
Manuscript Digitization: Overall Procedural Outline

Copyright© 2009 Wayne Torborg, HMML Director of Digital Collections and Imaging

The Hill Museum & Manuscript Library (HMML) has developed an efficient, cost-effective system to digitize manuscripts and rare printed books. HMML began its digital initiative in 2003 with a single project operating in Lebanon. As of this writing, HMML maintains over a dozen separate manuscript preservation projects and intends to expand its activities.

During this time, various “how-to” documents were created to provide training reference materials for people working on these projects. This document is an attempt to provide an overview of the entire digitization process with references to the documents already written that address specific aspects of the process.

These documents are available on the World Wide Web at the URL:

<http://www.hmml.org/wtorborg/downloads/downloads.htm>

The HMML Digitization Process in Brief:

Manuscript preservation is a photographic process; the technician photographs the pages of the manuscripts using a digital single-lens-reflex (dSLR) camera. The image files go directly into a personal computer rather than to the camera's internal memory card. Once an entire manuscript is photographed, the technician moves the files to a folder named for the manuscript and moves on to the next one.

At some point in the process, the manuscript image folders are copied to an external hard disk drive attached to the PC computer. This hard disk drive is then moved from the “photography” computer to the second computer, where images can be sorted, renamed, rotated into proper orientation, etc. If a second technician can work on this PC, the camera operator can continue to photograph manuscripts without delays. In some cases, two external hard disk drives are provided to allow the two computers to swap the drives as needed.

Once the image files are sorted, renamed, and structured properly, a set of of DVDs is produced for the holding library. The manuscript images are then copied to external hard disk drives provided by HMML; these are shipped to HMML when they are appropriately filled up.

The processes outlined here are somewhat generalized; the specific documents referenced will have the detailed information needed. In broad strokes, here are the important points:

Make Sure the Project is Ready to Begin

The Hill Museum & Manuscript Library acts as a partner to the institution holding the actual manuscript materials. As such, written agreements are created and signed before any manuscript digitization work commences. Before beginning, the technicians and scholars need to have a firm idea of the scope of the preservation work and what materials are to be digitized. For the photographic

technicians at a given project site, this generally will have been determined before the technical work is started.

Prepare the Workspace

Download and read this document:

<http://www.hmml.org/wtorborg/assets/HMML%20Digitization%20Studio%20Setup.pdf>

A working environment needs to be prepared for the digitization work. In most cases, this workspace will be in place for months or perhaps years, so strive for a semi-permanent setup. Adequate electrical power supplies are needed—two PC computers will be running, as well as a set of studio flash lighting units. If possible, it's best to have the computers plugged into an uninterruptible power supply (UPS) unit. This type of unit will provide battery power to run the computers for a short while in the event of an electrical power supply failure. This gives the technicians valuable time to suspend the work activity and properly shut down the computer systems without catastrophic crashes.

An area measuring about 3x2.5 meters in size will be needed for the photographic setup and the PC computer attached to the camera. The camera stand itself needs to be mounted on a very sturdy table (clamps are provided to allow the stand to be quickly attached without extra tools). A second table for the computer system is also needed in most cases.

In addition to this, it's usually good to have some additional space for a third table for the second computer system. Manuscripts awaiting photography can be sorted and queued up on this table. This second computer will be used for metadata entry, image file handling, and DVD recording.

Prepare the Supplies, Tools, Etc.

HMML provides the tools needed for manuscript photography. A book cradle system is provided, along with grayscale cards, plastic hold-down tabs, dusting brushes, tweezers, ruler, etc. In addition, there are some downloads you should get:

<http://www.hmml.org/wtorborg/assets/arrows.pdf>

<http://www.hmml.org/wtorborg/assets/metadata%20slate.pdf>

The documents can be printed out on any laser or inkjet printer. The “arrows.pdf” document has a bunch of arrow symbols that can be cut into individual arrows that are placed next to the manuscript photographed to indicate which way is up. The metadata slate document is likewise printed out and cut in half to create a bunch of forms that are filled out and photographed (you'll need a slate for each manuscript photographed).

PC Computer Configuration

HMML uses standard PC computers running Microsoft Windows XP. As of this writing, the software needed to control the digital cameras has not been tested on Vista, the newest Microsoft operating system. Users are encouraged to stick with XP until HMML tests the new operating system with the

software used.

HMML generally prepares the computers before they are deployed or sets them up properly when a project is started. Users shouldn't have to install software or reconfigure the computers once they are in place. Computers are supplied with the following software:

- Microsoft Windows XP (operating system)
- Microsoft Office (Excel, Access, Word, Powerpoint)
- Camera Control Software for Digital Camera (Canon or Nikon)
- Image Browser (Canon or Nikon)
- Flash Renamer (for renaming masses of image files)
- DVD Recording Software (to create DVD disks)
- Adobe Reader (for viewing PDF files)
- WinZip (for unzipping file archives)

There are often some other applications installed on the computers set up by HMML, but those listed above are the ones needed for HMML's work.

