PDF2DTP CHARACTERS EDITOR

Sometimes a PDF will use an embedded font which does not define specific unicode characters making it virtually impossible to convert the text. The PDF can be displayed or printed because the font uses various drawing commands (bezier curves) that renders a character (known as "glyph") as if it was a small "graphical image". Technically, a "text stream" inside the PDF is comprised of various "codes" and these values are sent to the embedded font which in turn looks up the commands in order to draw the glyphs. Most PDFs either have a specific font encoding for the codes or a "ToUnicode" table that can be used to lookup and map codes to unicodes. Unfortunately, some PDFs do not contain the additional information required to be able to convert the codes to unicode.

An interesting test you can perform to prove this problem is to open the "problem" PDF in an application such as Adobe Acrobat, then copy and paste the text into a Text Editor and you will see "garbage" characters displayed. This is because the characters are essentially raw codes from the text stream (which are either Glyph IDs or indices into the Glyph drawing tables). The only solution to convert (or extract) the text from such PDFs is to perform advanced OCR (Optical Character Recognition), but even then the accuracy of obtaining the exact unicode values may not be 100% accurate.

Consequently, when the unicode values are unknown, PDF2DTP will substitute the characters with either a tilde "~" or a space (depending on your PDT2DTP Preferences setting). The tilde characters are therefore used as "markers" so that after the conversion is complete you can then use Find/Change and search for the tilde characters in order to manually edit them. Sometimes it is difficult to locate a tilde character, especially if it exists within overset text. One helpful tip is to select "Edit in Story Editor" (under the Edit menu) and you should then be able to see its exact location in the story.

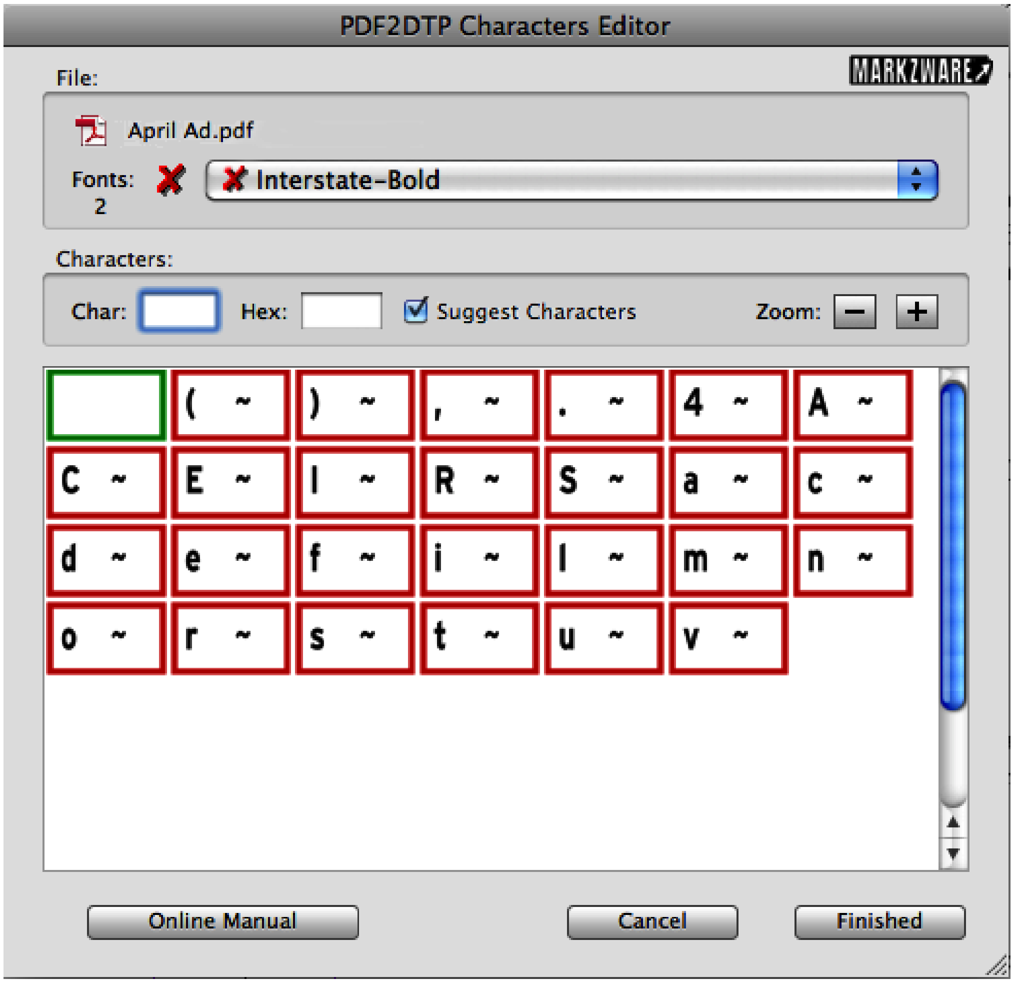
However, PDF2DTP offers a solution that allows you to define the unknown characters, thus avoiding the tilde substitutions. This process uses the **PDF2DTP Characters Editor**.

