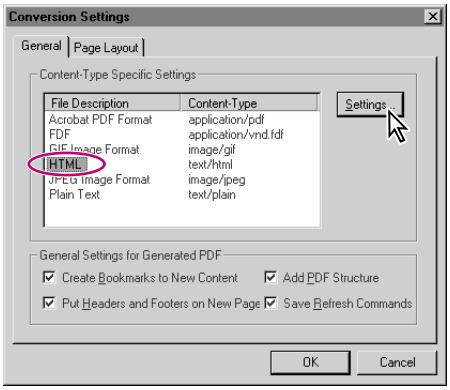
• Add PDF Structure to store a structure in the PDF file that corresponds to the HTML structure of the original Web pages.

• Put Headers and Footers on New Page (Windows) and Put Headers and Footers on New Content (Mac OS) to place a header with the Web page's title and a footer with the page's URL, page number in the downloaded set, and the date and time of download.

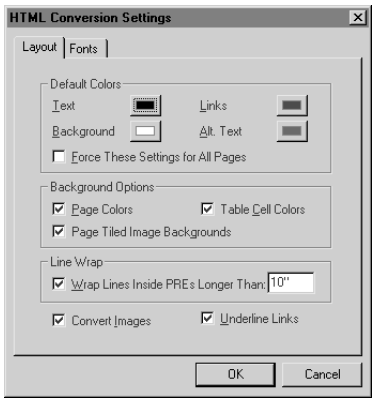
• Save Refresh Commands (Windows) or Save Update Commands (Mac OS) to save a list of all URLs in the PDF file for the purpose of refreshing pages.

5 Under Content-Type Specific Settings, select HTML and click Settings.

6 Click the Layout tab and look at the options available.



Conversion Settings dialog box



Layout options for HTML conversion

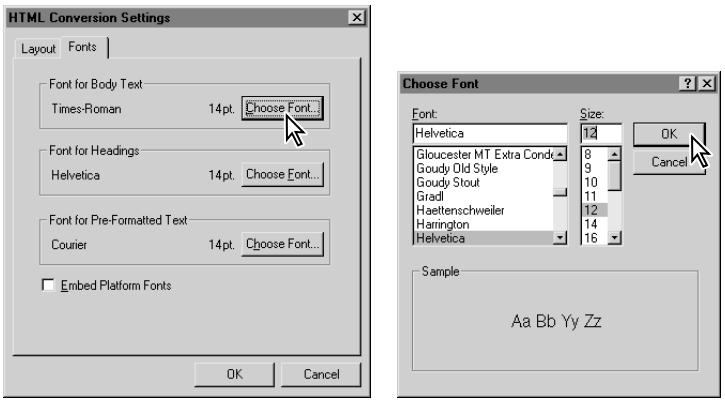
You can select colors for text, page backgrounds, links, and Alt text (the text that replaces an image on a Web page when the image is unavailable). You can also select background display options. For this lesson, you'll leave these options unchanged and proceed to selecting font options.

7 Click the Fonts tab.

8 Under Font for Body Text, do one of the following:

• In Windows, click Choose Font. In the Choose Font dialog box, choose a sans serif font from the Font list. (We chose Helvetica.) Choose 12 from the Size list, and then click OK.

• In Mac OS, choose a font and font size from the pop-up menus.



9 Under Font for Headings, do one of the following:

• In Windows, click Choose Font. In the Choose Font dialog box, choose a thick sans serif font from the Font list. (We chose Arial Black.) Choose 14 from the Size list.

• In Mac OS, choose a font and font size from the pop-up menus.

