A Proposed Standard Procedure to Define Minimum Scanning Attribute Levels for Hard Copy Documents

Mitchell Cochran
City of Monrovia
mcochran@ci.monrovia.ca.us

Abstract

Public agencies have been scanning hard copy documents into electronic systems for a number of years. Legally, the document needs to be a true copy of the original but no standards setting body has defined what specific attributes define when a document is a true archival copy. Documents can be scanned at high resolution levels but there is a trade off between resolution and file size. The goal is to meet the legal requirement but at the smallest file size that is reasonable. The paper will show that commonly used industry attributes of 200 dots per inch, (DPI) for text documents, 300 DPI and 8 bit gray scale colors for black and white photos and 300 DPI and 24 bit color for full color and or line work documents met the legal archival quality standard. With this recommendation, agencies can adhere to a common set of practical scanning attributes.

1. Introduction

A critical issue of this paper is defining when an electronic copy of a document is the equivalent of the original document. The Merriam-Webster dictionary defines a document as [1] "an original or official paper relied on as the basis, proof, or support of something" and "a writing conveying information". The Oxford dictionary defines it as [2] "a piece of written, printed, or electronic matter that provides information or evidence or that serves as an official record." In each case, the content or information that a document provides is important and not necessarily the physical medium.

The comparison of a document from an electronic media to that of a hard copy original should be made in relation to the comprehension of the content or meaning of a document. The criteria would be to decide if the content or meaning of a document has been altered by the reproduction process. A poorly scanned document would not provide the same information or meaning that a 'good' copy would provide. This paper will provide criteria that suggest quality thresholds based on the type of document.

Many organizations look to the Association for Image and Information Management (AIIM) or industry associations to provide information system standards. Specific attribute standards for document retention and scanning have not been universally accepted. Case law typically cannot help because it deals with specific factual situations and offers general guidelines in response. Currently, there are no widely cited cases to reference when attempting to establish organizational guidelines. There has not been any case law to help set expectations. Organizations have been scanning documents for years but it is not clear if those documents meet legal expectations. Many organizations have set their own policies but without clear legal guidance and or commonly accepted practices, procedures cannot be applied with confidence. As more documents are kept only in electronic form, there needs to be a commonly accepted format and resolution, This is especially true if you consider e-discovery issues and the additional requirements they entail.

There are a number of general legal guidelines for document scanning and retention. The problem is that these standards are relatively general and can be vague from a technical perspective. There are no specific numbers or attributes on each guideline. Similar to appreciating art, comedy, pornography, or other qualitative items, you may not be able to define it but you can evaluate it when you see it. The goal is to define a standard procedure so that industry proven attributes will meet the legal guidelines.

2. Legal Guidelines

California law does provide some guidance ont eh issue of true digital copies. California code includes a description of evidence in the California Legislative Information, Evidence Code, Division 11 in sections

978-1-4799-2504-9/14 $31.00 © 2014 IEEE
DOI 10.1109/HICSS.2014.258

1550 and 1553. Section 1550 provides for the admission of a unaltered reproduction into a court of law [3]. Section 1553 provides that the responsibility of the party providing evidence to ensure it is an accurate representation by a preponderance of the evidence. [4]

Similarly, the United States Federal Courts include rules as to the authentication of evidence. As part of the Rule 901 General Provision of the Federal Rules of Evidence [5] there is a requirement for authentication of evidence. Relevant subsections include:

> 4 - Distinctive characteristics and the like.—Appearance, contents, substance, internal patterns, or other distinctive characteristics, taken in conjunction with circumstances.
>
> 7 - Public records or reports.—Evidence that a writing authorized by law to be recorded or filed and in fact recorded or filed in a public office, or a purported public record, report, statement, or data compilation, in any form, is from the public office where items of this nature are kept.
>
> 8 - Ancient documents or data compilation.—Evidence that a document or data compilation, in any form, (A) is in such condition as to create no suspicion concerning its authenticity, (B) was in a place where it, if authentic, would likely be, and (C) has been in existence 20 years or more at the time it is offered.
>
> 9 - Process or system.—Evidence describing a process or system used to produce a result and showing that the process or system produces an accurate result.