An embedded font within a PDF is usually a sub-set of the characters of the original font and consists of only those characters which are actually used in the text stories. For example, if the font is used for the word "hello" then only four characters need to be defined: "h", "e", "l" and "o". There are usually only a small handful of characters that require editing, but sometimes there can be a fairly large quantity, and some fonts can use dozens of embedded fonts which do not define unicodes. While the task of editing all the characters may seem tedious, the process pales in comparison to having to retype the entire stories. Besides, once you've edited the characters the information will be saved to disk so that it can be used in subsequent PDFs that reference the same font.

Therefore, the PDF2DTP Characters Editor allows you to achieve highly accurate text conversion, that is of course, depending on how thorough and exact your editing is.

Edit Unknown Characters

To be able to edit the unknown characters before a conversion select the "Edit Unknown Characters" checkbox on the PDF2DTP Preferences window. This will instruct PDF2DTP to scan the PDF at the start of the conversion and if it encounters any unknown or undefined characters it will display the "**PDF2DTP Character Editor**":



PDF2DTP Characters Editor

File: April Ad.pdf

Fonts: 2 Interstate-Bold

Characters:

Char: Hex: ☑ Suggest Characters

The **PDF2DTP Characters Editor** is a very simple yet powerful tool to assist you in editing a PDF which has undefined or unknown unicodes.

File Area:

At the top-left of the Characters Editor window is the File area which will display the **name** of the PDF. To reveal the file in the Finder click on the PDF's name.

Underneath the file name is the **Fonts menu** which will list only those fonts which are embedded in the PDF that have undefined or unknown unicodes that require editing. Note that the **number** for the total amount of fonts on the menu will be conveniently displayed which is how you can tell when there is more than one font. (The red "X" will be explained in a moment).

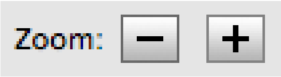
Characters Area:

The Characters area consists of a “**Char**” field, which is for typing in the desired character to substitute, and the “**Hex**” field, which is for entering a Hex unicode value in case you’d prefer that method.

Character Cells:

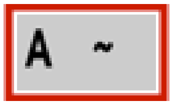
Although an embedded font may contain a large number characters, only those characters actually used in the text stories of the PDF which have no defined unicode will be displayed in the Character Cells. Each Character Cell will display the character (glyph) from the PDF on the left (which is guaranteed to be precise) and then on its right will be the displayed edited unicode character (which will subsequently be used in the final text for the document stories).