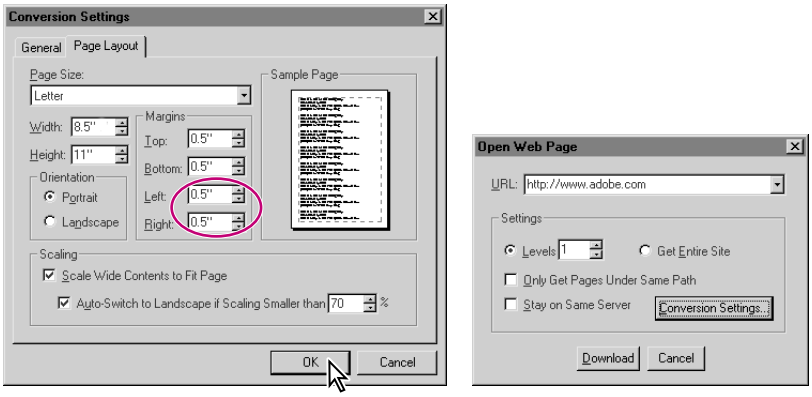
10 Click OK, and click OK again (Windows) to accept the HTML conversion settings.

11 In the Conversion Settings dialog box, click the Page Layout tab.

In Windows, a sample page with the current settings applied appears in the dialog box. You can choose from standard page sizes in the Page Size pop-up list, or define a custom page size. You can also define margins and choose page orientation.

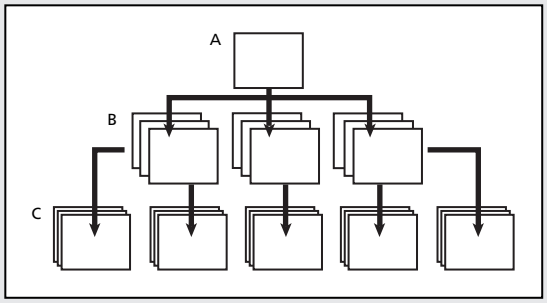
12 Under Margins, enter **0.5** for Left, Right, Top, and Bottom.

13 Click OK to accept the settings and return to the Open Web Page dialog box.



Tips for capturing Web pages

When you capture Web pages, it is important to remember that a Web site can have more than one level of pages. The opening page is the top level of the site, and any links on that page go to other pages at a second level. Links on the second-level pages go to pages at a third level, and so on. In addition, links may go to external sites (for instance, a link at a Web site on tourism may connect to a Web site for a travel agency). Most Web sites can be represented as a tree diagram that becomes broader as you move down the levels.



***A.** First level **B.** Second level **C.** Third level*

Be aware of the following when converting Web pages with Acrobat:

- *Acrobat can download HTML pages, JPEG and GIF graphics (including the first frame of animated GIFs), text files, image maps, and password-secured areas from a Web site.*
- *HTML pages can include tables, links, most types of frames (except certain complex frames, such as cascading stylesheets), background colors, text colors, and forms. HTML links are turned into Weblinks, and HTML forms are turned into PDF forms.*
- *JavaScript cannot be downloaded.*
- *To convert Japanese Web pages to PDF, you must have the Asian system files installed. You must select a Japanese encoding from the HTML conversion settings. This feature is not supported for the other Asian languages.*

***Important:** If you're converting Web sites to PDF, you need to be aware of the number and complexity of pages you may encounter when downloading more than one level at a time. In addition, downloading pages over a modem connection will usually take much longer than downloading them over a high-speed connection.*

–From the online Adobe Acrobat User Guide, Chapter 5

Converting a Web page with Acrobat

Now you'll enter a URL in the Open Web Page dialog box and capture some Web pages.

***Note:** If you are working from within a company network, you may encounter a firewall that limits your access to external Web pages from Acrobat. For instructions on configuring your system to bypass a company firewall, consult your network administrator.*

1 If the Open Web Page dialog box is not open, choose File > Open Web Page.

2 For URL, enter the address of the Web site you'd like to capture. (We used the NASA Mars Global Surveyor site at http://www.jpl.nasa.gov/mgs/overvu/overview.html.)

You control the number of captured pages by specifying the levels of site hierarchy you wish to capture, starting from your entered URL. For example, the top level consists of the page corresponding to the specified URL; the second level consists of pages linked from the top-level page, and so on.

3 Select Levels, and enter **2** to retrieve two levels of pages in the Web site.

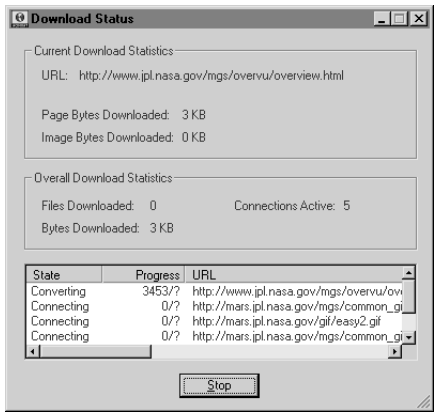
4 Select Only Get Pages Under Same Path to capture only pages that are subordinate to the URL you entered.

