Comparing Scanned PDF Documents: Adobe Acrobat Pro DC, Nuance PowerPDF, Foxit Phantom for Business and ABBYY PDF Transformer

Karen McCall, M.Ed.

Table of Contents

Documents used for this Tutorial 3

Potential Problems 3

The Salamanca Statement (Original) 4

Adobe Acrobat Pro DC (First Attempt) 4

ABBYY PDF Transformer (Second Attempt) 6

Starting from Scratch 6

Adobe Acrobat Pro DC (Colour Document) 8

ABBYY PDF Transformer (Colour Document) 11

Foxit Phantom for Business (Colour Document) 14

Nuance PowerPDF Advanced (Colour Document) 22

Switch to Greyscale! 26

Adobe Acrobat Pro DC (Greyscale Document) 26

ABBYY PDF Transformer (Greyscale Document) 27

Foxit Phantom for Business (Greyscale Document) 29

Nuance PowerPDF (Greyscale Document) 31

Summary 36

Appendix A: List of Attached Documents 38

TIFF Documents **Error! Bookmark not defined.**

Colour Copy of Page 6 38

Black and White (Greyscale) 38

Three pages of The Salamanca Statement 38

Black and White Versions of the Three Pages 38

Colour Versions of the Three Pages 38

The Complete Salamanca Statement Framework Document 39

Appendix B: Contact Information 40

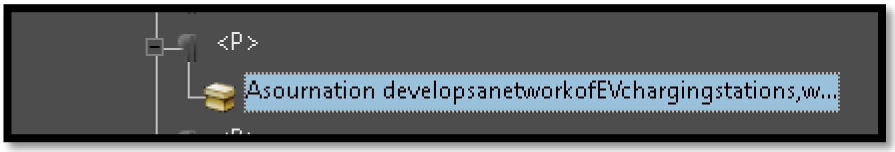
Documents used for this Tutorial

I've attached the original document entitled "The Salamanca Statement and Framework for Action " of 1994and all of my test files so you can explore them. There is a list of what is attached to this PDF tutorial in Appendix A. There is a link for the original document in Appendix A.

Note: After all of this testing, this, I still haven't read the entire document. I gave up. But it is a useful test document.

Potential Problems

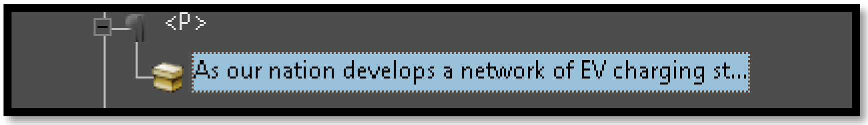
One of the most common problems we see in legacy PDF and/or scanned documents that go through OCR is that tagging tools will not put spaces between words. This is primarily due to the types of fonts used in the document, when the document was created and with what authoring tool was used. For example in a simple text document, the tagging tool in Adobe Acrobat Pro DC produced text where there were no spaces between words once the document was tagged.



Asournation developsanetworkofEVchargingstations,w...

Figure 1 Paragraph where all the words are together with no spaces between them.

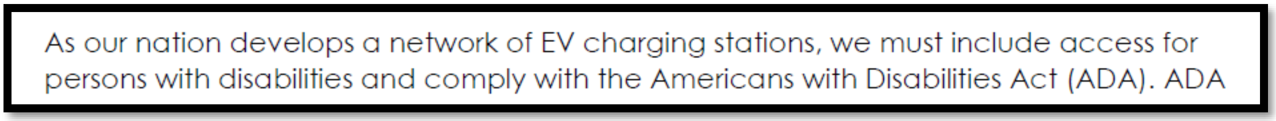
In the same simple text sample document, when I closed out of the document without saving, opened the same PDF document in ABBYY PDF Transformer and made the text searchable/performed their OCR, the results were that the words did have spaces between them once I took the now searchable PDF back to Adobe acrobat Pro DC. The document wasn't a scanned image, but making the text searchable in ABBYY PDF Transformer seemed to help make the content readable.



As our nation develops a network of EV charging st...

Figure 2 Same text in PDF document after going through ABBYY PDF Transformer and Acrobat.

The following image is of the text in the document showing that it is just plain text with nothing fancy and a pretty good quality scanned image.



As our nation develops a network of EV charging stations, we must include access for persons with disabilities and comply with the Americans with Disabilities Act (ADA). ADA

Figure 3 The text for the misrecognized text with no spaces between words.

Part of the problem we face is that we can’t test this text using a screen reader or Text-to-Speech tool until the document is tagged. It is the Tags that give us access to the text. Until we Tag the document and access it as someone who uses a screen reader or Text-to-Speech tool, we are only accessing the visual representation of the document. As we see from the previous image, visually this text can be read.

The Salamanca Statement (Original)

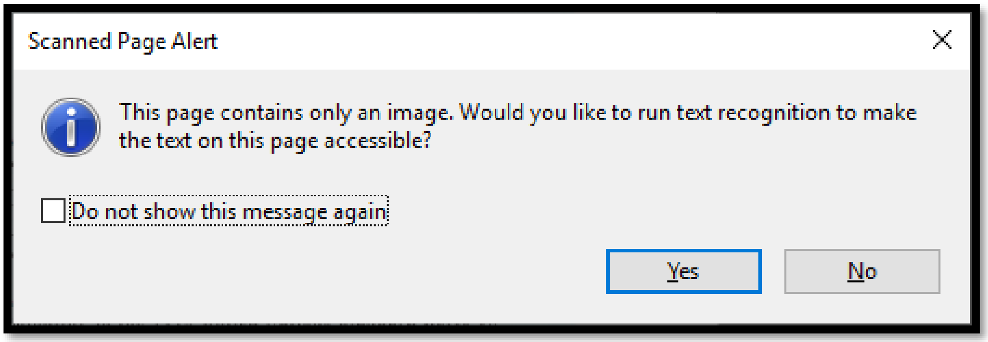
The Salamanca Statement makes the perfect document for this process as it was a PDF from 1994 and is composed of bits of other documents as well as original content. The Salamanca Statement, as a bit of background, was a global statement supporting student centred education on an international level. In 2019, there was a conference in the US to take a look at any progress we’ve made toward that goal. This IS a document I had to try and find a way to read in order to participate in the discussions.

Adobe Acrobat Pro DC (First Attempt)

When I downloaded the document, my screen reader told me it was an untagged document. I immediately pressed the Escape key as I didn’t want the “”this is an untagged PDF, do you want me to try and figure it out” to happen. As of Adobe Acrobat and Reader 6, we have the concept of “trusted adaptive technology” which is an on-board tool for legacy untagged PDF. To be clear, this is not a work around or end run to avoid making accessible PDF documents!

If a document is a scanned image, we get a message telling us this and “Virtual OCR” tries to make the document available for “Virtual Tags.” Neither of these are permanent. Each time you open a document, you can get a different result in terms of the level of access to the content of the document. Only performing “real OCR” and adding “real Tags” will make a scanned image of a document and/or an untagged document accessible and give consistent levels of access to the content.

The following images illustrate the messages we get if we are using adaptive technology.

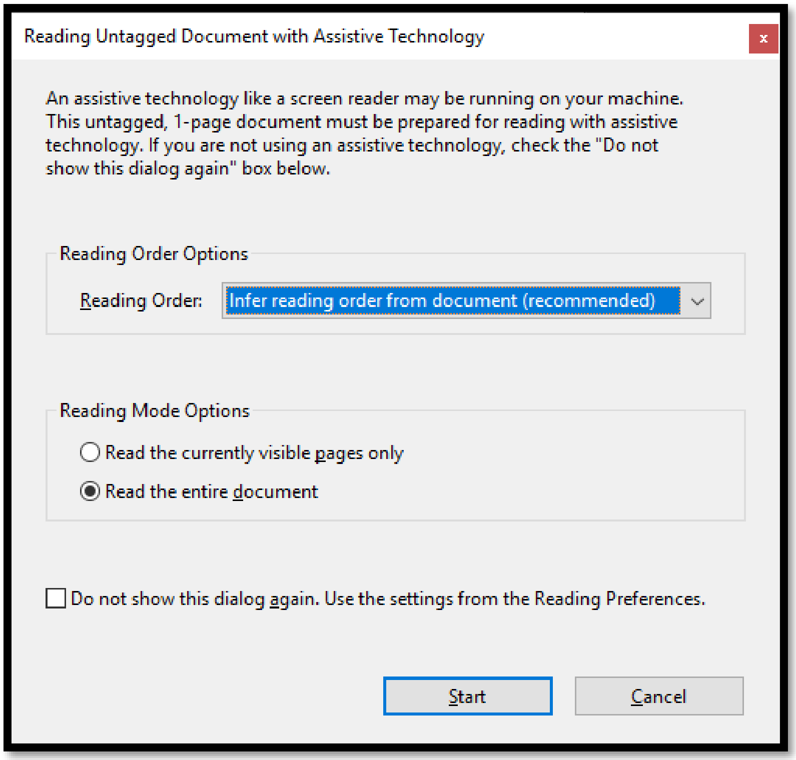

Scanned Page Alert

This page contains only an image. Would you like to run text recognition to make the text on this page accessible?

□ Do not show this message again

Figure 4 Notification of scanned image if you use adaptive technology.

If the document is a scanned image, we can let the "Virtual OCR" happen which will be followed by a message saying that the document is not tagged and asking if you want Adobe Acrobat or Reader to try and figure it out by adding what I call "Virtual Tags" because they really don't exist. If you open the Tags Tree, you won't see them. They are a band-aid for legacy PDF that are not tagged. They facilitate quick access to something that is not currently accessible.


Reading Untagged Document with Assistive Technology

An assistive technology like a screen reader may be running on your machine. This untagged, 1-page document must be prepared for reading with assistive technology. If you are not using an assistive technology, check the "Do not show this dialog again" box below.

Reading Order Options

Reading Order: Infer reading order from document (recommended)

Reading Mode Options

○ Read the currently visible pages only

● Read the entire document

□ Do not show this dialog again. Use the settings from the Reading Preferences.

Figure 5 Notification of an untagged document if you use adaptive technology.

If I am remediating a document, I dismiss both of these messages as I find that allowing either of these processes to continue can adversely affect my ability to correctly OCR and/or Tag a document. The original The Salamanca Statement was not a scanned image so

I dismissed the “do you want me to figure it out” dialog and then used Adobe Acrobat Pro DC to Tag the document.

I’ve extracted page 6 from the original document and added Tags using Adobe Acrobat Pro DC, Nuance’s PowerPDF Advanced and Foxit Phantom for Business. The experience of page 6 is representative of the entire document. I have, however, printed and scanned back into the computer pages 6, 7 and 8 in both greyscale and colour and taken the 3 page versions through the various tools.

ABBYY PDF Transformer (Second Attempt)

For the second try at getting a version of The Salamanca Statement I could read and understand, I used ABBYY PDF Transformer with page 6 of the original Salamanca Statement document. Although Transformer is a stand-alone OCR tool, I often find that if I open an unreadable PDF in Transformer, convert it to Word, I get something that is more readable.

For the second attempt at making the document readable and to see if I could figure out what was going on under the hood, I opened the original Salamanca Statement in ABBYY PDF Transformer and chose to send the document to Word. Normally this works if I have an untagged PDF or a particularly horrid PDF, even one that is tagged horridly.

I’ve attached the Word document that is the result of simply opening The Salamanca Statement in Transformer and sending it to Word. There is a conversion/sort of OCR process that goes on during the conversion.

As you can see from the Word document, although a bit clearer, still not readable or understandable.

Starting from Scratch

After trying my usual go to solutions, I decided to explore other ways of making the text readable. Keep in mind that I needed to read this document for work and there is a deadline for participation!

For my next attempts, and for most of the remainder of the content in this tutorial, I actually printed page 6 of the Salamanca Statement and scanned it back into the computer as a test to see which application would give me the best and most readable results. I used four applications:

- ABBYY PDF Transformer (no longer available for purchase but it is a light version of ABBYY FineReader which you can purchase).
- Adobe Acrobat Pro DC.
- Nuance PowerPDF Advanced.
- Foxit Phantom for Business.

Both the colour and greyscale versions of page 6 were scanned into the computer with a setting of 600 dpi or dots per inch.. Only the pages scanned directly into ABBYY PDF Transformer were scanned in at 300 dpi which was recommended and I kept that default.

I created two versions of page 6: one colour and one greyscale. I chose to try both because this document is typical of what we might find in a scanned PDF. The text on some of the pages of the document are in italics which can cause problems for OCR tools and the text is on a light blue background which may violate colour contrast checkpoints but can't be changed without sending the document back to the author. This document was created in 1994 and my guess would be that the document author is no longer available.

I've attached all of my test files to this tutorial and hopefully have given them meaningful filenames. They are, for the most part, "as is" and represent the initial output when using a specific tool. The exception is page 6 colour PowerPDF which only had a few remediations and I couldn't resist. All others required extensive remediation.

The Salamanca Statement document represents problems that can't be "seen" until the document is tagged and read using adaptive technology. When the initial OCR was done on the original Salamanca Statement document, many of the words had spaces between the characters in the words making reding impossible. In other instances, hidden hyphens were exposed and in parts of the words that didn't make sense if you were going to hyphenate them. This isn't discovered until you read the text with adaptive technology once it is tagged.

Adding the Tags before you add links, form controls and/or multimedia violates the Hierarchy of tasks set out by Adobe with the release of Acrobat Pro 6. The hierarchy of tasks identifies things you must do before you Tag a PDF document. However, if you go through and make all the remediations, spending time and resources to get a clean accessibility check only to find that the document is unreadable because of spacing and hyphenation problems, you are duplicating your work and wasting time. For this reason, I recommend that once you perform any type of OCR on a scanned PDF document, ad the Tags to it and go through it with a screen reader. You may want to use a non-phoneme synthesizer like Eloquence that ships with JAWS and can be purchased to work with NVDA so that you get a more accurate reading of the document. One of the issues with natural phoneme voices is that often subtle punctuation and symbols are missed. One natural phoneme voice reads the word "wasted" as "was ted" which complicates any type of accuracy detection in a scanned PDF that has gone through OCR.

All tests from this point in the tutorial document are with page 6 from The Salamanca Statement original document that was printed and scanned back into the computer. Page 6 remediation is done with both a colour version and a greyscale version. I used the Windows 10 Scan and Fax tool and the HP DeskJet 37000 to scan both versions of page 6 into my computer.

Adobe Acrobat Pro DC (Colour Document)

With the first attempt at using the Enhanced Scan tools in Adobe Acrobat Pro DC, I used the Recognize Text button. This time, I decided to try the Enhance button to see if there is any difference. There wasn't much. I would say try both and take the best result based on the quality of the scan. If there is no good readable document, use a stand-alone OCR tool that is dependable. I recommend ABBYY Fine Reader[1] but you can also use OmniPage Pro[2].

