Chapter 5

Scanning Print to PDF

Opportunities and Obstacles for Screen Reader Accessibility

Robert Browder*

Scanning print to PDF opens a world of opportunity for sharing, using, and reusing resource materials. Here at Virginia Tech's Newman Library, we've been able to bring previously unavailable publications to the web in PDF format, including out-of-print journals and historical documents. Making resources available online in an accessible format creates opportunities for patrons that were not there before. Patrons can have their own copy of a document at the touch of button. After being rendered as an accessible PDF, resources that previously existed only in print take on new utility; they can be read aloud by a computer. This is a wonderful opportunity for all patrons, but especially for those with visual impairments.

How is it that PDF has remained so popular with the emergence and maturity of other digital reading technologies? In 2008, following many years of practical use and popularity, Adobe Systems, creator of the PDF file format, released the file specification to the International Organization for Standardization (ISO) for management and expansion.[1] Adobe did this in response to heavy use of the format by governments and public organizations. Releasing the specification to the ISO brought the PDF format into the world of "open technology" and cemented the confidence of public institutions. For the typical user, PDF provides a reading experience that is "near-book" by providing an application interface that creates a firm boundary from all the distraction that is the modern web browsing experience. With the combination of focus and flexibility provided by PDF format, it's really no wonder that it continues to thrive.

The ability to use semi-automated processes to create PDF documents from printed materials has obvious time-saving advantages. With the right equipment, you can scan fifty to ninety printed pages per minute. However, merely scanning printed materials as images is not enough. While creating a digital image of text on a page is a great leap in preservation and "sharability," a wide variety of vision issues may affect any of us at some point in our lives, rendering visually oriented materials difficult or impossible to use. Making PDF documents accessible to those with visual disabilities via screen reader technology is well within the reach of our current technical abilities. However, scanning print to PDF is not a panacea to create accessibility for all types of content. While it is perfect for some types of content, more complex types of content prove to be remarkably difficult and time-consuming to render screen-reader-accessible in PDF.

Scanning print to PDF presents unique opportunities and challenges. The source material to be scanned will determine how much effort is required to make a PDF accessible. For complex content like large tables, graphs, charts, and equations, HTML often provides better opportunities for accessibility and production efficiency than is possible with PDF. For simpler content, such as text and images that can be described with ease verbally, scanning print to PDF is often the most streamlined approach to creating an accessible resource from printed materials. In nearly all cases,

* **Robert Browder** is a digital publishing specialist with VT Publishing, a service of Virginia Tech Libraries. Since obtaining his undergraduate degree in information science and systems in 2011 from Radford University, Browder has served in a variety of technology and publishing roles. His work currently focuses on managing resources and workflows associated with the publication of online open-access scholarly journals.

PDF makes a suitable "pass-through" and preservation format to bring print into digital format while avoiding manual transcription processes.

Understanding Visual Disability

When we think about *visual disability* as a general term, we are addressing a community of conditions that have different causes but often share similar functional limitations. Visual impairment includes everything from complete blindness to conditions that merely require corrective lenses. Conditions like low vision, color-blindness, and corneal opacities each have their own limitations.

The World Health Organization groups moderate to severe vision impairment under the term *low vision*.[2] The majority of conditions categorized as low vision can be improved with the use of corrective lens. However, in the absence of corrective lens, low vision can make it incredibly difficult for individuals to read and perform daily tasks.

Color-blindness results in perceptions of colors that differ from the way the majority of the population perceive them. Three forms of color blindness are currently documented: red appears as green, blue appears as yellow, and complete absence of color vision. According to the National Eye Institute, "As many as 8 percent of men and 0.5 percent of women with Northern European ancestry have the common form of red-green color blindness."[3] As you might imagine, color-blindness creates unique challenges for interpreting color-coded information.

Globally, most cases of blindness can fit within a few categories. Corneal opacities (CO), clouding of the cornea, are often the result of infections but can also result from injury. Age-related macular degeneration is a progressive degeneration of a person's main field of vision due to lesions of the retina. Glaucoma is caused by optic neuropathy, in which messages from the eye are either not conducted or poorly conducted to the brain. Cataract is a clouding of the lens that prevents light from entering the eye.[4]

The World Health Organization reports that 253 million people live with visual impairment of some kind.[5] While creating accessible digital materials does not solve the root problem, it does make information available to those who otherwise would not have it. Consider the benefit you get from reading an article you are interested in and multiply it by 253 million. That's real opportunity there.

