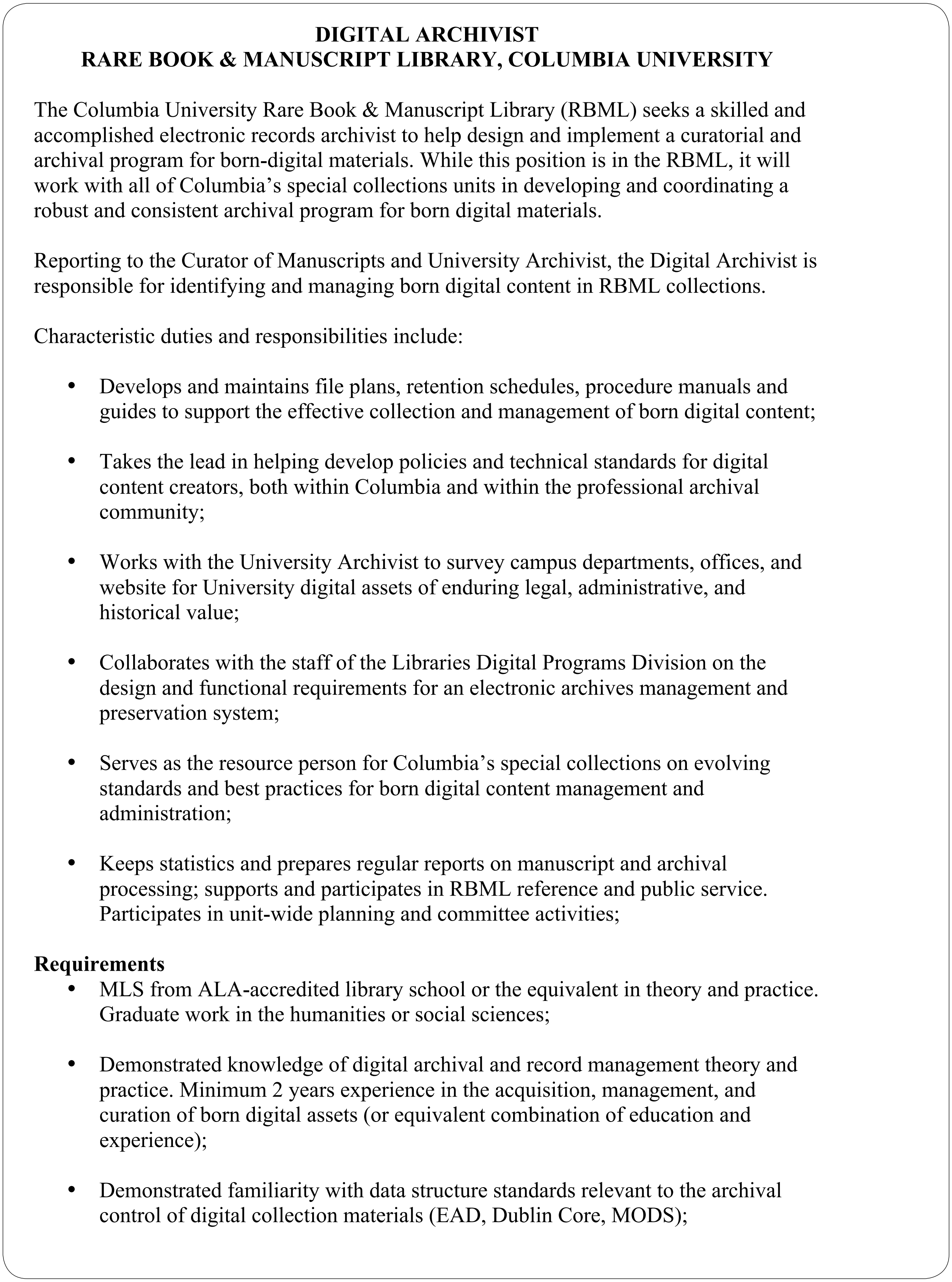
REPRESENTATIVE DOCUMENTS

Job Descriptions



DIGITAL ARCHIVIST
RARE BOOK & MANUSCRIPT LIBRARY, COLUMBIA UNIVERSITY

The Columbia University Rare Book & Manuscript Library (RBML) seeks a skilled and accomplished electronic records archivist to help design and implement a curatorial and archival program for born-digital materials. While this position is in the RBML, it will work with all of Columbia's special collections units in developing and coordinating a robust and consistent archival program for born digital materials.

Reporting to the Curator of Manuscripts and University Archivist, the Digital Archivist is responsible for identifying and managing born digital content in RBML collections.

Characteristic duties and responsibilities include:

- Develops and maintains file plans, retention schedules, procedure manuals and guides to support the effective collection and management of born digital content;

- Takes the lead in helping develop policies and technical standards for digital content creators, both within Columbia and within the professional archival community;

- Works with the University Archivist to survey campus departments, offices, and website for University digital assets of enduring legal, administrative, and historical value;

- Collaborates with the staff of the Libraries Digital Programs Division on the design and functional requirements for an electronic archives management and preservation system;

- Serves as the resource person for Columbia's special collections on evolving standards and best practices for born digital content management and administration;

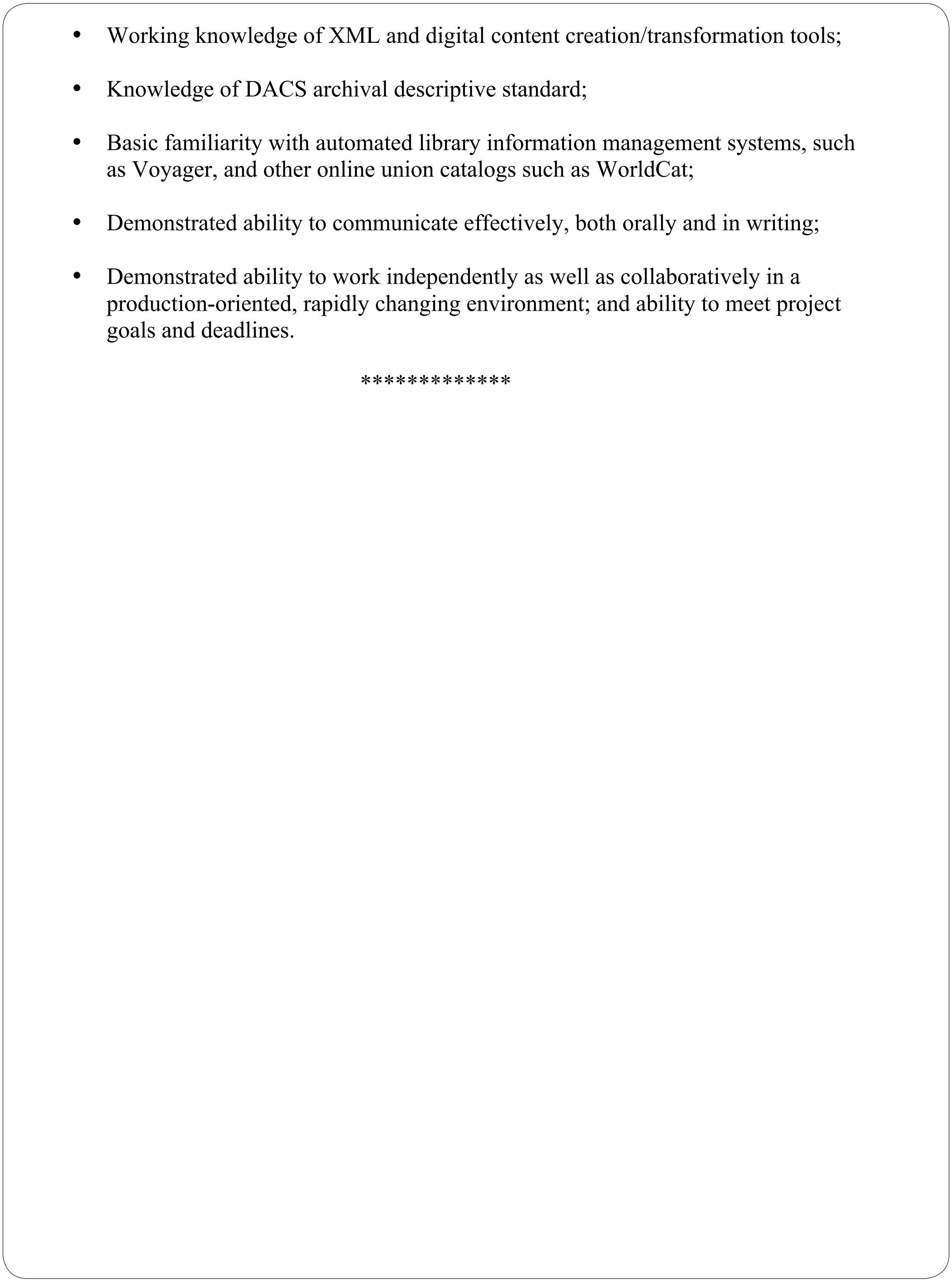
- Keeps statistics and prepares regular reports on manuscript and archival processing; supports and participates in RBML reference and public service. Participates in unit-wide planning and committee activities;

Requirements

- MLS from ALA-accredited library school or the equivalent in theory and practice. Graduate work in the humanities or social sciences;

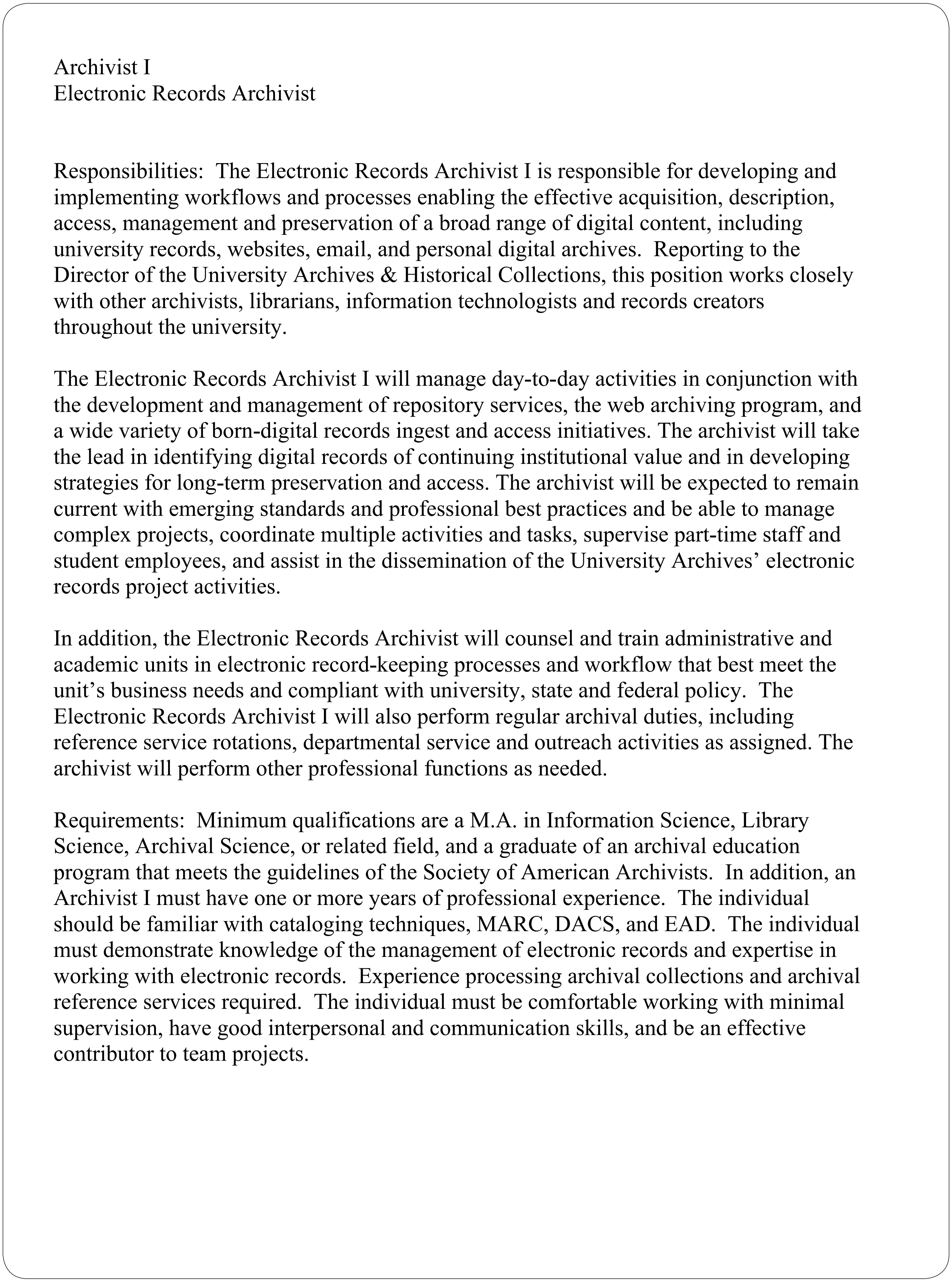
- Demonstrated knowledge of digital archival and record management theory and practice. Minimum 2 years experience in the acquisition, management, and curation of born digital assets (or equivalent combination of education and experience);

- Demonstrated familiarity with data structure standards relevant to the archival control of digital collection materials (EAD, Dublin Core, MODS);



- Working knowledge of XML and digital content creation/transformation tools;
- Knowledge of DACS archival descriptive standard;
- Basic familiarity with automated library information management systems, such as Voyager, and other online union catalogs such as WorldCat;
- Demonstrated ability to communicate effectively, both orally and in writing;
- Demonstrated ability to work independently as well as collaboratively in a production-oriented, rapidly changing environment; and ability to meet project goals and deadlines.

****



Archivist I
Electronic Records Archivist

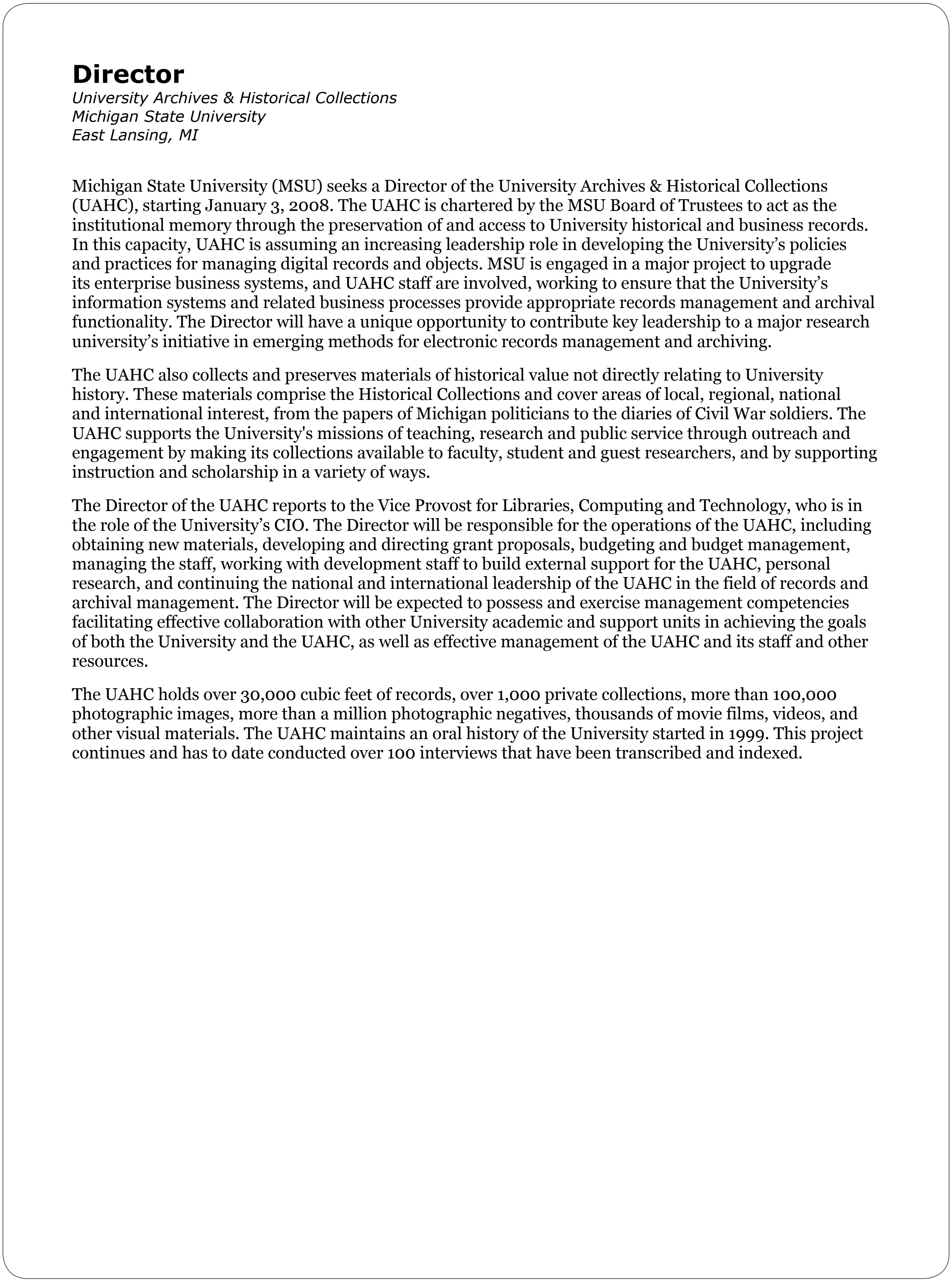
Responsibilities: The Electronic Records Archivist I is responsible for developing and implementing workflows and processes enabling the effective acquisition, description, access, management and preservation of a broad range of digital content, including university records, websites, email, and personal digital archives. Reporting to the Director of the University Archives & Historical Collections, this position works closely with other archivists, librarians, information technologists and records creators throughout the university.

The Electronic Records Archivist I will manage day-to-day activities in conjunction with the development and management of repository services, the web archiving program, and a wide variety of born-digital records ingest and access initiatives. The archivist will take the lead in identifying digital records of continuing institutional value and in developing strategies for long-term preservation and access. The archivist will be expected to remain current with emerging standards and professional best practices and be able to manage complex projects, coordinate multiple activities and tasks, supervise part-time staff and student employees, and assist in the dissemination of the University Archives' electronic records project activities.

In addition, the Electronic Records Archivist will counsel and train administrative and academic units in electronic record-keeping processes and workflow that best meet the unit's business needs and compliant with university, state and federal policy. The Electronic Records Archivist I will also perform regular archival duties, including reference service rotations, departmental service and outreach activities as assigned. The archivist will perform other professional functions as needed.

Requirements: Minimum qualifications are a M.A. in Information Science, Library Science, Archival Science, or related field, and a graduate of an archival education program that meets the guidelines of the Society of American Archivists. In addition, an Archivist I must have one or more years of professional experience. The individual should be familiar with cataloging techniques, MARC, DACS, and EAD. The individual must demonstrate knowledge of the management of electronic records and expertise in working with electronic records. Experience processing archival collections and archival reference services required. The individual must be comfortable working with minimal supervision, have good interpersonal and communication skills, and be an effective contributor to team projects.

MICHIGAN STATE UNIVERSITY
Director, University Archives & Historical Collections



Director

University Archives & Historical Collections
Michigan State University
East Lansing, MI

Michigan State University (MSU) seeks a Director of the University Archives & Historical Collections (UAHC), starting January 3, 2008. The UAHC is chartered by the MSU Board of Trustees to act as the institutional memory through the preservation of and access to University historical and business records. In this capacity, UAHC is assuming an increasing leadership role in developing the University's policies and practices for managing digital records and objects. MSU is engaged in a major project to upgrade its enterprise business systems, and UAHC staff are involved, working to ensure that the University's information systems and related business processes provide appropriate records management and archival functionality. The Director will have a unique opportunity to contribute key leadership to a major research university's initiative in emerging methods for electronic records management and archiving.

The UAHC also collects and preserves materials of historical value not directly relating to University history. These materials comprise the Historical Collections and cover areas of local, regional, national and international interest, from the papers of Michigan politicians to the diaries of Civil War soldiers. The UAHC supports the University's missions of teaching, research and public service through outreach and engagement by making its collections available to faculty, student and guest researchers, and by supporting instruction and scholarship in a variety of ways.

The Director of the UAHC reports to the Vice Provost for Libraries, Computing and Technology, who is in the role of the University's CIO. The Director will be responsible for the operations of the UAHC, including obtaining new materials, developing and directing grant proposals, budgeting and budget management, managing the staff, working with development staff to build external support for the UAHC, personal research, and continuing the national and international leadership of the UAHC in the field of records and archival management. The Director will be expected to possess and exercise management competencies facilitating effective collaboration with other University academic and support units in achieving the goals of both the University and the UAHC, as well as effective management of the UAHC and its staff and other resources.

The UAHC holds over 30,000 cubic feet of records, over 1,000 private collections, more than 100,000 photographic images, more than a million photographic negatives, thousands of movie films, videos, and other visual materials. The UAHC maintains an oral history of the University started in 1999. This project continues and has to date conducted over 100 interviews that have been transcribed and indexed.

PENNSYLVANIA STATE UNIVERSITY

Digital Records Archivist

Digital Records Archivist

The Pennsylvania State University Libraries seek applications and nominations for the position of Digital Records Archivist. The person appointed to this tenure-track, faculty position will manage the Eberly Family Special Collections Library's existing born digital archival holdings and expand its capacity to collect electronic records with the initial effort focused on university records.

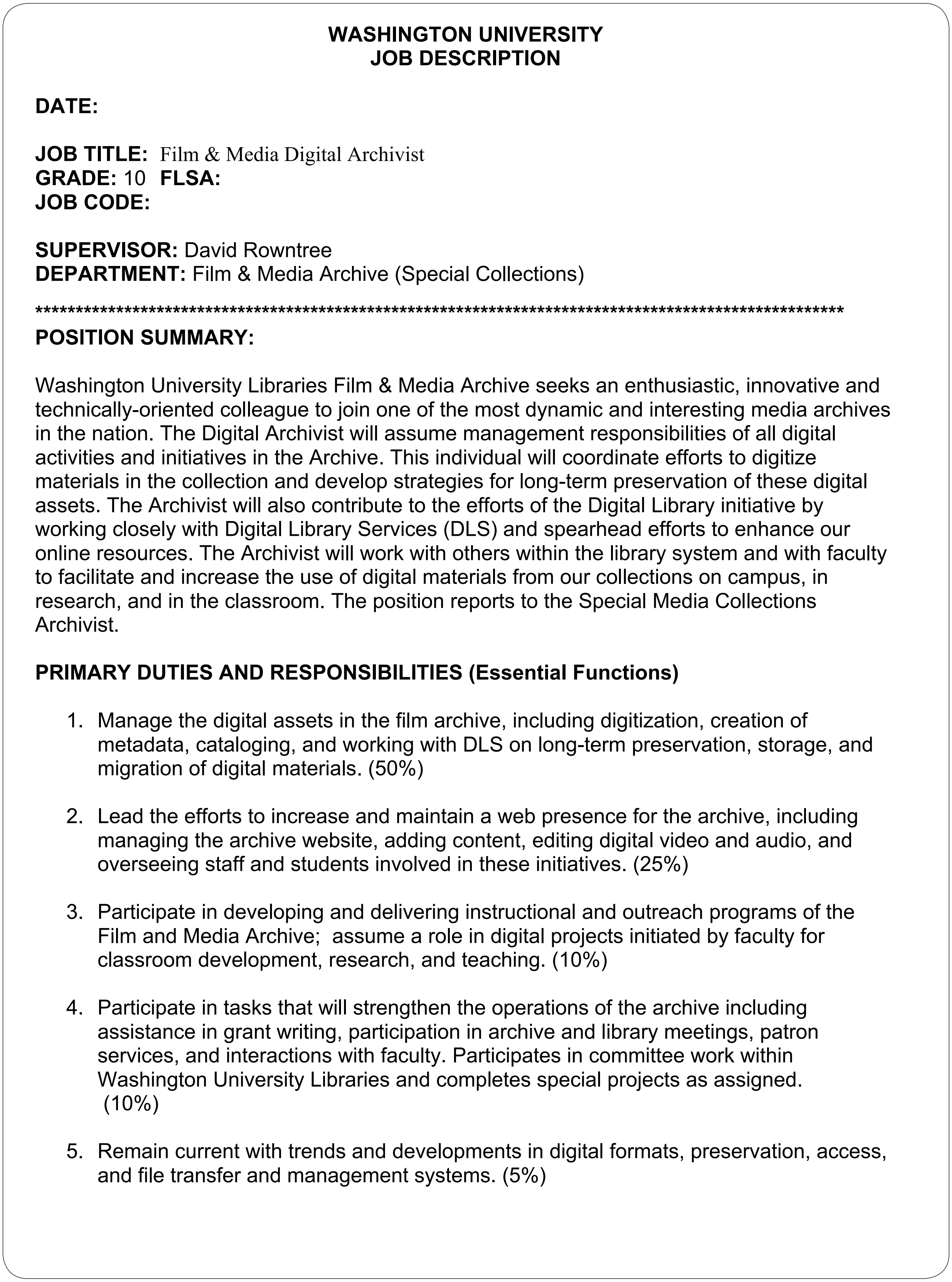
The Eberly Family Special Collections Library at University Park comprises three units: Historical Collections and Labor Archives, Rare Books and Manuscripts, and University Archives and Records Management, together including a total of 18 full-time faculty and staff. The University Archives oversees the University Records Management Program and Inactive Records Center, an extensive sports archives, photograph and audio-visual collections, as well as Fred Waring's America. More information about all special collections in the University Libraries is available online at http://www.libraries.psu.edu/psul/speccolls.html.

Responsibilities:
The Digital Records Archivist will help develop and implement workflows and processes enabling the effective acquisition, description, access, management and preservation of a broad range of digital content, including university records, websites, email, and personal digital archives. This position reports to the Head of the Eberly Family Special Collections Library and works closely with the University Archivist, other archival professionals, librarians, information technologists, and records creators throughout the University. The archivist will manage day-to-day activities in conjunction with the development and management of repository services, the web archiving program, and a wide variety of born-digital records ingest and access initiatives. The archivist will take the lead in identifying digital records of continuing institutional value and in developing strategies for long-term preservation and access. The archivist will be expected to remain current with emerging standards and professional best practices and be able to manage complex projects, coordinate multiple activities and tasks, supervise part-time and student employees, and curate electronic records and digital collections throughout the information lifecycle. The archivist will also perform regular archival duties, including reference service and outreach activities, and assist in the dissemination of best practices, trend reports, and operational guidelines. The archivist will perform other professional functions as needed.

Requirements:
Minimum qualifications are a MLS/MLIS from an ALA-accredited program (or equivalent), or a Masters in Information Science, Archival Science, or related field. Experience working with the curation of digital content in an archival repository. Familiarity with descriptive and data structure metadata standards such as MARC, DAS, EAD, Dublin Core, METS, MODS, and PREMIS. Familiarity with tools and workflows being developed to support the ingest and management of born digital records. Demonstrated knowledge of the management, preservation, and access of electronic records, and expertise in working with electronic records. Demonstrated knowledge of data storage methods, media, security, content management, and access. The candidate must have excellent analytical, interpersonal and communication skills, be an effective team contributor, have proven ability to manage projects and competing priorities with demonstrated ability to be flexible, to adapt to change, and to work successfully in a fast-paced, dynamic environment.

Preferred: Experience processing archival collections and providing archival reference services; experience working with tools that verify file authenticity, search for personal identity information, and harvest websites; programming/scripting skills in languages such as Java, PERL, and XSLT.

WASHINGTON UNIVERSITY IN ST. LOUIS
Film & Media Digital Archivist



WASHINGTON UNIVERSITY
JOB DESCRIPTION

DATE:

JOB TITLE: Film & Media Digital Archivist
GRADE: 10 **FLSA:**
JOB CODE:

SUPERVISOR: David Rowntree
DEPARTMENT: Film & Media Archive (Special Collections)

****

POSITION SUMMARY:

Washington University Libraries Film & Media Archive seeks an enthusiastic, innovative and technically-oriented colleague to join one of the most dynamic and interesting media archives in the nation. The Digital Archivist will assume management responsibilities of all digital activities and initiatives in the Archive. This individual will coordinate efforts to digitize materials in the collection and develop strategies for long-term preservation of these digital assets. The Archivist will also contribute to the efforts of the Digital Library initiative by working closely with Digital Library Services (DLS) and spearhead efforts to enhance our online resources. The Archivist will work with others within the library system and with faculty to facilitate and increase the use of digital materials from our collections on campus, in research, and in the classroom. The position reports to the Special Media Collections Archivist.

PRIMARY DUTIES AND RESPONSIBILITIES (Essential Functions)

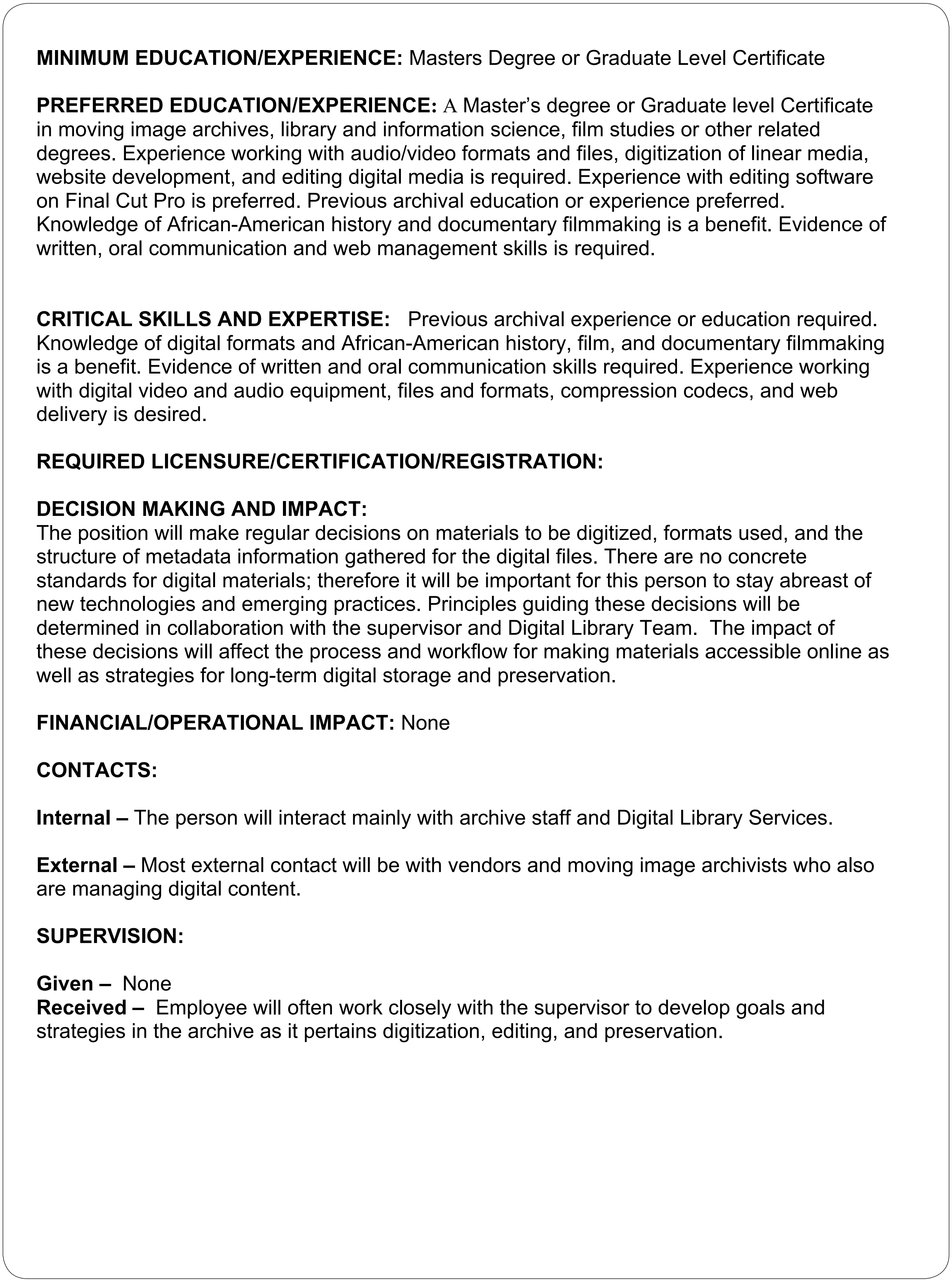
1. Manage the digital assets in the film archive, including digitization, creation of metadata, cataloging, and working with DLS on long-term preservation, storage, and migration of digital materials. (50%)

2. Lead the efforts to increase and maintain a web presence for the archive, including managing the archive website, adding content, editing digital video and audio, and overseeing staff and students involved in these initiatives. (25%)

3. Participate in developing and delivering instructional and outreach programs of the Film and Media Archive; assume a role in digital projects initiated by faculty for classroom development, research, and teaching. (10%)

4. Participate in tasks that will strengthen the operations of the archive including assistance in grant writing, participation in archive and library meetings, patron services, and interactions with faculty. Participates in committee work within Washington University Libraries and completes special projects as assigned. (10%)

5. Remain current with trends and developments in digital formats, preservation, access, and file transfer and management systems. (5%)



MINIMUM EDUCATION/EXPERIENCE: Masters Degree or Graduate Level Certificate

PREFERRED EDUCATION/EXPERIENCE: A Master's degree or Graduate level Certificate in moving image archives, library and information science, film studies or other related degrees. Experience working with audio/video formats and files, digitization of linear media, website development, and editing digital media is required. Experience with editing software on Final Cut Pro is preferred. Previous archival education or experience preferred. Knowledge of African-American history and documentary filmmaking is a benefit. Evidence of written, oral communication and web management skills is required.

CRITICAL SKILLS AND EXPERTISE: Previous archival experience or education required. Knowledge of digital formats and African-American history, film, and documentary filmmaking is a benefit. Evidence of written and oral communication skills required. Experience working with digital video and audio equipment, files and formats, compression codecs, and web delivery is desired.

REQUIRED LICENSURE/CERTIFICATION/REGISTRATION:

DECISION MAKING AND IMPACT:
The position will make regular decisions on materials to be digitized, formats used, and the structure of metadata information gathered for the digital files. There are no concrete standards for digital materials; therefore it will be important for this person to stay abreast of new technologies and emerging practices. Principles guiding these decisions will be determined in collaboration with the supervisor and Digital Library Team. The impact of these decisions will affect the process and workflow for making materials accessible online as well as strategies for long-term digital storage and preservation.

FINANCIAL/OPERATIONAL IMPACT: None

CONTACTS:

Internal – The person will interact mainly with archive staff and Digital Library Services.

External – Most external contact will be with vendors and moving image archivists who also are managing digital content.

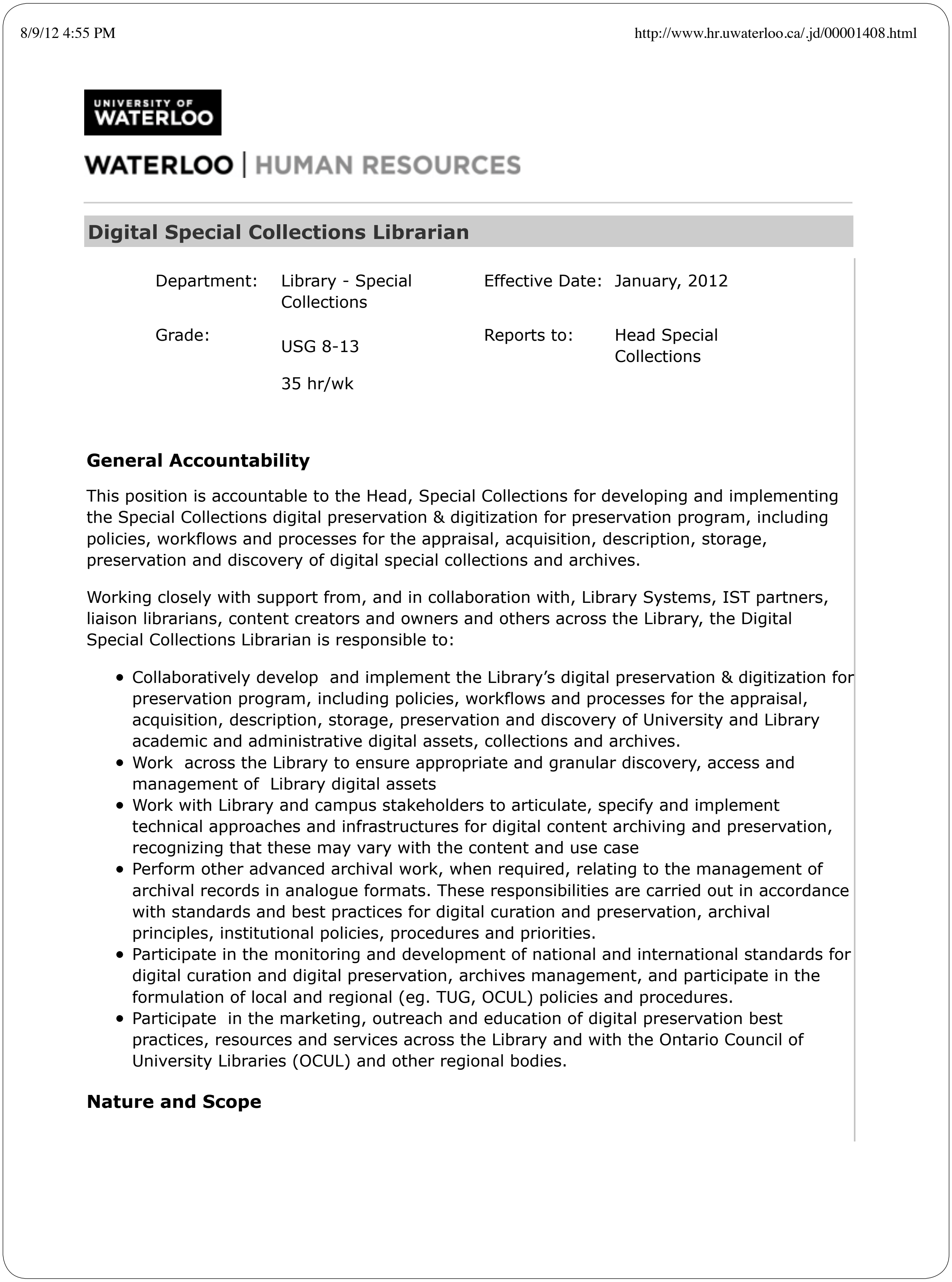
SUPERVISION:

Given – None
Received – Employee will often work closely with the supervisor to develop goals and strategies in the archive as it pertains digitization, editing, and preservation.

UNIVERSITY OF WATERLOO

Digital Special Collections Librarian

http://www.hr.uwaterloo.ca/.jd/00001408.html



8/9/12 4:55 PM http://www.hr.uwaterloo.ca/.jd/00001408.html

UNIVERSITY OF WATERLOO

WATERLOO | HUMAN RESOURCES

Digital Special Collections Librarian

Department:	Library - Special Collections	Effective Date:	January, 2012
Grade:	USG 8-13 35 hr/wk	Reports to:	Head Special Collections

General Accountability

This position is accountable to the Head, Special Collections for developing and implementing the Special Collections digital preservation & digitization for preservation program, including policies, workflows and processes for the appraisal, acquisition, description, storage, preservation and discovery of digital special collections and archives.

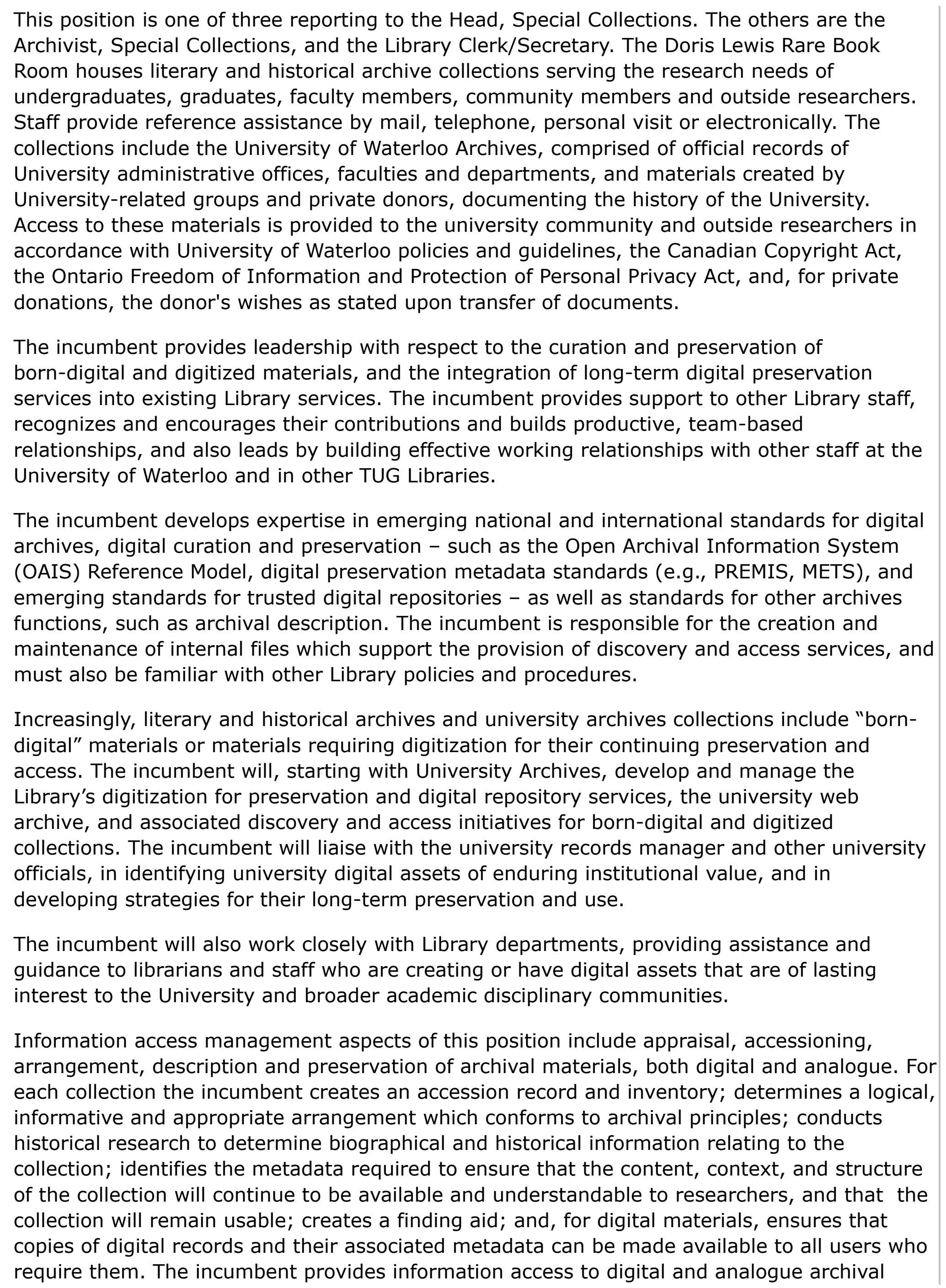
Working closely with support from, and in collaboration with, Library Systems, IST partners, liaison librarians, content creators and owners and others across the Library, the Digital Special Collections Librarian is responsible to:

- Collaboratively develop and implement the Library's digital preservation & digitization for preservation program, including policies, workflows and processes for the appraisal, acquisition, description, storage, preservation and discovery of University and Library academic and administrative digital assets, collections and archives.
- Work across the Library to ensure appropriate and granular discovery, access and management of Library digital assets
- Work with Library and campus stakeholders to articulate, specify and implement technical approaches and infrastructures for digital content archiving and preservation, recognizing that these may vary with the content and use case
- Perform other advanced archival work, when required, relating to the management of archival records in analogue formats. These responsibilities are carried out in accordance with standards and best practices for digital curation and preservation, archival principles, institutional policies, procedures and priorities.
- Participate in the monitoring and development of national and international standards for digital curation and digital preservation, archives management, and participate in the formulation of local and regional (eg. TUG, OCUL) policies and procedures.
- Participate in the marketing, outreach and education of digital preservation best practices, resources and services across the Library and with the Ontario Council of University Libraries (OCUL) and other regional bodies.