Click the plus “+” or minus “-” buttons to increase or decrease the display of the Character Cells:



Selecting a Character Cell:

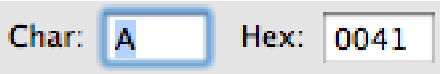
Simply click on a Character Cell to select it. The Cell will turn gray:



You can also select a Character Cell by holding down the command key and pressing the right or left arrow key, or the up or down key to move to the next or previous cell.

Editing a Character Cell:

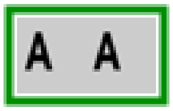
Select a Character Cell and type in the desired character. Below is an example of setting the Character Cell to an “A”. Or you can press the Tab key and type in the Hex field 0041 which is the equivalent for the letter “A”. (See the later discussion on Hex values).



Char: A Hex: 0041

To type an accented or alternate version of a character, hold a key down until its alternate characters are displayed. To choose one of the characters displayed, type the number that appears under the character, or click the character you want to use. (Note: You may need to check the Keyboard pane of the System Preferences to make sure that the Key Repeat slider is set to On).

When ready, press the Return key to accept the Character (which will now be displayed in the cell on the PDF character’s right). The Character Cell will then be displayed with a green frame which indicates it has been edited:



Note that using the Return key will conveniently move you to the next Character Cell. Use the Enter key instead to accept the Character and remain in the current cell.

Entering Characters or Hex Values:

One way to obtain the desired character is to search the internet, for example **http://www.unicode.org/charts**, and locate the value of the desired unicode and then enter its 4 numbers into the PDF2DTP Characters Editor **Hex** text edit field.

An even more helpful method, especially when you don't know how to type a specific character, is to go to the System Preferences and select "Keyboard" and "Show Keyboard and Emoji and Symbols in menu bar" in order to select the Viewers from the Finder.

Then you can bring up the System Keyboard Viewer, locate the desired character and single-click in order to insert it into the PDF2DTP Characters Editor **Char** text edit field, then press Return to accept.

Or, bring up the System Emoji and Symbols Viewer, locate the desired character and double-click in order to insert it into the PDF2DTP Characters Editor **Char** text edit field, then press Return to accept. Or, simply drag the character from the System Emoji and Symbols window to the PDF2DTP Characters Editor **Char** text edit field.

To obtain a Hex value, highlight the 4 numbers shown in the System Emoji and Symbols window in the "Character Info" section, then copy and paste into the PDF2DTP Characters Editor **Hex** text edit field (or drag the highlighted numbers).

If unicode values aren't showing on the System Emoji and Symbols window, locate the gear icon at the top-left, click and then select "Customize List". In the dialog that appears, scroll to the bottom and check "Unicode" under the "Code Tables" branch.

Color Status:

The best way to understand how the Characters Editor works is to always keep in mind it uses a convenient method of displaying a colored frame around a Character Cell to represent its "**Status**". When a PDF is scanned and unknown characters are discovered and the Characters Editor appears, each Character Cell will have an initial status drawn with a red frame with a tilde "~" (or space) for the suggested unicode replacement.

Green indicates "**Edited**": The Character Cell has been edited.

Red indicates an "**Unedited**": The Character Cell has not yet been edited.

Statuses:

Simply stated: the goal is to get all the of the Statuses indicators to be green.

A **Status icon** will be displayed for the overall status of the currently selected font in the File area at the top-left of the window.

✔ A green checkmark indicates all cells have been edited.

✘ A red "X" indicates there is at least one Character which has not yet been edited.

In other words, you can visually inspect the **Status icon** to know what additional editing needs to be addressed. For example, if the Fonts area shows a red "X" but the currently selected font has a green checkmark then you will know that one of the other fonts on the menu has Character Cells which have not yet been edited in which case you would need to switch to the other font to edit those Character Cells.