Adobe Acrobat Pro DC, Enhanced Scan, Recognize Text:

1. With the PDF document open, activate the Enhanced Scan tools from the Tools Task Pane.
    a. The Enhanced Scan Toolbar appears above the document.
2. From the Enhanced Scan tools, activate Recognize Text.

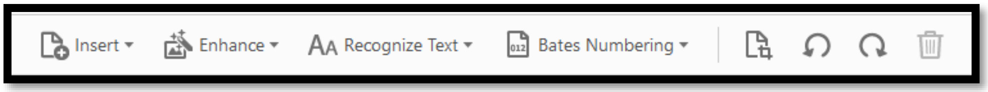
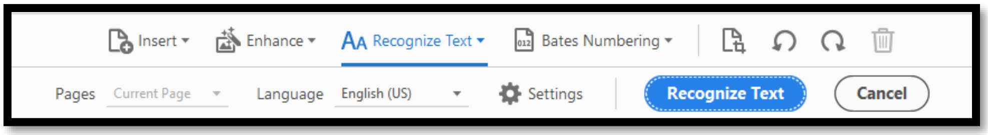


Figure 6 Enhanced Scan Toolbar.

3. Choose "From this File". This opens a sub-Toolbar just under the first one.



Pages Current Page

Language English (US)

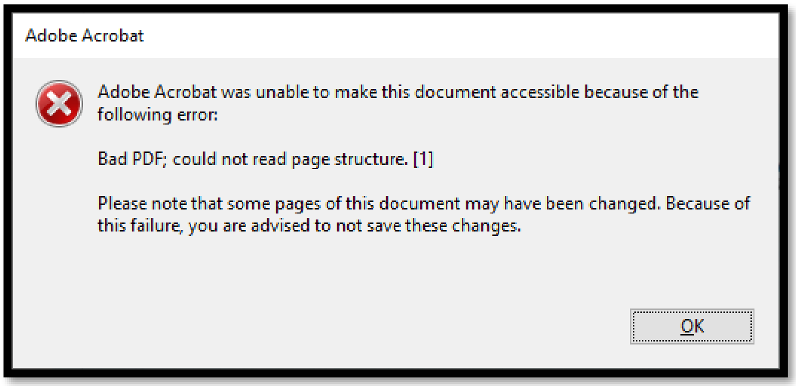
Figure 7 Recognize Text Toolbar in the Enhanced Scan tools.

4. The Settings are fine. The default setting is to make the document "Editable Text and Images."
5. Activate the Recognize Text button.
6. Tag the document and review the text with a screen reader.
7. If the document is readable, exit the document without saving, open the document again and preform the OCR, then continue with the Hierarchy of Tasks and any needed remediation.
    a. If you delete the tags and try to use the Enhanced Scan tools, you'll get the "Bad PDF" message. This is one of my favourite messages…"Bad PDF!"

[1] ABBYY Fine Reader: https://www.abbyy.com/en-ca/finereader/

[2] Nuance OmniPage Pro: https://shop.nuance.com/store/nuanceus/custom/pbpage.OmniPage-Standard?utm_medium=dr_nam_ps&utm_source=bing&utm_campaign=imaging&utm_term=nuance%20omnipage&cvokeywordid=32|571406&cvosrc=ps.Bing.nuance%20omnipage&matchtype=p&adid=77446878472443&addisttype=s&msclkid=58752cea06b91eb2c920701058c65467

If the document is NOT readable, close the document without saving, open it again and try using the Enhance button in the Enhanced Scan Toolbar instead of the Recognize Text button.



Adobe Acrobat

Adobe Acrobat was unable to make this document accessible because of the following error:

Bad PDF; could not read page structure. [1]

Please note that some pages of this document may have been changed. Because of this failure, you are advised to not save these changes.

Figure 8 "Bad PDF" message!

Let’s walk through the Enhanced Scan Toolbar, Enhance button tools.

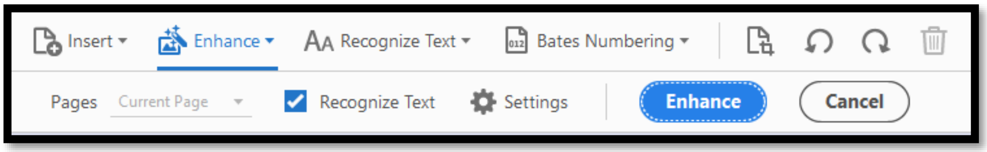


Figure 9 Enhanced Scan tools showing Enhance sub-Toolbar.

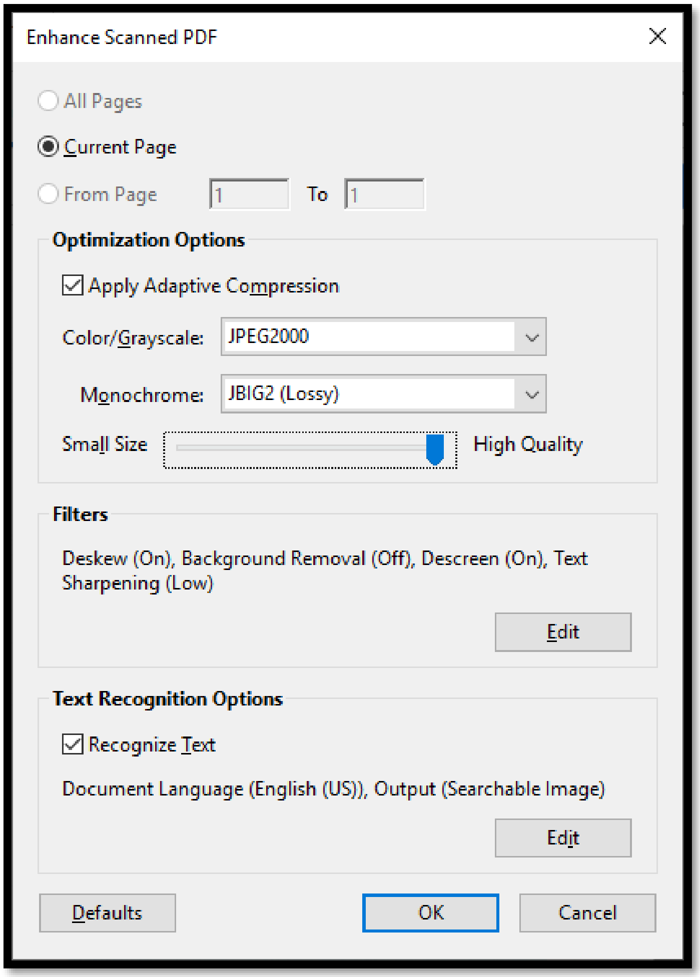
1. With the scanned PDF document open, activate the Enhanced Scan tools to show the Enhanced Scan Toolbar.

2. Activate the Enhance button tools on the Enhanced Scan Toolbar and choose Scanned Document from the drop-down list.

3. The Enhanced Scan Toolbar opens under the Enhanced Scan Toolbar just above your document.

4. I did adjust the Enhanced Scan settings to use High Quality instead of the default of “small file size.”

    a. This slider doesn’t make much sense as you would think you would have opposites like small size versus large size or low quality versus high quality, but we have small size versus high quality. While it is intuitive to think that small size means the file size, the wording is awkward at best. High Quality means a larger file size…I think…but a higher quality of recognition.



Enhance Scanned PDF

○ All Pages

● Current Page

○ From Page 1 To 1

Optimization Options

☑ Apply Adaptive Compression

Color/Grayscale: JPEG2000

Monochrome: JBIG2 (Lossy)

Filters

Deskew (On), Background Removal (Off), Descreen (On), Text Sharpening (Low)

Text Recognition Options

☑ Recognize Text

Document Language (English (US)), Output (Searchable Image)

Figure 10 Enhanced Scan, Enhanced Scan tools settings dialog.

Once the document was tagged, I saved it closed it and then opened it and read it using my screen reader.

Back to the drawing board.

Note: When you are testing files with adaptive technology you MUST launch the adaptive technology first! With the adaptive technology running, launching the application/document helps the adaptive technology choose the tools it will use to access the user interface and content. If you start the adaptive technology after you have the application/document open, the adaptive technology can get confused about where it is and what scripts/tools to use in order to give you the best possible outcomes. Always start the adaptive technology BEFORE you launch an application/document.

Note: I use the term “application/document” as sometimes I simply press Enter on a document which means that both the application and the document are “launched” or opened at the same time.

ABBYY PDF Transformer (Colour Document)

For this test, I scanned page 6 directly into ABBYY PDF Transformer making sure I checked the check box to recognize text in the scan settings.

Note to self: Make sure that your cats aren’t helping you too closely as their loose hairs can turn up on the scanned pages! If you look closely at the greyscale page 6, I think you’ll see a few strands of Olivia Zane’s fur!

Once I scanned page 6 into Transformer with the check box to recognize text checked, I saved it as a PDF and took it to Adobe Acrobat Pro DC. Using a screen reader, you can’t read anything in Transformer. You have to either send it to Word or another application.

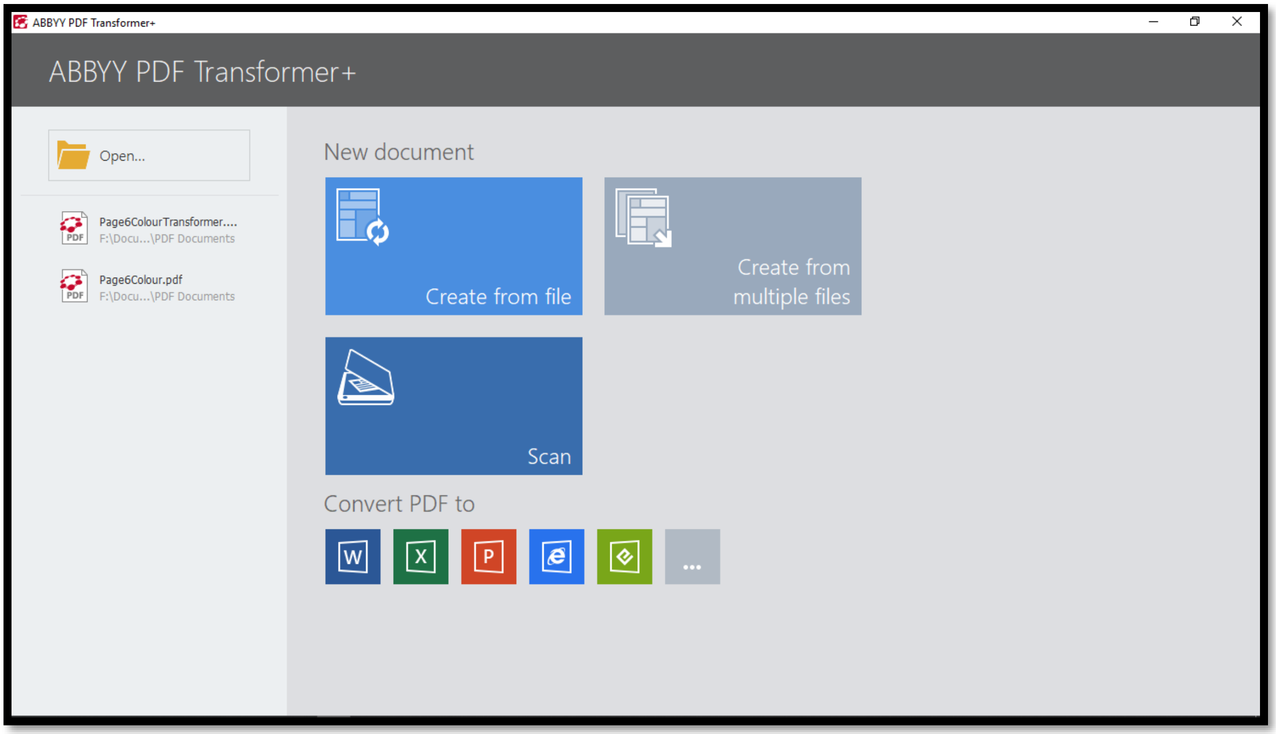
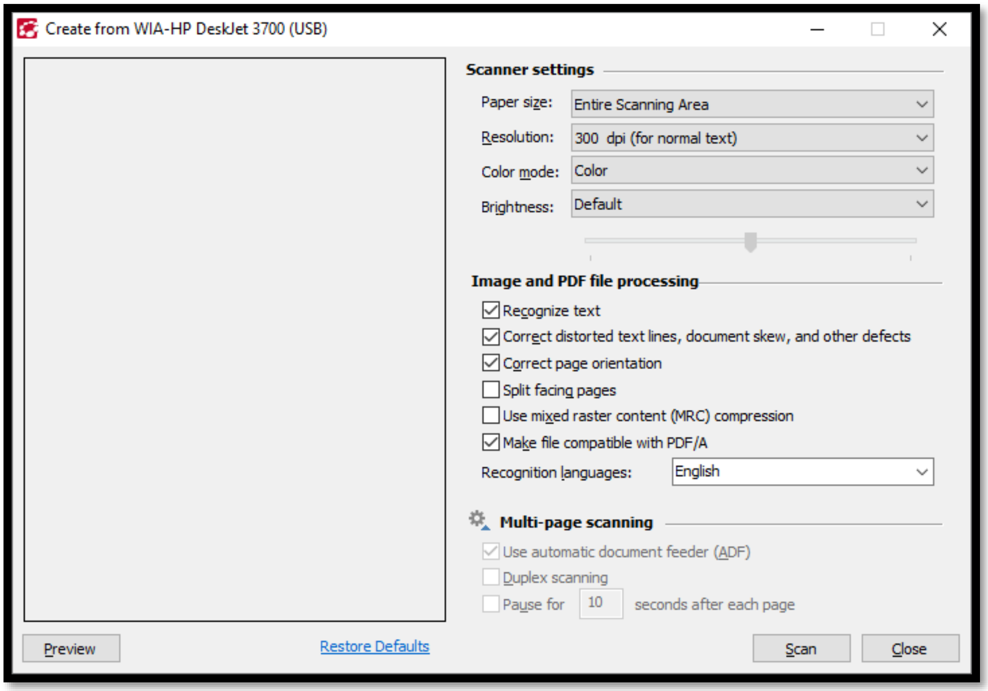


Figure 11 ABBYY PDF Transformer user interface showing Scan button..

When I scanned the printed colour page 6 into ABBYY PDF Transformer and saved it as a PDF document, when I opened it in Adobe Acrobat Pro DC I only had to do three things:

1. Add the appropriate document properties including language.
2. Make the words “The Salamanca Statement” a <H1> Tag.
3. Select the page and set the Tab Order to Page Structure in the Page Properties dialog.



