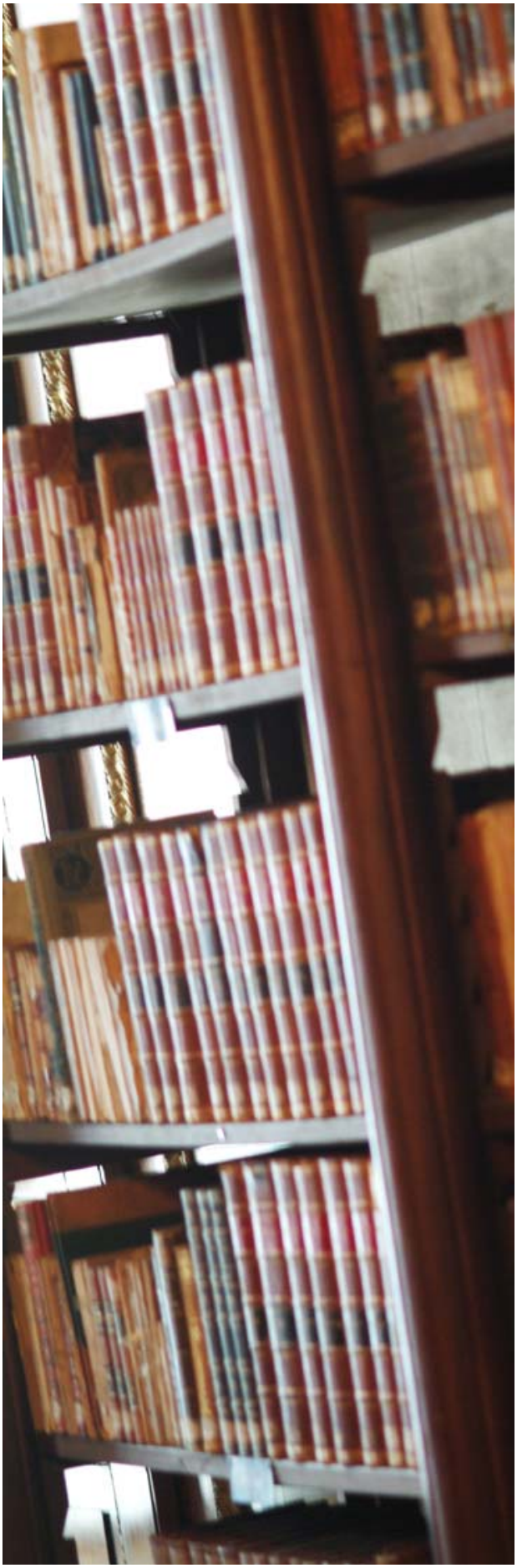


Internet Archive
Book Digitization Process

Robert Miller, Director of Books,
Internet Archive

Robert@archive.org
00-1-415-640-1092

May 29 , 2010 – Rev. M

Prepared by the
Internet Archive Operations Team

Table of Contents

APPENDIX

Introduction to Internet Archive

Background

Before settling upon the current operations workflow, engineers at Internet Archive (IA) tested and evaluated several of commercially available book-digitizing devices. After carefully reviewing the condition, variety and library requirements for materials to be digitized, it was decided that developing an in-house digitizing process, complete with equipment and software would provide an optimal balance of quality and efficiency while also ensuring that the materials being digitized would not be damaged. The IA-built equipment, software and process was reviewed with library preservation experts, field-tested and has subsequently been used in IA-run Scanning Centers around the world. To date, over 200 million pages have been digitized using IA's non-destructive digitization method.

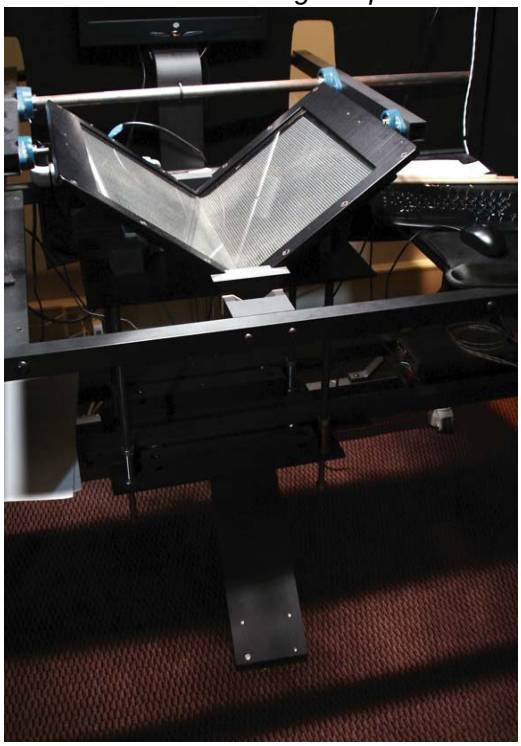
Periodically, new digitization equipment is reviewed to ensure that the most optimal software and hardware in use. The workflow is also reviewed frequently in order to incorporate the discovery of new "best practices". All Scanning Centers follow the same general workflow described in this document, with any minor deviations being based upon specific library requirements. With only a few exceptions, Scanning Centers are managed and staffed by trained IA personnel.

Non-Destructive Scanning Station

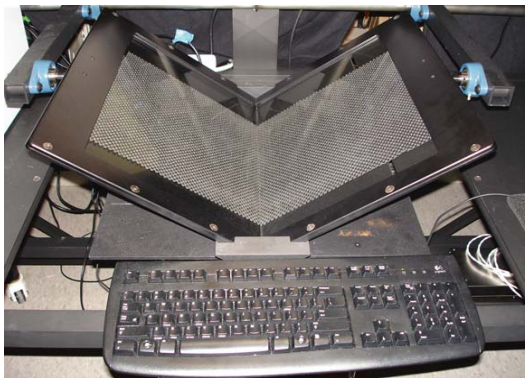
The Scribe workstation is comprised of a sturdy aluminum frame that supports two adjustable camera mounting rails, two color cameras that capture both recto and verso pages of each book, a floating V-shaped book-cradle whose angled design minimizes stress placed on materials, a glass platen that is raised and lowered by means of a foot pedal, two banks of museum qrade lights that illuminate the book and one computer that captures the color images from the camera and performs some of the pre-processing. Once the book is digitized and an on-site Quality Assurance process is completed, the captured images are uploaded via RSYNC to processing computers located in California.

Scribe Scanning Station

Book cradle and foot pedal which raises/lowers the glass platen



Book Cradle



1. Technical Details and Specifications

A. Mechanical, Electrical and Gross size dimensions

i. 7 amperes per Scribe, 800 watts of heat generated per Scribe (standard UK/USA voltage).
ii. Approximately 100 cubic feet of work area per Scribe.
iii. Scribe footprint when installed : 68" long x 37" wide x 79" high (172cm x 93cm x 200cm).
iv. Recommended door width – 34.5" (87.6 cm)
v. Width dimensions for delivery through doorways/into work area (when crate is removed): 60" long (152cm) x 33.5 " (85.1 cm) with only monitor and bracket and foot pedal holder on back side removed (33.5" (85.1 cm). If lifter arm has to be removed (not recommended and to only be done by IA staff) and the monitor and bracket and foot pedal holder are removed – 32" wide (81 cm)
vi. Dimensions for shipping (with reusable crate): 76" long x 38" wide x 88" high (193cm x 96cm x 224cm). IMPORTANT; for truck ensure that it has a lift gate and a door OPENING of 90 inches for clearance.
vii. Shipping Weight with crate: appx 668 pounds (276 kilos).