Nature and Scope

UNIVERSITY OF WATERLOO

Digital Special Collections Librarian
http://www.hr.uwaterloo.ca/.jd/00001408.html



This position is one of three reporting to the Head, Special Collections. The others are the Archivist, Special Collections, and the Library Clerk/Secretary. The Doris Lewis Rare Book Room houses literary and historical archive collections serving the research needs of undergraduates, graduates, faculty members, community members and outside researchers. Staff provide reference assistance by mail, telephone, personal visit or electronically. The collections include the University of Waterloo Archives, comprised of official records of University administrative offices, faculties and departments, and materials created by University-related groups and private donors, documenting the history of the University. Access to these materials is provided to the university community and outside researchers in accordance with University of Waterloo policies and guidelines, the Canadian Copyright Act, the Ontario Freedom of Information and Protection of Personal Privacy Act, and, for private donations, the donor's wishes as stated upon transfer of documents.

The incumbent provides leadership with respect to the curation and preservation of born-digital and digitized materials, and the integration of long-term digital preservation services into existing Library services. The incumbent provides support to other Library staff, recognizes and encourages their contributions and builds productive, team-based relationships, and also leads by building effective working relationships with other staff at the University of Waterloo and in other TUG Libraries.

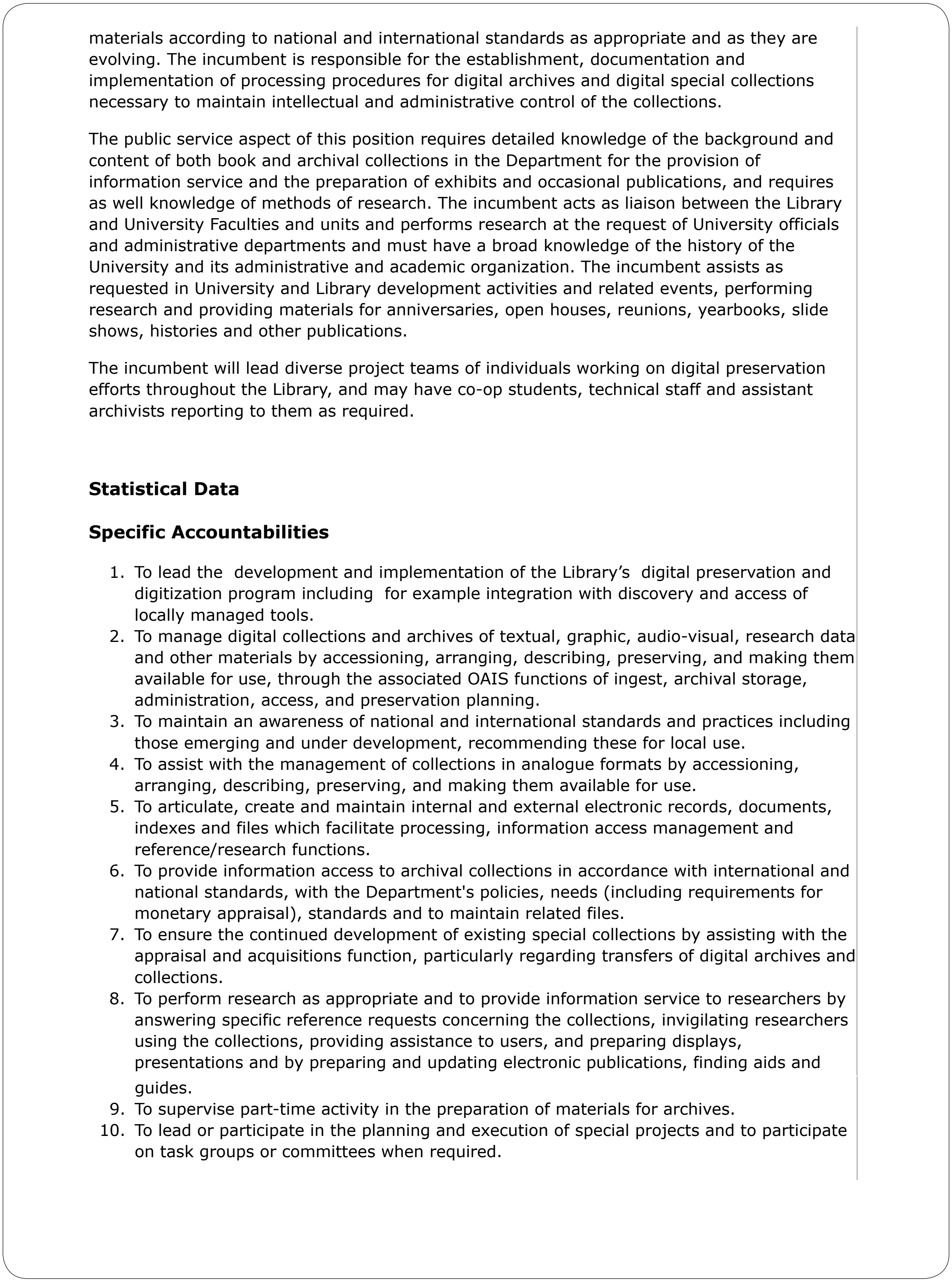
The incumbent develops expertise in emerging national and international standards for digital archives, digital curation and preservation – such as the Open Archival Information System (OAIS) Reference Model, digital preservation metadata standards (e.g., PREMIS, METS), and emerging standards for trusted digital repositories – as well as standards for other archives functions, such as archival description. The incumbent is responsible for the creation and maintenance of internal files which support the provision of discovery and access services, and must also be familiar with other Library policies and procedures.

Increasingly, literary and historical archives and university archives collections include "born-digital" materials or materials requiring digitization for their continuing preservation and access. The incumbent will, starting with University Archives, develop and manage the Library's digitization for preservation and digital repository services, the university web archive, and associated discovery and access initiatives for born-digital and digitized collections. The incumbent will liaise with the university records manager and other university officials, in identifying university digital assets of enduring institutional value, and in developing strategies for their long-term preservation and use.

The incumbent will also work closely with Library departments, providing assistance and guidance to librarians and staff who are creating or have digital assets that are of lasting interest to the University and broader academic disciplinary communities.

Information access management aspects of this position include appraisal, accessioning, arrangement, description and preservation of archival materials, both digital and analogue. For each collection the incumbent creates an accession record and inventory; determines a logical, informative and appropriate arrangement which conforms to archival principles; conducts historical research to determine biographical and historical information relating to the collection; identifies the metadata required to ensure that the content, context, and structure of the collection will continue to be available and understandable to researchers, and that the collection will remain usable; creates a finding aid; and, for digital materials, ensures that copies of digital records and their associated metadata can be made available to all users who require them. The incumbent provides information access to digital and analogue archival

UNIVERSITY OF WATERLOO
Digital Special Collections Librarian
http://www.hr.uwaterloo.ca/.jd/00001408.html



materials according to national and international standards as appropriate and as they are evolving. The incumbent is responsible for the establishment, documentation and implementation of processing procedures for digital archives and digital special collections necessary to maintain intellectual and administrative control of the collections.

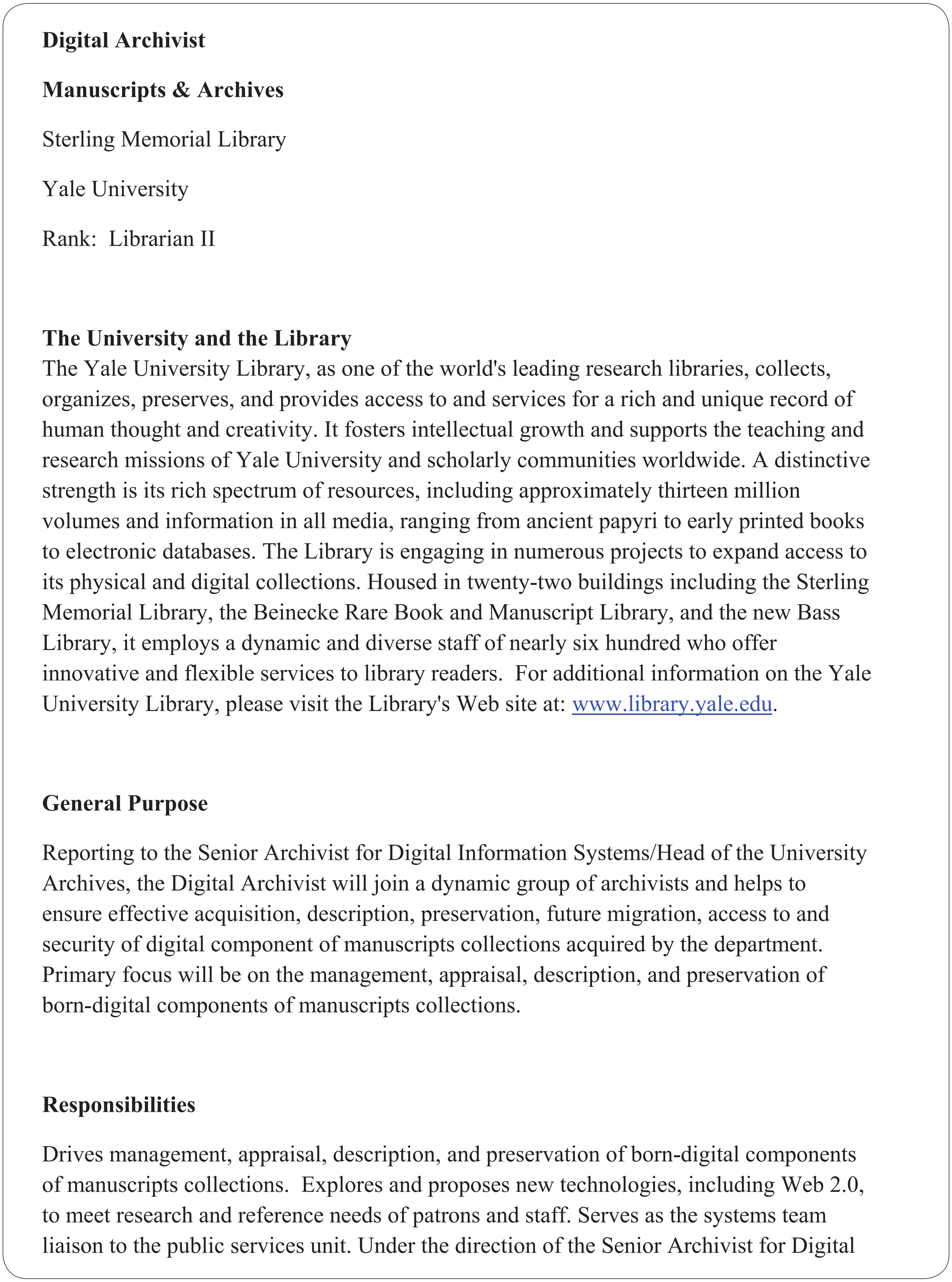
The public service aspect of this position requires detailed knowledge of the background and content of both book and archival collections in the Department for the provision of information service and the preparation of exhibits and occasional publications, and requires as well knowledge of methods of research. The incumbent acts as liaison between the Library and University Faculties and units and performs research at the request of University officials and administrative departments and must have a broad knowledge of the history of the University and its administrative and academic organization. The incumbent assists as requested in University and Library development activities and related events, performing research and providing materials for anniversaries, open houses, reunions, yearbooks, slide shows, histories and other publications.

The incumbent will lead diverse project teams of individuals working on digital preservation efforts throughout the Library, and may have co-op students, technical staff and assistant archivists reporting to them as required.

Statistical Data

Specific Accountabilities

1. To lead the development and implementation of the Library's digital preservation and digitization program including for example integration with discovery and access of locally managed tools.
2. To manage digital collections and archives of textual, graphic, audio-visual, research data and other materials by accessioning, arranging, describing, preserving, and making them available for use, through the associated OAIS functions of ingest, archival storage, administration, access, and preservation planning.
3. To maintain an awareness of national and international standards and practices including those emerging and under development, recommending these for local use.
4. To assist with the management of collections in analogue formats by accessioning, arranging, describing, preserving, and making them available for use.
5. To articulate, create and maintain internal and external electronic records, documents, indexes and files which facilitate processing, information access management and reference/research functions.
6. To provide information access to archival collections in accordance with international and national standards, with the Department's policies, needs (including requirements for monetary appraisal), standards and to maintain related files.
7. To ensure the continued development of existing special collections by assisting with the appraisal and acquisitions function, particularly regarding transfers of digital archives and collections.
8. To perform research as appropriate and to provide information service to researchers by answering specific reference requests concerning the collections, invigilating researchers using the collections, providing assistance to users, and preparing displays, presentations and by preparing and updating electronic publications, finding aids and guides.
9. To supervise part-time activity in the preparation of materials for archives.
10. To lead or participate in the planning and execution of special projects and to participate on task groups or committees when required.



Digital Archivist

Manuscripts & Archives

Sterling Memorial Library

Yale University

Rank: Librarian II

The University and the Library

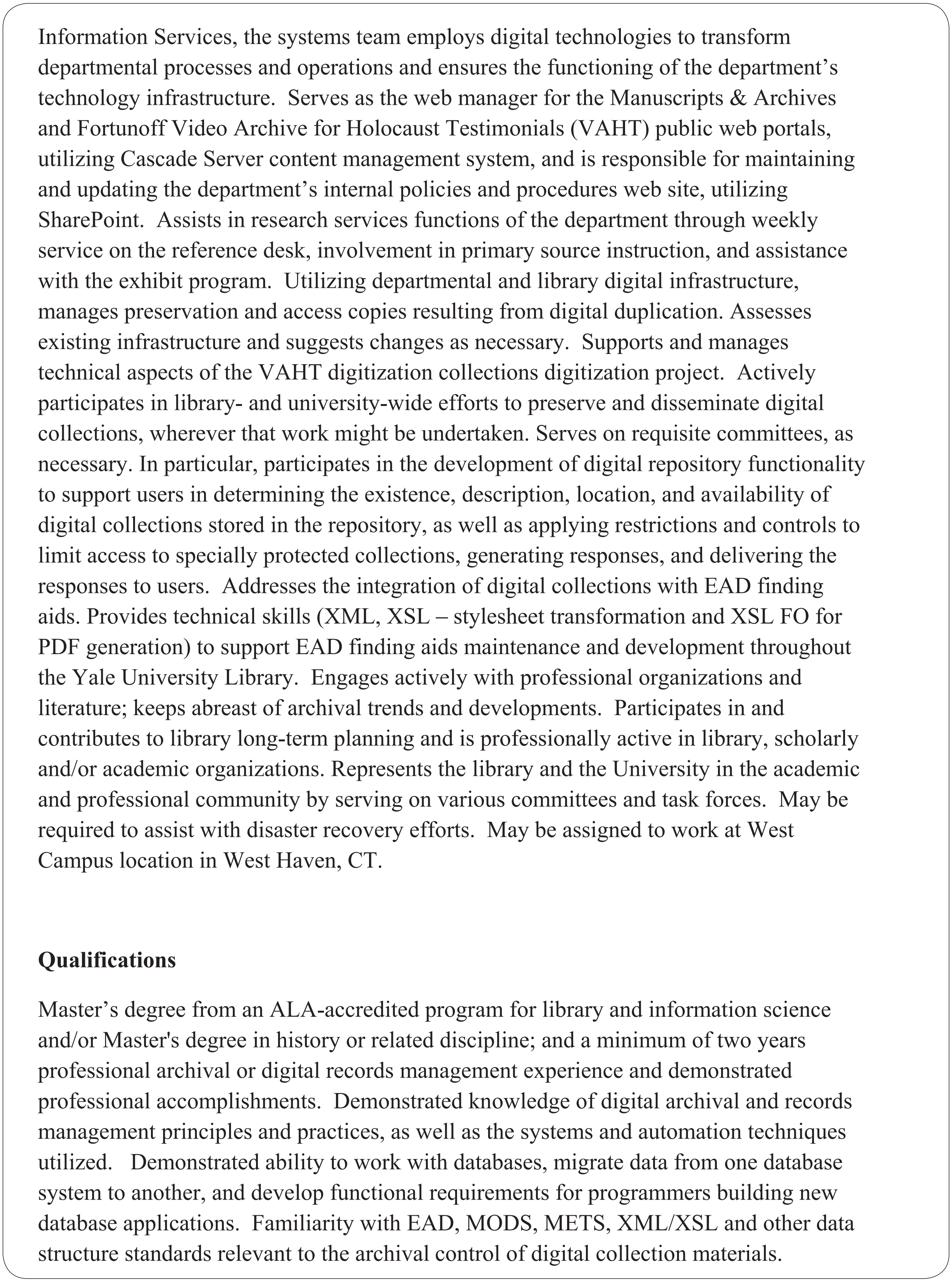
The Yale University Library, as one of the world's leading research libraries, collects, organizes, preserves, and provides access to and services for a rich and unique record of human thought and creativity. It fosters intellectual growth and supports the teaching and research missions of Yale University and scholarly communities worldwide. A distinctive strength is its rich spectrum of resources, including approximately thirteen million volumes and information in all media, ranging from ancient papyri to early printed books to electronic databases. The Library is engaging in numerous projects to expand access to its physical and digital collections. Housed in twenty-two buildings including the Sterling Memorial Library, the Beinecke Rare Book and Manuscript Library, and the new Bass Library, it employs a dynamic and diverse staff of nearly six hundred who offer innovative and flexible services to library readers. For additional information on the Yale University Library, please visit the Library's Web site at: www.library.yale.edu.

General Purpose

Reporting to the Senior Archivist for Digital Information Systems/Head of the University Archives, the Digital Archivist will join a dynamic group of archivists and helps to ensure effective acquisition, description, preservation, future migration, access to and security of digital component of manuscripts collections acquired by the department. Primary focus will be on the management, appraisal, description, and preservation of born-digital components of manuscripts collections.

Responsibilities

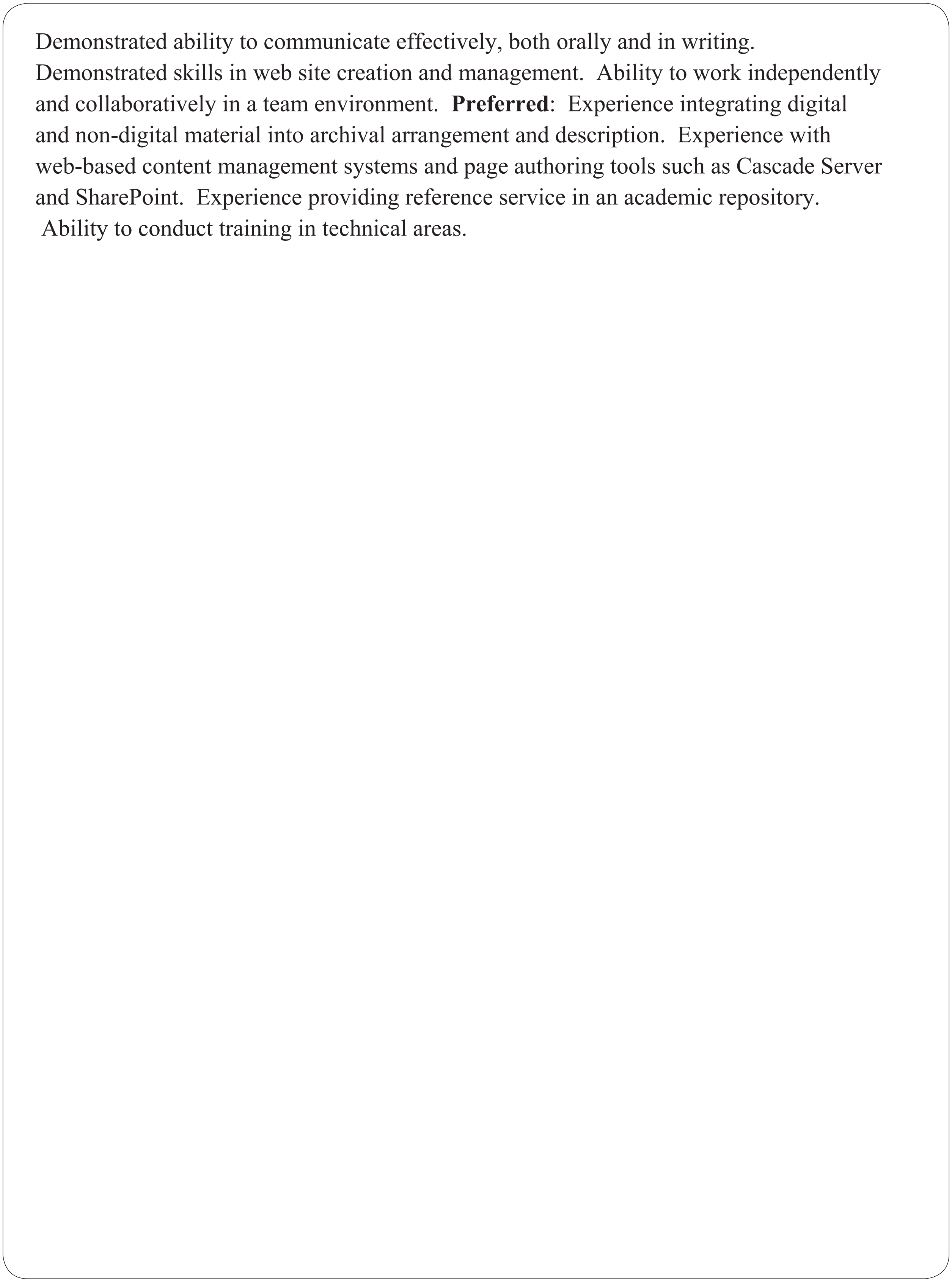
Drives management, appraisal, description, and preservation of born-digital components of manuscripts collections. Explores and proposes new technologies, including Web 2.0, to meet research and reference needs of patrons and staff. Serves as the systems team liaison to the public services unit. Under the direction of the Senior Archivist for Digital



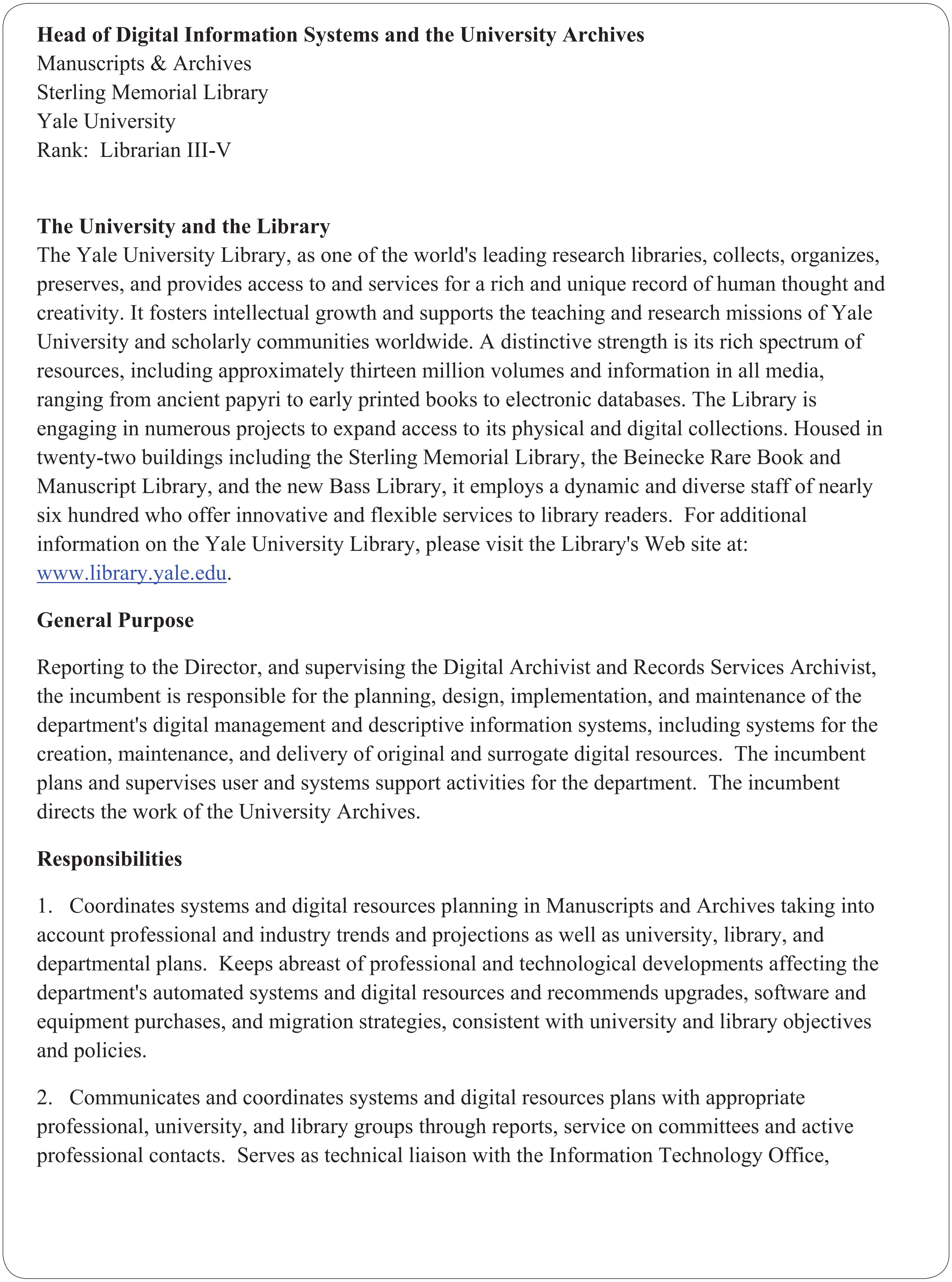
Information Services, the systems team employs digital technologies to transform departmental processes and operations and ensures the functioning of the department's technology infrastructure. Serves as the web manager for the Manuscripts & Archives and Fortunoff Video Archive for Holocaust Testimonials (VAHT) public web portals, utilizing Cascade Server content management system, and is responsible for maintaining and updating the department's internal policies and procedures web site, utilizing SharePoint. Assists in research services functions of the department through weekly service on the reference desk, involvement in primary source instruction, and assistance with the exhibit program. Utilizing departmental and library digital infrastructure, manages preservation and access copies resulting from digital duplication. Assesses existing infrastructure and suggests changes as necessary. Supports and manages technical aspects of the VAHT digitization collections digitization project. Actively participates in library- and university-wide efforts to preserve and disseminate digital collections, wherever that work might be undertaken. Serves on requisite committees, as necessary. In particular, participates in the development of digital repository functionality to support users in determining the existence, description, location, and availability of digital collections stored in the repository, as well as applying restrictions and controls to limit access to specially protected collections, generating responses, and delivering the responses to users. Addresses the integration of digital collections with EAD finding aids. Provides technical skills (XML, XSL – stylesheet transformation and XSL FO for PDF generation) to support EAD finding aids maintenance and development throughout the Yale University Library. Engages actively with professional organizations and literature; keeps abreast of archival trends and developments. Participates in and contributes to library long-term planning and is professionally active in library, scholarly and/or academic organizations. Represents the library and the University in the academic and professional community by serving on various committees and task forces. May be required to assist with disaster recovery efforts. May be assigned to work at West Campus location in West Haven, CT.

Qualifications

Master's degree from an ALA-accredited program for library and information science and/or Master's degree in history or related discipline; and a minimum of two years professional archival or digital records management experience and demonstrated professional accomplishments. Demonstrated knowledge of digital archival and records management principles and practices, as well as the systems and automation techniques utilized. Demonstrated ability to work with databases, migrate data from one database system to another, and develop functional requirements for programmers building new database applications. Familiarity with EAD, MODS, METS, XML/XSL and other data structure standards relevant to the archival control of digital collection materials.



Demonstrated ability to communicate effectively, both orally and in writing. Demonstrated skills in web site creation and management. Ability to work independently and collaboratively in a team environment. **Preferred**: Experience integrating digital and non-digital material into archival arrangement and description. Experience with web-based content management systems and page authoring tools such as Cascade Server and SharePoint. Experience providing reference service in an academic repository. Ability to conduct training in technical areas.



Head of Digital Information Systems and the University Archives
Manuscripts & Archives
Sterling Memorial Library
Yale University
Rank: Librarian III-V

The University and the Library

The Yale University Library, as one of the world's leading research libraries, collects, organizes, preserves, and provides access to and services for a rich and unique record of human thought and creativity. It fosters intellectual growth and supports the teaching and research missions of Yale University and scholarly communities worldwide. A distinctive strength is its rich spectrum of resources, including approximately thirteen million volumes and information in all media, ranging from ancient papyri to early printed books to electronic databases. The Library is engaging in numerous projects to expand access to its physical and digital collections. Housed in twenty-two buildings including the Sterling Memorial Library, the Beinecke Rare Book and Manuscript Library, and the new Bass Library, it employs a dynamic and diverse staff of nearly six hundred who offer innovative and flexible services to library readers. For additional information on the Yale University Library, please visit the Library's Web site at: www.library.yale.edu.

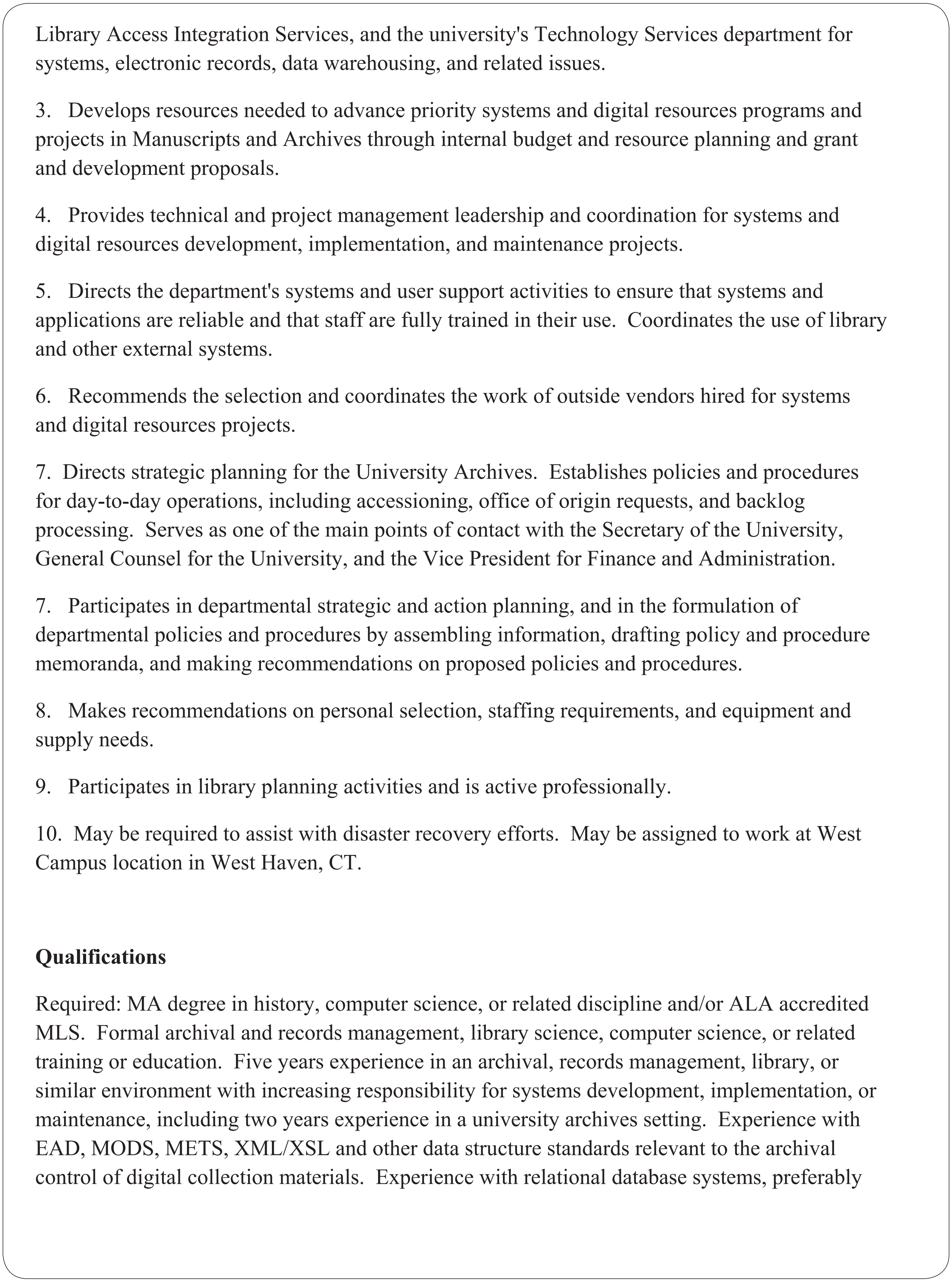
General Purpose

Reporting to the Director, and supervising the Digital Archivist and Records Services Archivist, the incumbent is responsible for the planning, design, implementation, and maintenance of the department's digital management and descriptive information systems, including systems for the creation, maintenance, and delivery of original and surrogate digital resources. The incumbent plans and supervises user and systems support activities for the department. The incumbent directs the work of the University Archives.

Responsibilities

1. Coordinates systems and digital resources planning in Manuscripts and Archives taking into account professional and industry trends and projections as well as university, library, and departmental plans. Keeps abreast of professional and technological developments affecting the department's automated systems and digital resources and recommends upgrades, software and equipment purchases, and migration strategies, consistent with university and library objectives and policies.

2. Communicates and coordinates systems and digital resources plans with appropriate professional, university, and library groups through reports, service on committees and active professional contacts. Serves as technical liaison with the Information Technology Office,



Library Access Integration Services, and the university's Technology Services department for systems, electronic records, data warehousing, and related issues.

3. Develops resources needed to advance priority systems and digital resources programs and projects in Manuscripts and Archives through internal budget and resource planning and grant and development proposals.

4. Provides technical and project management leadership and coordination for systems and digital resources development, implementation, and maintenance projects.

5. Directs the department's systems and user support activities to ensure that systems and applications are reliable and that staff are fully trained in their use. Coordinates the use of library and other external systems.

6. Recommends the selection and coordinates the work of outside vendors hired for systems and digital resources projects.

7. Directs strategic planning for the University Archives. Establishes policies and procedures for day-to-day operations, including accessioning, office of origin requests, and backlog processing. Serves as one of the main points of contact with the Secretary of the University, General Counsel for the University, and the Vice President for Finance and Administration.

7. Participates in departmental strategic and action planning, and in the formulation of departmental policies and procedures by assembling information, drafting policy and procedure memoranda, and making recommendations on proposed policies and procedures.

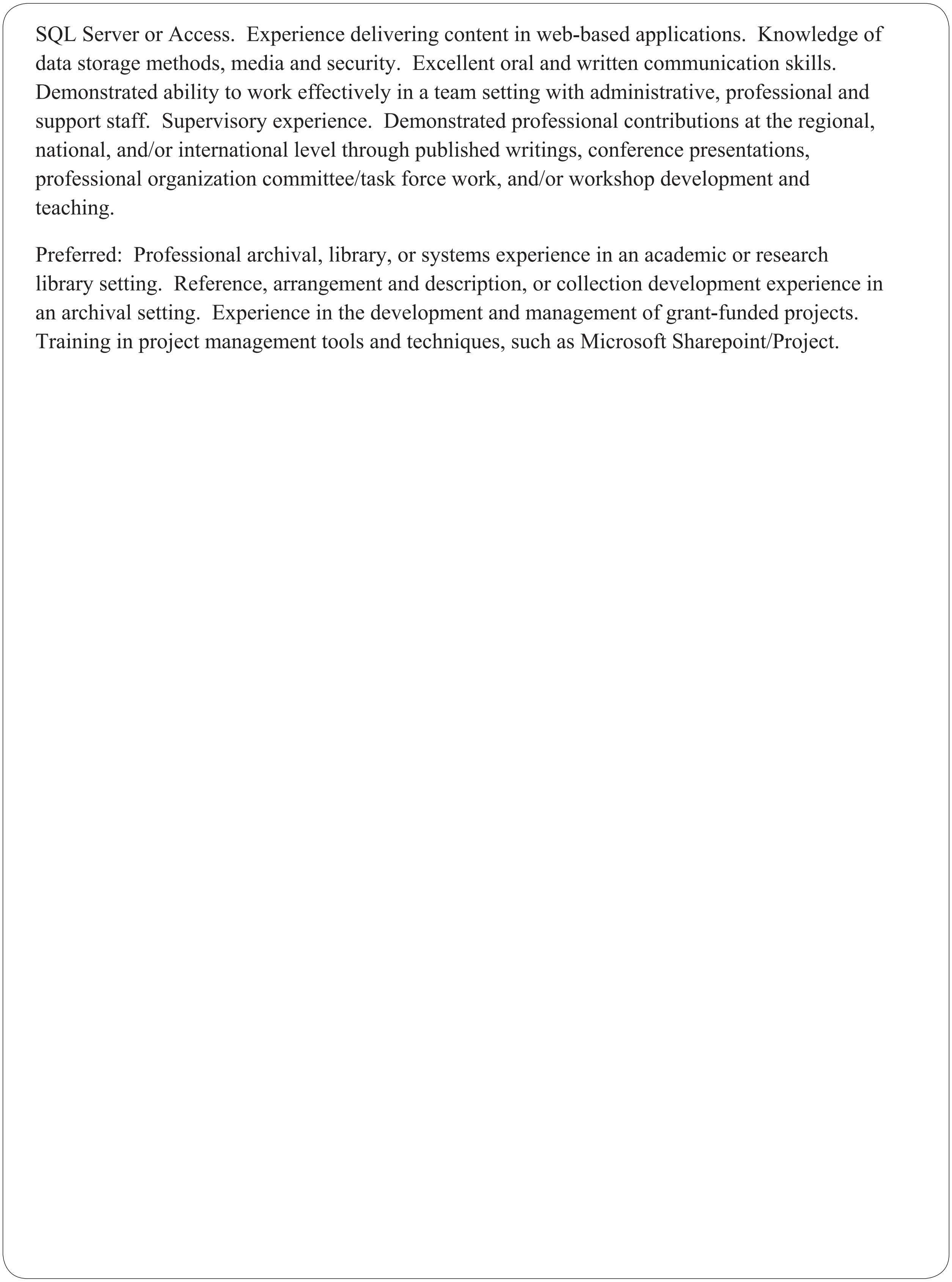
8. Makes recommendations on personal selection, staffing requirements, and equipment and supply needs.

9. Participates in library planning activities and is active professionally.

10. May be required to assist with disaster recovery efforts. May be assigned to work at West Campus location in West Haven, CT.

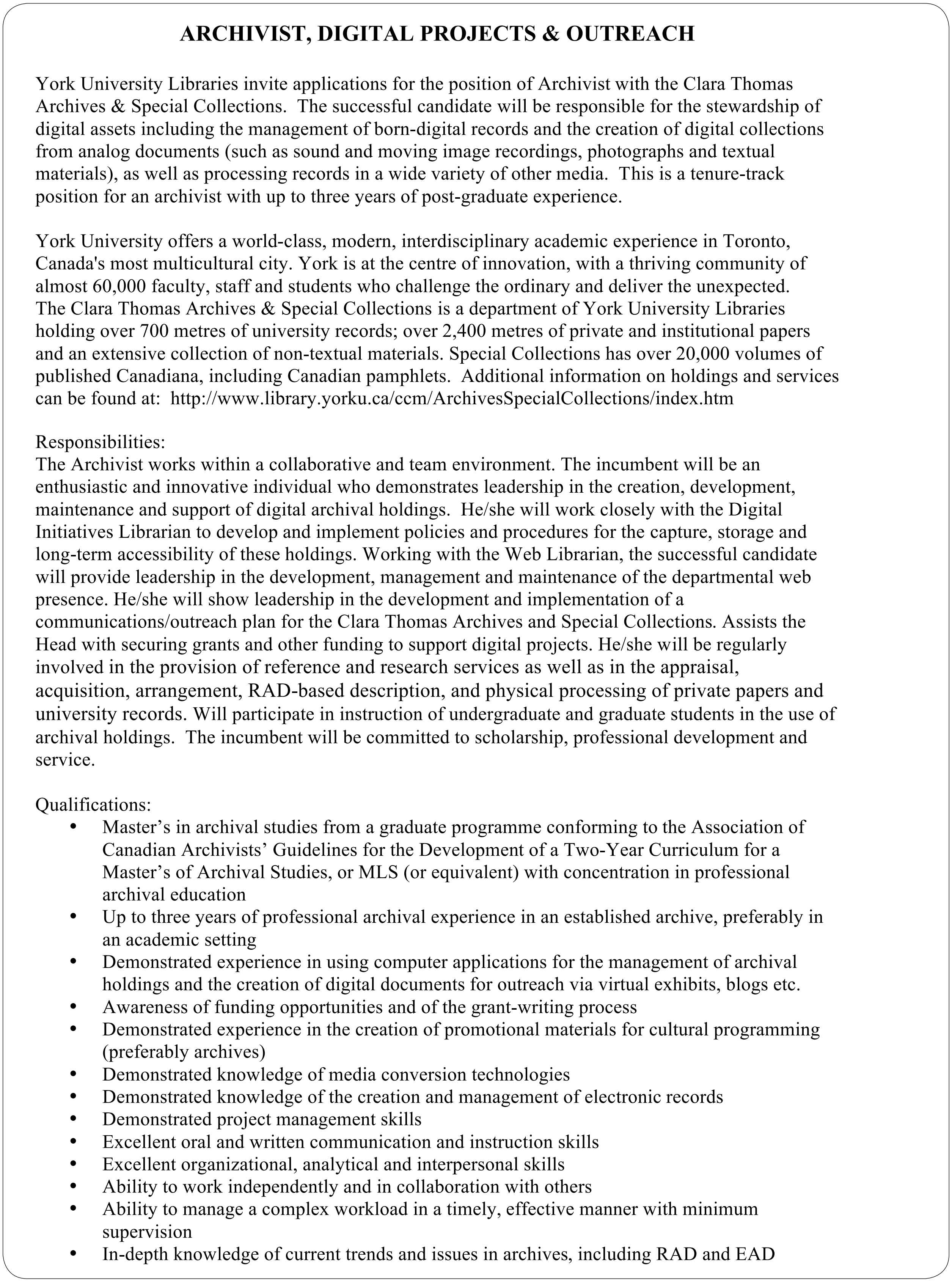
Qualifications

Required: MA degree in history, computer science, or related discipline and/or ALA accredited MLS. Formal archival and records management, library science, computer science, or related training or education. Five years experience in an archival, records management, library, or similar environment with increasing responsibility for systems development, implementation, or maintenance, including two years experience in a university archives setting. Experience with EAD, MODS, METS, XML/XSL and other data structure standards relevant to the archival control of digital collection materials. Experience with relational database systems, preferably



SQL Server or Access. Experience delivering content in web-based applications. Knowledge of data storage methods, media and security. Excellent oral and written communication skills. Demonstrated ability to work effectively in a team setting with administrative, professional and support staff. Supervisory experience. Demonstrated professional contributions at the regional, national, and/or international level through published writings, conference presentations, professional organization committee/task force work, and/or workshop development and teaching.

