Why 300ppi? Requirements-based methodology for determining digitization project specs

Matt Pearson; Stanford University Libraries; Stanford, CA

Abstract

This article shares some helpful advice and vocabulary for understanding how to determine digitization "resolution" based on need. It demystifies a key guideline established in many best practices and technical guidelines and explains why we sometimes do need 300ppi.

Introduction

Digital preservation projects typically require that materials be imaged at a "Best Practice" standard of either 300 or 600 pixels per inch (ppi) [1][2] for actual-size representation, often without a clear understanding of why or how these rigid guidelines are important. In many cases, imaging large quantities or oversize materials at these specifications is excessively demanding of staff and resources. There are opportunities for using a lower, more manageable ppi and there are situations when a very high number of ppi is necessary. Because of the difficulty in managing and delivering large files and to safeguard against overexposure to handling, for objects that are not adequately digitized, it is important that the choice of ppi be a rational and informed one.

The following article includes some helpful advice for understanding how to calculate sample frequency / ppi based on *need*. At a glance, the formulas may seem complicated, but rest assured, the math is actually quite simple. The sections that follow begin with a brief review of terminology, followed by background information, an overview of an approach for determining ppi needs, a case study in which 300 ppi is preferred, and some additional considerations.

Terminology

Pixels (picture elements) are the building blocks or fundamental units of digital image files. They are arranged in a grid and vary in brightness and/or color. Pixels can sometimes refer to sensors within digital cameras and camera backs. The points of color within a monitor's screen are also referred to as pixels.

Pixel size varies and is a characteristic of imaging devices as well as electronic image files. A particular 33 megapixel digital camera back might have sensor pixel sizes of 7.2 x 7.2 microns while a graphics LCD might have pixel sizes of 127 x 127 microns. Image file pixel sizes can vary: they are determined by sample frequency, which is established within imaging device driver software or image editing programs. The relationship of device pixels to image file pixels is somewhat similar to that of film grain to photographic enlargements and projected film.

Sample frequency is usually expressed as pixels per inch (ppi), the number of pixels assigned to one inch of the width or the height of an image file. The term "resolution", the ability to distinguish detail, is commonly misused to refer to sample frequency. For readability and clarity, this paper employs "sample frequency".

Dpi or dots per inch is sometimes used in place of ppi. Dpi is generally used to describe printed images, which are comprised of tiny 'dots' of ink.

Pixel Dimensions refers to the number of pixels in the length and width of an image file. Many Best Practices rely on long pixel dimension standards instead of sample frequency.

Top-down strategy is advocated for determining guidelines based on content analysis rather than bottom-up, which provides specifications based on end products. When determining sample frequency guidelines, top-down analysis begins with careful evaluation and measurement of the content to be digitized.

Web presentation images or access images are lower resolution files destined for the internet and for display by monitor or projector. They are often created from preexisting archival or master files. The New Jersey Digital Highway's (NJDH) Digital Imaging Specifications suggests reducing the sample frequency to "screen resolution, usually 72-150dpi" for access files [3].

Archival Masters, sometimes referred to as Preservation Masters serve as the parent files for access and production images. These files are intended for long-term and serve as optimized digital negatives. They are highest-quality in order to support preservation and are almost never altered once they have been deposited into an institution's digital collection. Most best practices and guidelines describe the creation of archival masters.

*An **overview of basic digitization terminology*** is available online in Cornell's tutorial, "Moving Theory into Practice"[4]. Richard Pearce-Moses has also developed a lexicon that can be accessed at The Society of American Archivists website [5].

Motivation and background

Whether building a preservation-oriented collection or digitizing for web access, it is important to establish standards that support project efficiency and safe handling of materials. Many programs, such as North Carolina's Exploring Cultural Heritage Online (ECHO), have adopted a "scan once methodology":

> *It is expensive for institutions to go back and re-digitize their holdings. Few ever do so. In addition, many originals could suffer from the handling and exposure to bright light required by digitization. Therefore, it is best to simply "scan once," create a master image, and make any future duplicates from it.* [6].

Common practice, described in the Bibliographical Center for Research's (BCR) CDP Digital Imaging Best Practices, is to produce high-quality master files that are used to generate multiple versions in smaller sizes or alternative formats for a variety of uses [7]. Occasionally, projects include benchmarking where specifications are verified with sample captures of a selection of materials.

