Workflows

Accessioning Workflow

1. Donor Agreement received
2. Media physically secured (create separation sheets if necessary to preserve original order)
3. Record accession information AT
4. Assign Barcode (use double barcodes for separation sheets as appropriate)
5. Photograph media
6. Acquire media content (disk image or copy)
7. Record checksums
8. Scan content for PII & Viruses
    Exceptions
        i. Check donor agreements for existing policies
        ii. If none apply: negotiate restriction, return, or destruction with donor
        iii. Comply with agreements
        iv. Record restrictions & actions taken
9. Move content to Dark Storage
10. Securely erase local copy

General Policies

- Electronic media received by RBMSCL should only be accessed in read-only mode
    Media with a USB interface must use the write-blocker
    Firewire & eSATA drives must be mounted in read-only mode
- No media received by RBMSCL shall be reused for any other purpose.
- Electronic media shall never leave the custody of RBMSCL/UA except for:
    Preservation activities (e.g. specialized data recovery services) under the direction of the Electronic Records Archivist (requires the use of a signed transfer form)
    Very large volume transfers copying to ITS secure network storage by ITS staff under the direction of the Electronic Records Archivist (requires the use of a signed transfer form)
- All media should be clearly marked with the accession number and/or collection name.
    Label bands or dedicated storage boxes with labels are preferred. Avoid directly labeling the media if possible.
- If there is an unavoidable delay in transferring the data to the secure network storage, a record for the data will be added to the electronic media transfer queue so that the need for the transfer is documented and attended to in a timely manner.
- RBMSCL/UA transfer drives (used by archivists visiting the donor):
    Shall be clearly labeled
    Used only by permission of the Electronic Records Archivist
    Shall be cleared only after transfer to ITS secure network storage has been verified and then only by the Electronic Records Archivist
    Archivists shall request the donor NOT purge copied files until transfer has been verified

Seth Shaw — last modified Jul 20, 2011 04:13 PM

Bentley Historical Library Digital Processing Manual

Table of Contents

Introduction

This processing manual provides guidance and instructions for the processing of digital materials at the Bentley Historical Library (BHL). Procedures, tools, and the overall digital processing workflow are subject to change due to advances in professional best practices, the development of resources in the digital curation community, and the Bentley Library's ongoing collaboration with the University of Michigan Library Information Technology division. In addition to revisions that take place as BHL Digital Curation Services implements digital processing procedures, this manual will be reviewed on an annual basis.

The BHL *Digital Processing Manual* details procedures that will take place from the initial transfer and appraisal of content to archival custody through the eventual deposit of material in a long-term digital repository. Digital Curation Services advocates a *More Product, Less Process* approach to handling digital records and emphatically notes that processing archivists and student processors will not be able to deal with content on an individual file level. The BHL digital processing workflow instead relies upon a number of micro-services that will perform batch operations on digital accessions. In addition to traditional archival procedures such as the appraisal, arrangement, separation, and description of content, digital processing for long-term preservation requires the following:

- Migration of content from removable media
- Capture
- Virus scans
- Renaming
- File format conversion
- Personally-identifiable information scans
- Creation of ZIP archives
- File characterization
- Message digest calculation

The various steps in the digital processing workflow produce log files that will be preserved as metadata for the digital accession. It is of the utmost importance that the steps and procedures outlined in this manual—from file naming conventions for log files to settings used in application—be strictly followed by all BHL employees engaged in the processing of digital content. In addition, processing archivists and student processors will produce descriptive, administrative, and preservation metadata that will permit the Bentley Library to generate a Metadata Encoding and Transmission Standard (METS) document for each digital deposit.

Progress on each digital deposit will be tracked with the Bentley Historical Library Digital Processing Checklist, a document that will reside in the *Metadata*\ folder.

Workflow: Overview

The basic workflow for processing digital records will involve the following steps:

1. UARP/MHC reaches an agreement with a donor/creator regarding the transfer of digital content to the Bentley Historical Library.

2. Archivists are provided access to digital materials (either remotely or via removable storage media).

3. A preliminary review of the digital materials will be performed (if it has not taken place prior to the transfer agreement) to determine if they warrant additional processing and long-term preservation by the Bentley Historical Library. Archivists will also confirm the presence of sensitive materials that will require restrictions under applicable laws and/or BHL policies.

4. Create an accession record in BEAL. If some/all of the digital content will not be processed for long-term preservation, note these materials in the separations.

5. Digital Curation Services will manage the creation of and access to appropriate processing directories in the Interim Repository.

6. Content will be migrated to the appropriate processing directory in the Interim Repository.
    a. Depending on the source/transfer method, archivists will use one of several tools identified and tested by Digital Curation Services.
    b. Processing directory will include a *Metadata*\ folder.
    c. Create a separations folder (titled: CollectionID_Name) in *bhl-root**Separations*\
    d. Note the unprocessed location in BEAL and record the capture of content in the PREMIS preservation event spreadsheet.

7. Following a *More Product, Less Process* approach, the archivist/student processor will conduct the following operations:
    a. Change filename to the Deposit ID (the collection ID plus a four digit number, i.e. 87134_0001)
    b. Virus scan (Save log file in the *Metadata*\ folder and record event in the PREMIS preservation event spreadsheet.)
    c. Backup of content
    d. Normalization of folder/file names (Save log file in the *Metadata*\ folder and record event in the PREMIS preservation event spreadsheet.)
    e. Scan for personally identifiable information (PII) (Save log file in the *Metadata*\ folder and record event in the PREMIS preservation event spreadsheet.)

f. Appraisal and analysis of content (If email is present, archivist may need to convert file format to MBOX to review messages in an MBOX viewer.)
g. Add file extensions to unidentified files with TRiD (Save log file in the *Metadata*\ folder and record event in the PREMIS preservation event spreadsheet.)
h. Separation of unnecessary or superfluous content
    i. Use TreeSize to identify and move content to appropriate folder in *bhl-root**Separations*\
    ii. Save log file in the Metadata folder and record event in the PREMIS preservation event spreadsheet.
i. Arrangement (only if needed)
j. Run *bhl_batch.bat* to create preservation copies of material in at-risk formats.
    i. Text/office documents: MS Word, Excel, and PowerPoint documents will be migrated to 2010 Office Open XML; PDF documents will be converted to PDF/A.
    ii. Raster Images: BMP, PSD, PCD, PCT, and TGA will be converted to TIFF.
    iii. Raw Camera Images: 3FR, ARW, CR2, DCR, MRW, NEF, ORF, PEF, RAF, RAW, and X3F will be converted to JPEG (for access)
    iv. Vector Images: AI, EMF, and WMF will be converted to SVG; PS and EPS will be converted to PDF/A.
    v. Audio files: WMA, RA, SND, and AU will be converted to WAV.
    vi. Video files: FLV, WMV, RMVB, and RV will be converted to MPEG4 (with H.264 encoding).
    vii. Email will be converted to MBOX.
    viii. Database files: ACCDB, MDB, SQL Server and Oracle DB will be converted to SIARD open XML.
k. Create ZIP archive files (if necessary) and finalize packaging of content for deposit in a long-term preservation repository
l. Content characterization with DROID

8. Content will be transferred to a post-processing location
    a. Restricted content: *bhl-archive*\ ("dark" storage location)
    b. Unrestricted content: *bhl-root**deepblue_deposits*\ in the Interim Repository