Preferred: Professional archival, library, or systems experience in an academic or research library setting. Reference, arrangement and description, or collection development experience in an archival setting. Experience in the development and management of grant-funded projects. Training in project management tools and techniques, such as Microsoft Sharepoint/Project.



ARCHIVIST, DIGITAL PROJECTS & OUTREACH

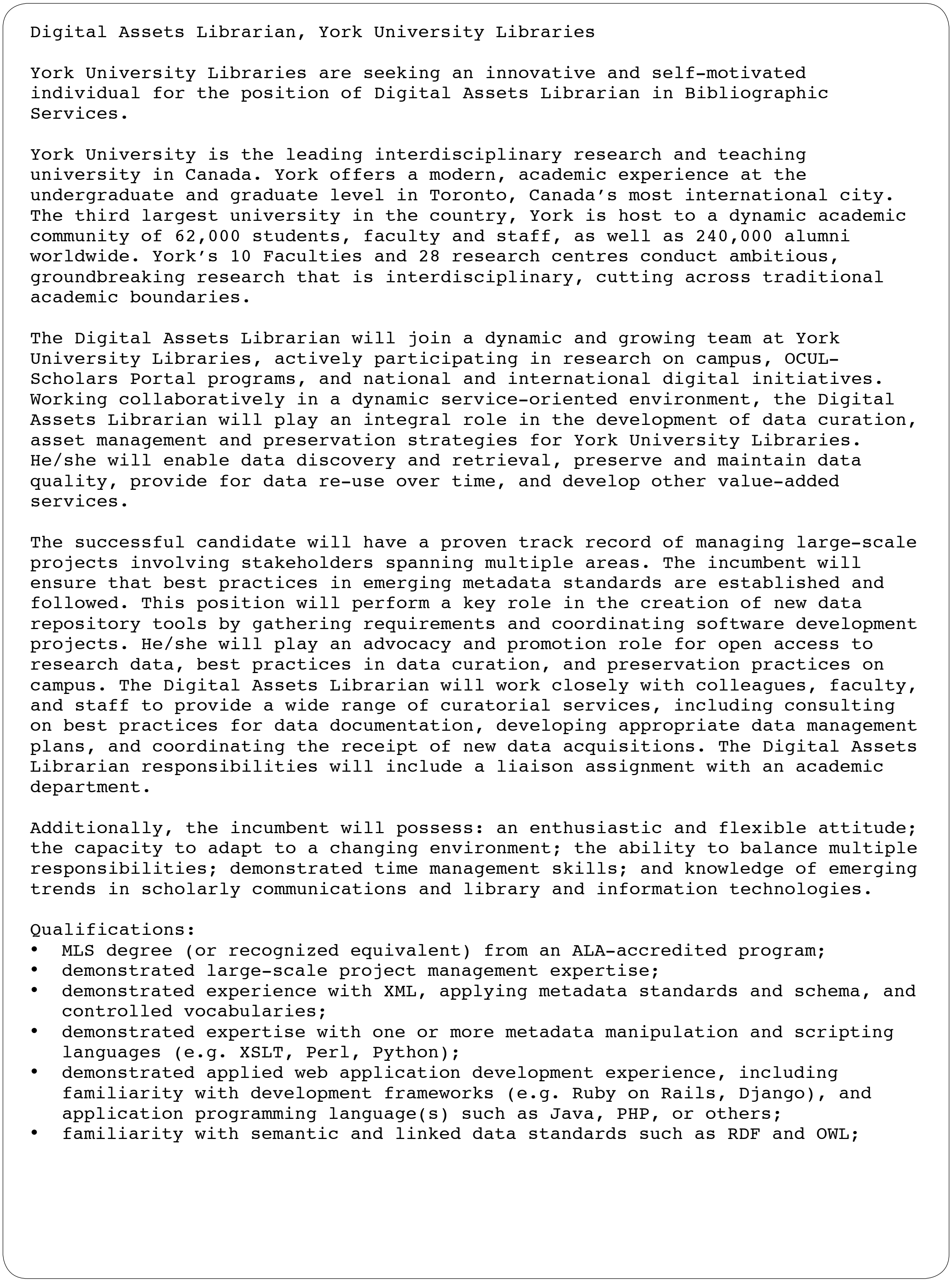
York University Libraries invite applications for the position of Archivist with the Clara Thomas Archives & Special Collections. The successful candidate will be responsible for the stewardship of digital assets including the management of born-digital records and the creation of digital collections from analog documents (such as sound and moving image recordings, photographs and textual materials), as well as processing records in a wide variety of other media. This is a tenure-track position for an archivist with up to three years of post-graduate experience.

York University offers a world-class, modern, interdisciplinary academic experience in Toronto, Canada's most multicultural city. York is at the centre of innovation, with a thriving community of almost 60,000 faculty, staff and students who challenge the ordinary and deliver the unexpected. The Clara Thomas Archives & Special Collections is a department of York University Libraries holding over 700 metres of university records; over 2,400 metres of private and institutional papers and an extensive collection of non-textual materials. Special Collections has over 20,000 volumes of published Canadiana, including Canadian pamphlets. Additional information on holdings and services can be found at: http://www.library.yorku.ca/ccm/ArchivesSpecialCollections/index.htm

Responsibilities:
The Archivist works within a collaborative and team environment. The incumbent will be an enthusiastic and innovative individual who demonstrates leadership in the creation, development, maintenance and support of digital archival holdings. He/she will work closely with the Digital Initiatives Librarian to develop and implement policies and procedures for the capture, storage and long-term accessibility of these holdings. Working with the Web Librarian, the successful candidate will provide leadership in the development, management and maintenance of the departmental web presence. He/she will show leadership in the development and implementation of a communications/outreach plan for the Clara Thomas Archives and Special Collections. Assists the Head with securing grants and other funding to support digital projects. He/she will be regularly involved in the provision of reference and research services as well as in the appraisal, acquisition, arrangement, RAD-based description, and physical processing of private papers and university records. Will participate in instruction of undergraduate and graduate students in the use of archival holdings. The incumbent will be committed to scholarship, professional development and service.

Qualifications:

- Master’s in archival studies from a graduate programme conforming to the Association of Canadian Archivists’ Guidelines for the Development of a Two-Year Curriculum for a Master’s of Archival Studies, or MLS (or equivalent) with concentration in professional archival education
- Up to three years of professional archival experience in an established archive, preferably in an academic setting
- Demonstrated experience in using computer applications for the management of archival holdings and the creation of digital documents for outreach via virtual exhibits, blogs etc.
- Awareness of funding opportunities and of the grant-writing process
- Demonstrated experience in the creation of promotional materials for cultural programming (preferably archives)
- Demonstrated knowledge of media conversion technologies
- Demonstrated knowledge of the creation and management of electronic records
- Demonstrated project management skills
- Excellent oral and written communication and instruction skills
- Excellent organizational, analytical and interpersonal skills
- Ability to work independently and in collaboration with others
- Ability to manage a complex workload in a timely, effective manner with minimum supervision
- In-depth knowledge of current trends and issues in archives, including RAD and EAD



Digital Assets Librarian, York University Libraries

York University Libraries are seeking an innovative and self-motivated individual for the position of Digital Assets Librarian in Bibliographic Services.

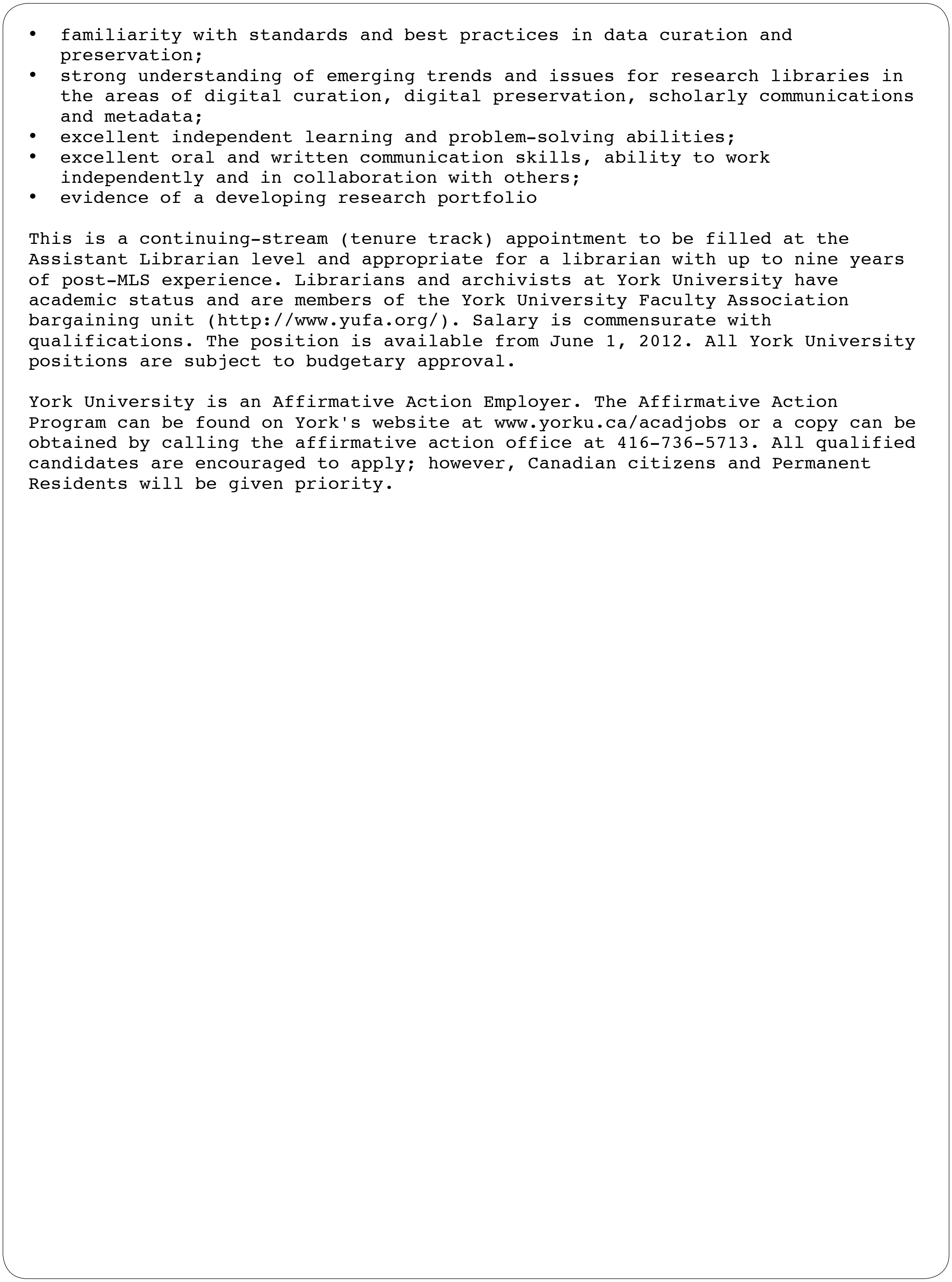
York University is the leading interdisciplinary research and teaching university in Canada. York offers a modern, academic experience at the undergraduate and graduate level in Toronto, Canada's most international city. The third largest university in the country, York is host to a dynamic academic community of 62,000 students, faculty and staff, as well as 240,000 alumni worldwide. York's 10 Faculties and 28 research centres conduct ambitious, groundbreaking research that is interdisciplinary, cutting across traditional academic boundaries.

The Digital Assets Librarian will join a dynamic and growing team at York University Libraries, actively participating in research on campus, OCUL-Scholars Portal programs, and national and international digital initiatives. Working collaboratively in a dynamic service-oriented environment, the Digital Assets Librarian will play an integral role in the development of data curation, asset management and preservation strategies for York University Libraries. He/she will enable data discovery and retrieval, preserve and maintain data quality, provide for data re-use over time, and develop other value-added services.

The successful candidate will have a proven track record of managing large-scale projects involving stakeholders spanning multiple areas. The incumbent will ensure that best practices in emerging metadata standards are established and followed. This position will perform a key role in the creation of new data repository tools by gathering requirements and coordinating software development projects. He/she will play an advocacy and promotion role for open access to research data, best practices in data curation, and preservation practices on campus. The Digital Assets Librarian will work closely with colleagues, faculty, and staff to provide a wide range of curatorial services, including consulting on best practices for data documentation, developing appropriate data management plans, and coordinating the receipt of new data acquisitions. The Digital Assets Librarian responsibilities will include a liaison assignment with an academic department.

Additionally, the incumbent will possess: an enthusiastic and flexible attitude; the capacity to adapt to a changing environment; the ability to balance multiple responsibilities; demonstrated time management skills; and knowledge of emerging trends in scholarly communications and library and information technologies.

Qualifications:
- MLS degree (or recognized equivalent) from an ALA-accredited program;
- demonstrated large-scale project management expertise;
- demonstrated experience with XML, applying metadata standards and schema, and controlled vocabularies;
- demonstrated expertise with one or more metadata manipulation and scripting languages (e.g. XSLT, Perl, Python);
- demonstrated applied web application development experience, including familiarity with development frameworks (e.g. Ruby on Rails, Django), and application programming language(s) such as Java, PHP, or others;
- familiarity with semantic and linked data standards such as RDF and OWL;



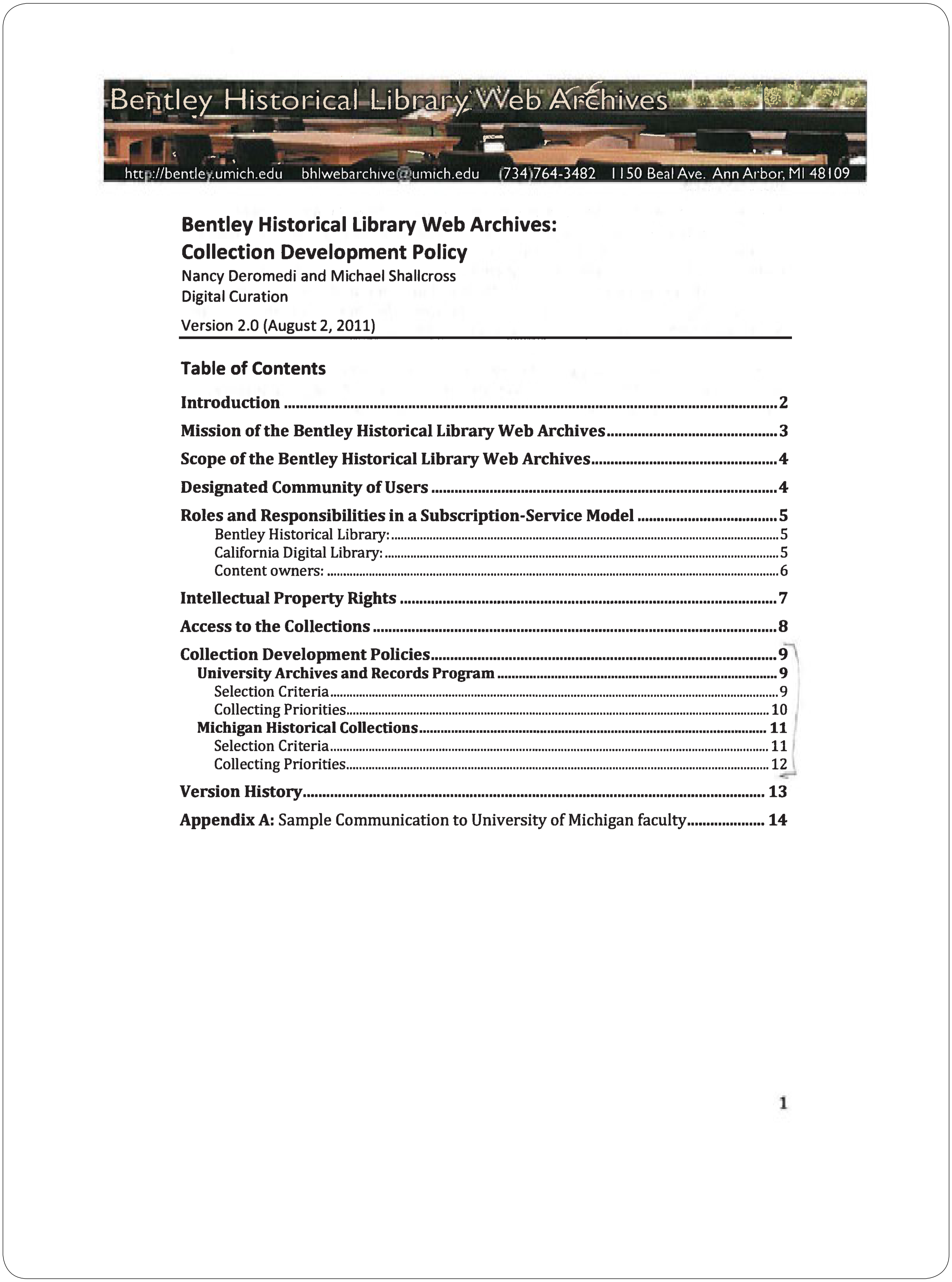
- familiarity with standards and best practices in data curation and preservation;
- strong understanding of emerging trends and issues for research libraries in the areas of digital curation, digital preservation, scholarly communications and metadata;
- excellent independent learning and problem-solving abilities;
- excellent oral and written communication skills, ability to work independently and in collaboration with others;
- evidence of a developing research portfolio

This is a continuing-stream (tenure track) appointment to be filled at the Assistant Librarian level and appropriate for a librarian with up to nine years of post-MLS experience. Librarians and archivists at York University have academic status and are members of the York University Faculty Association bargaining unit (http://www.yufa.org/). Salary is commensurate with qualifications. The position is available from June 1, 2012. All York University positions are subject to budgetary approval.

York University is an Affirmative Action Employer. The Affirmative Action Program can be found on York's website at www.yorku.ca/acadjobs or a copy can be obtained by calling the affirmative action office at 416-736-5713. All qualified candidates are encouraged to apply; however, Canadian citizens and Permanent Residents will be given priority.

Collection Policies

UNIVERSITY OF MICHIGAN
BHL Web Archives: Collection Development Policy
http://bentley.umich.edu/dchome/webarchives/BHL_WebArchives_Policy.pdf



Bentley Historical Library Web Archives

http://bentley.umich.edu bhlwebarchive@umich.edu (734)764-3482 1150 Beal Ave. Ann Arbor, MI 48109

Bentley Historical Library Web Archives: Collection Development Policy

Nancy Deromedi and Michael Shallcross
Digital Curation

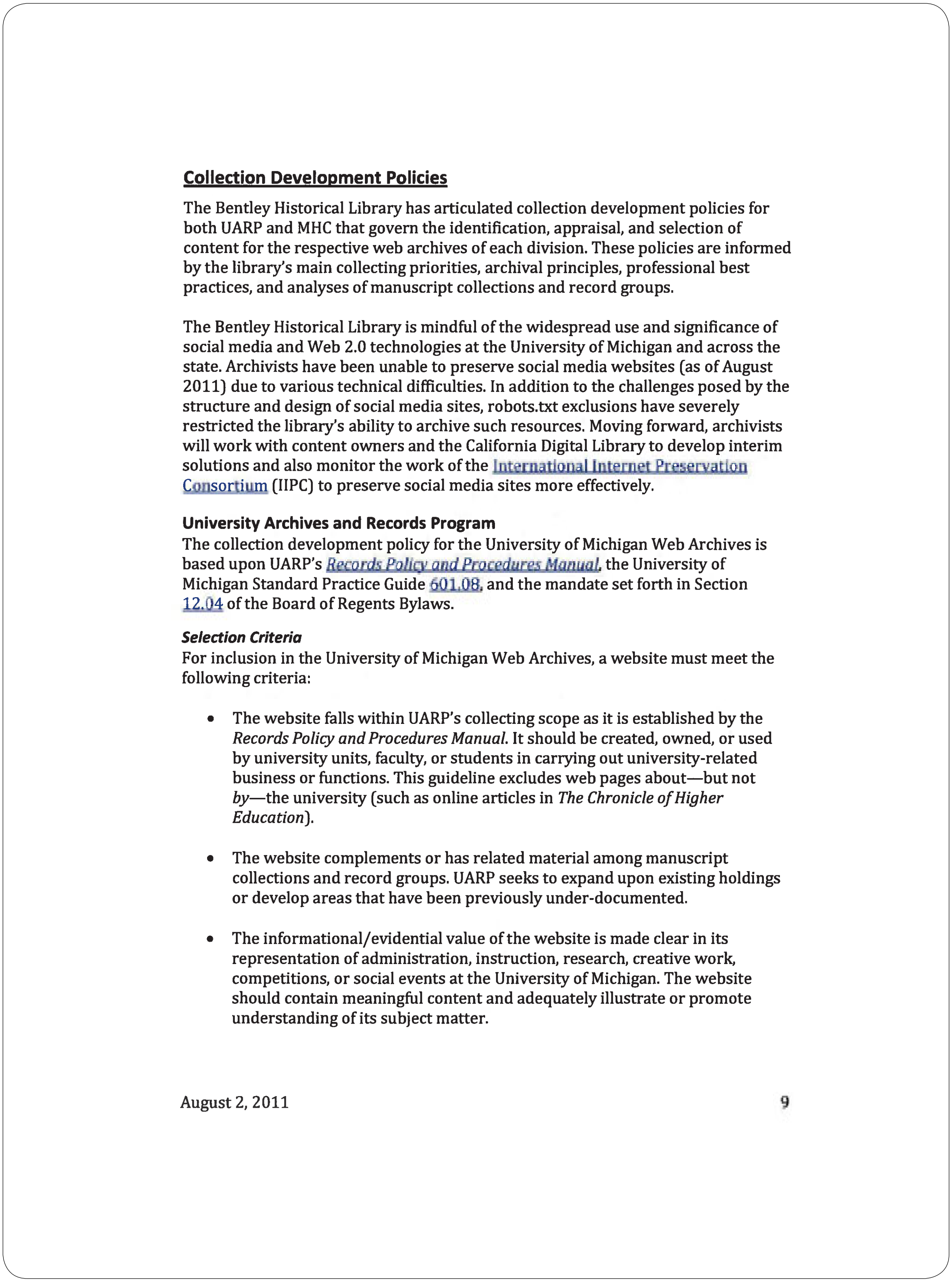
Version 2.0 (August 2, 2011)

Table of Contents

UNIVERSITY OF MICHIGAN

BHL Web Archives: Collection Development Policy

http://bentley.umich.edu/dchome/webarchives/BHL_WebArchives_Policy.pdf



Collection Development Policies

The Bentley Historical Library has articulated collection development policies for both UARP and MHC that govern the identification, appraisal, and selection of content for the respective web archives of each division. These policies are informed by the library's main collecting priorities, archival principles, professional best practices, and analyses of manuscript collections and record groups.

The Bentley Historical Library is mindful of the widespread use and significance of social media and Web 2.0 technologies at the University of Michigan and across the state. Archivists have been unable to preserve social media websites (as of August 2011) due to various technical difficulties. In addition to the challenges posed by the structure and design of social media sites, robots.txt exclusions have severely restricted the library's ability to archive such resources. Moving forward, archivists will work with content owners and the California Digital Library to develop interim solutions and also monitor the work of the International Internet Preservation Consortium (IIPC) to preserve social media sites more effectively.

University Archives and Records Program

The collection development policy for the University of Michigan Web Archives is based upon UARP's *Records Policy and Procedures Manual*, the University of Michigan Standard Practice Guide 601.08, and the mandate set forth in Section 12.04 of the Board of Regents Bylaws.

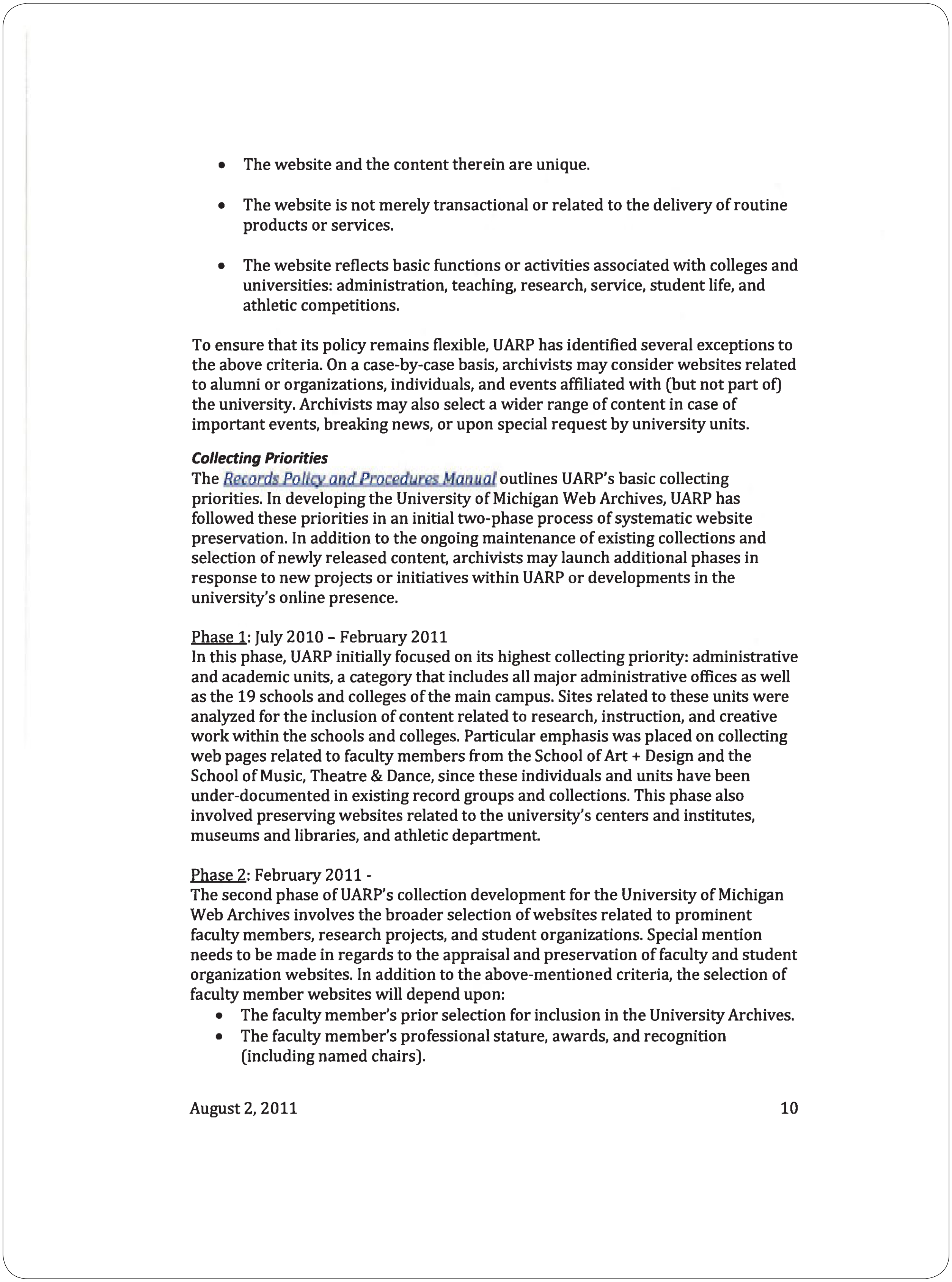
Selection Criteria

For inclusion in the University of Michigan Web Archives, a website must meet the following criteria:

- The website falls within UARP's collecting scope as it is established by the *Records Policy and Procedures Manual*. It should be created, owned, or used by university units, faculty, or students in carrying out university-related business or functions. This guideline excludes web pages about—but not *by*—the university (such as online articles in *The Chronicle of Higher Education*).

- The website complements or has related material among manuscript collections and record groups. UARP seeks to expand upon existing holdings or develop areas that have been previously under-documented.

- The informational/evidential value of the website is made clear in its representation of administration, instruction, research, creative work, competitions, or social events at the University of Michigan. The website should contain meaningful content and adequately illustrate or promote understanding of its subject matter.



- The website and the content therein are unique.

- The website is not merely transactional or related to the delivery of routine products or services.

- The website reflects basic functions or activities associated with colleges and universities: administration, teaching, research, service, student life, and athletic competitions.

To ensure that its policy remains flexible, UARP has identified several exceptions to the above criteria. On a case-by-case basis, archivists may consider websites related to alumni or organizations, individuals, and events affiliated with (but not part of) the university. Archivists may also select a wider range of content in case of important events, breaking news, or upon special request by university units.

Collecting Priorities

The *Records Policy and Procedures Manual* outlines UARP's basic collecting priorities. In developing the University of Michigan Web Archives, UARP has followed these priorities in an initial two-phase process of systematic website preservation. In addition to the ongoing maintenance of existing collections and selection of newly released content, archivists may launch additional phases in response to new projects or initiatives within UARP or developments in the university's online presence.

Phase 1: July 2010 – February 2011
In this phase, UARP initially focused on its highest collecting priority: administrative and academic units, a category that includes all major administrative offices as well as the 19 schools and colleges of the main campus. Sites related to these units were analyzed for the inclusion of content related to research, instruction, and creative work within the schools and colleges. Particular emphasis was placed on collecting web pages related to faculty members from the School of Art + Design and the School of Music, Theatre & Dance, since these individuals and units have been under-documented in existing record groups and collections. This phase also involved preserving websites related to the university's centers and institutes, museums and libraries, and athletic department.

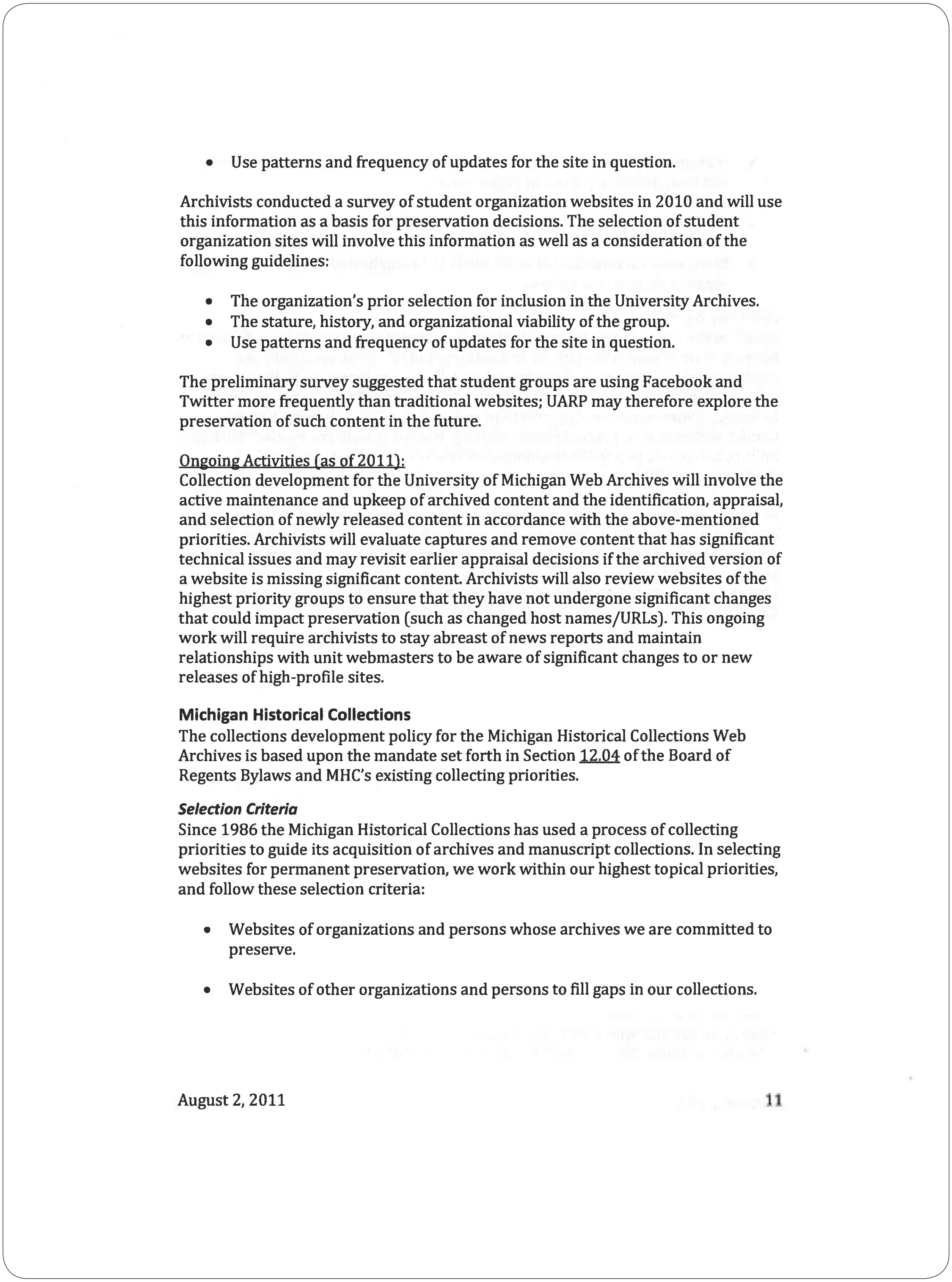
Phase 2: February 2011 -
The second phase of UARP's collection development for the University of Michigan Web Archives involves the broader selection of websites related to prominent faculty members, research projects, and student organizations. Special mention needs to be made in regards to the appraisal and preservation of faculty and student organization websites. In addition to the above-mentioned criteria, the selection of faculty member websites will depend upon:

- The faculty member's prior selection for inclusion in the University Archives.
- The faculty member's professional stature, awards, and recognition (including named chairs).

UNIVERSITY OF MICHIGAN

BHL Web Archives: Collection Development Policy

http://bentley.umich.edu/dchome/webarchives/BHL_WebArchives_Policy.pdf



- Use patterns and frequency of updates for the site in question.

Archivists conducted a survey of student organization websites in 2010 and will use this information as a basis for preservation decisions. The selection of student organization sites will involve this information as well as a consideration of the following guidelines:

- The organization's prior selection for inclusion in the University Archives.
- The stature, history, and organizational viability of the group.
- Use patterns and frequency of updates for the site in question.

The preliminary survey suggested that student groups are using Facebook and Twitter more frequently than traditional websites; UARP may therefore explore the preservation of such content in the future.

Ongoing Activities (as of 2011):
Collection development for the University of Michigan Web Archives will involve the active maintenance and upkeep of archived content and the identification, appraisal, and selection of newly released content in accordance with the above-mentioned priorities. Archivists will evaluate captures and remove content that has significant technical issues and may revisit earlier appraisal decisions if the archived version of a website is missing significant content. Archivists will also review websites of the highest priority groups to ensure that they have not undergone significant changes that could impact preservation (such as changed host names/URLs). This ongoing work will require archivists to stay abreast of news reports and maintain relationships with unit webmasters to be aware of significant changes to or new releases of high-profile sites.

Michigan Historical Collections

The collections development policy for the Michigan Historical Collections Web Archives is based upon the mandate set forth in Section 12.04 of the Board of Regents Bylaws and MHC's existing collecting priorities.

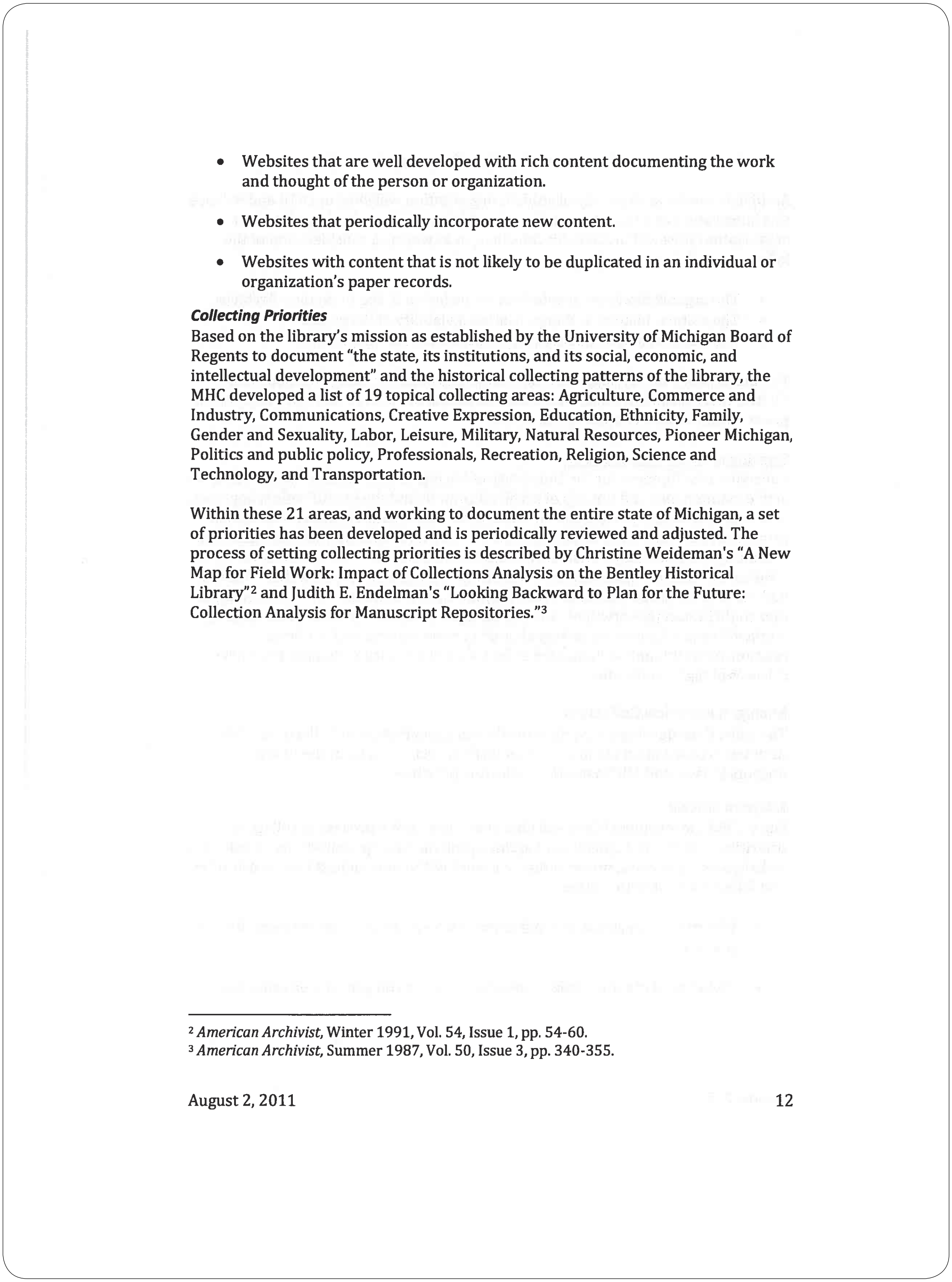
Selection Criteria

Since 1986 the Michigan Historical Collections has used a process of collecting priorities to guide its acquisition of archives and manuscript collections. In selecting websites for permanent preservation, we work within our highest topical priorities, and follow these selection criteria:

- Websites of organizations and persons whose archives we are committed to preserve.

- Websites of other organizations and persons to fill gaps in our collections.

UNIVERSITY OF MICHIGAN
BHL Web Archives: Collection Development Policy
http://bentley.umich.edu/dchome/webarchives/BHL_WebArchives_Policy.pdf

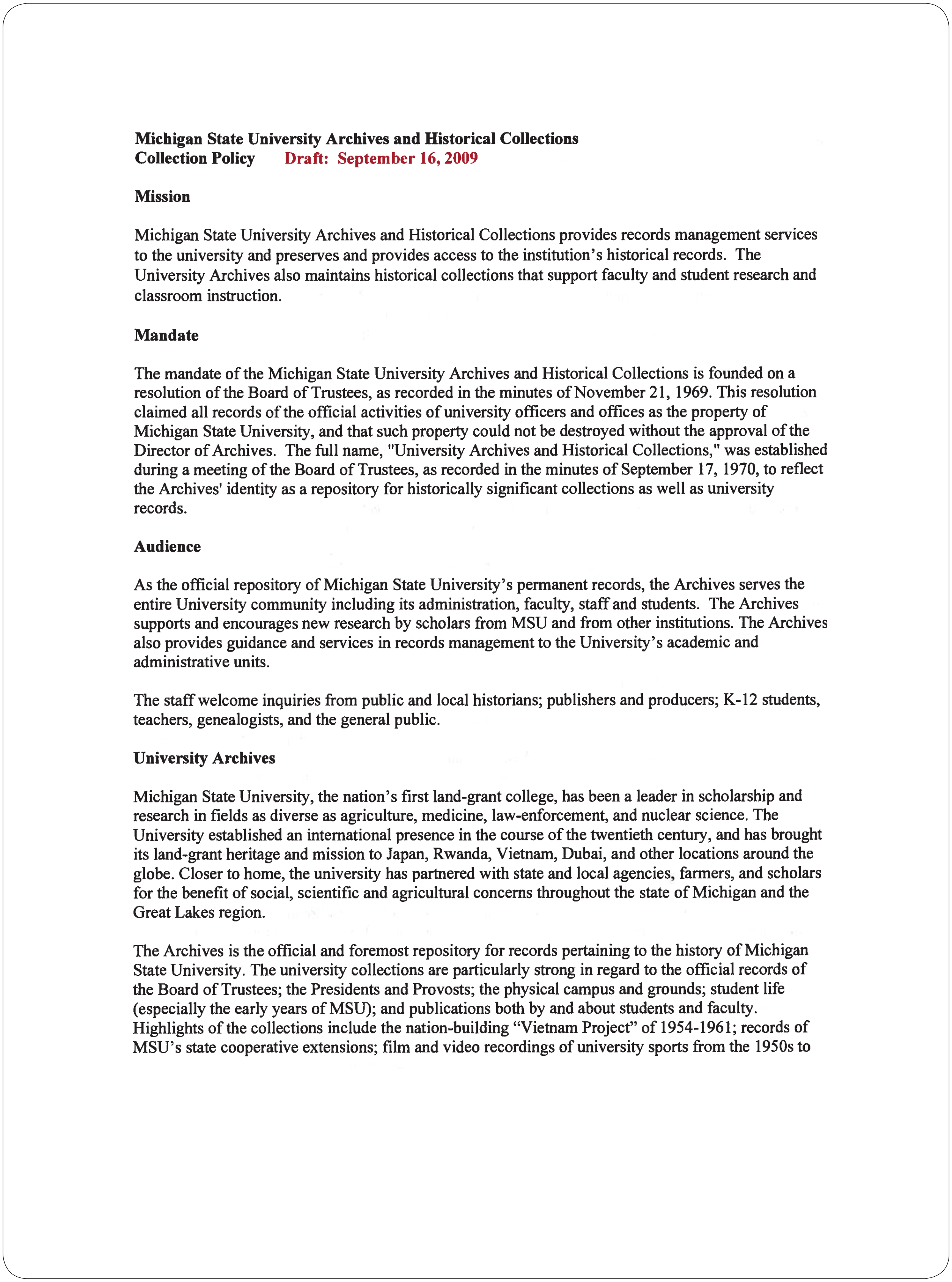


- Websites that are well developed with rich content documenting the work and thought of the person or organization.
- Websites that periodically incorporate new content.
- Websites with content that is not likely to be duplicated in an individual or organization's paper records.