Create from WIA-HP DeskJet 3700 (USB)

Scanner settings

Paper size: Entire Scanning Area

Resolution: 300 dpi (for normal text)

Color mode: Color

Brightness: Default

Image and PDF file processing

☑ Recognize text

☑ Correct distorted text lines, document skew, and other defects

☑ Correct page orientation

□ Split facing pages

□ Use mixed raster content (MRC) compression

☑ Make file compatible with PDF/A

Recognition languages: English

Multi-page scanning

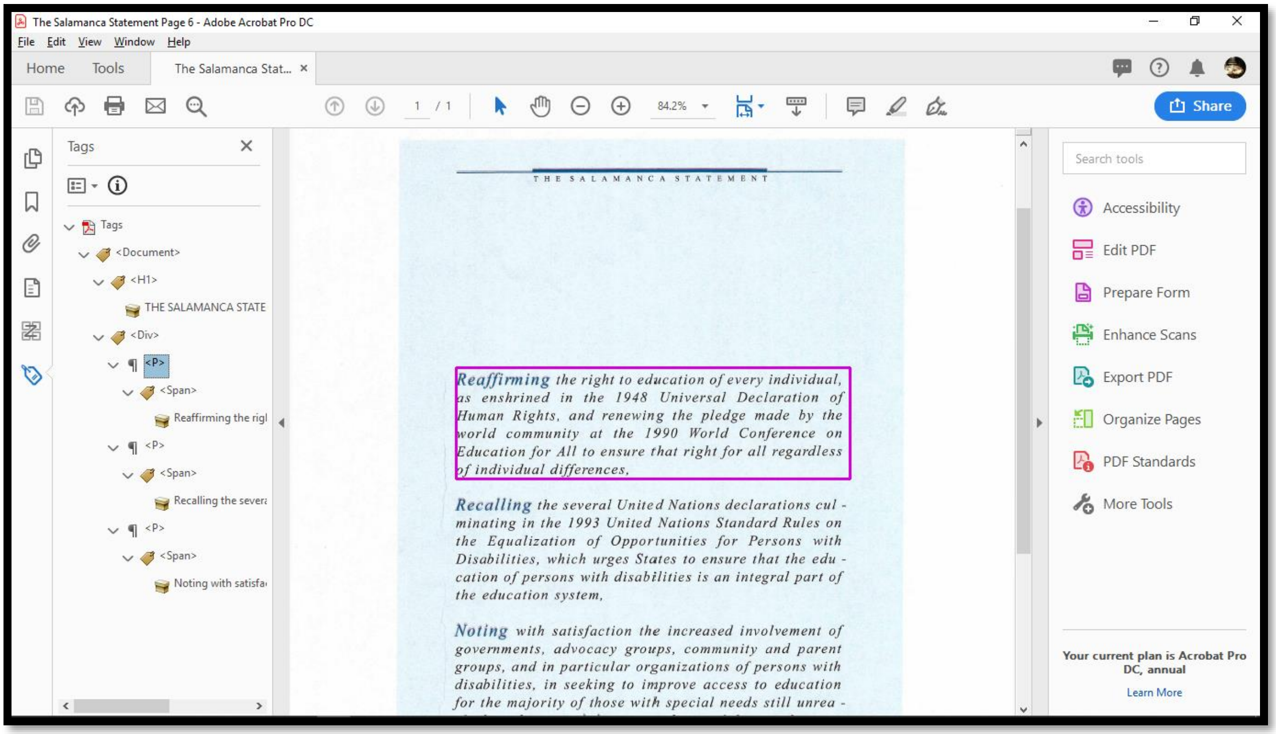
☑ Use automatic document feeder (ADF)

□ Duplex scanning

□ Pause for 10 seconds after each page

Figure 12 Scanning settings in ABBYY PDF Transformer.

When I opened the document in Adobe Acrobat Pro DC, the Tags were there and aside from a liberal use of the <Div> and <Span> Tags, the Tags Tree was clean.



THE SALAMANCA STATEMENT

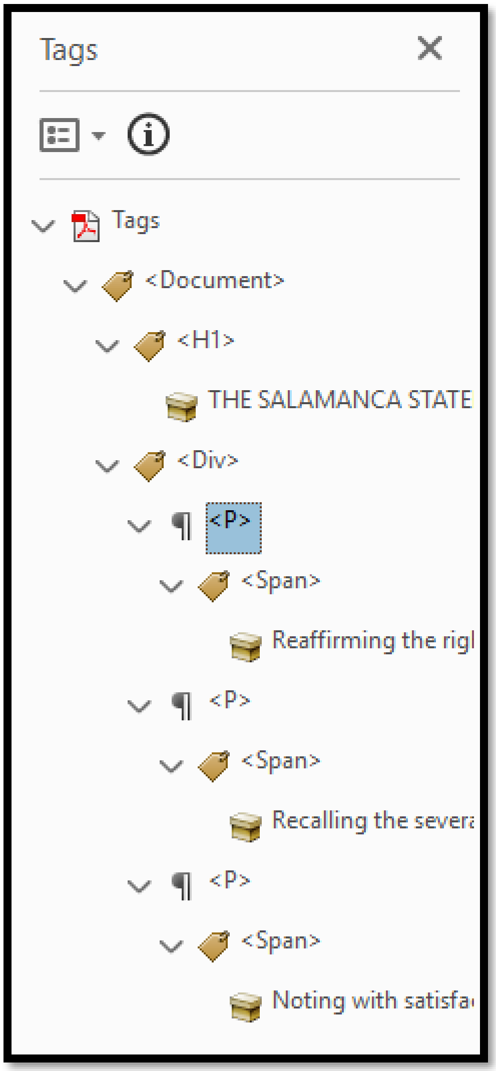
Reaffirming the right to education of every individual, as enshrined in the 1948 Universal Declaration of Human Rights, and renewing the pledge made by the world community at the 1990 World Conference on Education for All to ensure that right for all regardless of individual differences,

Recalling the several United Nations declarations culminating in the 1993 United Nations Standard Rules on the Equalization of Opportunities for Persons with Disabilities, which urges States to ensure that the education of persons with disabilities is an integral part of the education system,

Noting with satisfaction the increased involvement of governments, advocacy groups, community and parent groups, and in particular organizations of persons with disabilities, in seeking to improve access to education for the majority of those with special needs still unrea-

Figure 13 ABBYY PDF Transformer page 6 colour scanned directly into Transformer, saved and opened in Acrobat.

The following image is a close-up of the Tags Tree as it was generated by ABBYY PDF Transformer during the scan and recognize process. Once the document was scanned into Transformer, I just needed to save it and open it in Acrobat to see the Tags.



Tags

- Tags
  - <Document>
    - <H1>
      - THE SALAMANCA STATE
    - <Div>
      - <P>
        - <Span>
          - Reaffirming the rigl
      - <P>
        - <Span>
          - Recalling the severa
      - <P>
        - <Span>
          - Noting with satisfa

Figure 14 ABBYY PDF Transformer Tags Tree from scan and recognize process.

One of the things that this document (page 6) shows us is the effect that full justification of text has on readability. This is one reason that full justification is not used in digital content. Left justification should be the default for digital content. The hyphenated words are not identified as being hyphenated and sound like word fragments as you read through the document using a screen reader or Text-to-Speech tool. However, this is the result of all of the tools used to OCR the contents of The Salamanca Statement document..

Foxit Phantom for Business (Colour Document)

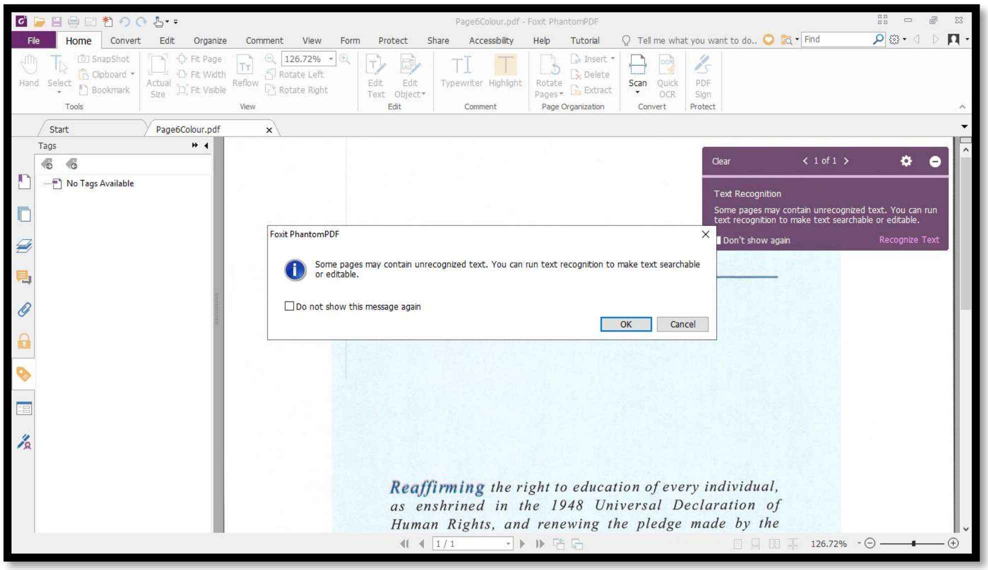
One of the nice things about Foxit is that once you open a scanned image of a document, you have a purple pop-up from which you can start the OCR process and you have a dialog that will also start the process.

One of the frustrating things about Foxit Phantom for Business is that I have to keep adding the Tags, Order and Content Panels to the Navigation Pane EVERY time I launch it!

If you are running any adaptive technology when you launch Foxit Phantom for Business, you will get two notices. The purple pop-up and a regular dialog. This can be confusing. At first I thought that if I chose the regular dialog, I would get something similar to the Acrobat “virtual OCR.” Actually, it was only when writing this up that it twigged that there were two messages about OCR and I thought of the Adobe band-aid solutions. I then tried the process with and without my screen reader running. Page 6 colour Foxit 01 is with my screen reader running and page 6 colour Foxit 02 is with my screen reader turned off.

The difference is that with my screen reader running, I was in an endless loop of finding the same suspects and as you can see from the test file, ended up with duplicate text under Tags. I haven’t tested to see whether removing one of these duplicates will result in an error in the accessibility check. At this point I am simply looking at the raw Tag output.

In terms of the process, without my screen reader running I didn’t have duplicate suspect text and content was only tagged once. However, as you will see from the Tags in both documents, the content is not really readable as words and phrases are truncated into individual paragraphs which means the adaptive technology will pause as it normally would at the end of a paragraph before moving on to the next paragraph. In both processes, there is a lot of remediation required.



Text Recognition

Some pages may contain unrecognized text. You can run text recognition to make text searchable or editable.

□ Don't show again   Recognize Text

Foxit PhantomPDF

Some pages may contain unrecognized text. You can run text recognition to make text searchable or editable.

□ Do not show this message again

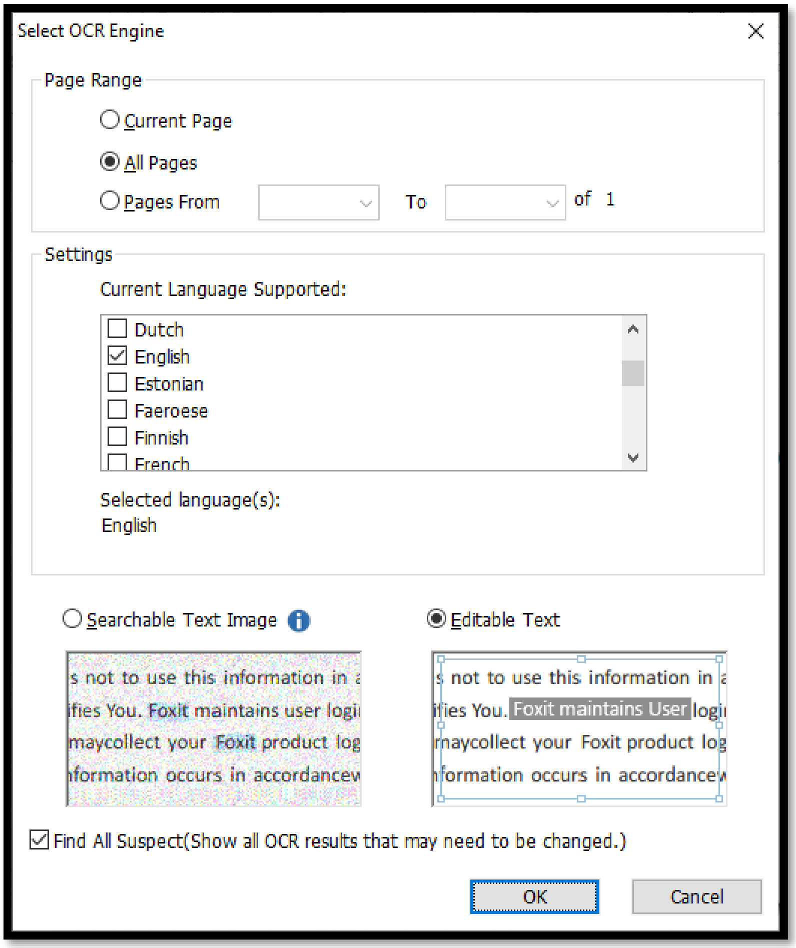
OK   Cancel

Reaffirming the right to education of every individual, as enshrined in the 1948 Universal Declaration of Human Rights, and renewing the pledge made by the

Figure 15 Foxit user interface showing both the purple pop-up and the regular dialog for recognizing text.

The other nice thing about Foxit is that as soon as you start the OCR process, you get the Settings dialog. The only changes I made were to choose Editable Text and check the check box to find all Suspect Text.

The OCR engine was set to English so I didn’t have to change that and all other settings were fine.