Some studios will choose to connect the two computers using Ethernet cabling to allow data to be transferred between them. It is not advised by HMML to connect these computers to the Internet, however.

Prepare Manuscripts for Photography

Generally, library specialists will gather the manuscripts to be photographed and bring them to the studio setup. Before they are photographed, data about the items needs to be entered into the second PC computer (the one not used for photography). HMML needs to have at least a basic set of metadata about a manuscript in order to have something to put into its master manuscript database.

HMML has created two different data-entry tools for manuscripts:

<http://www.hmml.org/wtorborg/assets/Metadata.zip>

This is a simple Microsoft Access database that allows the user to enter manuscript information into a form, creating a “flat-file” database of manuscripts. It has provisions for printing out a metadata sheet that can be printed out and photographed as part of the manuscript photography.

<http://www.hmml.org/wtorborg/assets/MASTER%20CW%202009.xls>

This is a more comprehensive cataloguing tool created in Microsoft Excel. This tool allows the user to enter more information than the one listed above, and is intended to be used by scholars or specialists who can provide detailed descriptions of manuscript contents. If one intends to use this tool, she or he should also download:

<http://www.hmml.org/wtorborg/assets/Cataloguing%20workbook%20manual.pdf>

This document explains how the Excel-based metadata system works. Further queries about the system can be directed to HMML.

Once data about a manuscript is entered in one of these two systems and a metadata sheet is either printed out or filled out by hand, the actual photography of the manuscript can take place.

Getting Ready to Photograph

To understand the rationale behind HMML's manuscript digitization methods, download and read:

<http://www.hmml.org/wtorborg/assets/Rationale.pdf>

This document outlines the basic reasoning behind HMML's choice of digitization methods, covering cameras, lighting systems, book holding devices, etc.

For information on some of the specifics of manuscript photography with digital SLRs, download and read:

<http://www.hmml.org/wtorborg/downloads/items/book%20photography.pdf>

This document outlines the book photography procedures to a greater degree. It's also essential to read and understand the document:

<http://www.hmml.org/wtorborg/assets/foiliation&filenames3.pdf>

This explains the methods and software used to create image files with the proper filenames—this will in turn allow a folder of images to be sorted by name, keeping the pages in the proper order.

Learn to Use the Digital Camera

A variety of digital single-lens-reflex cameras are suitable for manuscript photography. The following camera models are currently in use by HMML:

Canon 1Ds
Nikon D2x
Canon 5d
Nikon D200

Camera expertise is helpful but not absolutely necessary for technicians employed by HMML. User guides for these cameras can be found at:

<http://www.hmml.org/wtorborg/downloads/downloads.htm>

Guides have been written for specific camera models explaining how to set them up for manuscript photography. In addition, the official camera manuals from Nikon and Canon are available here. Once you know which camera model will be used, it's best to read the guides and familiarize yourself with the camera. Currently, HMML is purchasing the Canon 5d for its new projects.

Decide on a Project Prefix Code

The projects partnering with HMML are given a simple alphabetical code agreed upon by HMML and the holding institution. This is an important thing; this code is used in a number of ways:

It's used as the filename prefix for naming the image files
It is part of the HMML “source” number for a particular manuscript
Used to label folders to identify specific manuscripts

Some example of HMML project codes:

| Project Code | Code Translation | Location of Project |
|---------------------|-------------------------------------|----------------------------|
| OLM | Order of Lebanese Maronites | USEK, Kaslik, Lebanon |
| CFMM | Church of the Forty Martyrs, Mardin | Mardin, Turkey |
| MLRI | Muzeul Literaturii Române Iași | Iași, Romania |

Each individual manuscript is assigned a “source number” based on this prefix. Examples:

OLM 00256
CFMM 00015
MLRI 00006

The Prefix is written in capital letters, followed by a space. The number following the prefix is always written out with five “placeholders,” allowing up to 99,999 manuscripts to be identified for a specific collection. This is the information entered in the “source” field for HMML's metadata systems.

A particular file within a manuscript folder might have a name such as:

OLM_00357_005r.JPG

This file is from the OLM project, number 357. The image is of a single page, number 5 recto. It is a JPEG image file. The next file in the proper sequence would be “5v,” with a name as such:

OLM_00357_005v.JPG

Since “r” comes before “v” in the alphabet, this system ensures that files sorted by filename will fall into proper order.

Things that cannot be part of a HMML source number.

The source number is used as part of a database record for a number of interrelated systems at HMML. Programming languages and web scripting are used for querying and linking these records. Thus, there cannot be any alphabetic characters in the HMML source number that will cause the programming code to fail. Among these are:

() * & @ \ / < > | ^ \$

Generally, folks tend to create a source number that matches the actual *shelfmark* of the manuscript in question. This is a good thing to try to maintain, but if shelfmarks contain any of these sorts of alphabetic characters, problems will arise. Work around this problem so that the HMML source number won't contain “illegal” characters. An example:

Shelfmark of Manuscript
SCAA 005(16)

HMML Source Number Assigned
SCAA 00005 16

In this case, the parenthesis characters used in the actual shelfmark of the manuscript caused HMML's web programming to fail on these records. Simply replacing them with spaces fixed the problem. Note also that the primary source number has been padded out to five digits in the corrected version.