From a governmental perspective, the guidelines support documents that are assumed to be from the government using a normal process.

Rule 1001 sets definitions for original and reproduction documents: [5]

> Original.—An ''original'' of a writing or recording is the writing or recording itself or any counterpart intended to have the same effect by a person executing or issuing it. An ''original'' of a photograph includes the negative or any print there from. If data are stored in a computer or similar device, any printout or other output readable by sight, shown to reflect the data accurately, is an ''original''.
>
> Duplicate.—A ''duplicate'' is a counterpart produced by the same impression as the original, or from the same matrix, or by means of photography, including enlargements and miniatures, or by mechanical or electronic re-recording, or by chemical reproduction, or by other equivalent techniques which accurately reproduces the original.

Rule 1003, Admissibility of Duplicates, shows that scanned documents can be admissible in court:

> A duplicate is admissible to the same extent as an original unless (1) a genuine question is raised as to the authenticity of the original or (2) in the circumstances it would be unfair to admit the duplicate in lieu of the original.

Considering these rules, it should be possible for organizations to produce legally acceptable scanned documents using normal procedures.. The proposed standard follows industry practices in developing scanning procedures. .

There are two sources of legal requirements that pertain to the quality of a document. The first is the issue of admissibility and trusted systems s requirement. For admissibility requirements, the focus is on reproduction of content in a visible form.

The second part is the procedural requirements for e-Discovery. e-Discovery procedures require that documents be able to be searched, preferably by an automated method. It is important that documents that may need to be searched be stored in a format that is easily searchable using optical character recognition (OCR).

Using concepts from both the State of California code and the Federal code we can provide some simple scanning guidelines which include:

- Scanned images shall be a true copy of the archival quality
- Documents shall be reproducible in their original form matching both size and color.
- Imaged documents need to be verified as a true copy for image and index information prior to entry into a document management system.
- No paper record can be destroyed if its image contained within a document system cannot be reproduced with full legibility.

3. Spatial and Tonal Resolution

Kit Peterson lists two types of attributes [6]

- Spatial resolution – capturing detail (DPI)
- Tonal resolution – color, bit-depth, and dynamic range

The primary scanning attribute is the resolution which is measured in dots per inch or DPI. The higher the resolution, the better the quality and recognition of document content. The question is at what resolution level is sufficient to satisfy the users.

Typical document resolutions are 200DPI, 300DPI, 400 DPI, and 600DPI. The original laser printers, HP4 for example were printing at 300DPI in the 1980's. Compared to the quality of dot matrix printers, laser printed documents at a resolution of 300 DPI were considered originals. The 300DPI documents compared well to traditional to analog copiers. Today, digital copiers typically scan and printout at 600DPI. Figure one demonstrates how the resolution affects the quality of the image. The figure is intended as a general representation of the resolution differences for a printed image. The reader would need to examine images that are printed as originals in each resolution.

Figure One Spatial Resolution Examples [7]

AIIM, has created an industry accepted standard in the ARP1-2009 document, Analysis, Selection, and Implementation of Electronic Document, Management Systems. Section 5.4.2.1, Document Scanning, of that document lists general scanning requirements [8], 'the capability of the scanner to scan at the resolution meeting the specific image quality requirements of the system, such as 200, 300, or 400 DPI'.

One of the key discussions of resolution is the resulting file size of the scanned image. As the resolution increases, the file size increases. When a document is measured at 200DPI, it is 200DPI in both the horizontal and vertical directions. So a square inch at a 200DPI will be 200 by 200 dots or 40,000 pixels. Table One provides for a comparison of various scan resolutions.

