

Digitization of the Walt Whitman Manuscripts

Project Documentation

Compiled by Tina Kirkham and Michael Adamo

August 31, 2005 (*Revised 14 September 2005*)

Introduction.....	2
Personnel	2
Timeline.....	2
Challenges	3
Source and scope	4
Conservation measures.....	5
Imaging.....	5
Goals for digital capture	5
Specifications.....	6
Equipment.....	6
Book cradle	6
Metadata and organization	6
Quality control	7
Anomalies.....	8
Delivery	8
Appendices.....	9
Appendix A: Poems	9
Appendix B: Prose	9
Appendix C: Proofs	9
Appendix D: File naming scheme	9
Links.....	9

Introduction

In the Fall of 2004, the [Digital Production Center](#) (DPC) in Perkins Library was charged with digitizing the Trent Collection of Whitmaniana Writing Series. The project emerged from an agreement between Kenneth Price of [The Walt Whitman Archive](#) and [Duke University Libraries](#), in particular Paul Conway, former Director of Information Technology Services in Perkins Library, and Robert Byrd, Head of [Rare Books, Manuscripts, and Special Collections Library](#) (RBMSCL).

The Writing Series comprises approximately 1000 handwritten notes, drafts, and commentary made by Whitman during his lifetime. It includes jottings and outlines for *Leaves of Grass* as well as early drafts and portions of other major works. Also included are clippings and proofs with marginalia by Whitman.

Newly staffed and equipped, the DPC started production on the project in early April 2005. The resulting images were then provided to the staff of The Walt Whitman Archive for transcribing, encoding, and display on their web site at www.whitmanarchive.org. This document describes the digitization process.

Personnel

The Whitman Archive (University of Nebraska at Lincoln team)

Brett Barney	Project Manager
Kenneth M. Price	Co-Director
Amanda Gailey	Editorial Assistant
Stacey Provan	Graduate Student

Duke University Libraries team

Michael Adamo	Digitization Specialist (Photographer), DPC
Beth Doyle	Collections Conservator
Tina Kirkham	Manager, DPC
Linda McCurdy	Head, RBMSCL Research Services

Timeline

November 2004	Initial conversations. A cursory review of source materials is made.
December 2004	The Trent Collection finding aid is used to construct a preliminary item by item inventory for each subseries.
January 2005	Each item is pulled and examined. The number of digital images to be produced is estimated. Rough measurements for each item are made. Each inventory is expanded into a digitization guide, including call number, main entry or first line, size, format (bound or unbound), and estimated number of documents (i.e. pages) in each item. A preliminary file naming scheme is devised.

February 2005	When an item appears to have no corresponding physical volume or folder, archivist Sam Hammond is consulted and corrections to call numbers are made in the digitization guide where necessary. <i>Catalogue of the Whitman Collection in the Duke University Library</i> , (Durham, NC: 1945), compiled by Ellen Frances Frey, is consulted to clarify evident anomalies in the finding aid. The file naming scheme is approved.
March 2005	DPC's scanback is installed. A book cradle is built and approved. A procedure for the safe transfer of source materials from RBMSCL to the DPC is developed. The digitization guides are converted to Excel spreadsheets. Unique numbers to each item to be scanned using the file naming scheme.
April 2005	Scanning begins. As pages are counted and versos examined, each spreadsheet is expanded to include a discrete entry (row) for each digital image to be created. Sizes for each document are entered in the spreadsheets.
June 2005	Quality assurance on the completed images starts.
August 2005	Scanning concludes. Quality assurance is completed. Documentation is finalized. Images are delivered.

Challenges

When preparing to scan documents, common practice is to sort originals so that like sizes are scanned together. This practice requires fewer camera and focus adjustments than a series of dissimilar items, and thus, tends to be more efficient.

The Whitman manuscripts presented an unusual challenge. Nearly 90% of these documents are housed in bound volumes (octavos, quartos, and folios). Each volume is equivalent to an *item* in the finding aids and in our naming scheme. Prior to scanning, an initial pass through the material enabled us to sort the items by approximate size. Although the documents within a particular volume obviously fall within those size ranges, in many cases a single volume housed documents of multiple sizes and shapes. Thus, as the photographer moved through the documents in a particular volume, frequent refocusing was required in order to accommodate the unique size, shape, and visual characteristics of the documents it housed. The smallest single document in the project was 1 X 3 inches; the largest was 17 X 11 inches.