Consult with HMML if further information is needed on creating viable HMML source numbers.

You'll see that in the case of naming individual files, an underscore character (_) is often used instead of a space. For the filenames, this is perfectly fine. For the actual database source entry, the form is to use the project code, a space, and then the five-digit number.

Photograph the Manuscripts

The documents titled “Book photography. pdf” and “foliation&filenames3.pdf” contain the instructional information for manuscript photography. Rather than repeat this, it might help to see a couple of good examples of the final result.



Manuscripts can be photographed as single pages or two-page spreads, depending on whether the book can easily lie flat without straining the binding. Some aspects of these photographs:

- Images are nicely framed. The book page takes up most of the available frame space without being cut off.
- Non-distracting background. HMML supplies a black fabric base as part of the book cradle system. This provides an attractive background for the book.
- The grayscale indicates proper color balance and exposure. The gray patch in the center of the scale is within tolerances specified in other “how to” documents. The white patch doesn't exceed 240 on the RGB color code scale. Notice that the grayscale is positioned to be level with the surface of the book photographed, receiving illumination from both copy lights. In this way, the grayscale gets exactly the same lighting as the page.

- Both photographs have an “up” arrow positioned properly to allow people to know which way is up even if they are looking at the images at a reduced size or are unfamiliar with the script used in the manuscript.
- A small piece of paper with the HMML source number (described earlier) is included in the image. This way, if an image file is separated from the others in a given folder (or has its name changed), it is still possible to figure out where it came from.
- Clear plastic tabs are used to hold down the manuscript pages in the image on the right. The tabs allow the viewer to see the page contents underneath. They are unobtrusive and easy to retouch out of the image if need be.

One important note: Many books can and should be photographed as two-page spreads. This saves time and makes for less work afterwards. One thing for the camera operator to keep in mind is that if he or she photographs a book oriented “right side up” as they look at it while photographing it, the resulting images will actually be upside down when they are viewed from the computer! This is because the camera is usually oriented the opposite of how the technician views the book on the photography stage. The solution is to photograph two-page spreads “upside down;” this will produce images on disk that read correctly and won’t require rotation afterwards.

Foliate, Rotate and Rename Files, Record Images to Disk

In general, the technician operating the camera spends his or her time photographing pages in accordance with the instructional documents mentioned. When a number of books have been imaged, the images are usually transferred to the second PC computer for these further operations. This often is accomplished by simply transferring an external hard disk drive from one computer to the other. This is faster than sending the files over a network, and frees up the camera computer for more photography.

On the second computer, a number of actions are performed on the image files. Among them:

- RAW and JPEG images are separated and placed in different folders.
- RAW and JPEG files are renamed so their filenames give an indication of their position in the book. This way, the files will list out in the proper order.
- The JPEG images are rotated into proper orientation if need be. Two-page spreads shouldn’t need rotation if they were photographed correctly. Books photographed as recto and verso pages are easy to deal with if the recto pages are sent to one folder and the verso pages sent to another. Batch rotation can be used to turn the images right side up. The RAW image files generally aren’t rotated, as they aren’t immediately used for display purposes.
- DVD disks are made for the use of the holding library. From here, the holding institution can decide whether to also copy the files to a central server, or back them up in any other way.
- Images are copied to an external hard disk drive (supplied by HMML) to send to HMML.

To understand these activities, read the following documents:

<http://www.hmml.org/wtorborg/assets/foliation&filenames3.pdf>

<http://www.hmml.org/wtorborg/assets/CD%20and%20DVD%20Recording.pdf>

In the document, “foliation&filenames3,” pay particular attention to the folder structure HMML specifies for storing the RAW and JPEG images created. HMML will be copying the JPEG images to a large server array when the data arrives here; the structure outlined will make this efficient.

Send Finished Images to HMML

HMML is now using large-capacity external disk drives as the means for projects to send the finished manuscript images to HMML. Before sending images, take the time to make sure of the following:

- Check that the folder structure on the hard disk is in line with the recommendations in the document titled “foliation&filenames3.pdf.”
- The images in the JPEG folders are upright
- The image files sort by name in the proper order (representative of the structure of the manuscript itself).
- Each manuscript folder has an image file of the metadata sheet, along with photos of the spine, covers, and edges (the “additional” pictures, as described in the foliation and filenames document).
- Electronic metadata exists for the manuscript, either in the “simple” Microsoft Access database or in a more comprehensive Excel worksheet. HMML would prefer that an updated database file (or Excel workbooks) accompany the image files when sent back to HMML.
- It's a good idea to scan external hard drives for malware, computer viruses, etc., prior to sending them to HMML.