File Naming Standards for Digital Collections

Jessica Rogers

B Sides 2014

U of I SLIS Journal

Keywords: digitization, naming standards, metadata, collections management, special collections, digital humanities

File naming standards for digital collections are a much discussed topic among digital librarians for the simple reason that the standards do not exists. A "standard" implies a set of rules agreed upon by most members of a field; however when it comes to file naming for archival and digital collections, no set of rules exists. This is due largely in part to the fact that each library has different needs within their digital collection and that file names are based on who will be accessing the files. Many institutions create and publish their own set of conventions and there are generally two schools of thought on file naming standards; opaque and descriptive[1], and each has various pros and cons. When an excellent DAMS, or Digital Asset Management System is in place, there becomes less of a need for descriptive file naming, a file can be searched for and found easily by any number of its metadata assets. However, "without any other cataloging system in place, file names become *the* way to figure out what's inside the hundreds of thousands of files" (Beard) and thus, descriptive file naming can be a time saving necessity. We are constantly reminded of how quickly technology outdates itself and although it is unlikely that future digital collection platforms will not be able to interact with current DAMS, having a safety net in the form of descriptive file names is an easy way to ensure that digital images are accounted for within an institution's digital collection. The following paper will discuss what other university libraries are doing to establish a file naming convention within their own digital collections and will conclude with a file naming convention suggestion for the University of Iowa Library's Digital Collections.

[1] Shaw, Emily. Interview. 16th November 2012.

Although many institutions have differing opinions on file naming best practices, there is one thing that all agree on and that is the omission of special characters such as /\:*?"<>|[}&$[2]. Spaces should also be avoided and instead replaced with an underscore (_) or dash (-), this is to prevent web-based repositories from replacing the space in a file name with the characters %20[3]. Isaiah Beard suggests using PascalCase, also known as CamelCase, which involves the capitalization of the first letter of each word, instead of using an underscore to indicate a space between words. The omission of special characters and spaces is where any agreements on file naming standards stop and although many institution's self-imposed standards are similar, none are too similar to others to be considered a foundation for a universal file naming standard. For example, the number of characters in a file name is agreed to be no more than 8 by the Library of Congress[4] and University of Colorado at Boulder[5]. Hamilton College Library has a character limit of 27[6] and Robert Walker suggests a limit of 34 in his proposal for Southern Methodist University[7]. The University of Illinois at Urbana-Champaign tops the chart by placing a "limit [of] total path length to 207 characters"[8]. Typically, longer file names are discouraged on the basis of input-user error, however, renaming files using batch processing does much to eliminate inconsistent input. Although institutions have agreed on the omission of special characters, they are not agreed on

[2] Hamilton College. "Hamilton College Library Digital Collections." 6 December 2010. Hamilton College. Web site. 23 November 2012.

[3] Beard, Isaiah. "Using Consistent File Naming Conventions for Digital Preservation Projects." 21 October 2010. From Page to Pixel: One Blog. Defending Against the Digital Dark Age. Blog. 22nd November 2012.

[4] The Library of Congress. "Technical Standards for Digital Conversion of Text and Graphic Material." n.d. The Library of Congress. PDF. 24 November 2012.

[5] University Libraries Digital Project Advisory Group. "Guidelines on File Naming Conventions for Digital Collections." 4 March 2008. University of Colorado at Boulder. PDF. 16 November 2012.

[6] Hamilton College. "Hamilton College Library Digital Collections." 6 December 2010. Hamilton College. Web site. 23 November 2012.

[7] Walker, Robert. "SMU Central University Libraries Digitization Guidelines and Procedures: Best Practices for Digitization." 4 April 2011. Norwick Center for Digital Services, Central University Libraries, Southern Methodist University. PDF. 22 November 2012.

[8] University of Illinois at Urbana-Champaign. "2.0 Best Practices for File Naming." 2010. University of Illinois at Urbana-Champaign. Web site. 24 November 2012.