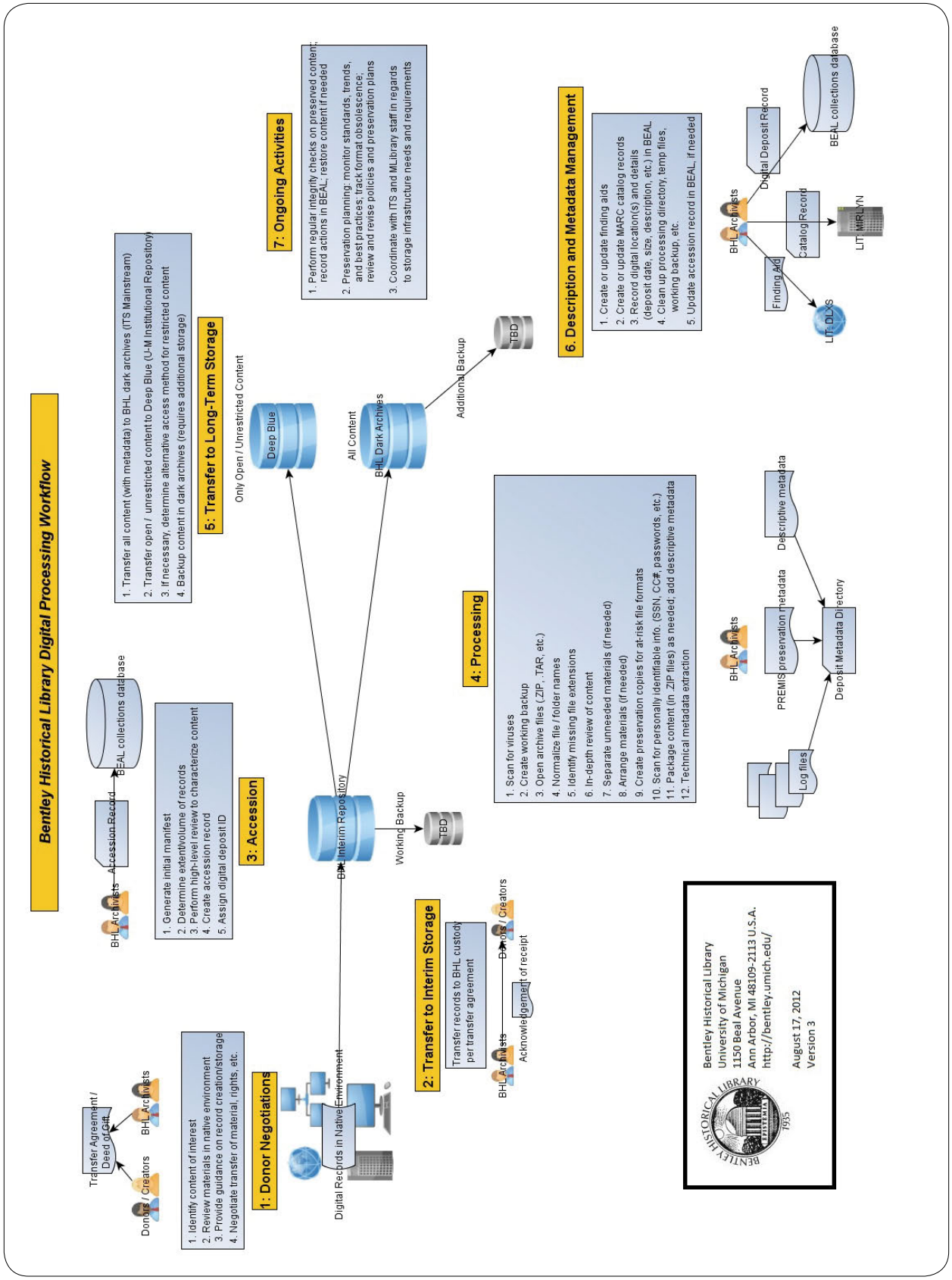
9. Complete metadata forms
    a. Deep Blue deposit spreadsheet
    b. PREMIS preservation event spreadsheet
    c. EAD descriptive and administrative metadata template

10. If the content is unrestricted, Digital Curation Services will coordinate its deposit in Deep Blue.

11. For unrestricted material, place a copy of the deposit (with *Metadata*\ folder) in *bhl-archive*\.

12. Description:
    a. Create/update finding aid
    b. Create/update catalog record
    c. Update BEAL record

13. Clean up:
    a. Manage disposition of separations, per the transfer agreement.
    b. Delete backup copy
    c. Delete version from 'Unprocessed' and *deepblue_deposits*\ directories (if applicable).


Bentley Historical Library Digital Processing Workflow
1: Donor Negotiations
1. Identify content of interest
2. Review materials in native environment
3. Provide guidance on record creation/storage
4. Negotiate transfer of material, rights, etc.
Transfer Agreement / Deed of Gift
Donors / Creators
BHL Archivists
Digital Records in Native Environment
3: Accession
1. Generate initial manifest
2. Determine extent/volume of records
3. Perform high-level review to characterize content
4. Create accession record
5. Assign digital deposit ID
BHL Archivists
Accession Record
BEAL collections database
2: Transfer to Interim Storage
Transfer records to BHL custody per transfer agreement
BHL Archivists
Donors / Creators
Acknowledgement of receipt
BHL Interim Repository
Working Backup
TBD
4: Processing
1. Scan for viruses
2. Create working backup
3. Open archive files (.ZIP, .TAR, etc.)
4. Normalize file / folder names
5. Identify missing file extensions
6. In-depth review of content
7. Separate unneeded materials (if needed)
8. Arrange materials (if needed)
9. Create preservation copies for at-risk file formats
10. Scan for personally identifiable info. (SSN, CC#, passwords, etc.)
11. Package content (in .ZIP files) as needed; add descriptive metadata
12. Technical metadata extraction
BHL Archivists
Log files
PREMIS preservation metadata
Descriptive metadata
Deposit Metadata Directory
5: Transfer to Long-Term Storage
Only Open / Unrestricted Content
1. Transfer all content (with metadata) to BHL dark archives (ITS Mainstream)
2. Transfer open / unrestricted content to Deep Blue (U-M Institutional Repository)
3. If necessary, determine alternative access method for restricted content
4. Backup content in dark archives (requires additional storage)
Deep Blue
All Content
BHL Dark Archives
Additional Backup
TBD
7: Ongoing Activities
1. Perform regular integrity checks on preserved content; record actions in BEAL; restore content if needed
2. Preservation planning: monitor standards, trends, and best practices; track format obsolescence; review and revise policies and preservation plans
3. Coordinate with ITS and MLibrary staff in regards to storage infrastructure needs and requirements
6. Description and Metadata Management
1. Create or update finding aids
2. Create or update MARC catalog records
3. Record digital location(s) and details (deposit date, size, description, etc.) in BEAL
4. Clean up processing directory, temp files, working backup, etc.
5. Update accession record in BEAL, if needed
BHL Archivists
Finding Aid
Catalog Record
Digital Deposit Record
BEAL collections database
LIT: DLXS
LIT: MIRLYN
Bentley Historical Library
University of Michigan
1150 Beal Avenue
Ann Arbor, MI 48109-2113 U.S.A.
http://bentley.umich.edu/
August 17, 2012
Version 3

Bentley Historical Library Web Archives: Methodology for the Acquisition of Content

Nancy Deromedi and Michael Shallcross
Digital Curation

Version 2.0 (August 2, 2011)

Table of Contents

Introduction

The Bentley Historical Library's Digital Curation Division has developed a methodology and workflow for the acquisition of content. These procedures are based on the available features of the California Digital Library (CDL)'s Web Archiving Service (WAS) as well as standard archival practices (such as appraisal and description). This document provides an overview of the Bentley Historical Library's methodology for website preservation.

The actual process of website preservation may be broken down into three main steps:

1. Identification of the crawl target
2. Configuration of the crawler settings
3. Contextualization of content

Guided by collecting priorities, surveys of relevant websites, and knowledge of significant individuals and organizations, archivists identify potential targets for preservation. By standardizing the configuration of web crawler settings and addition of metadata and descriptions, archivists are able to ensure that websites are preserved in a manner that is consistent, efficient, and cost-effective.

Given the fast pace of change in web archiving technology and ongoing development of features and functionalities in WAS, this methodology document will be reviewed on an annual basis and revised accordingly.

Identification of Content

The Bentley Historical Library employs the Heritrix web crawler (also known as a spider or robot) to copy and preserve websites. As a subscriber to WAS, the Bentley Library relies upon an implementation of Heritrix specially configured and maintained by the CDL. A web crawler is an application that starts at a specified URL and then methodically follows hyperlinks to copy html pages and associated files (images, audio files, style sheets, etc.) as well as the websites underlying structure. The initiation of a web capture requires the archivist to specify one or more seed URLs from which the web crawling application will preserve the target site.

