Quality Control: Basic Inspection and Verification Steps to Ensure Accurate and Correct Manuscript Digitization Work

March, 2018

Wayne Torborg, Director of Digital Collections and Imaging, HMML

The Hill Museum & Manuscript Library (HMML) has created a number of instructional documents outlining the methods for producing digitized manuscripts that meet HMML's standards. These documents can be found on the web at the following URL:

http://www.vhmml.us/Resource/Downloads/

Updates were made to the most important of these documents in 2016—the information contained in the first four of these documents are the starting point in achieving proper results in manuscript photography.

Even with the best intentions and reference resources, mistakes can occur anyway. At times, these errors are not detected until many years after photography has taken place, making correction difficult. Outlined here are some basic quality-control checks that can easily find common file-naming and sequencing errors. These checks should be performed by the digitization team at the studio site prior to sending manuscript materials to HMML—this way errors can be corrected right away.

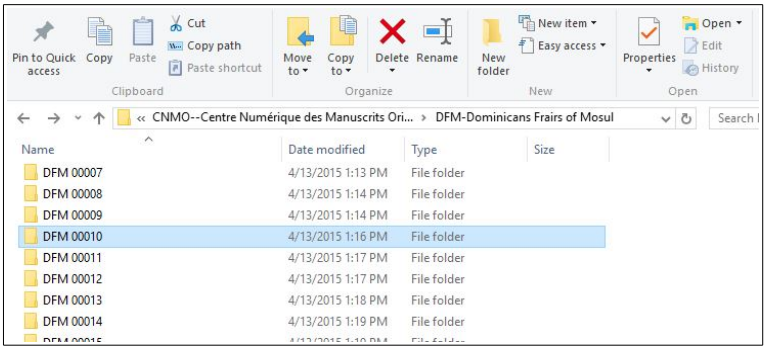
Is Everything Named Correctly?

HMML uses a format called "HMML Project Number" to identify a particular manuscript item. This is also often referred to as the "Source." The formation and syntax of these project numbers is exhaustively outlined in the *Foliation and Filenames* document found at:

http://www.vhmml.us/Resource/Downloads/2016%20Foliation%20and%20Filenames.pdf

Once a particular HMML project number is decided upon for a given manuscript item, there are a number of things that have to be in order and properly done. They are:

-Folder Containing Manuscript Images Must Have the Proper Name

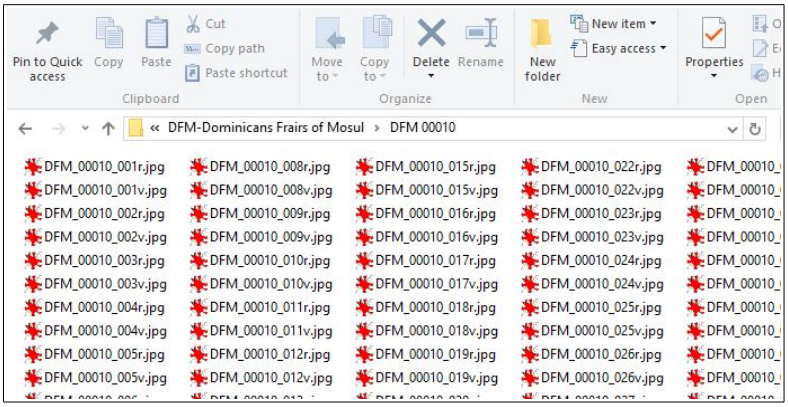


Properly-Named Folder. The HMML Project Number for this manuscript is "DFM 00010." Note the space between the alphabetic "project prefix" code and the five-digit number.

HMML uses a consistent project number syntax consisting of an alphabetic project prefix and a five-digit number. This number may or may not be the same as the original shelfmark or library number

assigned to the object at its holding library—it's nice if the number is reflective of the shelfmark, but not necessary.