B. Output Formats

(Please see Section 4, Part A, page 12 for a more detailed list of output files).

i. Color images in JPEG2000 format in pixels per inch listed below.
ii. Optical Character Recognition (OCR) in two XML formats: ABBYY and DjVu. (ABBYY 8.0 is currently used). As new versions and alternative vendors become available, a review will be conducted. If a new version or vendor is deemed as good as or better than the existing version or vendor it will be implemented. OCR XML character format is UTF-8.
iii. XML for metadata from MARC.
iv. XML for operational metadata collected during digitization.
v. Searchable PDF.
vi. XML structural metadata for monographs includes: pagination (when page numbers are printed on the book leaves), front/back cover, title page, copyright page and tissue paper (if found in the book.)
vii. The formats listed above will be delivered from the Internet Archive servers to the Internet via HTTP. Testing at the Library of Congress suggests that downloading an entire book file from the IA site takes approximately 90 seconds.
viii. The Library Partner may download as many copies of each file as they wish.

C. PPI and Size (for Canon 5D's Mark 1)

i. Chart of the PPIs used with different sized images. These are chosen to optimally capture a given size book.

PPI	Height (inch)	Width (inch)	Height (cm)	Width (cm)
300	16	9.25	40	23
400	10.75	6.75	27	17
500	8.75	5.5	22	14

D. Image Capture

i. The Scribe machine currently captures page images with two digital single lens-reflex (DSLR) cameras, specifically the Canon model 5D/12.8 mega-pixel camera (http://is.gd/IuQc) and the Canon EF 100mm f 2.8-macro lens (http://is.gd/IuRG). IA may in the future evaluate and test newer camera models as they come onto the market to determine if they will provide similar or better performance.

Digital cameras shoot from above and capture both verso and recto simultaneously

ii. The lighting system used for book illumination consists of eight (8) 5000 Kelvin, 36 degree, and 35-watt museum-grade Solux bulbs and provides a smooth daylight spectrum with a high color-rendering index. If future alternative lighting methods are found to provide similar or improved results, changes to the lighting system may result.

iii. Please note that since there are two independent cameras in use**,** there may be a detectable difference in lighting between the recto and verso images.

iv. Reference targets: a color target (ColorChecker 24) and a white card are shot at the end of each book as reference tools and may be used for ICC-based color management.

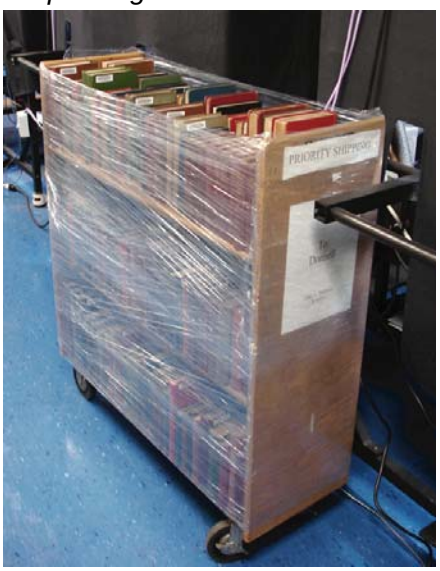
v. Image transfer: images are downloaded in real time to a Scribe management/image-processing computer. This computer is also responsible for running the camera management software that operates the camera shutters.

E. Equipment Calibration

i. Scribe station frames are calibrated and aligned before being put into use.

ii. Cameras are calibrated per manufacturer's specifications. Cameras that no longer perform within specifications are immediately sent to the manufacturer for repair.
iii. Kelvin light bulbs used in the digitization process are replaced as necessary. Lights are allowed to stabilize for 15 minutes before image capture.

Books are typically brought to the scanning center in packing tubs or book carts:



2. Project Setup and Shipment

A. Project Preparation and Special Handling Requirements

i. Prior to beginning a new digitization project, several forms including a Partner Form and WonderFetch-style Pick List must be completed and reviewed with IA staff. Please refer to www.archive.org/details/partnerdocs for the most current documents related to setting up a new scanning relationship between the Internet Archive and a Library Partner. Based on the volume of digitization to be undertaken, a Z39.50 connection may be established between the Library Partner and IA. If it is deemed unnecessary, IA will use the Library of Congress catalog or similar catalog to locate a corresponding MARC record. Correspondingly the Wonderfetch sheet (found at www.archive.org/details/partnerdocs) may be used to create library partner initiated meta data. Contact Robert@archive.org or the Site Coordinator for details.
ii. IA and the Library Partner will review the materials chosen for digitization and determine what can and cannot be included in the process. Questionable materials may be tested before being added to the digitization workflow. Selection of all materials will follow this guideline.
iii. Library preservation personnel may meet with IA to establish and agree on any special handling of materials to be digitized, as well as how to deal with rejected materials, how to indicate when material cannot be digitized, and similar workflow processes.

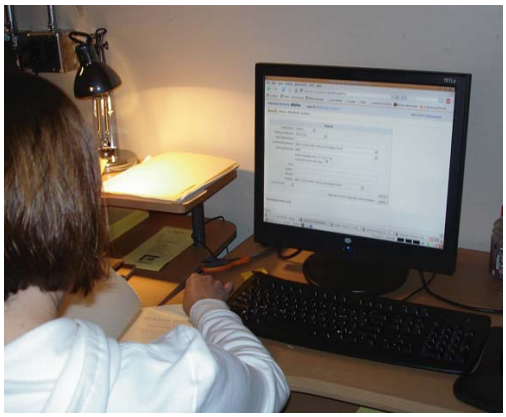
iv. “Special Handling” materials that are fragile, unique or not currently in a library’s general circulation must not be included in the general workflow alongside more robust materials. If special books, books with unique handling needs or the equivalent are to be digitized, IA must be notified beforehand. Deviations from this process will be in writing and, where possible, all steps will be documented with a visual images for reference. Again, it is important that any materials not part of a library’s general collection, or requiring special handling, be identified as such by the Library Partner and be sent in a separate, distinct shipment.

B. Shipment of Materials to the Scanning Center

i. Only materials meeting IA specifications of items to be digitized (See Appendix: Part 1, page 20, Digitization Criteria) will be delivered to the IA Scanning Centers.
ii. Materials will be packed on a library cart, unless another method is agreed upon, and safely wrapped for transport. This is to be done by the Library Partner. Any special procedures will be determined in advance.
iii. A paper copy of the item Pick List (see example in appendix) is to travel with each shipment of materials delivered to a Scanning Center. In addition, a digital copy (Excel) must be sent simultaneously to the Scanning Center Coordinator.
iv. When books are received in the Scanning Center, they are first inspected. Any exceptions such as damage, items that do not fit the digitization profile or similar anomalies, are noted. The scanning coordinator will notify the Library Partner of the materials safe arrival by email unless otherwise agreed to.
v. IA will also ensure that the Pick List item count matches the number of items in a shipment. If a discrepancy with the item count is discovered, the Scanning Center Coordinator will alert the Library Partner immediately and no scanning will occur until this is resolved.