The Scanning Process

Scanning print to PDF is a process that is used regularly at Virginia Tech's University Libraries. The scanning process is the heart of our print-to-PDF pipeline. Combined with a reliable optical character recognition (OCR) process, automated scanning provides extraordinary efficiency. Christy Stanley, Virginia Tech University Libraries scanning specialist, uses a process consisting of the following basic steps:

- Prepare for automated feed scanning
  - Organization of materials
  - Removal of spine for bound materials
- Scanning
  - Loading and monitoring the scanner
- Adjust scanned pages
  - Adjust for skew
  - Crop pages to remove ragged edges
- Compile scanned pages into PDF documents
- Color balance adjustments
  - Setting the text to black makes it much more legible for those with low vision and may improve the quality of OCR output.
- OCR
  - This step can be done either at the end of the scanning process or at the beginning of the read order editing process.

We have a couple of Fujitsu scanners, the 6240-Z and the 6770. Both have an auto-feed tray and a flatbed. The 6770 will handle larger pages and will scan more pages per minute. There are lots of options for scanners made by familiar brands like Kodak, Canon, and HP that provide functionality similar to these. If you're thinking about buying a scanner and your library is already invested in equipment from a particular vendor, it may make sense to get their stuff in the hope that all components will play together nicely. Most scanners come with software that may be helpful in building or refining the scanning process.

After pages are scanned, post-processing can be achieved using vendor software that came with the scanner or an open-source tool like Scan Tailor. Post-processing allows the technician to straighten crooked pages, adjust the color balance, remove unsightly edges, and group a collection of scanned pages into a single PDF document.

The importance of adjusting the color balance of a document should not be ignored. Color balance adjustments can often increase color contrast of typography, yielding notable improvements in readability for those with low vision. OCR processes may also benefit from color balance adjustments.

Color can be an important part of any visual communication and can a have serious impact on accessibility. Color contrast is important for users with low vision or color blindness. Color is often used to convey meaning and communicate essential information. Nowhere do we see this more clearly than in the

example of charts and graphs. Colors without appropriate contrast may render bar graphs and charts difficult to use. This situation must be considered carefully when scanning documents that contain graphics that use color to communicate. Alternative text (alt text) can be used to add meaning to images that have poor color contrast.

Setting Expectations for Output of the Optical Character Recognition Process

OCR is a process that uses computer algorithms to analyze and identify letter shapes and words. OCR can be achieved with Adobe Acrobat or, in some cases, with software that came with the scanner. An OCR process adds character encoding to the document so that screen readers can read the document to users with visual impairments. OCR also allows users to copy and paste text from the document. While current OCR technologies do pretty well with recognizing standard fonts, OCR algorithms will be confounded by poor quality scans, decorative typefaces, and handwriting. So the output of an OCR process can be only as good as the input. Keep this in mind when setting expectations for print-to-PDF projects.

Testing the Output

After OCR, the document must be tested. A screen reader such as JAWS, NVDA, or VoiceOver will be very useful. Adobe Acrobat will also prove indispensable. Using screen readers allows us to know something about the experience the document will provide to those with visual impairments. Using Acrobat will provide us with a window into the technical organization of the scanned document.

As a first step, it is always useful to test the document with a screen reader. Listen carefully as the application reads the text to you. Pay attention and make note of any inconsistencies. Since this document is newly scanned, we can expect that some content, such as images or other graphics, will be skipped over by the screen reader. We may also find that content is not always read in the correct order. We may find that artifacts of the document, such as running headers or footers, are read by the application when they should not be. While the scanning and OCR processes have saved a great deal of time, we may find that the resulting screen reader output is intelligible but not intelligent. Human intervention is typically required to organize the document in such a way that its full context can be conveyed via screen reader.

Abode Acrobat provides tools that can be used to analyze and edit the underlying structure of PDF documents. Developing familiarity with this tool kit is a marathon, not a sprint. Consistent time investment in developing skills with this tool set will yield best results. Acrobat offers various levels of automation for different tasks that are helpful in creating accessible documents, including tagging, accessibility checking, alternative text, and reading order.

Tagging PDF Documents

One of the most important steps in creating accessible PDF documents is tagging. Tagging allows us to define different elements within the document. Common elements that need to be tagged are headings, paragraphs, and images. Screen readers use these tags to assist the reader in using and navigating the document. Acrobat provides automation for this task that is marginally helpful. The automated process will often tag artifacts that should be ignored by the screen reader. Quality will be improved with manual review and editing.

Alternative Text for Images, Figures, Graphs, and Charts

Alternative text is descriptive text that can be added to a document to replace images, figures, math, graphs, and charts when the document is read by a screen reader. Alt text fills in the blanks that otherwise result from unseen images. How well alt text fills in those blanks is another story. Ideally, alt text would be supplied by the author of a text, but in the case of scanning, this is usually impossible. Alt text must be created by someone who understands the context and content of the images. With simple images and figures, filling in the alt text is a simple task. With complex charts and graphs, creating alt text that reliably communicates the information becomes a specialty that may require a subject specialist.