Select OCR Engine

Page Range

○ Current Page

● All Pages

○ Pages From To of 1

Settings

Current Language Supported:

☐ Dutch
☐ English ☑
☐ Estonian
☐ Faeroese
☐ Finnish
☐ French

Selected language(s):
English

○ Searchable Text Image ● Editable Text

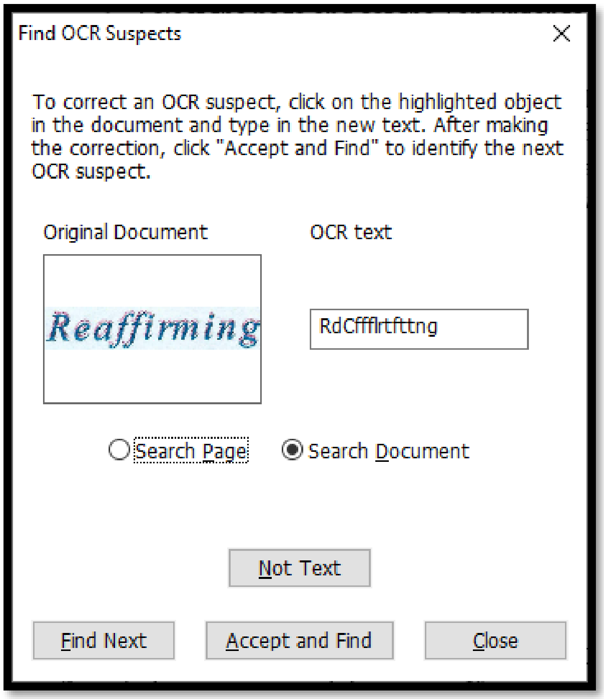
s not to use this information in a
ifies You. Foxit maintains user logi
maycollect your Foxit product log
formation occurs in accordancew
s not to use this information in a
ifies You. Foxit maintains User logi
maycollect your Foxit product log
formation occurs in accordancew

☑ Find All Suspect(Show all OCR results that may need to be changed.)

Figure 16 Select OCR Engine dialog in Foxit Phantom for Business.

With my screen reader running, I found that I often had to go back through the process in order for the Find OCR Suspects dialog to open once the OCR process completed. The process is supposed to be fluid, moving from OCR to finding suspect text automatically. This doesn't always happen if you have adaptive technology running.

However, you can use the Ask a Question tool (Alt + Q which is the same as it is in Office 365 desktop applications). I typed Suspect Text and several tools appeared in the results list. I chose OCR which has a sub-menu and one of the tools is Find OCR Suspects. If you lose this tool or if the process stalls for some reason, this is a good way to go back and launch the Find OCR Suspects dialog.


Find OCR Suspects

To correct an OCR suspect, click on the highlighted object in the document and type in the new text. After making the correction, click "Accept and Find" to identify the next OCR suspect.

Original Document

Reaffirming

OCR text

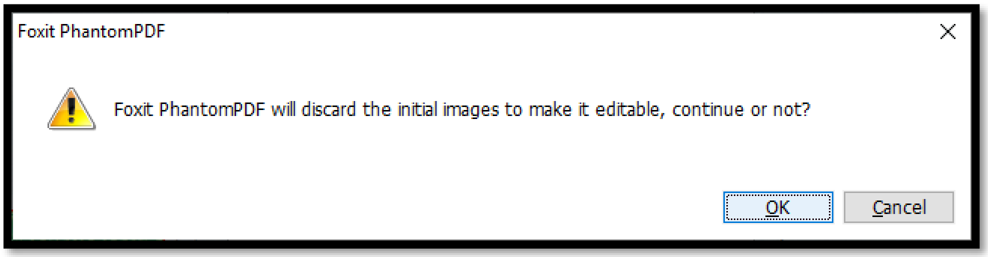
RdCffflrtftng

○ Search Page ● Search Document

Figure 17 Find OCR Suspects dialog in Foxit Phantom for Business.

In the Find OCR Suspects dialog you can type in the word that is correct (right) while looking at the OCR result on the left. You can designate something as “Not Text” and activate Accept and Find once you’ve typed the correct word or have checked that the OCR was correct. I did have JAWS running while going through this process and the Find OCR Suspects dialog is accessible although you can’t really “see” the suspect text. However, you can see the misspelling in the editable form control so you have an idea of what went wrong.

You can use the mouse to move from suspect text to suspect text within the document itself if you find you are in a loop, which I did find using my screen reader. As you click on text bordered by a red line representing suspect text, the text bordered by the red line is also shaded to let you know you were successful and information changes in the Find OCR Suspect dialog and you are able to make corrections if necessary.


Foxit PhantomPDF

Foxit PhantomPDF will discard the initial images to make it editable, continue or not?

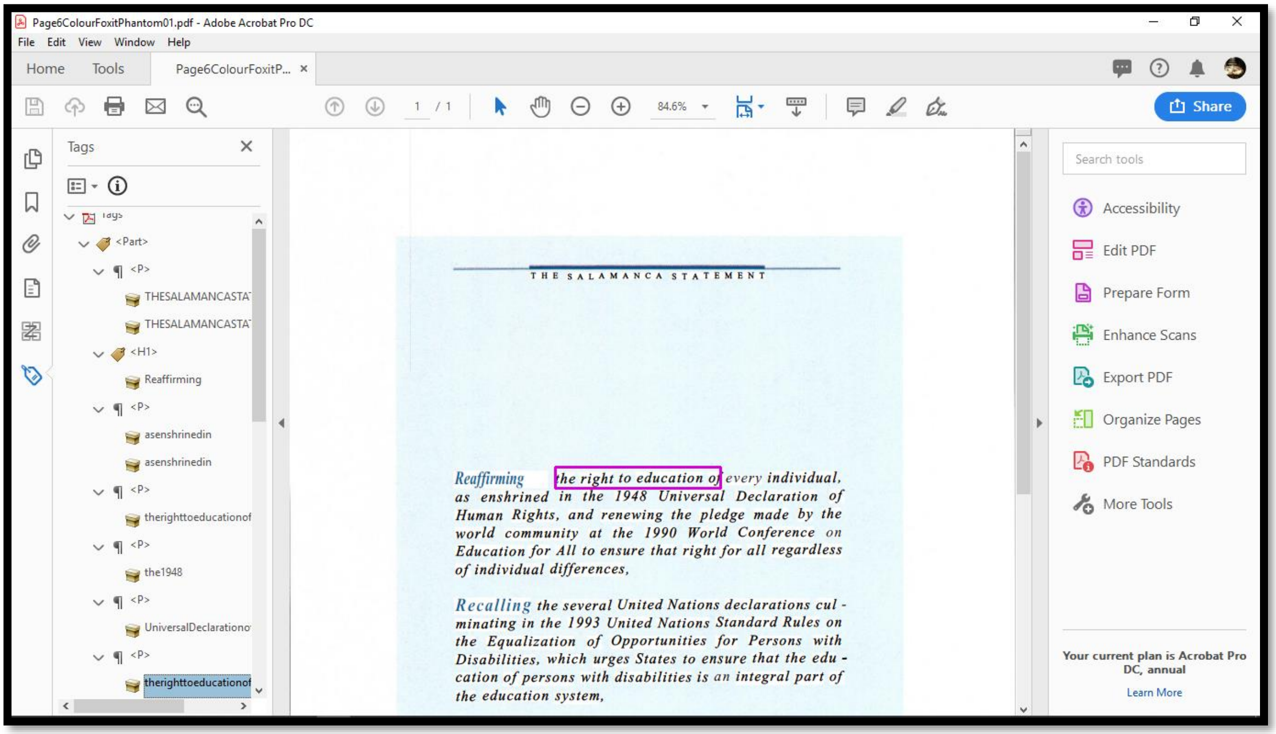
Figure 18 Foxit Phantom PDF message indicating that suspect items will be discarded from the document.

With my screen reader running, the only message I got once the OCR and Find OCR Suspects process completed was the message saying that all of the suspect items would be discarded from the document.

Without my screen reader running, there was a message saying that the Find OCR Suspect process was complete but for some reason it appeared behind the Find OCR Suspect dialog and I couldn't bring it to the front to dismiss it. I couldn't dismiss the Find OCR Suspect dialog and the only way to end the process was to activate the OK button in the "We're now going to discard all suspect items" dialog.

I then tagged the PDF document and reviewed the Tags. The results of both processes (with and without the screen reader running) are attached to this tutorial.

The following image is of the page 6 colour Foxit Phantom 01 document (process with screen reader running).



```
Tags
  <Part>
    <P>
      THESALAMANCASTA
      THESALAMANCASTA
    <H1>
      Reaffirming
    <P>
      asenshrinedin
      asenshrinedin
    <P>
      therighttoeducationof
    <P>
      the1948
    <P>
      UniversalDeclarationo
    <P>
      therighttoeducationof
```

THE SALAMANCA STATEMENT

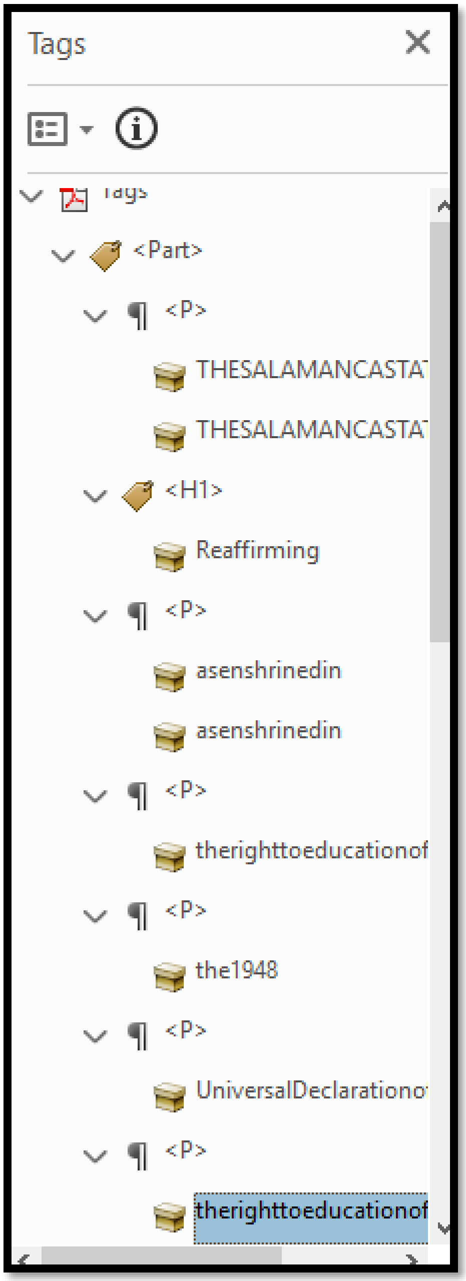
Reaffirming *the right to education of every individual, as enshrined in the 1948 Universal Declaration of Human Rights, and renewing the pledge made by the world community at the 1990 World Conference on Education for All to ensure that right for all regardless of individual differences,*

Recalling the several United Nations declarations culminating in the 1993 United Nations Standard Rules on the Equalization of Opportunities for Persons with Disabilities, which urges States to ensure that the education of persons with disabilities is an integral part of the education system,

Figure 19 PDF document after OCR and tagging with a screen reader running.

I've included the document with the view of the Tags Tree because the logical reading order is really out of order and individual pieces of the paragraphs are tagged as separate paragraphs.

The following is a closer look at the Tags Tre for page 6 colour Foxit Phantom 01.

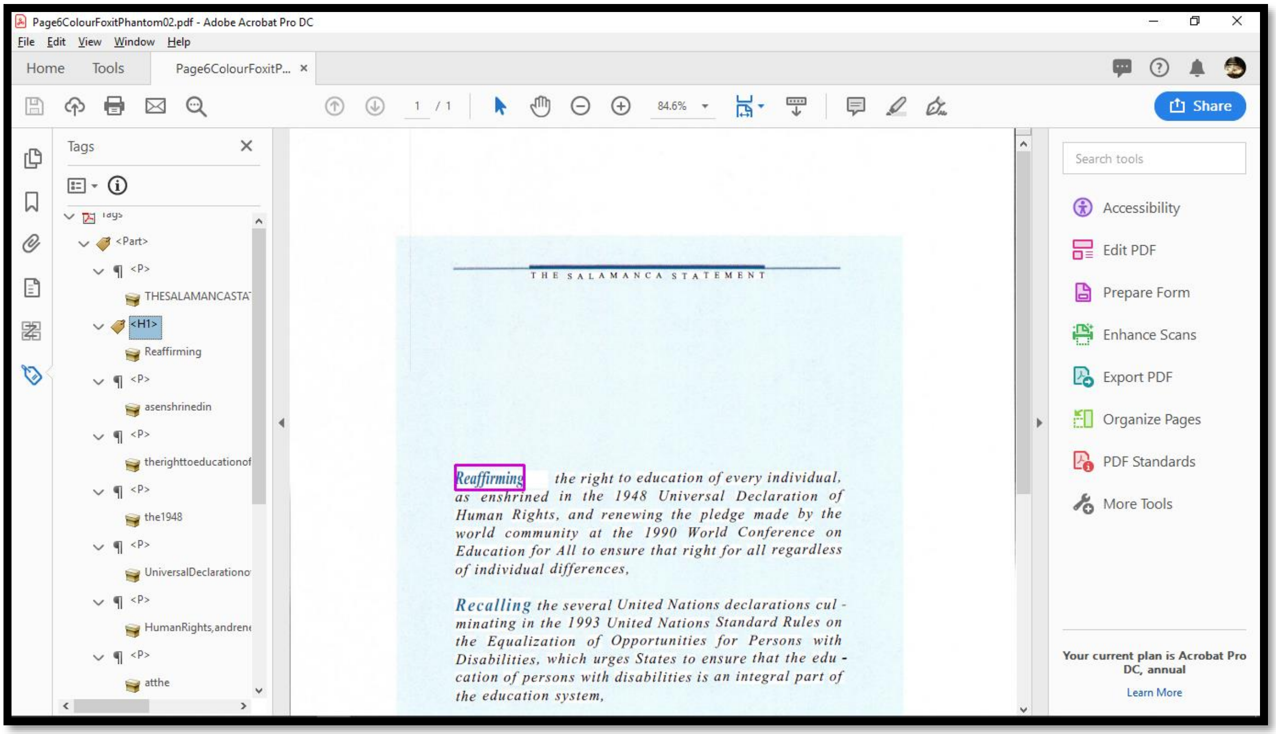


Tags

- Tags
  - <Part>
    - <P>
      - THESALAMANCASTAT
      - THESALAMANCASTAT
    - <H1>
      - Reaffirming
    - <P>
      - asenshrinedin
      - asenshrinedin
    - <P>
      - therighttoeducationof
    - <P>
      - the1948
    - <P>
      - UniversalDeclarationo
    - <P>
      - therighttoeducationof

Figure 20 Close up of the Tags Tree from Foxit created with a screen reader running.

If we compare the results of taking the same page 6 colour document and going through the process without any adaptive technology running, we get slightly different results. The document is still not readable, but the Tags Tre is a bit cleaner.



