Internet Archive Book Digitization Workflow and Quality Assurance Process

Prepared by the
Internet Archive Operations team
Robert Miller, Director of Books,
Internet Archive
Robert@archive.org

October 7, 2008 – Rev. F

DRAFT

Table of Contents

Internet Archive Book Digitization Workflow And Quality Assurance Process

Background

Before settling upon the current operations workflow, engineers at Internet Archive (IA) tested and evaluated a wide variety 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 machine would provide the safest and ultimately most cost-effective method by which to digitize materials. The IA-built equipment was reviewed with library preservation experts, field-tested and has subsequently been used in IA-run scanning centers around the world. To date over 100 million pages and over 300,000 books have been digitized using IA's non-destructive scanning method.

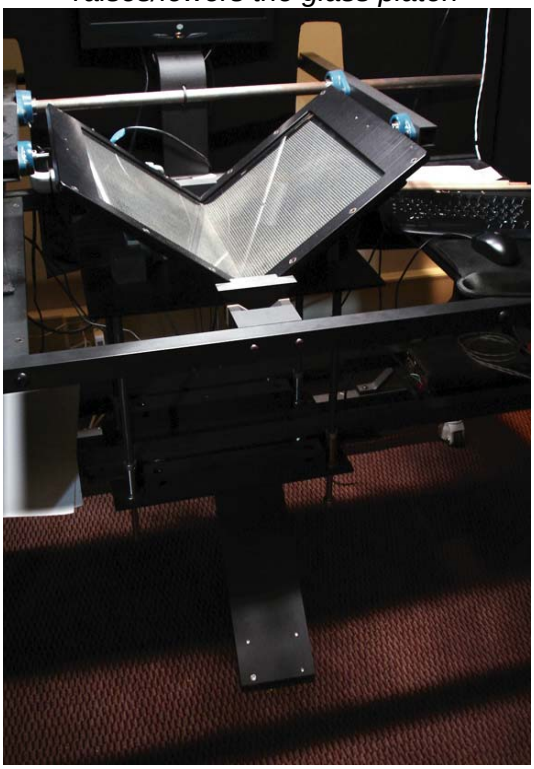
Periodically, new digitization equipment is reviewed to ensure that the most cost-effective and safest process is in use. The workflow is also reviewed frequently in order to incorporate the discovery of new "best practices". All Scanning Centers follow the same general workflow described in this document, with any minor exceptions being based upon specific library requirements. All 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 controlled 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 a 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

Overview: Technical and Operational Details

Output Formats

a. Color images in JPEG2000 format in pixels per inch listed below.
b. 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.
c. XML for metadata from MARC.
d. XML for operational metadata collected during scanning.
e. Searchable PDF.
f. 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.)
g. The formats listed above will be delivered from the Internet Archive servers to the Internet via HTTP, FTP, RSYNC, or OAI. Testing by the Library of Congress suggests that downloading an entire book file takes approximately 90 seconds.

PPI vs. Size

a. Example of PPI vs. size, chosen to optimally image a given size book.

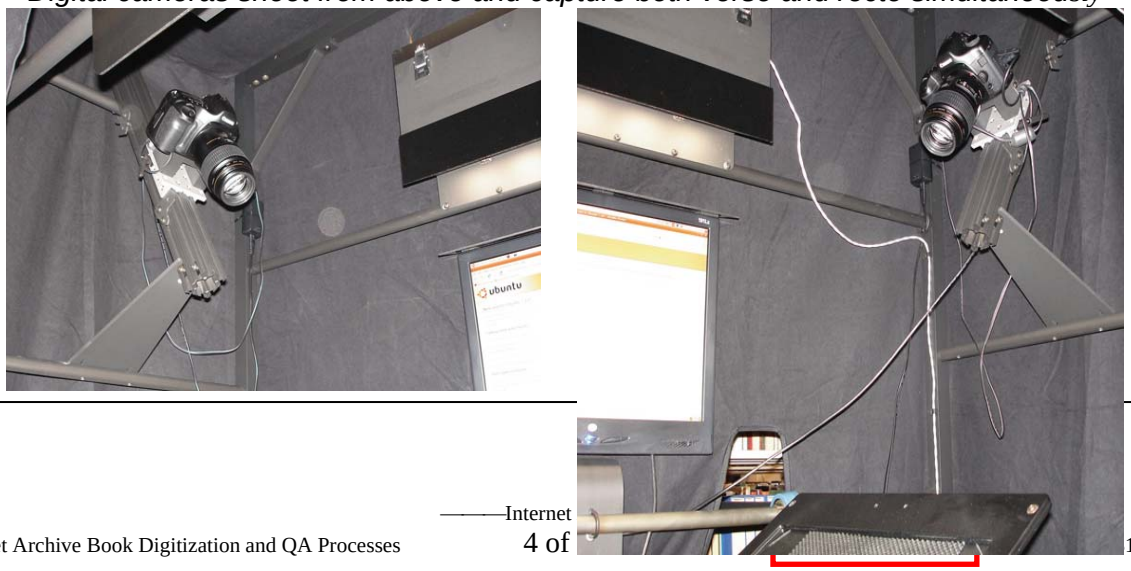
PPI	Height (inch)	Width (inch)
300	14.2	9.4
400	10.6	7
500	8.5	5.5

Internet Archive Digitization Workflow

Image Capture

1. 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://tinyurl.com/8hkhc) and the Canon EF 100mm f 2.8-macro lens (http://tinyurl.com/3ohfhp). 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



2. 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.