Accurate and thorough website preservation requires the archivist to become familiar with a site's content and architecture in order to define the exact nature of the target. This attention to detail is important because content may be hosted from multiple domains. For example, the University of Michigan's Horace H. Rackham School of Graduate Studies hosts the majority of its content at http://www.rackham.umich.edu/ but maintains information on academic programs at https://secure.rackham.umich.edu/academic_information/programs/. To completely capture the Rackham School's online presence, archivists needed to identify both domains as seed URLs.

At the same time, multiple domains present on a site may merit preservation as separate websites. For example, the University of Michigan's Office of the Vice President of Research (http://research.umich.edu/) maintains a large body of information related to research administration (http://www.drda.umich.edu/) and human research compliance (http://www.ohrcr.umich.edu/). Although these latter sites could be included as secondary seeds for the Vice President of Research's site, their scope and informational value led archivists to preserve them separately.

Once the target of the crawl has been identified and defined, the archivist enters the seed URL(s) and site name in the WAS curatorial interface (see Figure 1).

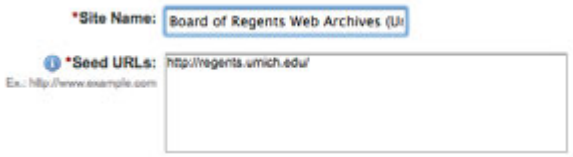


Figure 1

The Bentley Historical Library standardizes the names of preserved sites by using the title found at the top of the target web page or, in the absence of a formal/adequate title, the name of the creator (i.e. the individual or organization responsible for the intellectual content of the site). The library follows the best practices for collection titles as established by Describing Archives: a Content

Standard (DACS); to ensure that the nature of the collections are clear, archivists supply "Web Archives" in the final title. The University Archives and Records Program (UARP) furthermore includes "University of Michigan" in titles to highlight the provenance of websites. Complete names for sites in the University of Michigan Web Archives thus follow the pattern "Board of Regents Web Archives (University of Michigan)."

Configuration of Web Crawler Settings

WAS utilizes the open-source web crawler Heritrix to archive websites. As a command-line tool, this application allows for a wide range of user settings; the curatorial interface in WAS provides for a more-limited number of options. For each crawl, archivists may adjust the following settings:

- **Scope:** defines how much of the site will be captured. The archivist may elect to capture the entire host site (i.e. http://bentley.umich.edu/), a specific directory (i.e. http://bentley.umich.edu/exhibits/), or a single page (i.e. a letter written by Abbie Hoffman to John Sinclair, featured at http://bentley.umich.edu/exhibits/sinclair/ahletter.php) (see Figure 2).

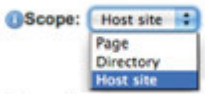


Figure 2

To thoroughly capture target websites, the Bentley Historical Library generally uses the "Host site" setting, unless the target is a single directory located on a more extensive host or a specific page.

Linked pages: determines whether or not content from other hosts/URLs will be captured; archivists have two options for this setting. If set to "No," the crawler will only archive materials on the seed URL entered by the archivist; if "Yes," the crawler will follow hypertext links one 'hop' to capture linked resources. Capturing linked pages will not result in an indefinite crawl (in which the robot follows link after link after link); instead, the crawler will only capture the page (and embedded content) that is specified by the hypertext link. No additional content on this latter site will be crawled.

To avoid preserving extraneous content, the Bentley Historical Library by default does not captures linked pages. Archivists will only capture linked pages if it required as a result of website design or if it is necessary to capture contextual information for a high priority web crawl.

Maximum time: specifies the maximum duration of a crawl. The archivist may select "Brief Capture (1 hour)" or "Full Capture (36 hours)" and the crawl will continue until all content has been preserved (in which case it may end early) or the allotted time period has elapsed. If a session times out before the crawler has finished, the resulting capture may be incomplete.

To avoid missing content due to time restrictions, the Bentley Historical Library uses the "Full Capture" option by default. Archivists use the "Brief Capture" if the target involves a limited amount of content and the additional

crawl time would result in unnecessary content (for instance, the archivist only wants to capture a blog's most recent posts and is not interested in the entire site).

- **Capture frequency:** designates how often a crawl will be repeated. The archivist may elect to crawl a site once or configure the robot to perform daily, weekly, monthly, or custom captures (see Figure 3).

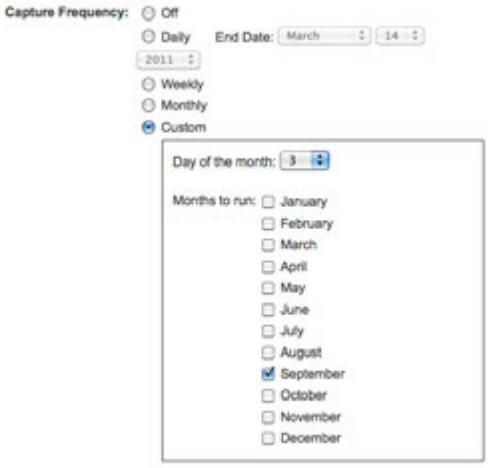


Figure 3

Archivists generally choose the "Custom" option and select an annual capture date, being mindful of important events/dates that might result in updates to the target site. (For instance, University of Michigan sites are captured near the beginning or end of the academic year.) This strategy is particularly effective with 'aggregative' websites in which new content is placed at the top/front of pages while older information is moved further down the page or placed in an 'archive' section. For high priority targets (such as the University of Michigan Office of the President) or sites with a large turnover of important content, captures may be scheduled on a more frequent basis.

As the foregoing discussion reveals, the accurate and effective configuration of crawl settings must be based on the archivist's appraisal of content and understanding of the target site's structure. The failure to consider these factors may lead to a capture that, on the one hand, is narrowly circumscribed and incomplete or, on the other, is unnecessarily broad and filled with superfluous information.

Contextualization of Content

After the configuration of crawl settings, archivists supply each website with a description, metadata, and tags to help contextualize the preserved content and facilitate access.

Description:

WAS provides a 'Site Description' field so that archivists may contextualize preserved websites with an overview of the creator and/or subject matter (see Figure 4).

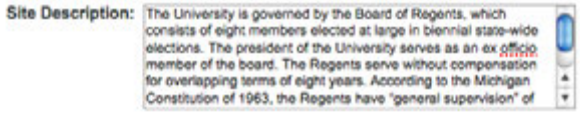


Figure 4

To ensure accurate descriptions, archivists often use text supplied by the websites in an "About Us" or "More Information" section, if it is available. Patrons have ready access to this information from each page in the web archives under the "Show Details" tab (see Figure 5).

Figure 5

Metadata

The WAS curatorial interface permits archivists to enter information related to the "Creator," "Publisher," "Subjects," and "Geographic coverage" of each site (see Figure 6).

Figure 6

Although WAS intended these metadata fields to mirror elements in the Dublin Core Metadata Set, the Bentley Historical Library needed to establish local definitions and conventions. After extensive discussions among archivists, the following practices were adopted:

- *Creator* denotes the individual or organization that generated or supplied the website's intellectual content (and not merely the web designer who created the page).

- *Publisher* refers to the entity ultimately responsible for the production and presentation of content. Although the publisher may often be identical to the creator, the Regents of the University of Michigan are recognized as the collective publisher for all sites affiliated with the university. Similar situations may arise with other archived websites.

- *Subjects* express Library of Congress subject authorities that correspond to MARC21 6XX fields. Due to the lack of formatting in this field (and the indeterminate status of their use within WAS), the Bentley Historical Library does not include indicators and subfield codes but instead simply enters the primary and secondary descriptors and separates them with double hyphens.

- *Geographic coverage* identifies where the activities described in the site took place. Archivists again utilized MARC21 conventions so that the main geographic entry is followed by the subdivision but did not (for reasons stated above) include the field codes themselves.