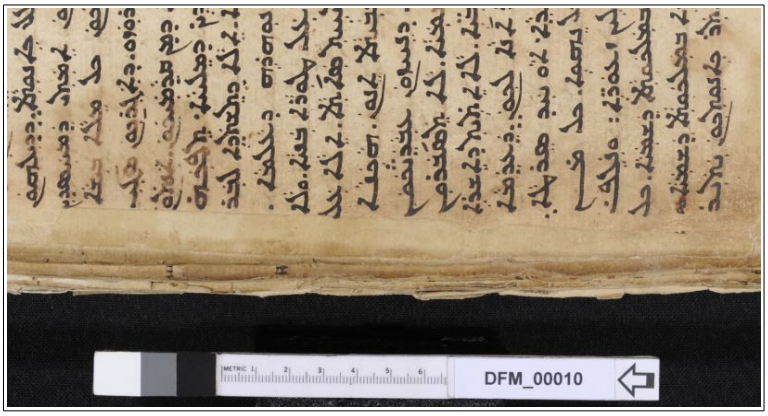
-Files in Folder Must Have the Same Project Prefix and Number



Files Contained in the "DFM 00010" Folder. Note that in filenames, HMML uses underscore characters instead of the space used in the folder name. This assures that such filenames will sort properly on all common computer platforms.

While the proper folder name uses a *space* between the project code prefix and the number, *underscore* characters are used in the filenames to ensure a filename without spaces. The proper form is a single underscore. Sometimes we see double underscores, hyphens, or combinations of underscores and actual spaces, underscores and hyphens, etc. These are all incorrect naming conventions.

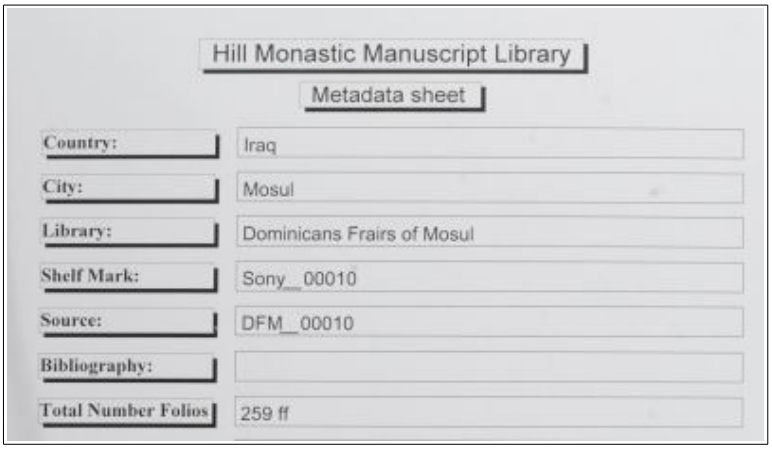
-Images Must Have Proper Paper Label



Small Paper Label. This is usually affixed to the metric/grayscale with sticky putty or wax. It must match the HMML project number for the manuscript being photographed.

In the manuscript images done for HMML, a metric/grayscale is included in the photograph. There is room on this metric scale to affix a small paper label (using sticky putty or wax) indicating the project number for the manuscript item. This can be created using a computer print-out or carefully lettered by hand. It's easy to forget about the label and leave in in place when one begins photographing the next manuscript; this creates a disagreement between the folder name, filenames, and the paper label. This should be avoided.

-Paper Metadata "Slate" Must Have Proper HMML Project Number



Paper Metadata Sheet or "Slate." This is photographed along with the manuscript itself and included with the image set. Here, the "Source" number is the project number which matches the folder, filenames, and small paper label. Note that in this case the shelfmark and project number are not the same—this is sometimes the case, and it's not a problem.

A paper sheet containing basic metadata is produced for each manuscript object photographed for HMML. Sometimes this is generated by database software used to create the manuscript inventory; other times it is simply typed on the computer and printed out. In some cases, it is a paper form filled out by hand. In any case, the Project Number (sometimes called the Source) has to match the folder name, the filenames, and the small paper label that appears in each photograph.

All is well when these things are in agreement with each other. If mismatches are encountered at HMML, it becomes a guessing game as to what is wrong and what is correct if these various elements do not match.

Is the Set of Files Complete?

The image files for a given manuscript object need to be named properly and should sort in proper reading order. If any pages are skipped or not photographed for whatever reason, the resulting set of files will not "read" correctly when sorted by filename. The text will "break" at the point of the missing page.