3. Book Digitization Workflow

A. Metadata Retrieval and Book Loading



i. The Scanning Center Book Loader will inspect each item for possible factors that will impact its ability to be digitized (See Appendix: Part 2, Section A, page 22). These items may be digitized at a later date as new processes become available and/or as different cost structures are put into place.
ii. In order to begin loading, a book ID or equivalent (e.g. a bibliographic ID) is loaded into the IA Biblio Tool in order to locate a corresponding MARC record. Concomitant with this, a Wonderfetch pick list may be used to create or load bibliographic data. A quality check is performed to ensure that the item in hand and the MARC record match exactly.
iii. Books with no locatable record will not be digitized and will be returned to the Library Partner pending further information. Any items that we are unable to digitize, both

due to an inability to locate an adequate MARC record or due to book condition issues, will receive an explanatory reject form and be returned to the Library Partner.

iv. If a Scanning Center receives a series of items that have been cataloged within one bibliographic ID but have no discernable set of volume numbering, or are part of a set that has been catalogued under multiple titles, or any other similar cataloging anomaly, IA will work with the Library Partner to determine the best method for uniquely identifying each item. This may include supplementing each record with metadata unique to each item or otherwise creating a volume numbering system for digitization purposes. IA will not delete or add any information to the description fields within the MARC record.

v. IA's Biblio Tool automatically creates a unique, persistent identifier for each item and the MARC record is attached to that identifier.

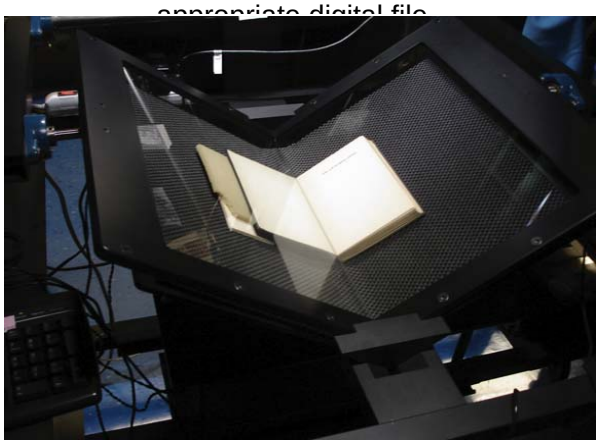
vi. Each book within a collection or project may be given a color-coded flag (to assist in book tracking), which also indicates the book identifier. Any special digitization instructions are included with the book. The book is now ready for digitization and is placed in a queue for the Scribe Operators.

B. Book Digitization Process

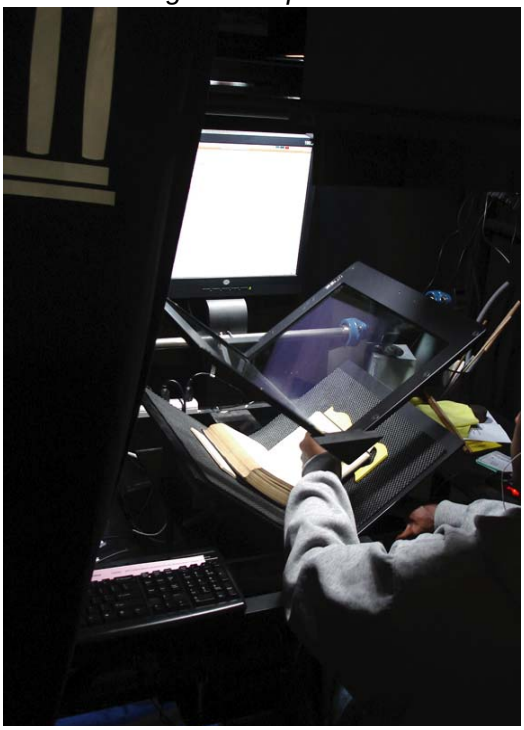
i. Scribe operators compare the color-coded flag containing the book identifier with the actual item in hand. This ensures that the file of digital images they create will be matched up with the correct corresponding metadata.

Books with flags (shown top) books with foldouts or maps are identified (shown bottom)

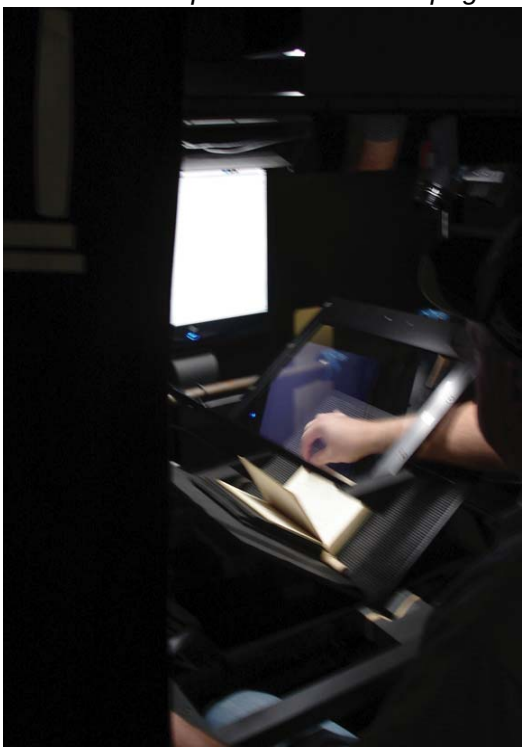
Images are digitized and placed in the appropriate digital file



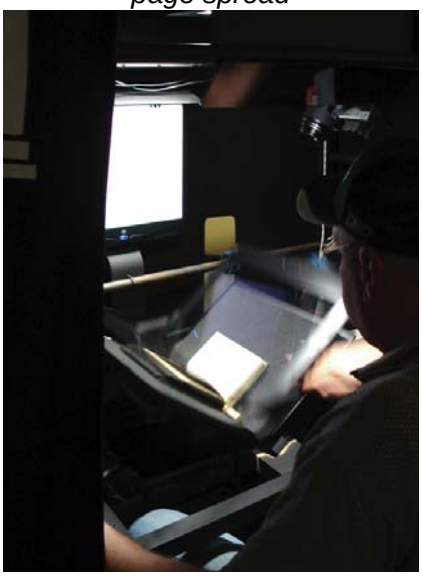
The Scribe operator may reposition the book as necessary during digitization process



The Scribe operator turns each page



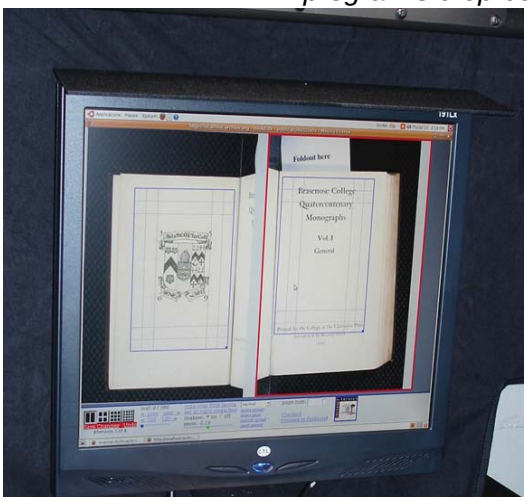
The glass platen is gently lowered onto the book before shooting the next page spread

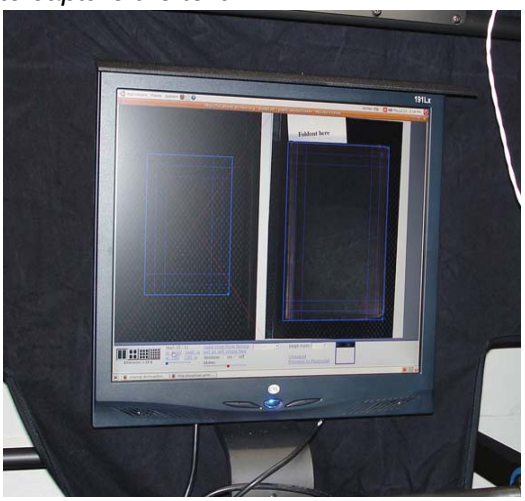


ii. Images are shot through anti-static, anti-glare coated glass to minimize distortion due to page curvature. This also aids in the OCR process.

iii. Images are quality checked throughout the digitization session and are adjusted for such things as text block cropping and image de-skewing in order to ensure proper preservation and presentation. NOTE: IA uses an algorithm to set the skew on a page. It works best on pages where there is a full text box. It does not do as good a job when there is a photo, title page, chart or text block without a lot of text. IA will attempt to manually adjust the skew in the first 10 pages of a book, but will in general not correct skew elsewhere in the book unless prior discussions have happened with the Library Partner.

