Demo Doc PDF Clean Up

Guide Overview

This is the second part of a multi-part guide that will guide you through the steps to clean up a PDF exported from Microsoft Word. All the guide parts and deliverables, listed below, can be downloaded from nyc.gov/accessibilityguides.

- Word to PDF-1 Accessible Doc Creation.pdf
- Word to PDF-2 Demo Doc (Deliverable).docx
- Word to PDF-3 Demo Doc--as Exported (Deliverable).pdf
- Word to PDF-4 Demo Doc PDF Clean Up Steps.pdf
- Word to PDF-5 Demo Doc--Cleaned Up (Deliverable).pdf

Acrobat Pro Basics

ⓘ First thing’s first. Sanity management. I strongly urge you to do a Save As and add an annotation like “--Cleaned” to the file name before you begin any remediation. This way you can always start over if you make an un-doable mistake. Also, I urge you to do frequent Saves. This will allow you to revert the file without losing too much. One thing to note about Acrobat, it only allows you to undo the last thing you did, sometimes, and only for a short time.

Within Acrobat, there is a Navigation Pane along the left side of the window and a Task Pane along the right side of the window. These Panes can be expanded and collapsed by clicking the left or right facing triangle about midway down the side of the window.

The Navigation Pane contains buttons that allow a user to navigate through the document in different ways. For our purposes, we will be using Content and Tags. If you do not see one, or both, of these buttons, right click on the expanded Pane and choose the missing item from the dropdown menu that appears.

Content shows a tree view of the content of the document. Whether the document contains tags or not, Content will always have items. Tags provides a tree view of the ‘tagged’ content of the document. Tagging content is what creates an accessible document.

The Task Pane allows users to interact with a document’s content. As I remediate PDFs, I generally use the Edit PDF and Create Forms tasks. You only need to use these if there is a small mistake in the text, or if you are creating a fillable PDF form. For our purposes, we will not need to use either.

Document Properties

Under the file menu, go to ‘Properties’ and check that the following document properties have been set. If the Word doc was created with accessibility in mind, the only thing that should need to be changed is item 3, change ‘File Name’ to ‘Document Title’.

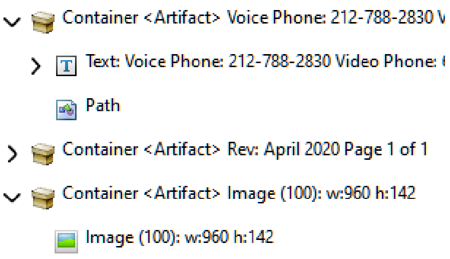
- Description --> Title: (Ex: Word to PDF Demo)
- Initial View --> Page Layout: Single Page Continuous
- Initial View --> Window Options, Show: Document Title
- Advanced --> Language: English

Artifacts and the Content Pane

The first place to start is to deal with the Artifacts in the document. This is done from the Content Pane. Click on Content in the Navigation Pane (left side of the window, look for the icon that looks like a piece of paper, you may have to right-click on the bar to find it). Then, expand the tree until you see all the tags under ‘Page 1’. This will show how all of the content in the document has been tagged (if it has been tagged). The first two Containers after Annotations are the Artifacted header and footer.

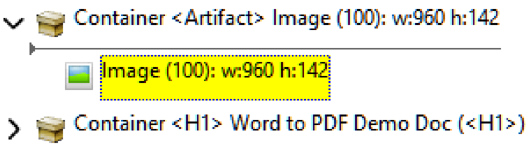
A note about headers and footers. Depending on the document, you’ll need to decide if any of the information in the header or footer should be included. Generally, page numbers and document titles do not need to be included. Also things like chapter or section names do not need to be included. Logos, contact information, and additional notes or revision dates should be included. In multipage documents, included information should only appear once. Generally, you would only need the header on the first page and the footer on the last page. All others can stay Artifacted.