5 Select Stay on Same Server to download only pages on the same server as the URL you entered.

6 Click Download. The Download Status dialog box displays the status of the download in progress. When downloading and conversion are complete, the captured Web site appears in the Acrobat document window, with bookmarks in the Bookmark palette. Structured bookmark icons differ from the plain icons for regular bookmarks.

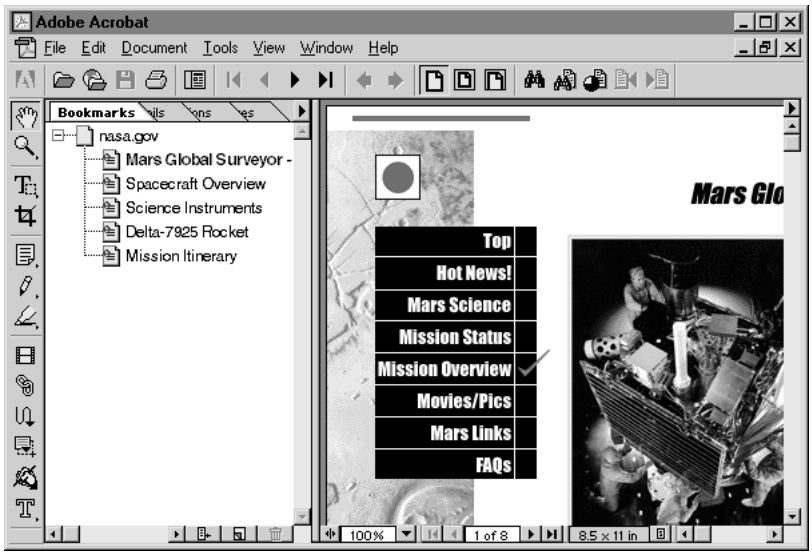
Note: *If you're downloading more than one level of pages in Windows, the Download Status dialog box moves to the background after the first level is downloaded. The globe in the Open Web Page button in the command bar continues spinning to show that pages are being downloaded. Choose Tools > Web Capture > Bring Status Dialogs To Foreground to see the dialog box again. (In Mac OS, the Download Status dialog box stays in the foreground in window shade mode.)*

Specifying URL to be downloaded



Downloading in progress

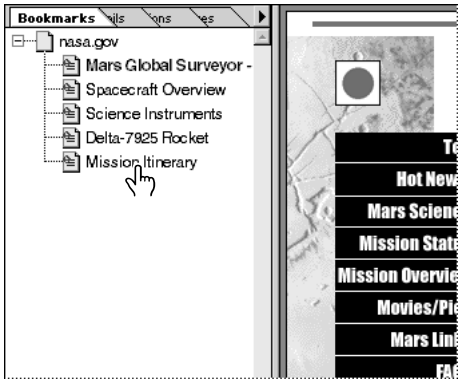
The captured Web site is navigable and editable just like any other PDF document. Acrobat formats the pages to reflect the page-layout conversion settings, as well as the look of the original Web site. Some of the longer Web pages may be spread across multiple PDF pages to preserve the integrity of the page content.



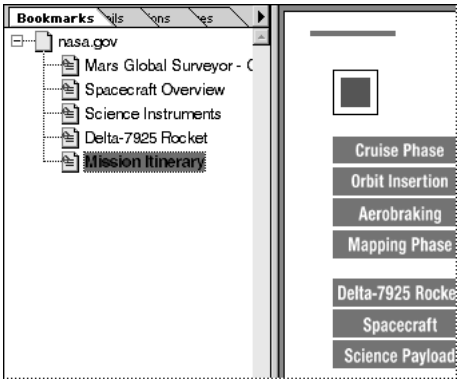
Now you'll use the Bookmarks palette to navigate to another captured page.