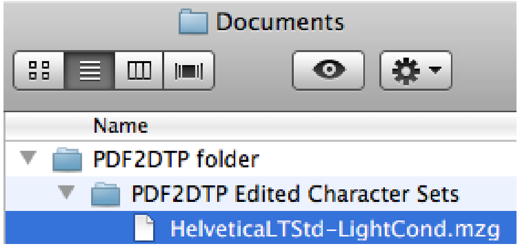
Suggested Characters:

The “**Suggest Characters**” checkbox will become enabled whenever PDF2DTP has determined there is a high confidence level in suggesting “best candidates” for the characters. This usually occurs when the font referenced by the PDF is active in the System, otherwise the checkbox will be disabled for the font.

When you select the “**Suggest Characters**” checkbox then the “best candidates” will be conveniently displayed instead of tildes. Keep in mind that these suggestions will then be used for the converted text stories. It is therefore important to always first examine each suggested character to see if it is correct (for example, **á** versus **à**) and if ok you can then click on the cell, edit the character if needed, then press Enter or Return to accept (which will set the frame to green). Because all of the displayed suggested characters will be used in the converted text stories, whether or not they have been “accepted”, and in order to help avoid using undesirable suggested characters, the default setting for the checkbox will be initially set to “off”.

Finished Button:

When ready, click the “Finished” button. Doing so will save the information to a file on disk which will be located in the “**/Users/{username}/Documents/PDF2DTP folder/PDF2DTP Edited Character Sets**” folder with the font type appended to the filename (ie: “T0” for Type0, “T1” for Type1, “TT” for TrueType, etc.).



When you perform a conversion in the future on a PDF and the associated Edited Characters Set file exists for the same font, then PDF2DTP will conveniently use the file’s edited unicodes (instead of replacing the unknown characters with tildes or spaces). If you discover you had entered a wrong value, as evidenced by a “typo” in the converted document, you can always reconvert the document, bringing up the editor and making the correction to the character. Keep in mind that if you enter unicodes not supported by either the font or a substituted font, the characters will appear as little “rectangles” in the text box. In this case you will need to either re-edit the characters to more suitable unicodes, or change the font in the document.

Always Show the Characters Editor

If PDF2DTP scans a PDF and determines there are unknown or undefined unicodes and the associated Edited Characters Set file exists for the font, then it will use the information in the file without bringing up the Characters Editor. However, if you select the “**Always Show the Characters Editor**” Preference then this will instruct PDF2DTP to always bring up the Characters Editor even when an Edited Character Set file exists. This allows you to review the characters before continuing the conversion.

Another way to force the Characters Editor to appear is to delete, move or rename the specific Edited Characters Set file in which case PDF2DTP will bring up the Editor allowing you to create a new set.

If you proof read the converted document and discover a character is incorrect then what you can do is make note of the font name (provided it wasn’t Substituted), re-convert the PDF with the “Always Show the Characters Editor” turned on (which will force the PDF2DTP Characters Editor to appear), select the font, find the Character Cell in question, edit the unicode, then click the “Finished” button to save the information. The conversion will then continue and use the corrected character.

Other Problem Fonts

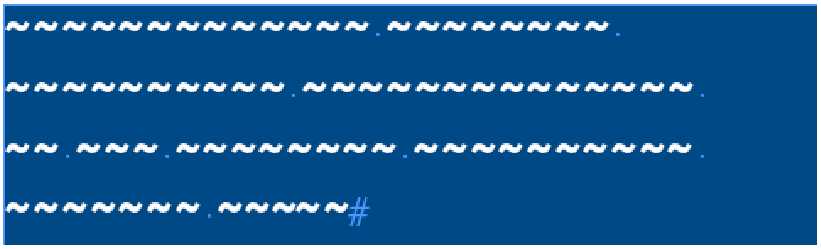
A Type3 font is a special case where the “glyph” is commonly a logo or barcode which makes it extremely difficult to know which unicode to enter in the Characters Editor. In this case, one special trick you can employ is to literally place the original PDF (by using InDesign’s File:Place... menu item and choosing the specific page number when importing) at the desired location on the page and then crop the picture box so that just the logo or barcode area of the PDF is viewable. Or you can edit the PDF in another application, such as Adobe Acrobat, Adobe Illustrator or Adobe Photoshop, by deleting unwanted pages and removing all of the other objects on the page, leaving just the logo or barcode, then save to a new PDF and place it in the document at the proper location.

