Yongli Zhou

Are Your Digital Documents Web Friendly?: Making Scanned Documents Web Accessible

The Internet has greatly changed how library users search and use library resources. Many of them prefer resources available in electronic format over traditional print materials. While many documents are now born digital, many more are only accessible in print and need to be digitized. This paper focuses on how the Colorado State University Libraries creates and optimizes text-based and digitized PDF documents for easy access, downloading, and printing.

To digitize print materials, we normally scan originals, save them in archival digital formats, and then make them Web-accessible. There are two types of print documents, graphic-based and text-based. If we apply the same techniques to digitize these two different types of materials, the documents produced will not be Web-friendly.

Graphic-based materials include archival resources such as historical photographs, drawings, manuscripts, maps, slides, and posters. We normally scan them in color at a very high resolution to capture and present a reproduction that is as faithful to the original as possible. Then we save the scanned images in TIFF (Tagged Image File Format) for archival purposes and convert the TIFFs to JPEG (Joint Photographic Experts Group) 2000 or JPEG for Web access. However, the same practice is not suitable for modern text-based documents, such as reports, journal articles, meeting minutes, and theses and dissertations. Many old text-based documents (e.g., aged newspapers and books), should be treated as graphic-based material. These documents often have faded text, unusual fonts, stains, and colored background. If they are scanned using the same practice as modern text documents, the document created can be unreadable and contain incorrect information. This topic is covered in the section "Full-Text Searchable PDFs and Troubleshooting OCR Errors."

Currently, PDF is the file format used for most digitized text documents. While PDFs that are created from high-resolution color images may be of excellent quality, they can have many drawbacks. For example, a multipage PDF may have a large file size, which increases download time and the memory required while viewing. Sometimes the download takes so long it fails because a time-out error occurs. Printers may have insufficient memory to print large documents. In addition, the Optical Character Recognition (OCR) process is not accurate for high-resolution images in either color or grayscale. As we know, users want the ability to easily download, view, print, and search online textual documents. All of the drawbacks created by high-quality scanning defeat one of the most important purposes of digitizing text-based documents: making them accessible to more users.

This paper addresses how Colorado State University Libraries (CSUL) manages these problems and others as staff create Web-friendly digitized textual documents. Topics include scanning, long-time archiving, full-text searchable PDFs and troubleshooting OCR problems, and optimizing PDF files for Web delivery.

Preservation Master Files and Access Files

For digitization projects, we normally refer to images in uncompressed TIFF format as master files and compressed files for fast Web delivery as access files. For text-based files, access files normally are PDFs that are converted from scanned images.

"BCR's CDP Digital Imaging Best Practices Version 2.0" says that the master image should be the highest quality you can afford, it should not be edited or processed for any specific output, and it should be uncompressed.[1] This statement applies to archival images, such as photographs, manuscripts, and other image-based materials. If we adopt the same approach for modern text documents, the result may be problematic. PDFs that are created from such master files may have the following drawbacks:

- Because of their large file size, they require a long download time or cannot be downloaded because of a timeout error.
- They may crash a user's computer because they use more memory while viewing.
- They sometimes cannot be printed because of insufficient printer memory.
- Poor print and on-screen viewing qualities can be caused by background noise and bleed-through of text. Background noise can be caused by stains, highlighter marks made by users, and yellowed paper from aged documents.
- The OCR process sometimes does not work for high-resolution images.
- Content creators need to spend more time scanning images at a high resolution and converting them to PDF documents.

Web-friendly files should be small, accessible by most users, full-text searchable, and have good

Yongli Zhou is Digital Repositories Librarian, Colorado State University Libraries, Colorado State University, Fort Collins, Colorado