3. IA currently has a precision lighting compensation procedure in place, which ensures that consistent, even lighting is used during the digitization process. 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.

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

5. 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.

6. Equipment Calibration

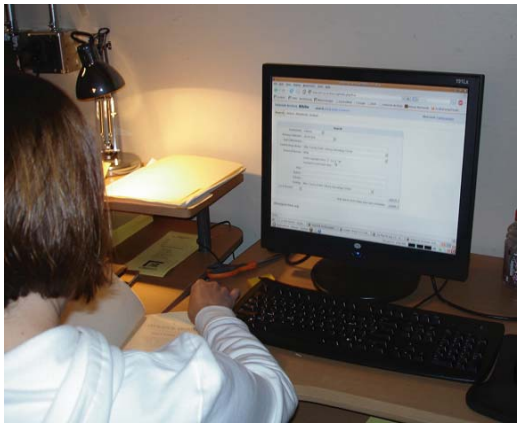
    a. Scribe station frames are calibrated and aligned before being put into use.

    b. Cameras are calibrated per manufacturer's specifications. Cameras that no longer perform within specifications are immediately sent to the manufacturer for repair.

    c. 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:

IA Loading Process



1. Prior to undertaking a new digitization project, an attribution form accompanied by a test "pick list" must be completed and reviewed with IA staff. Please contact Robert@archive.org or a site coordinator from the contact list on page 14 for further details.

2. Working from a book "pick list" provided by the Library Partner, a book ID or equivalent (e.g. a bibliographic ID) is loaded into the IA BiblioTool in order to locate a corresponding MARC record. A quality check is performed to ensure that the item being digitized and the MARC record match exactly.

3. 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 library record or condition issues, will receive an explanatory reject form and be returned to the Library Partner.

4. If IA receives a series of items that have been cataloged within one bibliographic ID, but have no discernable set of volume numbering, IA will work with the library partner to determine the best method for uniquely identifying each item, including 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.

5. IA's BiblioTool automatically creates a unique identifier for each item and the MARC record is attached to that identifier.

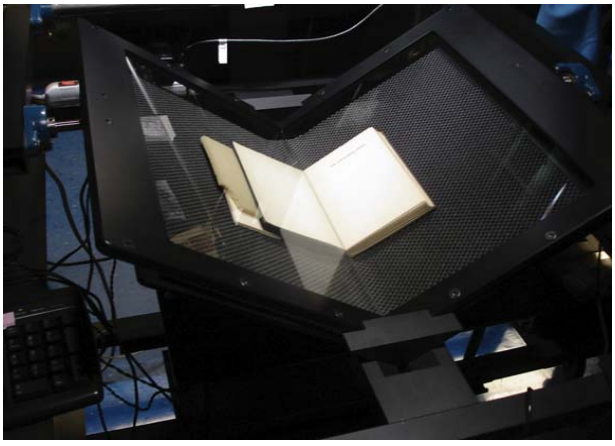
6. Each book within a collection or project is given a color-coded flag (to assist in book tracking), which also indicates the book identifier. Any special scanning instructions are included with the book. The book is now ready for digitization and is placed in a queue for the Scribe operators.

IA Scanning Process

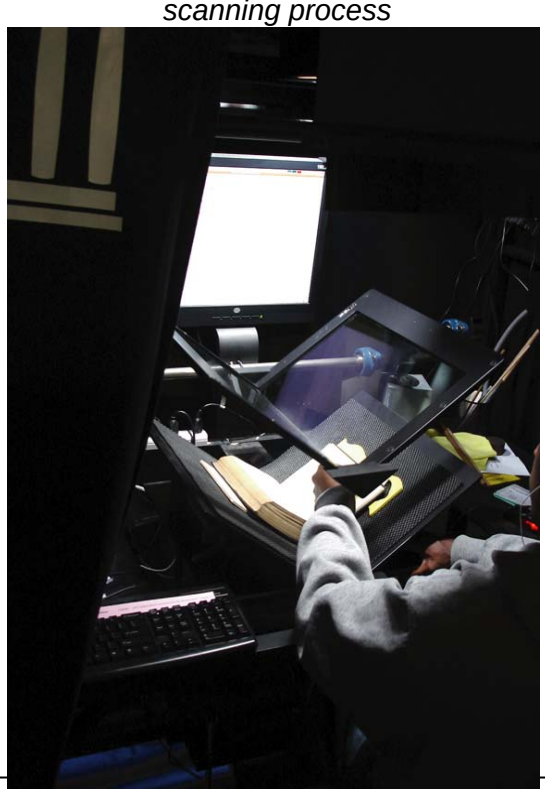
1. Scribe operators compare the color-coded flag containing the book identifier with the actual item in hand; to ensure that the digital file they create will be matched up with the correct corresponding metadata.

