

# Working with PDF documents in NVivo

The most recent versions of NVivo (9.1 and later) allow you to work with PDF documents in their original format—when you view PDFs inside the software, they'll look like they do in Adobe Reader. This document provides information about working with different types of PDFs in NVivo.

NVivo lets you directly import and work with PDF documents. This document shows you how.

You'll find out about:

- Working with PDF documents in their original format in NVivo (9.1 and later)
- Different types of PDF files (text-based or image only)
- Working with text-based PDF documents in NVivo
- Working with image-only PDF documents in NVivo
- Importing password-protected PDF files into NVivo
- Scanning documents and optical character recognition (OCR)



If you are using NVivo 9.0 (rather than NVivo 9.1 or later), we recommend you update your software. To update your software, click the **File** tab, point to **Help**, and then click **Check for Software Updates**.

## Working with PDF documents in their original format

In NVivo (9.1 and later), you can work with PDF documents in their original format, as PDF sources. Unlike in earlier versions of NVivo, imported PDFs are not converted to document sources.

Compared to earlier versions, you will notice significant improvements if you work with multi-column documents, or documents containing tables, charts and other graphics. Because your PDF file is not converted during import, its appearance does not change after you import it into NVivo.

When you work with PDF sources in NVivo, you can select text or regions of the page. Just like other sources, you can code, annotate or link selected content—for example, you might select a paragraph of text, or select an area of the page that contains an illustration.

### Note:

PDFs that you have imported into your project as document sources (in NVivo 9.0 or earlier) are not converted into PDF sources when you update to NVivo 9.1 or later. If you want to work with these documents in their original PDF format, you must re-import the PDF files into your NVivo project as PDF sources. If choose to you re-import your PDF files, all coding and other work completed on the documents will need to be done again.

## Different types of PDF files

PDF files can be created by:

- Scanning paper documents with or without optical character recognition (OCR); or
- Publishing an electronic document to PDF format.

What is ‘inside’ a PDF file varies, depending on how it was created:


- When you publish a Microsoft Word document to PDF, the resulting PDF contains text.
- When you scan a paper document, the scanner takes an ‘image’ of the page; each page in the resulting PDF contains a single image.
- When you scan a paper document and use OCR to ‘read’ the scanned image, the resulting PDF contains text.

We can therefore divide PDF files into two types:

Type	Each file contains
Text-based PDF	A series of text elements and (optionally) images
Image-only PDF	A single scanned image per page

You can import and work with both types of PDF files in NVivo, but when you import image-only PDFs you will not be able to code or query the textual content of the documents—see page 4 for more information.

To check whether you have a text-based or image-only PDF:

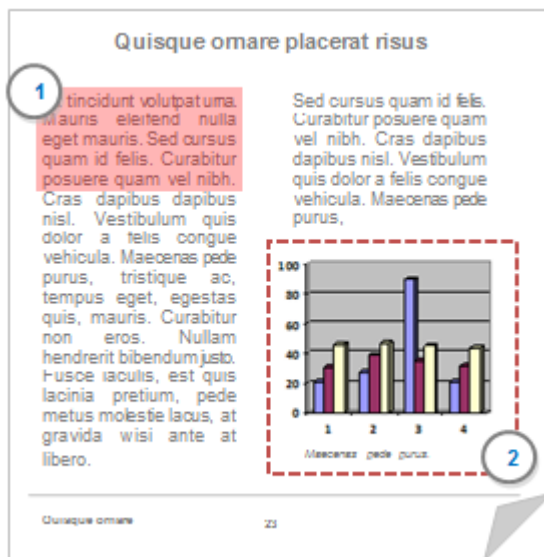
1. Open the PDF file in Adobe Reader.  
**Note:** Adobe Reader can be downloaded free from [www.adobe.com](http://www.adobe.com)
2. Position your cursor over a word, and then double-click. In a text-based PDF, the word will be selected and highlighted. In an image-only PDF, you cannot select an individual word by double-clicking. You may also notice the cursor changes to ‘cross hairs’ (that is; ) when you hover over the text.

### Note:

If you are working with NVivo 10 (or later) and you use NCapture to capture web pages, they are converted to text-based PDFs when you import them into NVivo.

## Working with text-based PDF documents in NVivo

When you are working with a text-based PDF document in NVivo, you can select text or regions of a page:



1. Select portions of text. By default, PDFs open in text selection mode—you can click and drag to select the text you want to code, link or annotate. You can also double-click to select a word and triple-click to select a line.
2. Select regions of a page. When you switch to region selection, you can click on an image to select it, or click and drag to select a rectangular region of the page. When you select a region, you are making an image selection, even if the region you select contains text.