Encrypted PDFs

Secured or encrypted PDFs which require entering a password to be able to convert, sometimes use “scrambled” definitions of unicode values to prevent copying the text. Unfortunately, there isn’t an easy way to convert the characters short of converting each page to a TIFF image and then using advanced OCR (Optical Character Recognition) to extract the text stories.

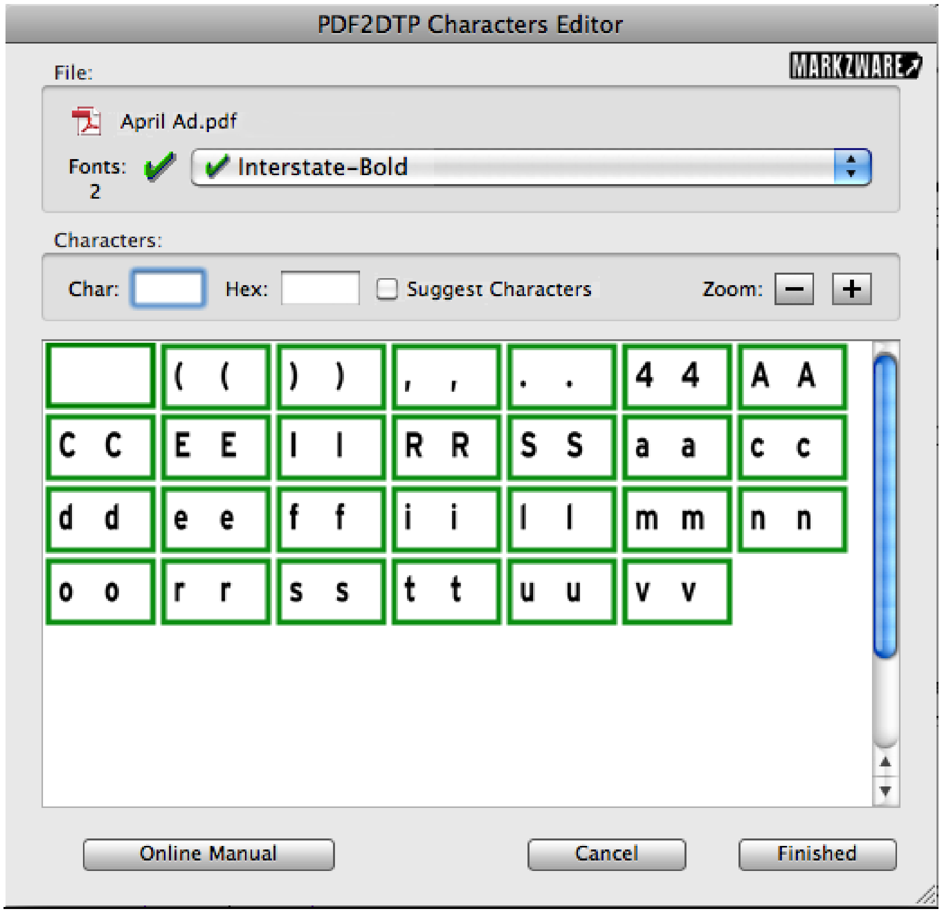
Example Conversion

Below is the result of converting a PDF with unknown or undefined unicodes. You will see that each character in the text box has been replaced with a tilde:



~~~~~~~~~~~~~ ~~~~~~~~
~~~~~~~~~~ ~~~~~~~~~~~~~~
~~ ~~~ ~~~~~~~~ ~~~~~~~~~~
~~~~~~~ ~~~~~#

After selecting the "Edit Unknown Characters" Preference and converting the PDF again, the Characters Editor will come up and then each Character Cell can been "Edited" until all cells are set to a green "Edited" status:



PDF2DTP Characters Editor

File:

April Ad.pdf

Fonts: ✓ 2 ✓ Interstate-Bold

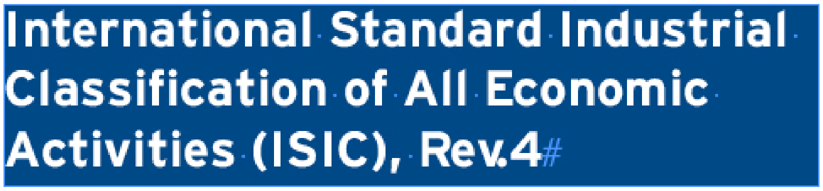
Characters:

Char: Hex: □ Suggest Characters Zoom: − +

	((	))	, ,	. .	4 4	A A
C C	E E	I I	R R	S S	a a	c c
d d	e e	f f	i i	l l	m m	n n