Tags

WAS also allows archivists to "tag" archived websites with one or more subject terms to facilitate user access to content. Archivists have therefore created tags that identified significant groups of interrelated content: for example, the "College of Engineering" tag identifies all archived websites that are created, maintained, or associated with this particular college. When browsing the site list of a public archives, a user may select a tag to review only those archived websites associated with a specific subject (see Figure 7).

Figure 7

Tags are currently employed in both the Bentley Historical Library Web Archives; additional ones will be created as the collections continue to expand and as archivists receive feedback from users. Management features in the curatorial interface allow archivists to modify or delete tags; all sites that are denoted by the affected tags will inherit these changes (see Figure 8).

Figure 8

Many sites in the web archives do not have tags because they do not fit into these established categories and tagging is only effective when there are a significant number (i.e. five or more) of related sites. Archivists may, however, add tags to existing archived websites should the need arise.

With the inclusion of description, metadata, and tags, the archivist may initiate the web crawl and successfully conclude the workflow for content acquisition. Archivists regularly meet to discuss the status of the web archives and review difficult appraisal and content management decisions.

Quality Assurance for Bentley Historical Library Web Archives: Guidelines and Procedures

Version 1.0 September 21, 2011

Michael Shallcross
Nancy Deromedi

Bentley Historical Library
Digital Curation Division

Table of Contents

Introduction

Quality assurance (QA) refers to the systematic evaluation of an activity or product "to maximize the probability that minimum standards of quality are being attained."[1] In performing QA on websites preserved by the University Archives and Records Program (UARP) and Michigan Historical Collections (MHC), the Bentley Historical Library (BHL) seeks to ensure the accuracy and integrity of its web archives collections.

BHL staff involved in the preservation and QA of archived websites should have a some understanding of the design and architecture of websites (including links, embedded content, web forms, navigational menus, etc.) as well as basic knowledge of HTML, Cascading Style Sheets (CSS), JavaScript (JS), and other significant web page features. A familiarity with the curatorial interface and basic functions of the California Digital Library (CDL)'s Web Archiving Service (WAS) is also important.

During this process, a BHL QA specialist will:

- Identify incomplete, inaccurate, or unsuccessful web captures
- Determine the underlying causes or issues that led to the substandard captures. This step may require the QA specialist to:
    - Verify crawl settings
    - Review crawl reports and logs
    - Inspect the content, layout, features, and source code of the target site
- Document:
    - Any technical limitations, robots.txt exclusions, or other issues that may have prevented a faithful and accurate capture of a website.
    - Contact information for webmasters (if necessary)
    - Recommendations to delete captures or initiate new crawls

Given the inherent challenges of various content types and the technical limitations of the WAS infrastructure, it is not feasible to perfectly preserve the content, appearance, functionality, and structure of all targeted websites. Although QA may not resolve all issues with a given archived website, careful documentation will help to establish the provenance of content and record actions taken by the archives. Information gathered during QA will also enable the library to revisit problematic captures as web archiving technology continues to mature.

The CDL's release of additional quality assurance tools and reporting features for WAS in late May/early June 2011 will require the revision of these guidelines and procedures. This document will also be reviewed on an annual basis to ensure that the information and procedures contained herein are current and applicable.

[1] "Quality assurance." *Wikipedia* (May 5, 2011). Retrieved on May 6, 2011 from http://en.wikipedia.org/wiki/Quality_assurance.

QA Procedures for Bentley Historical Library Web Archives

1. For each site, use the QA Spreadsheet to record:
    a. Your initials
    b. The date on which QA was conducted
    c. The number of captures currently held for the site
    d. The date range of the captures (may be a single date).

2. From the "Manage Sites" screen of the WAS curatorial interface, click on the site name to access the "Site Summary." (You may choose to right-click and open in a new tab.)
    a. Capture Settings
        i. Verify that the site name (i.e. "Department of Chemistry Web Archives (University of Michigan)") adheres to BHL conventions.
            1. BHL conventions for site titles may be found in the document: "Bentley Historical Library Web Archives: Methodology for the Acquisition of Content" (pp. 3-4).
            2. Modify site names as needed in step 8 (being sure to respect the original site's name, if possible).
        ii. Check if "linked pages" are being captured:
            1. For U of M content:
                a. Only "high priority" sites should include the capture of linked pages.
                b. For all other sites, linked pages should not be captured to avoid an excessive amount of content in the web archives.
            2. For MHC content, the QA specialist may need to verify if linked content should be captured. (See later steps.)
    b. Scheduling
        i. For U of M:

1. Only "high priority" sites will be scheduled for more than one capture a year (see list on p. 7).
2. Campus event websites (including the Arts Portal, Online Event calendar, etc.) and the Gateway may also be captured more frequently.
3. All other sites should only be captured on an annual basis.

ii. For MHC:

1. If there are multiple captures scheduled, conduct crawl comparisons to see if these are necessary.
2. Check with Project Administrators before adjusting schedule.

c. Descriptive Data

i. Check Description, Creator, Publisher, Subjects, and Geographic coverage elements to ensure that they follow BHL conventions.

1. BHL conventions for metadata entry may be found in the document: "Bentley Historical Library Web Archives: Methodology for the Acquisition of Content" (pp. 7-8).
2. Edit metadata as needed in step 8.

ii. Check "Site Tags" (on right hand side)to see if the archived website could be grouped with other relevant subjects. (This determination may require the QA Specialist to view the archived page.)

1. A full listing of tags for a specific project is available under the "Administration > Mange Tags" menu item.
2. BHL conventions for tagging may be found in the document: "Bentley Historical Library Web Archives: Methodology for the Acquisition of Content" (p. 9).
3. Only Project Administrators may add new tags to the current list. Please inform the appropriate administrator

if you believe that an additional tag (or tags) may be necessary.

d. Capture History
    i. Check general the following for potential issues:
        1. "Status": may reveal ongoing technical issues
        2. "Files": could be problematic if extremely low or high
        3. "Duration": could be problematic if extremely short or timed out

3. Click "View Results" link to access the Crawl Overview
    a. Check seed URL(s) for redirects
    b. In case of an extremely small number of files or short duration, check "Robot Exclusions" statistics to see if the crawler was blocked
    c. In case of an extremely large number of files or in the event that the crawler exceeded the 36 hour duration, check the "Hosts Report" to see how many URLs are remaining for the main seed URL(s)
    d. Pending the review of the archived content, it may be necessary to examine other crawl reports.

4. View archived website
    a. Verify that content is an archived resource (instead of a redirected 'live' web page).
    b. Verify that CSS files are present (i.e. pages are *not* text only)
    c. Click on main navigational links (depending upon crawl settings, additional content may or may not have been intended for capture).
    d. For high priority targets, click through the entire site to ensure that significant content and features have been captured.
    e. Troubleshooting:
        i. If a particular resource does not appear in the archive, conduct a search for the URL (search feature available from the main Results screen)

ii. Viewing the source code of the original page will help to identify web design features or resources that may not have been captured.
iii. Check live version of archived site (if available) to compare appearance of archived version.
iv. Check reports/crawl logs to understand issues with the crawl.
    1. Look up specific URLs to see if they were captured.
    2. Trace progress of crawl, identify where issues arose.

f. If (for MHC or high priority U of M sites) linked pages have been captured, determine if these contain significant information. This may require consulting the "Hosts" report (or others).

5. For sites with multiple captures:
    a. If there are more than 3 captures, only review a sample (i.e. the first, one in the middle, and the most recent).
    b. Check to see if content/features change significantly between captures. Are these frequent captures necessary? Does older content (such as course schedules or news stories) tend to stay on the site as it is updated? Will a less-frequent capture schedule allow us to preserve the same information?