Most of the best practices available online offer umbrella standards for image capture. In his much cited Best Practices for image Capture, Howard Besser suggests (bottom-up), standards of 3000 and 6000 pixel requirements for long dimensions of image files [8]. This guideline is based on file size and provides for capturing information at 300ppi or 600ppi for 8”x10” documents. Unfortunately, a 24” long document would be digitized with a sample frequency of 125ppi or 250ppi, which is probably too low for accurate recording. Conversely, a 7” long document would scan at 429ppi or 858ppi, resulting in an excessively large file and long capture time. Similar long dimension pixel requirements appear in the guidelines of North Carolina’s Exploring Cultural Heritage Online, The New Jersey Digital Highway’s Digital Imaging Specifications, and the BCR’s CDP Digital Imaging Best Practices (for text).

More recently, institutions have been establishing sample frequency ranges based on content. North Carolina’s Exploring Cultural Heritage Online offers a 600[ppi] option for scanning photographs and requires 200-300[ppi] for scans of text [9]; the BCR’s CDP Digital Imaging Best Practices recommends ranges of 400ppi to 800ppi for photographs and 600ppi to 800ppi for graphic material [10]; the National Archives and Records Administration recommends 400 to 600ppi for scanning text [11]; and the Library of Congress requires 300 to 400ppi for text and manuscripts, 400 to 600ppi for rare books, and 300ppi to device maximum for photographs [12].

Volume projects may necessitate adopting an umbrella range for sample frequency. An extreme example, most of the web access images produced by the Internet Archive’s Scribe workstations range from 300 to 600ppi [13]. They also generalize, for contributors, that books should be digitized using the maximum device settings (to insure adequate sample frequency). The Internet Archive’s mission is to build an ‘Internet Library’ and provide access to educational and cultural texts often found in physical libraries. The volume of material being digitized necessitates a reasonable range of ppi and the scale of most information being recorded requires 300 to 600ppi. There are special occasions when the Internet Archive adjusts this parameter to better match the content.

While these sample frequency oriented guidelines usually provide adequate quality, they can also result in overkill (and underkill): image files are much larger than necessary or, for materials with very fine detail, files are too small. Essentially, they are still part of a bottom-up approach. In the interests of efficiency and safe handling, successful projects of any size should incorporate top-down analysis.

Top-down: a very basic approach

Gather the tools used for determining project sample frequency: materials to be imaged, a ruled textile or typesetter’s loupe and a basic calculator.

Examine the details closely. Determine the smallest element to be recorded and measure its smallest dimension. For example: a mark made with a Staedtler Mars technical pencil might have a width of between .0118 and .063 inches (standard technical pencil lead sizes).

Divide 1 by the smallest element's smallest dimension (in inches) to determine the number of details that would fit side-by-side into an inch, the base unit for measuring sample frequency. Round the quotient up in order to convert to pixels per inch. The number of technical pencil marks that would measure up to 1 inch is between 84.74 and 15.87; the *absolute* sample frequency for recording any of them is between 85 and 16 ppi.

$$1 / .0118 \text{ inch} = 84.74 \text{ details / inch} = 85\text{ppi} \quad (1)$$

$$1 / .063 \text{ inch} = 15.87 \text{ details / inch} = 16\text{ppi}$$

PPI, practical sample frequencies, and extended usefulness

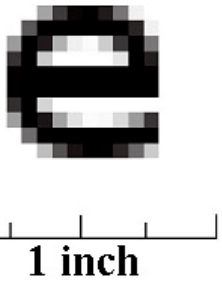


Figure 1. *A 10-pixel-wide letter “e”.*

In the illustration above, the magnified letter “e” measures ten pixels wide. If twenty pixels span one inch of the image file's width or height, its sample frequency is expressed as 20 ppi and the letter “e” is determined to be ½ inch wide. (Most image editors provide tools for measuring.)

“Lines per inch” is an extension of pixels per inch where a line represents a row or column that is one pixel wide. The 20ppi file’s image could portray 20 horizontal or vertical lines per inch.