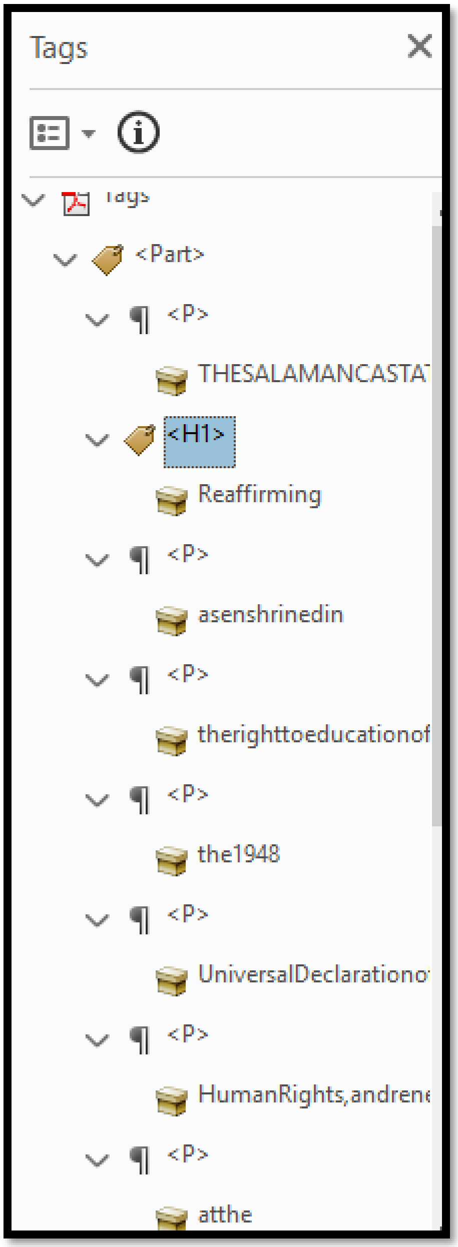
THE SALAMANCA STATEMENT

Reaffirming the right to education of every individual, as enshrined in the 1948 Universal Declaration of Human Rights, and renewing the pledge made by the world community at the 1990 World Conference on Education for All to ensure that right for all regardless of individual differences,

Recalling the several United Nations declarations culminating in the 1993 United Nations Standard Rules on the Equalization of Opportunities for Persons with Disabilities, which urges States to ensure that the education of persons with disabilities is an integral part of the education system,

Figure 21 Page 6 colour after OCR and tagging in Foxit Phantom without adaptive technology running.

The following image is a closer look at the tags Tree for Page 6 colour Foxit Phantom 02.



```
Tags
  <Part>
    <P>
      THESALAMANCASTA
    <H1>
      Reaffirming
    <P>
      asenshrinedin
    <P>
      therighttoeducationof
    <P>
      the1948
    <P>
      UniversalDeclarationo
    <P>
      HumanRights,andren
    <P>
      atthe
```

Figure 22 Tags Tree for tagged PDF created after OCR and tagging in Foxit Phantom without adaptive technology running.

The conclusion is that while the Foxit user interface provides some helpful directions in terms of the pop-up to ask if you want to perform text recognition, given the scanned from print copy of page 6 of the Salamanca Statement, the document remains unreadable. I did not perform any remediations as this would be counter productive to testing the process.

Nuance PowerPDF Advanced (Colour Document)

I used the same document, page 6 of The Salamanca Statement, when I tested the OCR capability in Nuance's PowerPDF Advanced. The difference between Adobe Acrobat Pro DC/Foxit Phantom for Business and Nuance's PowerPDF is night and day!

I completed this task WITH my screen reader running!

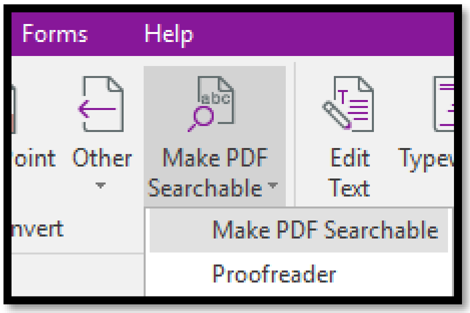
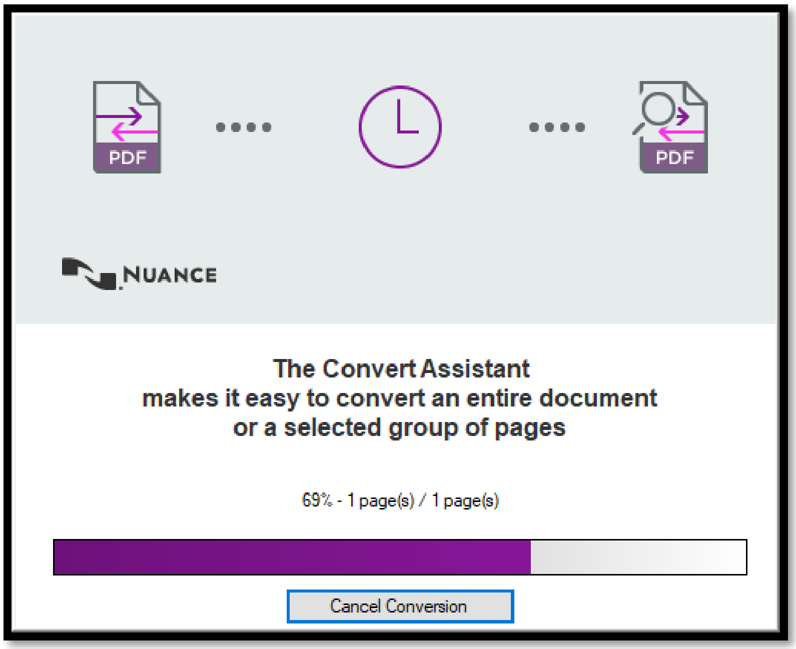


Figure 23 Nuance PowerPDF Make PDF Searchable button in the Home Ribbon.

The Make PDF Searchable tool is found on the Home Ribbon just under the Help Ribbon. It has a sub-menu of two items: Make PDF Searchable which is the OCR tool and Proofreading which is the "find suspects" tool. If you make the PDF searchable the proofreading tool starts automatically. However, you can run the proofreading tool independently.

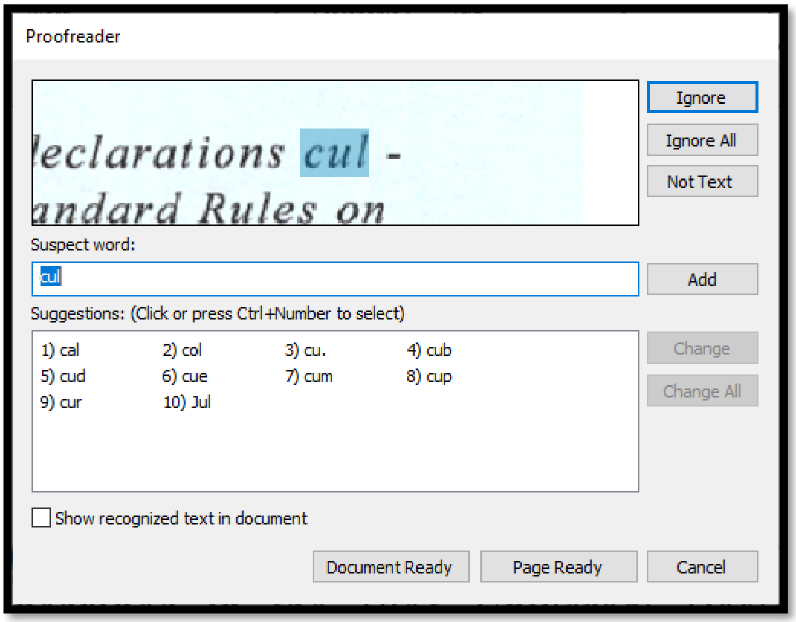


The Convert Assistant
makes it easy to convert an entire document
or a selected group of pages

69% - 1 page(s) / 1 page(s)

Figure 24 PowerPDF Advanced conversion dialog.

As soon as the conversion process or the OCR process is completed, the Proofreading dialog opens and the first item is found. Looking at the dialog, it is designed to work with Nuance's Dragon products. The possible choices are numbered for easy use with voice recognition. I also think that because PowerPDF is a Nuance product, it is leveraging the OmniPage Pro OCR tool/engine.



leclarations cul -
andard Rules on

Suspect word: cul

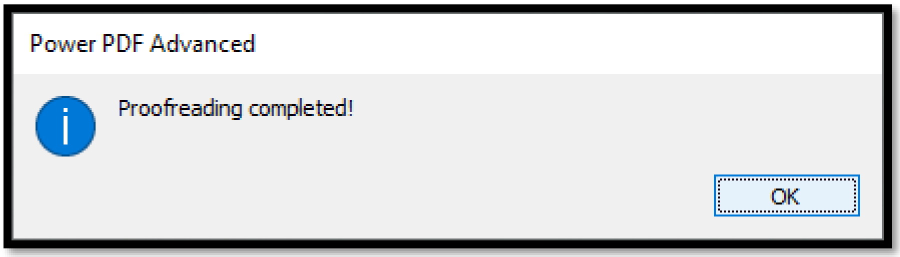
Suggestions: (Click or press Ctrl+Number to select)

1) cal 2) col 3) cu. 4) cub
5) cud 6) cue 7) cum 8) cup
9) cur 10) Jul

□ Show recognized text in document

Figure 25 PowerPDF Proofreading dialog.

In the Proofreading dialog you can identify elements that are not text, type in a corrected word and choose Add to add it to a dictionary or choose one of the choices in a list. There is also a check box to show recognized text in document but I found that the document shifted back and forth if I had this checked and it was difficult to see…/read.



Proofreading completed!

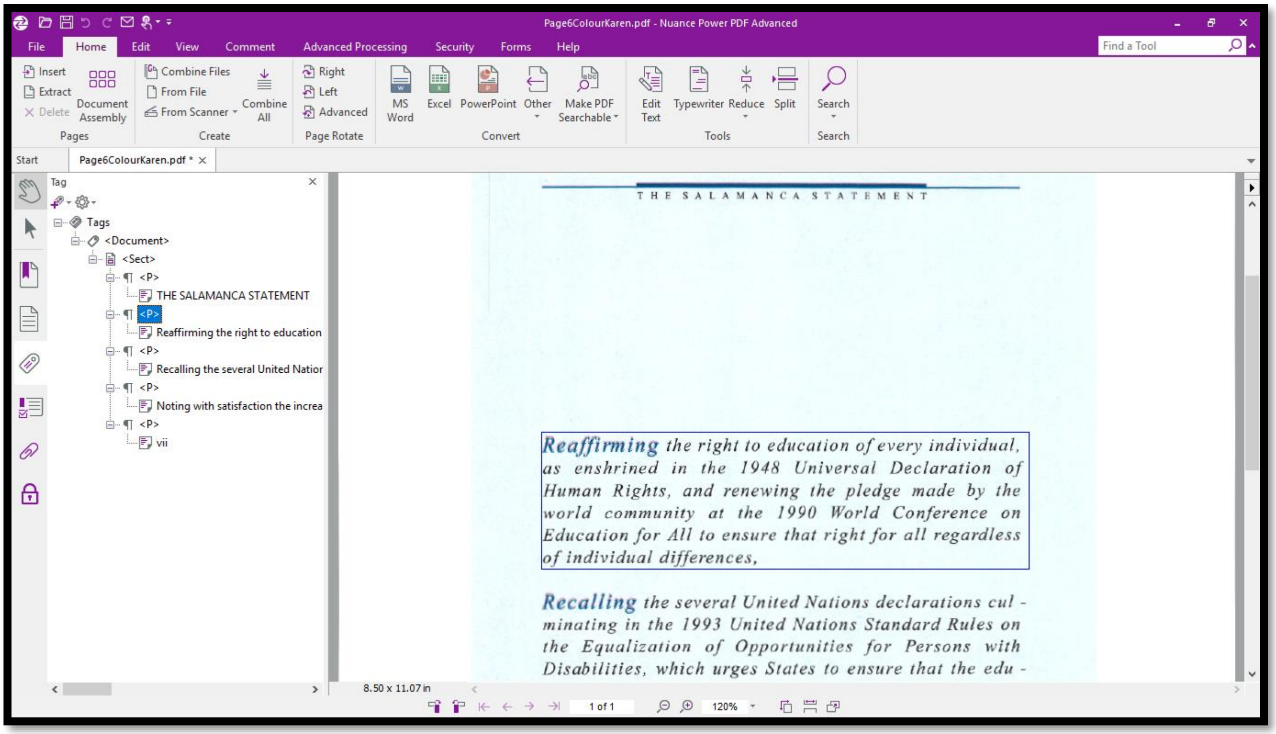
Figure 26 PowerPDF Advanced dialog indicating that Proofreading is complete.

The Proofreading dialog also has two buttons, one to indicate that the page is ready/no more suspect elements and one to indicate that the document is ready. Once you choose Document Ready, a dialog opens telling you that proofreading is complete.

You have to save the document before you can Tag it. The button to add Tags is in the Tags Panel to the immediate left of the Options button. The online Help documentation is out of date and the tools/method they talk about is either no longer available or buried so deep that I couldn't find it. I like the "easy" button which lets me either Tag or Retag the document.

Figure 27 PowerPDF Tag PDF button in the Tags Panel.

The following image is of page 6 colour PowerPDF. This is the only document I did remediate! I only had to change the <P> Tag to <H1> and add the document properties! The following image is of the document in PowerPDF before I made those changes.



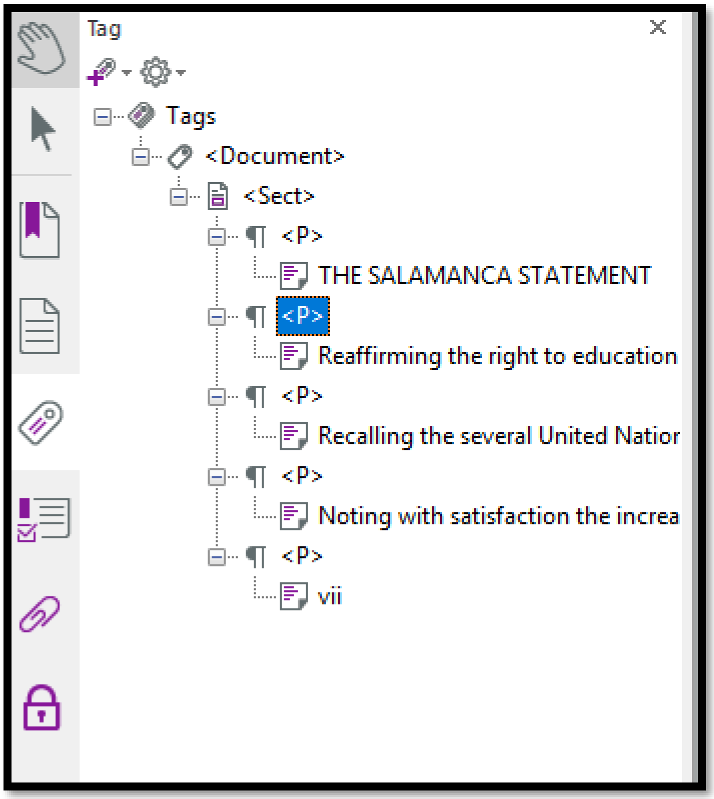
THE SALAMANCA STATEMENT

Reaffirming the right to education of every individual, as enshrined in the 1948 Universal Declaration of Human Rights, and renewing the pledge made by the world community at the 1990 World Conference on Education for All to ensure that right for all regardless of individual differences,

Recalling the several United Nations declarations culminating in the 1993 United Nations Standard Rules on the Equalization of Opportunities for Persons with Disabilities, which urges States to ensure that the edu -

Figure 28 PowerPDF Advanced showing PDF document after OCR and Tags were added.