Collecting Priorities
Based on the library's mission as established by the University of Michigan Board of Regents to document "the state, its institutions, and its social, economic, and intellectual development" and the historical collecting patterns of the library, the MHC developed a list of 19 topical collecting areas: Agriculture, Commerce and Industry, Communications, Creative Expression, Education, Ethnicity, Family, Gender and Sexuality, Labor, Leisure, Military, Natural Resources, Pioneer Michigan, Politics and public policy, Professionals, Recreation, Religion, Science and Technology, and Transportation.

Within these 21 areas, and working to document the entire state of Michigan, a set of priorities has been developed and is periodically reviewed and adjusted. The process of setting collecting priorities is described by Christine Weideman's "A New Map for Field Work: Impact of Collections Analysis on the Bentley Historical Library"[2] and Judith E. Endelman's "Looking Backward to Plan for the Future: Collection Analysis for Manuscript Repositories."[3]

[2] *American Archivist*, Winter 1991, Vol. 54, Issue 1, pp. 54-60.
[3] *American Archivist*, Summer 1987, Vol. 50, Issue 3, pp. 340-355.



Michigan State University Archives and Historical Collections
Collection Policy **Draft: September 16, 2009**

Mission

Michigan State University Archives and Historical Collections provides records management services to the university and preserves and provides access to the institution's historical records. The University Archives also maintains historical collections that support faculty and student research and classroom instruction.

Mandate

The mandate of the Michigan State University Archives and Historical Collections is founded on a resolution of the Board of Trustees, as recorded in the minutes of November 21, 1969. This resolution claimed all records of the official activities of university officers and offices as the property of Michigan State University, and that such property could not be destroyed without the approval of the Director of Archives. The full name, "University Archives and Historical Collections," was established during a meeting of the Board of Trustees, as recorded in the minutes of September 17, 1970, to reflect the Archives' identity as a repository for historically significant collections as well as university records.

Audience

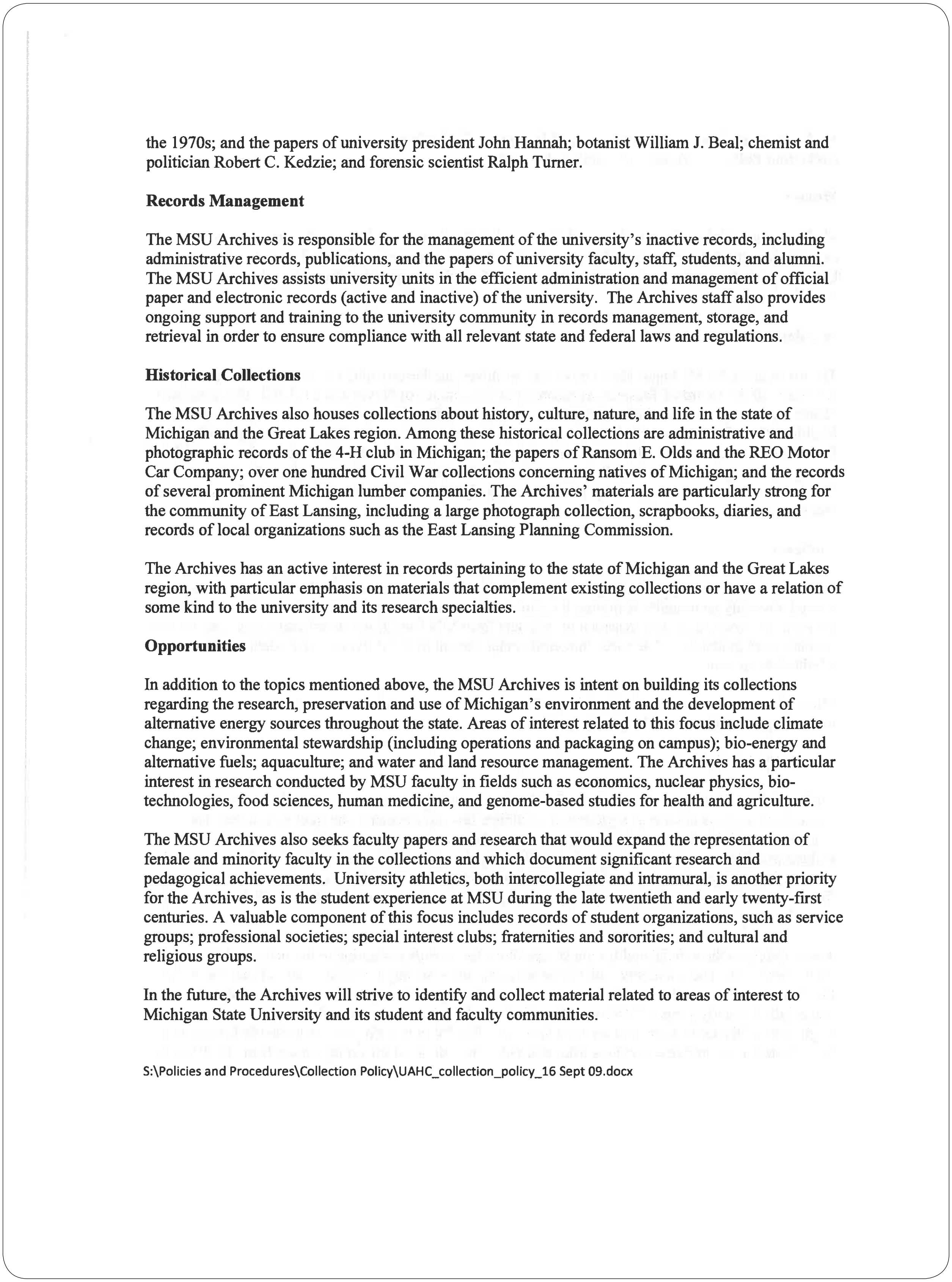
As the official repository of Michigan State University's permanent records, the Archives serves the entire University community including its administration, faculty, staff and students. The Archives supports and encourages new research by scholars from MSU and from other institutions. The Archives also provides guidance and services in records management to the University's academic and administrative units.

The staff welcome inquiries from public and local historians; publishers and producers; K-12 students, teachers, genealogists, and the general public.

University Archives

Michigan State University, the nation's first land-grant college, has been a leader in scholarship and research in fields as diverse as agriculture, medicine, law-enforcement, and nuclear science. The University established an international presence in the course of the twentieth century, and has brought its land-grant heritage and mission to Japan, Rwanda, Vietnam, Dubai, and other locations around the globe. Closer to home, the university has partnered with state and local agencies, farmers, and scholars for the benefit of social, scientific and agricultural concerns throughout the state of Michigan and the Great Lakes region.

The Archives is the official and foremost repository for records pertaining to the history of Michigan State University. The university collections are particularly strong in regard to the official records of the Board of Trustees; the Presidents and Provosts; the physical campus and grounds; student life (especially the early years of MSU); and publications both by and about students and faculty. Highlights of the collections include the nation-building "Vietnam Project" of 1954-1961; records of MSU's state cooperative extensions; film and video recordings of university sports from the 1950s to



the 1970s; and the papers of university president John Hannah; botanist William J. Beal; chemist and politician Robert C. Kedzie; and forensic scientist Ralph Turner.

Records Management

The MSU Archives is responsible for the management of the university's inactive records, including administrative records, publications, and the papers of university faculty, staff, students, and alumni. The MSU Archives assists university units in the efficient administration and management of official paper and electronic records (active and inactive) of the university. The Archives staff also provides ongoing support and training to the university community in records management, storage, and retrieval in order to ensure compliance with all relevant state and federal laws and regulations.

Historical Collections

The MSU Archives also houses collections about history, culture, nature, and life in the state of Michigan and the Great Lakes region. Among these historical collections are administrative and photographic records of the 4-H club in Michigan; the papers of Ransom E. Olds and the REO Motor Car Company; over one hundred Civil War collections concerning natives of Michigan; and the records of several prominent Michigan lumber companies. The Archives' materials are particularly strong for the community of East Lansing, including a large photograph collection, scrapbooks, diaries, and records of local organizations such as the East Lansing Planning Commission.

The Archives has an active interest in records pertaining to the state of Michigan and the Great Lakes region, with particular emphasis on materials that complement existing collections or have a relation of some kind to the university and its research specialties.

Opportunities

In addition to the topics mentioned above, the MSU Archives is intent on building its collections regarding the research, preservation and use of Michigan's environment and the development of alternative energy sources throughout the state. Areas of interest related to this focus include climate change; environmental stewardship (including operations and packaging on campus); bio-energy and alternative fuels; aquaculture; and water and land resource management. The Archives has a particular interest in research conducted by MSU faculty in fields such as economics, nuclear physics, bio-technologies, food sciences, human medicine, and genome-based studies for health and agriculture.

The MSU Archives also seeks faculty papers and research that would expand the representation of female and minority faculty in the collections and which document significant research and pedagogical achievements. University athletics, both intercollegiate and intramural, is another priority for the Archives, as is the student experience at MSU during the late twentieth and early twenty-first centuries. A valuable component of this focus includes records of student organizations, such as service groups; professional societies; special interest clubs; fraternities and sororities; and cultural and religious groups.

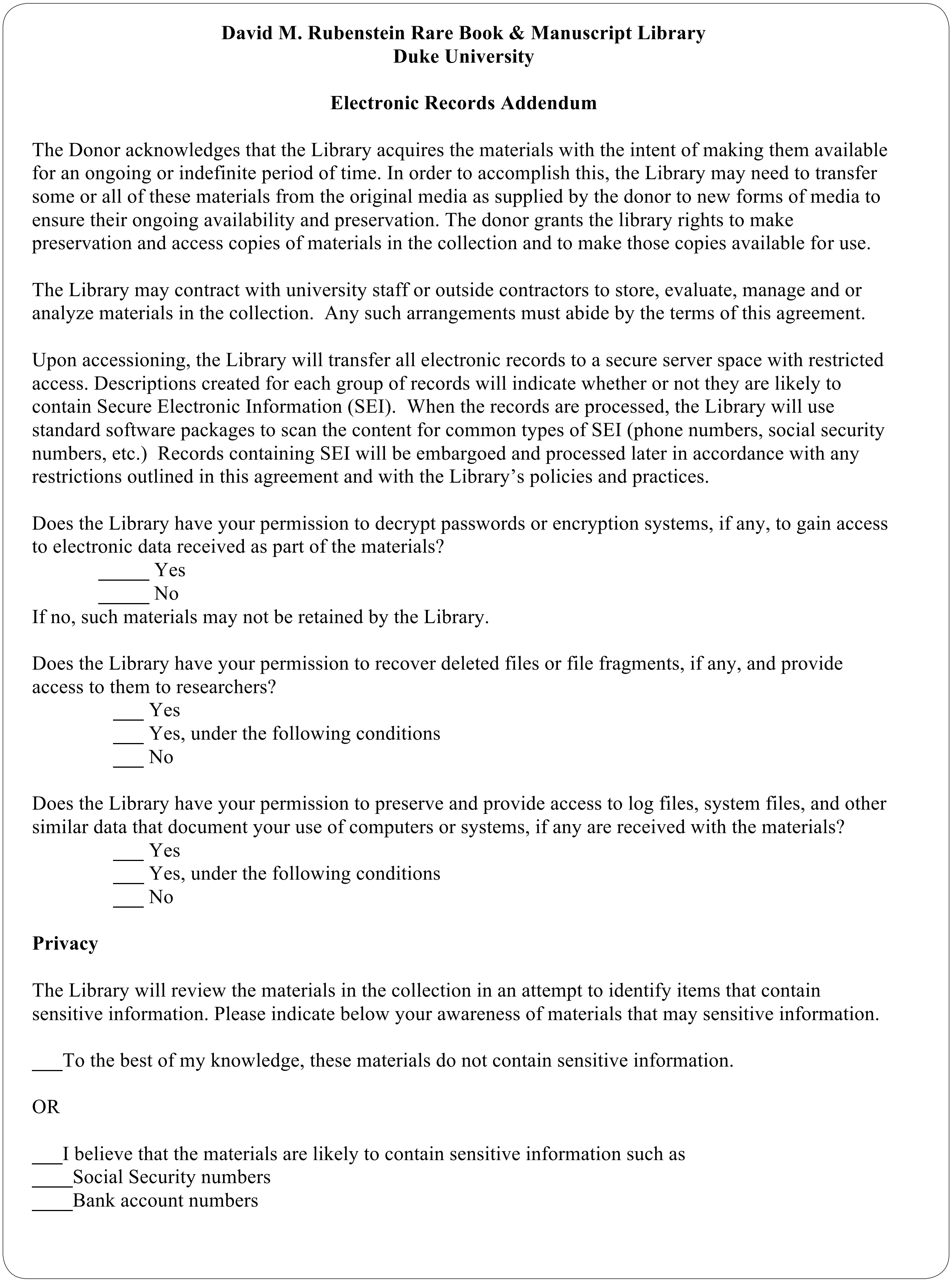
In the future, the Archives will strive to identify and collect material related to areas of interest to Michigan State University and its student and faculty communities.

S:\Policies and Procedures\Collection Policy\UAHC_collection_policy_16 Sept 09.docx

Gift/Purchase Agreements

DUKE UNIVERSITY

Donor Agreement Addendum for Electronic Records



David M. Rubenstein Rare Book & Manuscript Library
Duke University

Electronic Records Addendum

The Donor acknowledges that the Library acquires the materials with the intent of making them available for an ongoing or indefinite period of time. In order to accomplish this, the Library may need to transfer some or all of these materials from the original media as supplied by the donor to new forms of media to ensure their ongoing availability and preservation. The donor grants the library rights to make preservation and access copies of materials in the collection and to make those copies available for use.

The Library may contract with university staff or outside contractors to store, evaluate, manage and or analyze materials in the collection. Any such arrangements must abide by the terms of this agreement.

Upon accessioning, the Library will transfer all electronic records to a secure server space with restricted access. Descriptions created for each group of records will indicate whether or not they are likely to contain Secure Electronic Information (SEI). When the records are processed, the Library will use standard software packages to scan the content for common types of SEI (phone numbers, social security numbers, etc.) Records containing SEI will be embargoed and processed later in accordance with any restrictions outlined in this agreement and with the Library's policies and practices.

Does the Library have your permission to decrypt passwords or encryption systems, if any, to gain access to electronic data received as part of the materials?

_____ Yes
_____ No

If no, such materials may not be retained by the Library.

Does the Library have your permission to recover deleted files or file fragments, if any, and provide access to them to researchers?

___ Yes
___ Yes, under the following conditions
___ No

Does the Library have your permission to preserve and provide access to log files, system files, and other similar data that document your use of computers or systems, if any are received with the materials?

___ Yes
___ Yes, under the following conditions
___ No

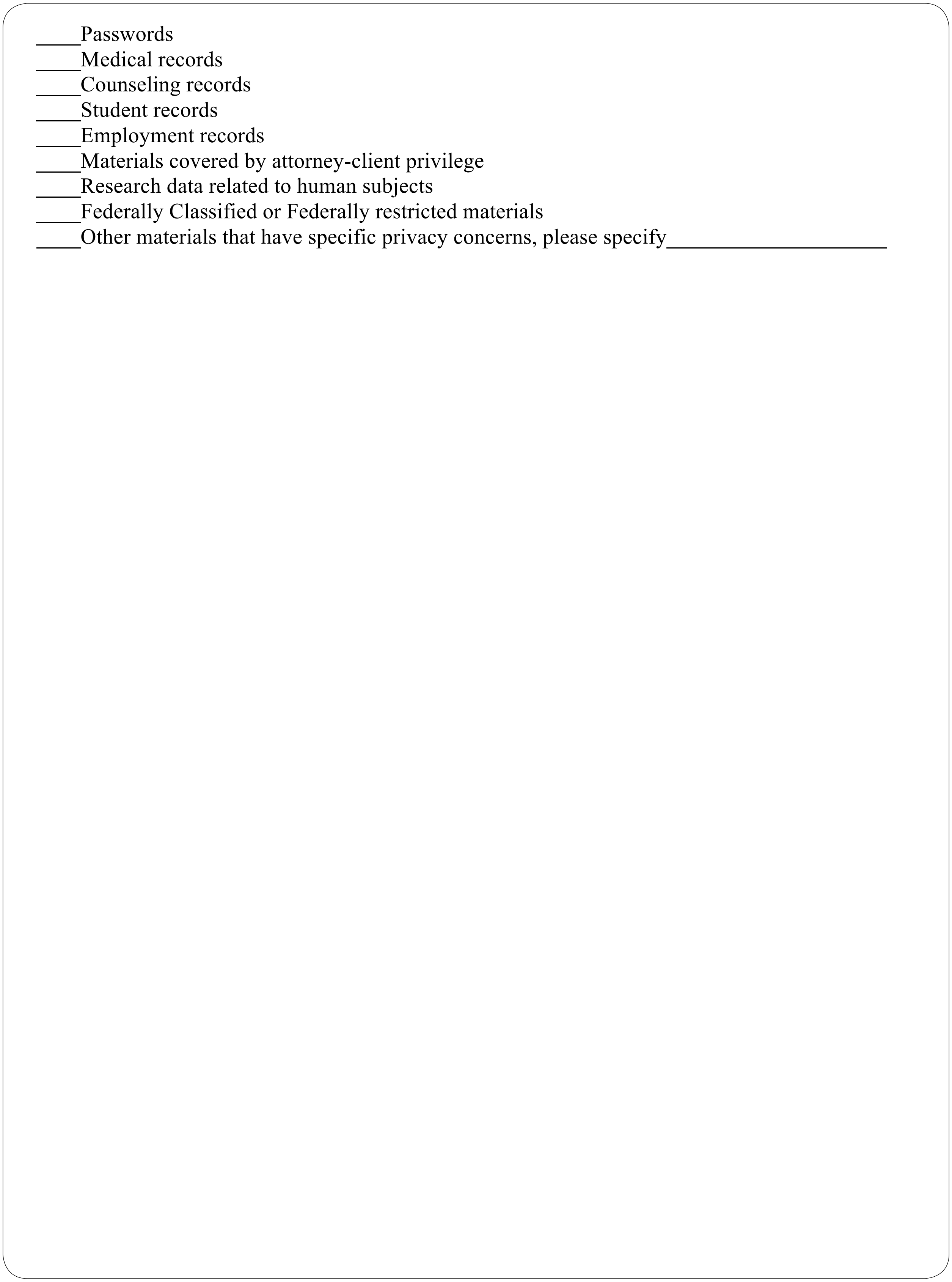
Privacy

The Library will review the materials in the collection in an attempt to identify items that contain sensitive information. Please indicate below your awareness of materials that may sensitive information.

___To the best of my knowledge, these materials do not contain sensitive information.

OR

___I believe that the materials are likely to contain sensitive information such as
____Social Security numbers
____Bank account numbers



____Passwords

____Medical records

____Counseling records

____Student records

____Employment records

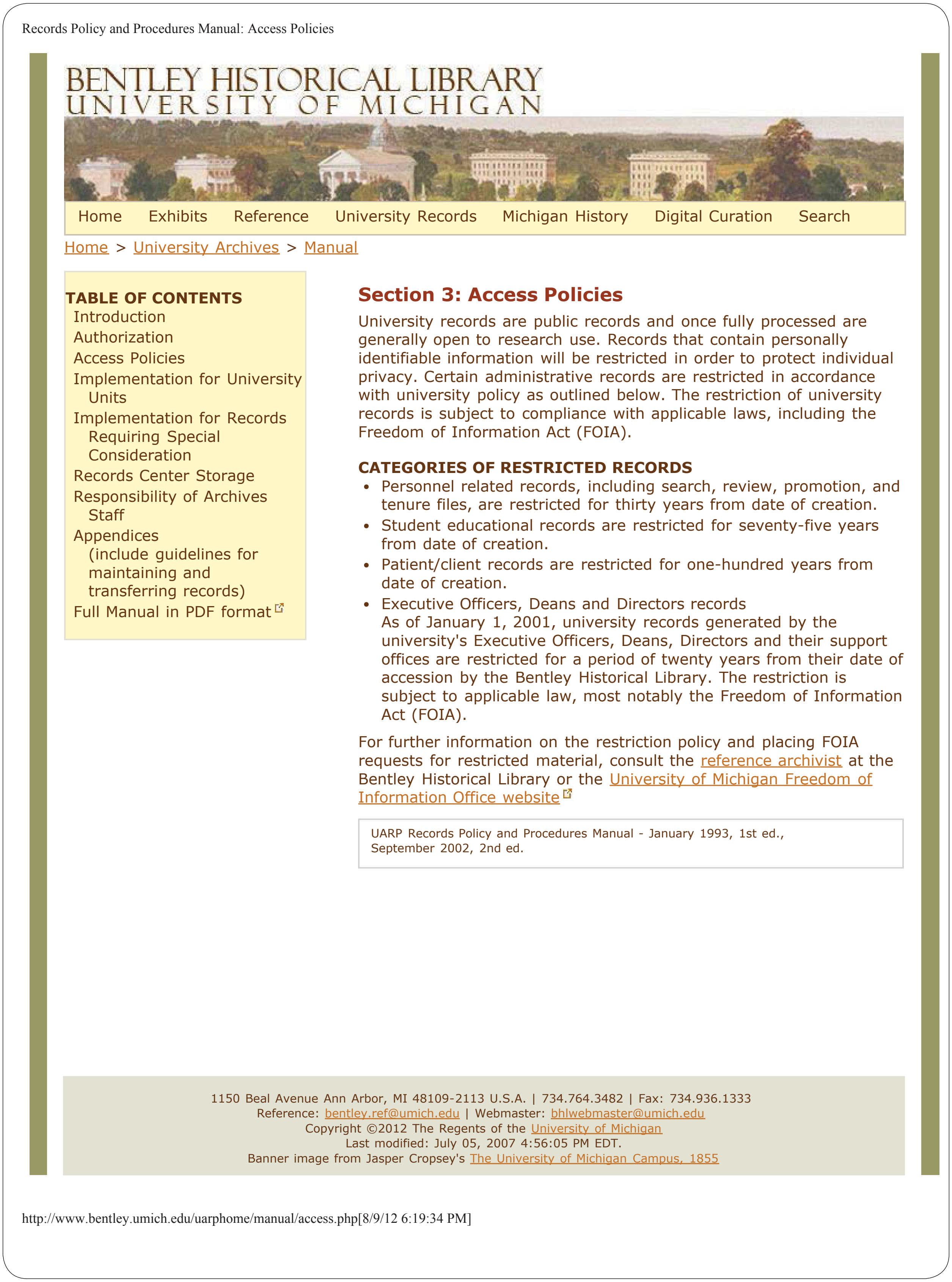
____Materials covered by attorney-client privilege

____Research data related to human subjects

____Federally Classified or Federally restricted materials

____Other materials that have specific privacy concerns, please specify____________________

UNIVERSITY OF MICHIGAN
Section 3: Access Policies
http://www.bentley.umich.edu/uarphome/manual/access.php



Records Policy and Procedures Manual: Access Policies

Home > University Archives > Manual

TABLE OF CONTENTS

Section 3: Access Policies

University records are public records and once fully processed are generally open to research use. Records that contain personally identifiable information will be restricted in order to protect individual privacy. Certain administrative records are restricted in accordance with university policy as outlined below. The restriction of university records is subject to compliance with applicable laws, including the Freedom of Information Act (FOIA).

CATEGORIES OF RESTRICTED RECORDS

- Personnel related records, including search, review, promotion, and tenure files, are restricted for thirty years from date of creation.
- Student educational records are restricted for seventy-five years from date of creation.
- Patient/client records are restricted for one-hundred years from date of creation.
- Executive Officers, Deans and Directors records
  As of January 1, 2001, university records generated by the university's Executive Officers, Deans, Directors and their support offices are restricted for a period of twenty years from their date of accession by the Bentley Historical Library. The restriction is subject to applicable law, most notably the Freedom of Information Act (FOIA).

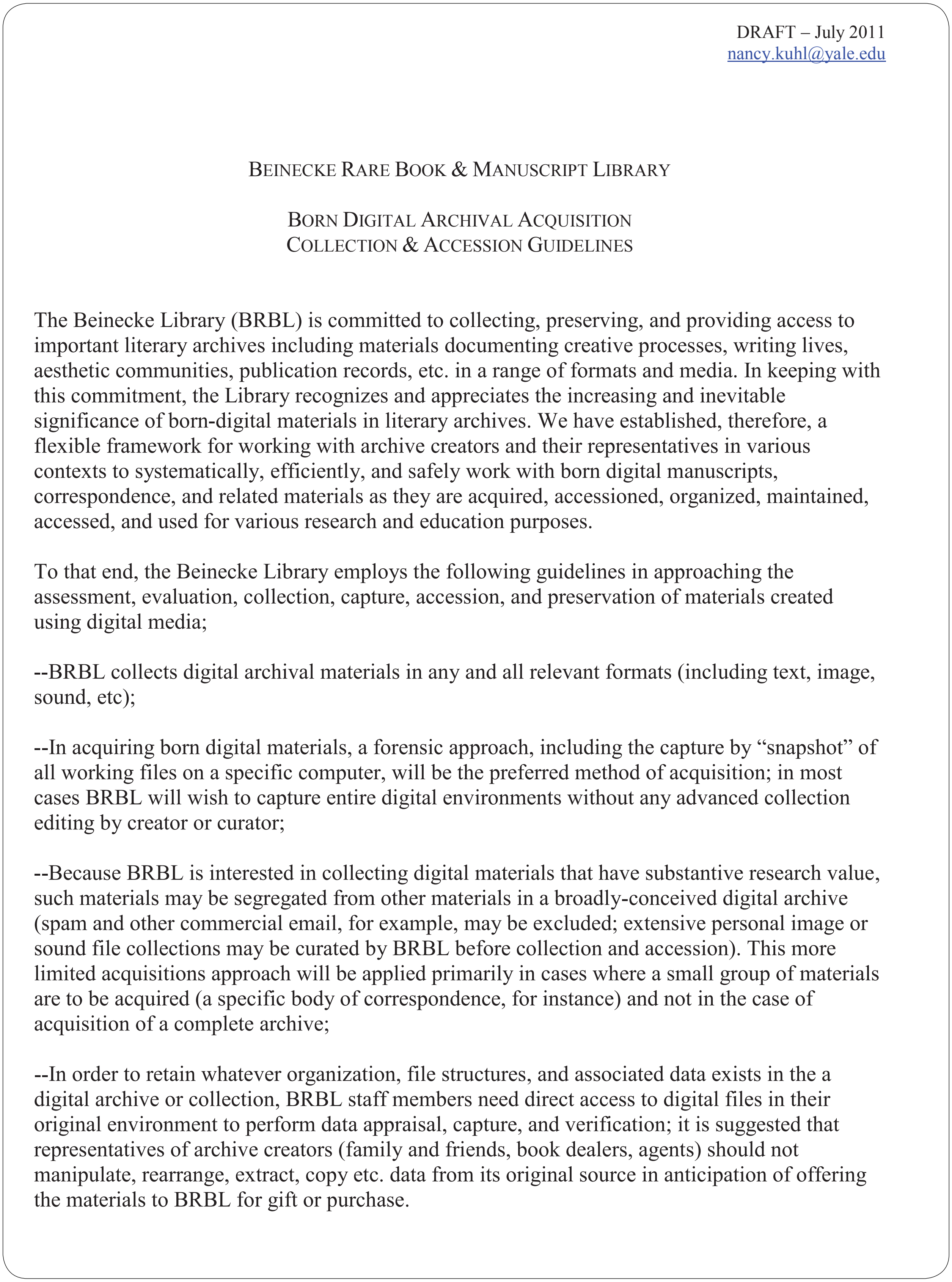
For further information on the restriction policy and placing FOIA requests for restricted material, consult the reference archivist at the Bentley Historical Library or the University of Michigan Freedom of Information Office website

UARP Records Policy and Procedures Manual - January 1993, 1st ed., September 2002, 2nd ed.

1150 Beal Avenue Ann Arbor, MI 48109-2113 U.S.A. | 734.764.3482 | Fax: 734.936.1333
Reference: bentley.ref@umich.edu | Webmaster: bhlwebmaster@umich.edu

Last modified: July 05, 2007 4:56:05 PM EDT.
Banner image from Jasper Cropsey's The University of Michigan Campus, 1855

http://www.bentley.umich.edu/uarphome/manual/access.php[8/9/12 6:19:34 PM]



DRAFT – July 2011
nancy.kuhl@yale.edu

BEINECKE RARE BOOK & MANUSCRIPT LIBRARY

BORN DIGITAL ARCHIVAL ACQUISITION
COLLECTION & ACCESSION GUIDELINES

The Beinecke Library (BRBL) is committed to collecting, preserving, and providing access to important literary archives including materials documenting creative processes, writing lives, aesthetic communities, publication records, etc. in a range of formats and media. In keeping with this commitment, the Library recognizes and appreciates the increasing and inevitable significance of born-digital materials in literary archives. We have established, therefore, a flexible framework for working with archive creators and their representatives in various contexts to systematically, efficiently, and safely work with born digital manuscripts, correspondence, and related materials as they are acquired, accessioned, organized, maintained, accessed, and used for various research and education purposes.

To that end, the Beinecke Library employs the following guidelines in approaching the assessment, evaluation, collection, capture, accession, and preservation of materials created using digital media;

--BRBL collects digital archival materials in any and all relevant formats (including text, image, sound, etc);

--In acquiring born digital materials, a forensic approach, including the capture by “snapshot” of all working files on a specific computer, will be the preferred method of acquisition; in most cases BRBL will wish to capture entire digital environments without any advanced collection editing by creator or curator;

--Because BRBL is interested in collecting digital materials that have substantive research value, such materials may be segregated from other materials in a broadly-conceived digital archive (spam and other commercial email, for example, may be excluded; extensive personal image or sound file collections may be curated by BRBL before collection and accession). This more limited acquisitions approach will be applied primarily in cases where a small group of materials are to be acquired (a specific body of correspondence, for instance) and not in the case of acquisition of a complete archive;

--In order to retain whatever organization, file structures, and associated data exists in the a digital archive or collection, BRBL staff members need direct access to digital files in their original environment to perform data appraisal, capture, and verification; it is suggested that representatives of archive creators (family and friends, book dealers, agents) should not manipulate, rearrange, extract, copy etc. data from its original source in anticipation of offering the materials to BRBL for gift or purchase.

Beinecke Deed of Gift section applying to curation of born-digital material

6. Terms and Conditions

Yale has accepted Donor’s gift of the Property, subject to the following terms and conditions:

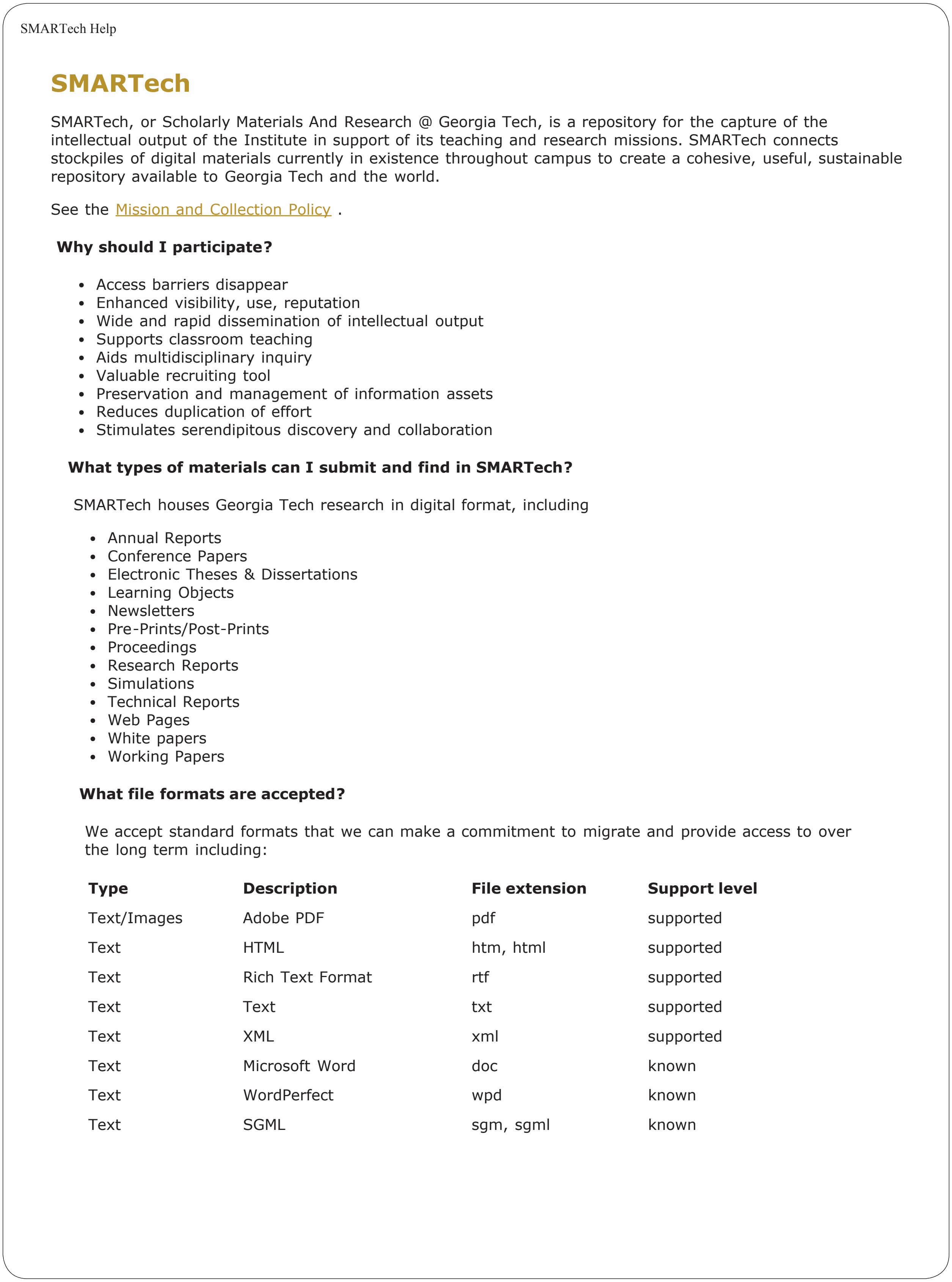
B. Donor acknowledges and agrees that upon execution of this Deed of Gift, the Property shall irrevocably become the property of Yale. Donor further acknowledges and agrees that the administration, use, physical display, care, treatment, preservation, conversation, and/or maintenance of the Property, including without limitation any conversion or transferral of the Property into microform, digital format, or any other format or medium now existing or hereinafter devised, shall be at Yale’s sole discretion, unless otherwise provided for in this agreement.

Format Policies

GEORGIA TECH

SMARTech Help

http://smartech.gatech.edu/about



SMARTech Help

SMARTech

SMARTech, or Scholarly Materials And Research @ Georgia Tech, is a repository for the capture of the intellectual output of the Institute in support of its teaching and research missions. SMARTech connects stockpiles of digital materials currently in existence throughout campus to create a cohesive, useful, sustainable repository available to Georgia Tech and the world.

See the Mission and Collection Policy .

Why should I participate?

- Access barriers disappear
- Enhanced visibility, use, reputation
- Wide and rapid dissemination of intellectual output
- Supports classroom teaching
- Aids multidisciplinary inquiry
- Valuable recruiting tool
- Preservation and management of information assets
- Reduces duplication of effort
- Stimulates serendipitous discovery and collaboration

What types of materials can I submit and find in SMARTech?

SMARTech houses Georgia Tech research in digital format, including

- Annual Reports
- Conference Papers
- Electronic Theses & Dissertations
- Learning Objects
- Newsletters
- Pre-Prints/Post-Prints
- Proceedings
- Research Reports
- Simulations
- Technical Reports
- Web Pages
- White papers
- Working Papers

What file formats are accepted?

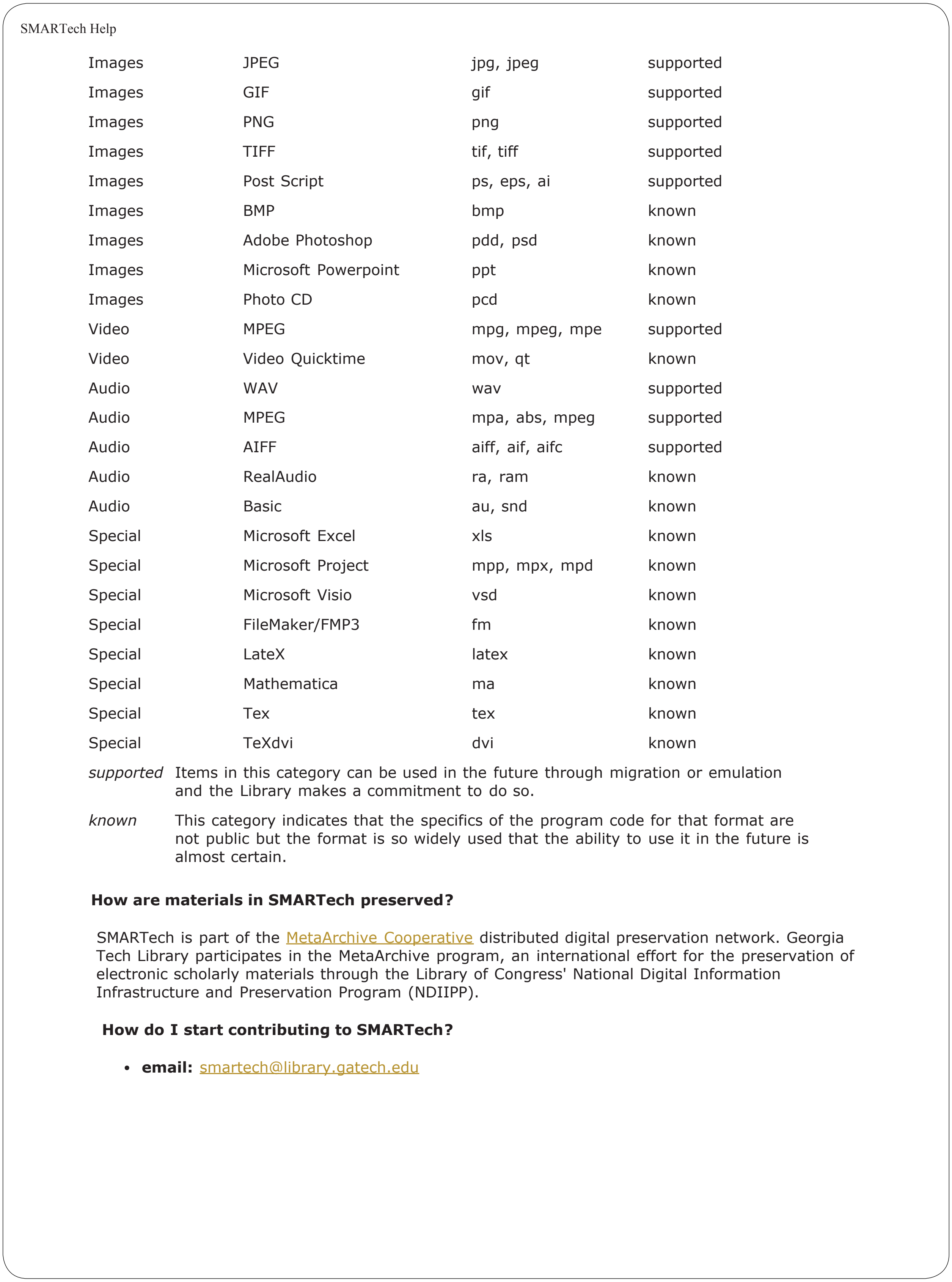
We accept standard formats that we can make a commitment to migrate and provide access to over the long term including:

Type	Description	File extension	Support level
Text/Images	Adobe PDF	pdf	supported
Text	HTML	htm, html	supported
Text	Rich Text Format	rtf	supported
Text	Text	txt	supported
Text	XML	xml	supported
Text	Microsoft Word	doc	known
Text	WordPerfect	wpd	known
Text	SGML	sgm, sgml	known

GEORGIA TECH

SMARTech Help

http://smartech.gatech.edu/about



SMARTech Help

Images	JPEG	jpg, jpeg	supported
Images	GIF	gif	supported
Images	PNG	png	supported
Images	TIFF	tif, tiff	supported
Images	Post Script	ps, eps, ai	supported
Images	BMP	bmp	known
Images	Adobe Photoshop	pdd, psd	known
Images	Microsoft Powerpoint	ppt	known
Images	Photo CD	pcd	known
Video	MPEG	mpg, mpeg, mpe	supported
Video	Video Quicktime	mov, qt	known
Audio	WAV	wav	supported
Audio	MPEG	mpa, abs, mpeg	supported
Audio	AIFF	aiff, aif, aifc	supported
Audio	RealAudio	ra, ram	known
Audio	Basic	au, snd	known
Special	Microsoft Excel	xls	known
Special	Microsoft Project	mpp, mpx, mpd	known
Special	Microsoft Visio	vsd	known
Special	FileMaker/FMP3	fm	known
Special	LateX	latex	known
Special	Mathematica	ma	known
Special	Tex	tex	known
Special	TeXdvi	dvi	known

supported Items in this category can be used in the future through migration or emulation and the Library makes a commitment to do so.

known This category indicates that the specifics of the program code for that format are not public but the format is so widely used that the ability to use it in the future is almost certain.

How are materials in SMARTech preserved?

SMARTech is part of the MetaArchive Cooperative distributed digital preservation network. Georgia Tech Library participates in the MetaArchive program, an international effort for the preservation of electronic scholarly materials through the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP).