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

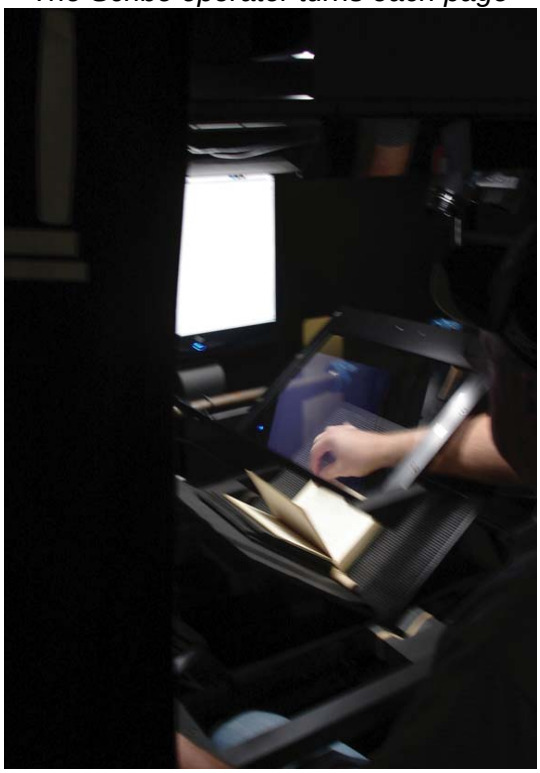
Images are scanned into the appropriate digital file



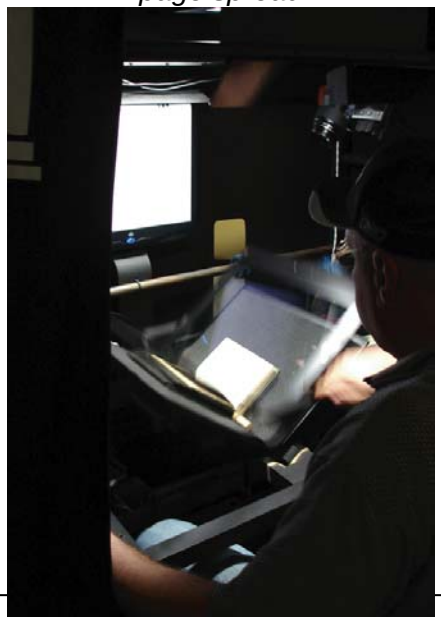
The Scribe operator may reposition the book as necessary during scanning process



The Scribe operator turns each page



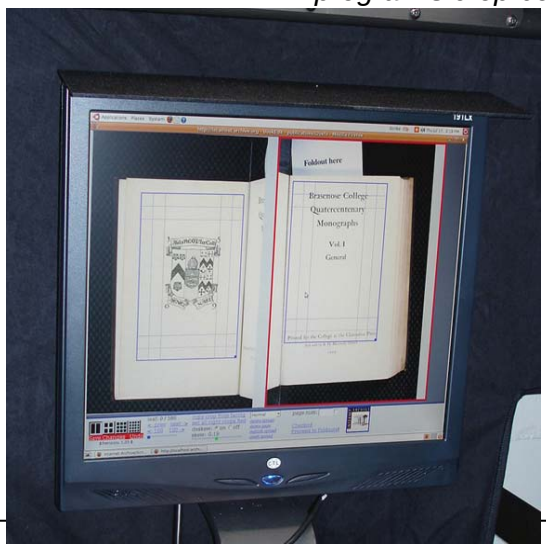
The glass platen is gently lowered onto the book before shooting the next page spread

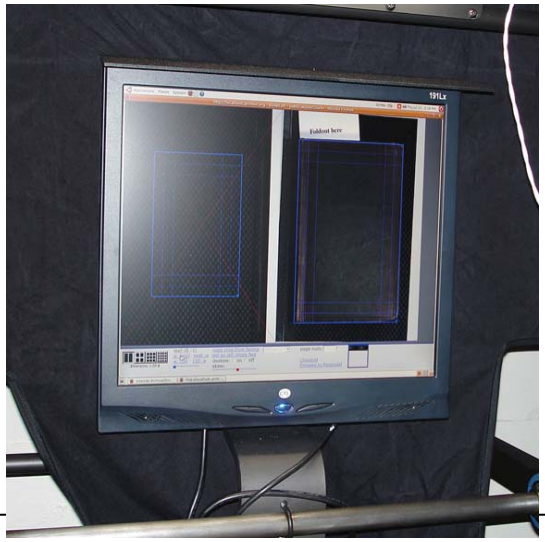


2. Images are shot through glass to minimize distortion due to page curvature. This also aids in the OCR process.

3. Images are quality checked throughout the digitization session and are adjusted for such things a such things as text block cropping and image de-skewing in order to ensure proper preservation and presentation.

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





4. The digital file is 100% Quality Inspected for missing pages, crop/de-skew issues and general presentation.

5. The completed files are then uploaded to IA's processing center.

IA Image Processing

1. Uploaded images are processed to create storage and access files.

2. File formats are covered in the Technical and Operational Details section on page 4. The Library partner, or anyone else, may download as many copies as they want.