regular characters. Hamilton College and the University of Colorado state that 0-9 and 'a'-'z' are acceptable characters, whereas capitalized characters, 'A'-'Z', are not. Hamilton further suggests avoiding the letters 'l' and 'o' as they may be mistaken for '1' and '0'. All of the library guidelines stated that the file type extension should never be more than three characters: .doc, .pdf, .jpg, . tif, .mov, .wav[9], however Isaiah Beard talks about how increasingly sophisticated computers rely "less on file extensions to identify file types, and more on embedded metadata in the file itself to figure out how to best treat it." In addition "new file formats, such as documents made with newer versions of Microsoft Office, have mandatory four-letter file extensions by default (.docx, .xlsx, .pptx)." So, even the file type character length agreement has been undone in a way. Finally, one feature that is very useful and implemented by the University of Colorado and suggested by the Library of Congress is an accompanying "Filename Pattern Legend"[10]. In the University of Colorado's standards a file name legend is used in their digital map collection, in which the file name incudes a 3 letter city or county code. Some kind of 'read_me' information .doc, possibly located in the highest digital directory of the map images folder, contains a breakdown of 'feature codes' such as:

ala	Alamosa
arv	Arvada
yam	Yampa
yum	Yuma
un	Unnumbered sheet

[11]

In general, the basics of file naming will be consistent across collections, however, based on the oddities of some items within

[9] University of Oregon. "Recommended File Naming Conventions." n.d. University of Oregon Libraries: Digital Content Coordinators. Web site. 24 November 2012.

[10] The Library of Congress. "Technical Standards for Digital Conversion of Text and Graphic Material." n.d. The Library of Congress. PDF. 24 November 2012.

[11] University Libraries Digital Project Advisory Group. "Guidelines on File Naming Conventions for Digital Collections." 4 March 2008. University of Colorado at Boulder. PDF. 16 November 2012.

specific collections, an accompanying Filename Pattern Legend could prove very helpful.

It has been demonstrated that file naming standards vary from library to library, as it should be, based on each library's individual needs. It is important for each library to establish a file naming practice and implement it moving forward, as retroactively applying a new standard is not recommended[12]. Currently the University of Iowa uses a mixture of opaque and descriptive file names, based on the collection. For example, digital images of Civil War Diaries are typically named using a barcode identifier and journal entry date: 31858267589_1863-12-05—1863-12-10 or barcode_yyyy-mm-dd(double dash)yyyy-mm-dd. Although this process can be time consuming as it requires manual input, the descriptive nature of the file name ensures that anyone who is searching for a specific page or date will have some ease in locating it from among the other images. The Nile Kinnick scrapbook collection uses an opaque file naming system; all three scrapbooks are image named by barcode and image order number or, 31858627799_001. Due to the nature of the scrapbooks, i.e. the lack of pagination and varying date structure, an opaque file name for the digital images is all that is necessary. The collection which was the motivation for this paper, and which is the exemplar for the proposed file naming system, is the University of Iowa's Szathmary Culinary Manuscript collection. The Szathmary collection is an excellent representation of the variety of file naming issues one can encounter when using only an opaque naming system and shows how descriptive file naming can clear up confusion. The current opaque file naming system used in the collection is: barcode_### or, 31858266753_001.

The first abnormality observed was Box 19 Item 130c*. The physical item itself contained a bound manuscript and many loose pages of ephemera. It is important to note that these loose page ephemera were not interleaved among the manuscript pages but were separate, yet still considered and catalogued as one item, item 130c. The initial resulting digital image files were named as follows:

---

[12] University Libraries Digital Project Advisory Group. "Guidelines on File Naming Conventions for Digital Collections." 4 March 2008. University of Colorado at Boulder. PDF. 16 November 2012.

* The University of Iowa Libraries Special Collections is currently in the process of renumbering and relabeling item box numbers and the items themselves. The information represented within this paper is based on the previous numbering system.

Bound manuscript:

31858055131845_001
31858055131845_002
31858055131845_003
…
31858055131845_073
31858055131845_074
31858055131845_075

Loose pages:

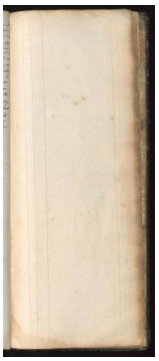
31858055131845_076r
31858055131845_076v
31858055131845_077r
31858055131845_077v
31858055131845_078
31858055131845_079r
31858055131845_079v

What is demonstrated here is that within one file folder the three digit number following the underscore changes from representing the *image* number, as is the case with the bound manuscript, to representing the *item* number, or counting the ephemera. The ephemera consisted of paper scraps that nearly all looked the same, thus it became important to indicate 'r' and 'v' in order clarify when two sides of one item were being captured instead of just one side of an item. It is important to note here that the current image capturing process does not include the scanning of blank or out of scope pages in the interest of saving time and file size. Thus, opaque file naming becomes an issue with the combination of loose and bound pages within one item digital folder.