Without variation in tonality or color, 20 side-by-side lines would display as an amorphous mass. For this reason, alternating black and white lines (line pairs) are traditionally used in evaluating the resolving power of optical systems. Representing 20 line pairs (40 lines in all) requires 40ppi, double the absolute sample frequency. Likewise, our small pencil mark would require a *minimum* sample frequency of 170ppi.

The problem of “scattered pixels” provides an argument for using twice the minimum sample frequency for representing discernible diagonal line pairs. Pixel-wide vertical and horizontal lines are represented more accurately than pixel-wide diagonals. Regardless of capture / scanning device capabilities and limitations, minute details are vulnerable to a degree of aliasing or pixel scattering during capture and processing phases.

Figure 2. *Line pairs imaged at a minimum sample frequency with greater aliasing (left) with line pairs imaged at a practical sample frequency with minimized aliasing (right).*

The top left illustration shows a potential effect of rotating an image file. Although the lines can (arguably) be discerned, there are not enough pixels for well-defined edges.

The top right illustration's line pairs were captured with a *practical* sample frequency of twice as many pixels per inch. Some tools (Adobe Photoshop, ImageMagick, etc.) are very good at resampling for basic image manipulations such as rotation. However, there is no way to guarantee which applications will be used in future work with the image file. Providing the smallest important details with two or more pixels instead of one is a good rule of thumb for ensuring readability. When doubling, multiply the minimum sample frequency by 2.

Brief review

Figure 3. *Absolute sample frequency: one pixel or line represents one image detail; minimum sample frequency: two pixels or lines represent one image detail; practical sample frequency: four pixels or lines represent one image detail.*

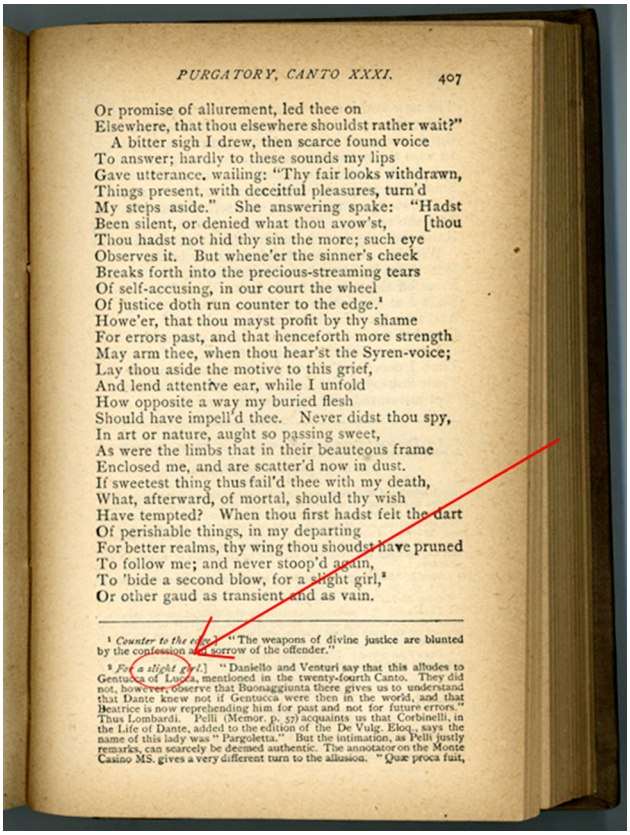
Theory into practice: Cases for 300ppi

Scanning Texts

This first example illustrates an image file of a page from a book scanned using an Epson Perfection 4490 photo scanner. The approach would also apply to scanning using book cradles, overhead book scanners and other flatbed scanner models.

The letter "h" within the footnotes had a smallest dimension typical of the overall volume. Using a scaled loupe, the width was determined to be approximately .01 inches or .25mm. The corresponding sample frequencies were: an absolute of 100ppi; a minimum of 200ppi; and a practical of 400ppi.

While the clear choice of 400ppi was used for digitizing the volume, a scan at 200ppi was made for illustrative purposes.