The following image is a closer look at the Tags Tree of page 6 colour PowerPDF



```
Tags
  <Document>
    <Sect>
      <P>
        THE SALAMANCA STATEMENT
      <P>
        Reaffirming the right to education
      <P>
        Recalling the several United Natior
      <P>
        Noting with satisfaction the increa
      <P>
        vii
```

Figure 29 PowerPDF Tags Tree for page 6 colour PowerPDF.

The file is called Page 6 colour PowerPDF.

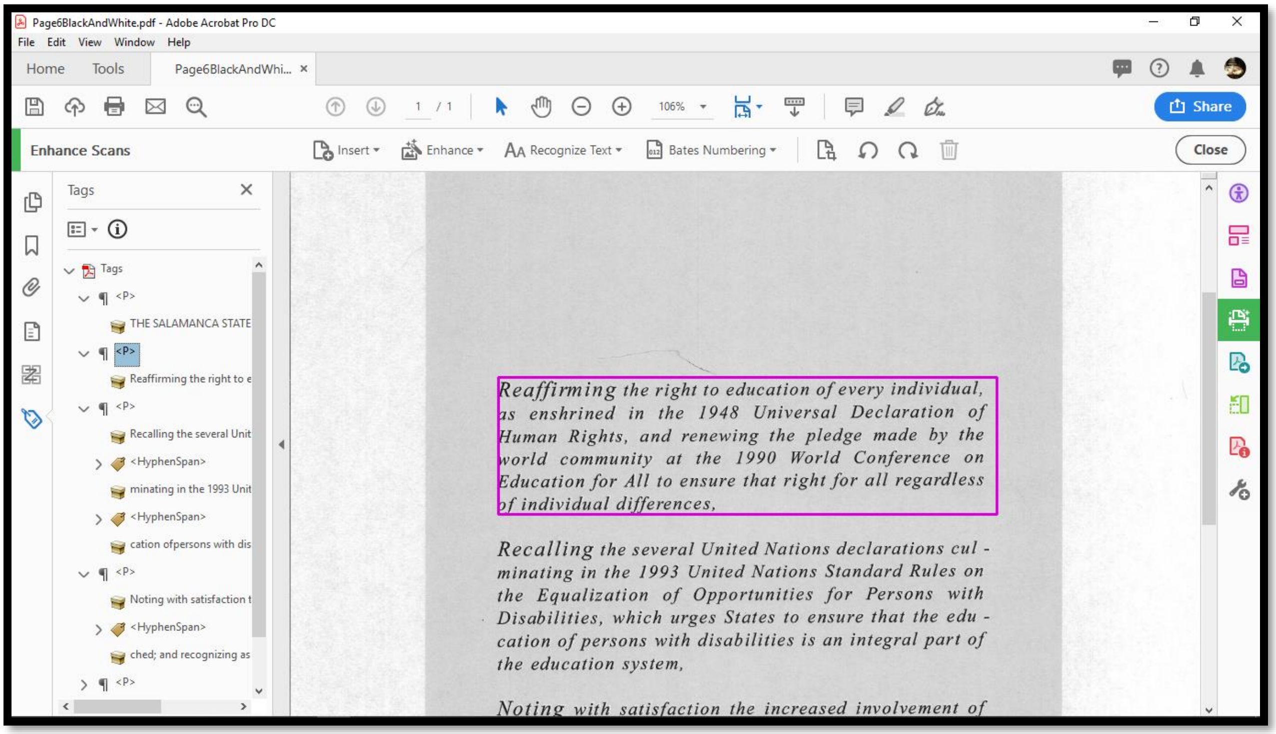
Aside from printing and scanning a document back into the computer, Nuance's PowerPDF performed the best and most accurate in taking page 6 colour scanned document and recognizing the text then adding the correct Tags.

Switch to Greyscale!

For comparison I printed a greyscale copy of page 6 and ran the same tests. The exception was that I ran the Foxit Phantom for Business test without my screen reader running.

Adobe Acrobat Pro DC (Greyscale Document)

Using the Enhanced Scan tools and the Enhance button with the settings still at High Quality, I got the same result with the greyscale as I got with the colour version of page 6.


Page6BlackAndWhite.pdf - Adobe Acrobat Pro DC
File Edit View Window Help
Home Tools Page6BlackAndWhi... ×
1 / 1
106%
Share
Enhance Scans
Insert
Enhance
Recognize Text
Bates Numbering
Close
Tags
Tags
<P>
THE SALAMANCA STATE
<P>
Reaffirming the right to e
<P>
Recalling the several Unit
<HyphenSpan>
minating in the 1993 Unit
<HyphenSpan>
cation ofpersons with dis
<P>
Noting with satisfaction t
<HyphenSpan>
ched; and recognizing as
<P>

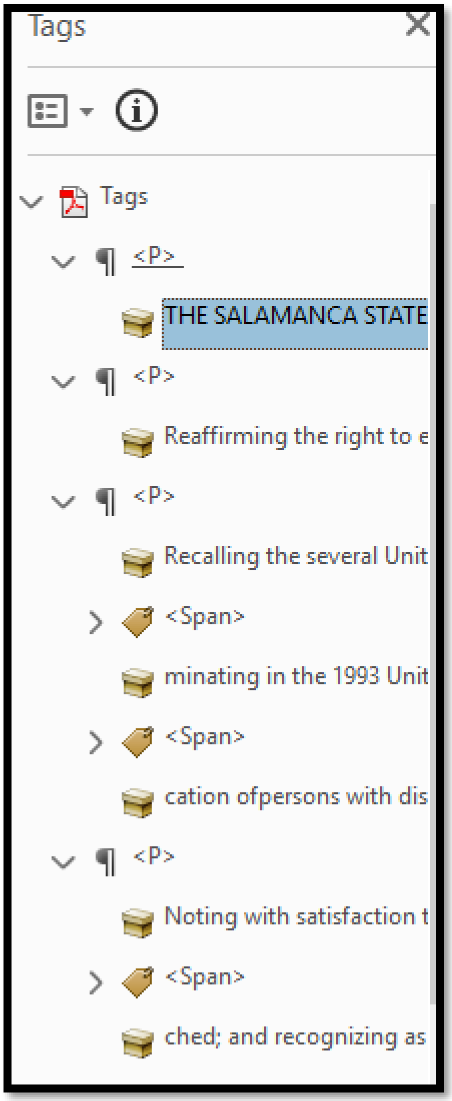
Reaffirming the right to education of every individual, as enshrined in the 1948 Universal Declaration of Human Rights, and renewing the pledge made by the world community at the 1990 World Conference on Education for All to ensure that right for all regardless of individual differences,

Recalling the several United Nations declarations culminating in the 1993 United Nations Standard Rules on the Equalization of Opportunities for Persons with Disabilities, which urges States to ensure that the education of persons with disabilities is an integral part of the education system,

Noting with satisfaction the increased involvement of

Figure 30 Adobe Acrobat Pro DC greyscale version of OCR and tagged page 6.

The following is a close-up of the Tags Tree for this document.



Tags

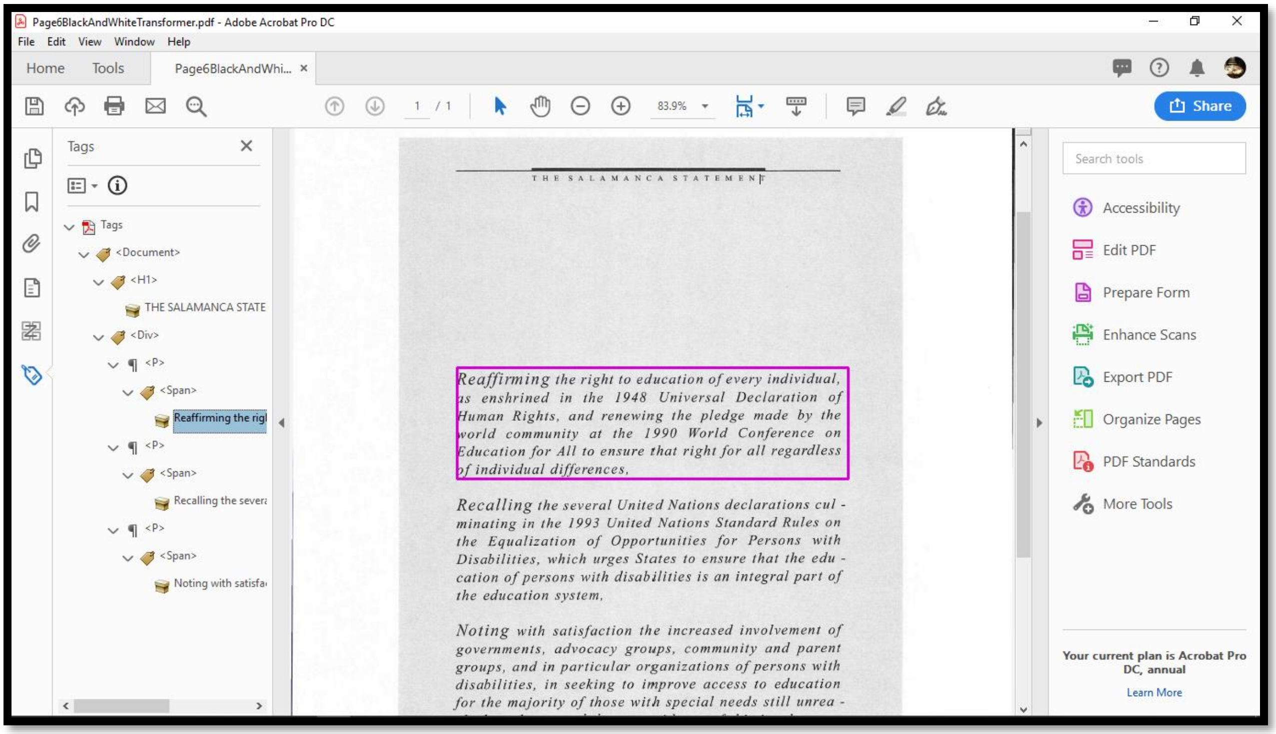
- Tags
  - <P>
    - THE SALAMANCA STATE
  - <P>
    - Reaffirming the right to e
  - <P>
    - Recalling the several Unit
    - <Span>
    - minating in the 1993 Unit
    - <Span>
    - cation ofpersons with dis
  - <P>
    - Noting with satisfaction t
    - <Span>
    - ched; and recognizing as

Figure 31 Close-up of the Tags Tree for the greyscale version of page 6 tagged in Acrobat.

ABBYY PDF Transformer (Greyscale Document)

While the page 6 greyscale was scanned in at 600 dpi, Transformer recommended 300 dpi or dots per inch so I kept the default.

When the scanning and text recognition was finished, I saved the document and opened it in Adobe Acrobat Pro DC to review the tags. The title of the document "The Salamanca Statement" was not tagged so I tagged it as an <H1>. All other Tags in the attached greyscale version are "as is" out of the box/process. The document will not have any document properties either…but then again, none of the documents in this test have document properties unless I manually added them.



THE SALAMANCA STATEMENT

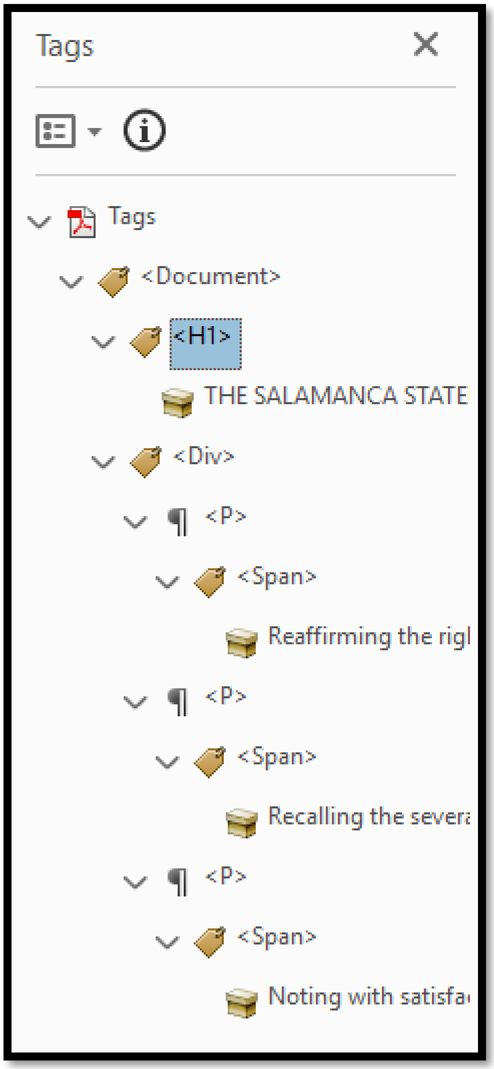
Reaffirming the right to education of every individual, as enshrined in the 1948 Universal Declaration of Human Rights, and renewing the pledge made by the world community at the 1990 World Conference on Education for All to ensure that right for all regardless of individual differences,

Recalling the several United Nations declarations culminating in the 1993 United Nations Standard Rules on the Equalization of Opportunities for Persons with Disabilities, which urges States to ensure that the education of persons with disabilities is an integral part of the education system,

Noting with satisfaction the increased involvement of governments, advocacy groups, community and parent groups, and in particular organizations of persons with disabilities, in seeking to improve access to education for the majority of those with special needs still unrea -

Figure 32 Tagged PDF from scanned page 3 greyscale through ABBYY PDF Transformer and viewed in Acrobat..