3. Detailed list of files
    a. ID.pdf
    b. ID_jp2.zip
        - zipped folder of the book
        - [ID]_nnnn.jp2
    c. ID_lib_jp2.zip
        - zipped folder of the book (note: this folder could be included if a bookplate or watermark is applied)
        - [ID]_lib_nnnn.jp2
        - ID_marc.xml
    d. ID_meta.mrc

e. ID_meta.xml
f. ID_metasource.xml
g. ID_raw_jp2.zip - unprocessed storage format (no watermark/bookplate)
h. ID_wm_jp2.zip - this would be where a watermarked/bookplate version would be located.
i. Scandata.zip

4. Metadata will reside in meta.xml file, and will include the following:
    a. Identifier
    b. Identifier/Bib ID (IA identifier and local ID from pick list)
    c. Contributor
    d. Title
    e. Volume
    f. Creator (if in MARC record)
    g. Publisher (if in MARC record)
    h. h Collection (possibly multiple collection fields)
    i. Operator
    j. Scanner
    k. Scandate
    l. Identifier/Access (URL for accessing this book)

5. Processing Background

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.

- Storage format – raw JPG2000 is a compressed,high-quality lossy, un-cropped, non-rotated, non-de-skewed, camera-calibrated JPG2000 file. 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.)

- Processed master – high-quality lossy, cropped, rotated, de-skewed, camera-calibrated 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).

- 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 PDF and DjVu, both of which are OCRed.

- Note: all compression ratios might vary based on which version of ABBYY is used and specific software parameters. These numbers are for reference purposes only.

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 PDF compressor. This setting was determined based upon user surveys. As improved software becomes available, vendor selection may change.

Book Plate and Watermark

1. 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.

2. A watermark may be digitally inserted on every image page near the bottom. In extremely rare instances, when the text has afforded a very small or non-existent margin along the bottom, the watermark may overwrite the digitized text. A sample of what this looks is available upon request. Any decisions regarding the insertion of a bookplate or a watermark will be agreed upon before commencement of digitization

Turnaround for Processing by IA

For scanning 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 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.

1. IA's goal is to derive and upload books for web access within 24-48 hours after digitization.

2. An internal Quality Assurance process is performed inside the scanning center after the books are made available online. Criteria for Quality Assurance are outlined in the Quality Assurance section on page 15.

3. 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.

4. 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. The checkout process is discussed below.

IA Scanning Center Material Return and Check-Out Process

1. 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.

2. The Scanning Center Coordinator may generate an updated pick list that now contains corresponding IA book identifiers and indications of any rejected materials.

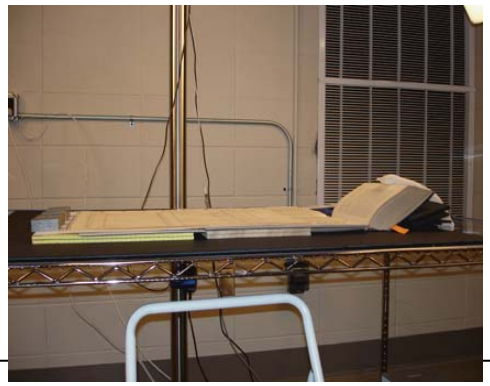
3. The books are then shipped or returned to the Library Partner.

Foldouts/Maps

1. Details regarding foldouts may be found in the Appendix, section 3, page 24. In general, images up to 18x24 inches may be captured in a single shot. Images larger than this will require multiple shots.
2. Photos of the foldout workstation are listed below.

Foldouts are digitized using an overhead camera stand, in a process step after the regular book digitization

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



LIBRARY PARTNER materials will be scanned at one of eight 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 9 smaller digitization centers located in California, Illinois, Maryland, North Carolina, Canada, Scotland, Guatemala and the United Kingdom.

Microfilm digitization is done at the San Francisco facility and Microfiche is digitized in Alberta, Canada. A separate workflow for Microform is available upon request.

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 at the New York Public Library branch, New York City, New York

"Scribes" at NYPL branch

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"

Site Coordinator contact details:

San Francisco CA	Damian@archive.org
Los Angeles, CA	KrisBrix@archive.org
Fort Wayne, IN	JeffS@archive.org
Princeton, NJ	Melissa@archive.org
Boston, MA	PaulN@archive.org
NYC, NY	Stacy@archive.org
Washington, DC	Ronnie@archive.org
Toronto, Ontario	Gabe@archive.org

Internet Archive Staff Contact List
Note: some contacts may change based upon the location of each specific scanning center.
Robert Miller is the primary point of contact for initial program review.

Name	Role	Phone	Email	Fax
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	
(see list above), Scanning Center Coordinator	Daily Operations (will receive and scan 10 initial test books)		(see list above)	
Eric Ostlund, Systems Engineer	Information Technology	(415) 561-6767	EricO@archive.org	
Julie Lefevre, Quality Manager	Quality Assurance	(415) 561-6767	Julie@archive.org	

Internet Archive Quality Assurance (QA) Process

Overview -There are four major phases within the QA process:

1. 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.)

2. After the images are on-line: 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.) Some elements may change slightly to reflect process improvements.

3. Before the invoice is sent: A monthly curation is performed and an internal random audit is conducted at IA headquarters.

4. After the invoice is sent: Errors brought to IA's attention will be reviewed (an on-line work ticket may be found on each book details page) 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 normally avoided, as it is usually the most expensive solution. The timeframe for the library to identify errors that will be fixed by the IA at no charge shall be detailed in the digitization plan.