PURGATORY, CANTO XXXI. 407

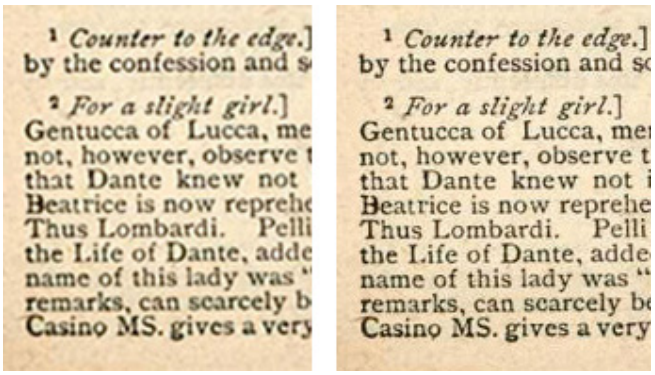
Or promise of allurement, led thee on
Elsewhere, that thou elsewhere shouldst rather wait?"
A bitter sigh I drew, then scarce found voice
To answer; hardly to these sounds my lips
Gave utterance, wailing: "Thy fair looks withdrawn,
Things present, with deceitful pleasures, turn'd
My steps aside." She answering spake: "Hadst
Been silent, or denied what thou avow'st, [thou
Thou hadst not hid thy sin the more; such eye
Observes it. But whene'er the sinner's cheek
Breaks forth into the precious-streaming tears
Of self-accusing, in our court the wheel
Of justice doth run counter to the edge.[1]
Howe'er, that thou mayst profit by thy shame
For errors past, and that henceforth more strength
May arm thee, when thou hear'st the Syren-voice;
Lay thou aside the motive to this grief,
And lend attentive ear, while I unfold
How opposite a way my buried flesh
Should have impell'd thee. Never didst thou spy,
In art or nature, aught so passing sweet,
As were the limbs that in their beauteous frame
Enclosed me, and are scatter'd now in dust.
If sweetest thing thus fail'd thee with my death,
What, afterward, of mortal, should thy wish
Have tempted? When thou first hadst felt the dart
Of perishable things, in my departing
For better realms, thy wing thou shouldst have pruned
To follow me; and never stoop'd again,
To 'bide a second blow, for a slight girl,[2]
Or other gaud as transient and as vain.

[1] *Counter to the edge.*] "The weapons of divine justice are blunted by the confession and sorrow of the offender."

[2] *For a slight girl.*] "Daniello and Venturi say that this alludes to Gentucca of Lucca, mentioned in the twenty-fourth Canto. They did not, however, observe that Buonaggiunta there gives us to understand that Dante knew not if Gentucca were then in the world, and that Beatrice is now reprehending him for past and not for future errors." Thus Lombardi. Pelli (Memor. p. 57) acquaints us that Corbinelli, in the Life of Dante, added to the edition of the De Vulg. Eloq., says the name of this lady was "Pargoletta." But the intimation, as Pelli justly remarks, can scarcely be deemed authentic. The annotator on the Monte Casino MS. gives a very different turn to the allusion. "Quæ proca fuit,

Figure 4. *An image file of a page with smallest detail/text of interest identified in footnotes.*

In the following 200ppi scan, some passages are not well recorded. The text is readable, but some of the letters must be inferred because of aliasing and an insufficient sample frequency. Some of the character of the typeface appears lost or changed. In short, 200ppi works, but the scans are somewhat "hard on the eyes" and not entirely reliable.



[1] *Counter to the edge.*]
by the confession and s

[2] *For a slight girl.*]
Gentucca of Lucca, me
not, however, observe t
that Dante knew not
Beatrice is now repreh
Thus Lombardi. Pelli
the Life of Dante, adde
name of this lady was "
remarks, can scarcely b
Casino MS. gives a very

[1] *Counter to the edge.*]
by the confession and s

[2] *For a slight girl.*]
Gentucca of Lucca, me
not, however, observe t
that Dante knew not
Beatrice is now repreh
Thus Lombardi. Pelli
the Life of Dante, adde
name of this lady was "
remarks, can scarcely b
Casino MS. gives a very

Figure 5. *100% magnification of a 200ppi and 400ppi scan.*

The 400ppi scan offers better representation and is somewhat more readable.