7 Select the hand tool (), and click a bookmark in the Bookmarks palette.

The page corresponding to the bookmark appears in the Acrobat window.



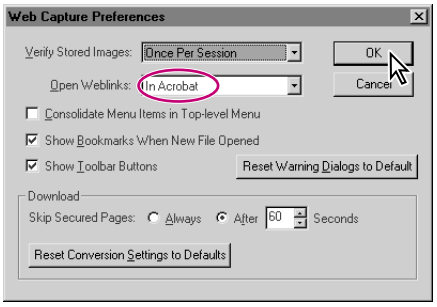
Clicking a structured bookmark . . .



. . . links to the corresponding page.

You can also click a Web link in the document that links to an unconverted page to download and convert that page to PDF. In order to convert linked pages to PDF, you must set Web Capture preferences to open Weblinks in Acrobat (the default setting) rather than in your default browser.

8 Choose File > Preferences > Web Capture.

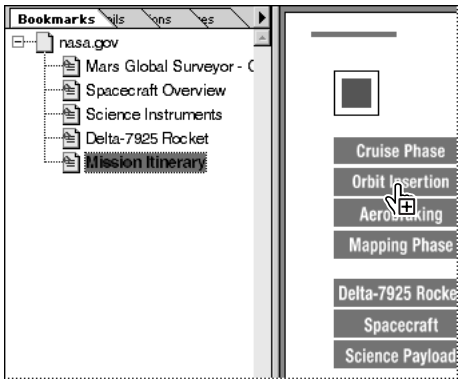


9 For Open Weblinks, choose In Acrobat. Then click OK.

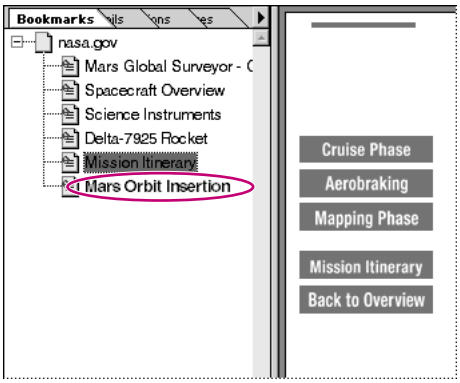
10 Navigate through the captured Web site until you find a Weblink to an unconverted page, and click the link. (The pointer changes to a pointing finger with a plus sign when positioned over a Weblink.)

Note: *If the Specify Weblink Behavior dialog box appears, make sure that Open Weblink to Acrobat is selected, and click OK.*

The Download Status dialog box again displays the status of the download. When download and conversion are complete, the linked page appears in the Acrobat window, with a bookmark for the page added to the Bookmarks list.



Clicking a Weblink . . .



. . . converts the target page to PDF.

11 Choose File > Save As, rename the file **Web1.pdf**, and save it in the Lesson10 folder. Then close the file.

Review questions

1 What are the properties of a PDF Image Only file? A PDF Normal file?

2 How do you make a scanned page searchable and editable in Acrobat?

3 What is a "suspect" word?

4 How do you control the number of Web pages captured by Acrobat?

5 How do you convert destinations of Web links automatically to PDF?

Review answers

1 In a PDF Image file, all the pictures and text are treated as images. You cannot edit the text using the touchup text tool or search the text using the Find or Search command. A PDF Normal file contains searchable and editable text.

2 To convert a scanned PDF Image file to a searchable, editable PDF Normal file, apply the Capture Pages command. This command uses optical character recognition to convert bitmap text to searchable, editable text.

3 A suspect is a word that has probably been recognized incorrectly by the Capture Pages command. Acrobat provides its best guess for the characters in the suspect word and lets you correct its mistakes.

4 You can control the number of captured Web pages by specifying the following options:

- The Levels option lets you specify how many levels in the site hierarchy you want to capture.
- The Only Get Pages Under Same Path option lets you download only pages that are subordinate to the specified URL.
- The Stay on Same Server option lets you download only pages that are stored on the same server as the specified URL.

5 To convert the destination of a Web link to PDF, first choose File > Preferences > Web Capture, and choose In Acrobat for Open Weblinks. Then click the Web link in the PDF file to convert the link's destination to PDF.