Initially, for the Poems subseries unbound items were separated and scanned first (laid flat on the copy table). Then, all bound items in the subseries were scanned (using a book cradle). This strategy introduced extra complications into the workflow and provided few benefits, so it was quickly abandoned. For the Prose subseries, bound and unbound items were scanned in the order in which they appeared in the spreadsheet, sorted by approximate size. (All items in the Proofs subseries are unbound.)

Average speed for imaging through the project approached eight minutes per scan, not including image adjustments, quality checks, and save and upload time. This relative slowness was due to a combination of factors: special handling requirements; the need to use a book cradle; the unpredictable variations in size and shape of the documents; and the challenging organization of the physical collection and resulting overhead to keep track of images and other data.

Source and scope

The Trent Collection Writing Series encompasses the following subseries:

- Manuscript Poems, ca. 1855 and n.d.
- Manuscript Prose, 1852-1891 and n.d.
- Proofs, 1874-1891 and n.d.

In total, 1027 digital images were produced from 194 items (folders or volumes). The table below shows image counts per subseries and the corresponding file names.

<i>Subseries</i>	<i>Number of items</i>	<i>Number of images</i>	<i>File ranges</i>
Poems	55	410	1001_010 to 1056_011
Prose	117	445	2001_010 to 2118_060
Proofs	22	172	3001_010 to 3022_010
Total	194	1027	

The DPC agreed to deliver uncompressed, uncropped TIF images. Because the Archive planned to create working TIFs according to their standards, JPG derivatives were not created.

Insofar as possible without damaging the source material, all recto sides were scanned. Versos were scanned when the presence of marks, embossing, printing, or any other visual data was detected. Interestingly, we found it difficult to make this judgment in some cases.

As discussed with Ken Price, we scanned all rectos in the *November Boughs* proof, even if no editorial marks were present on the page. In the context of the few pages that do have markings, the uncorrected pages are potentially meaningful to Whitman scholars.

In collaboration with Conservator Beth Doyle, attempts were made to shoot underwriting, pasted (or partially pasted) scraps, newspaper clippings, and ephemera (e.g. a lock of Whitman's hair). Photos and drawings encountered as we moved through the material were also scanned if we could not readily locate other, high-quality digital reproductions of them.

Conservation measures

The majority of these documents are one-of-a-kind. Whitman's stature in American letters combined with the uniqueness of the documents ensured that first priority would be given to protecting the source materials.

The excellent condition and value of the 20th-century bindings in which most of the manuscripts are housed eliminated disbinding as an option. It was determined that the volumes could be safely opened to 90 degrees. The Buhl lights (see "Equipment" in the **Imaging** section), which throw little UV or heat on the objects, were also approved. Safe handling procedures were agreed upon, including: use only the approved book cradle in the approved manner; no glass should be placed over source materials; no gloves will be used when handling source materials; approved weighted "snakes" or mylar tape are acceptable for holding pages in place.

Prior to scanning, every document in every volume or folder was examined by Beth Doyle to determine the feasibility of digital capture. During her review, she requested that seven items be sent to Preservation for repairs following scanning. Through her assistance and guidance, including hands-on support for some scans, all 194 items in the Writing Series were digitized. In two cases, she devised a method for safely detaching a pasted corner to allow the verso side of the page to be captured.

Imaging

Goals for digital capture

The bulk of the Trent Collection has never been scanned. In conversations with the Whitman Archive team, we agreed to capture the entire object, i.e. we would include the edges of the document, not merely its textual content. In addition, we adopted the "scan once" methodology as described in the [North Carolina Exploring Cultural Heritage Online guidelines](#), with the intention of achieving the highest possible information capture that our equipment and the practical limits of file size allowed. These master images are unmanipulated images created at high resolution and stored in an uncompressed format (TIF). These "production masters" are intended to serve as digital reproductions of the originals, able to withstand close examination by scholars and supplanting to some extent the need to examine the analog objects.