The following image is a close-up view of the Tags Tree for the greyscale page 6 I scanned directly into Transformer, saved as PDF and opened in Acrobat Pro DC. The only remediation I did was to add the <H1> Tag and then use Create Tag From Selection to add “The Salamanca Statement” as the title of the document.



```
Tags
  <Document>
    <H1>
      THE SALAMANCA STATE
    <Div>
      <P>
        <Span>
          Reaffirming the rigl
      <P>
        <Span>
          Recalling the severa
      <P>
        <Span>
          Noting with satisfa
```

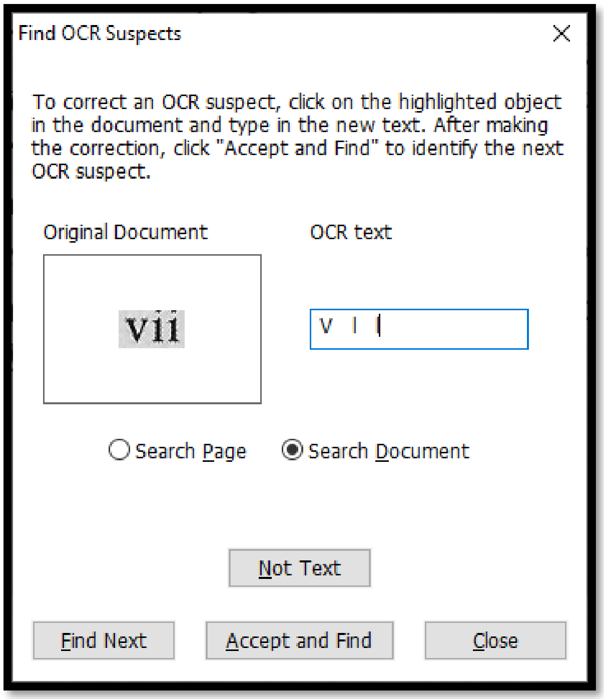
Figure 33 ABBYY PDF Transformer scanned and recognized as well as tagged page 6 greyscale viewed in Acrobat.

Foxit Phantom for Business (Greyscale Document)

One of the biggest problems with Foxit is that I have to show the Tags Panel/Tags Tree EVERY time I launch the application. It doesn't save my settings!

Once the OCR tool had gone through the document, there was only one suspect, the page number. I fixed it and continued on with the process.

Some of the bullets were identified as suspect and I activated the Not Text button to make them Artifacts.


Find OCR Suspects

To correct an OCR suspect, click on the highlighted object in the document and type in the new text. After making the correction, click "Accept and Find" to identify the next OCR suspect.

Original Document

vii

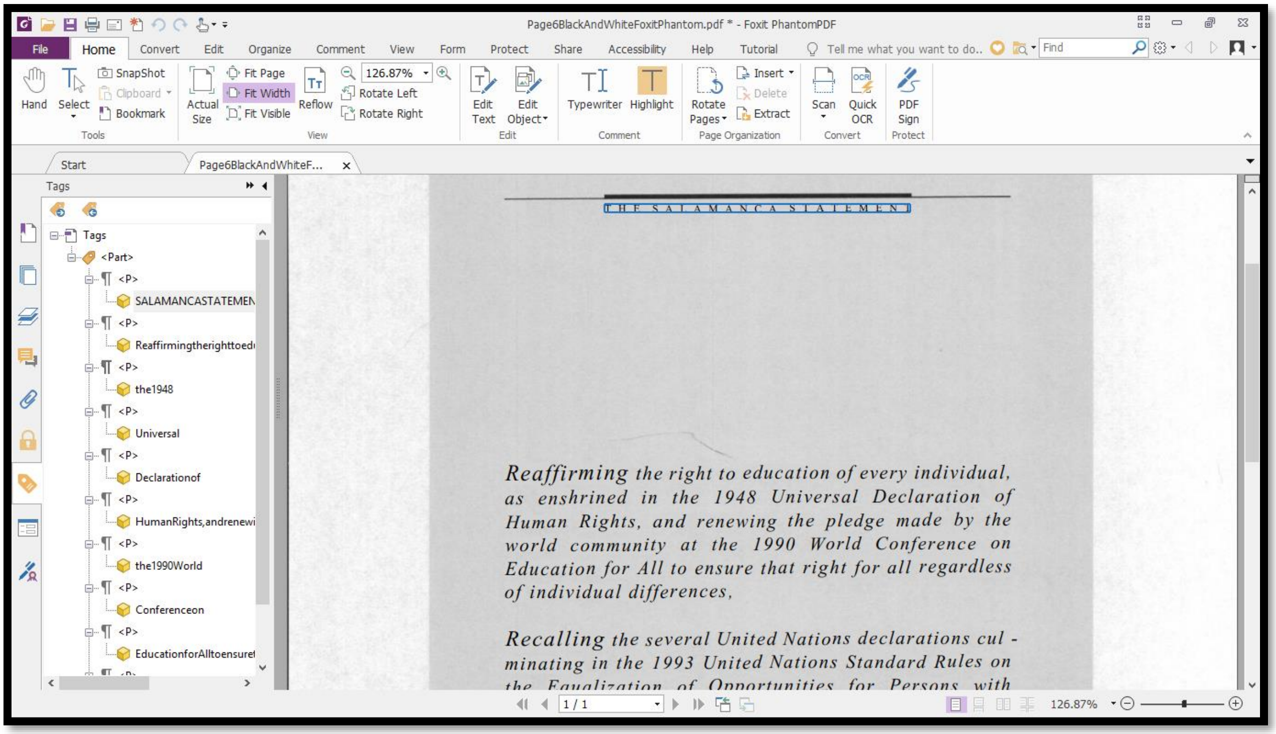
OCR text

V I I

○ Search Page ● Search Document

Figure 34 Foxit Phantom Find OCR Suspects for greyscale PDF.

I saved the document and then added the Tags. I had to add the Tags Panel to the Navigation Pane AGAIN because I had exited Foxit Phantom for Business.


Page6BlackAndWhiteFoxitPhantom.pdf * - Foxit PhantomPDF

Tags
- <Part>
  - <P>
    - SALAMANCASTATEMEN
  - <P>
    - Reaffirmingtherighttoedı
  - <P>
    - the1948
  - <P>
    - Universal
  - <P>
    - Declarationof
  - <P>
    - HumanRights,andrenewi
  - <P>
    - the1990World
  - <P>
    - Conferenceon
  - <P>
    - EducationforAlltoensuret

THE SALAMANCA STATEMENT

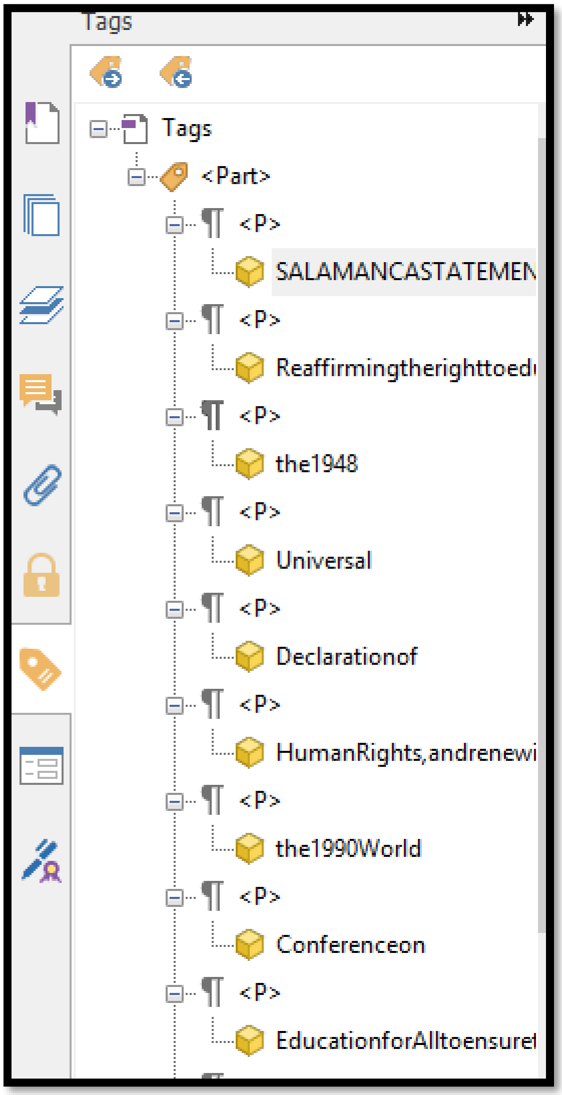
Reaffirming the right to education of every individual, as enshrined in the 1948 Universal Declaration of Human Rights, and renewing the pledge made by the world community at the 1990 World Conference on Education for All to ensure that right for all regardless of individual differences,

Recalling the several United Nations declarations cul - minating in the 1993 United Nations Standard Rules on the Equalization of Opportunities for Persons with

Figure 35 Foxit Phantom greyscale document after OCR and tagging.

The other problem with Foxit Phantom for Business is that the text is truncated into individual paragraphs which mean more remediation is required. Some of the text/Tags are also out of their logical reading order.

The following image is a close-up of the Tags Tree for page 6 greyscale Foxit Phantom



Tags
<Part>
<P>
SALAMANCASTATEMEN
<P>
Reaffirmingtherighttoed
<P>
the1948
<P>
Universal
<P>
Declarationof
<P>
HumanRights,andrenewi
<P>
the1990World
<P>
Conferenceon
<P>
EducationforAlltoensuret

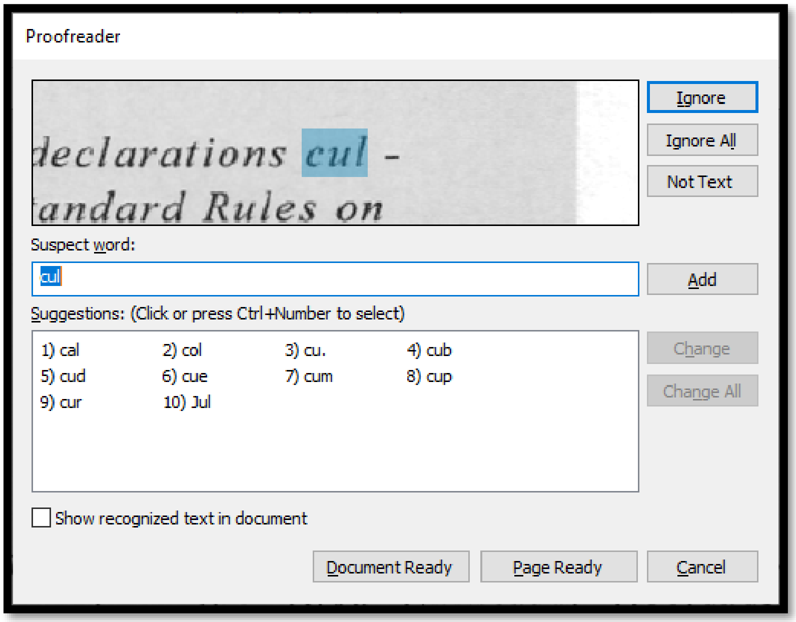
Figure 36 Foxit Phantom Tags Tree for greyscale PDF after OCR and tagging.

The results were consistent with what were the results of doing the OCR and tagging on the colour representation of page 6 with Foxit Phantom for Business.

Nuance PowerPDF (Greyscale Document)

While annoying, PowerPDF doesn't keep Highlight Content turned on! I'm so used to just seeing the corresponding text highlighted when working with PDF documents in other applications that I begin to question whether the document is tagged or not…until I remember I have to turn on Highlight Content! You can turn on Highlight Content by right clicking in the Tags Tree and clicking on Highlight Content or by pressing the AppKey in the Tags Tree and pressing Enter on Highlight Content.

As with the colour version of this document, the hyphenated words are identified by the Proofreading tool as suspect text.



Suspect word: cul

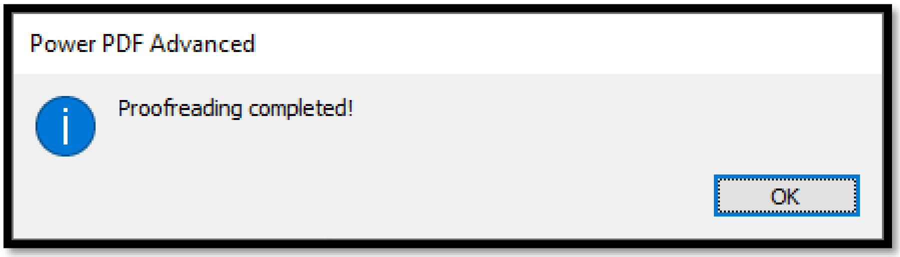
Suggestions: (Click or press Ctrl+Number to select)

1) cal 2) col 3) cu. 4) cub
5) cud 6) cue 7) cum 8) cup
9) cur 10) Jul

□ Show recognized text in document

Figure 37 PowerPDF Advanced Proofreading dialog.

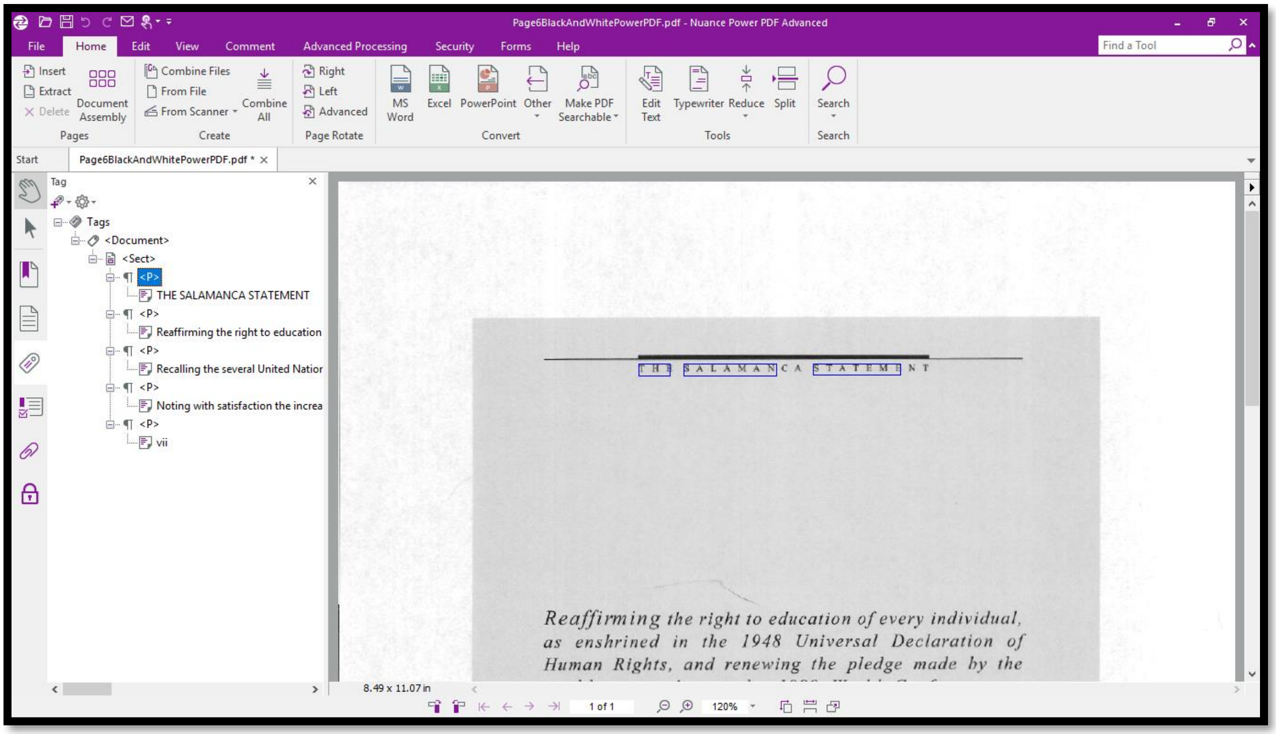
Once all of the proofreading/suspects have been dealt with, the Proofreading is complete dialog appears.