Once images are captured, the scribe operator adjusts the program's crop box to capture the text





iv. The digital file is inspected for missing pages, crop/de-skew issues and general presentation, as part of the Quality Assurance process.
v. The completed files are then uploaded to IA's Data Center for processing.

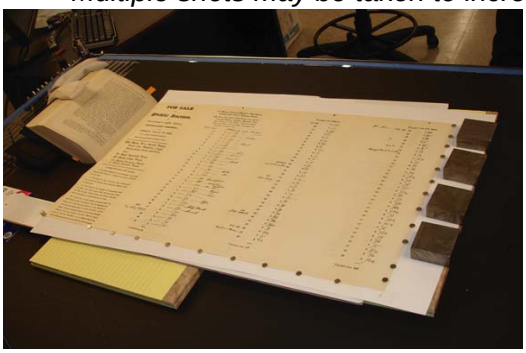
C. Foldouts/Maps

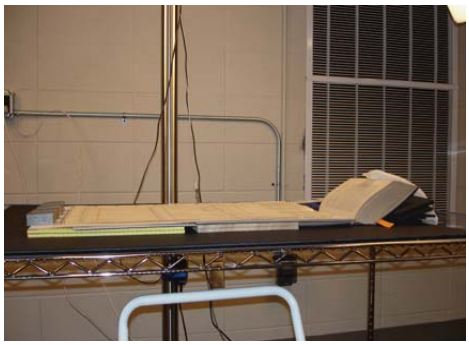
i. Foldouts and maps of appropriate size may be digitized in full color and can be inserted seamlessly into the leaf images of any digitized item.
ii. To preserve acceptable resolution, normal foldouts should be no larger than: 30" wide x 20" high (76cm x 50cm).
iii. For items larger than 30" wide x 20" high (76cm x 50cm), a single, lower resolution reference shot of the entire foldout is taken and is supplemented by a number of higher resolution sectional images. For example, a large map may be shot as a single reference shot PLUS 3 additional shots- 1 of 3, 2 of 3 and 3 of 3 shots.
iv. The maximum size foldout allowable is 41" wide x 27" high (104cm x 70cm). (For this size, the single reference image is shot at a reduced resolution of 107 PPI.
v. NOTE: IA also has the capability to "stitch" a client's image into the leaf images of any digitized item. This image may be virtually any size, as long as it is an uncompressed TIFF image. With this capability, a Library Partner may use their own camera to capture a map or foldout at a very high resolution. IA can then insert this image into the series of images that comprises the digitized book.

Camera Height	PPI	Width & length (up to)
21.3 in. (54.1cm)	248	18" x 12" (46cm x 30cm)
24.2 in. (61.5cm)	208	21" x 14" (53cm x 36cm)
27.4 in. (69.6cm)	180	24" x 16" (70cm x 41cm)
30 in. (76.2cm)	146	30" x 20" (76cm x 50cm)
36 in. (91.4cm)	**121**	**36" x 24"** (91cm x 70cm) - shot in sections as well
41 in (104.1cm)	**107**	**41"x 27"** (104cm x 70cm) - shot in sections as well

Foldouts are digitized using an overhead camera after book digitization at the Scribe

Items are "flattened" using magnets and are positioned to minimize curvature. Multiple shots may be taken to increase the PPI if the image is large.





D. Other capabilities (Please contact IA for more details)

i. Microform - IA can accommodate standard 35mm and 16mm microfilm and standard microfiche. Microfilm digitization is presently performed at the San Francisco, California facility and microfiche digitization is performed at the Alberta, Canada facility. Images are captured as gray scale, turned into JPG 2000 images and then processed as books. Pricing and specs are available upon request.
ii. Folios - IA is able to digitize bound, single-sided, folio-sized books, an example being a large art folio. In order to retain acceptable resolution, the ideal maximum size for a folio would be 12" wide x 18" high (30cm x 46cm) and a thickness of about 150 pages. A test digitization of potential material is recommended. Pricing and examples are available upon request.
iii. Archival Collections - IA is able to accommodate special digitization needs, such as those arising from collections of papers or ephemera. We can work with the special needs of a collection and ask only that the materials are flat and do not contain staples or clasps that must be removed in order to digitize. The Library Partner should perform any collation or organization necessary.
iv. Photographic materials - IA is able to accommodate special digitization needs, such as those arising from yearbooks, photo albums and loose photos.
v. Loose-leaf materials - Mass digitization of single sheet documents is a capability currently under development; please contact IA for details.
vi. Video, Audio and LP conversion - IA is working to accommodate special digitization needs, such as those arising from audio and video materials.

4. Post-Digitization Processing

A. Image Processing

Uploaded images are processed to create storage and access files. (See Section 1, Part B, page 4 for a more general list of output files).

i. Detailed list of files
   1. ID.pdf
   2. ID_jp2.zip
      - zipped folder of the book
      - [ID]_nnnn.jp2
   3. ID_lib_jp2.zip

- zipped folder of the book (note: this folder could be included if a bookplate is applied)
- [ID]_lib_nnnn.jp2
- ID_marc.xml

4. ID_meta.mrc
5. ID_meta.xml
6. ID_metasource.xml
7. ID_raw_jp2.zip - unprocessed storage format (no bookplate)
8. ID_wm_jp2.zip - this is where a bookplate highlighting attribution of sponsor or contributor would be located.
9. Scandata.zip

ii. Metadata will reside in meta.xml file, and will include the following
1. Identifier
2. Identifier/Bib ID (IA identifier and local ID from Pick List)
3. Contributor
4. Title
5. Volume
6. Creator (if in MARC record)
7. Publisher (if in MARC record)
8. Collection/s
9. Operator
10. Scanner
11. Scandate
12. Identifier/Access (URL for accessing this book is found by using www.archive.org/details/identifer)

B. Background Processing

i. Initially the digitized image is captured as a camera raw file (CR2). This is run through a JPG2000 compression to generate a raw JPG2000 for storage. The raw JPG2000 is then turned into a processed master, which is used to generate the access formats.

ii. Storage format – raw JPG2000 is a compressed, lossy, un-cropped, non-rotated, non-de-skewed, non-light compensated JPG2000 file. There is a contrast enhancement process step that is done here. It expands the color-value range of 30-240 to 0-255. This means that the darkest -12% of the total range gets flattened to black and the lightest -6% gets flattened to white, with everything else in-between stretched accordingly. Image sizes vary depending on the complexity of the page, but are typically in the 900 KB range, yielding an approximate compression ratio of 15:1 relative to the camera raw image (CR2 is approximately 15MB/image.)

iii. Processed master – lossy, cropped, rotated, de-skewed, light compensated JPG2000. Image sizes may vary depending on complexity of the page, but are typically in the 800 KB range, yielding an approximate compression ratio of 20:1 relative to the camera raw image (CR2 is approximately 15MB/image).

iv. Access format – the processed JPG2000 masters are compressed in a JPG2000 format, which feeds into the OCR and book generation tools. Image sizes may vary depending on the complexity of the page, but are typically in the 760 KB range, yielding an approximate compression ratio of 120:1 relative to the camera raw image (CR2 is approximately 15MB/image). Also ReadOnline view, PDF and DjVu, both of which are OCRed.

v. Note: all compression ratios might vary based on which version of ABBYY is used and on specific software parameters. These numbers are for reference purposes only.
vi. Quality settings will vary based on vendor tools used. For example, a quality setting of 50 on a scale of 1-100 was used for the LuraTech (or equivalent) PDF compressor. This setting was determined based upon user surveys. As improved software becomes available, vendor selection may change.