6. If there is a notable problem with the crawl, identify the underlying cause and document the issue on the QA spreadsheet.
    a. Robots.txt exclusions
    b. Crawl limits (timed out)
    c. Display errors:
    d. Seed redirect
    e. 'Live links'—rendering error
    f. Missing .css files
    g. Resources not in archive (partial)
    h. Seed issues: did not capture (at all)

i. Crawl of unusual size
j. Adjust crawl frequency

7. Make recommendations on the QA Spreadsheet in regards to:
    a. Back up spreadsheet while working on it
    b. The deletion of a previous crawl.
        i. Deletions should be reserved for crawls that were misdirected, erroneous, or never completed (due to robots.txt or technical issues).
        ii. In some cases, excessively large captures (i.e. greater than 4 GB) may need to be deleted to preserve space.
    c. The initiation of a new crawl.
    d. Reducing the crawl frequency of high-priority sites
    e. Communication with the contact owner if it will be necessary to request a modification of the robots.txt file or resolve another issue with the site. Try to identify and record the name/email address of the site's webmaster or main contact.

8. Edit crawl settings:
    a. "Capture Linked Pages"
        i. For U of M content:
            1. Only "high priority" sites should include the capture of linked pages.
            2. For all other sites, the capture linked pages setting should be changed to "**No"** to avoid an excessive amount of content in the web archives**.**
        ii. For MHC content, the QA specialist may need to
    b. If you determine that the web archives need to capture a smaller/wider range of content, make one (or more) of the following changes (and note in the QA Spreadsheet):
        i. Decrease/increase scope (host, directory, or page)

ii. Decrease/increase maximum crawl time (1 or 36 hours)
iii. Recommend the deletion/addition of additional seed URLs on the QA Spreadsheet.

c. While crawl schedules should be accurately set at the time of capture, check with an archivist if the frequency for a site seems too low/high.

Common Issues and Problems with Web Captures

- Crawler traps: These are essentially infinite loops from which a robot is unable to escape. Online calendars are among the most common examples. The crawler will start with the present date and capture page after page of the calendar until the crawl expires without preserving more meaningful site content. The resulting capture may have a very large number of files and will likely reach the maximum time setting before finishing.

- Unexpected seed redirects: The web crawler may be unexpectedly redirected from the target seed URL and begin the crawl on a random page (sometimes completely unassociated with the original seed URL). The redirection may truncate the crawl, cause important content (such as a home page) to be missed, or may lead to a crawler trap.

- Inaccurate seed URLs: Some sites require the crawler to start at a specific web page instead of a basic domain name. For instance, the accurate capture of the U of M Law School required http://www.law.umich.edu/Pages/default.aspx to be included as a seed (instead of just http://www.law.umich.edu/). Other sites will require the crawler to start at ".../home" or "…/index.html." Failure to include accurate seeds may result in a failed crawl, unexpected redirect, or a crawler trap. The BHL QA specialist may need to visit the live website to identify the exact URL from which the crawler should begin.

- Robots.txt files: A "robots.txt" file is an Internet convention used by webmasters to prevent all or certain sections of websites from being captured by a web crawler. The robots.txt must reside in the root of the site's domain and its presence may be verified by typing '/robots.txt' after the root URL (i.e. http://umich.edu/robots.txt). By convention, a web crawler or robot will read the robots.txt file of a target site before doing anything else. This text file will specify what sections of a site the robot is forbidden to crawl. A typical robots.txt exclusion statement is as follows:

  User-agent: *
  Disallow: /

  User-agent' refers to the crawler; * is a wildcard symbol that indicates the exclusion applies to all robots; and / applies the exclusion to all pages on the

DRAFT

How to Accession Electronic Records to the Spartan Archive Storage Vault

1. Receive transfer from unit, including transmittal form and inventory

2. Create Archivists' Toolkit record
    - Assign 'A' accession #
    - Enter accession date (indicates accession created)
    - Link to Resource (MSU unit/record group)

3. Provide accession # to unit

4. Add accession # to transmittal form

5. If transmittal form and inventory are paper, scan as PDFs. If transmittal form and inventory are digital files, print a copy.

6. File paper version of transmittal form in records management files. Inventory should be stapled to transmittal form, if available.

7. If necessary, create a folder in the Storage Vault for the record group. The name of the folder for the record group should be the official UAHC record group number.

For electronic records coming in on hard drives or removable media:

8. Label hard drive or media with accession #. If more than one piece of hardware or media is in the accession, label each with the accession # plus a sequential number. For example, Axxxxxx-1, Axxxxxx-2, etc.

9. Write protect hard drive or media, if possible.

10. Connect hard drive or insert media on electronic records processing workstation.

11. Check for viruses on hard drive or media using the Kaspersky virus scanning utility.
    - Connect hard drive or insert removable media as necessary
    - Open Kaspersky.
    - Select disk to scan.
    - Click "Start Scan" button.
    - If viruses are present, Kaspersky will identify the infected files and ask to quarantine them. Agree to quarantine. (Steps in this case TBD.)

3/15/12

DRAFT

12. Accession files on hard drive or media into the Digital Shelf using Duke Data Accessioner.
    - Open Duke Data Accessioner
    - Under "Adapters" menu, select DROID and JHOVE adapters.
    - Under "Metadata Managers, select Duke PREMIS.
    - Enter your name, the accession number assigned, and the collection number.
    - Click the button labeled "Accession Directory" and select the accession's record group folder in the Storage Vault.
    - Click the disk icon and select the drive or media to be accessioned.
    - Ensure that a logical name is entered into the Disk Name text box. For example, if the accession includes several CDs, the first might be named CD-1, the second CD-2, and so on.
    - Click the "Disk Label" tab. Transcribe any appropriate label text into the box.
    - Click on the "Additional Notes" tab and enter any pertinent information that a processor might need to know about the original disk or the data. For example, file formats found of the preserved files. Any restrictions could be noted here as well.
    - Click the "Migrate" button. This will create a folder labeled with the accession number in the record group folder. The new folder will contain a folder labeled with the assigned disk name containing two files: (1) the contents of the media and (2) an XML file that includes checksums, creation dates, and other metadata for the files on the media.
    - Verify the creation of the new folder and files in the record group folder.
    - Repeat the steps above for each hard drive or media in the accession. Each addition to the accession will result in a new folder containing the contents of that media. Additional XML markup will concatenate in the original XML file.
    - For more on using the Duke Data Accessioner, refer to the Duke University Data Accessioner guide, http://www.duke.edu/~ses44/downloads/guide.pdf.

13. Remove media. Place all media related to the accession in a folder/envelope labeled with date of accession and accession number and store in the electronic records accession file drawer.

14. What to do with hard drive? (TBD)

15. Complete accession record in Archivists' Toolkit
    - Title – Unit ID, Unit Name
    - Extent—in GB
    - Summary (if needed)
    - Date range
    - Location—R Drive and/or Digital Accessions Drawer (for original media)
    - Retention Rule ("Permanent")
    - Description of records. Include information about transfer mechanism, original media, and any viruses in original transfer, if applicable.
    - Link to external document (transmittal form). Use inventory field only if needed

3/15/12

3. University of Virginia: Cheuse Papers Processing Plan

University of Virginia
Processing Plan
Collection 10726, The Papers of Alan Cheuse

Collection Name:	The Papers of Alan Cheuse
Collection Date:	Ca. 1950 – 2009
Collection Number:	10726; accessions _ through al
Extent (pre-processing):	83 disks (3.5" and CD) approx. 5.31 MB; ca. 80 linear feet
Types of materials:	3.5" disks and CDs, video cassettes and DVDs, paper manuscripts
Custodial History:	Alan Cheuse placed the papers on loan to the Library beginning in 1987. Earlier accessions were then purchased in 2003 with a commitment to purchase further groups.
Restrictions from Donors:	Explicit digital rights have yet been discussed. Four series (Accessions 17, 18, 20, and 21) are restricted from access until 2012.
Separated Materials:	Disks have been separated from the manuscript drafts and are stored with the other media and a/v.
Related Materials:	None
Preservation Concerns:	None
Languages other than English:	None
Overview of Contents:	This collection consists of the papers of the American author, book reviewer, and George Mason University professor, Alan Cheuse. These papers include manuscripts for articles, speeches, interviews, and short stories; book reviews; screen plays; cassette tape recordings; computer disks; video cassette & DVD; printed material; contracts and royalties; passports; photographs and drawings; correspondence; research material; short stories by other authors; appointment calendars; short stories and book manuscripts.
Existing Order and description:	Sixteen of the thirty-two accessions have been processed separately, as per institutional practices. They are described in both EAD finding aids and MARC records. They are each organized by type of writing (correspondence, topical files, novel manuscripts, review manuscripts, etc.) to the folder level. The other 16 accessions are recorded in MARC records at varying degrees of detail, some with no more than a title, date, and generic note. All computer media has been separated, numbered, and is referenced in finding aids and records, but has mostly not been processed. The contents of some disks were printed and filed with paper manuscripts. Seven of the accessions contain computer disk materials. Only one of these accessions has been described in an EAD finding aid.