So, the first thing to do is to 'unartifact' the header image and contact info. Start by clicking the greater than symbol next to each Artifact Container for the header so that it expands. Notice within the first Artifact, there is the text of the contact info and a 'Path' item. 'Path' items represent vector graphics within the document. In this case, it is the underline from the website link. Any time there is underlined text or the outline of a table, there will be Paths. These will almost always need to be Artifacted.


Container <Artifact> Voice Phone: 212-788-2830 V
Text: Voice Phone: 212-788-2830 Video Phone: (
Path
Container <Artifact> Rev: April 2020 Page 1 of 1
Container <Artifact> Image (100): w:960 h:142
Image (100): w:960 h:142

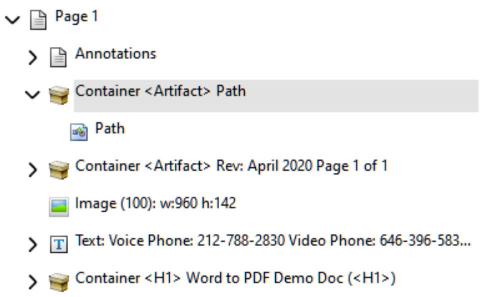
ⓘ Nore: Some logos or graphics will be made of Paths, so, unless they are decorative, they will need to be grouped into a <Figure> tag.

Now, drag the image element up so that the position indicator is just below its container and aligned with the container's icon. This moves the element out from under its parent container. You can then delete the empty Artifact container.

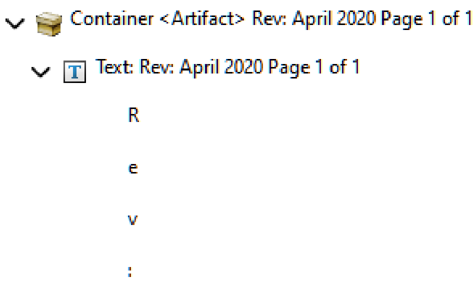

Container <Artifact> Image (100): w:960 h:142
Image (100): w:960 h:142
Container <H1> Word to PDF Demo Doc (<H1>)

Next, drag the text element with the contact info down from the Artifact container so that it is just above the <H1> container. These are now untagged elements. We will tag them later in Tags.

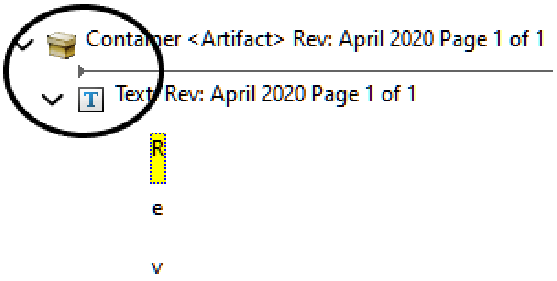
The Content tree should look like this. The first Artifact container should only have a Path element; the second Artifact container is untouched; the third Artifact container was emptied and then deleted. Leaving the untagged image and text elements just above the <H1> container.


Page 1
Annotations
Container <Artifact> Path
Path
Container <Artifact> Rev: April 2020 Page 1 of 1
Image (100): w:960 h:142
Text: Voice Phone: 212-788-2830 Video Phone: 646-396-583...
Container <H1> Word to PDF Demo Doc (<H1>)

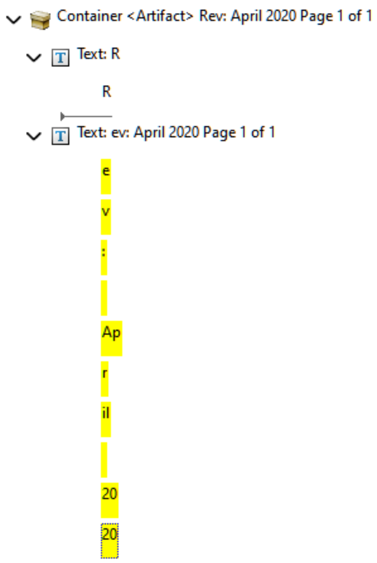
The next Artifact issue is in the footer text. If you click on it, you'll notice that everything in the footer is highlighted (look at the document to the right). If the text isn't highlighted, right-click on any tag in the Content Tree and make sure 'Highlight Content' is checked. So, we need to separate "Rev: April 2020" from the rest of the text in the page footer (a space and the page number). To do this, expand the 'Text: Rev…' element to expose its contents.