Manuscripts can be photographed as single pages or as two-page spreads. With either method, it's possible to mistakenly skip or repeat a particular page or spread.

When photographing using the "two-page spreads" method, a missed spread image will cause a "gap" in the text; after the gap, however, the text will read correctly. A spread that is accidentally imaged twice simply creates a "repeat" in the text—not a serious error, but one to avoid nonetheless.

In a manuscript imaged as single pages, a missed page will generally scramble the continuity of the text as read in successive pages. This is because the content becomes "shifted," in a manner whereby the images *identified* as being on the same folio via filename (say, 024r and 024v) are actually on *different* folios due to a missing image (which could be anywhere *on or before* folio 24 in this example). Further errors mess things up even more.

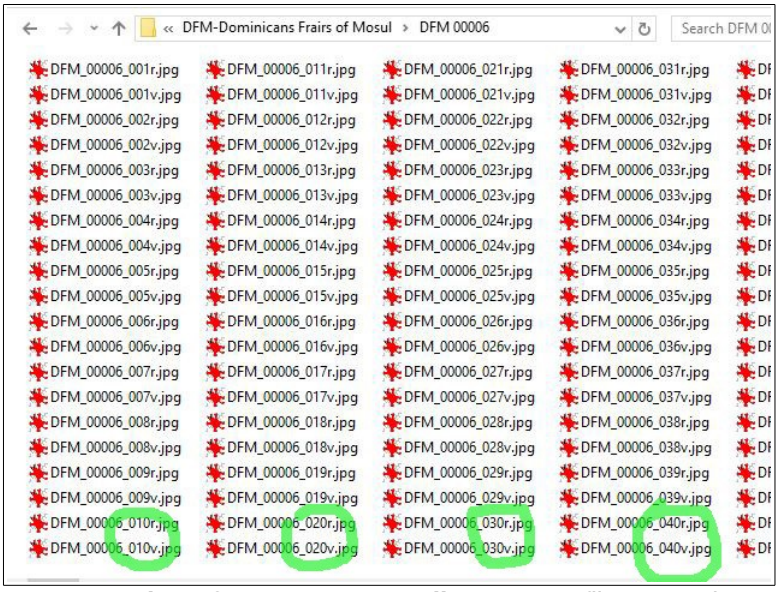
So how to prevent such problems? The first line of defense against errors is proper preparation. Before any photography proceeds on a manuscript to be imaged as single pages, the following should be done:

-Count the folios. At HMML, we consider the first "free" leaf (not pasted to the binding board) to be the first folio, continuing to the last free leaf. It may be prudent, when working with a manuscript that has pages that want to stick together, to count them *twice* to make sure what the number is. After photography, one should have this number of recto and verso images. No more or less, the same number for both recto and verso, matching the count made beforehand.

-Number the folios. Often, this will require the services of the official librarian, archivist, etc. as well as permission to do so. The book may already be foliated, but often this is where the numbering begins at the first page containing the main text rather than the first *physical* folio present in the book. It is also the case that "old" foliation can contain errors.

-Check the photography as you proceed. Say you reach the 100th recto image. Take a look at the contents of the *reverse* (verso) side of that particular page. Write down notes on what is there (the first words, first sentence, any decoration, etc.). Then, when you reach number 100 during the "verso" photography, it should be the page you identified by those notes. Do this several times during the course of photographing a book and it becomes much easier to produce perfect work.