Desired Processing:	All computer media should be processed. Additionally, all accessions should be combined into a single finding aid. Where EAD exists, these records will be combined into a single <archdesc> and <dsc> with each accession being represented as a series. The accessions represented by MARC records will be converted to series components. In addition, subject headings, which were not included in the original EAD, should be added from all MARC records. No further work will be done with paper materials at this time. The processor will create disk images of the disks and then process using FTK. Disks containing commercial works that were used for research purposes should not be imaged or stored at this time. Individual files will be labeled with the disk number so that they may later be associated with the correct container element in the EAD. Titles of individual works will be added to the finding aid so that some reference to the works available on the disks is present. This is to match the level of processing of the paper manuscripts, which are indicated by name within the collection descriptions. Files containing confidential information will be completely restricted at this time. Obsolete file formats will not be migrated at this time, but this work should be considered in the future. Access to materials on the disk will be at the individual file level. After imaging the disk a copy of the image will be transferred to the StoreNext preservation store. Copies of the unrestricted files will be added to the Hypatia repository for public access. The disk images will be referenced by identifier number within the ead. They will exist as individual subcomponents of the accession or sub-series (if it exists) and the disk number will be referenced in a "unitid" attribute. The finalized finding aid will also be uploaded into the Hypatia repository and the individual files will be linked to the accession or container they belong to.
Next steps	Reprocessing all accessions into one collection arranged intellectually, rather than intellectually within individual accessions, is recommended for the future when the collection is deemed "complete." As technology and infrastructure develop, migration of obsolete formats and redaction within restricted files in order to make them available should also be undertaken.
Notes to Processors:	Examine the contents of the CDs later in the series to determine which are simply copies of commercially produced works and do not need to be imaged.
Anticipated Time for Processing:	5 days

4. Yale University: Tobin Collection Processing Plan

Processing Work Plan

Institution: MSSA
Archivist: Mark A. Matienzo
Date: June 7, 2011
Collection title: James Tobin papers
Creator: Tobin, James
Current call number(s): MS 1746, **Accession 2004-M-088**
Provenance: Gift of Elizabeth Tobin, 2004.
Extent: 8.75 linear feet; 27 3.5" inch diskettes (35.7 MB)

Overview:
Research strengths: correspondence regarding professional activities; working and final drafts of conference papers, periodical columns, and other publications.

Types of electronic records present: Correspondence (e-mail and computer-written letters); writings; spreadsheets and graphs; office files (biographical statements, calendars, publication lists, etc.), course materials. Files are primarily WordPerfect and Lotus 1-2-3; some Quicken files exist; e-mail is in text form, either in Eudora mailboxes individually saved text files.

Significant preservation concerns: See file formats above. Most significant concern is Lotus 1-2-3 files; several should be considered compound objects with graphs and formatting information.

Description:

Current: Minimal. Labels from individual diskettes have been transcribed as component titles within finding aid.

Proposed enhancement: Description should follow executed organization as specified below.

Recommended description work for later: see under organization.

Organization:

Current: Hard to determine. Paper records do not seem to have a coherent overall organization, with the exception of the correspondence; however, correspondence is still scattered between "Letters to Jim," "Professional Correspondence", "Nobel Prize Correspondence," and "Personal Correspondence." Writings are very disorganized;

Diskettes appear to be used as transfer media for files between his office, his home, and his cottage in Wisconsin. A few disks, or sets thereof, show some grouping based on type of records, such as "office files" (publication lists, telephone lists/address books) and letters that Tobin wrote in WordPerfect. Writings are not grouped together thematically.

Proposed arrangement: Arrangement should be based on record types. Within the electronic records for this accession, logical groupings and subgroupings are as follows:

- Correspondence, 1992-2001 and undated
    - Correspondence written using WordPerfect, 1992-2000
    - E-mail, 1996-2001 and undated
- Course materials for Economics 480B, 1998
    - Lotus 1-2-3 spreadsheets, 1992-1997

  - "Primer" spreadsheets and graphs, 1996-1997
- Office files, 1995-2001
  - Biographical statements
  - Calendars
  - Lists of Tobin's publications
  - Quicken files
  - Recommendation letters and lists of recommendations
  - Telephone lists
- Writings, 1992-2001

Of all groupings, the Writings grouping would need the most considerable organization and description. In the short term I recommend either not listing individual files, or listing individual files with filename and date only.

Recommended arrangement work for later: Combine paper records and electronic records into a common arrangement. Considerable attention to Tobin's personal papers is needed, especially those related to his military service. Arrange writings alphabetically by title, identify explicit drafts, and reconcile against publication lists included in this accession as available from the Cowles Foundation. In the long term, we should plan to process the collection as a whole and integrate all the accessions into a common arrangement.

Appraisal:

Diskettes 1-3, 11, and 17 should be discarded; #1-3 contain printer drivers; #11 contains modem software; and #17 contains many deleted files and is mostly blank.

Some of Tobin's "office files" are of uncertain or low research value, such as the Quicken files, biographical statements and telephone lists. The publication lists are of questionable value as the Cowles Foundation has a detailed publication list in PDF form; however, Tobin has some topic-specific publication lists that may be helpful. Some of the office files also appear to be inventories of paper files, which may or may not be reflected in the paper records previously acquired.

Restrictions:

Other (paper) correspondence within this accession is restricted. E-mail contains both personal and professional correspondence; personal/family correspondence includes reference to health issues. Consider restricting e-mail under similar conditions. Most letters written using WordPerfect are professional in nature. Recommendation letters and Quicken files (which deal with Tobin's personal finances) should be restricted.

Preservation:

Proposed action now: Investigate migration options for Lotus 1-2-3 files, particularly those that reference graphs.

Recommended for later: Migrate WordPerfect files to PDF/A; migrate e-mail to a different format.

Access:

See Preservation. Files should be extracted into a storage option such as the YUL Rescue Repository so they can be paged on request. This collections does not have a high level use, so there is probably not an immediate need to create use copies.

Beinecke Rare Book & Manuscript Unit, Manuscript Unit, Processing Manual section on Electronic Files

5.6 Electronic Files

Computer media containing electronic or born digital files are sometimes found in manuscript collections and, like other collection material, should be accounted for in the arrangement and description of the archive.

Disks and other media are logged and pulled when a collection is accessioned and acknowledged in the AT Accession module. The content is captured for preservation, appraisal, and access, and the original media is returned to the collection and placed in Restricted Fragile Papers.

5.6.1 Security & Access

MS Unit and selected Beinecke staff members have access to use copies of disk images on a YU network directory.

Library guidelines for research use of eletronic files in manuscript collections are posted on the Beinecke website under Research Services at Ordering Copies / Photographs / Scans.

5.6.2 Collection Development

Library guidelines for collecting born digital manuscript material are maintained on the network directory under Curatorial\YCAL\Born Digital Docs.

5.6.3 Accessioning

Accessioning of computer media is defined by the library as capture of the content off the source media. Computer media should be removed from manuscript collections upon receipt or during baseline processing of new accessions in order to be

accessioned. The Manuscript Unit pursues a strategy of bit-level capture through disk imaging.

Documentation on the accessioning (i.e. disk imaging) of each piece of media is captured on an “Electronic Records Media Log”. The logs are maintained by accession number on the department’s Accessioning webpages at https://collaborate.library.yale.edu/BeineckeLibrary/MsUnit/accessioning/Lists/Electronic%20Records%20Media%20Log/AllItems.aspx.