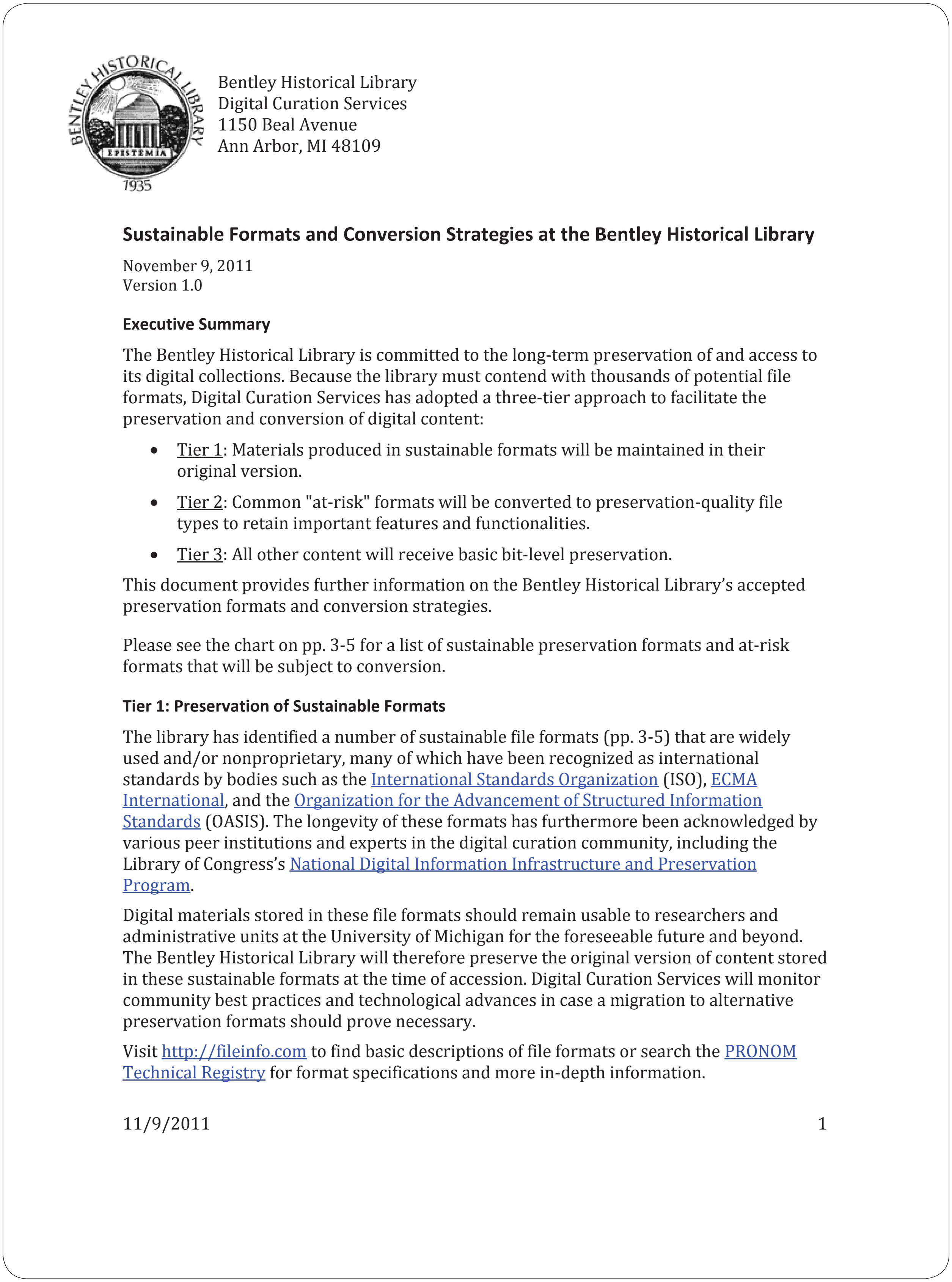
How do I start contributing to SMARTech?

- **email:** smartech@library.gatech.edu

UNIVERSITY OF MICHIGAN

Sustainable Formats and Conservation Strategies at the BHL

http://bentley.umich.edu/dchome/resources/BHL_PreservationStrategies_v01.pdf



Bentley Historical Library
Digital Curation Services
1150 Beal Avenue
Ann Arbor, MI 48109

Sustainable Formats and Conversion Strategies at the Bentley Historical Library

November 9, 2011
Version 1.0

Executive Summary

The Bentley Historical Library is committed to the long-term preservation of and access to its digital collections. Because the library must contend with thousands of potential file formats, Digital Curation Services has adopted a three-tier approach to facilitate the preservation and conversion of digital content:

- Tier 1: Materials produced in sustainable formats will be maintained in their original version.
- Tier 2: Common "at-risk" formats will be converted to preservation-quality file types to retain important features and functionalities.
- Tier 3: All other content will receive basic bit-level preservation.

This document provides further information on the Bentley Historical Library's accepted preservation formats and conversion strategies.

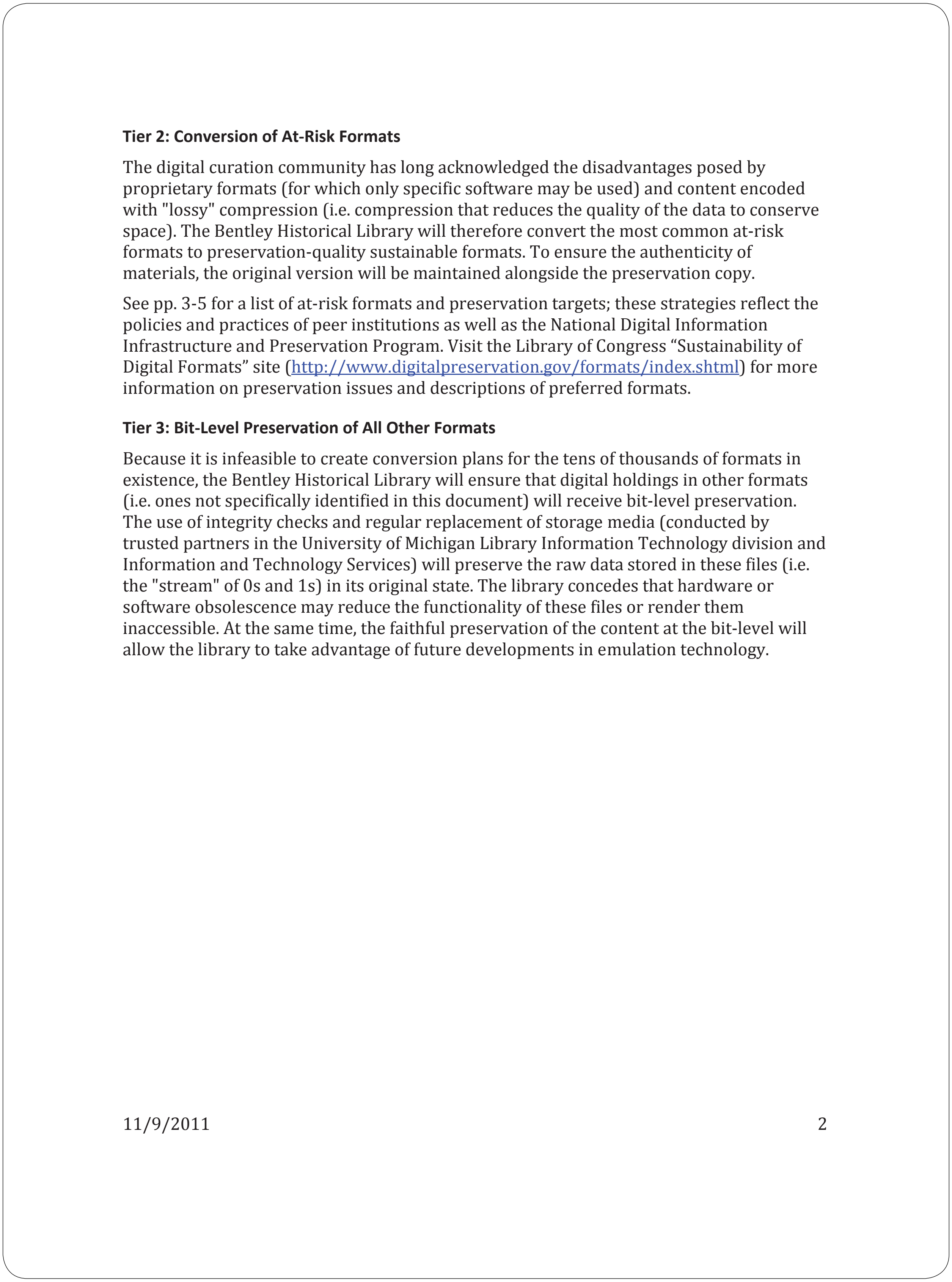
Please see the chart on pp. 3-5 for a list of sustainable preservation formats and at-risk formats that will be subject to conversion.

Tier 1: Preservation of Sustainable Formats

The library has identified a number of sustainable file formats (pp. 3-5) that are widely used and/or nonproprietary, many of which have been recognized as international standards by bodies such as the International Standards Organization (ISO), ECMA International, and the Organization for the Advancement of Structured Information Standards (OASIS). The longevity of these formats has furthermore been acknowledged by various peer institutions and experts in the digital curation community, including the Library of Congress's National Digital Information Infrastructure and Preservation Program.

Digital materials stored in these file formats should remain usable to researchers and administrative units at the University of Michigan for the foreseeable future and beyond. The Bentley Historical Library will therefore preserve the original version of content stored in these sustainable formats at the time of accession. Digital Curation Services will monitor community best practices and technological advances in case a migration to alternative preservation formats should prove necessary.

Visit http://fileinfo.com to find basic descriptions of file formats or search the PRONOM Technical Registry for format specifications and more in-depth information.



Tier 2: Conversion of At-Risk Formats

The digital curation community has long acknowledged the disadvantages posed by proprietary formats (for which only specific software may be used) and content encoded with "lossy" compression (i.e. compression that reduces the quality of the data to conserve space). The Bentley Historical Library will therefore convert the most common at-risk formats to preservation-quality sustainable formats. To ensure the authenticity of materials, the original version will be maintained alongside the preservation copy.

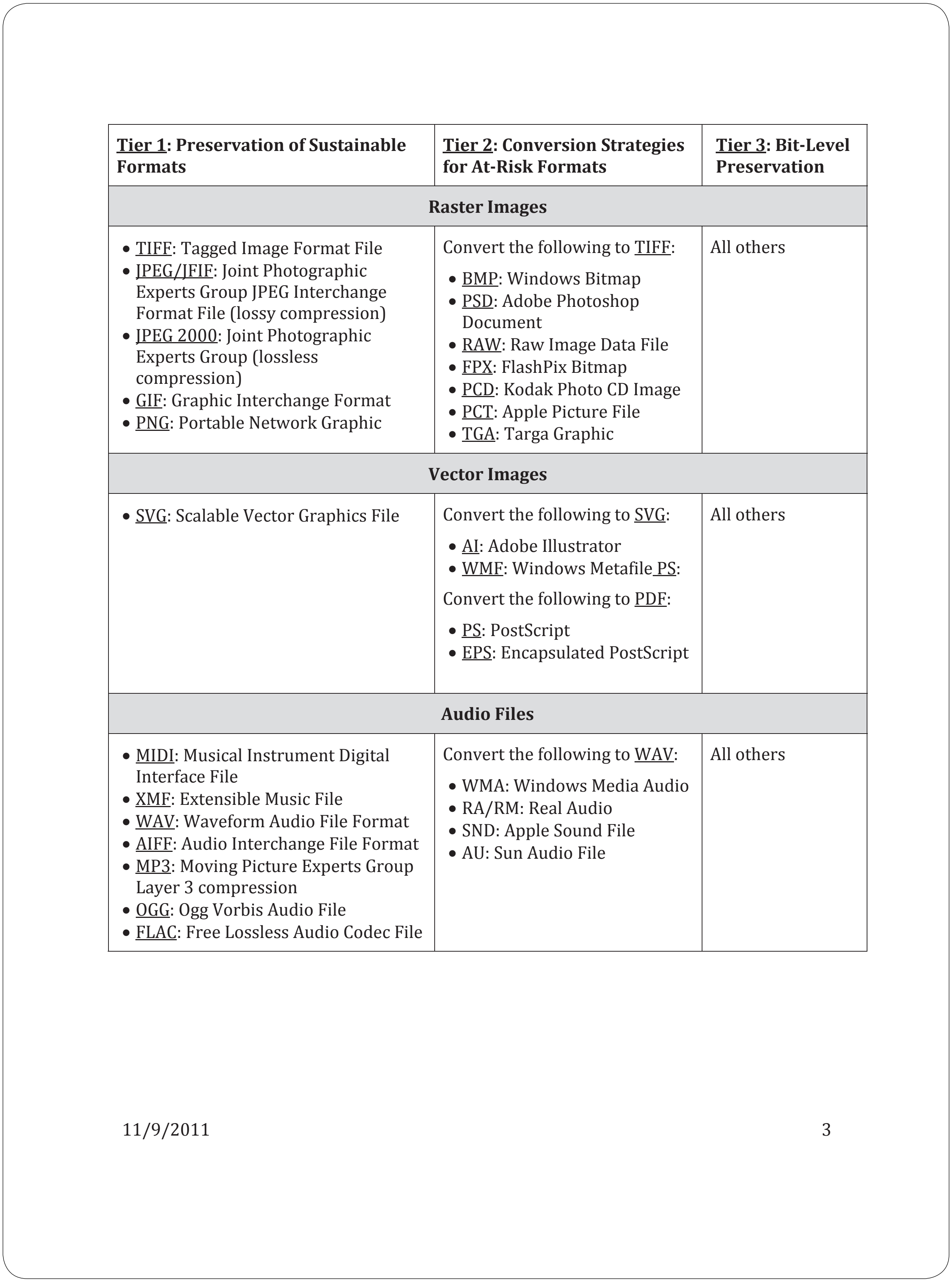
See pp. 3-5 for a list of at-risk formats and preservation targets; these strategies reflect the policies and practices of peer institutions as well as the National Digital Information Infrastructure and Preservation Program. Visit the Library of Congress “Sustainability of Digital Formats” site (http://www.digitalpreservation.gov/formats/index.shtml) for more information on preservation issues and descriptions of preferred formats.

Tier 3: Bit-Level Preservation of All Other Formats

Because it is infeasible to create conversion plans for the tens of thousands of formats in existence, the Bentley Historical Library will ensure that digital holdings in other formats (i.e. ones not specifically identified in this document) will receive bit-level preservation. The use of integrity checks and regular replacement of storage media (conducted by trusted partners in the University of Michigan Library Information Technology division and Information and Technology Services) will preserve the raw data stored in these files (i.e. the "stream" of 0s and 1s) in its original state. The library concedes that hardware or software obsolescence may reduce the functionality of these files or render them inaccessible. At the same time, the faithful preservation of the content at the bit-level will allow the library to take advantage of future developments in emulation technology.

UNIVERSITY OF MICHIGAN

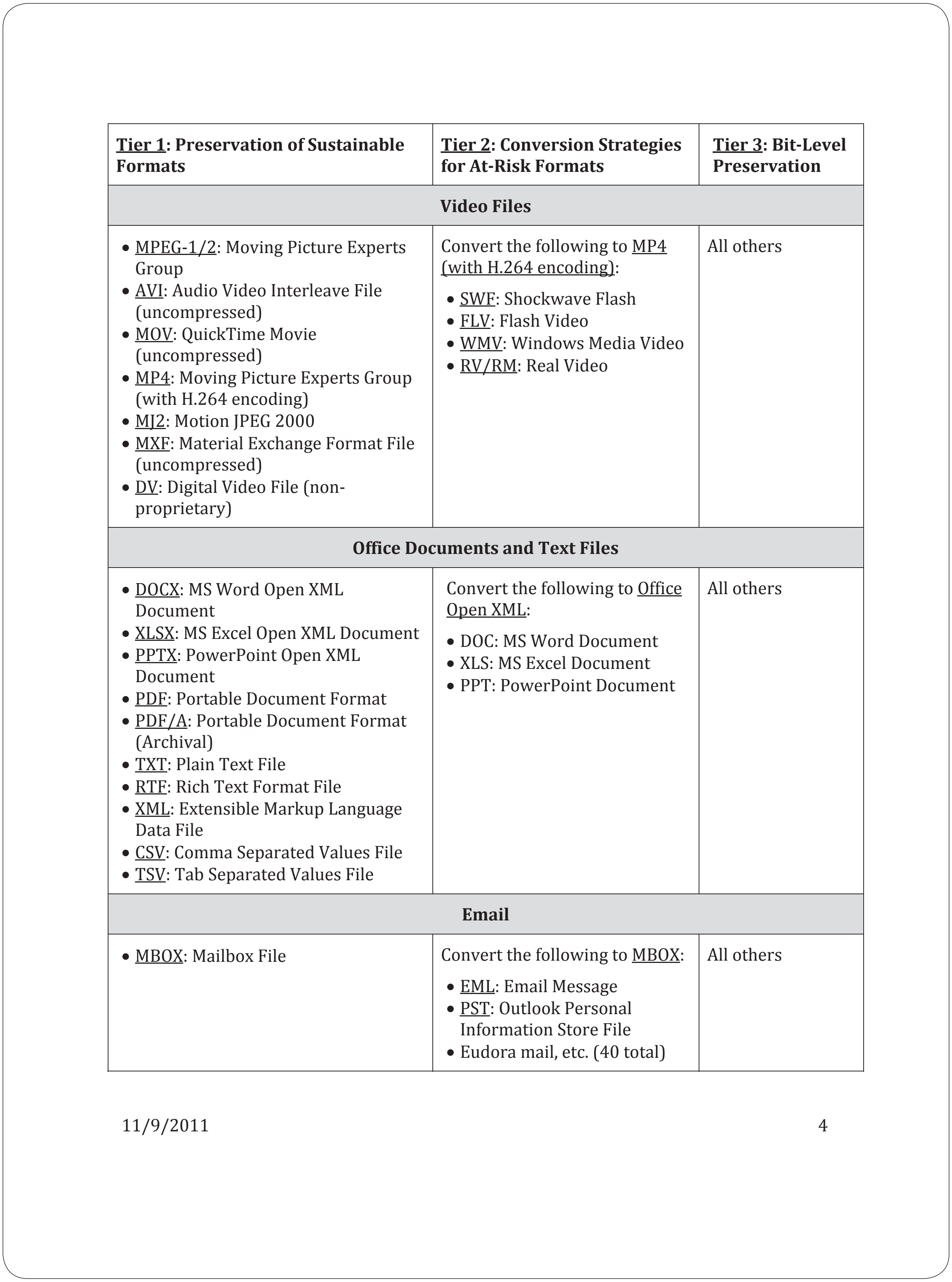
Sustainable Formats and Conservation Strategies at the BHL
http://bentley.umich.edu/dchome/resources/BHL_PreservationStrategies_v01.pdf



Tier 1: Preservation of Sustainable Formats	Tier 2: Conversion Strategies for At-Risk Formats	Tier 3: Bit-Level Preservation
Raster Images		
• TIFF: Tagged Image Format File • JPEG/JFIF: Joint Photographic Experts Group JPEG Interchange Format File (lossy compression) • JPEG 2000: Joint Photographic Experts Group (lossless compression) • GIF: Graphic Interchange Format • PNG: Portable Network Graphic	Convert the following to TIFF: • BMP: Windows Bitmap • PSD: Adobe Photoshop Document • RAW: Raw Image Data File • FPX: FlashPix Bitmap • PCD: Kodak Photo CD Image • PCT: Apple Picture File • TGA: Targa Graphic	All others
Vector Images		
• SVG: Scalable Vector Graphics File	Convert the following to SVG: • AI: Adobe Illustrator • WMF: Windows Metafile PS: Convert the following to PDF: • PS: PostScript • EPS: Encapsulated PostScript	All others
Audio Files		
• MIDI: Musical Instrument Digital Interface File • XMF: Extensible Music File • WAV: Waveform Audio File Format • AIFF: Audio Interchange File Format • MP3: Moving Picture Experts Group Layer 3 compression • OGG: Ogg Vorbis Audio File • FLAC: Free Lossless Audio Codec File	Convert the following to WAV: • WMA: Windows Media Audio • RA/RM: Real Audio • SND: Apple Sound File • AU: Sun Audio File	All others

UNIVERSITY OF MICHIGAN

Sustainable Formats and Conservation Strategies at the BHL
http://bentley.umich.edu/dchome/resources/BHL_PreservationStrategies_v01.pdf



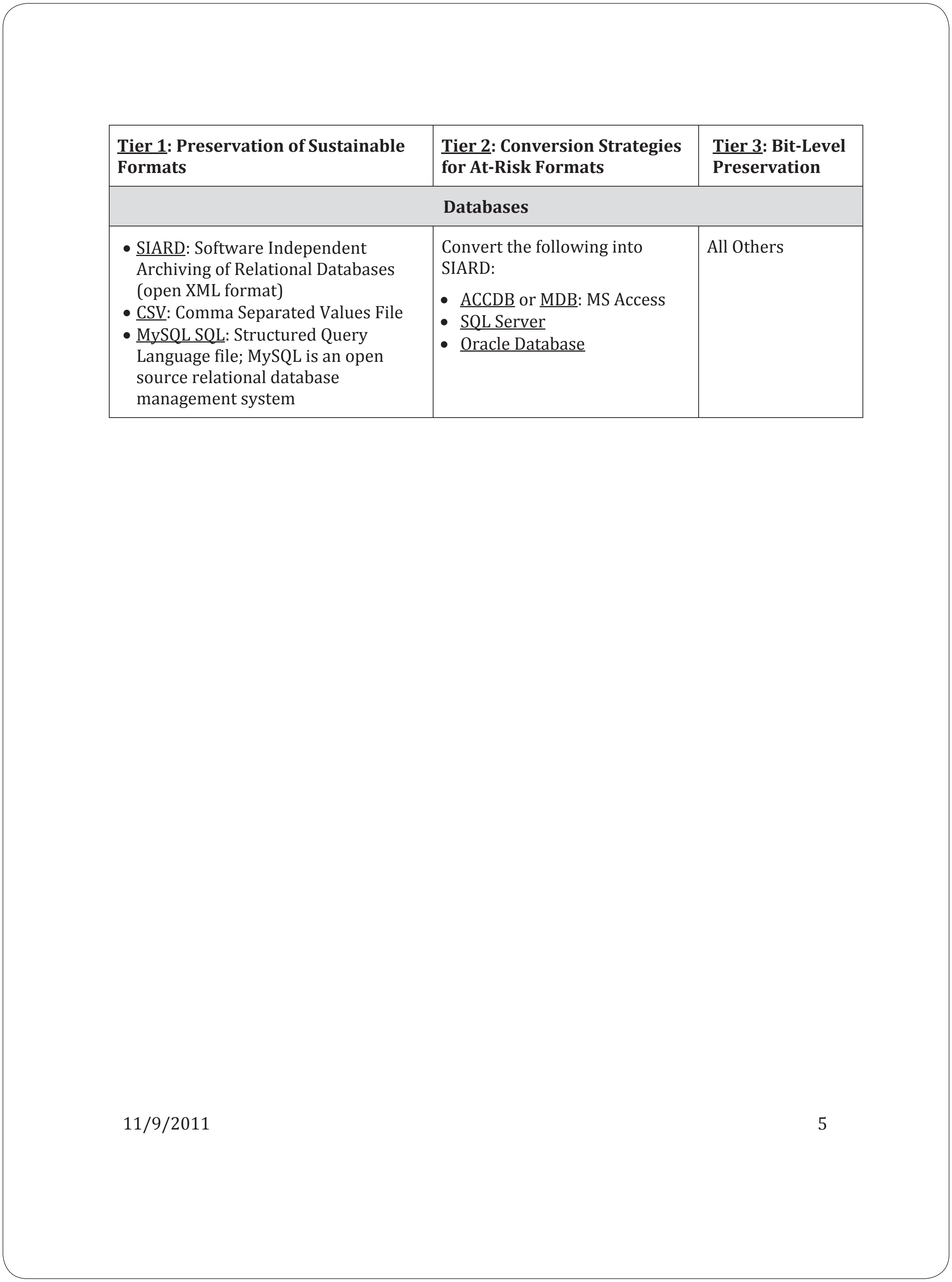
Tier 1: Preservation of Sustainable Formats	**Tier 2: Conversion Strategies for At-Risk Formats**	**Tier 3: Bit-Level Preservation**
Video Files		
• MPEG-1/2: Moving Picture Experts Group • AVI: Audio Video Interleave File (uncompressed) • MOV: QuickTime Movie (uncompressed) • MP4: Moving Picture Experts Group (with H.264 encoding) • MJ2: Motion JPEG 2000 • MXF: Material Exchange Format File (uncompressed) • DV: Digital Video File (non-proprietary)	Convert the following to MP4 (with H.264 encoding): • SWF: Shockwave Flash • FLV: Flash Video • WMV: Windows Media Video • RV/RM: Real Video	All others
Office Documents and Text Files		
• DOCX: MS Word Open XML Document • XLSX: MS Excel Open XML Document • PPTX: PowerPoint Open XML Document • PDF: Portable Document Format • PDF/A: Portable Document Format (Archival) • TXT: Plain Text File • RTF: Rich Text Format File • XML: Extensible Markup Language Data File • CSV: Comma Separated Values File • TSV: Tab Separated Values File	Convert the following to Office Open XML: • DOC: MS Word Document • XLS: MS Excel Document • PPT: PowerPoint Document	All others
Email		
• MBOX: Mailbox File	Convert the following to MBOX: • EML: Email Message • PST: Outlook Personal Information Store File • Eudora mail, etc. (40 total)	All others

11/9/2011 4

UNIVERSITY OF MICHIGAN

Sustainable Formats and Conservation Strategies at the BHL

http://bentley.umich.edu/dchome/resources/BHL_PreservationStrategies_v01.pdf

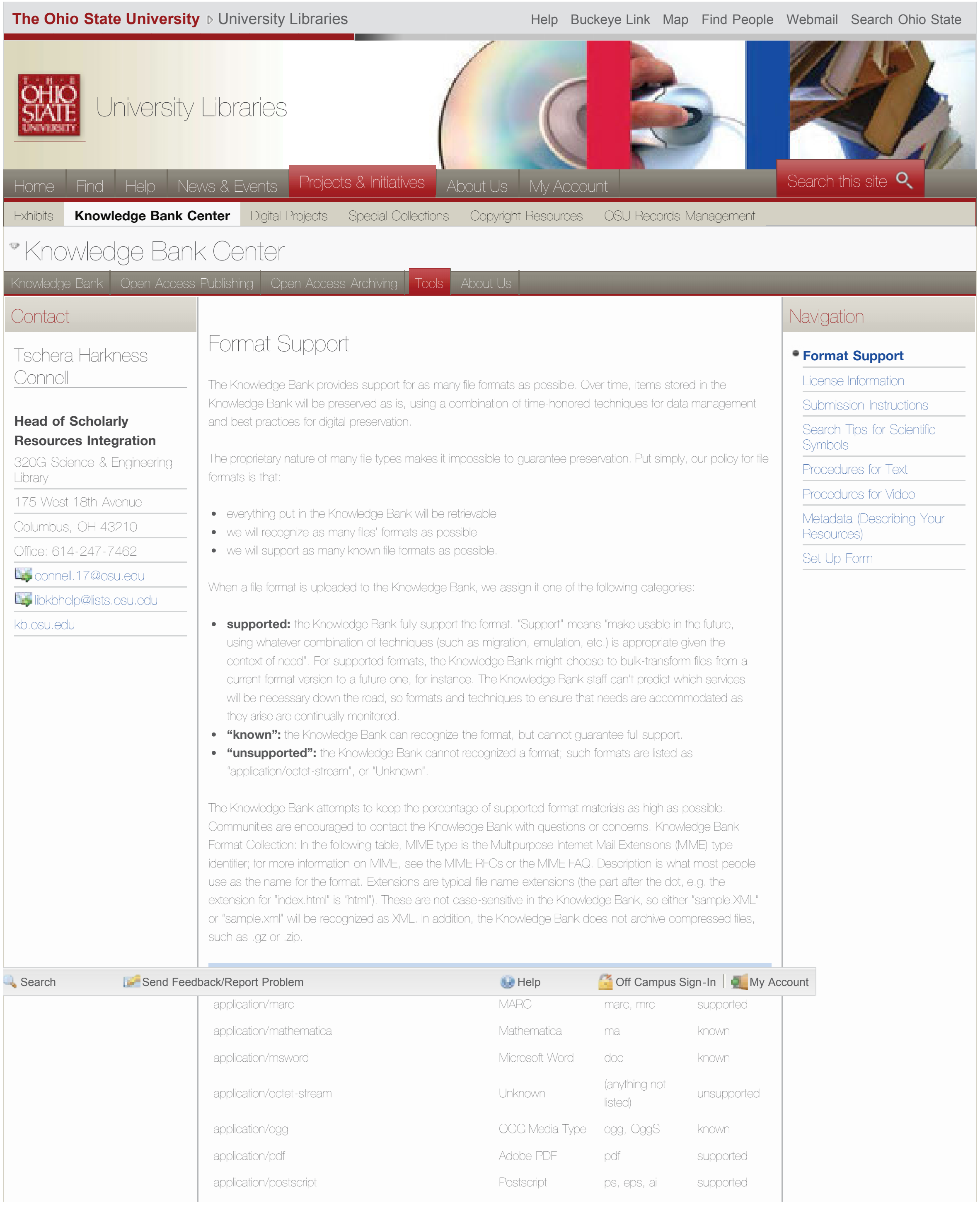


Tier 1: Preservation of Sustainable Formats	**Tier 2: Conversion Strategies for At-Risk Formats**	**Tier 3: Bit-Level Preservation**
	Databases	
• SIARD: Software Independent Archiving of Relational Databases (open XML format) • CSV: Comma Separated Values File • MySQL SQL: Structured Query Language file; MySQL is an open source relational database management system	Convert the following into SIARD: • ACCDB or MDB: MS Access • SQL Server • Oracle Database	All Others

OHIO STATE UNIVERSITY

Format Support

http://library.osu.edu/projects-initiatives/knowledge-bank/tools/format-support/


The Ohio State University ▹ University Libraries
Help Buckeye Link Map Find People Webmail Search Ohio State
University Libraries
Home Find Help News & Events Projects & Initiatives About Us My Account
Search this site
Exhibits Knowledge Bank Center Digital Projects Special Collections Copyright Resources OSU Records Management
Knowledge Bank Center
Knowledge Bank Open Access Publishing Open Access Archiving Tools About Us
Contact
Navigation
Format Support
License Information
Submission Instructions
Search Tips for Scientific Symbols
Procedures for Text
Procedures for Video
Metadata (Describing Your Resources)
Set Up Form
Search Send Feedback/Report Problem Help Off Campus Sign-In My Account

Tschera Harkness Connell

Head of Scholarly Resources Integration

320G Science & Engineering Library
175 West 18th Avenue
Columbus, OH 43210
Office: 614-247-7462
connell.17@osu.edu
libkbhelp@lists.osu.edu
kb.osu.edu

Format Support

The Knowledge Bank provides support for as many file formats as possible. Over time, items stored in the Knowledge Bank will be preserved as is, using a combination of time-honored techniques for data management and best practices for digital preservation.

The proprietary nature of many file types makes it impossible to guarantee preservation. Put simply, our policy for file formats is that:

- everything put in the Knowledge Bank will be retrievable
- we will recognize as many files' formats as possible
- we will support as many known file formats as possible.

When a file format is uploaded to the Knowledge Bank, we assign it one of the following categories:

- **supported:** the Knowledge Bank fully support the format. "Support" means "make usable in the future, using whatever combination of techniques (such as migration, emulation, etc.) is appropriate given the context of need". For supported formats, the Knowledge Bank might choose to bulk-transform files from a current format version to a future one, for instance. The Knowledge Bank staff can't predict which services will be necessary down the road, so formats and techniques to ensure that needs are accommodated as they arise are continually monitored.
- **"known":** the Knowledge Bank can recognize the format, but cannot guarantee full support.
- **"unsupported":** the Knowledge Bank cannot recognized a format; such formats are listed as "application/octet-stream", or "Unknown".

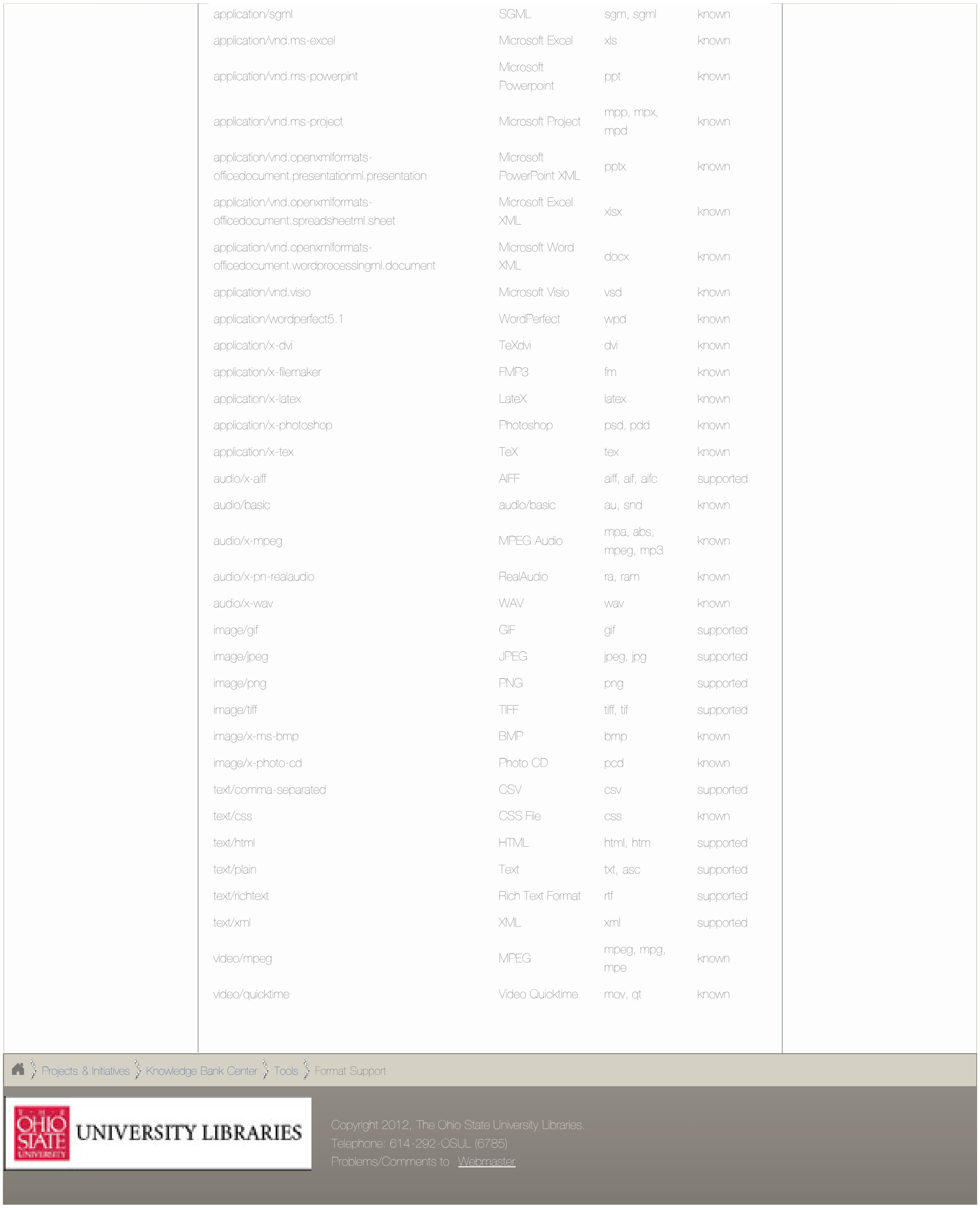
The Knowledge Bank attempts to keep the percentage of supported format materials as high as possible. Communities are encouraged to contact the Knowledge Bank with questions or concerns. Knowledge Bank Format Collection: In the following table, MIME type is the Multipurpose Internet Mail Extensions (MIME) type identifier; for more information on MIME, see the MIME RFCs or the MIME FAQ. Description is what most people use as the name for the format. Extensions are typical file name extensions (the part after the dot, e.g. the extension for "index.html" is "html"). These are not case-sensitive in the Knowledge Bank, so either "sample.XML" or "sample.xml" will be recognized as XML. In addition, the Knowledge Bank does not archive compressed files, such as .gz or .zip.

application/marc	MARC	marc, mrc	supported
application/mathematica	Mathematica	ma	known
application/msword	Microsoft Word	doc	known
application/octet-stream	Unknown	(anything not listed)	unsupported
application/ogg	OGG Media Type	ogg, OggS	known
application/pdf	Adobe PDF	pdf	supported
application/postscript	Postscript	ps, eps, ai	supported

OHIO STATE UNIVERSITY

Format Support

http://library.osu.edu/projects-initiatives/knowledge-bank/tools/format-support/



application/sgml	SGML	sgm, sgml	known
application/vnd.ms-excel	Microsoft Excel	xls	known
application/vnd.ms-powerpint	Microsoft Powerpoint	ppt	known
application/vnd.ms-project	Microsoft Project	mpp, mpx, mpd	known
application/vnd.openxmlformats-officedocument.presentationml.presentation	Microsoft PowerPoint XML	pptx	known
application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	Microsoft Excel XML	xlsx	known
application/vnd.openxmlformats-officedocument.wordprocessingml.document	Microsoft Word XML	docx	known
application/vnd.visio	Microsoft Visio	vsd	known
application/wordperfect5.1	WordPerfect	wpd	known
application/x-dvi	TeXdvi	dvi	known
application/x-filemaker	FMP3	fm	known
application/x-latex	LateX	latex	known
application/x-photoshop	Photoshop	psd, pdd	known
application/x-tex	TeX	tex	known
audio/x-aiff	AIFF	aiff, aif, aifc	supported
audio/basic	audio/basic	au, snd	known
audio/x-mpeg	MPEG Audio	mpa, abs, mpeg, mp3	known
audio/x-pn-realaudio	RealAudio	ra, ram	known
audio/x-wav	WAV	wav	known
image/gif	GIF	gif	supported
image/jpeg	JPEG	jpeg, jpg	supported
image/png	PNG	png	supported
image/tiff	TIFF	tiff, tif	supported
image/x-ms-bmp	BMP	bmp	known
image/x-photo-cd	Photo CD	pcd	known
text/comma-separated	CSV	csv	supported
text/css	CSS File	css	known
text/html	HTML	html, htm	supported
text/plain	Text	txt, asc	supported
text/richtext	Rich Text Format	rtf	supported
text/xml	XML	xml	supported
video/mpeg	MPEG	mpeg, mpg, mpe	known
video/quicktime	Video Quicktime	mov, qt	known

Projects & Initiatives › Knowledge Bank Center › Tools › Format Support

UNIVERSITY LIBRARIES

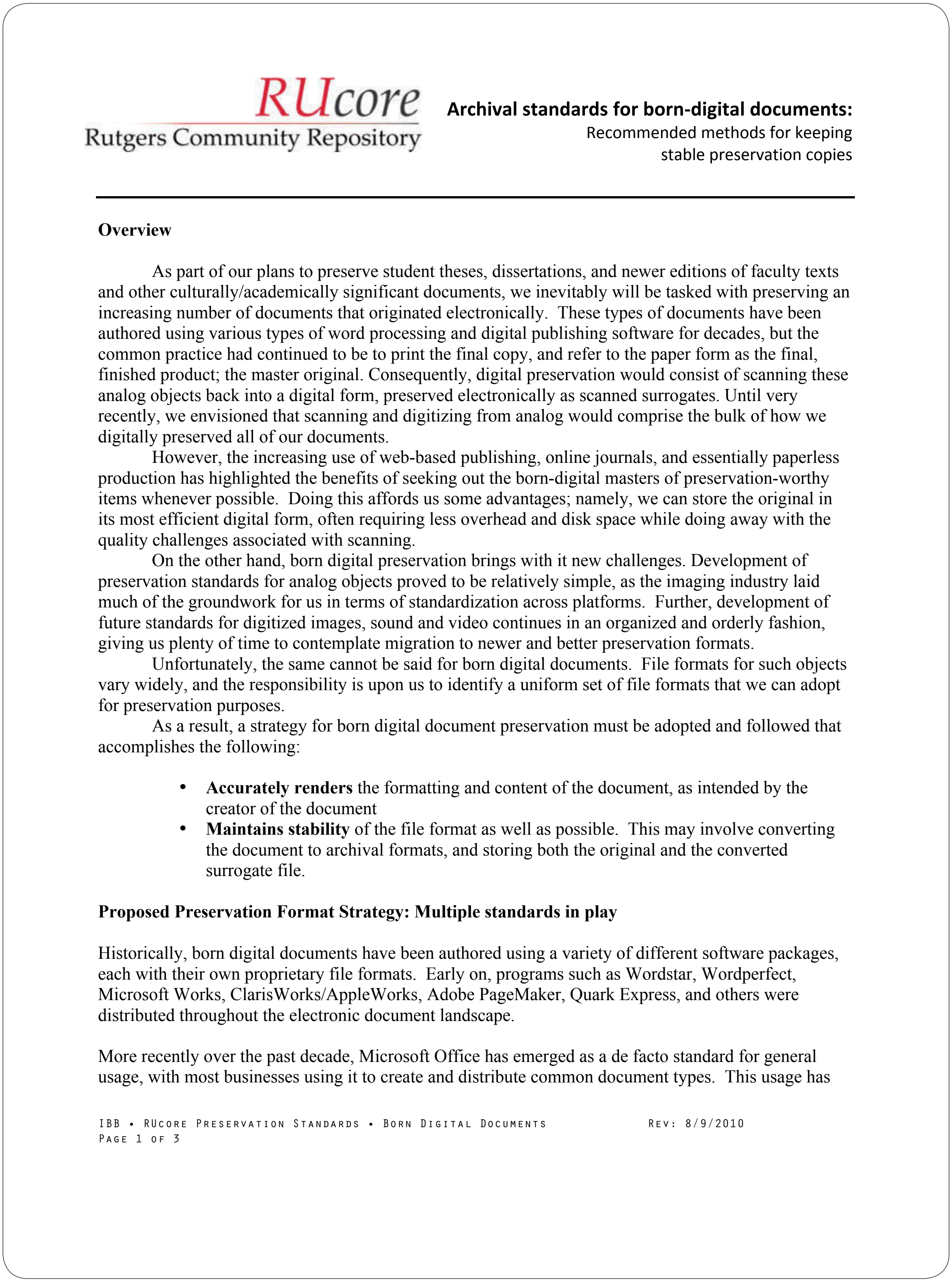
Telephone: 614-292-OSUL (6785)

Problems/Comments to Webmaster

RUTGERS UNIVERSITY

RUcore. Archival Standards for Born-Digital Documents

http://odin.page2pixel.org/standards/latest/RUcoreStandards-BornDigitalDocuments.pdf



RUcore

Rutgers Community Repository

Archival standards for born-digital documents:

Recommended methods for keeping
stable preservation copies

Overview

As part of our plans to preserve student theses, dissertations, and newer editions of faculty texts and other culturally/academically significant documents, we inevitably will be tasked with preserving an increasing number of documents that originated electronically. These types of documents have been authored using various types of word processing and digital publishing software for decades, but the common practice had continued to be to print the final copy, and refer to the paper form as the final, finished product; the master original. Consequently, digital preservation would consist of scanning these analog objects back into a digital form, preserved electronically as scanned surrogates. Until very recently, we envisioned that scanning and digitizing from analog would comprise the bulk of how we digitally preserved all of our documents.

However, the increasing use of web-based publishing, online journals, and essentially paperless production has highlighted the benefits of seeking out the born-digital masters of preservation-worthy items whenever possible. Doing this affords us some advantages; namely, we can store the original in its most efficient digital form, often requiring less overhead and disk space while doing away with the quality challenges associated with scanning.

On the other hand, born digital preservation brings with it new challenges. Development of preservation standards for analog objects proved to be relatively simple, as the imaging industry laid much of the groundwork for us in terms of standardization across platforms. Further, development of future standards for digitized images, sound and video continues in an organized and orderly fashion, giving us plenty of time to contemplate migration to newer and better preservation formats.

Unfortunately, the same cannot be said for born digital documents. File formats for such objects vary widely, and the responsibility is upon us to identify a uniform set of file formats that we can adopt for preservation purposes.