Table One Spatial Resolution Examples

200 DPI	40,000 pixels	
300 DPI	90,000 pixels	2.25x
400 DPI	160,000 pixels	4x
600 DPI	360,000 pixels	8x

Most discussions assume that a document is a standard 8.5 by 11 inch document. But a page size could be as large as a typical blueprint, an E-size drawing at roughly 3 feet by 4 feet. An E-size document is equal to roughly 12 traditional pages. Cornel lists 5 different types of documents [9]

- Printed text or simple line art
- Manuscripts
- Halftones
- Continuous tones
- Mixed

Documents such as printed text or simple line art could use a lower resolution since they are generally not complex. As the document gets more complex or adds color, they need a scan with a higher resolution to capture all of the details.

As pictures or colors are included in the document, it requires the person scanning to consider the color of the document. Cornel lists additional reasons to scan a document in color [9]:

- Pages are badly stained
- Paper has darkened to the extent that it is difficult to threshold the information to pure black and white pixels.
- Pages contain complex graphics or important contextual information (e.g., embossments, annotations)
- Pages contain color information (e.g., different colored inks)

A resulting scan color could be just black and white, shades of grade for black and white photographs, or in colors. The bit-depth is that amount of information that is captured at each point.

- 1 bit bi-tonal – black or white

- 8 bit gray scale – 256 shades of gray
- 24 bit RGB color – 16.8 million colors

AIIM, provides a file size comparison for a standard size document [10]

- 10 page b/w document .8mb
- 10 page 24 bit color – range from 10 to 60mb.

A document's dynamic range is the difference from the lighter areas to the darker areas of a graphic or photo.

Most scanning systems compress the image as part of the scanning process. The implementer and operator need to decide if the system will be 'loss less' which will not lose any graphical information or 'lossy' which will lose some graphical details as the image is compressed. Typically, two different file compression methods are used:

- Group 4
- JPG or JPEG

The State of California Secretary of State has accepted the AIIM document Analysis, Selection, and Implementation of Electronic Document, Management Systems as a standard [8]. This document lists acceptable formats as JPEG, JBIG, JPEG 2000, PDF-A. TIFF documents are acceptable only if the header information is maintained in the document management system and no proprietary information is added to create a variant TIFF.

The custodian needs to compare the compression method along with the options for that method. For example, a PDF stored document has a sliding scale for compression from high quality and large file size to low quality and small file size. The higher compression, the smaller the file size will be. However, it is a lossy method so some information can be lost, such as fine details. Table Two shows the file size of a standard 8 1/2 x 11 text document (compressed and uncompressed) at various DPI and use of gray scale uncompressed:

Table Two: Spatial Resolution Examples

DPI	200	300	300	300	400
Color			4 bit gray	8 bit gray	
Not Compressed	500 K	1.05 MB	4.08 MB	8.3 MB	2 MB
Group IV	50K	105K	400K	800K	220K

4. Proposed Standard Procedures

This section proposes a set of scanning resolutions standards that should generally support the creation of a true copy while balancing costs and operational efficiencies. The recommended scanning resolutions include:

- Black and white text only documents – 200 DPI in only black or white
- Black and white documents with black and white photos or complex diagrams – 300 DPI with 8 bit gray scale
- Documents with color photos or line work. – 300 DPI with 24 bit color

The goal is to capture as much detail as possible at the lowest resulting file size. Documents such as blue prints, building plans,& diagrams, and other complex or graphically detailed documents should be scanned at 300DPI. The color palette would be selected based on whether color provides additional content. For example, a blueprint would be scanned at an 8 bit gray scale since there are only two colors. Documents such as plans that use colors for layer differentiation would be scanned with a palette of 24 bit color. .

5. Common Practices

A number of standards from state archive organizations support the recommendations from the paper [11], [12], [13]:

- Georgia
- Kentucky
- Indiana
- Michigan
- Mississippi
- New York
- Oregon
- Washington

The archive for the state of Mississippi states that they require a resolution of 300DPI or better. [14] The Courts of the State of Arizona require that documents be scanned at a resolution of 200DPI or better [15]

The standard for Kentucky considers whether there is text that is smaller than 6 points [16]. It requires that instead of scanning at 200 DPI, those

documents would be scanned at 240DPI or 300DPI. This paper's recommendation of 300DPI for complex documents would meet this criterion. The one exception is for text only documents that are scanned at 200 should be scanned at 300.