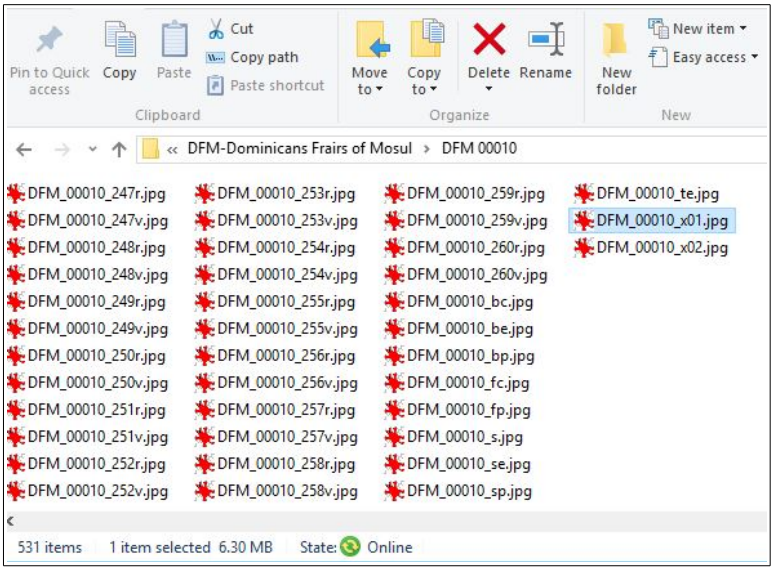
Additional information about the proper photography of books imaged as single pages can be found at length in the *Foliation and Filenames* document (link found on first page). When finished, you can verify that things are in order by simply sorting the files by filename. Then, choose to view the folder in "list" format and size the window so that 20 files fill each column.



Proper sorting of "Recto-Verso" Files. Twenty files per column. Note that the bottom of each column has the proper recto and verso file in multiples of 10.

Count the files in the column to verify that there are indeed 20 of them. After doing this, one should have the recto and verso for each file with a number that is a multiple of 10 at the bottom of each column. This pattern should continue until the last numbered file—whatever that particular number is, there should be a recto and verso for it.

The images of the front and back covers, spine, edges, etc. have their own special syntax outlined in the *Foliation and Filenames* document. Any additional things found in the book (notes, loose fragments, etc.) are given filenames to ensure that they sort at the end of the list.



Additional Files. After the last verso file (260v), the “additional” images, including the covers, edges and such are listed. Finally, two additional fragments found in the book are listed. The use of the letter “x” forces these “eXtra” files to the end of the list.

Examples of Errors

It’s often illuminating to see examples of things that aren’t correct in order to understand *why* they’re the wrong way to do things.

-Improper Placement of “r” and “v” in Recto-Verso Filenames

Improper Placement of “r” In Filenames. By placing the “r” or “v” before the folio number, the files sort incorrectly.

In the example depicted above, the technician improperly named the recto and verso images in the manuscript, putting the “r” and “v” *before* the folio number for each. This causes an incorrect sort,

where the spine, covers, and other additional images are listed first, all the rectos follow (in numeric) order, and finally all the versos in numeric order.

The solution is to get rid of the “r” and “v” characters and place them *after* the folio numbers, which will cause a proper filename sort. There are a number of ways to do this using a variety of tools, but the basic steps would be:

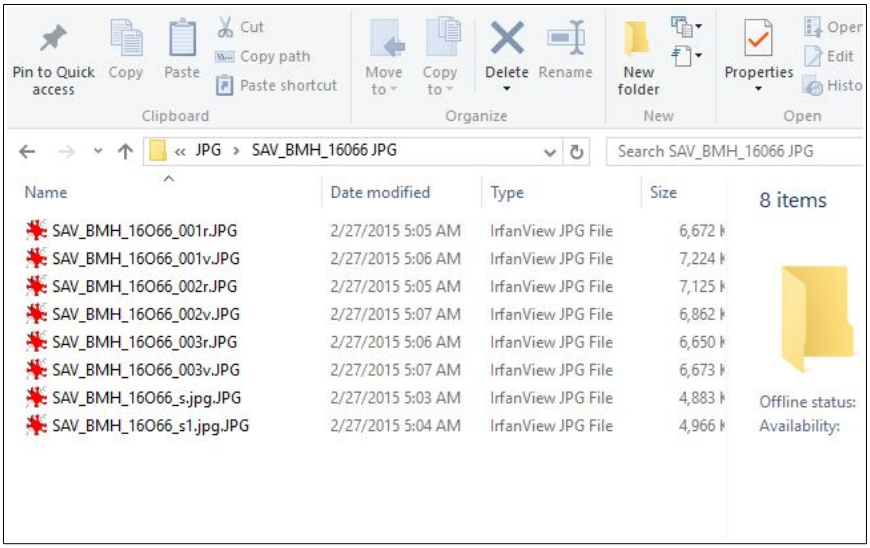
-Separate the files into recto and verso folders
-Remove the “r” from the recto files
-Insert “r” after the folio number in the recto folder
-Remove the “v” from the verso files
-Insert “v” after the folio number in the verso folder
-Put all the files back into the parent folder