o o	r r	s s	t t	u u	v v	

Online Manual Cancel Finished

Upon clicking the "Finished" button, the conversion results in the text using the edited unicodes:



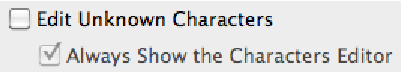
International Standard Industrial
Classification of All Economic
Activities (ISIC), Rev.4#

Therefore, the accuracy of the text conversion will depend upon the thoroughness and accuracy of your editing.
~~~~~~~~~~~~~

Characters Editor Tutorial

Here is a step-by-step example of how to use the Characters Editor:

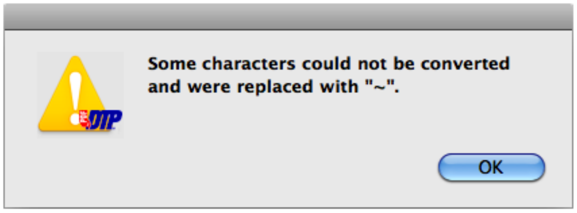
1. First, bring up the PDF2DTP Preferences and turn off the “Edit Unknown Characters” checkbox:



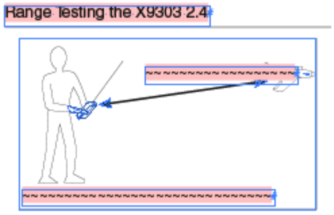
☐ Edit Unknown Characters

☑ Always Show the Characters Editor

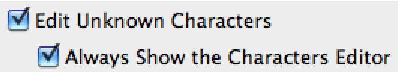
2. Convert the “CHARACTERS EDITOR TEST.pdf”. You will receive a warning that there were unknown characters replaced with a tilde (which you can see in the text box):



Some characters could not be converted and were replaced with "~".



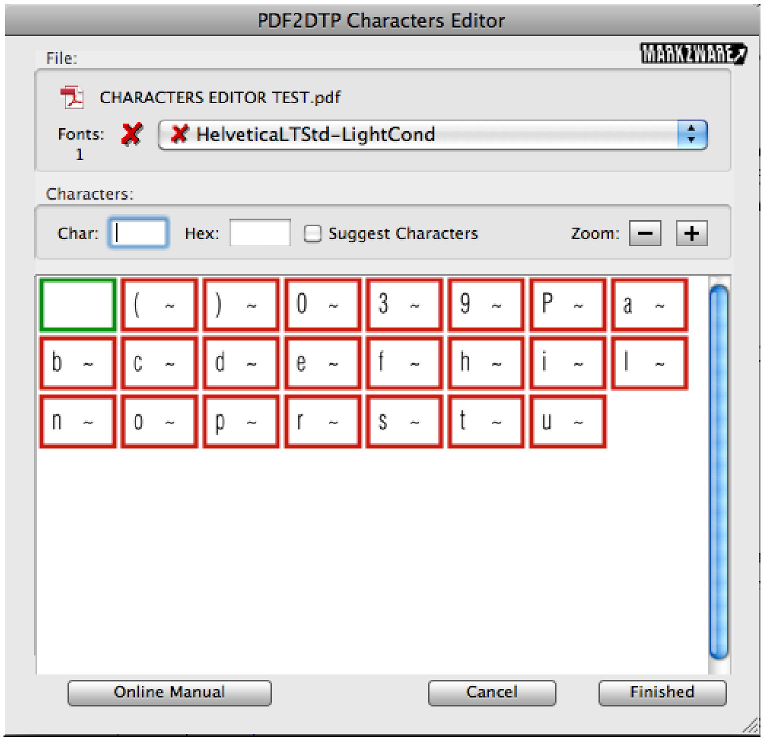
3. Bring up the PDF2DTP Preferences again and turn on the “Edit Unknown Characters” checkbox:



☑ Edit Unknown Characters

☑ Always Show the Characters Editor

4. Convert the “CHARACTERS EDITOR TEST.pdf” again and the Characters Editor window will appear:



PDF2DTP Characters Editor

File:

CHARACTERS EDITOR TEST.pdf

Fonts: 1 HelveticaLTStd-LightCond

Characters:

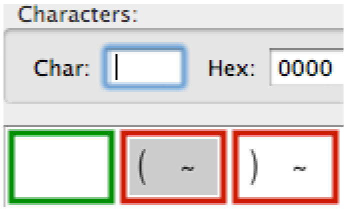
Char: Hex: ☐ Suggest Characters Zoom: − +

	(~	) ~	0 ~	3 ~	9 ~	P ~	a ~
b ~	c ~	d ~	e ~	f ~	h ~	i ~	l ~
n ~	o ~	p ~	r ~	s ~	t ~	u ~	

5. At the top-left you will notice there is a red “X” next to the font name. This indicates there are unknown characters for at least one of the fonts on the menu that need to be edited.

6. Notice the first Character Cell at the left is already framed in green. This indicates an “Edited” status because the character is actually a space and is therefore OK.

7. Notice the second Character Cell is framed in red indicating the unicode is unknown. Click on the cell to select it and it will turn gray:

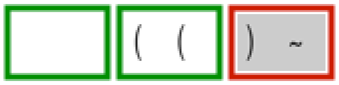


Characters:

Char: | Hex: 0000

(~) ~

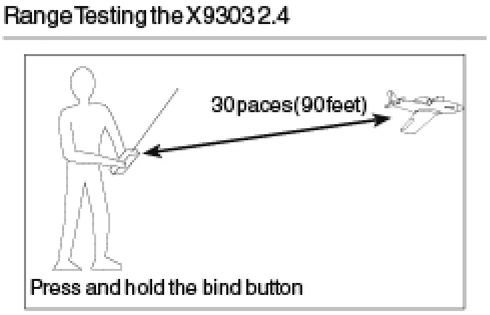
8. With the Character Cell selected, type a left parenthesis character in the Char field and press the Return key. The cell will become framed in green indicated it has been “Edited” and then the next cell on the right will become conveniently selected.



(( ) ~

9. Continue in this manner of entering the correct replacement character and pressing Return until all of the Character Cells have been edited, that is all cells have green frames.

10. Once all the cells have been edited, click the “Finished” button to continue the PDF conversion process and you will see the text in the document will use the replacement characters that you had entered:



RangeTestingtheX93032.4

30paces(90feet)

Press and hold the bind button

11. The edited unicodes information will be saved to disk so that subsequent conversions for PDFs that use the same font will use the edited characters. To test this, try converting the test PDF again. The Characters Editor will appear and show all Character Cells framed in green. Click “Finished” to continue and you will see the text stories are correctly converted. Then, as a final test, go to the PDF2DTP Preferences and uncheck “Always Show the Characters Editor” and convert the test PDF again. The document will again appear with the correct text, but this time without the Characters Editor ever coming up.