A. Quality Assurance Process in use within IA 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

- Each day the scanning center will review a set of books from the previous day's scanning 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

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

- The online digital books are then inspected using the criteria below (see Quality Assurance codes B1-B3). Errors or defects, if found, 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

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:

- If a "fail" is generated, the Scanning Center Coordinator will review the errors/defects to ascertain if the errors were generated from outside the Scanning Center (for example, a missing access file error would be sent to engineering for review) or from within the Scanning Center (for example, a missing page).

- If an error is generated from within the Scanning Center, the Coordinator will follow a pre-determined set of process steps ultimately culminating in a recommendation to deviate or 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.

B. Codes Used and Displayed on the QA Report

B1. FREEZE CODES, part I

Items receiving codes 101-113 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 Book deleted from scribe before upload completed
    - 112 Missing files(s)
    - 113 Cr2.tar file is malformed

  Resolution: *Material is rescanned.*

- Metadata
    - 120 Book is not in public domain
    - 121 Date is 1923 or later
    - 122 Date is unclear
    - 123 Date is 1923 or later

  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
    - 131 Blurred page(s)
    - 132 Missing page(s) (IA's goal is to digitize every page; if a book is inherently missing a page/es, a note will be added to the details page.)
    - 133 Front cover missing
    - 134 Back cover missing
    - 135 Book was scanned twice (Identified copy is darkened and removed from search engine.)
    - 136 Text is washed out or overly dark (poor lighting)
    - 137 Evidence of scanner (fingers, shadows, etc) visible on page
    - 138 Glass not centered in gutter; text is distorted or cropped

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

B2. **FREEZE CODES, part II.**

These codes are used for books that have fixable problems, but are not yet in billable condition.

- 140 Book and metadata do not match
- 141 n/a
- 142 Tissue pages marked incorrectly
- 143 Anomaly in image format is under investigation
- 144 Left/right pages are reversed

Resolution*: For items receiving codes 140-142 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.*

B3. INFORMATIONAL CODES

- 150 Bibliographic data missing:
- 151 De-commissioned
- 152 Copyright evidence reported incorrectly
- 153 Bibliographic record from library is truncated
- 154 Possible error in bibliographic record from library
- 155 Foreign language character encoding is incorrect
- 156 Incorrect or missing collection-library or bibliographic ID
- 160 Light/dark pages (intermittent)
- 161 Light/dark pages (throughout)
- 162 Pages skewed
- 163 Color cards show in access formats
- 164 White cards show in access formats
- 165 Both white cards and color cards show in access formats
- 166 Image of cradle is visible at front or back
- 167 Different crop-box sizes within same spread
- 168 Bad crop at page edges
- 169 Duplicate page spreads scanned
- 170 Page types not marked or marked incorrectly
- 171 Title page not marked because book does not have title page
- 172 Scan factors not marked or noted
- 198 This would be a good display book
- 199 Approved with no problems noted

Resolution: *For items 150, 152 and 155, post derive correction is possible. For errors 153 and 154, errors must be reviewed with Library Partner. For errors 160, 161, 162, 166, 167 and 168, a review is conducted to see if the material has the ability to be OCRed. If the material can't be OCRed based on current IA software, the book is rejected for errors unless the Library Partner has different instructions. For items 166, 170 and 171, a post-derive correction is attempted. If unsuccessful, materials may be rescanned. 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*.

IA may attempt to correct the following errors through post-derive treatment: 110, 111, 112, 113, 140, 142, 144, 150, 152, 155, 160, 161, 162, 166, 167, 168, 170 and 171.

The following errors may require rescanning:
130, 131, 132, 133, 134, 136, 137 and 138.

The following errors require consultation with the Library Partner:
120, 121, 122, 123, 153, and 154.

B4. RESCANNING PROCESS

- For materials that are 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.

B5. ERROR CODES & CLASSES

To track errors found by Library, a bug tracking system may be created. Response times for error resolution will be determined based upon type of error and time of response. In general, IA will attempt to resolve errors brought to its attention within 30 days of an error being identified.

Class/ID	Description	Type
Images		
130	Cropped text	major
131	Blurred page(s)	major
132	Missing page(s)	major
133	Front cover missing	major
134	Back cover missing	major
135	Book scanned twice; this copy darkened	minor
136	Washed-out text (bad light-comp)	minor
137	Evidence of scanner visible on page	minor
138	Glass not centered, text distorted or cropped	minor
140	Book and metadata do not match	minor
142	Tissue pages marked incorrectly	major
143	Anomaly in image format is under investigation	major
144	Left/right pages are reversed	major
Bibliographic		
150	Bibliographic data missing: MetaFetch not run in scanning center (post MF done in QA)	minor
151	Bibliographic data missing: MetaFetch was run but did not merge (post MF done in QA)	minor
152	Copyright evidence reported incorrectly, info corrected in QA.	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
156	Incorrect/missing collection	major
Consistency		
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
165	White cards & color cards visible in access formats	minor
166	Image of cradle is visible at front or back	minor
167	Different crop-box sizes in same spread	minor
168	Bad crop at page edges	minor
169	Duplicate page spreads scanned	minor
170	Page types not marked or marked incorrectly	minor
171	Title page not marked, absent from book	minor
172	Scan factors not marked or noted	minor