Proofreading completed!

Figure 38 PowerPDF Advanced Proofreading is complete dialog.

The following image is of the Tags Tree and greyscale page 6 PDF document in Nuance's PowerPDF Advanced.



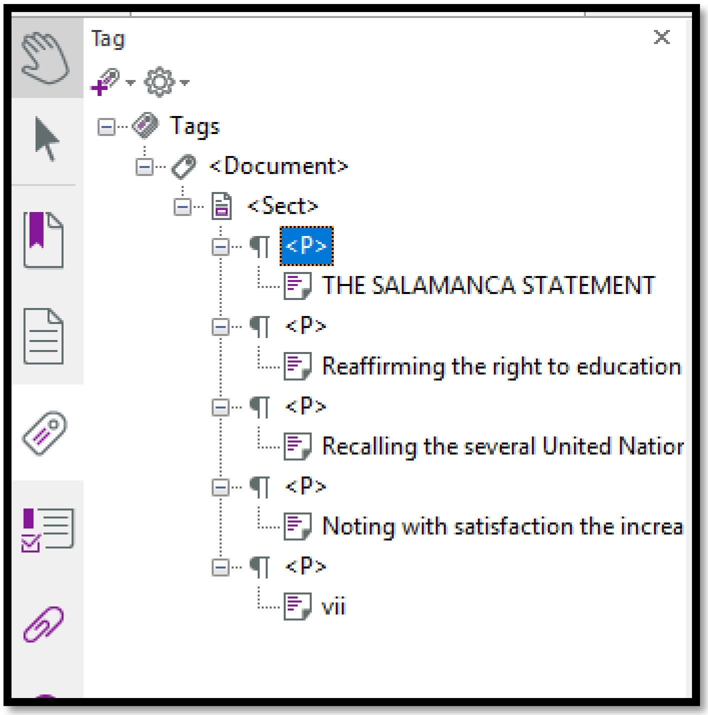
Tags
<Document>
<Sect>
<P>
THE SALAMANCA STATEMENT
<P>
Reaffirming the right to education
<P>
Recalling the several United Natior
<P>
Noting with satisfaction the increa
<P>
vii

THE SALAMANCA STATEMENT

Reaffirming the right to education of every individual, as enshrined in the 1948 Universal Declaration of Human Rights, and renewing the pledge made by the

Figure 39 PowerPDF Advanced page 6 greyscale document after OCR and tagging.

The following image is a closer look at the Tags Tree for page 6 greyscale PowerPDF. I did not remediate this document.



Tags
<Document>
<Sect>
<P>
THE SALAMANCA STATEMENT
<P>
Reaffirming the right to education
<P>
Recalling the several United Natior
<P>
Noting with satisfaction the increa
<P>
vii

Figure 40 PowerPDF Advanced Tags Tree for page 6 greyscale after OCR and tagging.

Using PowerPDF Advanced gave the best results of the specific PDF remediation taking a greyscale scan of page 6 from the Salamanca Statement of 1994. This is consistent with the results of page 6 colour PowerPDF. The Tags Tree is clean, the text is readable and only a few remediations are needed.

One can perhaps assign this to the fact that Nuance also owns OmniPage Pro and is using their OCR engine or a light version of it in this application. This presumption would fall in line with the quality of output from ABBYY PDF Transformer since ABBYY produces FineReader.

Three Pages of the Salamanca Statement

For the final OCR test, I used pages 6 through 8 of the Salamanca Statement in both greyscale and colour. The results were similar to those already identified in this tutorial. However, there were a few new things.

The first new document element was numbered parts of the statement.

The second new document element were lists.

The comments in this topic relate to Tags other than <P> Tags which were prevalent throughout the previous tests. What I'm looking for are Headings and lists and any other Tags. For the page 6 example, only <P> Tags were added to the Tags Tree with the exception of Adobe Acrobat Pro DC which invented a new Tag for the hyphens!

Adobe Acrobat Pro DC (3 Page Document)

I used the JAWS screen reader to read both versions of the documents in Acrobat.

For both the colour and greyscale version of pages 6 through 8, Acrobat did correctly Tag the lists. The large coloured numbers for the sections were incorporated into the paragraph text and, for the most part, paragraphs were not broken up into separate paragraphs or separate lines within a single <P> Tag. No attempt was made by Acrobat to identify any headings.

The bulleted item at the top of page 8 is tagged as a paragraph. On the whole, the document is readable but there are parts where words are either missing spaces between them or there are spaces between the characters in the words.

In reading paragraph by paragraph, some content was reread and focus shifted back to something I'd previously read and I had to jog JAWS out of its circular reading to continue. Not sure if this is a problem with the Tags or the screen reader. It shouldn't happen.

Foxit Phantom for Business (3 Page Document)

I used the JAWS screen reader to read both documents in Foxit Phantom.

Foxit Phantom for Business did add Headings, but it also included the words with spaces between the characters and mistagged content so that you got a piece of the first line in a paragraph followed by a piece of the second or third line in a paragraph and this continued throughout the greyscale and colour versions of the test document. This made the document unreadable using a screen reader (JAWS).

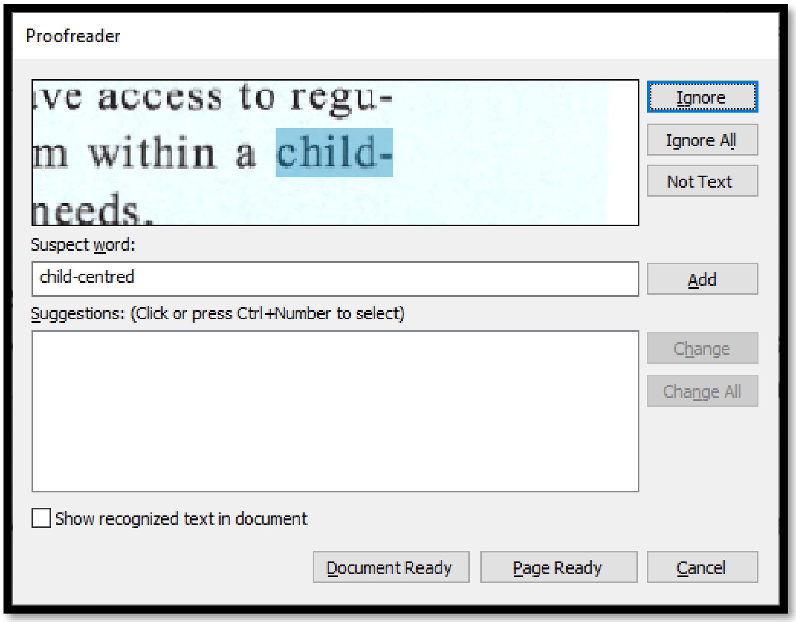
Foxit did not Tag lists properly but did separate the large blue numbers and made them Headings.

This document is not readable using a screen reader.

Nuance PowerPDF (3 Page Document)

I used the JAWS screen reader to read both documents in PowerPDF.

PowerPDF, which had been identifying the hyphenated words in other test documents, decided to identify a compound word with a hyphen. The word in the text is child-centered. The Proofreading dialog/tool identified “child” and suggested “child-centred” but the word “centered was not highlighted in the actual document. I didn’t know if that meant that if I chose the complete “child-centred” that I would end up with something like “child-centered – centered” or not. I chose to edit the suggestion to just “child” and see what happened.



Proofreader

ve access to regu-
m within a child-
needs.

Suspect word: child-centred

Suggestions: (Click or press Ctrl+Number to select)

☐ Show recognized text in document

Figure 41 PowerPDF Proofreading dialog showing hyphenated word correction.

It turns out that I made the wrong decision. The text at the bottom of page 7 now reads simply “child pedagogy” instead of “Child-centred pedagogy.” I would need to start over again with the non-recognized file in order to correct this.

PowerPDF did not identify any lists or Headings in the document. Everything is a paragraph or <P> Tag. Of the three applications (Acrobat, Foxit and PowerPDF) this document was the most readable even though there were no Headings or lists. The text was not either missing spaces between words or having spaces between characters in words.

Summary

The two applications/methods that gave me a better OCR and tagged PDF document were Nuance's PowerPDF Advanced and ABBYY PDF Transformer. With PowerPDF I could use the page that I had already scanned into the computer. With ABBYY PDF Transformer, I had to scan the page directly into Transformer in order to have the text recognized and tagged correctly. This is tedious if you have a document like this one that is 57 pages long.

Foxit gave me a pop-up asking if I wanted to perform OCR as soon as I opened a scanned document. Adobe Acrobat Pro DC and PowerPDF had me go into their recognize text tools to begin the process. Foxit Phantom "won" on this as I didn't have to figure out whether the document was scanned or not if I didn't use adaptive technology.

However, I couldn't use my screen reader with Foxit as I ended up in a loop of suspects and duplicate content in the Tags Tree. Foxit Phantom for Business also won't let me keep the Tags, Order and Content Panels in the Navigation Pane…which is a pain.

Nuance's PowerPDF makes you turn on the highlight Content each time you launch the program or open a file which is also annoying. The OCR tool was better and the process more consistent.

Adobe Acrobat Pro DC has two levels of OCR, "Recognize Text" and "Enhanced." However, neither of these tools gave me the improved results that PowerPDF and ABBYY PDF Transformer gave me. You have to try both Acrobat tools to see which is the best result. If one tool is better than the other in Adobe Acrobat, it is not clear why there are the two choices…wouldn't you always want to use the Enhance/better OCR?

The Adobe Acrobat pro DC OCR tools didn't let me find suspect text. This has been a problem since Adobe Acrobat Pro 8.

The original Salamanca Statement had spaces between characters in words which I hadn't seen before. This made the text unreadable. I have seen words with no spaces which is becoming more common when performing OCR using Adobe Acrobat Pro DC although I'm not sure why. I do realize that this might be a problem with the source file, but the example I used in the potential problems part of this tutorial was created in Word using the default font. There shouldn't have been any problem when the document was printed and scanned back into the computer as a sample OCR case. The printed copy did not have any blemishes.

Once I combined pages and examined the Tags Tree and read the document with the JAWS screen reader, Foxit Phantom was the worst in terms of including the spaces between characters and no spaces between words, then tagging individual pieces of paragraphs as separate <P> Tags. The order in which content was tagged was random so when using adaptive technology, you bounce around the document quite a bit.

PowerPDF produced only <P> Tags and couldn't even Tag a list in a document when you begin with a scanned image of a document. There were few, if any, characters in words wit spaces between them or no spaces between words. This was the most readable text of the test documents in either greyscale or colour but lacked any structure other than plain paragraphs..

Adobe Acrobat did not Tag anything as a Heading but did get the lists tagged correctly. However, it also included spaces between characters in words and no spaces between some words.

I'm not happy with any of the results. Each application has its strengths and weaknesses. However, it is worth mentioning that I have purchased 4 applications that can produce tagged PDF and there are no consistent results across them. Additionally, it is the applications that have access to powerful OCR engines that produce the best OCR results while Adobe Acrobat Pro DC and Foxit Phantom for business lag way behind.

Appendix A: List of Attached Documents

The following documents are attached to this PDF tutorial so you can explore the differences in accessibility yourself with adaptive technology. Do NOT use Adobe Read Out Loud! This is a tool that was added in Acrobat 6 to demonstrate what it might be like if you were using a Text-to-Speech tool and was never intended to be the sole adaptive technology used to access PDF documents! It either reads all of the document or the current page and has no capability to navigate by heading, paragraph, list or list item or table or table cell.

Colour Copy of Page 6

- Page 6 Colour.
- Page 6 Colour Acrobat.
- Page 6 Colour Foxit Phantom 01 Phantom (using a screen reader).
- Page 6 Colour Foxit Phantom 02 (without using a screen reader).
- Page 6 Colour PowerPDF.
- Page 6 Colour Transformer (page 6 scanned directly into Transformer).

Black and White Copy of Page 6 (Greyscale)

- Page6BlackAndWhiteAcrobat
- Page6BlackAndWhiteFoxit Phantom.
- Page6BlackAndWhitePowerPdF.
- Page6BlackAndWhiteTransformer (page 6 scanned directly into Transformer).

Three pages of The Salamanca Statement

Black and White Versions of the Three Pages

- Black and White Pages.
- Black and White Pages Acrobat.
- Black and White Pages Foxit Phantom.
- Black and White Pages PowerPDF.
- Black and White pages Transformer.

Colour Versions of the Three Pages

- Colour 3 Pages.
- Colour 3 Pages Acrobat.

- Colour 3 Pages Foxit Phantom.
- Colour 3 Pages PowerPDF.
- Colour 3 Pages Transformer (OCR done in Transformer and tagging in Acrobat).

The Complete Salamanca Statement Framework Document

The original Salamanca Statement and Framework was downloaded from the Internet[3].

The Salamanca Statement and Framework Original PDF document.

The Salamanca Statement and Framework original document taken through ABBYY PDF Transformer and converted to a Word document.

[3] Salamanca Statement and Framework for Action: https://www.right-to-education.org/resource/salamanca-statement-and-framework-action-special-needs-education

Appendix B: Contact Information

Karen McCall
info@karlencommunications.com

The Karen McCall School on Teachable[4] has for purchase items such as self-paced online courses and books on accessible Word, PowerPoint and PDF. The free tutorials are for Office 365 subscription (Office 2016 and 2019). You have to sign into the “school” but I only use the ability to send you e-mail to let you know that there is new content…no marketing.

The Karlen Communications website[5] has older tutorials pre Office 365 and older conference handouts as well as links to webinar archives.

All material, free or for purchase is in accessible tagged PDF. Videos in the self-paced courses are captioned.

[4] Karen McCall School on Teachable: https://karen-mccall.teachable.com/

[5] Karlen Communications website: https://www.karlencommunications.com/