A number of file-naming utilities can be used to do these things. At HMML, we make extensive use of an application called Bulk Rename Utility, which is free for non-profit use. It can be downloaded from:

http://www.vhmml.us/Resource/Downloads/

The user manual for Bulk Rename Utility can be download from this page as well and is well worth reading.

-Mistakes in File Naming Syntax

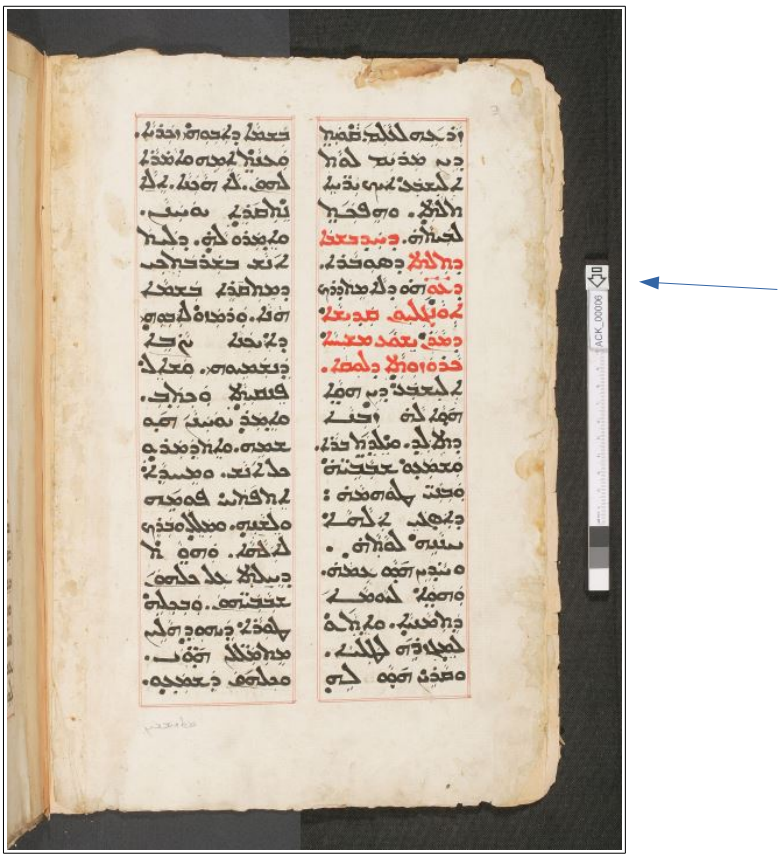


Mistakes in File-Naming. Here, two errors are on view. First, the uppercase letter “O” has been used in place of a proper numeric zero. In the last two files, the technician has unnecessarily put “jpg” before the file extension. Since the system automatically adds the file extension, this text is repeated unnecessarily.

Over the course of years, HMML has seen a number of file-naming errors, including:

-Improper use of hyphens instead of underscore characters
-Spaces in filenames
-Double underscores or combinations of underscores and hyphens
-The manual addition of file-type extensions, such as “jpg” to filenames
-Typographical errors, such as transposing numbers in file-names (this is quite common)

-Improperly Positioned "Up" Arrow



Improperly-Positioned "Up" Arrow. The small paper arrow is supposed to indicate the upright positioning of the page image.

HMML has a downloadable sheet of arrow symbols available at:

http://www.vhmml.us/Resource/Downloads/arrows.pdf

In practice, a small paper arrow symbol is cut out of this sheet and affixed to the metric scale that is in each page photograph (along with the small paper project number label). This is supposed to point "up" with relation to the orientation of the manuscript page. When photographing books as recto and verso pages, it is common to *forget to switch* the orientation of this paper arrow after finishing the recto pages and proceeding with the verso pages. The result is that the verso images end up with the arrow pointing downwards with relation to the proper text orientation.