C. Book Plate

i. A bookplate may be digitally inserted in the beginning pages of each digitized item. This decision will be made before commencement of digitization. The bookplate algorithm will find a blank image or a near blank image within the first 10 pages of the book. This algorithm is fairly conservative, so if it does not find a sufficiently blank page, no bookplate will be inserted. This is to avoid obscuring text on a page. It is rare that an item does not have at least one blank page within the front matter, so it is rather infrequently that we will dispense with the bookplate insertion altogether.

D. Completion of Book Processing

i. For digitization performed on-site in a partner library, the typical turnaround of a book cart is 72 hours, from arrival to return. A full Scanning Center consisting of 10 Scribes running 2 shifts will complete approximately 1,000 monographs per week. For shipments that will be sent to a regional center, turnaround time will estimated prior to shipping of the materials.
ii. IA's goal is to derive and upload books for web access within 24-48 hours after digitization.
iii. An internal Quality Assurance process is performed inside the Scanning Center after the books are made available online. (The Quality Assurance process is outlined in Section 5,).
iv. Items that pass the Quality Assurance process are ready to be "checked out" and returned to the Library Partner. Return shipping details shall be established prior to the commencement of the digitization process. Any rejected materials are also returned at this time.
v. At month's end, a final curation by IA staff is undertaken and a bill or invoice is issued. Materials will be available online prior to this and may be downloaded, but until a final invoice is issued and the curation occurs, there is a chance that minor changes to metadata or the files may still occur. If the scanning invoice is less than $1000 in a month, a bill may only be sent quarterly.

E. Check-Out Process and Material Return

i. The Scanning Center staff performs a final count of items to ensure all Library Partner material is included.
ii. The Scanning Center staff packs the books onto library book carts, shipping containers or other such receptacles per the guidelines established between the Library Partner and IA.
iii. The Scanning Center Coordinator may generate an updated Pick List that now contains corresponding IA book identifiers and indications of any rejected materials.
iv. The books are then shipped or returned to the Library Partner.

5. Quality Assurance (QA) Process

A. There are Four Major Phases of the QA Process:

i. At the Scribe Station - Before any digitized images are uploaded, the scribe operator will quality-inspect each image. The scribe operator will review the images for missing spreads, crop/de-skew issues, accurate page labeling (i.e. title page, covers, TOC) and will add any notes regarding defects in the book to the file (i.e. missing pages, tight binding, torn pages.)

ii. After the Images are Online - A statistical sampling and Quality Assurance process is conducted within the Scanning Center. IA currently adheres to American National Standard Institute (ANSI) guidelines, as published in ANSI/ASQ z1.4-2003, Table 1, General Level 2 (Details below in Section B). Some elements may change slightly to reflect process improvements.

iii. Before the Invoice is Sent - A monthly curation is performed and an internal random audit is conducted at IA headquarters.

iv. After the Invoice is Sent - Errors brought to IA's attention will be reviewed and dealt with in a timely manner. IA and the Library Partner will decide if any materials need to be re-digitized or if the problems can be resolved within the digital file, post-derive. Rescanning is avoided when possible, as it is usually the most expensive solution.

B. QA Process in the Scanning Centers

Please refer to ANSI/ASQ z1.4-2003, Table 1, General Level 2 which may be viewed at the following URL: http://www.proqc.com/ansi_sampling_tables.php

i. Each day the Scanning Center will review a set of books from the previous day's digitization production. The number of books to be reviewed will depend upon the total number of books in the set.

Books in set	9-15	16-25	26-50	51-90	91-150	151-280
Number to QA	3	5	8	13	20	32

ii. The Scanning Center Coordinator will select a representative sampling, which reflects a broad combination of scanners and machines and conforms to the statistical chart.

iii. The online digital books are then inspected using specific criteria. (See Appendix: Part 2, Section D,). If found, errors or defects are noted and added to an IA tracking form. An automatic scoring is then performed and a "pass/fail" grade is assigned to the lot.

bin	min	max	level	sample	major_accept	major_reject	minor_accept	minor_reject
1	9	15	2	3	0	1	0	1
2	16	25	2	5	0	1	0	1
3	26	50	2	8	0	1	1	2
4	51	90	2	13	1	2	1	2
5	91	150	2	20	1	2	2	3
6	151	280	2	32	2	3	2	3
7	281	1,000,000	2	99	7	10	7	10

iv. Example: If 125 books have been digitized in the previous day's production, bin 5 will be selected and 20 books will be 100% inspected. Based on the truth table above, if there is 1 major error or less and 2 minor errors or less, the lot is passed. If there are 2 or more major defects or 3 or more minor rejects, the lot fails. The following steps are then followed:

1. If a "fail" is generated, the Scanning Center Coordinator will review the errors or defects to ascertain if they were generated from outside the Scanning Center or from within. Engineering is notified of any problems out of the Scanning Centers' control (i.e. a missing access file) and Scribe operators are notified if any problems are direct a result of their work (i.e. missing page spreads).
2. If an error is generated from within the Scanning Center, the Scanning Center Coordinator will follow a pre-determined set of process steps, ultimately culminating in a recommendation to approve the lot, or a portion of the lot, with appropriate corrective actions identified. At this stage the Director of Books or the Quality Assurance Librarian may be involved and would approve a deviation. A corrective action report will be generated for rejected lots. This will be reviewed with engineering and operations management for longer-term solutions or corrective action.

(Please refer to Appendix: Part 2, Section B for a more detailed list of QA codes, their definitions and how they are used)

6. Scanning Center Locations

Materials will be digitized at one of 8 regional Internet Archive Scanning Centers in San Francisco, California; Los Angeles, California; Ft Wayne, Indiana; Boston, Massachusetts; Jersey City, New Jersey; Princeton, New Jersey; Washington, DC and Toronto, Canada. There are also 10 + smaller digitization centers located in California, the District of Columbia, Illinois, North Carolina, Utah, Canada, Scotland, Guatemala and the United Kingdom.

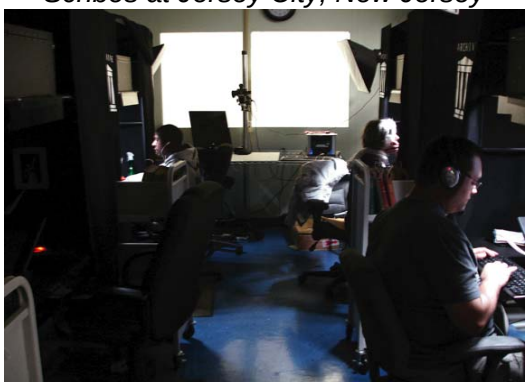
Below are pictures of several of the various centers:

Internet Archive Scanning Center at the Allen County Public Library, Fort Wayne, Indiana

Scribes in action at the Allen County Public Library

Internet Archive Scanning Center near the Jersey City Public Library, Jersey City, New Jersey

Scribes at Jersey City, New Jersey



Internet Archive Scanning Center at the Princeton Theological Seminary, Princeton, New Jersey

Scribes at the Princeton Theological Seminary

Scribes at the University of Toronto Scanning Center

7. Internet Archive Contact Information

Note: Robert Miller is the primary point of contact for initial program review.