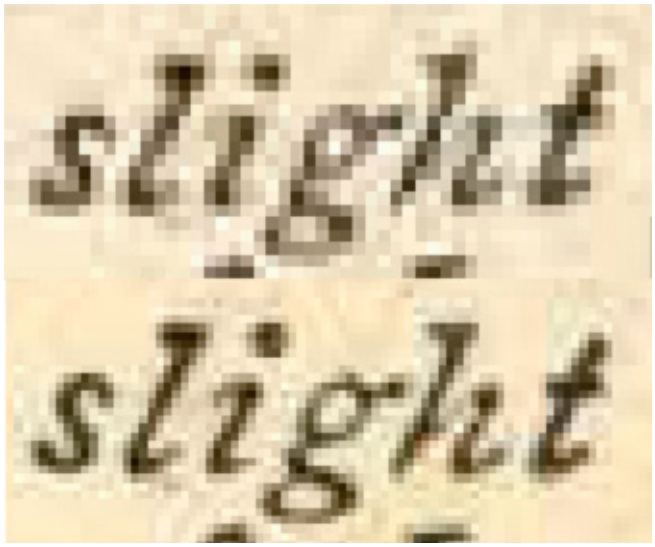


Figure 6. *200ppi and 300ppi equivalents.*

The next illustration compares magnifications of the two scans. OCR software could experience some minor difficulty reading the 200ppi scan, particularly for the letters “h” and “t”. However, good OCR software should be able to recognize the letters or learn to recognize them without much effort. In the 400ppi scan, the letters are clear and most readable. Since this text was scanned for both OCR and to be displayed (and read) on-screen, 400ppi was preferred. In a high volume production setting, 200ppi would probably have offered as much functionality and close to the same human readability.

Digitizing graphic content

For this next example, an image file with “machine readable” markings was the primary goal for digitizing the blueprint illustrated below.

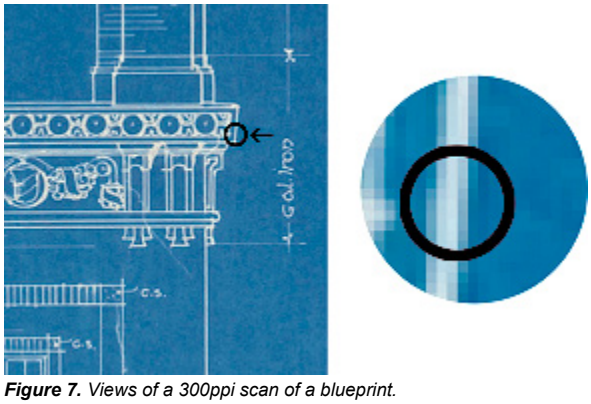


Figure 7. *Views of a 300ppi scan of a blueprint.*

Markings indicated by the red circles suggest the scale of typical important details. The lines’ smallest dimensions measured .014 inches, requiring an absolute sample frequency of close to 75ppi. At a minimum sample frequency of 150ppi, line edges were better defined, improving legibility and the overall appearance of the image. The sample frequency was further doubled to a practical 300ppi, safeguarding against aliasing and extending the usefulness of the image file.

Doubling the minimum sample frequency introduced a second level of information, which included minor paper damage and a suggestion of paper texture, useful for research and preservation purposes. The final representation of the .014 inch line was 3 to 4 pixels wide. At 150ppi, the same line would have been recorded with a width of only 1 or 2 pixels. The 300ppi image file can be resized or downsampled as needed; greater detail, however, can only be achieved by re-digitizing the blueprint, subjecting it to more handling and harmful light exposure.

Tables

Table 1: Sample frequencies for detail sizes in millimeters

Absolute sample frequency / ppi	Minmum sample frequency / ppi	Practical sample frequency / ppi	Appx. dimension of smallest recorded detail
18 ppi	36 ppi	72 ppi	1.42mm
36 ppi	72 ppi	136ppi	.71mm
50 ppi	100 ppi	200 ppi	.51mm
75 ppi	150 ppi	300 ppi	.34mm
100 ppi	200 ppi	400 ppi	.26mm
150 ppi	300 ppi	600 ppi	.17mm
200 ppi	400 ppi	800 ppi	.13mm
300 ppi	600 ppi	1200 ppi	.09 mm
400 ppi	800 ppi	1600 ppi	.06 mm