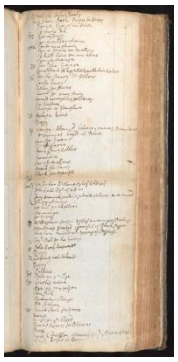
The next oddity came in the form of an alter-oriented book, or a book that has been written from front to back, then turned 'upside-down' and written back to front, with the old 'back' cover acting as the new 'front' cover. In some cases all recto pages will be right-oriented and all verso pages will be alter-oriented, but more often than not the two stop before meeting in the middle or near the 'right-side' end. Opaque file naming can often lead to confusion when viewing the images, particularly once they have been 'righted', or rotated, for user convenience. For example, a closer look at the digital images of barcoded item 31858055132264 from Box 11 shows how opaque file naming applied to an alter-oriented item which has been

'righted' can lead to conveying misinformation about the analogue item:

31858055132264_212



31858055132264_213



31858055132264_214

31858055132264_215

A user looking at this digital image may not notice that two recto pages are represented back-to-back, and if they did notice this, it might be assumed that the person capturing and ordering the images had made a mistake. What is more troubling than what the above combination of file naming and image rotation shows, it what it does NOT convey. Not only is the alter-orientation of the item not represented but there are also 34 blank leaves in-between the two orientations. These two points of information about the physical structure of the book are lost between image numbers 213 and 214. Also, due to an extensive index at the beginning of the manuscript the pagination is actually 189, 190, 222, 223, respectively. A person searching for page 189 would have to view various and random images before discovering that the page 189 correlated with image number 212. The above example shows how deeply lacking opaque file naming can be in an as-accurate-as-possible representation of the analogue item.

The solution to clarifying both aforementioned scenarios is to apply a descriptive file naming structure to the images. Below is an outline of basic tenants I think are most important in descriptive file

naming and, when applied to the above examples, work in tandem with the image to inform the user of the physical properties of an item that are otherwise lost in digital capture.

File Naming Standard Implementation

As Applied to the University of Iowa Library's Szathmary Collection

These guidelines are intended to adapt a basic file naming system which will represent and describe the digital image of an analogue item, make archival file names searchable, and be consistent so as to be understood and searchable by users not familiar with the collection.

This system is divided into two components, basic and additional. Every file, without exception, will have a basic file naming component structure. Additional file naming components are used to describe the image and vary from one image to the next.

Basic File Naming Components – Object Identification Number (OIN) & Image Order Number (ION)

Object Identification Number: **31858055134228**_010_216Va

This serves to indicate that all images in the folder belong to the same item. If any file image goes astray from the item folder, the OIN will act as a location marker. It will also link the digital images to the analogue item and finding aid.

Image Order Number: 31858055134228_**010**_216Va

Very important! The ION properly orders digital images in relation to the physical item and serves as collation for the digital surrogate. Sequential image ordering indicates to users and future users that this is the order in which the item was found and captured from "front" to "back", despite potential image inconsistencies in pagination, recto/verso order, or text orientation. It is possible to have more ION's than page numbers of the item.

Additional File Naming Components

Page Number: 31858055134228_010_**216V**a

Page numbering will follow actual item whenever possible and will otherwise be represented in modern format. Page numbering

may vary rarely be unnecessary when basic file naming (OIN & ION) is sufficient.

Recto & Verso: 31858055134228_010_216**V**a, 31858055134228_015_002**R**

Is helpful when orientation is "flipped" to show that although the righted image may appear to the user as the right side (recto) of the item, it is in fact the left side (verso) as the item was captured 'front' to 'back'. Recto/Verso also serves as a quality control check.

Page Image Number: 31858055134228_011_001R-**01**, 31858055134228_012_001R-**02**, 31858055134228_013_001R-**03**

This feature allows for one page to be captured multiple times if necessary and is how ephemera and foldouts found tucked between pages will be named.

Lowercase duplicate pagination identifier: 31858055134228_010_216V**a**, 31858055134228_016_216V**b**

The lowercase letter represents a distinction between two similarly numbered pages. This occurs when the person(s) responsible for original pagination accidently lost count and repeated a number, but not to such an extreme degree that original pagination could not be followed. Additionally, pagination will start over mid-way through the item, and this too can be represented by a lowercase identifier in the file name.

Ex:
…
31858055134228_035_022Ra
31858055134228_036_022Va
31858055134228_037_001Rb
31858055134228_038_002Vb
…

Following are further explanations and examples of the above components:

1. Basic file naming will include 1.) Object Identification Number (OIN), in most cases represented by a barcode, and 2.) Image Order Number (ION).

