Scanning for Digitization Projects

Larry Wentzel

Librarians and archivists find themselves facing the prospect of digitization. Everyone is doing it, everyone needs it. Discussions rage nationally and internationally concerning what to digitize and the best means to present and retain digital objects. Yet newcomers may seek answers to simpler questions. What is digitization? What does digitization equipment do? What do digitization standards mean?

Digitization is the act of making something digital, expressing a physical object "in numerical form, especially for use by a computer" (The American Heritage, 2000). In the similar manner that paintings express landscapes and faces through colored pigments, digital files re-present objects as numbers. It is, to misuse a phrase, painting by numbers.

Why should an object be represented by numbers? What is gained by doing so? The aim of expressing an object in numbers is that it can be stored and manipulated by computers. Computers are number crunchers, performing millions of calculations per second. By digitizing an original and placing a digital copy of it on a computer, the file can be manipulated, transferred, and stored with ease. Storage and distribution are the primary factors behind the national push for digitization. Storing a numeric or digital representation of photographs on a CD takes up far less physical space than the boxes and photo albums stored on shelves. Being able to access digital copies of material across the web allows patrons greater access to the content without increased wear and tear on the original.

Flatbed scanners, digital cameras, and other digitization equipment

Digitization equipment – such as flatbed scanners, digital cameras, and digital audio/video recorders – generates digital copies of physical objects. Flatbed scanners are, in short, desktop photocopiers and they are mostly mechanical devices[1,2]. A lamp moves slowly across the face of the original. The lamp shines light onto the original, and the reflected light is focused through a series of mirrors and lens onto the recording medium. In flatbed scanners, the medium is a compact light sensor, either a CCD (charged coupling device) or CIS (contact image sensor), each of which is composed of hundreds or thousands of elements. When light strikes each element, the intensity of the light is assigned a number. The numeric reading of light intensity and element position are recorded in sequence into a file, which forms the digital version of the original.

Additional hardware can enhance the scanning process, but do not affect the basic function. Transparency adapters make it possible to digitize slides, negatives, and transparencies. Rather than reflecting light off the surface of a transparent original, a transparency adapter shines light through the original and onto the CCD/CIS. Automatic document feeders increase the speed of the scanning process by handling the placement and removal of paper originals from the glass plate, reducing the delay between scans. The downside to automatic document feeders is that the originals must be loose (in the case of books, pages must be disbound) and able to withstand the physical stress of being run through the feeder.

Regrettably, manufacturers of consumer-grade scanners also adorn them with buttons displaying icons for email, photos, text, and the printer. Pressing each button activates a preset scanner setting, causing the scanner to scan an original and format it for email, pictures, text documents, or the printer. These buttons make the process of digitization easy; they remove any need for the scanning technician to understand the process. Push the button, it's made to order. Unfortunately, none of the buttons installed on flatbed scanners are preset for "digital archive" quality. In order to get the quality recommended by the Digital Library Federation (www.diglib.org/, scanning operators need to understand what the buttons do and how they can do it better.

Scanning process

While the flatbed scanner does the actual digitization, the device has no understanding what operators want for output nor does it store the files or perform alterations. Similarly, software like Adobe Photoshop and Microsoft Word do not operate the scanner; they are used to create and edit image files or documents.

Specialized software – the scanner driver – is needed to bridge the gap: scan drivers operate the scanner and transfer the digitized file to the hard drive or software. The scan driver may be a standalone driver or a plugin, a specialized version of the driver that is accessible through Photoshop, Word, or other programs. The standalone driver runs the scanner without involving other software and saves the file to the hard drive. Plug-ins are opened within Photoshop or Word (by choosing to Import or Insert pictures from the scanner) and after scanning, hand the files directly to Photoshop or Word for immediate use (without saving it first).

Scan drivers fall into two groups: native and third-party. Flatbed scanner manufacturers provide their own native driver for their scanner and provide updates for the driver through their web site. Native scan drivers are as sketchy or robust as the manufacturer deems necessary, and features can vary from model to model. In the case of specialized scanners, such as overhead book scanners or digital cameras, the native driver is the only driver available.

Third-party software makers focus entirely on creating scan drivers which are oftentimes more robust and offer better control over the scanner and scanned image than the native drivers. Third-party scan drivers must be purchased separately from the scanner, unless the scanner manufacturer has bundled it with their scanner as an incentive to consumers. An exception to this rule is the Windows Image Acquisition (WIA), a third-party scan driver provided in Microsoft Windows XP. Unlike other third-party scan drivers, WIA offers the most commonly available features used by all flatbed scanners (specialized scanners cannot be operated with WIA). WIA relies on the operator's needs being very basic: scanning a photo to be printed, sending via e-mail, etc. For archival digitization, the WIA interface does not provide enough control over the scanning process to be useful.

Regardless of the scan driver that is used, the key to using a flatbed scanner successfully is by knowing what settings are used to get results. Becoming familiar with the settings means any errors in scanning can be corrected.

Understanding resolution, bit depth, color space, and file formats

Some of the most commonly asked questions that newcomers ask is "what resolution should I scan at?" "What file format should I use?" The Digital Library Federation list their recommendations – 300 dpi 24-bit color TIFF for images, 600dpi 1-bit bitonal TIFF for pages of text (www.diglib.org/standards/bmarkfin.htm#benchmark). The best way to understand these recommendations is to break the specifications down into their components.