C. Post-Scanning Reporting Tool: The IA Advanced Search Engine

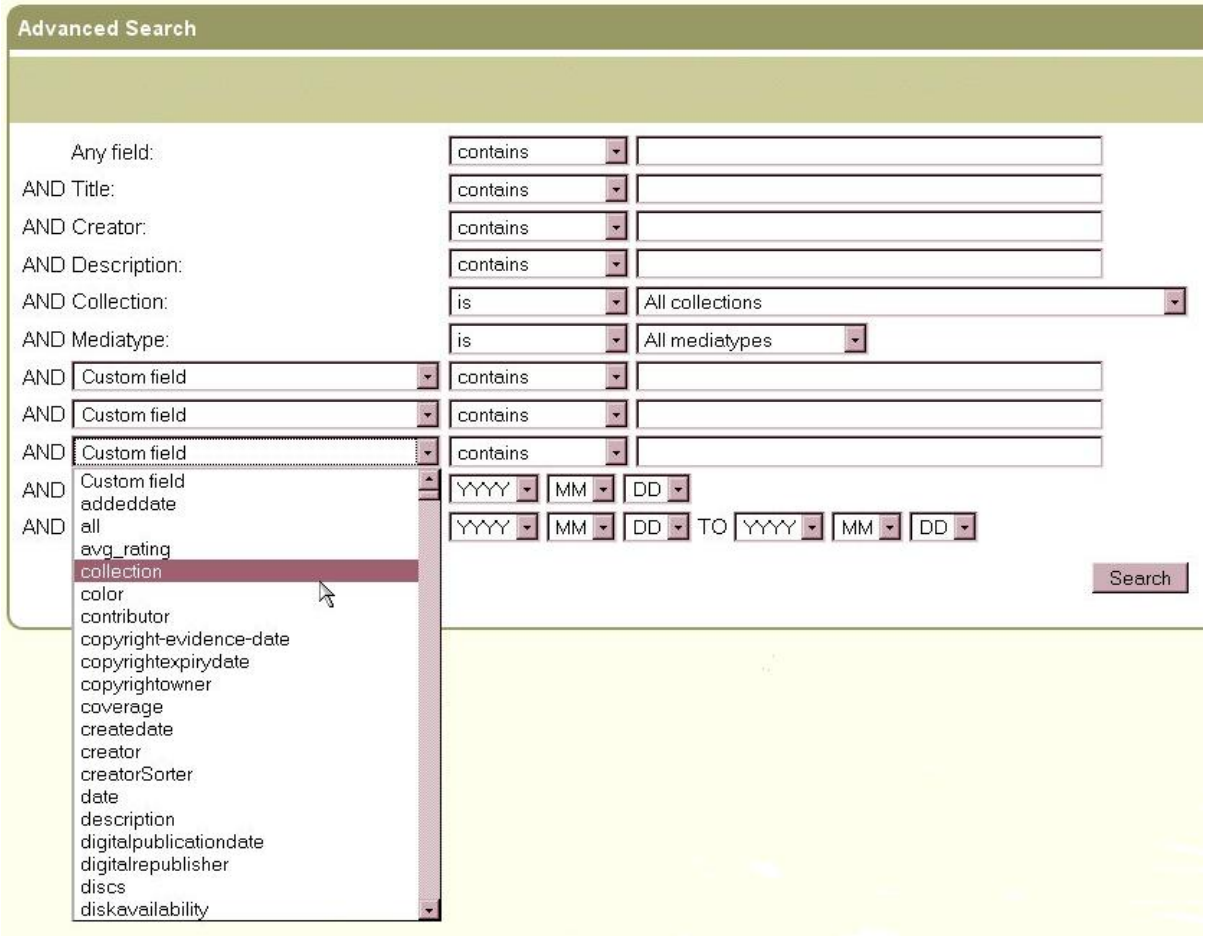
All Library Partners have access to the IA Advanced Search Engine, found at: http://www.archive.org/advancedsearch.php, which is a helpful reporting tool that may be used to search and review books that have been digitized, uploaded, QAed and curated. Fields that are viewable in the Advanced Search Engine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

Instructions for Using the Advanced Search Engine & Advanced XML Search

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.



The Advanced XML Search 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:

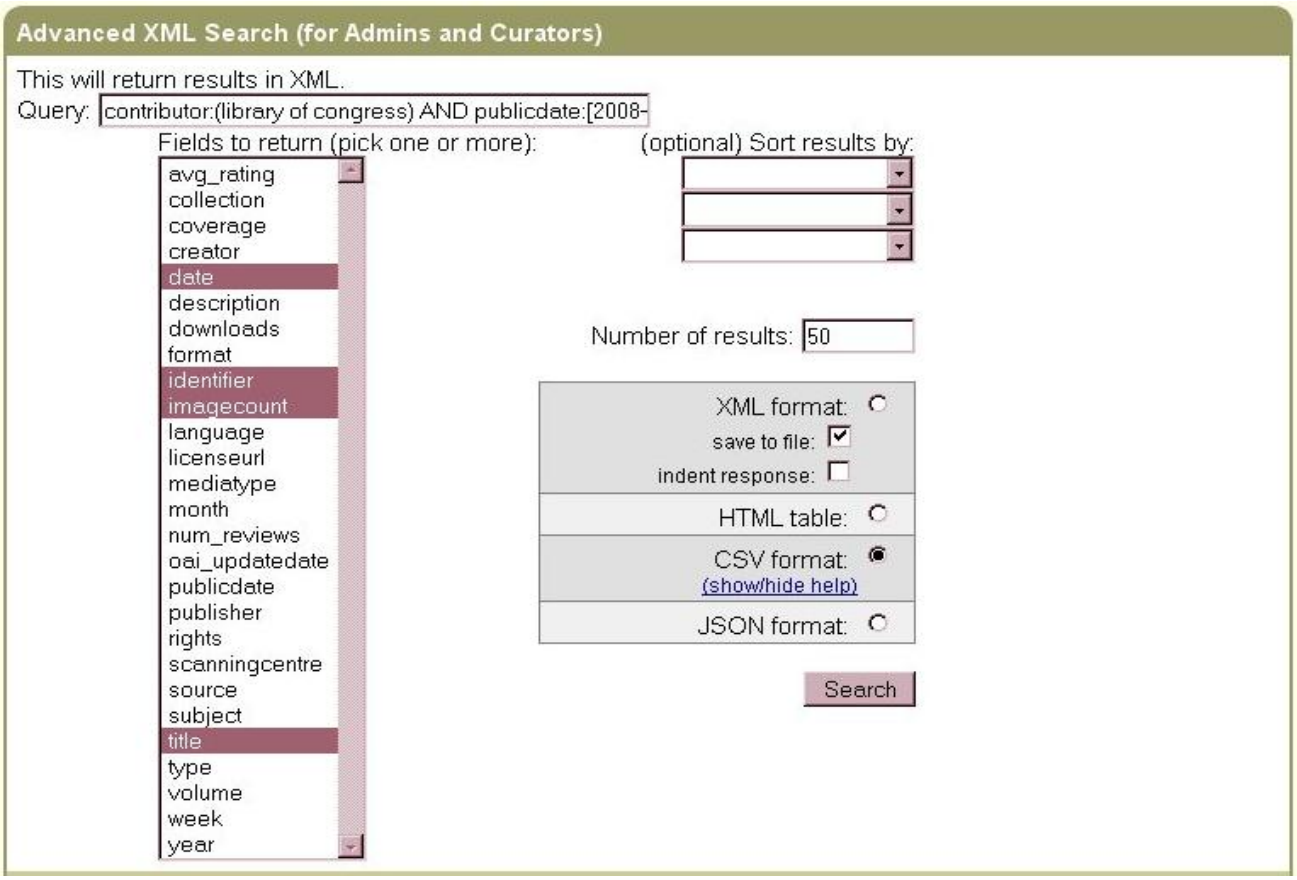
- In the Advance XML Search "Query" field, type:
  contributor:(library of congress) AND publicdate:[2008-08-01 TO 2008-08-30]

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

- Then, by holding down the "Shift" Key, highlight the fields you want to export to Excel. For example: "date, identifier, imagecount and title"

- 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.

- 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.



Appendix

Rejection Codes

Rejection codes for books that may be sent back to Member Institution include**:**

Code	Definition
BI	Fragile or no binding (includes items in clam shells or phase boxes)
CAT	Cataloging error
DAM	Damaged
DAT	Still in copyright
FO	Foldouts
LG	Too large
MAR	Margins too tight
MIS	Missing pages
MUL	Multiple titles bound together
NA	Not available
LAN	Outside language parameters
LIST	Pick list error
LINK	Unsuccessful link to metadata
NOS	Not on shelf (missing/lost)
OUT	Not on shelf (checked out)
PAG	Pagination problems (section[s] bound out of order or upside down)
PAP	Brittle paper, tissue paper
SKW	Skewed text (to the point of being unreadable)
SM	Too small
UNC	Uncut pages (more than 5)
FOR	Non-book format
VEL	Vellum
WD	Withdrawn
SPH	Requires special handling
DUP	Exact duplicate of another item on list

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:

1. Special preservation standards that are communicated by the Library Partner and accompany specific shipments/collections of books.

2. Materials that have multiple titles or multiple physical volumes (e.g. 75 volumes without analytics, containing 40 pamphlets per volume) will be reviewed to ensure all proper metadata is understood.

3. Materials will be screened for size and materials not fitting the requirements shown below will be returned un-digitized.

    - 9.5” wide x 14.5” high maximum. 3” wide x 3” high minimum.

- Books less than 3” thick may be acceptable and will be reviewed for binding/gutter before digitizing. Items thicker than 4” will be reviewed in order to determine if the book cradle can accommodate it.

- 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.

- IA has the capacity to digitize foldouts and maps. Testing and decisions about specifications should occur prior to digitization. Foldouts/maps greater than 18”x24” may be shot in a multipart, step-and-repeat fashion unless another method has been previously agreed upon. Foldouts and maps will have a lower resolution, or PPI, than regular image captures. Foldouts and maps are also priced higher than regular image capture. PPI settings vary according to size of image captured.

4. Book Style:

    - Ideal candidates for digitization are side-bound monographs, no single sheets and no top-bound books.

    - Soft cover books are acceptable as long as they are bound.

5. Book condition:

    - Covers that are separating from the body of the book, or otherwise appear fragile, may be rejected unless a decision to the contrary has been made by IA and the Library Partner prior to digitization.