Container <Artifact> Rev: April 2020 Page 1 of 1
Text: Rev: April 2020 Page 1 of 1
R
e
v
:

You'll notice Acrobat breaks text up in funny ways, so don't be surprised if yours doesn't look exactly like this. Click on the first thing in the text element (in this case the capitol R), and drag it up so it is above the text element, but still under the Container Tag. Be mindful of the location of the placement indicator, you'll want the left edge to be aligned with the 'C' in 'Container'. This will create a new Text element that contains just the 'R'.


Container <Artifact> Rev: April 2020 Page 1 of 1
Text Rev: April 2020 Page 1 of 1
R
e
v

Expand that new element and click on the next part of the expanded footer text ('e' in this example). While holding the ***Shift*** Key, click on the last chunk of the remaining text we want to separate, the second '20'. This will select all of the elements between the first selection and the last. Now, you can click and drag them up so that they are just below the 'R'. Again, be mindful that the placement indicator is aligned with the 'C'. Once the revision text has been isolated in its own Text element, you can collapse all of the text elements.


Container <Artifact> Rev: April 2020 Page 1 of 1
Text: R
R
Text: ev: April 2020 Page 1 of 1
e
v
:
Ap
r
il
20
20

Now, drag the 'Page 1 of 1' text element up to the Artifact container with the Path element. Then, drag the 'Rev…' text element down to the bottom so that it is the last item in the Content Tree. The empty Artifact container can be deleted. The first few containers under 'Page 1' should be as follows: Annotations, Artifact, Image…, Text: Voice…, H1, and H2. And the last one should be the 'Rev: April…' Text element.

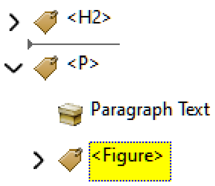
Finally, all of the Path elements within the document need to be placed within the Artifact container. Do this by expanding any Container with underlined text (sometimes the Path element is in the container after the underlined text, or untagged in the Content Tree) and dragging the Path element up so that is within the Artifact container. You will find these Path elements in the <H3>s and link text of this document. Once done, collapse the Artifact Container.

Order of Content

Before leaving the Content Pane, you'll want to make sure the order of the Containers and other untagged elements matches the flow of the document. The first two items will always be Annotations followed by the Artifact Container. Down arrow through the list of items starting after Artifact and ensure the proper flow.

For instance, the Figure Container near the bottom should be just before the paragraph that wraps around it, not after it.

Tags



The Tags Pane (on the left side of the window, the icon that looks like a hanging price tag) is where most of the accessibility magic happens. A quick note, the order of the Container tags in the Content Pane should match the order of content in the Tags Pane. The Content Pane only shows the lowest-level tags, so don't worry if it doesn't have the same amount of nesting. There is one issue with tag order, in the last <P> tag. If you expand it, you'll notice there is a <Figure> tag snuck in there. To fix, simply drag the <Figure> tag up so that it is right above the <P> tag, and then collapse the expanded tag.

Because of all the work done in the creation phase in Word, all that needs to be done with the Tags is to add the untagged elements from the header and footer to the tag tree, to get rid of excess <Span> tags, and to ensure the proper tag structure within the nested elements of the list.

Adding New Tags

When tagging content in the tag structure, you will be moving between the Content and Tags Panes. When adding a new tag, it will be inserted at the same level as the highlighted tag, and just below it.