This approach eschews overt image adjustment or color manipulation. Digitization Specialist Michael Adamo chose camera settings that most faithfully captured the visual traits of the source item. In ViewFinder™ (the BetterLight imaging software) he customized the reproduction curve to accommodate the unique traits of our camera/scanback combination.

Prior to scanning, the camera was focused for each image, and the photographer confirmed in the imaging software that the black and white values were within the acceptable ranges (see "Color fidelity" in the **Quality control** section). After scanning, no adjustments or alterations were made to the images except, in rare cases,

slight rotations to square a crooked image in the frame. Since any image adjustment involves the loss of visual data, we attempted to retain insofar as possible 100% of the image data. This strategy enables the greatest possible range of future uses for the image.

Specifications

All images were captured at 600 PPI in 24-bit color. The sRGB color profile was embedded. No minimum or maximum file size was set, and the resulting sizes covered a broad range. The smallest image file is 28.5 MB, and the largest is 238.7 MB. The average file size is 63 MB.

Equipment

The materials were scanned from above using a 4 X 5 Linhoff camera and a Rodenstock 125 mm lens mounted on a Kaiser 30 X 40-inch reprographic station. The BetterLight Super8K HS scanback was connected via USB to an Apple PowerMac G5. The source materials were lit by four freestanding Buhl 150-watt Softcube 4200-Kelvin lights. Completed images were transferred to a 1.2 TB Apple xServer for long-term storage, and copied to a secure web server for transfer to The Whitman Archive team.

Book cradle

Although the Writing Series comprises hundreds of discrete documents, approximately 85% of the items are hardbound in narrow, sturdy volumes. This necessitated the use of a cradle to position each volume under the camera. In March 2005, a functional book cradle was improvised and approved for use by the Head of Preservation, Winston Atkins, and the Head of Research Services, Linda McCurdy. The cradle's design is detailed [here](#).

Metadata and organization

Our use of the term "item" refers to the unit corresponding to a single call number in the [Trent Collection finding aid](#). In this collection, an item refers to a folder containing unbound documents, a bound volume housing one or more documents, or a box housing all proof pages for a particular work, for example, *November Boughs*. In the Trent Collection, these items are the smallest units of organization, an intellectual structure that is reflected in both the Trent Collection finding aid and in The Whitman Archive's [enhanced finding aid](#). Although item-level processing has of course been done on the collection, the finding aid does not describe each document individually. As a result, in addition to assigning a unique item number, our naming scheme also required us to indicate which page/side an image corresponds with. Thus, we counted the first document in each item as a page. For example, the item *In Paths Untrodden* has the call number *MS 4to 195*. We assigned a unique number (1002) to this item. It contains one page with recto and verso sides. Thus, the resulting digital images are numbered 1002_010 and 1002_011. (Per our discussions with Ken Price, if more than one item had been affixed to a single page, we shot the page as a whole.)

Quality control

Acceptable focus

An adjustment to focus was required when changing from one size document to another and whenever a significant change occurred in the plane of the object to be scanned, for instance, when moving from a bound to an unbound item or when moving through the pages of a bound volume. (In this project, refocusing was required generally every fourth page in a volume.) In addition, changes in page curvature often forced an adjustment to aperture. The photographer encountered a few cases in which a series of documents to be scanned shared like characteristics and was thus completed relatively quickly, for example, the unbound proof of *November Boughs*. However, for the bulk of the project, frequent refocusing of the camera was required.

There are instances where the focus might be a slightly soft in a corner or in an area with no information, but this fall off is typically only noticeable when viewing the image at 100%. Focus was limited to the documents. As a result, some of the black backgrounds may appear soft.

Color fidelity

No manipulation of the color or other visual attributes was performed. A Q13 color target was included in each image. The target acts as a visual sample with known characteristics. It provides information about the tonality and scale of the image, and it functions as a standard against which the image can be compared, both by this project team and by future users of the image. The white and black numerical values in each target were verified to ensure that the specified range of values was achieved. After tests of the material using the BetterLight scanback, the acceptable range for black was established at 10-25; for white, 235-245.