    - Materials should be of similar condition or quality to items currently in public circulation. Materials not deemed robust enough to go into circulation must be reviewed with the Scanning Center Coordinator.

    - If a special book, rare book or the equivalent is to be scanned, IA MUST be notified beforehand if special handling process is used. TO REPEAT: IA MUST BE NOTIFIED IF A BOOK IS RARE, SPECIAL, FRAGILE OR REQUIRES EXTRA CARE.

6. Paper style:

    - 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 we will conduct a review of the amount of extra time needed for digitizing these materials.

    - Pages should not be excessively dusty or have excessive mildew or mold.

- All pages should have the bolts pre-cut. Unless otherwise instructed, books with bolts (uncut pages) will not be scanned.

- Scribe operators should be able to lifted and turn the pages with normal effort. Pages with adhesive or that are otherwise tacky will not be digitized.

- 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.

7. Gutters/Margins:

- All books, rebound books in particular, will be reviewed to check for tight gutters and narrow margins along all sides of the text block.

- Tight bindings that will not lay open for digitization per IA specification limits will be rejected.

- If the page text is less than a quarter inch away from the gutter, the book will not be digitized.

- Text that runs to the edge of the page or margin may be digitized but the final presentation may be poor. The decision to digitize material of this nature will be made prior to digitization.

8. Bibliographic data:

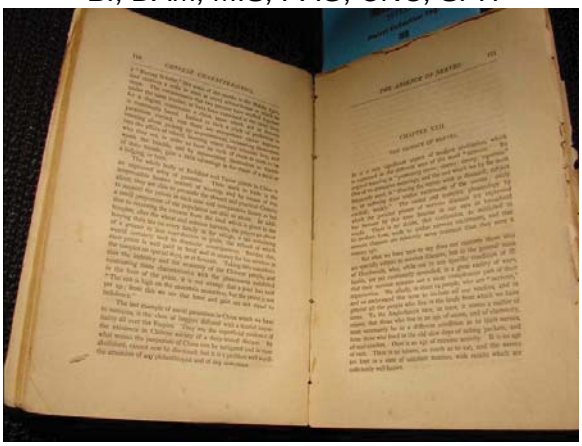
- When two or more titles are bound within the same volume (bound-widths) each item will need to be designated with its own pull slip and library identifier (i.e. Bib ID). Each title will be considered a separate book and will be scanned as such. IA and the Library Partner must discuss any deviations from this rule.

- Books that are out of approved copyright range may be digitized but should be discussed by IA and the Library Partner prior to digitization.

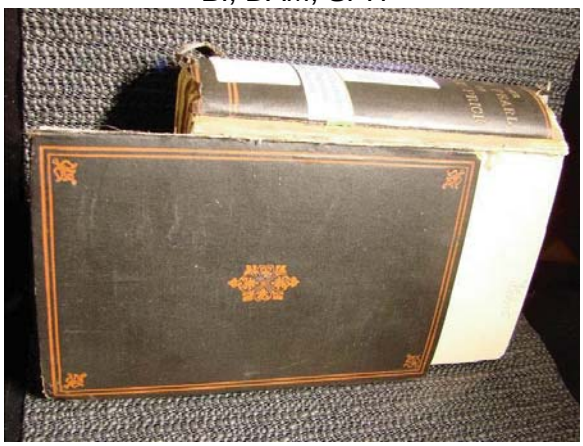
- If IA is unable to locate an item's MARC record, the book will be returned to the library partner pending further information.

Examples of "Problem" Books

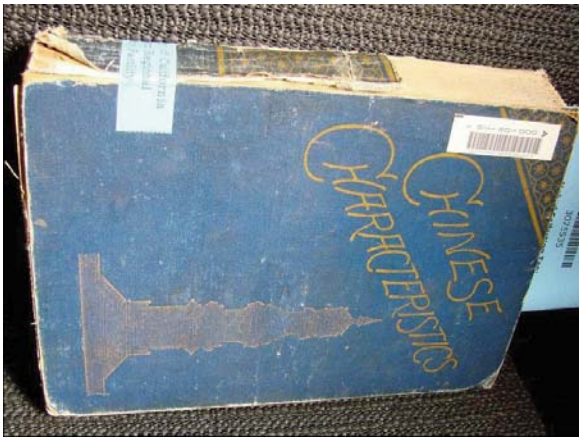
Example with the following problems:
BI, DAM, MIS, PAG, UNC, SPH



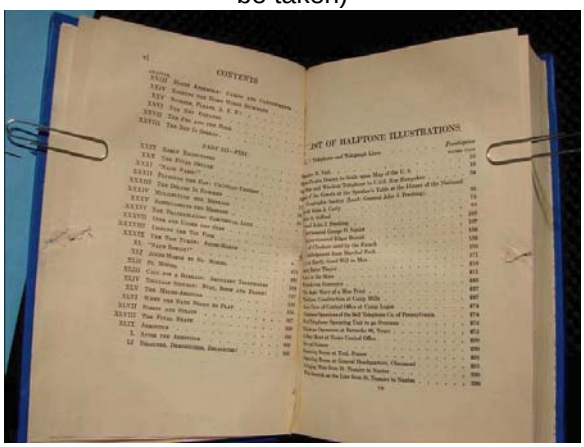
Example with the following problems:
BI, DAM, SPH



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