Table 2: Sample frequencies for detail sizes in inches

Absolute sample frequency / ppi	Minmum sample frequency / ppi	Practical sample frequency / ppi	Appx. dimension of smallest recorded detail
18 ppi	36 ppi	72 ppi	1/18 inch
36 ppi	72 ppi	136ppi	1/36 inch
50 ppi	100 ppi	200 ppi	1/50 inch
75 ppi	150 ppi	300 ppi	1/75 inch
100 ppi	200 ppi	400 ppi	1/100 inch
150 ppi	300 ppi	600 ppi	1/150 inch
200 ppi	400 ppi	800 ppi	1/200 inch
300 ppi	600 ppi	1200 ppi	1/300 inch
400 ppi	800 ppi	1600 ppi	1/400 inch

Putting the details into greater perspective: when too much is not enough

Consider scans / captures made at 300ppi. The smallest recorded detail could (theoretically) measure .007 inches (about half the width of a fine technical pencil mark). Sometimes it is important to record information of that scale: for example, in cases with object deterioration (staining, cracking, etc.) where the damage will be visually analyzed. However, a 300ppi scan / capture does not come close to recording detail on the scale of paper fiber.

Typical paper fiber widths range from 15 micrometers to 30 micrometers. A conservator wishing to record details of this scale, for a digitized A4-size object, would ask for an RGB file size of 200Mb to 1.5Gb, based on the following:

- detail measurements of 15 micrometers to 30 micrometers
- minimum sample frequency of 1,695 ppi to 3390ppi
- image file dimensions of 14,069 pixels by 19,832 pixels; a 16-bit file size of greater than 1.5 gigabytes

For recording paper fiber-scale detail, a very strong case can be made for capturing a portion of the object at very high magnification or for waiting a few years for imaging technology to catch up with demand.

Additional considerations

File characteristics - This paper does not cover many of the other important characteristics of image files. Considerations such as file format, bit depth, and file compression are well detailed by the New Jersey Digital Highway [14] and the Collaborative Digitization Program [15]. The University of Maryland Libraries also offers a clear overview of popular file naming conventions [16].

Optical Character Recognition (OCR) - For more than a decade, it has been accepted that a minimum of 300 ppi is necessary for image files destined for OCR, the process of converting graphic content into editable and searchable text [17]. In recent years, researchers have tested the improvements in imaging techniques and in software design, concluding that successful OCR can be accomplished with much lower standards [18].

Performance and resolution - Lens or optical issues and device performance factor into the resolution of digitizing systems. Since the aim of this article is to provide a basic understanding of how and why sample frequency is considered in digital reformatting, lengthy discussions of the impact of lens resolution and approaches for compensating have regretfully been sidestepped. It is important to note that the ability to record detail is also dependent on the quality of the optical systems employed.

Recording texture - Lighting is an important contributor in recording detail for textured and three-dimensional objects. Using a very high sample frequency will not always provide satisfactory detail. Nuanced lighting is often necessary for revealing textures and details that might be lost with the uniform illumination of flatbed scanners. The complexities of successful lighting for photography are within the domain of a skilled and experienced photographer.

Preserving edges - Image borders are often an unavoidable necessity for archival imaging. Since perfectly square objects are a rarity, uniform image borders are often tolerated for the sake of preserving object shape and edge detail. When incorporating borders, sample frequency may need to be adjusted for digitization projects where pixel dimensions are a determining project guideline.

Conclusion

Lately, the focus of digital preservation has shifted to asset management and sustainability. For content creation, institutions have grown to rely on guidelines from institutions such as the National Archives and Records Administration and the Library of Congress. The Definitions of Digital Preservation, provided by the Association for Library Collections and Technical Services (ALCTS) simply defines the goals of content creation as producing "accurate rendering" and including "clear and complete technical specifications, production of reliable master files, sufficient descriptive, administrative and structural metadata to ensure future access, and detailed quality control of processes"[19]. The information in this paper is intended to facilitate meeting the ALCTS content creation goals by arming project managers with a sound, top-down methodology.

When selecting materials and establishing parameters for a digitization projects, the details must not be overlooked. It is important to examine the materials, to identify the smallest information that needs to be recorded, and to determine sample frequency case by case. If your object details aren't finer than 1/50 inch, and you are recording at 600 ppi, your digital objects are 3 times larger than necessary. Conversely, if your fragile journals include 1/100 inch details and they are imaged at 200ppi, you may end up re-scanning or photographing the entire project down the road.