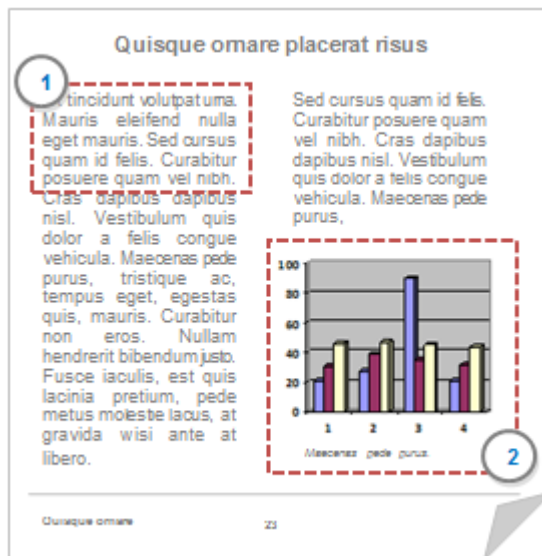
To switch between text and region selection—on the **Home** tab, in the **Editing** group, under **PDF Selection**, click **Text** or **Region**.

### Note:

- Each PDF page consists of text and/or image elements and each element has a specific position on the page. NVivo tries to determine the order of text on the page, however when you extend a text selection (for example, up or down the page), you may find that text is not sequenced as you expect.
- PDF documents can contain custom fonts—for example, a custom font might be used to display a company logo. Text using a custom font may display as red squares when you view the PDF in NVivo.

## Working with image-only PDF documents in NVivo

When you are working with an image-only PDF document in NVivo, you can select and code regions of a page:



1. You can only select text by selecting a region of the page, because each page consists of a single image. When you select a region, you are making an image selection, even if the region you select contains text.
2. You can use region select to select charts and other graphics on the page. When you are in region selection mode, you can click on an image element to select it, or click and drag to select a rectangular region of the page.

By default, PDF sources open in text selection mode—you must switch to region selection, before you can select anything in an image-only PDF. To switch to region selection—on the **Home** tab, in the **Editing** group, under **PDF Selection**, click **Region**.

**IMPORTANT:** You cannot use Text Search or Word Frequency queries to explore the textual content of an image-only PDF, because the PDF does not contain any text. If this is not satisfactory, and you prefer to work with text (rather than images of text), then you could:

- Use optical character recognition (OCR) to convert the image only PDF into a text-based PDF (or a Microsoft Word document) which you can import into NVivo. See the following pages for further information on using OCR.
- Find a text-based version of the document—for example, in Microsoft Word or some other digital format—and import it into NVivo.
- Create a linked memo and type the text that you want to code into the memo—you can then code from the memo (rather than from the PDF source).

## Importing password-protected PDF files into NVivo

PDF files can be secured with a *Document Open* password. When a *Document Open* password is set, the PDF can only be opened in Adobe Reader with the correct password. You will also be prompted to enter this password when you import the document into NVivo. If you do not know the password, you cannot import the document.

To check the security settings of a PDF file:

1. Open the PDF file in Adobe Reader.
2. On the **File** menu, click **Properties**, and then click on the **Security** tab.
3. On the **Security** tab, click **Show Details**.  
The **Document Security** settings are displayed.

## Scanning documents and optical character recognition (OCR)

Many scanners create PDF files by default. You may decide to scan a large volume of documents, with the intention of importing the output PDF files into NVivo. However, before you start scanning documents, you should consider whether you want to use OCR to convert the scanned images into editable text. If you do not use OCR, then the scanner will create image-only PDFs, and you will not be able to code or work with the individual text characters in NVivo.

Some scanners are sold with 'bundled' OCR software or you can purchase the software separately.

If you use OCR software you can:

- Save the output to a variety of file formats, including text-based PDF files.
- Choose to exclude certain portions of the document from the OCR process (for example, the headers and footers, or the table of contents).
- Edit the output before you import it into NVivo.

Because OCR recognition rates vary, it is important to make sure you are satisfied with the results **before** you start scanning large numbers of documents and importing them into NVivo.

OCR technology works best with typewritten, laser printed or typeset text. Neat hand-written text may be recognized reasonably well, but OCR tools cannot handle cursive (joined) writing.

OCR software can also be used to convert existing image-only PDF files into editable text files.

OCR can give very good results, but is dependent on the:

- Print quality of the original document
- Quality of the scanning
- Legibility of any handwriting in the document

OCR products will usually highlight ‘questionable’ words, which might have been incorrectly recognized—you should review and correct these **before** saving and importing the document into NVivo. Of course, you can edit the text in NVivo after importing the document, but it’s best to check the OCR results prior to import.

To get best results when scanning with OCR, you may need to:

- Adjust your scan settings to get a higher quality scan.
- Exclude areas of the scanned document from OCR. For example, handwritten margin notes can be excluded and treated as images.
- Adjust settings in your OCR software to achieve the best recognition results. For example, there may be a trade-off between speed and accuracy, or other ways to improve recognition performance.
- Review the OCR text output to check suspect words and correct recognition errors.

**Note:**

Microsoft OneNote (included with some editions of Office 2007, 2010 and 2013) provides OCR functionality that allows you to extract text from pictures—this may be useful if you are working with a small number of scanned pages.

Microsoft Office (2003 and 2007) includes the document scanning/OCR tool *Microsoft Office Document Imaging*—for more information refer to: <http://office.microsoft.com/en-us/help/HP010771031033.aspx>