Additional documentation on the "Electronic Files Workflow for New Accessions" and "Electronic File Log" can be found on the department's Accessioning webpages.

5.6.4 Appraisal

There are tools for appraising/analyzing content on disk images and electronic files. For appraising or analyzing content of files in disk images, commercial forensic tools (FTK Imager and AccessData FTK) are available. Consult the appropriate staff member regarding use of these tools in planning for processing. For appraising/analyzing the content of electronic files, the library has file viewing software (Quick View Plus) on some staff workstations, public workstations, and laptops. See the Quick View Plus website for comprehensive list of file types supported by the current version of the viewer.

5.6.5 Intellectual Arrangement

General note

When computer media is found in a collection it should be routed into the computer media accessioning workflow--see step 2 of the "Electronic Files Workflow for New Accessions".

When we receive computer media for which we have the technical infrastructure in place in the digital preservation lab to accession it, we will attempt to accession it in time for staff working on the paper component of the collection to analyze the records contained on the media and possibly integrate them into the collection. This will depend on various factors, including the volume of media in the accession and staff availability. This may enable staff to complete processing for some collections.

Because baseline processing of new accessions was implemented prior to disk imaging, collections dating from roughly 2008-2011 were processed before the policy above was in place. The result in most cases is that media was routed into the computer media accessioning workflow and documented in the finding aid only (as media and not records) in Restricted Fragile Papers. This represents a group of collections for which additional processing should be done in order to integrate the born digital content.

In baseline processing, staff should first consult the accessioning and baseline project documentation to determine if selected projects contain computer media. In the ACQ record, see the TTL, MAT, and LNO fields, and in the backlog files, see the Notes field. If collections contain computer media, staff should then consult the "Electronic Records Media Log" documenting accessions and/or contact the appropriate staff person to determine if the computer media has been fully accessioned and the records can be appraised/analyzed. If born digital materials are ready for processing, staff can consult about documentation, tools, and strategies.

5.6.5.1 Computer Disks

Most electronic files in manuscript collections accessioned before 2008 came on the standard data storage devices in use since the mid 1970s: 5 ¼ and 3 ½ inch disks, zip disks, and compact discs (CDs). When evaluating files on these media formats, the following instructions may best apply.

The number of disks and electronic files in a collection may determine whether you can conduct item-level analysis. Most files on these media formats include drafts of writings

or material relating to writing projects and correspondence (in word processing formats). When possible, respect context and original order in arrangement. When original order cannot be established, in general, small numbers of disks and files lend themselves to item or file-level analysis and arrangement by content. With larger numbers of disks and files, and disks with mixed files (e.g. writings, correspondence, etc.), other factors will probably also need to to considered in order to determine whether to arrange material by content or format. In baseline processing, media may also be listed where found (disks should be housed in Restricted Fragile).

As of December 2011, several collections containing computer media have been processed to varying levels, providing us with some useful examples:

For an example of a hybrid collection in which the electronic and paper materials were fully integrated and arranged to the file/item level, see the James Welch Papers (YCAL MSS 248).

For a baseline processing project example of a collection containing a moderate number of disks (33) in which some analysis of the content allowed the born digital and paper material to be integrated and arranged at the file level, see the Caryl Phillips Papers (GEN MSS 793).

For a baseline project example of a collection containing a smaller number of disks (22) in which context alone allowed for arrangement at the subseries/file level, see the Howard Roberts Lamar Papers (WA MSS S-2639).

For an example of a collection in which the electronic files were arranged by format, see the George Whitmore Papers (YCAL MSS 274).

One way to keep track of electronic files when doing item-level arrangement is to create a dummy folder, labeled with information about the file, and incorporate the folder into the sorting of like material. For example, when arranging material for a particular title

in a writing series, place a dummy folder for an electronic draft (see “Hotel Christobel” example in section 5.6.7.6) in the sequence of materials relating to the title.

Other types of text files can be treated the same way, placing them in the appropriate intellectual and sequential location of related files.

5.6.5.2 Snapshot Accessions, Computers, External Hard-drives

When dealing with digital records acquired directly from record creators through snapshot accessions or on retired media, such as computers (and possibly external hard drives), respect context and original order as recommended in the "Paradigm Exemplars for Arrangment," *Workbook of Digital Private Papers*, available at http://www.paradigm.ac.uk/workbook/cataloguing/ead-exemplars.html.

5.6.5.3 Special Cases

Some electronic files may not lend themselves to the management and access strategies outlined above. In these cases, other strategies may be desirable or necessary to provide staff and research access to the files.

For difficult-to-access files, files prone to corruption, and relational files, it might be preferable to print a copy of the file, rather than rely on the electronic copy, for reference and research use. These copies would go into the archival boxes just as Preservation Photocopies do, and would be clearly marked as printouts from electronic files.

For relational files, such as databases and hyperlinked documents, it may be better to recreate a mini-environment with the original software. For example, a suite of web pages could be copied to a folder that also contains a simple version of an HTML browser. Or a database file could be coupled with a viewing version of the database program.

For graphic files, Quick View Plus and other file viewers can open and display most types of images formats. Dynamic image data (e.g., motion picture files), however, will need to be viewed on software that can properly sequence them.

For batch files that we might describe at a finer level (e.g. Eudora e-mail folders containing e-mail from numerous correspondents, accessible in the original Eudora software), the access methods could take two forms: Arrange the file at the end of the Correspondence series as a general correspondence file (e.g. "Work Letters 1997") and include important names in a note. Use the original software, if available, to access the individual components, print them out, and file them as you would paper-based correspondence. Printouts must be marked to show that they are copies of material received in electronic form.

5.6.6 Physical Arrangement

Computer media should be placed in Restricted Fragile.

5.6.7 Description

In the finding aid, the existence, quantity, technical specifications and requirements, and conservation relating to computer media and electronic files can be described in the following EAD elements: Physical Description, Description of the Papers, Information About Access, and Notes.

5.6.7.1 Physical Description <extent>

The extent of computer media and/or electronic files may be documented at the collection, series/accesion, and folder level as appropriate.

When a collection or series/accession consists solely of digital records, record extent in terms of file storage size and, in some cases, number of files. Though *DACS* does not offer an example of digital extent recorded in terms of size, the general rule at 2.5.3

seems to allow for it. See also *RAD* 9.5B2, *ISAD(G)* 3.1.5 and the Paradigm fonds-level description recommendations, available at http://www.paradigm.ac.uk./workbook/cataloguing/ead-fonds.html. As of April 2010, recent professional practice and recommendations indicate use of gigabytes and megabytes. That said, use the most appropriate file storage size per *RAD* 9.5B2. For example:

Physical Description: 3.71 megabytes

In accordance with *DACS* 2.5.7, extent may be further defined through a parallel statement. This could be used to record a large number of files. For example:

Physical Description: 227 megabytes (2,215 files)

Alternately, when the file storage size is not available, describe the quantity in terms of material type(s) in accordance with *DACS* 2.5.5. See also *RAD* 9.5B3. This will be the case when some or all formats are unreadable or, in baseline processing, if media has not yet been fully accessioned. For example:

Physical Description: 57 computer disks

Similarly, in baseline processing, when the file storage size is not yet available, qualify the statement to highlight the existence of the material type in accordance with *DACS* 2.5.6. For example:

Physical Description: 7 folders, including 3 computer disks

EAD allows for multiples statements of extent. When the digital records make up a significant part of a hybrid collection or series/accession, provide two parallel expressions of extent, one for the physical content and one for the digital content. For example:

Physical Description: 4.17' (10 boxes)

Physical Description: 227 megabytes (2,215 files)

5.6.7.2 Description of the Papers <scopecontent>

The existence of computer media or electronic files can be noted in the Description of the Papers. Otherwise, if electronic files are arranged at the series level, this can be discussed in the series scope and content note.

If electronic files have been printed out, rather than left in electronic form, this should be noted. If they have been printed out because the electronic file was damaged or otherwise problematic, be sure to note that the file was "salvaged" from the electronic version. If some files are printed out and others are left in electronic form, provide the rationale for this decision.