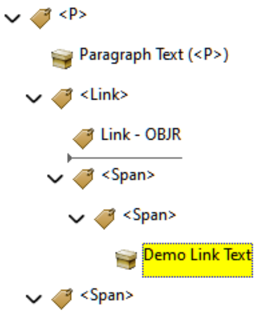
The first tag we are going to add will be located just above the <H1> tag. Since the <H1> is the first tag, we'll insert the new tag below, then drag the tags so that they are in the correct order. So, right click on the <H1> and choose 'New Tag'. From the drop down list in the dialog box that appears, choose Figure, and click Ok. Now, drag the <H1> tag so that it is below the new <Figure> tag, keeping it aligned with the <Figure>, and not within it.

Next, we will choose the content we want in that new empty tag. First, be sure that the <Figure> tag is highlighted. Then, click on the Content Pane button and select the untagged Image element just below the Artifact container. Next, go back to the Tags Pane and right click on the new <Figure> tag and choose 'Create Tag from Selection'. You'll notice the image element from the header is added under that new tag. Since the MOPD Logo was Artifacted upon conversion from Word, the alt text was lost. So, right click on the new <Figure> tag and choose Properties at the bottom. In the Alternate Text field enter "M.O.P.D. Logo". The periods force the screen reader to pronounce each letter, rather than trying to pronounce it as a word.

Follow the same procedure to add a new paragraph <P> tag after the <Figure> tag and add the contact info to it. Lastly, add a <P> tag to the bottom of the tag tree and fill it with the revision text.

Excessive Tagging

Word has a tendency to over tag, particularly when it comes to links and specially formatted text—like superscripts. You'll see this in the first paragraph after the <H3>. Hold down the ***Ctrl*** key and click on the greater than symbol next to that <P> tag after the <H3>. This will expand all of the tags within that Paragraph tag. Notice there are two sets of nested <Span> tags. To get rid of these unnecessary tags, start by dragging the link text ('Demo Link Text') up so that it is just below the 'Link - OBJR' element. Now, delete the empty <Span> tags and collapse the <Link> tag. In the second set of nested <Span> tags, do the same. Move the text element up so that ir is just below the <Link> tag and delete the now empty <Span> tags. The <P> tag can now be collapsed.



- <P>
  - Paragraph Text (<P>)
  - <Link>
    - Link - OBJR
    - <Span>
      - <Span>
        - Demo Link Text
  - <Span>