As a result, a strategy for born digital document preservation must be adopted and followed that accomplishes the following:

- **Accurately renders** the formatting and content of the document, as intended by the creator of the document
- **Maintains stability** of the file format as well as possible. This may involve converting the document to archival formats, and storing both the original and the converted surrogate file.

Proposed Preservation Format Strategy: Multiple standards in play

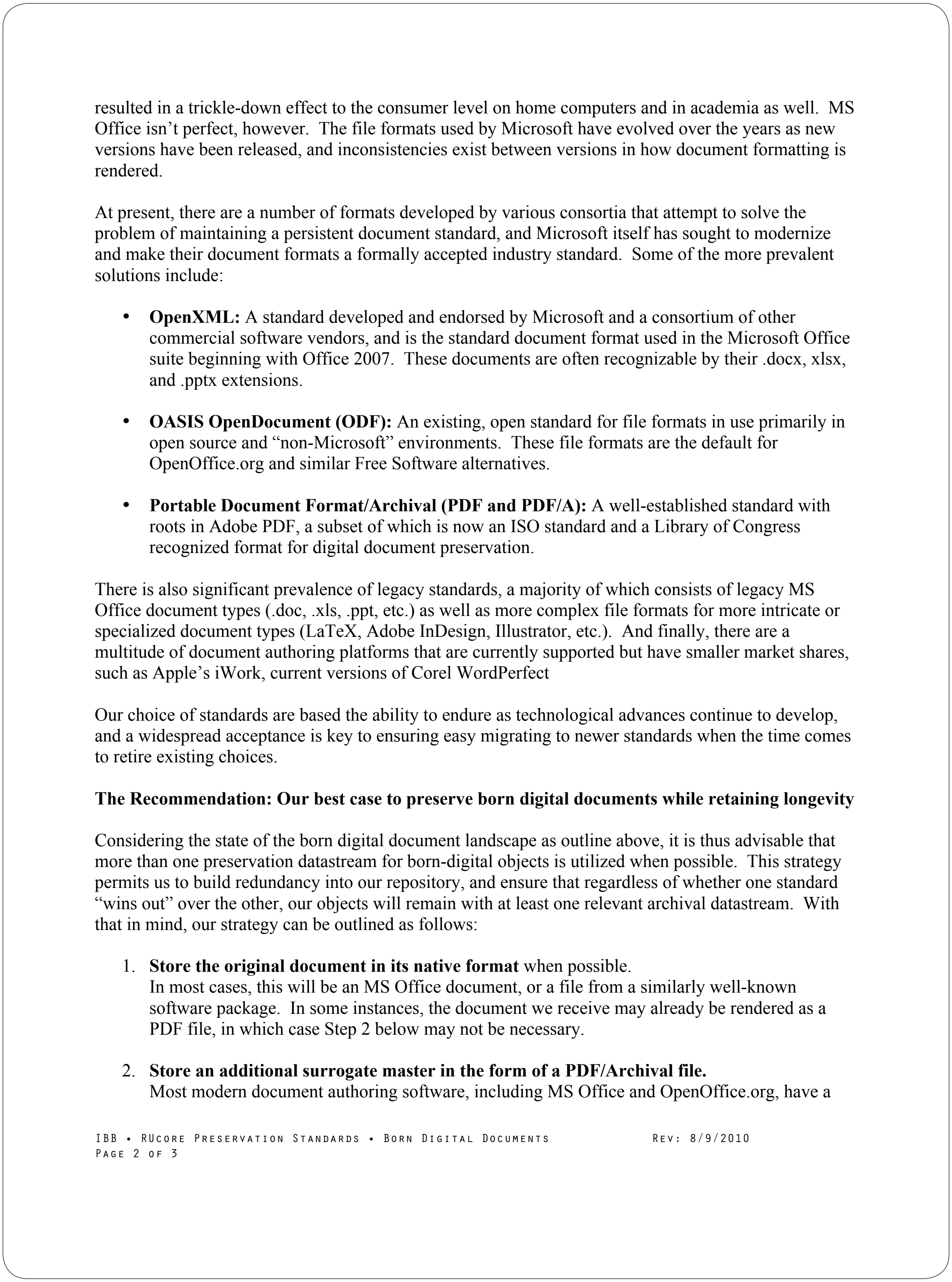
Historically, born digital documents have been authored using a variety of different software packages, each with their own proprietary file formats. Early on, programs such as Wordstar, Wordperfect, Microsoft Works, ClarisWorks/AppleWorks, Adobe PageMaker, Quark Express, and others were distributed throughout the electronic document landscape.

More recently over the past decade, Microsoft Office has emerged as a de facto standard for general usage, with most businesses using it to create and distribute common document types. This usage has

RUTGERS UNIVERSITY

RUcore. Archival Standards for Born-Digital Documents

http://odin.page2pixel.org/standards/latest/RUcoreStandards-BornDigitalDocuments.pdf



resulted in a trickle-down effect to the consumer level on home computers and in academia as well. MS Office isn't perfect, however. The file formats used by Microsoft have evolved over the years as new versions have been released, and inconsistencies exist between versions in how document formatting is rendered.

At present, there are a number of formats developed by various consortia that attempt to solve the problem of maintaining a persistent document standard, and Microsoft itself has sought to modernize and make their document formats a formally accepted industry standard. Some of the more prevalent solutions include:

- **OpenXML:** A standard developed and endorsed by Microsoft and a consortium of other commercial software vendors, and is the standard document format used in the Microsoft Office suite beginning with Office 2007. These documents are often recognizable by their .docx, xlsx, and .pptx extensions.

- **OASIS OpenDocument (ODF):** An existing, open standard for file formats in use primarily in open source and "non-Microsoft" environments. These file formats are the default for OpenOffice.org and similar Free Software alternatives.

- **Portable Document Format/Archival (PDF and PDF/A):** A well-established standard with roots in Adobe PDF, a subset of which is now an ISO standard and a Library of Congress recognized format for digital document preservation.

There is also significant prevalence of legacy standards, a majority of which consists of legacy MS Office document types (.doc, .xls, .ppt, etc.) as well as more complex file formats for more intricate or specialized document types (LaTeX, Adobe InDesign, Illustrator, etc.). And finally, there are a multitude of document authoring platforms that are currently supported but have smaller market shares, such as Apple's iWork, current versions of Corel WordPerfect

Our choice of standards are based the ability to endure as technological advances continue to develop, and a widespread acceptance is key to ensuring easy migrating to newer standards when the time comes to retire existing choices.

The Recommendation: Our best case to preserve born digital documents while retaining longevity

Considering the state of the born digital document landscape as outline above, it is thus advisable that more than one preservation datastream for born-digital objects is utilized when possible. This strategy permits us to build redundancy into our repository, and ensure that regardless of whether one standard "wins out" over the other, our objects will remain with at least one relevant archival datastream. With that in mind, our strategy can be outlined as follows:

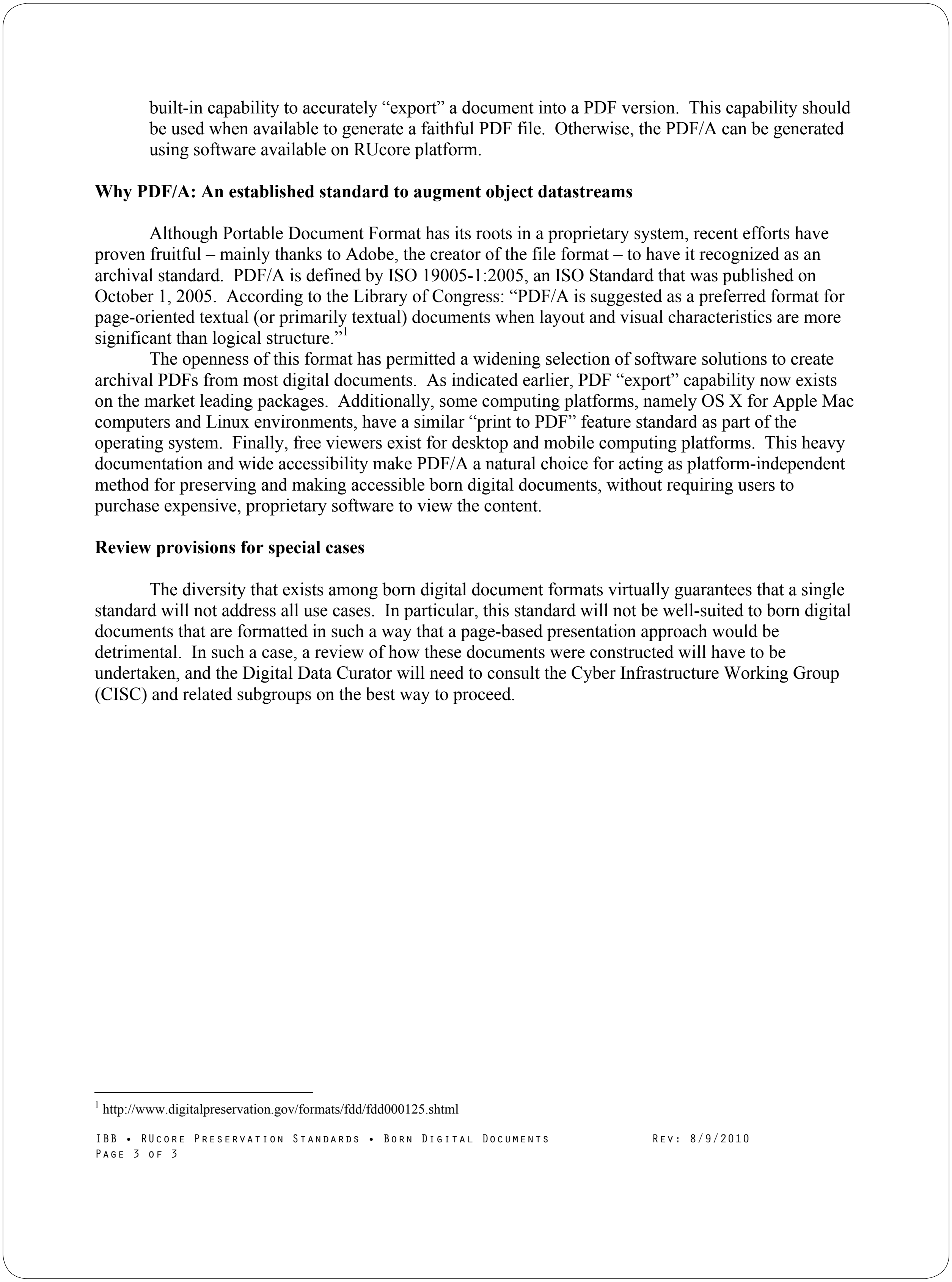
1. **Store the original document in its native format** when possible.
   In most cases, this will be an MS Office document, or a file from a similarly well-known software package. In some instances, the document we receive may already be rendered as a PDF file, in which case Step 2 below may not be necessary.

2. **Store an additional surrogate master in the form of a PDF/Archival file.**
   Most modern document authoring software, including MS Office and OpenOffice.org, have a

RUTGERS UNIVERSITY

RUcore. Archival Standards for Born-Digital Documents

http://odin.page2pixel.org/standards/latest/RUcoreStandards-BornDigitalDocuments.pdf



built-in capability to accurately "export" a document into a PDF version. This capability should be used when available to generate a faithful PDF file. Otherwise, the PDF/A can be generated using software available on RUcore platform.

Why PDF/A: An established standard to augment object datastreams

Although Portable Document Format has its roots in a proprietary system, recent efforts have proven fruitful – mainly thanks to Adobe, the creator of the file format – to have it recognized as an archival standard. PDF/A is defined by ISO 19005-1:2005, an ISO Standard that was published on October 1, 2005. According to the Library of Congress: "PDF/A is suggested as a preferred format for page-oriented textual (or primarily textual) documents when layout and visual characteristics are more significant than logical structure."[1]

The openness of this format has permitted a widening selection of software solutions to create archival PDFs from most digital documents. As indicated earlier, PDF "export" capability now exists on the market leading packages. Additionally, some computing platforms, namely OS X for Apple Mac computers and Linux environments, have a similar "print to PDF" feature standard as part of the operating system. Finally, free viewers exist for desktop and mobile computing platforms. This heavy documentation and wide accessibility make PDF/A a natural choice for acting as platform-independent method for preserving and making accessible born digital documents, without requiring users to purchase expensive, proprietary software to view the content.

Review provisions for special cases

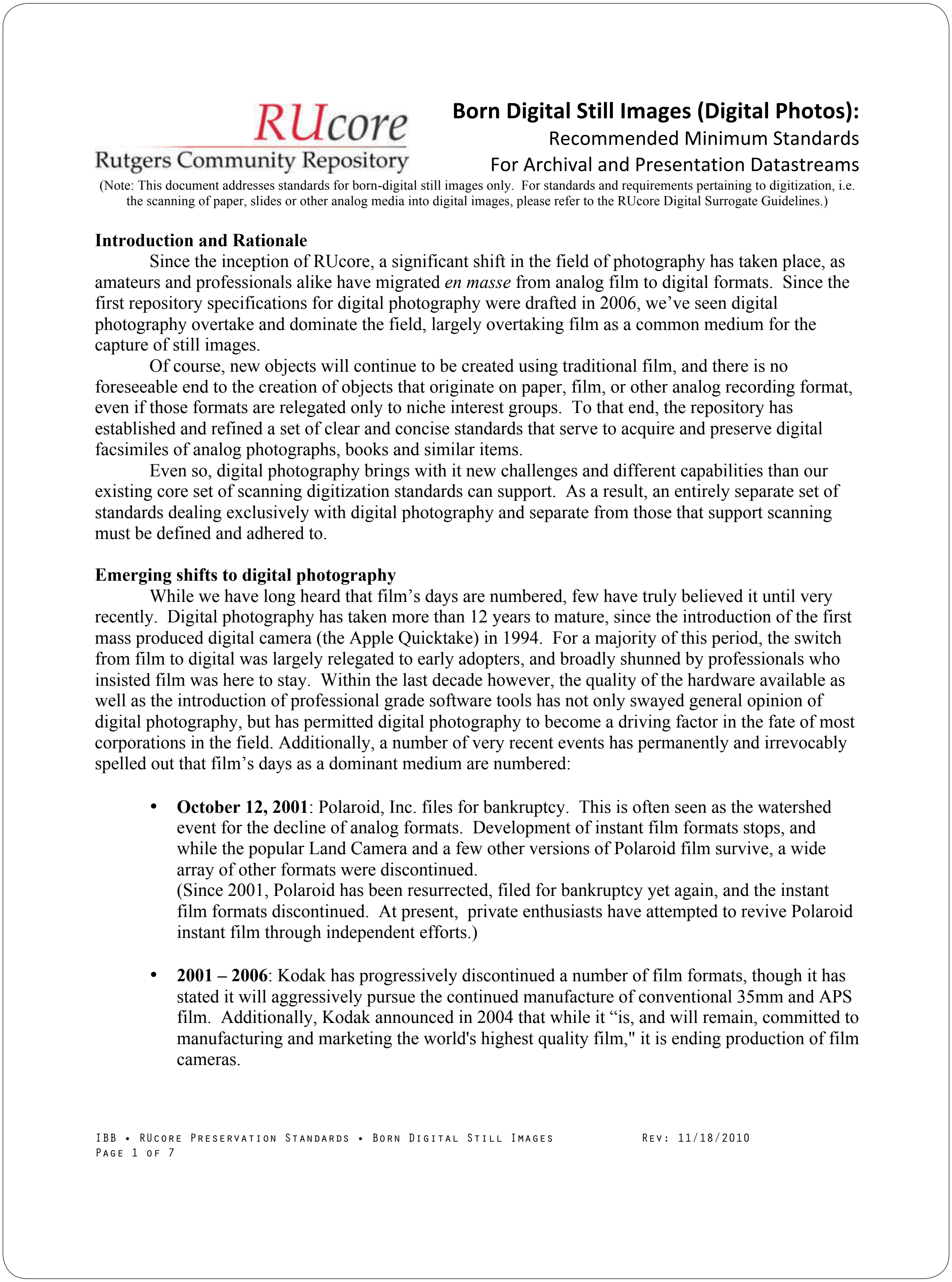
The diversity that exists among born digital document formats virtually guarantees that a single standard will not address all use cases. In particular, this standard will not be well-suited to born digital documents that are formatted in such a way that a page-based presentation approach would be detrimental. In such a case, a review of how these documents were constructed will have to be undertaken, and the Digital Data Curator will need to consult the Cyber Infrastructure Working Group (CISC) and related subgroups on the best way to proceed.

[1] http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml

RUTGERS UNIVERSITY

RUcore. Born Digital Still Images

http://odin.page2pixel.org/standards/latest/RUcoreStandards-BornDigitalPhotos.pdf



RUcore
Rutgers Community Repository

Born Digital Still Images (Digital Photos):

Recommended Minimum Standards
For Archival and Presentation Datastreams

(Note: This document addresses standards for born-digital still images only. For standards and requirements pertaining to digitization, i.e. the scanning of paper, slides or other analog media into digital images, please refer to the RUcore Digital Surrogate Guidelines.)

Introduction and Rationale

Since the inception of RUcore, a significant shift in the field of photography has taken place, as amateurs and professionals alike have migrated *en masse* from analog film to digital formats. Since the first repository specifications for digital photography were drafted in 2006, we've seen digital photography overtake and dominate the field, largely overtaking film as a common medium for the capture of still images.

Of course, new objects will continue to be created using traditional film, and there is no foreseeable end to the creation of objects that originate on paper, film, or other analog recording format, even if those formats are relegated only to niche interest groups. To that end, the repository has established and refined a set of clear and concise standards that serve to acquire and preserve digital facsimiles of analog photographs, books and similar items.

Even so, digital photography brings with it new challenges and different capabilities than our existing core set of scanning digitization standards can support. As a result, an entirely separate set of standards dealing exclusively with digital photography and separate from those that support scanning must be defined and adhered to.

Emerging shifts to digital photography

While we have long heard that film's days are numbered, few have truly believed it until very recently. Digital photography has taken more than 12 years to mature, since the introduction of the first mass produced digital camera (the Apple Quicktake) in 1994. For a majority of this period, the switch from film to digital was largely relegated to early adopters, and broadly shunned by professionals who insisted film was here to stay. Within the last decade however, the quality of the hardware available as well as the introduction of professional grade software tools has not only swayed general opinion of digital photography, but has permitted digital photography to become a driving factor in the fate of most corporations in the field. Additionally, a number of very recent events has permanently and irrevocably spelled out that film's days as a dominant medium are numbered:

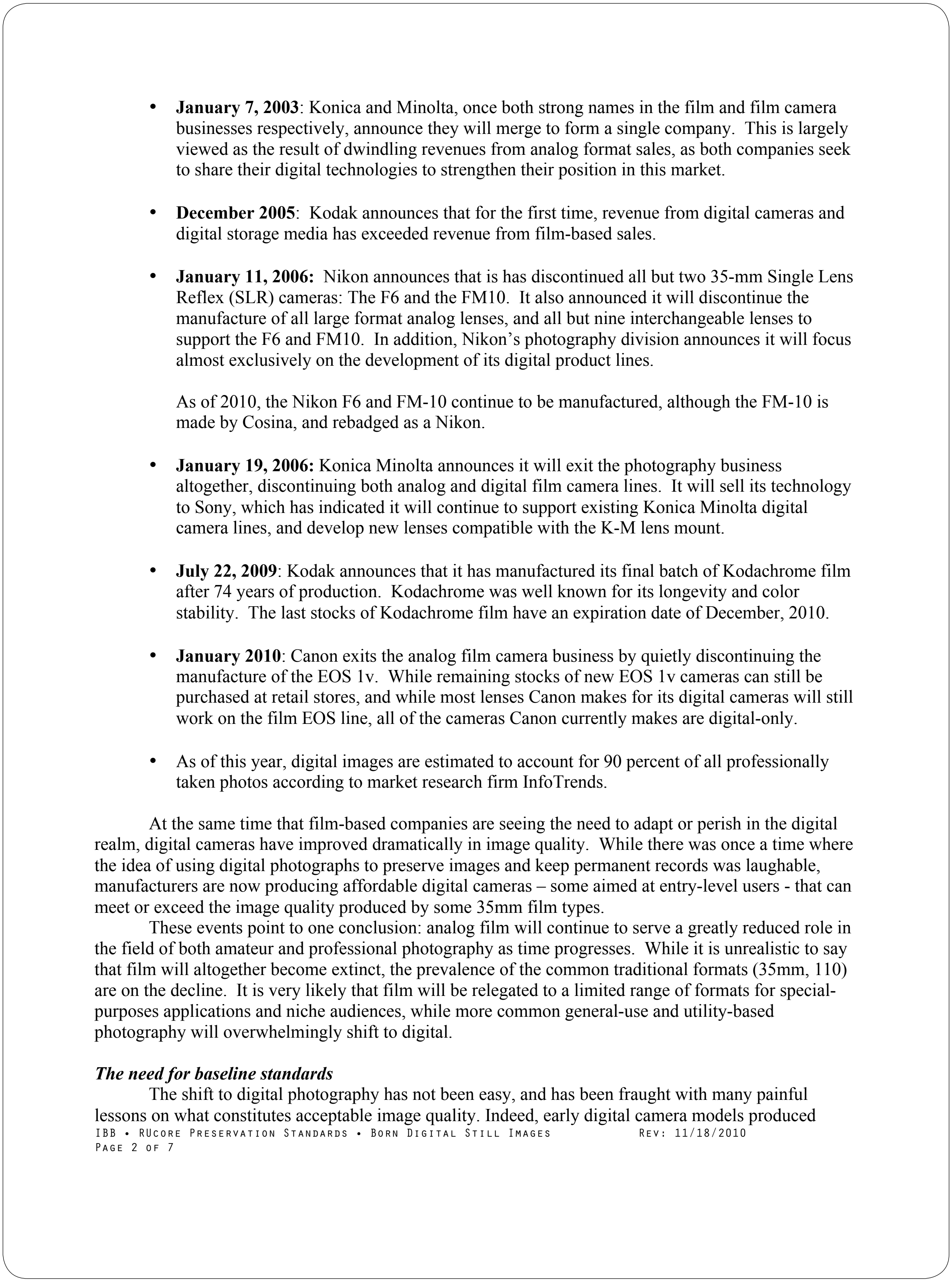
- **October 12, 2001**: Polaroid, Inc. files for bankruptcy. This is often seen as the watershed event for the decline of analog formats. Development of instant film formats stops, and while the popular Land Camera and a few other versions of Polaroid film survive, a wide array of other formats were discontinued.
  (Since 2001, Polaroid has been resurrected, filed for bankruptcy yet again, and the instant film formats discontinued. At present, private enthusiasts have attempted to revive Polaroid instant film through independent efforts.)

- **2001 – 2006**: Kodak has progressively discontinued a number of film formats, though it has stated it will aggressively pursue the continued manufacture of conventional 35mm and APS film. Additionally, Kodak announced in 2004 that while it "is, and will remain, committed to manufacturing and marketing the world's highest quality film," it is ending production of film cameras.

RUTGERS UNIVERSITY

RUcore. Born Digital Still Images

http://odin.page2pixel.org/standards/latest/RUcoreStandards-BornDigitalPhotos.pdf



- **January 7, 2003**: Konica and Minolta, once both strong names in the film and film camera businesses respectively, announce they will merge to form a single company. This is largely viewed as the result of dwindling revenues from analog format sales, as both companies seek to share their digital technologies to strengthen their position in this market.

- **December 2005**: Kodak announces that for the first time, revenue from digital cameras and digital storage media has exceeded revenue from film-based sales.

- **January 11, 2006:** Nikon announces that is has discontinued all but two 35-mm Single Lens Reflex (SLR) cameras: The F6 and the FM10. It also announced it will discontinue the manufacture of all large format analog lenses, and all but nine interchangeable lenses to support the F6 and FM10. In addition, Nikon's photography division announces it will focus almost exclusively on the development of its digital product lines.

  As of 2010, the Nikon F6 and FM-10 continue to be manufactured, although the FM-10 is made by Cosina, and rebadged as a Nikon.

- **January 19, 2006:** Konica Minolta announces it will exit the photography business altogether, discontinuing both analog and digital film camera lines. It will sell its technology to Sony, which has indicated it will continue to support existing Konica Minolta digital camera lines, and develop new lenses compatible with the K-M lens mount.

- **July 22, 2009**: Kodak announces that it has manufactured its final batch of Kodachrome film after 74 years of production. Kodachrome was well known for its longevity and color stability. The last stocks of Kodachrome film have an expiration date of December, 2010.

- **January 2010**: Canon exits the analog film camera business by quietly discontinuing the manufacture of the EOS 1v. While remaining stocks of new EOS 1v cameras can still be purchased at retail stores, and while most lenses Canon makes for its digital cameras will still work on the film EOS line, all of the cameras Canon currently makes are digital-only.

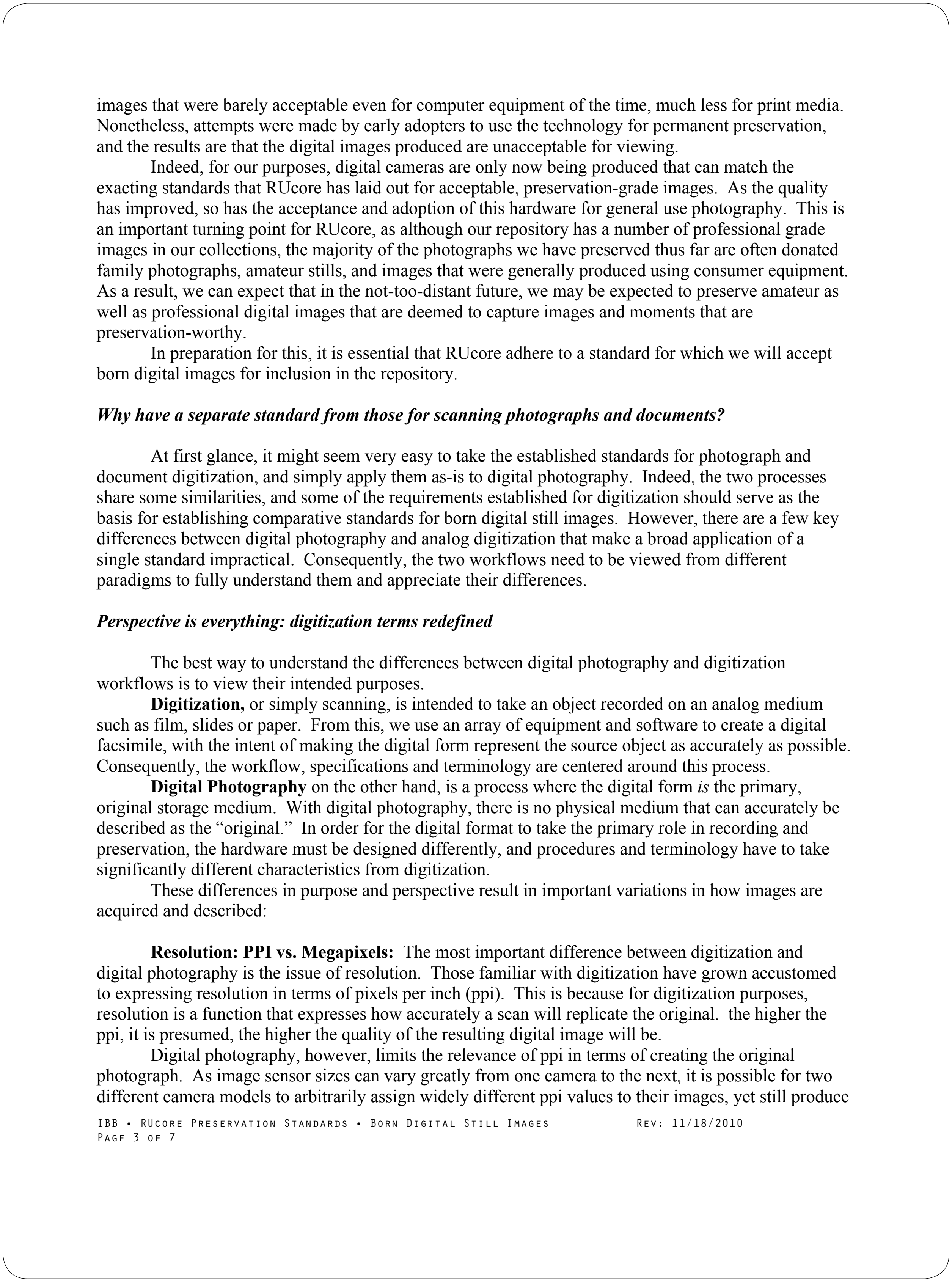
- As of this year, digital images are estimated to account for 90 percent of all professionally taken photos according to market research firm InfoTrends.

At the same time that film-based companies are seeing the need to adapt or perish in the digital realm, digital cameras have improved dramatically in image quality. While there was once a time where the idea of using digital photographs to preserve images and keep permanent records was laughable, manufacturers are now producing affordable digital cameras – some aimed at entry-level users - that can meet or exceed the image quality produced by some 35mm film types.

These events point to one conclusion: analog film will continue to serve a greatly reduced role in the field of both amateur and professional photography as time progresses. While it is unrealistic to say that film will altogether become extinct, the prevalence of the common traditional formats (35mm, 110) are on the decline. It is very likely that film will be relegated to a limited range of formats for special-purposes applications and niche audiences, while more common general-use and utility-based photography will overwhelmingly shift to digital.

The need for baseline standards

The shift to digital photography has not been easy, and has been fraught with many painful lessons on what constitutes acceptable image quality. Indeed, early digital camera models produced



images that were barely acceptable even for computer equipment of the time, much less for print media. Nonetheless, attempts were made by early adopters to use the technology for permanent preservation, and the results are that the digital images produced are unacceptable for viewing.

Indeed, for our purposes, digital cameras are only now being produced that can match the exacting standards that RUcore has laid out for acceptable, preservation-grade images. As the quality has improved, so has the acceptance and adoption of this hardware for general use photography. This is an important turning point for RUcore, as although our repository has a number of professional grade images in our collections, the majority of the photographs we have preserved thus far are often donated family photographs, amateur stills, and images that were generally produced using consumer equipment. As a result, we can expect that in the not-too-distant future, we may be expected to preserve amateur as well as professional digital images that are deemed to capture images and moments that are preservation-worthy.

In preparation for this, it is essential that RUcore adhere to a standard for which we will accept born digital images for inclusion in the repository.

Why have a separate standard from those for scanning photographs and documents?

At first glance, it might seem very easy to take the established standards for photograph and document digitization, and simply apply them as-is to digital photography. Indeed, the two processes share some similarities, and some of the requirements established for digitization should serve as the basis for establishing comparative standards for born digital still images. However, there are a few key differences between digital photography and analog digitization that make a broad application of a single standard impractical. Consequently, the two workflows need to be viewed from different paradigms to fully understand them and appreciate their differences.

Perspective is everything: digitization terms redefined

The best way to understand the differences between digital photography and digitization workflows is to view their intended purposes.

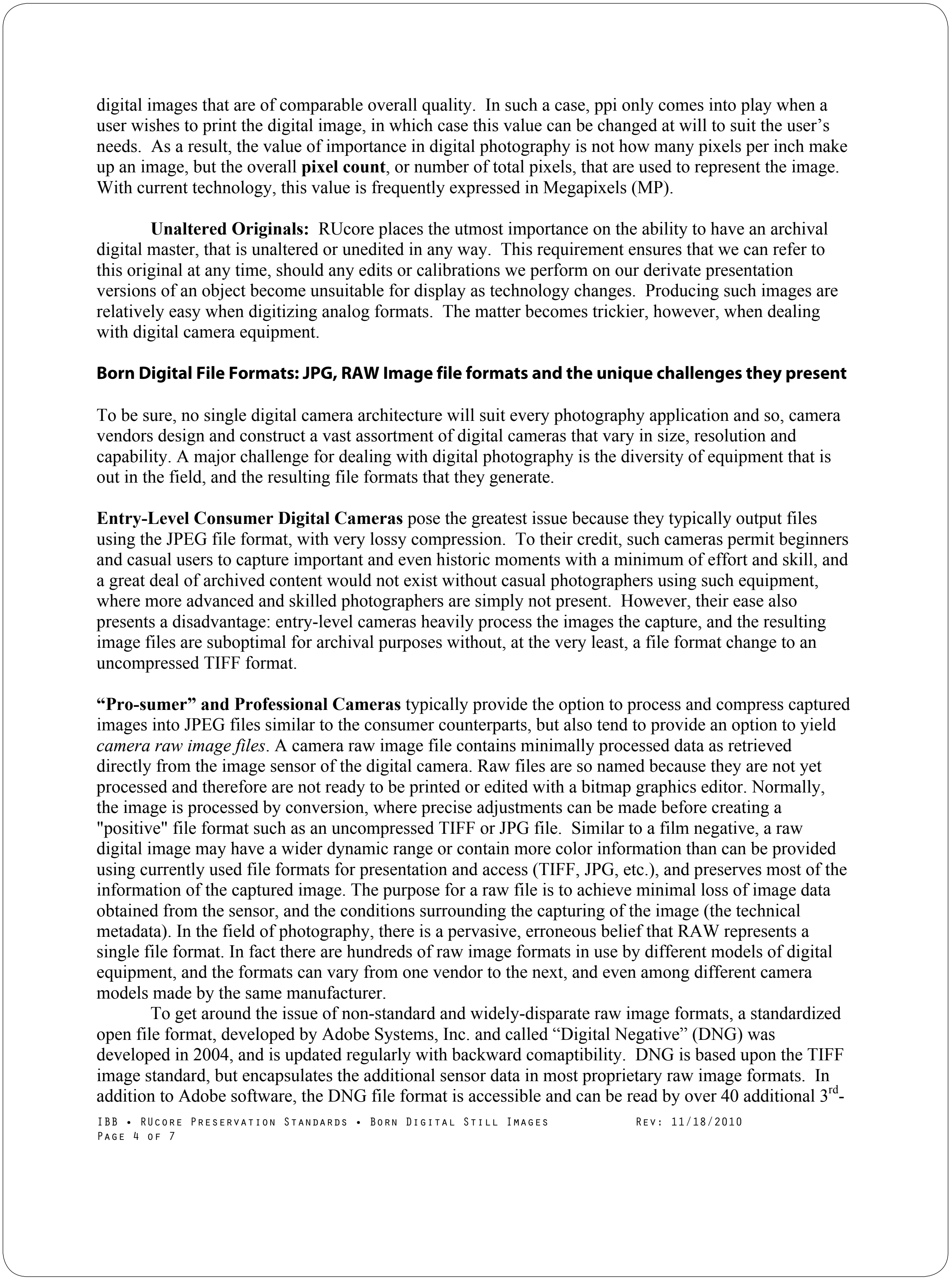
Digitization, or simply scanning, is intended to take an object recorded on an analog medium such as film, slides or paper. From this, we use an array of equipment and software to create a digital facsimile, with the intent of making the digital form represent the source object as accurately as possible. Consequently, the workflow, specifications and terminology are centered around this process.

Digital Photography on the other hand, is a process where the digital form *is* the primary, original storage medium. With digital photography, there is no physical medium that can accurately be described as the “original.” In order for the digital format to take the primary role in recording and preservation, the hardware must be designed differently, and procedures and terminology have to take significantly different characteristics from digitization.

These differences in purpose and perspective result in important variations in how images are acquired and described:

Resolution: PPI vs. Megapixels: The most important difference between digitization and digital photography is the issue of resolution. Those familiar with digitization have grown accustomed to expressing resolution in terms of pixels per inch (ppi). This is because for digitization purposes, resolution is a function that expresses how accurately a scan will replicate the original. the higher the ppi, it is presumed, the higher the quality of the resulting digital image will be.

Digital photography, however, limits the relevance of ppi in terms of creating the original photograph. As image sensor sizes can vary greatly from one camera to the next, it is possible for two different camera models to arbitrarily assign widely different ppi values to their images, yet still produce



digital images that are of comparable overall quality. In such a case, ppi only comes into play when a user wishes to print the digital image, in which case this value can be changed at will to suit the user's needs. As a result, the value of importance in digital photography is not how many pixels per inch make up an image, but the overall **pixel count**, or number of total pixels, that are used to represent the image. With current technology, this value is frequently expressed in Megapixels (MP).

Unaltered Originals: RUcore places the utmost importance on the ability to have an archival digital master, that is unaltered or unedited in any way. This requirement ensures that we can refer to this original at any time, should any edits or calibrations we perform on our derivate presentation versions of an object become unsuitable for display as technology changes. Producing such images are relatively easy when digitizing analog formats. The matter becomes trickier, however, when dealing with digital camera equipment.

Born Digital File Formats: JPG, RAW Image file formats and the unique challenges they present

To be sure, no single digital camera architecture will suit every photography application and so, camera vendors design and construct a vast assortment of digital cameras that vary in size, resolution and capability. A major challenge for dealing with digital photography is the diversity of equipment that is out in the field, and the resulting file formats that they generate.

Entry-Level Consumer Digital Cameras pose the greatest issue because they typically output files using the JPEG file format, with very lossy compression. To their credit, such cameras permit beginners and casual users to capture important and even historic moments with a minimum of effort and skill, and a great deal of archived content would not exist without casual photographers using such equipment, where more advanced and skilled photographers are simply not present. However, their ease also presents a disadvantage: entry-level cameras heavily process the images the capture, and the resulting image files are suboptimal for archival purposes without, at the very least, a file format change to an uncompressed TIFF format.

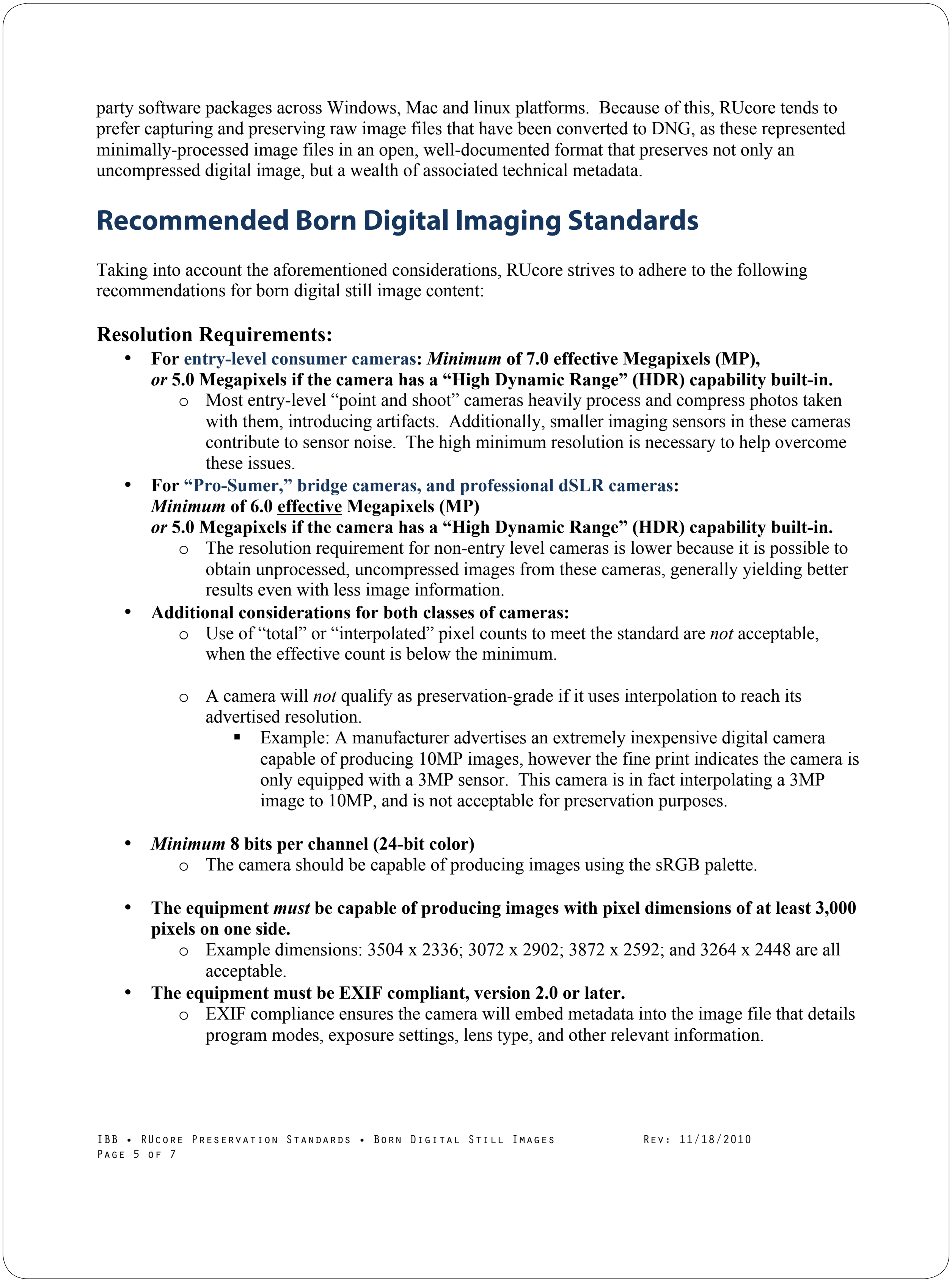
"Pro-sumer" and Professional Cameras typically provide the option to process and compress captured images into JPEG files similar to the consumer counterparts, but also tend to provide an option to yield *camera raw image files*. A camera raw image file contains minimally processed data as retrieved directly from the image sensor of the digital camera. Raw files are so named because they are not yet processed and therefore are not ready to be printed or edited with a bitmap graphics editor. Normally, the image is processed by conversion, where precise adjustments can be made before creating a "positive" file format such as an uncompressed TIFF or JPG file. Similar to a film negative, a raw digital image may have a wider dynamic range or contain more color information than can be provided using currently used file formats for presentation and access (TIFF, JPG, etc.), and preserves most of the information of the captured image. The purpose for a raw file is to achieve minimal loss of image data obtained from the sensor, and the conditions surrounding the capturing of the image (the technical metadata). In the field of photography, there is a pervasive, erroneous belief that RAW represents a single file format. In fact there are hundreds of raw image formats in use by different models of digital equipment, and the formats can vary from one vendor to the next, and even among different camera models made by the same manufacturer.

To get around the issue of non-standard and widely-disparate raw image formats, a standardized open file format, developed by Adobe Systems, Inc. and called "Digital Negative" (DNG) was developed in 2004, and is updated regularly with backward comaptibility. DNG is based upon the TIFF image standard, but encapsulates the additional sensor data in most proprietary raw image formats. In addition to Adobe software, the DNG file format is accessible and can be read by over 40 additional 3rd-

RUTGERS UNIVERSITY

RUcore. Born Digital Still Images

http://odin.page2pixel.org/standards/latest/RUcoreStandards-BornDigitalPhotos.pdf



party software packages across Windows, Mac and linux platforms. Because of this, RUcore tends to prefer capturing and preserving raw image files that have been converted to DNG, as these represented minimally-processed image files in an open, well-documented format that preserves not only an uncompressed digital image, but a wealth of associated technical metadata.