5.6.7.3 Information about Access <accessrestrict> and <phystech>

Access restrictions on original media and files should be noted in the Access Restrict element in accordance with *DACS* 4.2. Use the following format: [Container type] [number or span] ([type of media]): Restricted Fragile Material. Reference copies [may be requested/are available]. Consult Access Services for further information. For example:

Box 4 (computer disks): Restricted Fragile Material. Reference copies of electronic files may be requested. Consult Access Services for further information.

Box 14 (laptop computer): Restricted Fragile Material. Reference copies of electronic files are available. Consult Access Services for further information.

Technical requirements for patron access to copies are meant to be noted in the Physical Characteristics/Technical Requirements element in accordance with *DACS* 4.3.5 but, because the this element is rarely used in YUL finding aids, this information will be added to the access restriction in the Access Restrict element if appropriate.

5.6.7.4 Notes <notes>

Preservation actions that results in changes to the file, such as migration, should be documented in a note element in accordance with *DACS* 7.1.4 See also *RAD* 9.8B10b. For example:

> Electronic files migrated by National Data Conversion from the original word-processing software (WordStar for CP/M) to Wordstar 4.0 for DOS and to ASCII to maintain readability of data. Technical specifications are filed with media in Restricted Fragile.

Expanding on *DACS* 7.1.4, in an effort to be more transparent about the reproduction process, document refreshment or ingest into the local digital repository. For example:

Electronic files migrated by National Data Conversion from the original word-processing software (WordStar for CP/M) to Wordstar 4.0 for DOS and to ASCII to maintain readability of data. Wordstar 4.0 for DOS and ASCII files refreshed into the Yale University Library Rescue Repository. Technical specifications are filed with media in Restricted Fragile.

5.6.7.5 Series and Subseries Headings

Local practice is to apply the term "Electronic Files" to series and subseries headings. Electronic Files is preferred to Computer Files, the *AACR2* GMD (*AACR2* 1.1C1), as a broader and ostensibly more accurate term, one, for example, that can encompass electronic or born-digital files created on contemporary portable devices (such as digital cameras, cell phones, PDAs, etc.) not commonly identified as computers. Electronic Files is preferred to Electronic Records in order to distinguish materials created or received by individuals common to personal papers from records created or received in the course of institutional activity. Electronic is also preferred to Digital as a broader term, encompassing both analog and digital formats.

At this time Beinecke does not apply headings by specific format (e.g. text files, image files).

See George Whitmore Papers (YCAL MSS 274).

5.6.7.6 Folder Headings and Folder Notes

The recommended chief source of information for electronic files is the title screen (*AACR2* 9.0B1). Transcribe the title screen of the file when applying item-level analysis and arrangement. Other prescribed sources of information include the physical carriers or labels. When applying disk-level analysis, transcribe information from the physical carrier (e.g. disk or jewel case) or label. See the George Whitmore Papers (YCAL MSS 274).

When transcribing or supplying folder headings for files arranged at the item level, such as a draft, add the term "electronic," as you would the GMD. When electronic files are arranged intellectually, outside of an "Electronic Files" series/accession, always include the following folder note in an Access Restrict element <accessrestrict> in accordance with *DACS 4*.2:

Computer disks are restricted. Copies of electronic files may be requested through Access Services:[Accession #, Disk #, Disk label]

For example:

Series I. Writings

PLAYS

"Hotel Christobel"

4	21	Research notes	1990

22 Preliminary sketches 1990 Oct 1

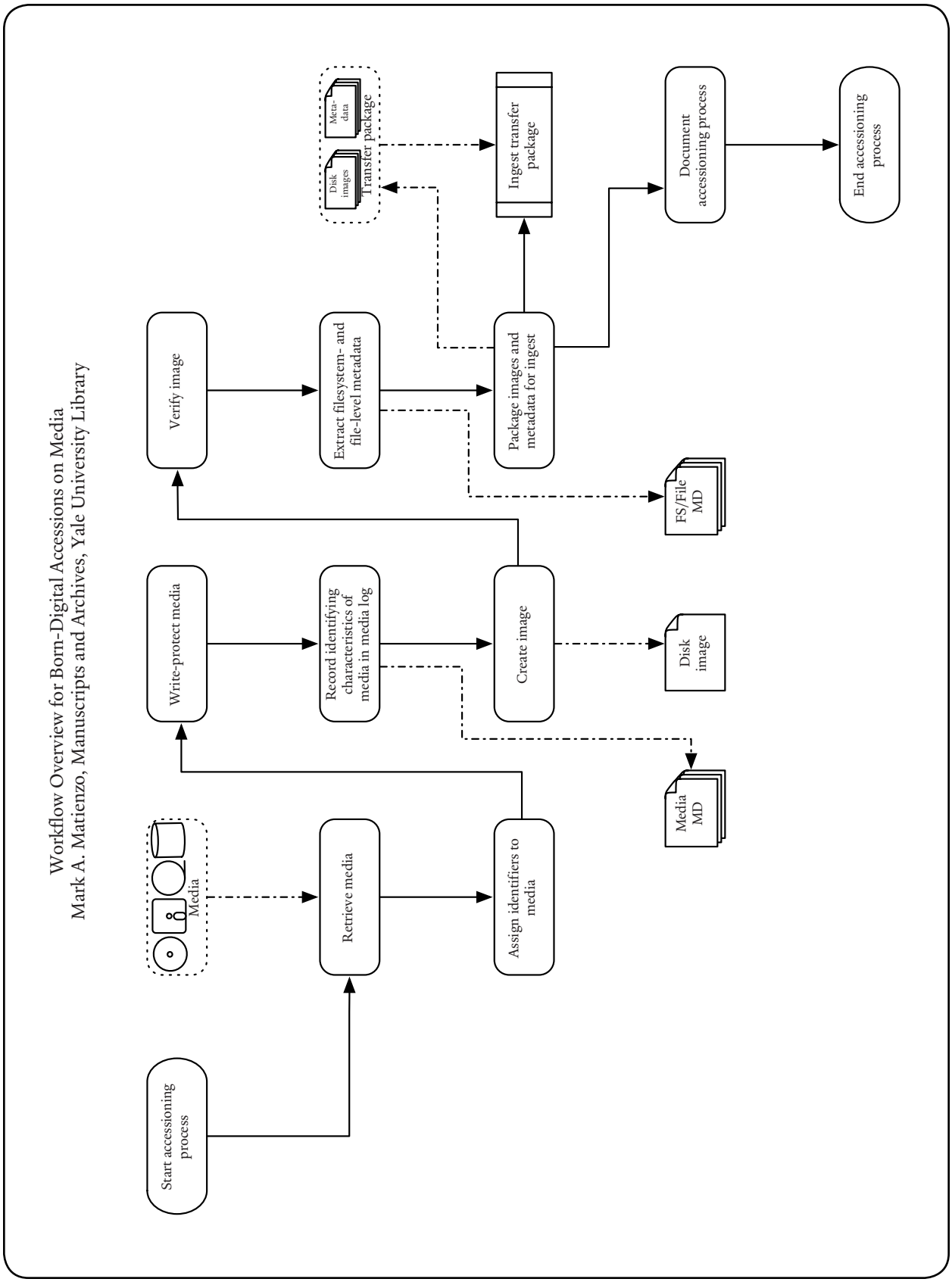
Draft, electronic 1990 Nov

Computer disks are restricted. Copies of electronic files may be requested through Access Services: [Accessions #], Disk#17, Hotel.doc

23 Galley proof 1990 Dec

See James Welch Papers (YCAL MSS 248).

Item-level description might also include the original file format.


Workflow Overview for Born-Digital Accessions on Media
Mark A. Matienzo, Manuscripts and Archives, Yale University Library
Start accessioning process
Media
Retrieve media
Assign identifiers to media
Write-protect media
Record identifying characteristics of media in media log
Create image
Verify image
Extract filesystem- and file-level metadata
Package images and metadata for ingest
Transfer package
Disk images
Meta-data
Ingest transfer package
Document accessioning process
End accessioning process
Media MD
Disk image
FS/File MD