The National archive notes [17] that engineering drawings, maps or other documents with fine drawings might have to be scanned at 600 DPI or greater. The organization states that the appropriate resolution should be determined on a case by case basis. This procedure follows the proposed standard as laid out in the paper. Use the lower standard scanning resolution until it is needed in specialized cases. The important point to recognize is which documents are specialized and need a higher resolution versus the trade off of file size when scanning everything at a higher resolution.

The National Archives state in 2003 that records must be 8 or 16 bit continuous tone gray scale or color raster images in section 5.3.1 [17] It follows in the next section, 5.3.2, that color images must be produced in RGB as either 24 or 48 bit files.

A higher resolution will give better results when an image is being processed for optical character recognition (OCR). The system will try to interpret what characters that the dot patterns represent. The higher a resolution is, the more accurate the interpretation will be. If the system sees two ovals, and the scan is at a low resolution, it may not be able to accurately differentiate between a pattern that is a 'B' or one that is an '8'.

Both industry and academic sources have found that 300DPI represents a good trade off for balancing resolution versus file size. Fujitsu, a leading scanner manufacturer, recommends that 300DPI be used as the default standard [18]. The University of Illinois at Urbana-Champlain Library also recommends images be scanned at 300DPI for an OCR process [19]. It does point out the OCR accuracies are typically between 97% and 99%. However, that is for character accuracy not accuracy of the words. Only 75% of the words may be accurate. One solution is to then run the resultant file through a word processor or spell check function to identify the miss-spelled words. The operator can then ensure that the spellings are correct or that the word is the correct word. The University Library also points out those documents with smaller than a 6 point font would need a higher resolution, 600 DPI, to allow for an OCR process.

A United States Printing office study showed that scanning older documents in RGB mode meets their 99% accuracy requirement [20] The legal reference site, www.lexbe.com, confirms that scanning in resolutions higher than 300 DPI does not appreciably increase the accuracy. [21] The site also points out that scanned documents may need to be used in litigation, so they need to meet minimum standards: 'To be non-specific is to invite an adversary to return documents scanned at 150DPI without OCR, that may be unsearchable, illegible and unintelligible!'

An older study by Rice et all, [22] shows error rates for documents scanned at 200DPI, 300DPI and 400DPI. The error rates for 300DPI were 30 to 50% better than at a 200DPI resolution. Documents scanned at a 400DPI resolution generally had between a 2 to 10% accuracy improvement rate over documents with a 300DPI resolution. The user needs to decide if the higher OCR accuracy of a document with a 400DPI resolution is worth storing documents that are close to twice the size of a document scanned at 300DPI. In making such a decision, the user should also consider how well indexed the document is.

An AIIM general document [10] provides a recommendation for scanning resolutions:

- Black/White – lowest at 200, conventional 300
- Color – lowest 150, conventional, 200, highest, 300. archival 600

6. Procedures

What has been accepted for electronic documents is not only just the specific attributes but also the process. The legal process includes a quality control method. Each image needs to be checked as it is scanned.

For the State of Kentucky, KRS 171.660 authorizes optical imaging if the criteria of the Kentucky Department for Library and Archives and 725 KAR 1:020 are met [16]. The document points out that it may be challenged in court but the organization would need to show that appropriate controls and procedures are in place.

Quality control refers to the methodology and techniques used to ensure consistency of procedures and output. Rigorous quality control procedures shall be used to ensure that the recorded images are of

acceptable quality and can be accurately retrieved with the indexing method employed. [16]

The operator needs to fix normal scanning errors such as a crooked or deformed image. If the image does not match quality expectations then it needs to be rescanned. If needed, the rescan process can use a higher resolution than 300 DPI or 24 bit color. Should the image be an archival quality image, it might need a higher resolution to maintain the high quality of the original. The 300 DPI guideline is intended for use with what can be considered general business documents. These documents consist of text, photos and line work.