Recommended Born Digital Imaging Standards

Taking into account the aforementioned considerations, RUcore strives to adhere to the following recommendations for born digital still image content:

Resolution Requirements:

- **For entry-level consumer cameras: *Minimum* of 7.0 effective Megapixels (MP),**
  ***or* 5.0 Megapixels if the camera has a "High Dynamic Range" (HDR) capability built-in.**
  - Most entry-level "point and shoot" cameras heavily process and compress photos taken with them, introducing artifacts. Additionally, smaller imaging sensors in these cameras contribute to sensor noise. The high minimum resolution is necessary to help overcome these issues.
- **For "Pro-Sumer," bridge cameras, and professional dSLR cameras:**
  ***Minimum* of 6.0 effective Megapixels (MP)**
  ***or* 5.0 Megapixels if the camera has a "High Dynamic Range" (HDR) capability built-in.**
  - The resolution requirement for non-entry level cameras is lower because it is possible to obtain unprocessed, uncompressed images from these cameras, generally yielding better results even with less image information.
- **Additional considerations for both classes of cameras:**
  - Use of "total" or "interpolated" pixel counts to meet the standard are *not* acceptable, when the effective count is below the minimum.
  - A camera will *not* qualify as preservation-grade if it uses interpolation to reach its advertised resolution.
    - Example: A manufacturer advertises an extremely inexpensive digital camera capable of producing 10MP images, however the fine print indicates the camera is only equipped with a 3MP sensor. This camera is in fact interpolating a 3MP image to 10MP, and is not acceptable for preservation purposes.
- ***Minimum* 8 bits per channel (24-bit color)**
  - The camera should be capable of producing images using the sRGB palette.
- **The equipment *must* be capable of producing images with pixel dimensions of at least 3,000 pixels on one side.**
  - Example dimensions: 3504 x 2336; 3072 x 2902; 3872 x 2592; and 3264 x 2448 are all acceptable.
- **The equipment must be EXIF compliant, version 2.0 or later.**
  - EXIF compliance ensures the camera will embed metadata into the image file that details program modes, exposure settings, lens type, and other relevant information.

RUTGERS UNIVERSITY

RUcore. Born Digital Still Images

http://odin.page2pixel.org/standards/latest/RUcoreStandards-BornDigitalPhotos.pdf

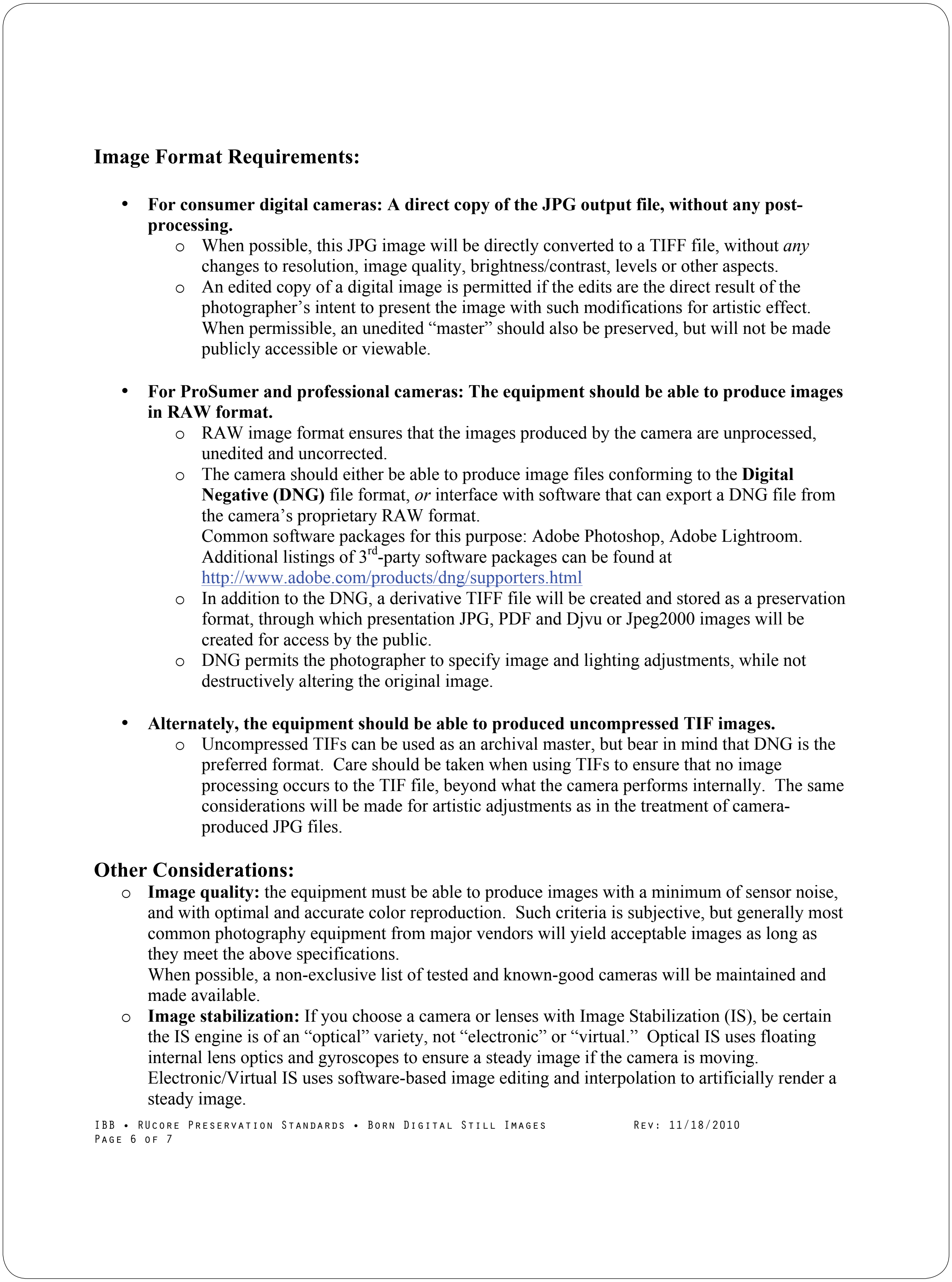


Image Format Requirements:

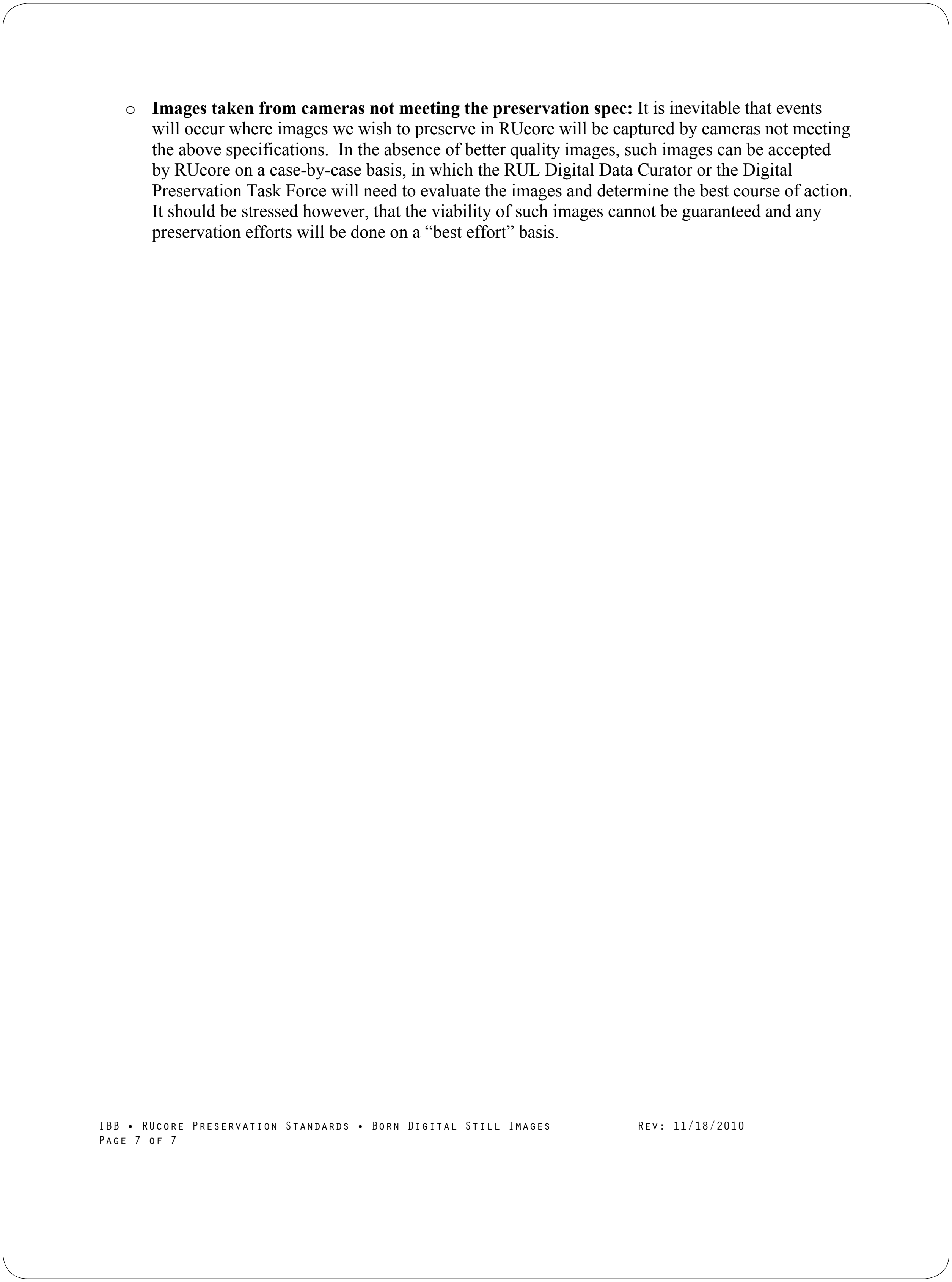
- **For consumer digital cameras: A direct copy of the JPG output file, without any post-processing.**
    - When possible, this JPG image will be directly converted to a TIFF file, without *any* changes to resolution, image quality, brightness/contrast, levels or other aspects.
    - An edited copy of a digital image is permitted if the edits are the direct result of the photographer's intent to present the image with such modifications for artistic effect. When permissible, an unedited "master" should also be preserved, but will not be made publicly accessible or viewable.

- **For ProSumer and professional cameras: The equipment should be able to produce images in RAW format.**
    - RAW image format ensures that the images produced by the camera are unprocessed, unedited and uncorrected.
    - The camera should either be able to produce image files conforming to the **Digital Negative (DNG)** file format, *or* interface with software that can export a DNG file from the camera's proprietary RAW format.
      Common software packages for this purpose: Adobe Photoshop, Adobe Lightroom.
      Additional listings of 3rd-party software packages can be found at
      http://www.adobe.com/products/dng/supporters.html
    - In addition to the DNG, a derivative TIFF file will be created and stored as a preservation format, through which presentation JPG, PDF and Djvu or Jpeg2000 images will be created for access by the public.
    - DNG permits the photographer to specify image and lighting adjustments, while not destructively altering the original image.

- **Alternately, the equipment should be able to produced uncompressed TIF images.**
    - Uncompressed TIFs can be used as an archival master, but bear in mind that DNG is the preferred format. Care should be taken when using TIFs to ensure that no image processing occurs to the TIF file, beyond what the camera performs internally. The same considerations will be made for artistic adjustments as in the treatment of camera-produced JPG files.

Other Considerations:

- **Image quality:** the equipment must be able to produce images with a minimum of sensor noise, and with optimal and accurate color reproduction. Such criteria is subjective, but generally most common photography equipment from major vendors will yield acceptable images as long as they meet the above specifications.
  When possible, a non-exclusive list of tested and known-good cameras will be maintained and made available.
- **Image stabilization:** If you choose a camera or lenses with Image Stabilization (IS), be certain the IS engine is of an "optical" variety, not "electronic" or "virtual." Optical IS uses floating internal lens optics and gyroscopes to ensure a steady image if the camera is moving. Electronic/Virtual IS uses software-based image editing and interpolation to artificially render a steady image.

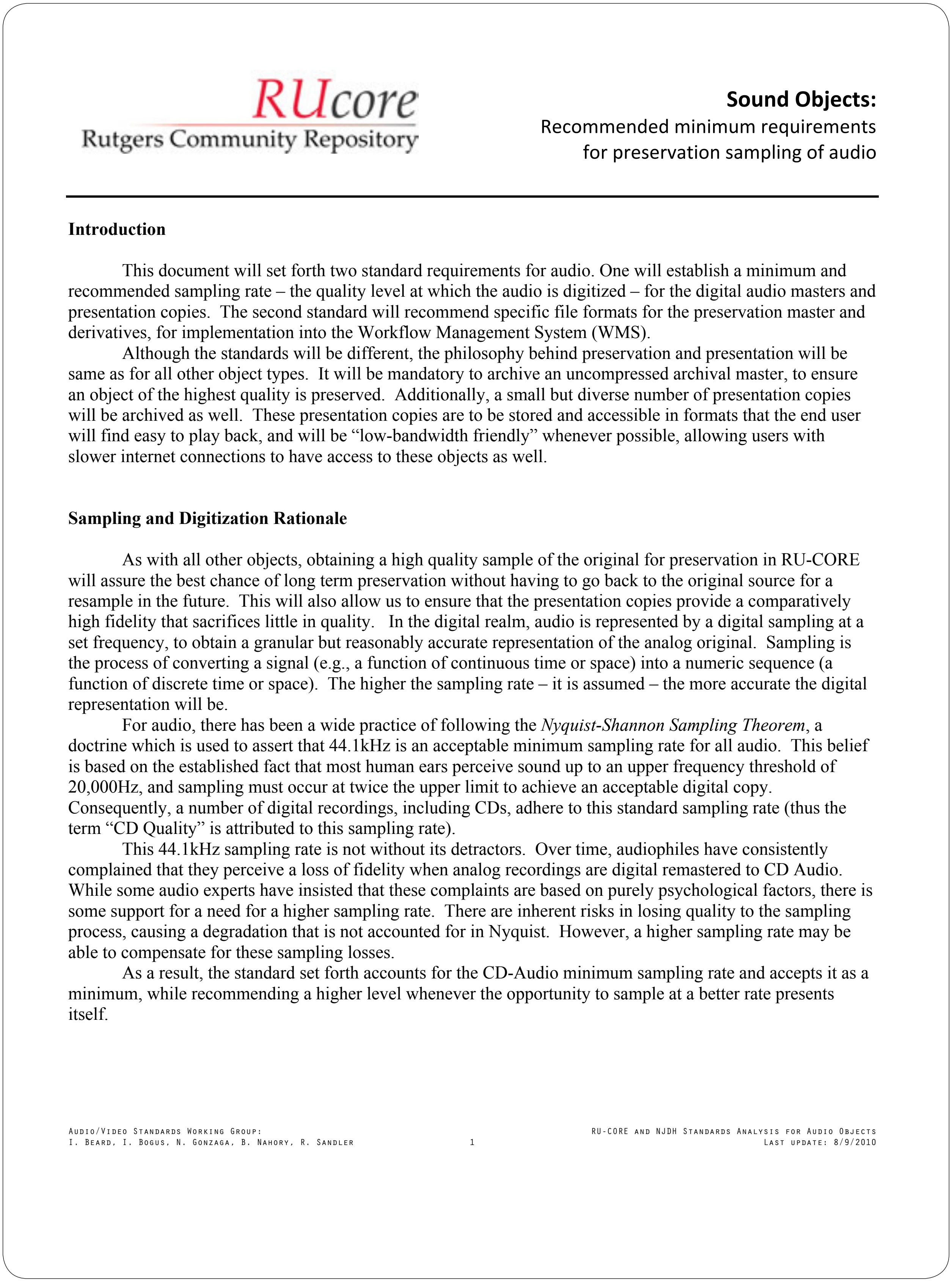


- **Images taken from cameras not meeting the preservation spec:** It is inevitable that events will occur where images we wish to preserve in RUcore will be captured by cameras not meeting the above specifications. In the absence of better quality images, such images can be accepted by RUcore on a case-by-case basis, in which the RUL Digital Data Curator or the Digital Preservation Task Force will need to evaluate the images and determine the best course of action. It should be stressed however, that the viability of such images cannot be guaranteed and any preservation efforts will be done on a “best effort” basis.

RUTGERS UNIVERSITY

RUcore. Sound Objects

http://odin.page2pixel.org/standards/latest/RUcoreStandards-Audio.pdf



RUcore
Rutgers Community Repository

Sound Objects:

Recommended minimum requirements
for preservation sampling of audio

Introduction

This document will set forth two standard requirements for audio. One will establish a minimum and recommended sampling rate – the quality level at which the audio is digitized – for the digital audio masters and presentation copies. The second standard will recommend specific file formats for the preservation master and derivatives, for implementation into the Workflow Management System (WMS).

Although the standards will be different, the philosophy behind preservation and presentation will be same as for all other object types. It will be mandatory to archive an uncompressed archival master, to ensure an object of the highest quality is preserved. Additionally, a small but diverse number of presentation copies will be archived as well. These presentation copies are to be stored and accessible in formats that the end user will find easy to play back, and will be "low-bandwidth friendly" whenever possible, allowing users with slower internet connections to have access to these objects as well.

Sampling and Digitization Rationale

As with all other objects, obtaining a high quality sample of the original for preservation in RU-CORE will assure the best chance of long term preservation without having to go back to the original source for a resample in the future. This will also allow us to ensure that the presentation copies provide a comparatively high fidelity that sacrifices little in quality. In the digital realm, audio is represented by a digital sampling at a set frequency, to obtain a granular but reasonably accurate representation of the analog original. Sampling is the process of converting a signal (e.g., a function of continuous time or space) into a numeric sequence (a function of discrete time or space). The higher the sampling rate – it is assumed – the more accurate the digital representation will be.

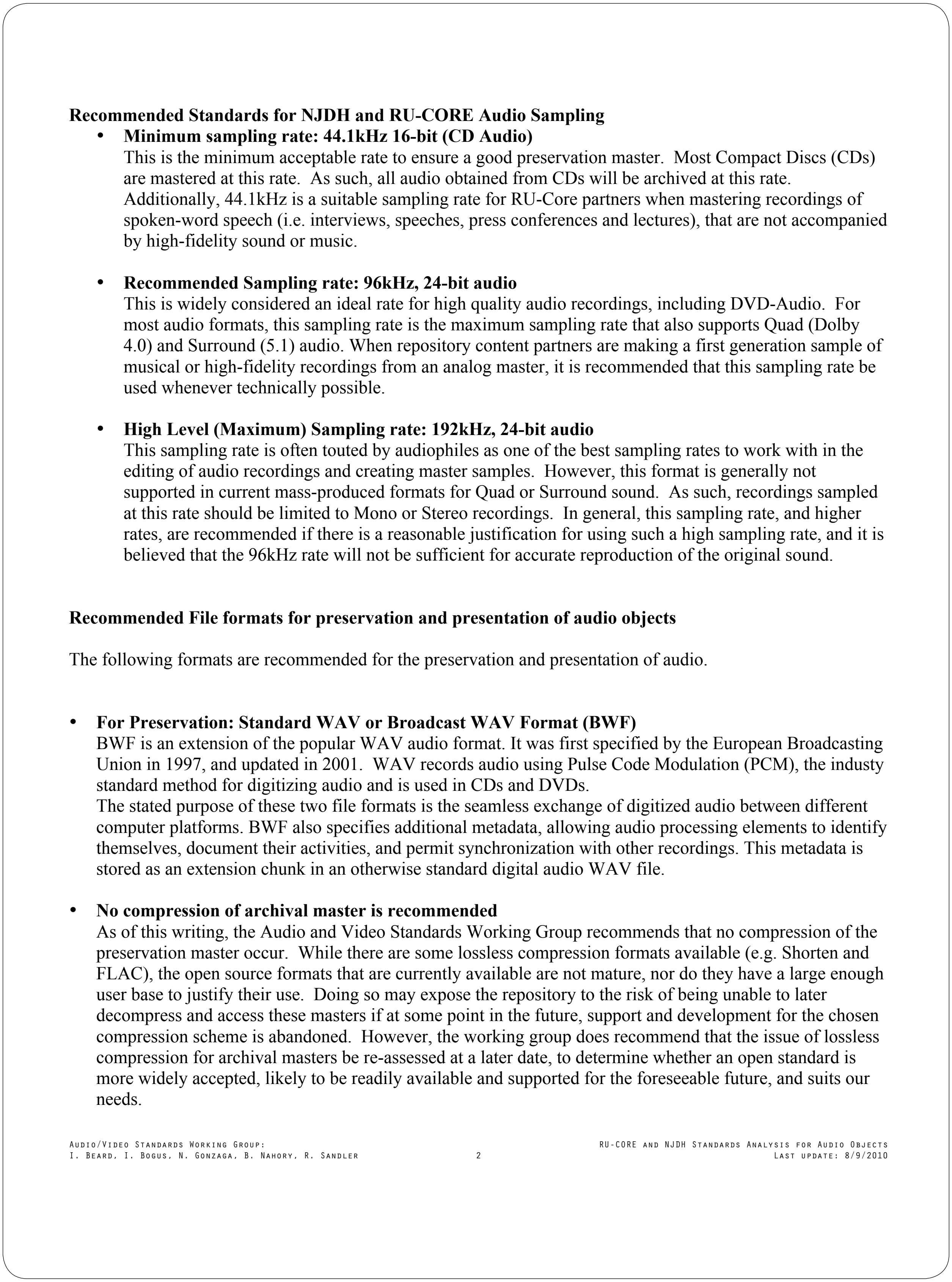
For audio, there has been a wide practice of following the *Nyquist-Shannon Sampling Theorem*, a doctrine which is used to assert that 44.1kHz is an acceptable minimum sampling rate for all audio. This belief is based on the established fact that most human ears perceive sound up to an upper frequency threshold of 20,000Hz, and sampling must occur at twice the upper limit to achieve an acceptable digital copy. Consequently, a number of digital recordings, including CDs, adhere to this standard sampling rate (thus the term "CD Quality" is attributed to this sampling rate).

This 44.1kHz sampling rate is not without its detractors. Over time, audiophiles have consistently complained that they perceive a loss of fidelity when analog recordings are digital remastered to CD Audio. While some audio experts have insisted that these complaints are based on purely psychological factors, there is some support for a need for a higher sampling rate. There are inherent risks in losing quality to the sampling process, causing a degradation that is not accounted for in Nyquist. However, a higher sampling rate may be able to compensate for these sampling losses.

As a result, the standard set forth accounts for the CD-Audio minimum sampling rate and accepts it as a minimum, while recommending a higher level whenever the opportunity to sample at a better rate presents itself.

RUTGERS UNIVERSITY

RUcore. Sound Objects
http://odin.page2pixel.org/standards/latest/RUcoreStandards-Audio.pdf



Recommended Standards for NJDH and RU-CORE Audio Sampling

- **Minimum sampling rate: 44.1kHz 16-bit (CD Audio)**
  This is the minimum acceptable rate to ensure a good preservation master. Most Compact Discs (CDs) are mastered at this rate. As such, all audio obtained from CDs will be archived at this rate. Additionally, 44.1kHz is a suitable sampling rate for RU-Core partners when mastering recordings of spoken-word speech (i.e. interviews, speeches, press conferences and lectures), that are not accompanied by high-fidelity sound or music.

- **Recommended Sampling rate: 96kHz, 24-bit audio**
  This is widely considered an ideal rate for high quality audio recordings, including DVD-Audio. For most audio formats, this sampling rate is the maximum sampling rate that also supports Quad (Dolby 4.0) and Surround (5.1) audio. When repository content partners are making a first generation sample of musical or high-fidelity recordings from an analog master, it is recommended that this sampling rate be used whenever technically possible.

- **High Level (Maximum) Sampling rate: 192kHz, 24-bit audio**
  This sampling rate is often touted by audiophiles as one of the best sampling rates to work with in the editing of audio recordings and creating master samples. However, this format is generally not supported in current mass-produced formats for Quad or Surround sound. As such, recordings sampled at this rate should be limited to Mono or Stereo recordings. In general, this sampling rate, and higher rates, are recommended if there is a reasonable justification for using such a high sampling rate, and it is believed that the 96kHz rate will not be sufficient for accurate reproduction of the original sound.

Recommended File formats for preservation and presentation of audio objects

The following formats are recommended for the preservation and presentation of audio.

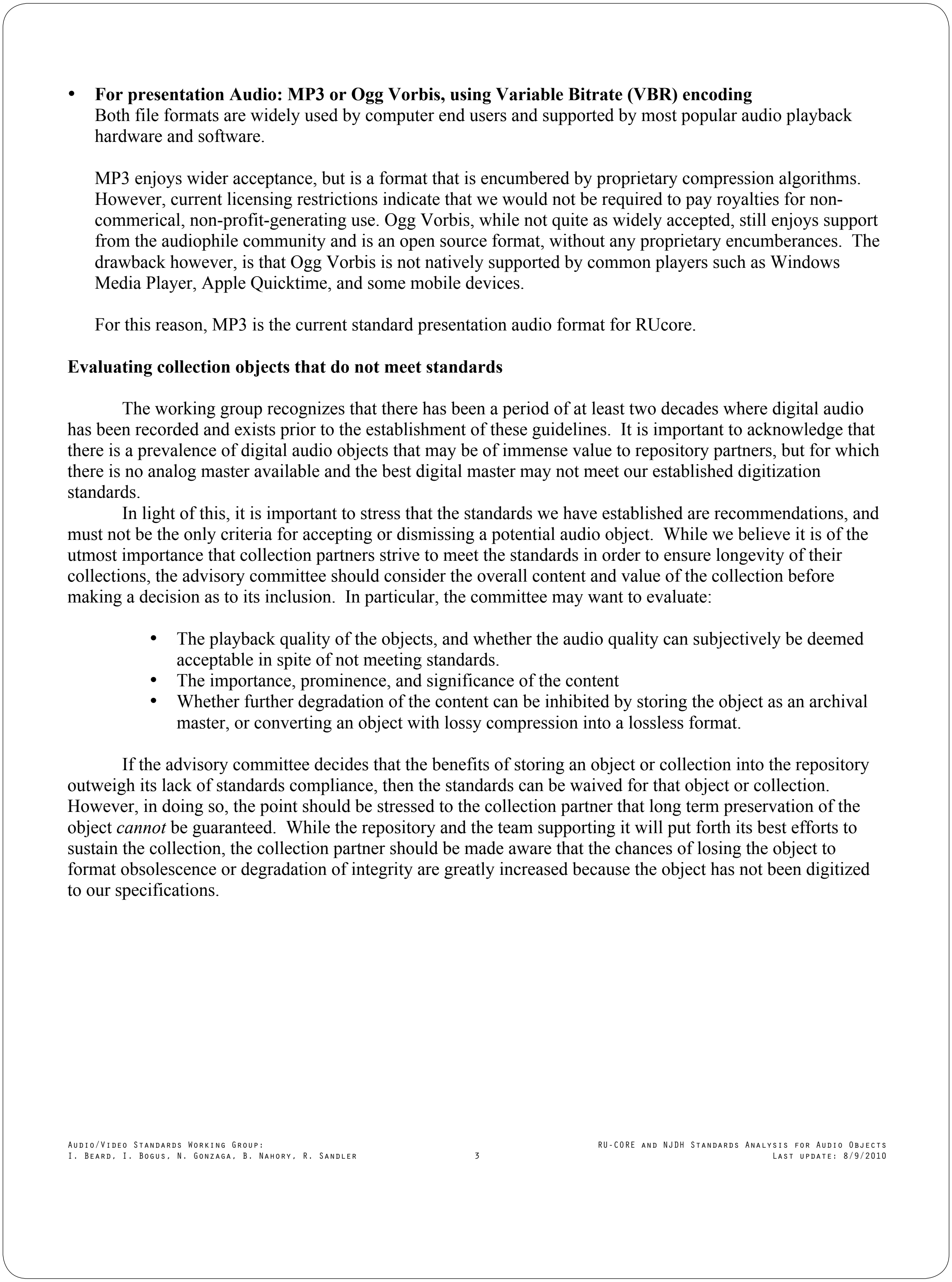
- **For Preservation: Standard WAV or Broadcast WAV Format (BWF)**
  BWF is an extension of the popular WAV audio format. It was first specified by the European Broadcasting Union in 1997, and updated in 2001. WAV records audio using Pulse Code Modulation (PCM), the industy standard method for digitizing audio and is used in CDs and DVDs.
  The stated purpose of these two file formats is the seamless exchange of digitized audio between different computer platforms. BWF also specifies additional metadata, allowing audio processing elements to identify themselves, document their activities, and permit synchronization with other recordings. This metadata is stored as an extension chunk in an otherwise standard digital audio WAV file.

- **No compression of archival master is recommended**
  As of this writing, the Audio and Video Standards Working Group recommends that no compression of the preservation master occur. While there are some lossless compression formats available (e.g. Shorten and FLAC), the open source formats that are currently available are not mature, nor do they have a large enough user base to justify their use. Doing so may expose the repository to the risk of being unable to later decompress and access these masters if at some point in the future, support and development for the chosen compression scheme is abandoned. However, the working group does recommend that the issue of lossless compression for archival masters be re-assessed at a later date, to determine whether an open standard is more widely accepted, likely to be readily available and supported for the foreseeable future, and suits our needs.

RUTGERS UNIVERSITY

RUcore. Sound Objects

http://odin.page2pixel.org/standards/latest/RUcoreStandards-Audio.pdf



- **For presentation Audio: MP3 or Ogg Vorbis, using Variable Bitrate (VBR) encoding**
  Both file formats are widely used by computer end users and supported by most popular audio playback hardware and software.

  MP3 enjoys wider acceptance, but is a format that is encumbered by proprietary compression algorithms. However, current licensing restrictions indicate that we would not be required to pay royalties for non-commerical, non-profit-generating use. Ogg Vorbis, while not quite as widely accepted, still enjoys support from the audiophile community and is an open source format, without any proprietary encumberances. The drawback however, is that Ogg Vorbis is not natively supported by common players such as Windows Media Player, Apple Quicktime, and some mobile devices.

  For this reason, MP3 is the current standard presentation audio format for RUcore.

Evaluating collection objects that do not meet standards

The working group recognizes that there has been a period of at least two decades where digital audio has been recorded and exists prior to the establishment of these guidelines. It is important to acknowledge that there is a prevalence of digital audio objects that may be of immense value to repository partners, but for which there is no analog master available and the best digital master may not meet our established digitization standards.

In light of this, it is important to stress that the standards we have established are recommendations, and must not be the only criteria for accepting or dismissing a potential audio object. While we believe it is of the utmost importance that collection partners strive to meet the standards in order to ensure longevity of their collections, the advisory committee should consider the overall content and value of the collection before making a decision as to its inclusion. In particular, the committee may want to evaluate:

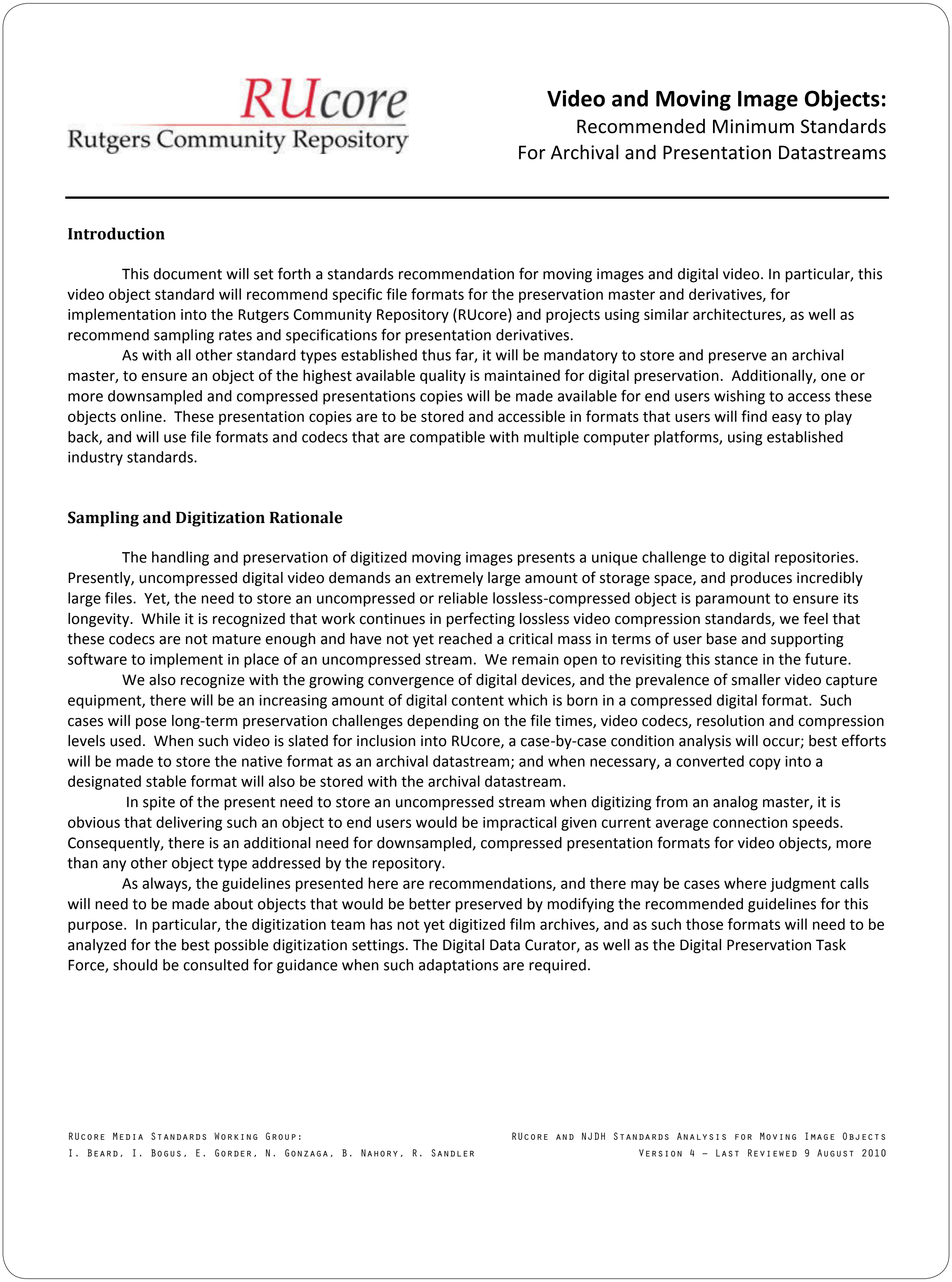
- The playback quality of the objects, and whether the audio quality can subjectively be deemed acceptable in spite of not meeting standards.
- The importance, prominence, and significance of the content
- Whether further degradation of the content can be inhibited by storing the object as an archival master, or converting an object with lossy compression into a lossless format.

If the advisory committee decides that the benefits of storing an object or collection into the repository outweigh its lack of standards compliance, then the standards can be waived for that object or collection. However, in doing so, the point should be stressed to the collection partner that long term preservation of the object *cannot* be guaranteed. While the repository and the team supporting it will put forth its best efforts to sustain the collection, the collection partner should be made aware that the chances of losing the object to format obsolescence or degradation of integrity are greatly increased because the object has not been digitized to our specifications.

RUTGERS UNIVERSITY

RUcore. Video and Moving Image Objects

http://odin.page2pixel.org/standards/latest/RUcoreStandards-Video.pdf



RUcore
Rutgers Community Repository

Video and Moving Image Objects:

Recommended Minimum Standards
For Archival and Presentation Datastreams

Introduction

This document will set forth a standards recommendation for moving images and digital video. In particular, this video object standard will recommend specific file formats for the preservation master and derivatives, for implementation into the Rutgers Community Repository (RUcore) and projects using similar architectures, as well as recommend sampling rates and specifications for presentation derivatives.

As with all other standard types established thus far, it will be mandatory to store and preserve an archival master, to ensure an object of the highest available quality is maintained for digital preservation. Additionally, one or more downsampled and compressed presentations copies will be made available for end users wishing to access these objects online. These presentation copies are to be stored and accessible in formats that users will find easy to play back, and will use file formats and codecs that are compatible with multiple computer platforms, using established industry standards.

Sampling and Digitization Rationale

The handling and preservation of digitized moving images presents a unique challenge to digital repositories. Presently, uncompressed digital video demands an extremely large amount of storage space, and produces incredibly large files. Yet, the need to store an uncompressed or reliable lossless-compressed object is paramount to ensure its longevity. While it is recognized that work continues in perfecting lossless video compression standards, we feel that these codecs are not mature enough and have not yet reached a critical mass in terms of user base and supporting software to implement in place of an uncompressed stream. We remain open to revisiting this stance in the future.

We also recognize with the growing convergence of digital devices, and the prevalence of smaller video capture equipment, there will be an increasing amount of digital content which is born in a compressed digital format. Such cases will pose long-term preservation challenges depending on the file times, video codecs, resolution and compression levels used. When such video is slated for inclusion into RUcore, a case-by-case condition analysis will occur; best efforts will be made to store the native format as an archival datastream; and when necessary, a converted copy into a designated stable format will also be stored with the archival datastream.

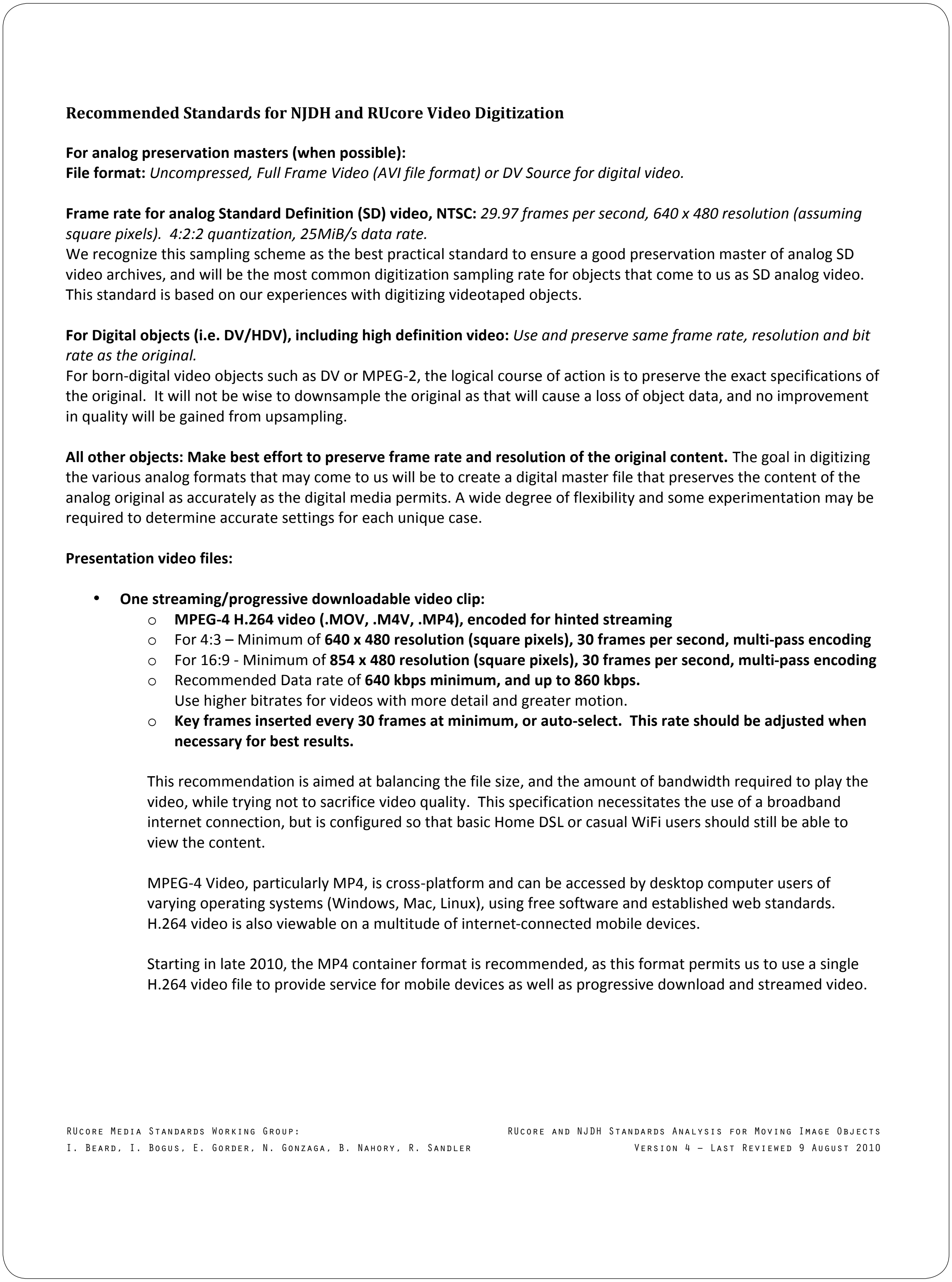
In spite of the present need to store an uncompressed stream when digitizing from an analog master, it is obvious that delivering such an object to end users would be impractical given current average connection speeds. Consequently, there is an additional need for downsampled, compressed presentation formats for video objects, more than any other object type addressed by the repository.

As always, the guidelines presented here are recommendations, and there may be cases where judgment calls will need to be made about objects that would be better preserved by modifying the recommended guidelines for this purpose. In particular, the digitization team has not yet digitized film archives, and as such those formats will need to be analyzed for the best possible digitization settings. The Digital Data Curator, as well as the Digital Preservation Task Force, should be consulted for guidance when such adaptations are required.

RUTGERS UNIVERSITY

RUcore. Video and Moving Image Objects

http://odin.page2pixel.org/standards/latest/RUcoreStandards-Video.pdf



Recommended Standards for NJDH and RUcore Video Digitization

For analog preservation masters (when possible):
File format: *Uncompressed, Full Frame Video (AVI file format) or DV Source for digital video.*

Frame rate for analog Standard Definition (SD) video, NTSC: *29.97 frames per second, 640 x 480 resolution (assuming square pixels). 4:2:2 quantization, 25MiB/s data rate.*
We recognize this sampling scheme as the best practical standard to ensure a good preservation master of analog SD video archives, and will be the most common digitization sampling rate for objects that come to us as SD analog video. This standard is based on our experiences with digitizing videotaped objects.

For Digital objects (i.e. DV/HDV), including high definition video: *Use and preserve same frame rate, resolution and bit rate as the original.*
For born-digital video objects such as DV or MPEG-2, the logical course of action is to preserve the exact specifications of the original. It will not be wise to downsample the original as that will cause a loss of object data, and no improvement in quality will be gained from upsampling.

All other objects: Make best effort to preserve frame rate and resolution of the original content. The goal in digitizing the various analog formats that may come to us will be to create a digital master file that preserves the content of the analog original as accurately as the digital media permits. A wide degree of flexibility and some experimentation may be required to determine accurate settings for each unique case.

Presentation video files:

- **One streaming/progressive downloadable video clip:**
  - **MPEG-4 H.264 video (.MOV, .M4V, .MP4), encoded for hinted streaming**
  - For 4:3 – Minimum of **640 x 480 resolution (square pixels), 30 frames per second, multi-pass encoding**
  - For 16:9 - Minimum of **854 x 480 resolution (square pixels), 30 frames per second, multi-pass encoding**
  - Recommended Data rate of **640 kbps minimum, and up to 860 kbps.**
    Use higher bitrates for videos with more detail and greater motion.
  - **Key frames inserted every 30 frames at minimum, or auto-select. This rate should be adjusted when necessary for best results.**

  This recommendation is aimed at balancing the file size, and the amount of bandwidth required to play the video, while trying not to sacrifice video quality. This specification necessitates the use of a broadband internet connection, but is configured so that basic Home DSL or casual WiFi users should still be able to view the content.