For many projects, 300ppi or 600ppi makes perfect sense. There are situations in which a lower sample frequency is appropriate. If the project merits it, there are techniques for achieving much higher resolution as well. For any digitization project, a basic understanding and appreciation of sample frequency is essential for making the most practical, informed decisions when establishing guidelines and for evaluating during the quality assurance phase of content creation.

References

[1] U.S. National Archives and Records Administration, "Technical guidelines for digitizing archival materials for electronic access: creation of production master files – raster images", (2004) pg. 51. http://www.archives.gov/preservation/technical/guidelines.pdf.

[2] Library of Congress, "Technical standards for digital conversion of text and graphic materials", (1995) pg. 5. http://memory.LC.gov/ammem/about/techStandards.pdf.

[3] New Jersey Digital Highway, "NJDH Digital Imaging Specifications", (2007). http://www.njdigitalhighway.org/image_requirements_libr.php.

[4] Cornell University Library. "Moving theory into practice: digital imaging tutorial", (2003). http://www.library.cornell.edu/preservation/tutorial/intro/intro-02.html.

[5] Richard Pearce-Moses. "A glossary of archival and records terminology", Society of American Archivists, (2005). http://www.archivists.org/glossary/index.asp.

[6] North Carolina Exploring Cultural Heritage Online, "Digital Guidelines", (2007). http://www.ncecho.org/dig/guide_4production.shtml#4.2.

[7] Bibliographical Center for Research, "CDP Digital Imaging Best Practices Version 2.0", CDP Digital Imaging Best Practices Working Group, (2008) pg. 23. http://www.bcr.org/dps/cdp/best/digital-imaging-bp.pdf.

[8] Howard Besser, "Best Practices for Image Capture", (1999) Pg.22. www.gseis.ucla.edu/~howard/MOA2/bp90.doc.

[9] North Carolina Exploring Cultural Heritage Online, "Digital Guidelines", (2007). http://www.ncecho.org/dig/guide_4production.shtml#4.2.
[10] Bibliographical Center for Research, "CDP Digital Imaging Best Practices Version 2.0", CDP Digital Imaging Best Practices Working Group, (2008). pg. 26-28. http://www.bcr.org/dps/cdp/best/digital-imaging-bp.pdf.
[11] U.S. National Archives and Records Administration, "Technical guidelines for digitizing archival materials for electronic access: creation of production master files – raster images", (2004) pg. 51. http://www.archives.gov/preservation/technical/guidelines.pdf.
[12] Library of Congress, "Technical standards for digital conversion of text and graphic materials", (1995) pg. 11-12. http://memory.LC.gov/ammem/about/techStandards.pdf.
[13] The Internet Archive, "Frequently Asked Questions", (2009). http://www.archive.org/about/faqs.php#195.
[14] New Jersey Digital Highway, "NJDH Digital Imaging Specifications", (2007). http://www.njdigitalhighway.org/image_requirements_libr.php.
[15] Bibliographical Center for Research, "CDP Digital Imaging Best Practices Version 2.0", CDP Digital Imaging Best Practices Working Group, (2008) pg. 26-28. http://www.bcr.org/dps/cdp/best/digital-imaging-bp.pdf.
[16] University of Maryland Libraries, "Best Practice Guidelines for Digital Collections", (2007) pg. 12. http://www.lib.umd.edu/dcr/publications/best_practice.pdf.
[17] Library of Congress, "Technical standards for digital conversion of text and graphic materials", (1995) pg. 6. http://memory.LC.gov/ammem/about/techStandards.pdf.
[18] Poposki Dimitar, "Impossible and advanced optical character recognition using a cheap point and shoot digital camera", (2006) pg. 1. http://www.ncd.matf.bg.ac.yu/seedi/events/Prezentacije/Poposki_pres.pdf.
[19] American Library Association, Association for Library Collections & Technical Services, "Preservation and Reformatting Section, Definitions of Digital Preservation", (2007). http://www.ala.org/ala/mgrps/divs/alcts/resources/preserv/defdigpres0408.cfm.

Author Biography

Matt Pearson is the Quality Assurance Specialist for Stanford University Libraries and an active member of the digital preservation community. Address: 71A, Green Library, 557 Escondido Mall, Stanford, CA 94305-6004; Phone: (650) 319-5639; email: matp@stanford.edu