It is important to realize that the quality of an image will degrade the more times it is scanned or compressed in a lossy compression. A poor quality image will never improve after scanning. The New York State Archives specify that documents need to be at least 200 DPI with a loss-less compression [23]. The New York archive lists criteria that a quality control operator should look for and what an operator can do or not do:

Inspection:

a. Correct image filename (unique identifier)
b. Correct file format for each image type (master and access)
c. Image scanned at appropriate unenhanced DPI for each image type
d. Image oriented properly, whether landscape or portrait
e. Image is correct size (in pixels along both dimensions)
f. Image is not skewed
g. Image is not rotated or flipped
h. Image is neither too light nor too dark
i. Appropriate contrast exists within the image
j. No distortion of the image
k. No extraneous materials (fingers or fasteners) obscure the image
l. No noise or other problems in image file
m. Appropriate indexing terms are associated with the scanned image
n. Monitor where images are verified is calibrated and is operated under controlled viewing conditions
o. Image viewer used to view and evaluate the images is indicated
p. DPI verified by an independent software program

Allowable corrections:

a. Correcting image filename
b. De-skewing, rotating, or flipping the image to correct its orientation
c. Adjusting brightness, contrast, or tone through rescanning
d. Cropping that does not remove any information in the document
e. Rescanning, followed by a re-inspection of the new image
f. Updating index database to correct errors

Unacceptable alterations:

a. Sharpening the image
b. Retouching or de-speckling
c. Dithering or quantization
d. Removing information from the images
e. Adding information to the images
f. Burning annotations or "sticky notes" onto the image file itself

For the City of Monrovia, California, the building officials have found that 300DPI is acceptable for reading blueprints [24]. The City has been scanning blueprints and associated drawings for over ten years.

7. Conclusion

Organizations generally follow industry accepted practices that have been codified into governing body guidelines. At this point, no industry wide guidelines have been developed or accepted for the definition of a document. The legal definition is vague and left to the users' interpretation.. Even an industry accepted AIIM document only lists 200DPI, 300DPI and 400DPI as standards without any specific direction. The document states that the resolution should meet the needs of the image. There needs to be a tradeoff between the quality of the document and the amount of storage that is necessary. This paper proposes that organizations should have the following scanning attributes as a minimum:

- Black and white documents – 200 DPI in only black or white
- Black and white documents with black and white photos – 300 DPI with 8 bit gray scale
- Documents with color photos or line work. – 300 DPI with 24 bit color

Both state archive organizations and the National Archives had defined scanning standards that are met by this criterion.

The paper also introduces four general guidelines for implementing a document imaging process:

- Scanned images shall be a true copy of the archival quality
- Documents shall be reproducible in their original form matching both size and color.
- Imaged documents need to be verified as a true copy for image and index information prior to entry into a document management system.
- No paper record can be destroyed if its image contained within a document system cannot be reproduced with full legibility

8. Limitations and Future Research

File compression will have a large effect on storage requirements but since specific systems may require specific file types, the selection of a compression technology is not proposed in this paper.

The paper is focused on hard copy documents that need to be entered back into an electronic system. Some documents exist as entries in database without necessarily a hard copy form. An entry in a database could be easily imported into a records retention system with almost perfect clarity since the document is imported as text instead of an image of the text.

The paper develops thresholds for hard copy documents. Threshold criteria should also be developed for other document mediums such as audio or video recordings.

The California Secretary of State has promulgated standards and guidelines using AIIM standards as the basis for recording, storing, and reproducing permanent and nonpermanent documents or records in electronic media and those guidelines are applicable to certain public entities in California.

9. References

[1] Merriam-Webster, http://www.merriam-webster.com/dictionary/document, acquired June 2013.

[2] Oxford Dictionary, http://oxforddictionaries.com/definition/english/document, acquired June 2013.