Here at HMML, we often employ student workers to rotate page images into proper reading orientation. In many cases, these students aren't familiar with the languages or scripts present in these manuscripts, so they go solely by the orientation arrow. If the arrow is wrong, the image will often be rotated incorrectly.

-Improper Folder Structure on Hard Drives Containing Manuscripts

This is outlined in detail in the *Foliation and Filenames* document, but needs to be reiterated here. At HMML, we copy all the JPEG versions of photographed manuscripts to an in-house network server. The easiest way to do this is to have all the JPEG versions of each manuscript in a master "JPEG"

folder, completely separate from the RAW versions of the manuscript images. This way, all the folders can be selected from the JPEG folder at once and copied *en masse* to the server. This involves structuring the hard drive as follows ("TEST" being the sample project prefix):

JPEG
- **TEST 00001**
- **TEST 00002**
- **and so on...**

RAW
- **TEST 00001**
- **TEST 00002**
- **and so on...**

Sometimes people put the RAW and JPEG folders together under a main project number folder. In this case, HMML workers have to copy the JPEG folders to the server individually, which is tedious.

Sometimes, we find a project number folder containing both RAW and JPEG images together. In this case, HMML workers have to sort the files in the folder by file type and manually separate them before copying them them individually to the server, which is even more work.

Implementing Quality Control

Quality control is not magic. There is no process or method by which people suddenly gain the ability to *never make mistakes*—that's not possible. The key to quality control is to check the work as it's being done and make whatever necessary corrections are needed as the work proceeds. The result looks like flawless work but is in fact the product of continuous vigilance on the part of all team members involved.

A quality-control checklist has been devised by HMML and is available for downloading at:

http://www.vhmml.us/Resource/Downloads/QC_Checksheet.xlsx

The user of this sheet can add HMML project numbers into the list and check off (with "yes," "x" or "y," etc.) each of the columns describing a quality control criterion. Depending on the work, the column titled:

For Single-Page Image Sets, Equal Number of Recto and Verso Images

may have NA ("Not Applicable) entered if the manuscript in question was photographed as two-page spreads instead of as single pages.

The last column, titled:

Manuscript Reads Correctly

is the final quality check on the photography work. Here, someone who can read the particular language and script for a given manuscript should sort the image files in filename order. Then, she or he should pick a spot near the beginning of the text and read a short passage that spans a page break; this ensures that the text makes sense and has not been "broken." Do this again somewhere in the

middle, and one more time somewhere near the end. If all three checks result in logical text continuity, it can be assumed that nothing was missed and that files are in proper order.

If need be, this checksheet can be made available as an online Google document for use by project technicians.

By understanding, implementing, and verifying the quality-control criteria described in this and other HMML documents, HMML's manuscript imaging work will achieve a higher level of accuracy and consistency.

Addendum: Autumn 2018

Since the first writing of this document, I have corresponded with a number of HMML technicians regarding the quality-control checking of completed manuscript photography. One thing that I often hear is that folks are checking the manuscripts on a given hard drive before sending it to HMML. That is, they are checking many manuscripts at a time, often weeks or months after the actual photography has taken place.

My opinion is that this is not the optimal way to do this. Here at HMML, Mary Hoppe is our book photography technician. Her procedural method is to check the completeness and continuity of a photographed book or manuscript *immediately after completing the photography*. All quality checks are completed *before* proceeding on to the next book. This is a superior way of doing this because:

-All knowledge of the book is still fresh in the photographer's mind rather than being a distant memory. Any notes taken by the photographer during the imaging of the book are also right at hand.

-The checking of the page images can be done with the actual book or manuscript at hand. Images can be checked against actual book pages. If one waits until a later date to do this quality checking, it may be necessary to retrieve the book, which may have been returned to a secure storage location (or may not be available at all).

-If errors are found, they can be corrected immediately. If the checking is done right after the photography, the camera is most likely still at the correct column height, the lights are set up properly, and everything is as it was for the photography of that particular book.

I would encourage all of HMML's projects to employ this quality-control methodology if at all possible.