While alt text is intended to create an experience that is comparable to interacting with the document visually, whether or not it actually does is, in many cases, debatable. HTML is often a better format for complex graphs and charts. Tactile graphics can create a truly comparable experience for the visually impaired.

Read Order Editing

Read order is the order in which a screen reader will read the contents of a PDF to a human listener. While a human reader will evaluate a page using visual cues, a screen reader needs to have the read order explicitly defined. Acrobat can automate the process of

assigning read order to a PDF document. But it cannot determine which elements add meaning to the work or the correct reading order for content found in complex layouts. Ideally, read order should be comparable to the way a human would read a text.

For example, let's consider a typical page that contains a running header in the top right corner of the page with page number and several paragraphs of text in the body of the page. Even with just these few elements on the page, there is possibility for improperly assigned read order to disrupt the flow of the text and its meaning. Let's suppose that an automated read order assignment has defined the running header as the first element on the page and the paragraphs in the body text as the second, third, and fourth elements. At first glance this may seem fine, but what if the first sentence on the page is a continuation of the last sentence on the previous page? If the screen reader reads the running header first, it will break the flow of the sentence and possibly confuse or distort its meaning. This is a serious quality issue that can create problems for users of the resource. The solution to this problem is to manually edit the read order of the document. The correct action in this case would be to define all of the running headers with page numbers throughout the entire document as background and assign the first paragraph in the body text as the first element on the page. This approach maintains the flow and meaning of the content.

Tables

Tables are a special challenge for the print-to-PDF process. Simple tables are easy enough to tag and use with a screen reader. The simplest of tables can even be tagged as a figure and amply described with alt text. Larger tables are challenging and time-consuming to tag in PDF. I argue that even the most detailed tagged tables do not provide a comparable experience for those with visual impairments. Let's take a moment to remember what a table is and what it is supposed to do. A table is a tool that creates a matrix that allows the user to explore data relationships in a two-dimensional format, columns and rows. The matrix functionality that makes a table such a valuable tool for presenting information can be severely diminished by representing it verbally. Trends and patterns that are obvious when using the table either visually or tactilely may be much more difficult to identify when attempting to explore the information verbally. The goal is to share information revealed by looking at the relationships of data organized within the matrix. HTML, braille, and tactile graphics are often better formats for this type of complex content.

What about Math?

While an OCR process can interpret characters and group them into words and sentences, generating a consistent screen reader experience for mathematical equations is a bit beyond what can reasonably be expected from the OCR process. If equations are included as part of a sentence, they may or may not come through reliably. In the case that equations are presented on their own apart from the text, they can be tagged as figures and have a verbal description added as alt text. This approach is especially helpful with complex multilevel equations that use special characters and symbols, such as Greek letters.

Tagging equations as figures and adding alt text is a reasonable way to treat equations in the print-to-PDF process; however, it raises another issue. The alt text must be *meaningful*. A subject specialist who understands how to correctly communicate the equation with a text description will be required.

Opportunity or Obstacle? It's All about the Content

While scanning print to PDF is a great opportunity for preserving and sharing printed materials of all types over the web and is a great entry point for bringing print into the digital space, the ability to produce a PDF that is highly accessible for screen readers is not always straightforward and often requires excessive inputs of time and specialized skills. The complexity of content in the source material is the deciding factor in how well a PDF document can meet the rigorous demands of screen reader accessibility. PDF offers wonderful opportunities for plain text and images. We start to run into obstacles when documents contain more complex types of content like tables, graphs, and complex math equations. While all of these content types perform fine visually, developing the PDF document to the point that it provides a comparable experience delivered verbally via screen reader is time-consuming and often requires input of specialized skills.

To Scan or Not to Scan?

Yes, by all means, scan. But know your goals, know the limitations of a print-to-PDF scanning process, and set expectations accordingly. PDF readily satisfies goals for preservation and dissemination of visually accessibly materials. PDF also performs well as a pass-through format to aid in avoiding manual transcription. PDF can sometimes satisfy the needs of accessible documents, depending on the types of content found in the document. Before a scanning project

begins, it is important to consider whether or not the complexity of the content can be faithfully communicated via PDF with screen reader technology and how much effort will be required to organize PDF documents for screen reader accessibility.

Notes

1. "PDF Format Becomes ISO Standard," International Standards Organization, July 2, 2008, https://www.iso.org/news/2008/07/Ref1141.html.
2. "Vision Impairment and Blindness," fact sheet, World Health Organization, last updated October 2017, www.who.int/mediacentre/factsheets/fs282/en/.
3. "Facts about Color Blindness," National Eye Institute, National Institutes of Health, last updated February 2015, https://nei.nih.gov/health/color_blindness/facts_about.
4. "Priority Eye Diseases," World Health Organization, accessed March 8, 2018, www.who.int/blindness/causes/priority/en/index8.html.
5. "Vision Impairment and Blindness."