Ex: barcode_img#
31858055134228_001

The image number will increase with every image and will never be repeated. This guarantees that what is viewed is the most accurate representation of the item as captured from FRONT to BACK.

2. Pagination of original document will be followed if it is mostly (at least 50% of text) consistent. Pagination will not begin with table of contents or fly leaves. Therefore, table of contents, indexes and all other front, back, and non-paginated matter, will be represented with basic (opaque) file naming structures. Recto and verso will always be represented by 'R' and 'V'

Ex: barcode_img#_pg#R/V

If pagination is modern (increases with every turn of *page* i.e. recto/verso):
31858055134228_001 (front cover)
31858055134228_002 (inside front cover)
31858055134228_003 (title page)
31858055134228_004 (table of contents)
31858055134228_005_001R
31858055134228_006_002V

If pagination increases with each *leaf*:
31858055134228_007_003R
31858055134228_008_003V

Note: The image number always increases, so even if "R" and "V" are not defined the image order will not be upset.

3. Whenever original pagination is not consistent or is discontinued midway, modern pagination standards will be applied, i.e. recto to verso numeration is increased by one. This may mean that file naming pagination will change within one item folder which is why image numbering is important.

Ex:
31858055134228_097_093R
31858055134228_098_093V

31858055134228_099_094R
31858055134228_100_095V

4. When orientation of text is reversed midway (alter-orientation), the item will continue to be scanned "front" to "back", imaged ordered and file named accordingly. Post file naming all alter-oriented images will be rotated 180 degrees using Photoshop Creative Suite 6 batch processing, for user legibility. An accompanying read_me.tif will precede the re-oriented section. Basic file naming will apply to read_me.tif's, which will be image numbered according to relevance. read_me.tif images will be further explained in guideline #9.

Ex: barcode_img#_read_me
31858055134228_009_read_me

5. If alternate pagination is used for alter-orientated pages and is consistent then this too will be represented in the file naming.

Ex:
31858055134228_007_001R
31858055134228_008_217V
31858055134228_009_002R
31858055134228_010_216V
31858055134228_011_003R
31858055134228_012_215V

Notice how the image order reminds the user that the images have been correctly captured from front to back, and indicates that the manuscript is alter-oriented.

6. If alter-oriented pages are less than half of the text block and have no original pagination then they will be paginated from 'front' to 'back' before rotation. This will appear consistent with 'right-side-up' file naming pagination.

Ex:
31858055134228_088_149V
31858055134228_089_148R

31858055134228_090_read_me ('Note: pages are alter-oriented')
31858055134228_091_149V (alter-oriented page, rotated for legibility)
31858055134228_092_150R
31858055134228_093_151V (alter-oriented page, rotated for legibility)

7. Multiple captured images of one page will be represented with a '-', this will be useful for fold-outs and inserted ephemera. Page specific image numbering will begin at '1' every time.

Ex: barcode_img#_pg#R/V-pgimg#
31858055134228_010_008V-01
31858055134228_011_008V-02
31858055134228_012_009R
31858055134228_013_010V-01
31858055134228_014_010V-02

8. If item pagination is mainly consistent but displays a few repeats, these will be represented by a lowercase letter beginning each time with 'a'.

Ex: barcode_img#_pg#R/Va
31858055134228_015_009Ra
31858055134228_016_009Va
31858055134228_017_009Rb
31858055134228_018_009Vb
OR
31858055134228_015_009Ra
31858055134228_016_010Va
31858055134228_017_009Rb
31858055134228_018_010Vb
OR
31858055134228_015_009Ra
31858055134228_016_010V
31858055134228_017_009Rb
31858055134228_018_011V

There can be many variations within each item, but maintaining consistent 'basic file naming' will insure correct image order and therefore item representation.

9. read_me.tif's will be used sparingly but will be helpful to explain discrepancies within the item, such as missing pages, an unexplained jump in pagination, and any points of interest that cannot be fully represented within the digital image and file name. For example, if original pagination is present but in no way consistent then modern pagination standards would instead be applied to the file naming process. In this case the page number represented in the file name will likely not match with the page number as seen in the image. This issue would be addressed by image 001 being a read_me.tif.

Ex: 318580551342228_001_read_me

"Note to user: Pagination of original item is inconsistent and for that reason has not been applied to the file naming of the digital images. Please understand that what you see is an exact representation of the item and although page numbers in the images are not sequential, the page images are displayed in the order in which they were bound."