Name	Role	Phone	Email
Robert Miller, Director of Books	Project Manager, Engineering Questions, Attribution Form & Pick List	(415) 561-6767 (office) (415) 640-1092 (mobile)	robert@archive.org
Hank Bromley, Engineering	Software Updates, Z39.50 Connection (If approved)		hank@archive.org
Scanning Center Coordinator (See list below)	Daily Operations (will receive and digitize 10 initial test books)		(See list below)
Ken LeTran Systems Engineer	Information Technology	(415) 561-6767	Ken@archive.org
Calvin Quality	Quality Assurance	(415) 561-6767	Calvin@archive.org

Scanning Center Coordinator Contact Info

San Francisco CA	paul.n@archive.org
Los Angeles, CA	KellyCritch@archive.org
Fort Wayne, IN	jeffs@archive.org
Princeton, NJ	judec@archive.org
Boston, MA	melissa@archive.org
Jersey City, NJ	stacy@archive.org
Washington, DC	ronnie@archive.org
Toronto, Ontario	gabe@archive.org

APPENDIX

1. Digitization Criteria

Criteria used to determine if materials may be digitized are listed below and will include, but not necessarily be limited to, the following:

A. Book Condition/Preservation Standards

i. Library preservation personnel will meet with IA to establish and agree on any special handling of materials to be digitized. If special books, rare books or the equivalent are to be digitized, IA must be notified beforehand if special handling processes are to be used. TO REPEAT: IA MUST BE NOTIFIED IF A BOOK IS consider NOT GENERAL CIRCULATION, SPECIAL, FRAGILE OR REQUIRES EXTRA CARE.

ii. Non general circulation materials that are fragile, unique or not currently in library's general circulation must not be included in the general IA digitization workflow. It is important that any materials that require special handling, be identified as such by the Library Partner and be sent in a separate, distinct shipment.

iii. Books with separating covers or that otherwise appear fragile, may be rejected unless a decision to the contrary, prior to digitization, has been made by IA and the Library Partner.

B. Materials with Multiple Titles/Multiple Volumes

i. If a Scanning Center receives a series of items that have been cataloged within one bibliographic ID but have no discernable set of volume numbering, or are part of a set that has been catalogued under multiple titles, or any other similar cataloging anomaly, IA will work with the Library Partner to determine the best method for uniquely identifying each item. This may include supplementing each record with metadata unique to each item or otherwise creating a volume numbering system for digitization purposes. IA will not delete or add any information to the description fields within the MARC record.

C. Size and Shape Requirements

(Materials not fitting the requirements shown below will be returned, un-digitized.)

i. 9.5" wide x 14.5" high (24cm x 36.8cm) maximum, 3" wide x 3" high (7.6cm x 7.6cm) minimum.

ii. Items thicker than 3" (7.6cm) will be reviewed in order to determine if the book cradle can accommodate them and to ensure the glass platen can reach into the gutters.

iii. On average, books should be approximately 200 pages or greater. If a collection consists mainly of items with fewer than 100 pages, IA and the Library Partner will review in order to ensure that the quoted price per digitized page can be maintained.

iv. IA has the capacity to digitize foldouts and maps. (See Section 3, Part C) Testing and decisions about specifications should occur prior to digitization. Foldouts and maps will have a lower resolution, or PPI, than regular image captures. Foldouts and maps have a higher price point than regular image captures.

D. Book Style

i. Ideal candidates for digitization are side-bound monographs - no single sheets and no top-bound books.
ii. Soft cover books are acceptable as long as they are bound.

E. Paper Style/Print Quality

i. Most paper styles may be digitized except highly acidic paper that has already degraded and is disintegrating. Please note that if any archivally fragile paper is to be digitized, IA will conduct a review of the amount of extra time needed for digitizing these materials.
ii. Pages should not be excessively dusty or have excessive mildew or mold.
iii. All pages should have the bolts pre-cut. Unless otherwise instructed, books with bolts (uncut pages) will not be digitized.
iv. Scribe operators should be able to lift and turn the pages with normal effort. Pages with adhesive or that are otherwise tacky will not be digitized.
v. Microfilmed reproductions should be reviewed with the Scanning Center before being digitized. If the resulting text resembles a film negative or photocopy, we will not digitize it. If microfilm reproductions have more than one page on each leaf, these will also be rejected. Any deviation must be agreed to by IA and the Library Partner.

F. Gutters/Margins

i. All books, rebound books in particular, will be reviewed to check for tight gutters and narrow margins along all sides of the text block.
ii. Tight bindings that will not lay open for digitization per IA specification limits may be rejected.
iii. If the page text is less than a quarter inch away from the gutter, the book will not be digitized, unless agreed to with the Library Partner and IA.
iv. Text that runs to the edge of the page or margin may be digitized, but the final presentation may be less than optimal. The decision to digitize material of this nature will be made prior to digitization.

G. Bibliographic Metadata

i. When two or more titles are bound within the same volume (bound-widths) each item will need to be designated with its own pull slip and library identifier (i.e. Bib ID). The Library Partner will prepare this. Each title will be considered a separate book and will be digitized as such. IA and the Library Partner must discuss any deviations from this rule.
ii. Books that are post 1923 may be digitized, but should be discussed with IA and the Library Partner prior to digitization.
iii. If IA is unable to locate an item's MARC record, the book will be returned to the Library Partner pending further information.

2. Rejection/Error Codes and Resolutions

A. Rejection Codes

Any items that we are unable to digitize, due to an inability to locate an adequate MARC record or due to condition issues, may receive an explanatory reject form and be returned to the Library Partner. The following codes may be contained on the form.

Code	Definition
BI	Item is fragile or has no binding
CAT	Item record is unreadable
DAM	Item is damaged
DAT	Item is still in copyright
DUP	Item is an exact duplicate of another item on list
FO	Item has foldouts that are too large for scanning
FOR	Item is not a book
LAN	Item is outside the language parameters
LG	Item is too large for scanning
LINK	Unsuccessful link to metadata
LIST	Pick List error
MAR	Items margins/gutters are too tight for the Scribe
MIS	Item is missing more than five pages
MUL	Item has an unacceptable number of volumes bound together
NA	Not available
NOS	Item is not on shelf (missing/lost)
OUT	Item is not on shelf (checked out)
PAG	Pagination issues (section[s] bound out of order or upside down)
PAP	Item has brittle pages, deemed unsuitable for scanning
SKW	Item has skewed text (to the point of being unreadable)
SM	Item is too small for scanning
SPH	Item requires special handling
UNC	Item has more than five uncut pages (pages must be separated)
VEL	Vellum
WD	Withdrawn

B. Error Codes - Definitions and Resolutions

IA uses a variety of codes to define and describe errors that may be found in digitized items. Error resolution falls into three forms of correction: post-derive, rescanning or a consultation with the Library Partner. If IA is unable to correct problems using any of these methods, then the book is rejected and rescanned if possible. In addition to the thorough in-house QA performed within the Scanning Centers, if the Library Partner discovers any problems with an online item, IA will attempt to resolve the error within 30 days of being notified.

i. FREEZE CODES, part A
Items receiving codes 101-113, 115 and 130-138 will be rescanned using the same identifier and URL. If we are unable to rescan or otherwise correct any problems post-derive, the Library Partner will not be billed for this item.