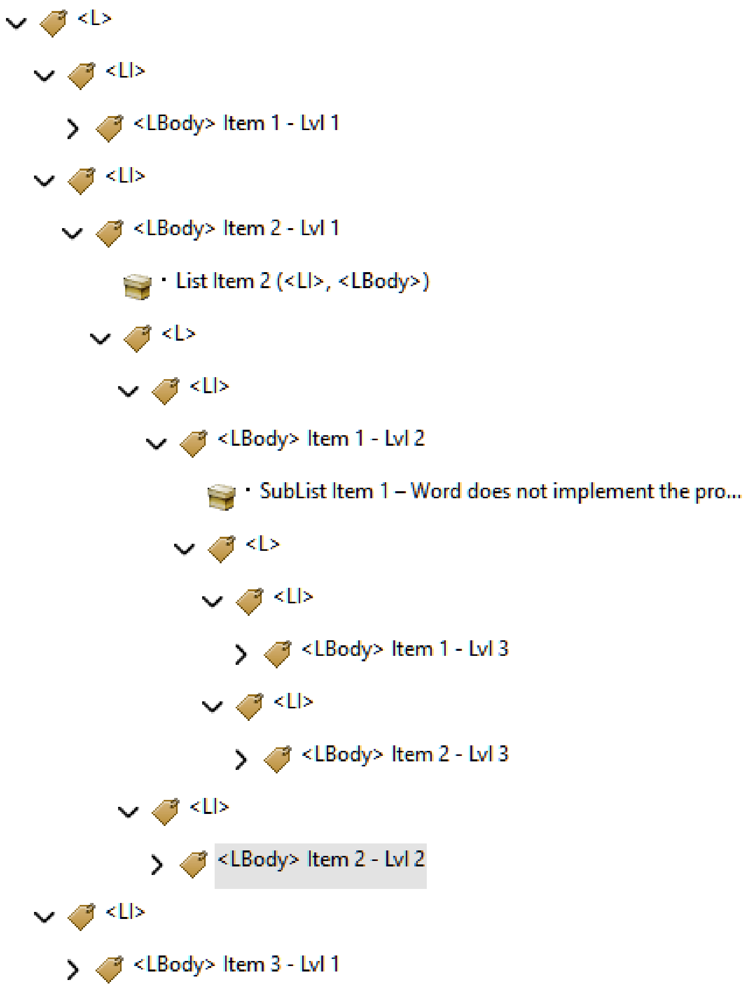
Links

Most links will be automatically tagged when converting the Word document to a PDF. The one exception will be any links in the header or footer. So next, you'll need to create the link for the website in the header. Start by clicking on an empty space in the document to the right, to ensure nothing is highlighted. Then, click and highlight the website address in the header, right-click over the highlighted text, and choose copy. Right-click again over the highlighted text and now choose 'Create Link'. Within that dialog, most settings can be left alone, under 'Link Action' choose 'Open a web page' and click 'Next'. In the next dialog, simply paste the copied web address, appending "http://" to the beginning of the copied text. Finally, click 'Ok'. You'll now see a tagged <Link> in the first <P> tag.

Lists

Word will tag lists for you, though it is imperfect. For instance, when there are nested lists, the correctness of the tags is hit or miss and depends largely on the way the list was created in Word. In this document, Word's native list formatting was used, with a tab to indicate lower levels of nesting. Below is the way this document's nested list was tagged. This is the correct tagging.

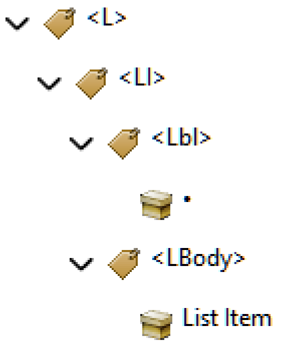
Correct Nested List Tagging



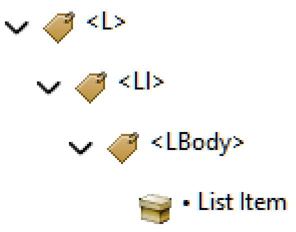
```
<L>
  <LI>
    <LBody> Item 1 - Lvl 1
  <LI>
    <LBody> Item 2 - Lvl 1
      • List Item 2 (<LI>, <LBody>)
      <L>
        <LI>
          <LBody> Item 1 - Lvl 2
            • SubList Item 1 – Word does not implement the pro...
            <L>
              <LI>
                <LBody> Item 1 - Lvl 3
              <LI>
                <LBody> Item 2 - Lvl 3
        <LI>
          <LBody> Item 2 - Lvl 2
  <LI>
    <LBody> Item 3 - Lvl 1
```

A quick note: I added the labels next to each <LBody> tag to provide clarity, it is not needed otherwise. Notice, the general tagging structure Word uses for lists is as follows: <L>, <LI>, <LBody>. There is an additional tag that can be used, <Lbl>, to 'house' the bullet or number. Both <Lbl> and <LBody> are optional tags.

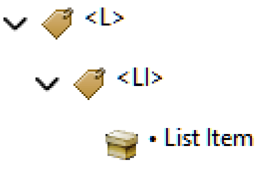
Here are the three valid PDF list tagging structures:



```
<L>
  <LI>
    <Lbl>
      •
    <LBody>
      List Item
```



```
<L>
  <LI>
    <LBody>
      • List Item
```



```
<L>
  <LI>
    • List Item
```

Wrapping Up

The last thing to do is a check of the tag order. Select the <Figure> tag and arrow down while looking at the document to the right. Make sure everything flows as it should. If something is out of order, just adjust its order in the Tag Tree.