The first part of the recommendation – 300dpi, 600dpi – addresses the resolution at which the image should be scanned. Resolution is a grid pattern that the original image is segmented into. The number equals the number of pixels (picture elements) captured within an inch, abbreviated as dpi (dots per inch) or ppi (pixels per inch). The higher the resolution, the finer the grid used to segment the image. Resolution has a proportional effect on file size. The higher the resolution used, the more numbers are used to describe it, increasing the size of the file. When purchasing flatbed scanners, consumers may note two different numbers – optical and interpolated resolution. Why are there two? The difference lies in how they are generated. Optical resolution is the maximum number of pixels a scanner is physically capable of capturing. Interpolated resolution is artificially generated; software takes pixels captured by the scanner, expands the grid pattern, and guestimates pixels that lie in between the pixels that were captured by the scanner.

The second part – 1-bit, 24-bit – covers bit depth. "Bit" is an abbreviation for binary digit (www.webopedia.com/TERM/B/bit.html). Bits have two values, 0 and 1. Bits are used to describe the range of shades between pure black and pure white. For example, black and white files are called 1-bit because there are only two shades, black and white. With grayscale images, the number of bits has been increased to describe additional shades from black to white, 8-bit grayscale means there are 256 gradations (2 the power of 8, 2 × 2 × 2 × 2 × 2 × 2 × 2 × 2). 16-bit grayscale is 216 or 65, 536 gradations from light to dark.

Color images are created through the combination of three colors: red, green, and blue[3]. When objects are scanned, the reflected light is separated into red, green, and blue channels. Each color is then described in bits. The specification above lists color as 24-bit. When you described each color – red, green, and blue – in 8 bits, the total generates 16.7 million colors. If the specification listed 48-bit color, 16 bits are being used for each color, creating a total 281 trillion colors[4]. As with resolution, the higher the bit depth is, the larger the digital file becomes.

The final part of the specification – TIFF – is the file format. TIFF stands for Tagged Image File Format and is accepted as the standard for archival files (i.e. most programs will open TIFFs). When scanning images, the recommended practice is to save the file directly after scanning as a TIFF before any image editing or compression is performed. Other common file types include GIF, JPEG, and JPEG2000, all of which are display file formats, meaning the emphasis is quick image display rather than data fidelity. GIF stands for Graphic Image File, a low resolution file format used primarily for web graphics and icons. GIFs are suited for image display because the color palette of GIFs is limited to 256 colors, whether that's shades of gray or individual colors (you can have GIFs with 2, 4, 16, 32, 64, or 128 shades/colors as well, depending on the bit-depth used).

JPEG is another well known display file format. JPEG is somewhat misnamed, because it is not the name of the file format, but the name of the compression used. JPEG stands for Joint Photographic Experts Group, an international coalition of photographic experts who created an image compression routine to make photographs small enough for displaying on the Internet, while retaining image quality. The actual name for the file format is JFIF, JPEG File Interchange Format. One of the drawbacks of JPEG is that the compression is lossy[5]; to address this the JPEG consortium developed JPEG 2000, which incorporates improved lossy and lossless compression as well as pan-and-zoom capabilities.

Recommendations and final tips

The Digital Library Federation recommendations mentioned above are minimum standards, set to ensure that the digital images captured are of sufficient quality to be useful in the future. In all digitization work, the aim is to subject all original material to the rigors of digitization as few times as possible; once is the ideal. Thereafter, additional work should concentrate on reuse of the digital files to generate derivative files (print and display quality files) that meet the current needs of the institution.

As for derivative file settings, the following settings have proven to be useful in my experience:

- Normal web image – 72dpi GIF/JPEG.
- Minimum gray/color print setting – 150dpi JPEG.
- Optimal color print setting – 300dpi TIFF.

- Optimal setting for running pages of text through OCR – 300dpi TIFF.
- Best black and white print setting – 600dpi TIFF.
- Archival setting (all colors) – 600dpi TIFF.

Reference

The American Heritage (2000), *Dictionary of the English Language*, 4th ed., Houghton Mifflin Company, Boton, MA, available at: http://dictionary.reference.com/search?q=digital

Notes

1. Digital photocopiers are in essence scanners which digitize the image and immediately transfer it to paper, rather than storing it as a file on a hard drive.
2. http://home.howstuffworks.com/scanner.htm The site has a flash animation of the interior of a flatbed scanner as it scans a page of text.
3. The primary colors learned in school – yellow, red, and blue – are the color palette created with pigments. Red, green, and blue are the color palette created by separating light. Blend RGB together and you get white light. This is why televisions and computer monitors use red, green, and blue phosphors for color images.
4. The human eye can distinguish approximately 410 million colors – experts are divided about the exact number and the range varies from person to person. However, human color perception differs from computer rendering, as the human eye does not see equal gradations of color and shade in the same way as a scanner. More colors are needed from scanning to approximate the colors seen by the human eye. For further reading about color palettes and the science of color, the following web sites provide an excellent overview: RGB World – Understanding Color (www.rgbworld.com/color.html); The Science of Color (www.webopedia.com/DidYouKnow/Computer_Science/2002/Color.asp)
5. Data compression is either lossy or lossless. With lossless compression, the routine saves space by reducing redundant data. When the lossless compressed file is opened, it matches the file before compression; no data are lost. With lossy compression, the routine saves space by throwing away data. When the lossy compressed file is opened, it does not match the file before compression; data have been lost.

Larry Wentzel (*lrw5@psulias.psu.edu*) is the Digital Preservation Coordinator at Pennsylvania State University Libraries, University Park, PA.