Cropping and deskewing

In general, the pages were rectangular and cropping was not problematic. When pages were not square, one of two methods was used. (1) The image was oriented so that the text was square to the bottom edge, or (2) if the page was odd-shaped, it was oriented in order to achieve a balance between the text orientation and the page borders.

In some cases, the strap used to hold the book on the cradle is visible in the image. Due to preservation concerns and the nature of the cradle, this was at times unavoidable. Judicious cropping when making the working TIF will eliminate the strap in most cases.

Digital dusting

To enable accurate future readings of their values, the color targets in the TIFs were digitally “cleaned” using the *Rubber Stamp* feature in Photoshop CS. The texture of the velvet covering the cradle attracted dust and caught the light in a way as to appear dusty. Because most of these black backgrounds will not appear in the working TIFs, we did not attempt to digitally dust them.

QC pass

Following initial scanning, Digitization Specialist Michael Adamo performed a quality check on all images. He checked for edge-to-edge focus and sharpness; confirmed the correct profile; deskewed if necessary; verified white and black target values; checked for border uniformity; and looked for dropped pixels, banding, and blocking.

By comparing each image to its corresponding entry in the spreadsheets, the Digitization Specialist confirmed the accuracy of the file names and flagged anomalies in file sequencing and recto/verso correspondences. In some cases, the physical item was reexamined to ensure accuracy in file naming. In addition, the files on the server were compared to those in the spreadsheets and production database to ensure that the sequence was uniform and that all files adhered to the naming scheme.

Rescans

Four items required rescanning because the initial scan did not achieve acceptable focus or because a portion of the target was inadvertently cropped.

Anomalies

Known gaps in file sequence

Due to an early error in assigning file names (which we chose not to correct), there are no items 1014 or 2085. We opted to allow these gaps in the file sequence rather than risk introducing new errors by renumbering all succeeding image files.

Verso with no recto

Because of the way it was bound, item 1017 has a verso side but no corresponding recto. Thus, the file sequence for this item begins with 1017_011.tif.

November Boughs file names

Some image files for the *November Boughs* proof (item 3002) deviate from the naming scheme. This item contained more than 99 pages. In order to assign file names for the images of pages 100-136, we made use of the placeholder in the naming scheme, replacing the underscore with the first digit in the page number. Thus, in the file sequence, the *November Boughs* file 3002_990.tif is followed by the non-*November Boughs* files 3004 through 3022. The remaining *November Boughs* files resume with 30021000.tif.

Delivery

In total, the project comprised approximately 70 GB of data, making the prospect of delivery via CDs daunting. As a result, Perkins Library IT staff set up a secure web server with a Novell NetStorage interface to provide remote access for the Archive team. Because the maximum data that a WinZip file can contain is 4 GB, the image files were grouped into 3.9-GB batches and zipped. Three ZIP files at a time were placed on the server for downloading. Delivery commenced on August 22 and is expected to conclude on or about August 31, 2005.

Appendices

Note: In the spreadsheets (Appendices A-C), the entries in the Size column are not consistent with regard to whether the long or short dimension appears first.

Appendix A: Poems

See *AppendixA_Poems.xls*.

Appendix B: Prose

See *AppendixB_Prose.xls*.

Appendix C: Proofs

See *AppendixC_Proofs.xls*.

Appendix D: File naming scheme

See *AppendixD_FileNames.doc*.

Links

The Walt Whitman Archive

<http://www.whitmanarchive.org/introduction/>

Duke University Libraries

<http://library.duke.edu/>

Rare Books, Manuscripts, and Special Collections Library

<http://scriptorium.lib.duke.edu/>

Digital Production Center

<http://www.lib.duke.edu/its/dpc/index.html>

Trent Collection finding aid

http://scriptorium.lib.duke.edu/dynaweb/findaids/whitmaniana/@Generic_BookView?DwebQuery=Trent

North Carolina Exploring Cultural Heritage Online guidelines

<http://www.ncecho.org/Guide/production.htm#4.3>

DPC book cradle

<http://www.lib.duke.edu/its/dpc/cradle/instructions.html>

Whitman Archive enhanced finding aid

http://www.whitmanarchive.org/manuscripts/manuscriptsframeset_files/findaidsindex.html