[3] State of California Legislative, Legislative Information:, evidence code, division 11, Writings, Secondary Evidence of Writings, Section 1550, [http://leginfo.legislature.ca.gov/faces/codes_displayexpandedbranch.xhtml], accessed June 2013

[4] State of California Legislative, Legislative Information:, evidence code, division 11, Writings, Secondary Evidence of Writings, Section 1553, [http://leginfo.legislature.ca.gov/faces/codes_displayexpandedbranch.xhtml], accessed June 2013

[5] Committee on the Judiciary, House of Representatives, Federal Rules of Evidence, 111th Congress, 2nd Session, December 2010.

[6] Peterson, Kit A. Introduction to Basic Measures of a Digital Image for Pictorial Collections, Prints & Photographs Division, Library of Congress, Washington, D.C. 20540-4720

[7] Illustration by Phil Michel from photograph by Jack Delano, "Cars of Migratory Tomato Wrappers. Kings Creek, Maryland," July 1940. Farm Security Administration Collection, LC-DIG-ppmsca-08927.

[8] Association for Information and Image Management (AIIM), Analysis, Selection, and Implementation of Electronic Document, Management Systems, ARP1-2009, June 2009.

[9] Cornel University Library, Moving Theory into Practice, Digital Imaging Tutorial, 2003

[10] Association for Information and Image Management, Scanning Business Documents To PDF Best Practices, www.aim.org, May 6, 2012, accessed June 2013.

[11] Georgia Secretary of State, Georgia Archives, Electronic Document Imaging Systems Guidelines – Part Four, Technical Guidelines, http://www.sos.ga.gov/archives/InformationForGovernmentAgencies/Records_advice/TechnicalLeaflets/Imaging/electronic_document_imaging_systems_guidelines4.htm , acquired June 2013.

[12] Michigan Department of History, Arts, and Libraries, Technical Standards for Capturing Digital Images from Paper or Microfilm, http://www.michigan.gov/documents/hal_mhc_rms_st_for_digitizing_125531_7.pdf, July, 2005.

[13] Washington State Legislature, WAC 434-663-305, Scanning density, http://apps.leg.wa.gov/wac/default.aspx?cite=434-663-305, acquired June 2013.

[14] Mississippi Department of Archives and History, Public Records – Standards, Destruction of Original Records after Imaging., acquired June 2013.

[15] Arizona Code of Judicial Administration, Part 1: Judicial Branch Administration, Chapter 5: Automation, Section 1-504: Electronic Reproduction and Imaging of Court Records, January 2001.

[16] Kentucky Department for Libaries and Archives, Public Records Division, Ensuring Long-term Accessibility of Textual Records Stores as Digital Images: Guidelines for State and Local Government Officials, Revised January, 2010

[17] National Archives and Records Administration, Tips for Scheduling Potentially Permanent Digital Photographic Records, acquired June 2013.

[18] Neal, Kevin, Why is OCR at 300 DPI, Fujitsu, http://scansnaptest.fcpa.fujitsu.com/tips-tricks/1652-why-is-ocr-at-300-DPI-a-standard/, January 13, 2010.

[19] University of Illinois at Urbana-Champlain Library, http://www.library.illinois.edu/dcc/bestpractices/chapter_05_ocr.html, acquired May, 2013.

[20] Booth, Jon, M, Gelb, Jeremy, Optimizing OCR Accuracy on Older Documents: A Study of Scan Mode, File Enhancement, and Software Products, Office of Innovation and New Technology, U.S. Government Printing Office, Washington DC, June 2006.

[21] Lexbe, http://www.lexbe.com/hp/lititips-scanning-to-searchable-PDF-for-lawyers-and-litigation.aspx, acquired May 2013.

[22] Rice, Stephen V., Kanai, Junichi, Nartker, Thomas A., The Third Annual Test of OCR Accuracy, http://www.stephenvrice.com/images/AT-1994.pdf, 1994.

[23] New York State Archives, Imaging Production Guidelines http://www.archives.nysed.gov/a/records/mr_erecords_imgguides_accessible.html, 2006.

[24] Cervantes, Encarnacion, City of Monrovia, California, Building Official, personal interview, June 2013.