Raster, Vector and Text – What’s Really in My PDF?

Summary

How to determine whether your PDF is raster- or vector-based, and how this affects the ability to snap to an object or select text.

Problem

You’re unable to snap to an object when taking measurements.
You cannot edit text using Edit > PDF Content > Edit Text .
You cannot select or search for text.

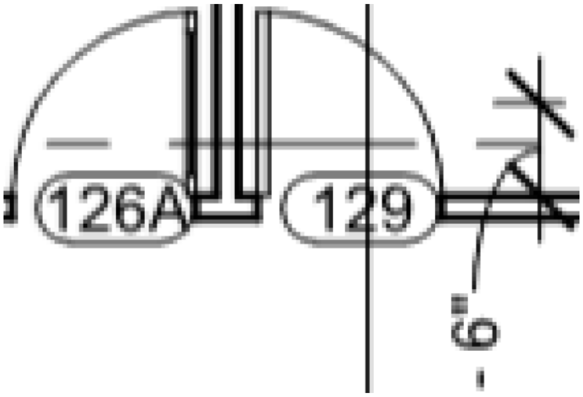
Why does this happen?

The reason both of these occur is that PDFs aren’t all created in the same way. Some PDFs contain more information than others, even though they seem indistinguishable at first. The page may appear to contain lines and characters, but the underlying elements that represent them in the PDF may not be vector lines and text elements, which are needed to snap to content and search and select text.

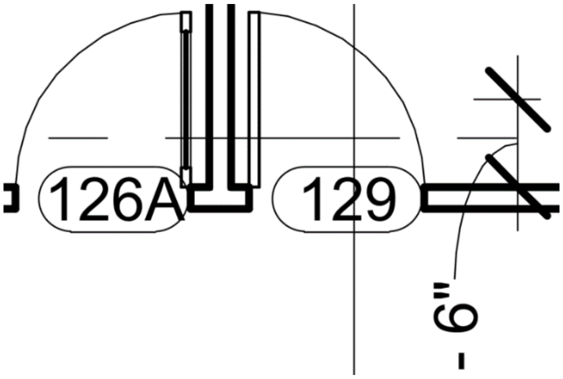
Raster vs. Vector Content

Let’s look at the difference between raster and vector content in a PDF.

RASTER PDF



VECTOR PDF



Raster

A raster image is created from a series of square dots called pixels. One example of a raster PDF is a file created from scanning a paper. A scanned PDF is created by making a bitmap image (like a JPEG or TIFF) of the page, and placing that image on the PDF page. This means that a scanned or raster PDF only contains a grid of dots that represent lines and text; it does not actually contain lines or text that a computer can recognize. Therefore, there are no lines for the Snap to Content function to snap to, and no text to select or search.

To determine if a PDF is a raster image, or scan, just zoom in very closely. The lines and characters on the page will either change to a grid of square dots or become blurry.

Vector

A vector-based PDF uses line segments to define all of the geometry on the page. Most PDFs created from CAD (Computer-Aided Design) are vector-based. Vector PDFs are usually preferred to raster PDFs because they contain more data that make it easier to work with. You should always try to work with vector PDFs created from the source instead of creating PDFs from scans.

The benefit of working with a vector PDF is that the display of the geometry remains sharp when you zoom in to see details of the drawing. As such, measurements and takeoffs (as well as their calibration) are precise in a vector PDF because you can use Snap to Content to snap to the vector lines in the PDF.

Text

Text is an independent type of content in PDFs. You may see text characters in the PDF, but those characters are not necessarily PDF text elements. Instead, it might be defined by raster dots or vector line segments. Although these elements appear to be text, they do not have the data that allows a computer to recognize it as text. As such, this type of "text" is essentially an image that cannot be selected, searched, or edited.