- Formats
  - 101 Test book
  - 102 DjVu is missing or corrupt
  - 103 PDF is missing or corrupt
  - 104 Flip book is missing or corrupt

Resolution: *Material is re-derived. If the problem cannot be corrected in this manner, the material is rescanned.*

- Uploading or piping problems
  - 110 Truncated file(s)
  - 111 Incomplete Scan
  - 112 Missing files(s)
  - 113 Cr2.tar file is malformed
  - 115 Cameras assigned incorrectly

Resolution: *Material is rescanned.*

- Metadata
  - 123 Possibly not in public domain
  - 124 Removed by request of copyright holder or library

Resolution: *If material is in copyright, the item is removed from the IA search engine. If material is in question, the Library Partner is consulted and appropriate action is taken.*

- Images
  - 130 Cropped text [major]
  - 131 Blurred page(s) [major]
  - 132 Missing page(s) [major] (If a book is inherently missing a page/es, a note will be added to the details page.)
  - 133 Front/back cover missing [major]
  - 134 Overexposure in image obscures text (white streak) [major]
  - 135 Book was digitized twice (Identified copy is darkened and removed from search engine.) [minor]
  - 136 Text is washed out or overly dark (poor lighting) [major]
  - 137 Evidence of scanner (fingers, shadows, etc) visible on page [minor]

- 138 Glass not centered in gutter; text is distorted or cropped [major]

Resolution: *Items receiving codes 130-134 will be rescanned. For items receiving codes 136-138, a decision to accept or rescan will be made by IA*

ii. FREEZE CODES, part B
These codes are used for books that have fixable problems, but are not yet in billable condition.
- 140 Book and metadata do not match [minor]
- 141 Call number is missing or incorrect [minor]
- 142 Tissue pages marked incorrectly [minor]
- 143 Anomaly in image format [major]
- 144 Left/right pages are reversed [major]

Resolution*: For items receiving codes 140-143 a post-derive correction is attempted. For items receiving codes 144, if a post-derive correction doesn't fix the problem, the material is rescanned.*

iii. INFORMATIONAL CODES
- 150 Bibliographic data missing [minor]
- 151 Bookplate or watermark missing or corrupt [minor]
- 152 Copyright evidence reported incorrectly [minor]
- 153 Bibliographic record from library is truncated [minor]
- 154 Possible error in bibliographic record from library [minor]
- 155 Foreign language character encoding is incorrect [minor]
- 160 Light/dark pages (intermittent) [minor]
- 161 Light/dark pages (throughout) [minor]
- 162 Pages skewed [minor]
- 163 Color cards show in access formats [minor]
- 164 White cards show in access formats [minor]
- 166 Image of cradle is visible [minor]
- 167 Different crop-box sizes within same spread [minor]
- 168 Bad crop at page edges/gutter [minor]
- 169 Duplicate page spreads digitized [minor]
- 170 Page types not marked or marked incorrectly [minor]
- 171 Title page not marked [minor] (book does not have title page)
- 172 Scan factors not marked or noted [minor]

Resolution: *For items 150-152, 155, 168, 170 and 172, post-derive correction is possible. Errors 153 and 154 must be reviewed with the Library Partner. For errors 160-162 and 168 a review is conducted to see if the material can be OCRed. If the material can't be OCRed based on current IA software, the book is rejected and re-requested for rescan. Codes 163-167 and 171 are information only and do not affect the readability of the item.*

C. Rescanning Process

i. For materials to be rescanned, a request for the re-delivery of those items is submitted to the Library Partner approximately once a month. Materials are rescanned into the existing URL.

3. Post-Digitization Reporting Tools

All Library Partners have access to the IA Advanced Search Engine, found at: http://www.archive.org/advancedsearch.php. This is a helpful reporting tool that may be used to search and review books that have been digitized, uploaded, QAed and curated.

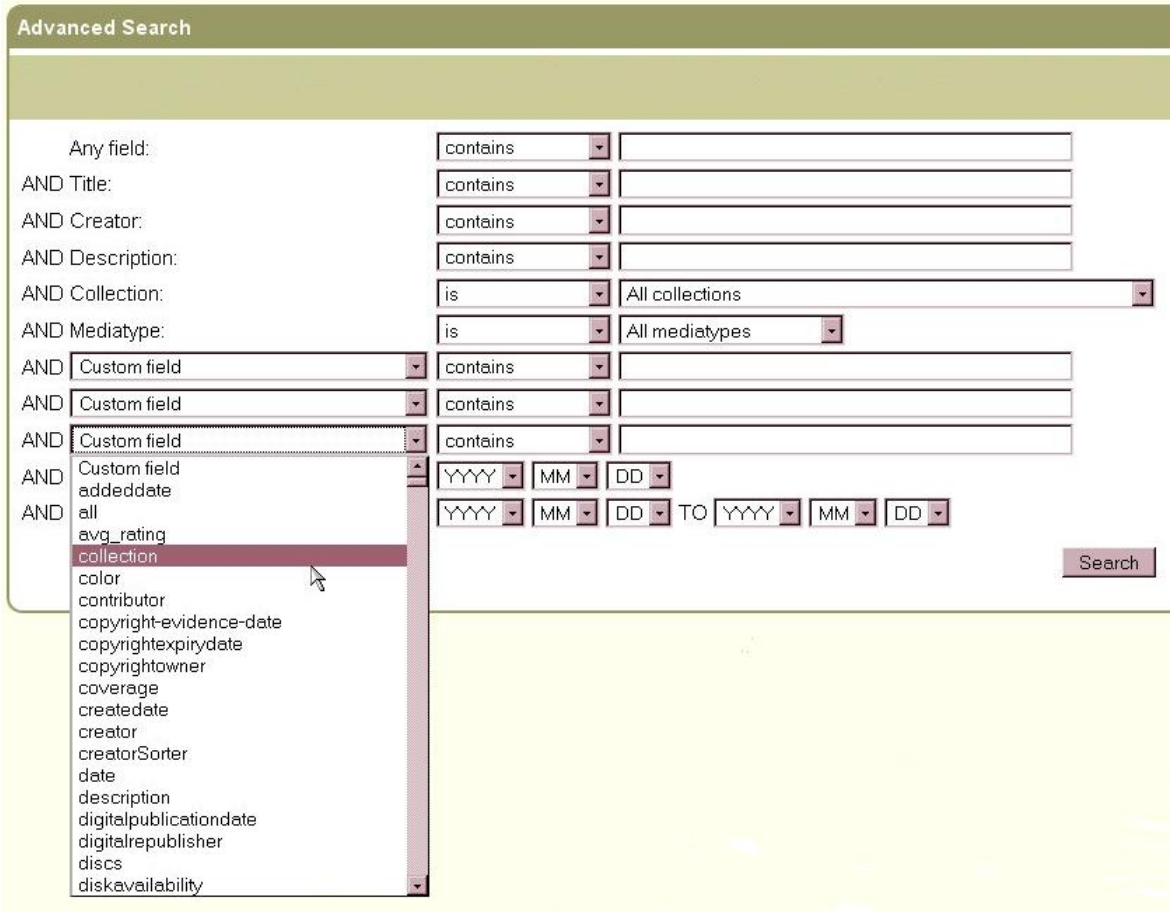
Fields that are viewable in the Advanced Search Engine include:

- title
- creator
- collection
- contributor
- sponsor
- image count
- public date

Note: The curation stage is the last stage in the IA process, during which the books are made viewable online. This may happen on a non-scheduled basis but is typically done several times a month.

A. The Advanced Search Engine

The Advanced Search form is relatively simple to operate. Users only need to enter one search term into any search field in order to produce results. Note: If you select "not" as your match criteria, you must select one other field.