Each read_me,tif will end with this disclaimer:

"This is a memo clarifying the digital image for the benefit of the user and is not part of the original artifact."

10. Blank pages are not to be captured, but instead counted by leaf and represented with a read_me.tif. Pagination will pick up with the addition of the blank pages.

Ex:

31858055134228_087_102R
31858055134228_088_103V
31858055134228_089_104R
31858055134228_090_read_me (22 blank leaves = 44 pages)
31858055134228_091_149V

11. Should ephemera not be interleaved but separate from the item it will be represented using basic file naming components only *if only one side is captured*. If both sides of the ephemera are

digitized then they will be represented with an 'R' and 'V' at the digital technician's discretion, meaning that because ephemera is not paginated there is not always a distinct front or back.

Ex:
31858055134228_011_R (front of ephemera #1)
31858055134228_012_V (back of ephemera #2)
31858055134228_013 (ephemera #3)
31858055134228_014_R (front of ephemera #4)
31858055134228_015_V (back of ephemera #4)
31858055134228_016 (ephemera #5)

Further examples:

31858055134228_011_001R-01
31858055134228_012_001R-02
31858055134228_013_001R-03
31858055134228_014_216Va
31858055134228_015_002R
31858055134228_016_215V
31858055134228_017_003R
31858055134228_018_216Vb

The above indicates that pagination begins 11 images into the item, the first 10 images are front cover and other front matter. There is likely some ephemera loosely interleaved between page 1 and the opposite verso. The three images will capture the recto and verso of the ephemera and page 1 cleared of the ephemera. On the verso of page 1 we find the end of the alter-oriented text. The alphabetic addendum indicates that the initial writer lost count during pagination. Both the image order number and the V next to the page number inform the user that although the page appears to be recto due to rotation for legibility, it is actually the verso of the previous image.

Possible variations:
Basic and additional file naming components be separated by a dash (-) instead of an underscore (_):31858055134228_016-216Vb. This distinguishes the foundational digital information from item representative information.

Pagination represented by the indicator 'p': 31858055134228_016-**p**216Vb.

Use appropriate amount of leading '0's to represent number of pages.

Let us now return to Box 19 item 130c, the bound manuscript with loose ephemera. Using the above mentioned guidelines the file naming will appear as follows:

Bound manuscript:

31858055131845_001
31858055131845_002
31858055131845_003_01R
31858055131845_004_02V
…
31858055131845_078_80V
31858055131845_079_81R
31858055131845_080

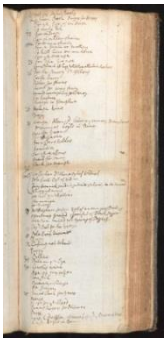
Loose ephemera:

31858055131845_087_R
31858055131845_088_V
31858055131845_089
31858055131845_090_R
31858055131845_091_V
31858055131845_092

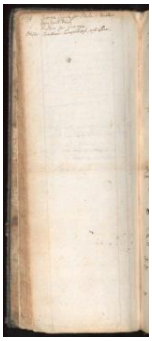
When combined into one digital file folder the bound and loose items transition smoothly from one to the other, representing the items as complimentary to each other and better representing the analogue item.

Further, applying the above guidelines to barcoded item 31858055132264 from Box 11 yields a much more comprehensive and closer-to-accurate representation of the actual item.

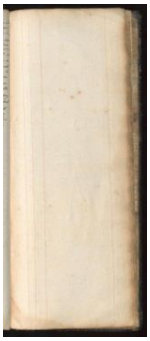
31858055132264_212_189V



31858055132264_213_190R



31858055132264_215_222V



31858055132264_216_223R

From this point 34 blank pages follow. Script was upside down due to the back of the book being used as a new front. The images have been righted for reader convenience.

This memo is added for conveying information about the item and is not a part of the original artifact.

31858055132264_214_read_me

The file names of the images are now descriptive of the analogue item and where file naming alone does not suffice, the read_me.tif fills in to give any extra information.

A descriptive file naming system is both important and useful in that it guides users to specific images while also conveying information about what the image is and it's relation to the other images. A captured digital image is a surrogate of an analogue item which, once captured, is removed from its origional context. Librarians and preservationist need to insure that these out of context images are re-constituted in a manner that closest represents the analogue item. Although it is hard to accurately represent digitally what is physical, it is better done with descriptive, meaningful, and informative file names.