To avoid confusion, "characters" will refer to text in general while "text" will refer to PDF text elements, or "real text." Before going into details, there is a quick test to determine if your PDF contains text. From the Menu Bar, go to Edit > PDF Content > Select All Text. All text in the PDF should highlight in blue. If the characters don't highlight, they are either a raster or vector image.

Highlighted Text

PDF Text Elements (or Real Text) – Always preferred for PDFs because it results in more responsive content. PDFs created from character-based programs (e.g., Word® and Excel®) almost always create PDFs that contain real text. When you zoom in on the text, the edges of the characters always look sharp and crisp – no matter how close you zoom in. The text is searchable and can always be selected.
Optical Character Recognition (OCR) Text – Running OCR (for Revu eXtreme only) allows for the translation of raster and vector images into searchable data. In other words, OCR interprets the images on a scanned PDF and creates an invisible text layer on top of them. This layer is what allows you to search, select, and highlight images that don't have real text.

Vector Characters – Created by line segments that are used to draw the shape of each character. This usually occurs when the PDF has been created from CAD (often AutoCAD®) or a non-TrueType font is used.

Why doesn’t CAD use TrueType fonts to create real text? The answer is because AutoCAD predates Macintosh®, Windows®, and TrueType fonts. They needed to create their own system of fonts, called SHX fonts. SHX fonts are defined using line segments. Those line segments are translated into the PDF instead of text data.

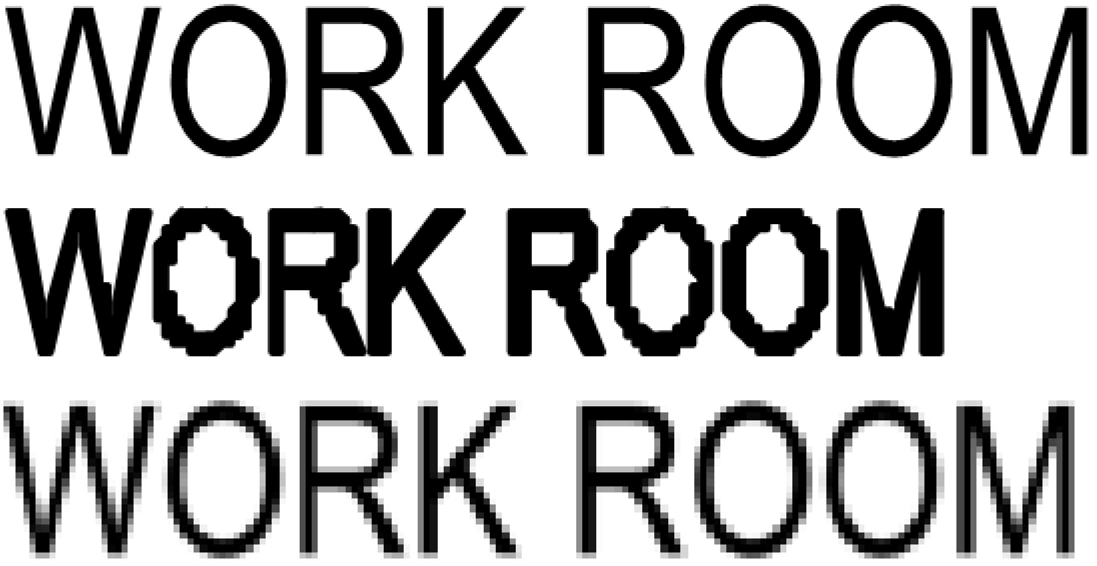
Using TrueType fonts in CAD is preferable for creating PDFs.

Vector characters are distinguished by their lumpy appearance when zoomed in. These lumps are created by the line segments that make up each character.

Graphic design programs (e.g. Adobe Illustrator®), also create vector characters. However, these vector characters have clear, sharp edges when zoomed in.

Raster Characters – As mentioned earlier, individual pixels are used to define each character.

Examples of characters that are text, vector and raster, respectively.


WORK ROOM
WORK ROOM
WORK ROOM