B. The Advanced XML Search

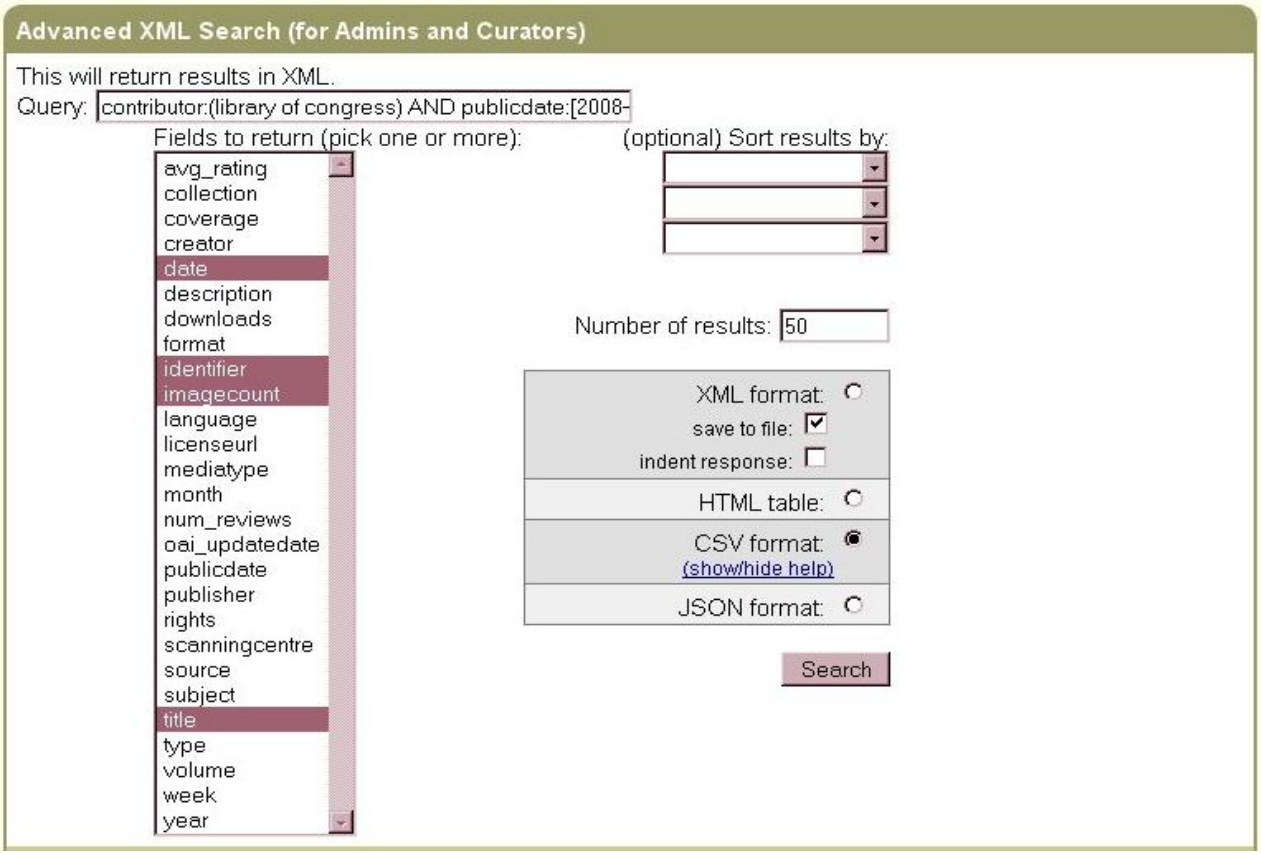
This is used similarly but has some significant differences. This search tool is also located at: http://www.archive.org/advancedsearch.php. An example of how it is used follows. Please refer to the "Help with CSV and Excel" section in the previous URL for tips on searching and information about known bugs.

Example: If a Library Partner wished to see how many pages were digitized in a given month, August 2008 for instance, they would perform the following search:

i. In the Advance XML Search "Query" field, type:
contributor:(library of congress) AND publicdate:[2008-08-01 TO 2008-08-30]

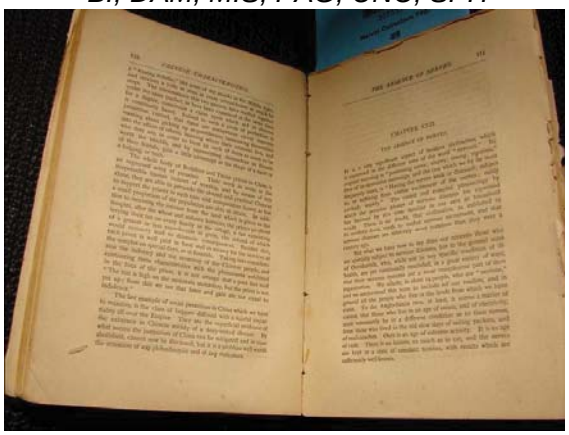
> ***Important:*** *Parenthesis "()" must be around the contributor name.*
> *No spaces after the colon ":".*

ii. Then, by holding down the "Shift" Key, highlight the fields you want to export to Excel. For example: "date, identifier, imagecount and title"
iii. Click the radio button for "CSV" and click "Search". There might be a slight delay as the search is executed. In this query, approximately 2,143 results will be returned.
iv. When prompted, save the ".csv" file to your hard drive and note the location saved. If you open this document in Excel, the data will be parsed into columns so that you may sort or otherwise manipulate the data.

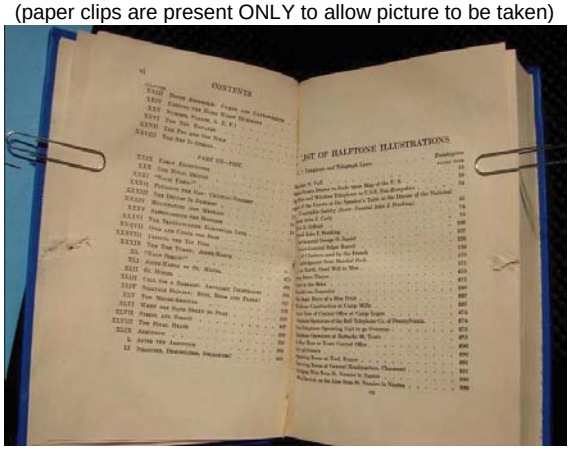


4. Examples of “Problem” Books

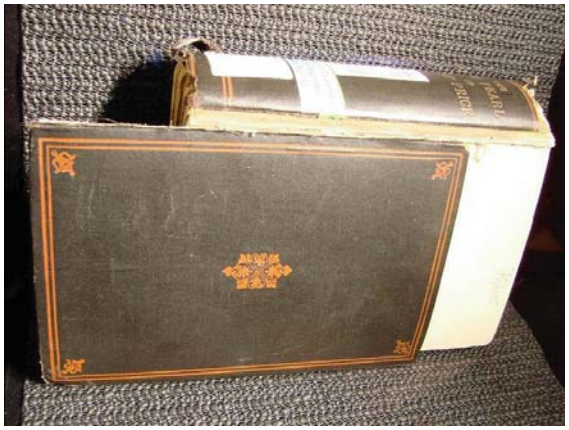
Example with the following problems:
BI, DAM, MIS, PAG, UNC, SPH



Example with the following problems:
Glued pages – DAM, MIS, PA, PAG, SPH
(paper clips are present ONLY to allow picture to be taken)



Example with the following problems:
BI, DAM, SPH



Example with the following problems:
Tight Binding – MAR, PA, PAG, SPH