  MPEG-4 Video, particularly MP4, is cross-platform and can be accessed by desktop computer users of varying operating systems (Windows, Mac, Linux), using free software and established web standards. H.264 video is also viewable on a multitude of internet-connected mobile devices.

  Starting in late 2010, the MP4 container format is recommended, as this format permits us to use a single H.264 video file to provide service for mobile devices as well as progressive download and streamed video.

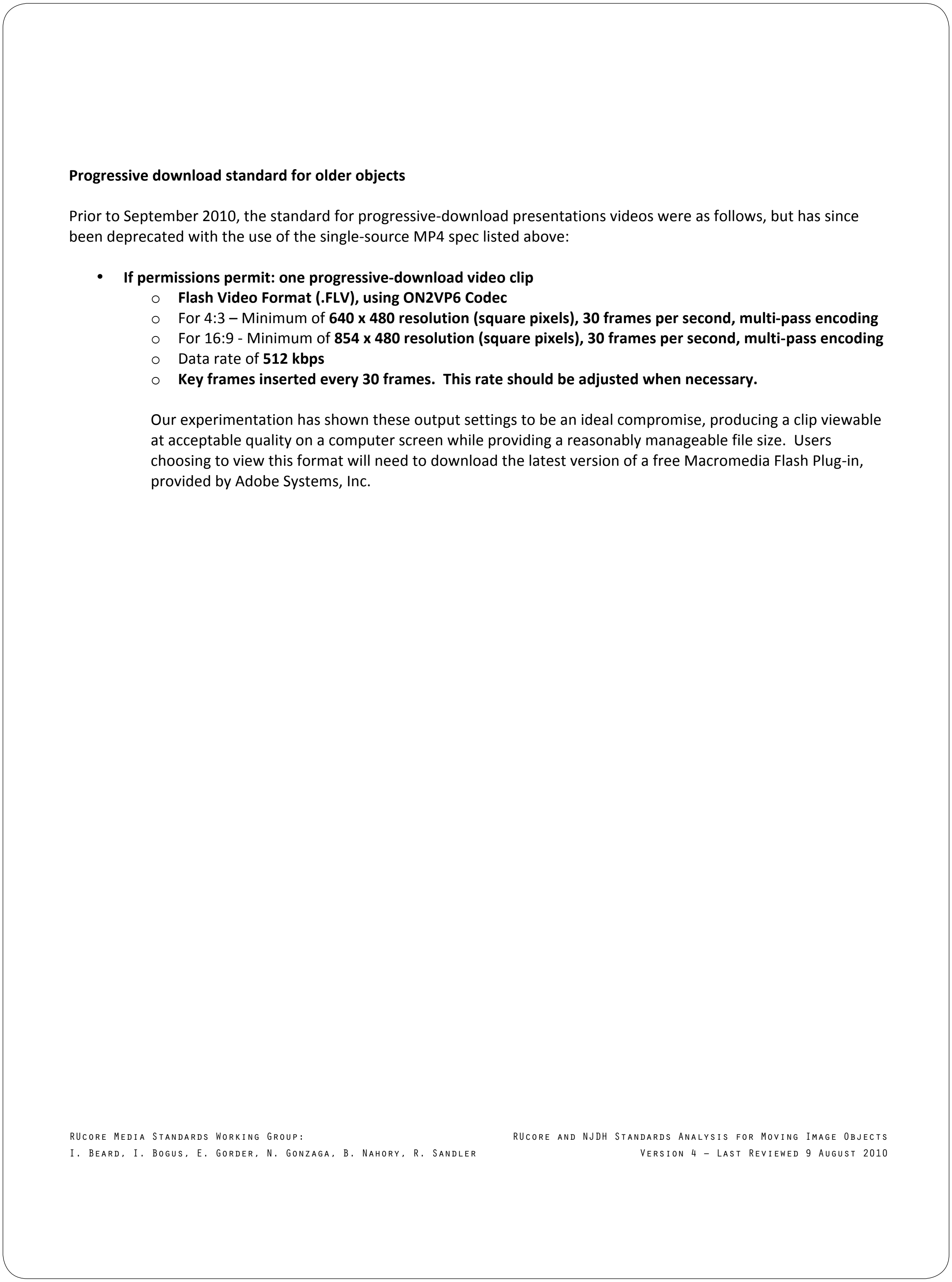
RUcore Media Standards Working Group:
I. Beard, I. Bogus, E. Gorder, N. Gonzaga, B. Nahory, R. Sandler

RUcore and NJDH Standards Analysis for Moving Image Objects
Version 4 – Last Reviewed 9 August 2010

RUTGERS UNIVERSITY

RUcore. Video and Moving Image Objects

http://odin.page2pixel.org/standards/latest/RUcoreStandards-Video.pdf



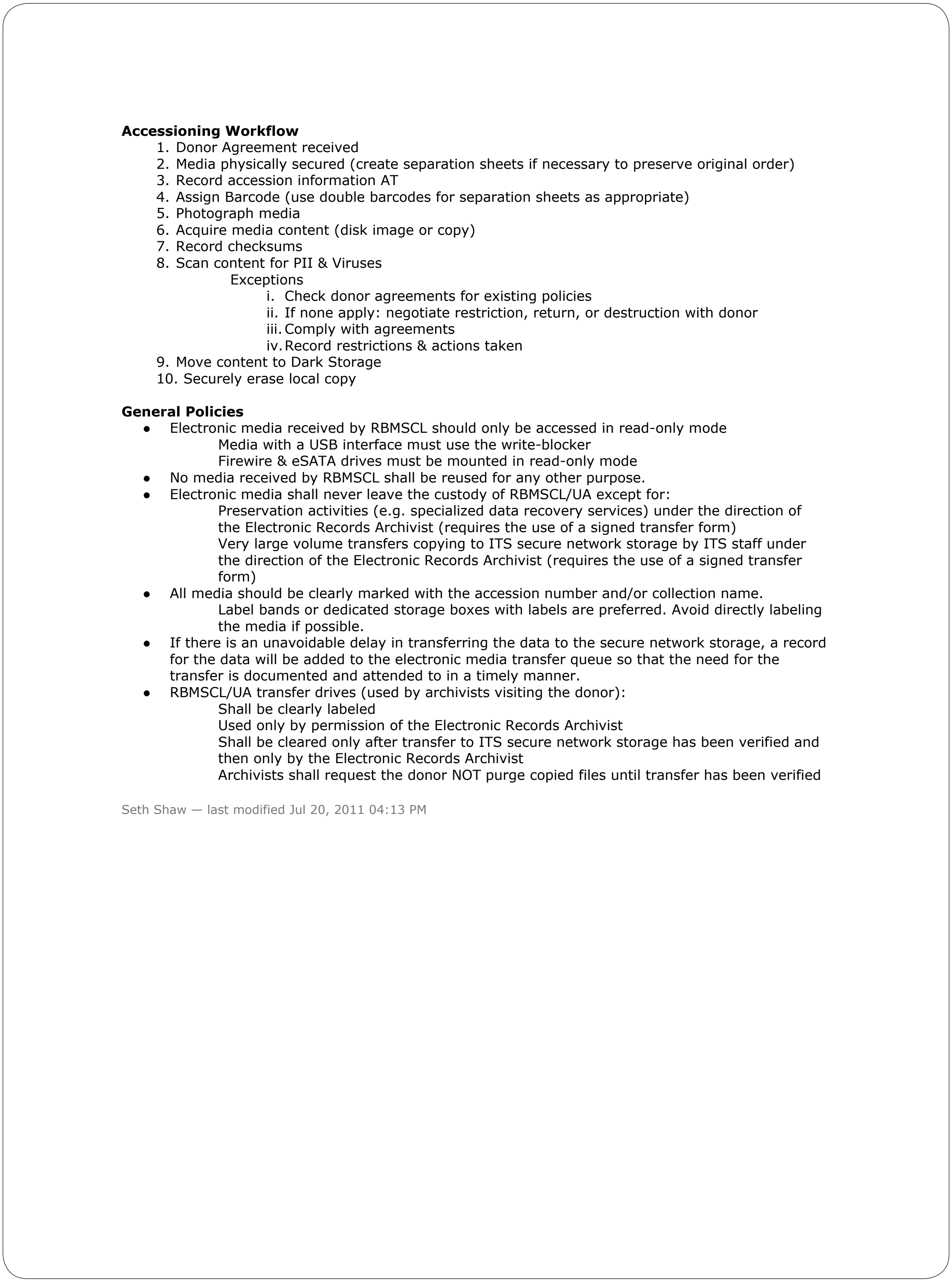
Progressive download standard for older objects

Prior to September 2010, the standard for progressive-download presentations videos were as follows, but has since been deprecated with the use of the single-source MP4 spec listed above:

- **If permissions permit: one progressive-download video clip**
    - **Flash Video Format (.FLV), using ON2VP6 Codec**
    - For 4:3 – Minimum of **640 x 480 resolution (square pixels), 30 frames per second, multi-pass encoding**
    - For 16:9 - Minimum of **854 x 480 resolution (square pixels), 30 frames per second, multi-pass encoding**
    - Data rate of **512 kbps**
    - **Key frames inserted every 30 frames. This rate should be adjusted when necessary.**

    Our experimentation has shown these output settings to be an ideal compromise, producing a clip viewable at acceptable quality on a computer screen while providing a reasonably manageable file size. Users choosing to view this format will need to download the latest version of a free Macromedia Flash Plug-in, provided by Adobe Systems, Inc.

Workflows



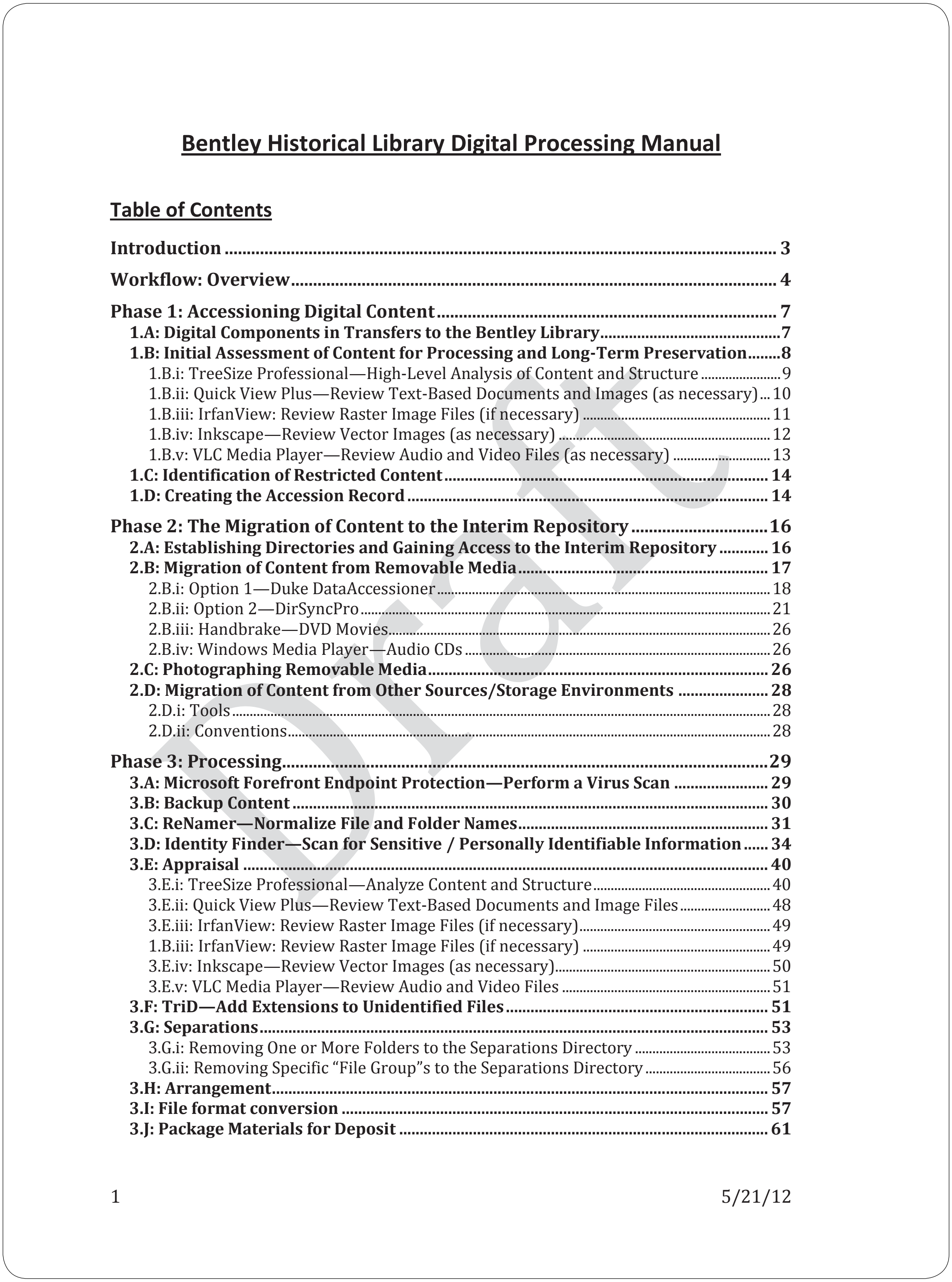
Accessioning Workflow

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

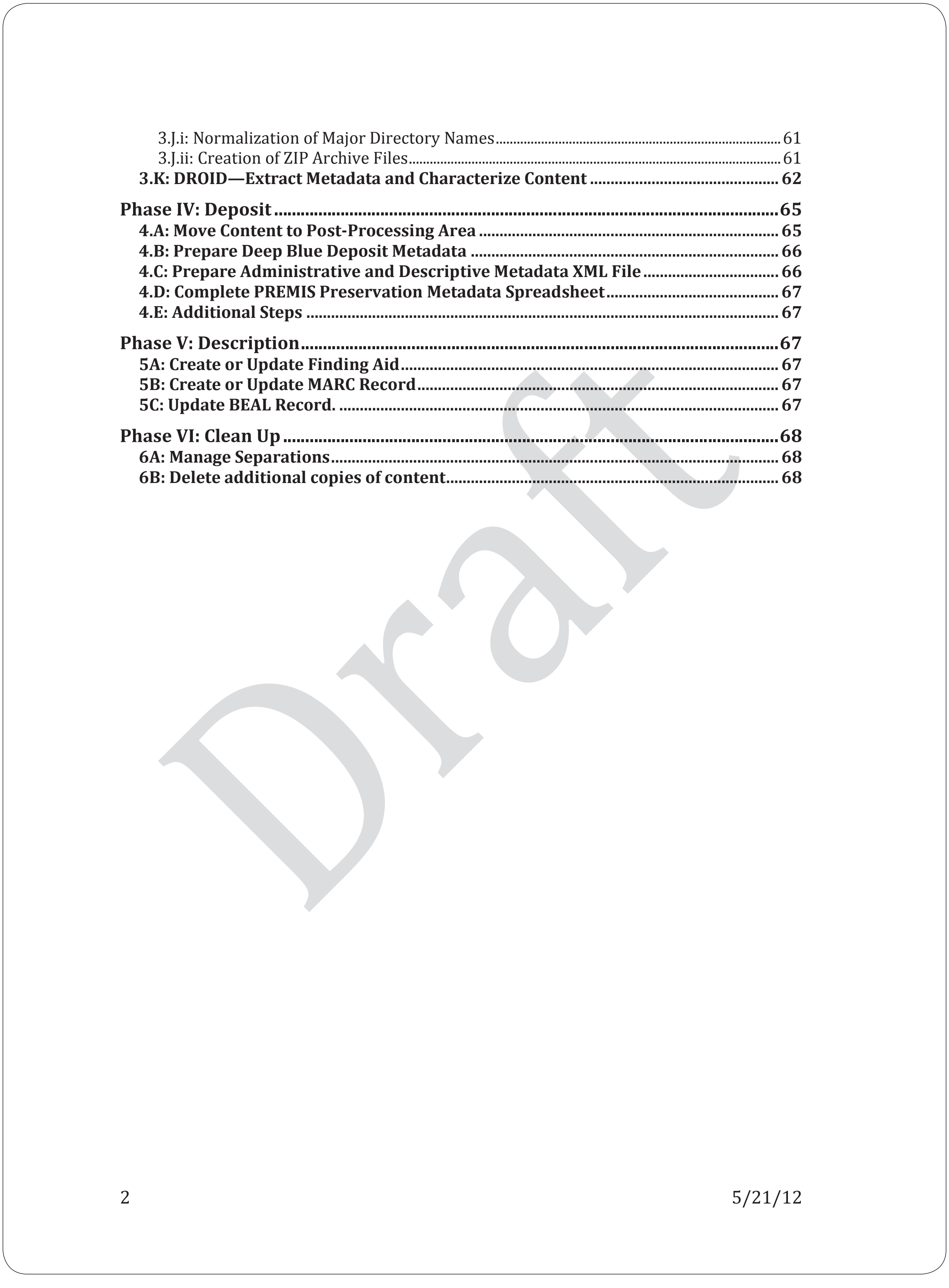
- Electronic media received by RBMSCL should only be accessed in read-only mode
  Media with a USB interface must use the write-blocker
  Firewire & eSATA drives must be mounted in read-only mode
- No media received by RBMSCL shall be reused for any other purpose.
- Electronic media shall never leave 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

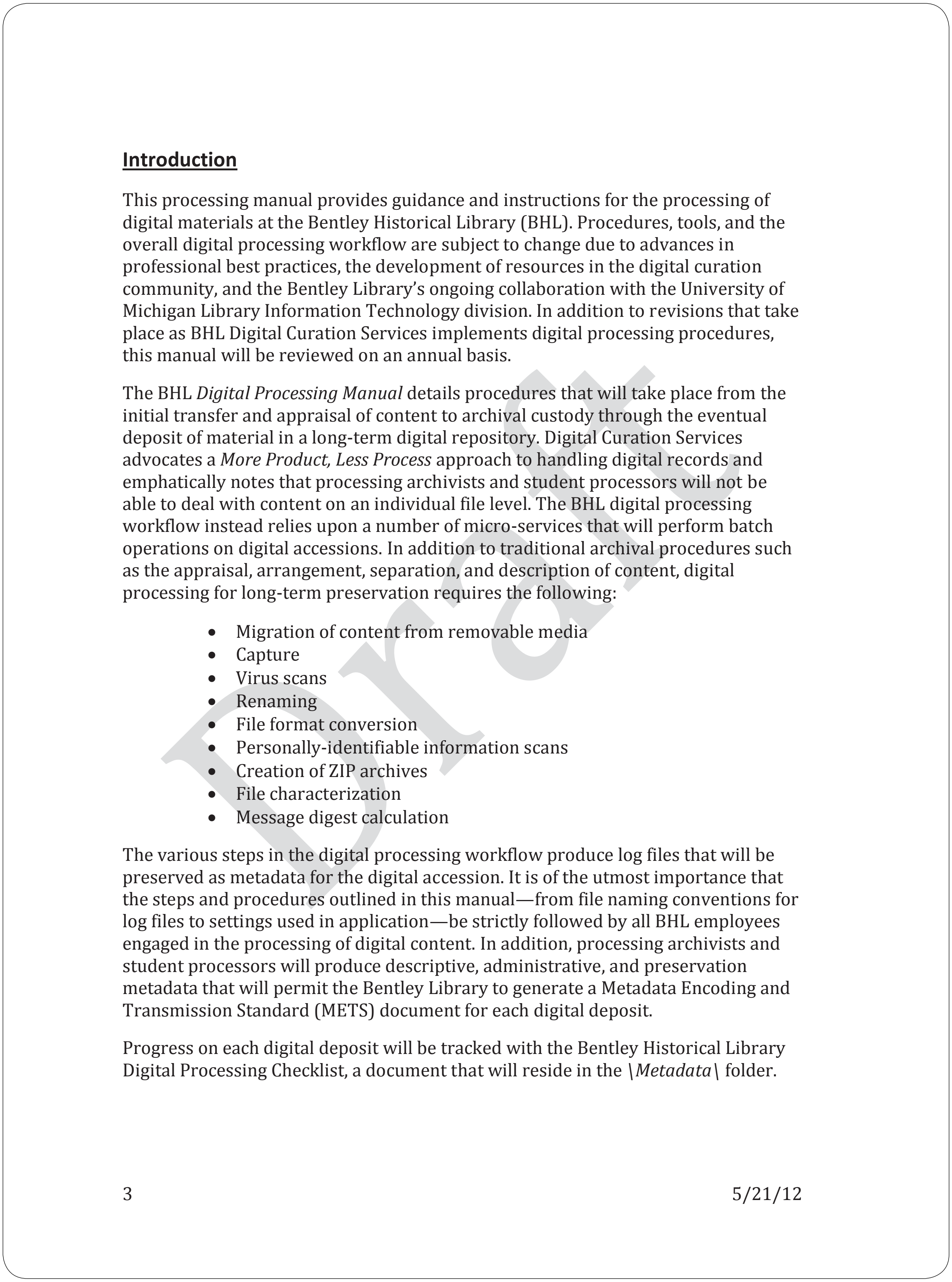
Seth Shaw — last modified Jul 20, 2011 04:13 PM



Bentley Historical Library Digital Processing Manual

Table of Contents





Introduction

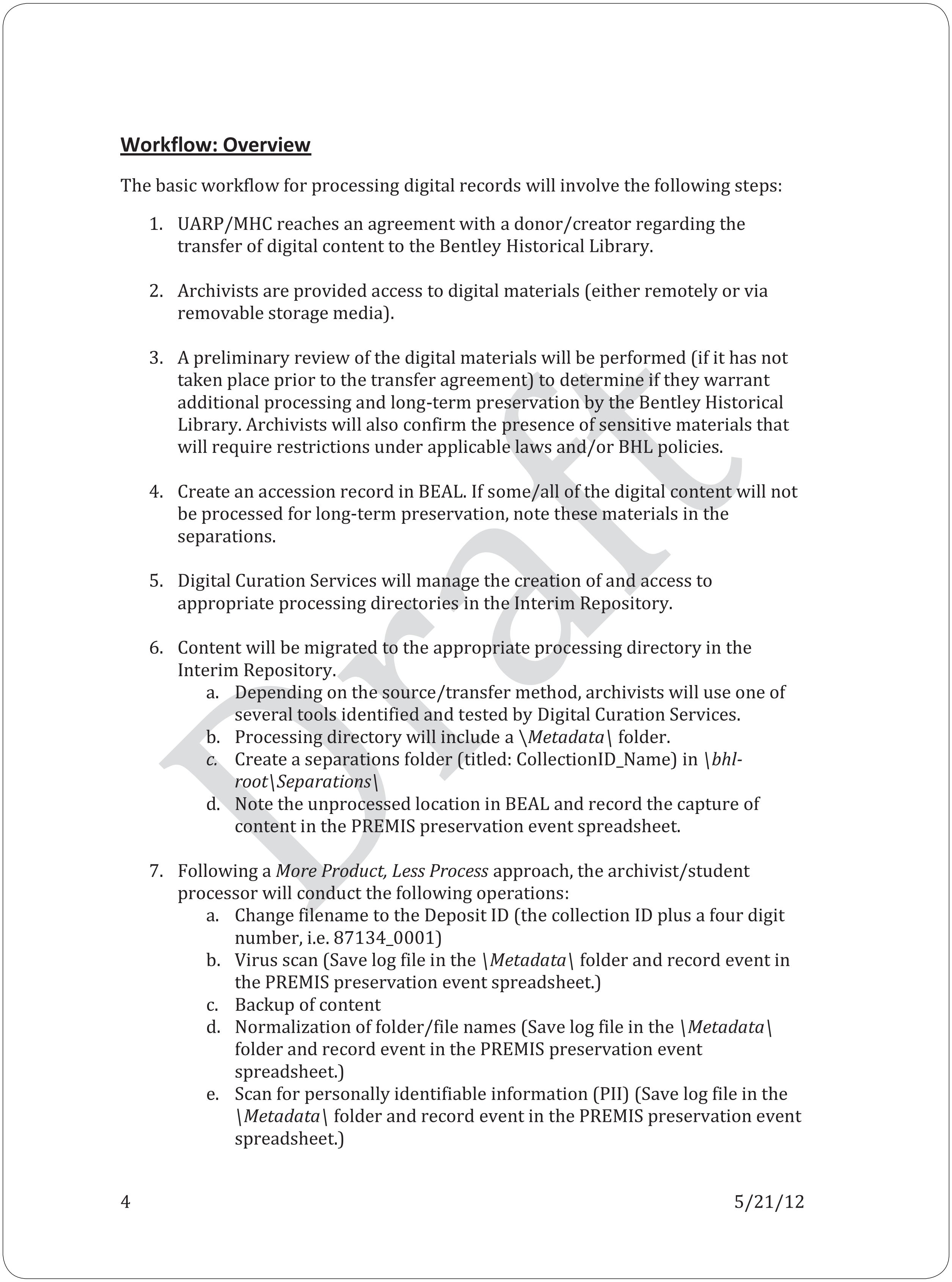
This processing manual provides guidance and instructions for the processing of digital materials at the Bentley Historical Library (BHL). Procedures, tools, and the overall digital processing workflow are subject to change due to advances in professional best practices, the development of resources 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).

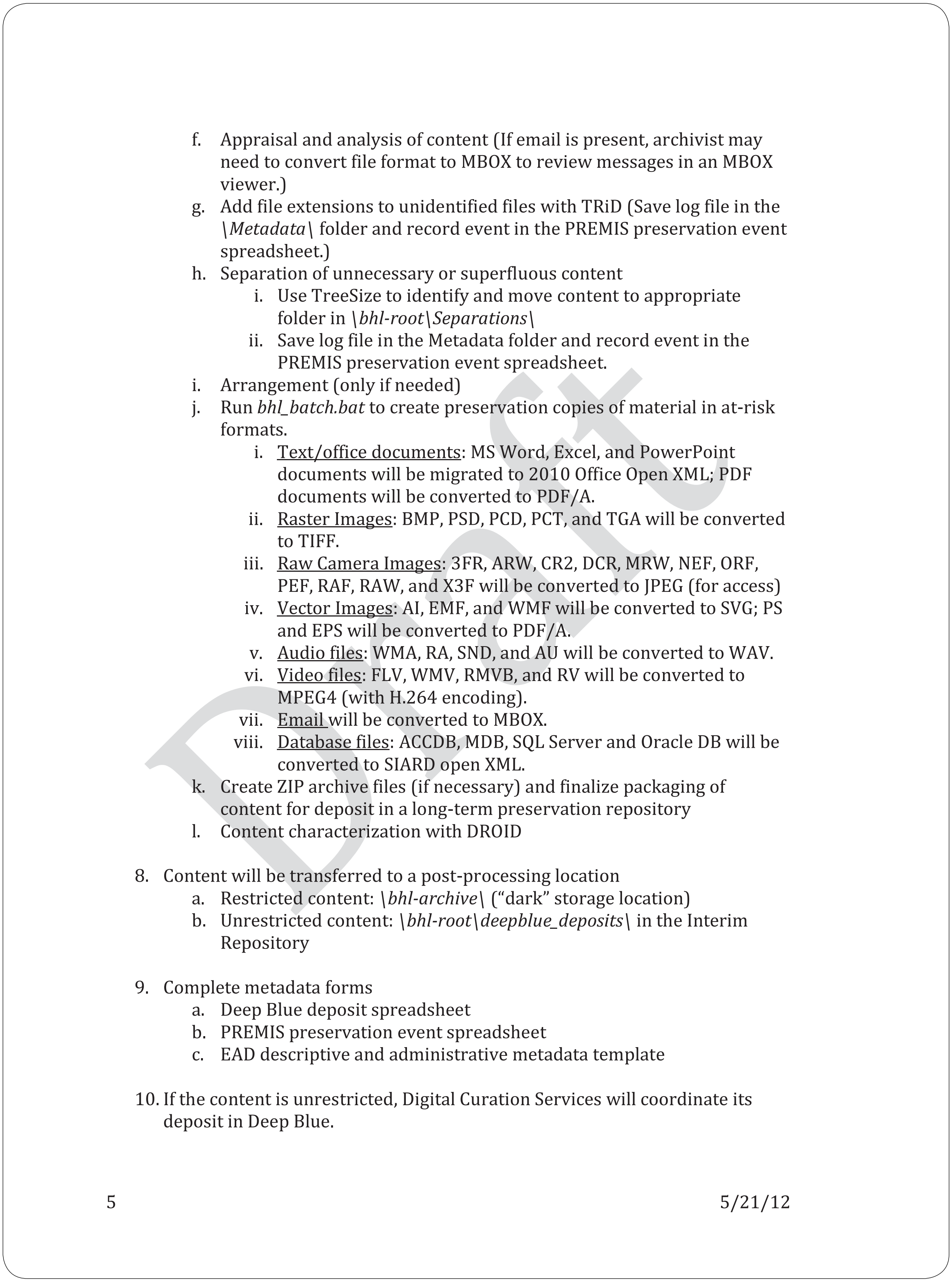
3. A preliminary review of the digital materials will be performed (if it has not taken place prior to the transfer agreement) to determine if they warrant additional processing and long-term preservation by the Bentley Historical Library. Archivists will also confirm the presence of sensitive materials 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. <u>Text/office documents</u>: MS Word, Excel, and PowerPoint documents will be migrated to 2010 Office Open XML; PDF documents will be converted to PDF/A.

   ii. <u>Raster Images</u>: BMP, PSD, PCD, PCT, and TGA will be converted to TIFF.

   iii. <u>Raw Camera Images</u>: 3FR, ARW, CR2, DCR, MRW, NEF, ORF, PEF, RAF, RAW, and X3F will be converted to JPEG (for access)

   iv. <u>Vector Images</u>: AI, EMF, and WMF will be converted to SVG; PS and EPS will be converted to PDF/A.

   v. <u>Audio files</u>: WMA, RA, SND, and AU will be converted to WAV.

   vi. <u>Video files</u>: FLV, WMV, RMVB, and RV will be converted to MPEG4 (with H.264 encoding).

   vii. <u>Email</u> will be converted to MBOX.

   viii. <u>Database files</u>: 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

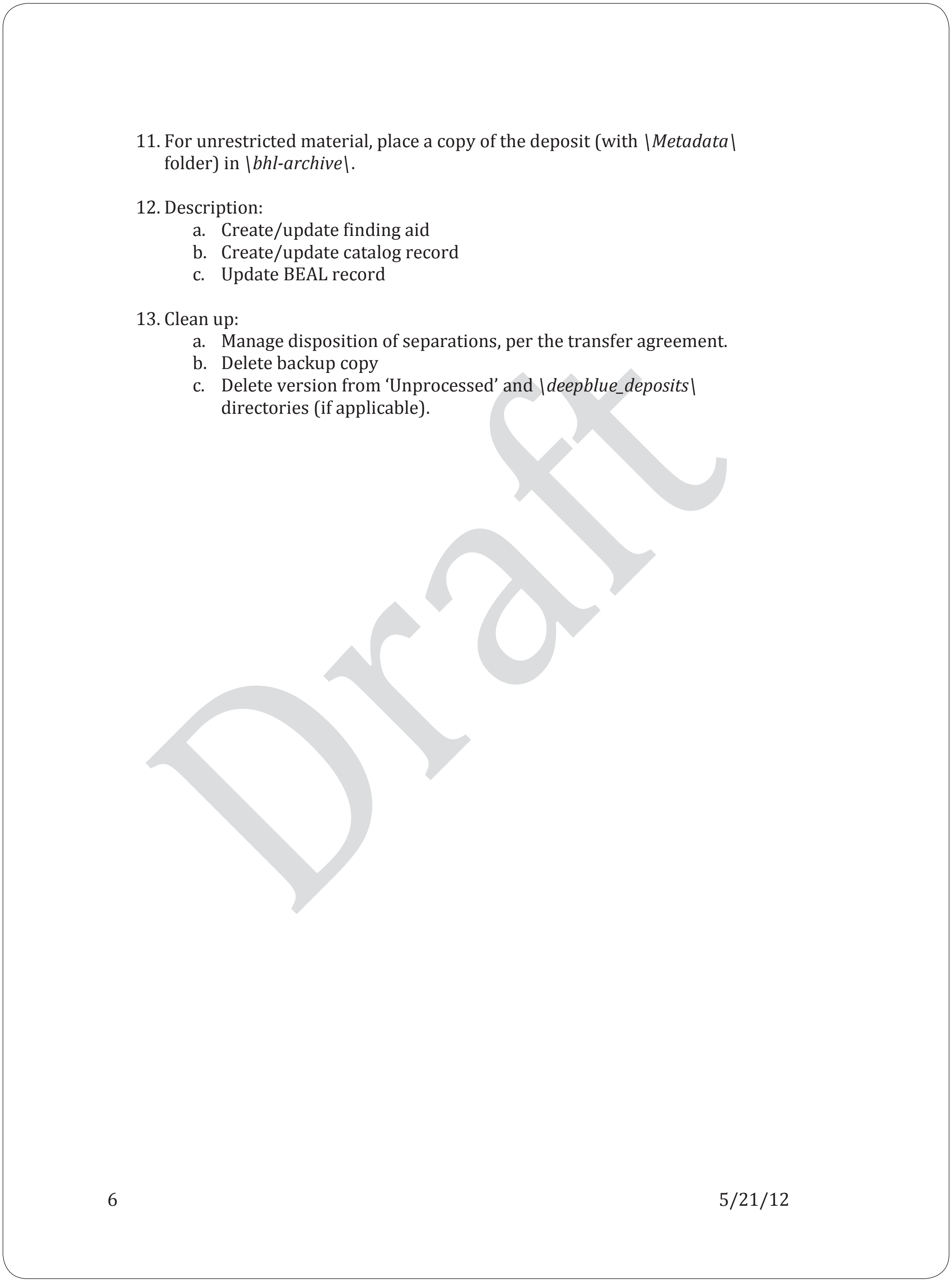
9. Complete metadata forms

   a. Deep Blue deposit spreadsheet

   b. PREMIS preservation event spreadsheet

   c. EAD descriptive and administrative metadata template

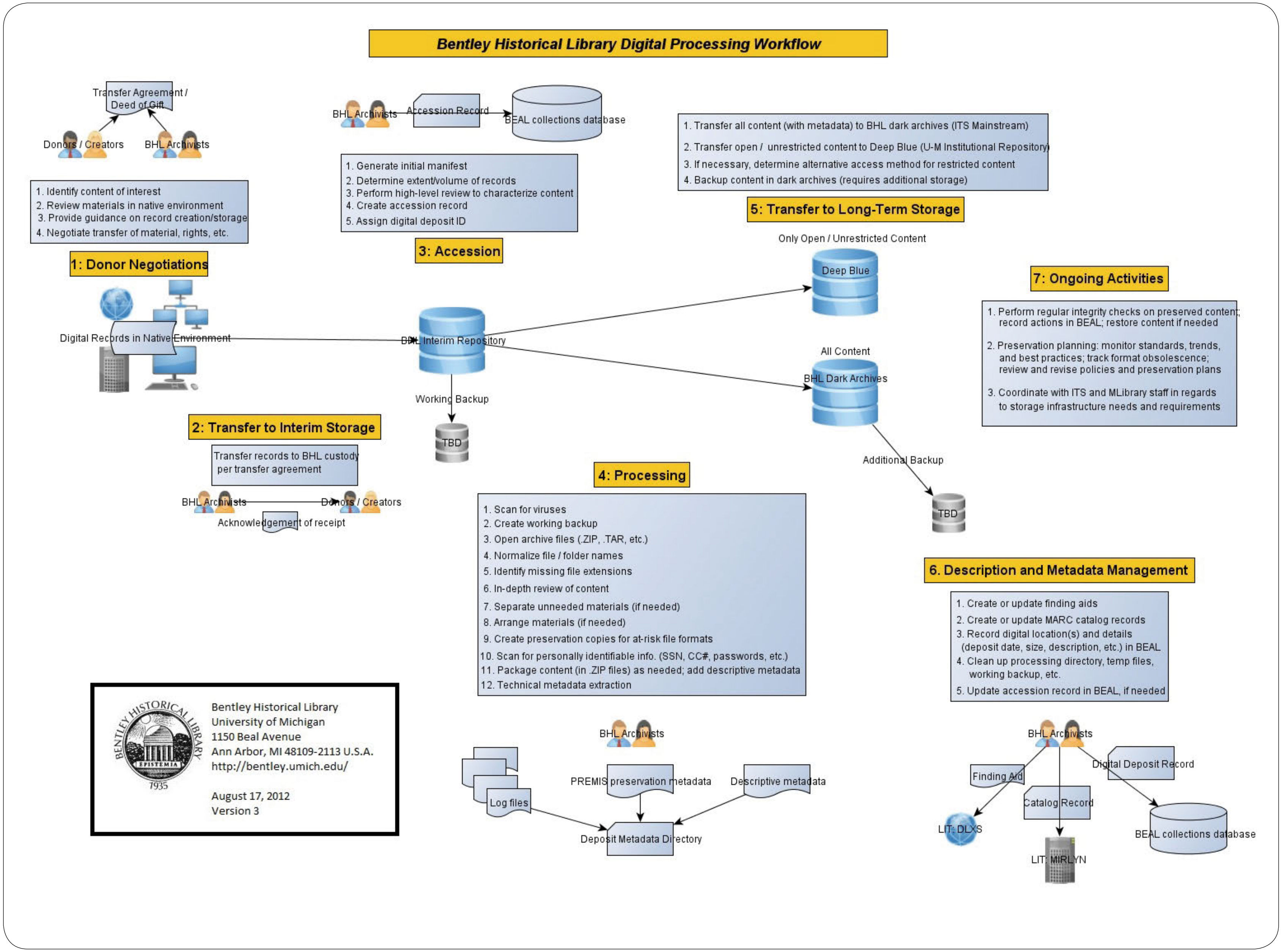
10. If the content is unrestricted, Digital Curation Services will coordinate its deposit in Deep Blue.



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
Transfer Agreement / Deed of Gift
Donors / Creators
BHL Archivists
1: Donor Negotiations
Digital Records in Native Environment
2: Transfer to Interim Storage
BHL Archivists
Donors / Creators
Acknowledgement of receipt
BHL Archivists
Accession Record
BEAL collections database
3: Accession
BHL Interim Repository
Working Backup
TBD
5: Transfer to Long-Term Storage
Only Open / Unrestricted Content
Deep Blue
All Content
BHL Dark Archives
Additional Backup
TBD
4: Processing
BHL Archivists
PREMIS preservation metadata
Descriptive metadata
Log files
Deposit Metadata Directory
7: Ongoing Activities
6. Description and Metadata Management
BHL Archivists
Finding Aid
Catalog Record
Digital Deposit Record
LIT: DLXS
LIT: MIRLYN
BEAL collections database

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.

2: Transfer to Interim Storage

Transfer records to BHL custody per transfer agreement

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

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

5: Transfer to Long-Term Storage

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)

6. Description and Metadata Management

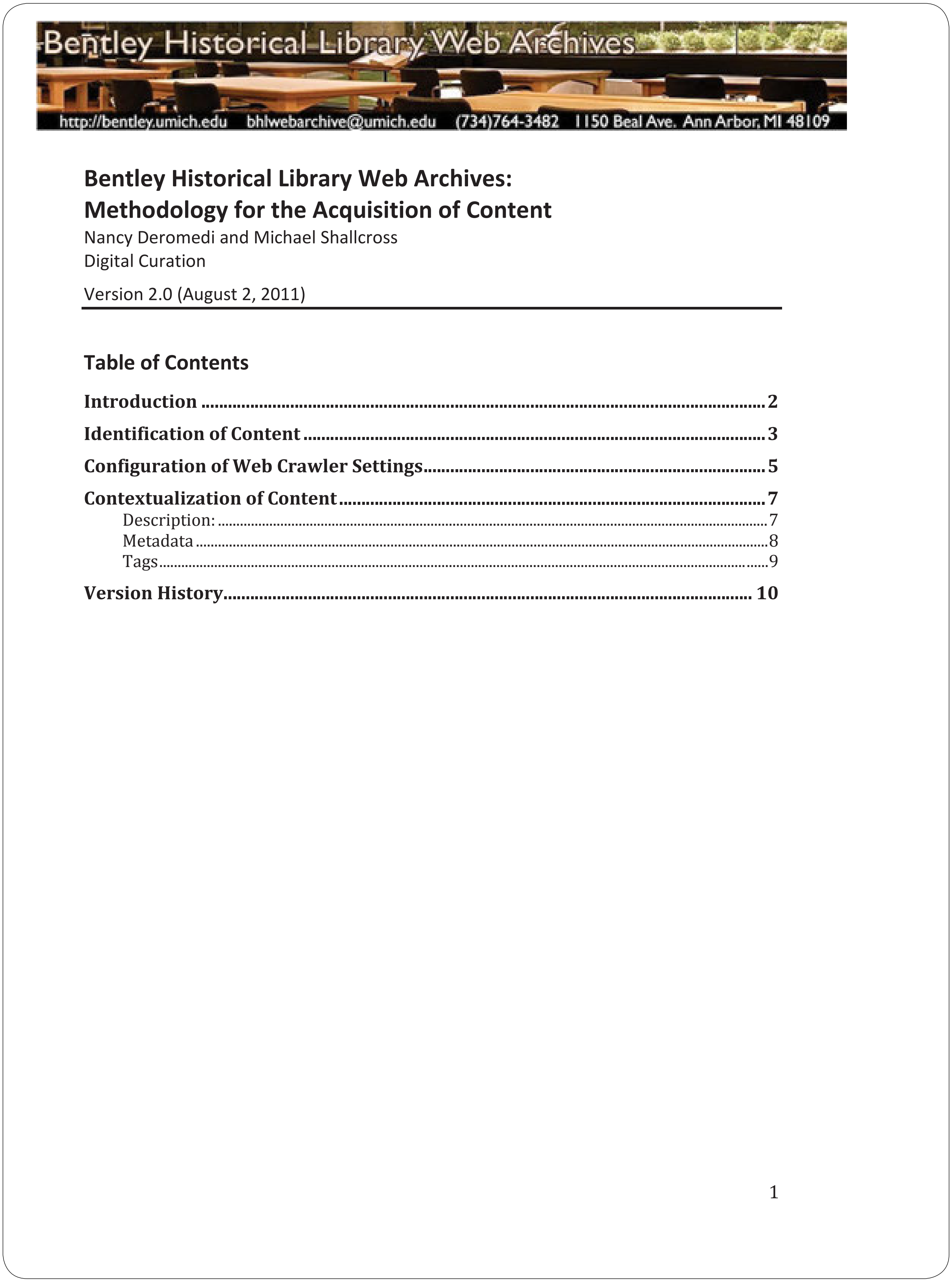
1. Create or update finding aids
2. Create or update MARC catalog records
3. Record digital location(s) and details (deposit date, size, description, etc.) in BEAL
4. Clean up processing directory, temp files, working backup, etc.
5. Update accession record in BEAL, if needed

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

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

http://bentley.umich.edu bhlwebarchive@umich.edu (734)764-3482 1150 Beal Ave. Ann Arbor, MI 48109

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

Nancy Deromedi and Michael Shallcross
Digital Curation

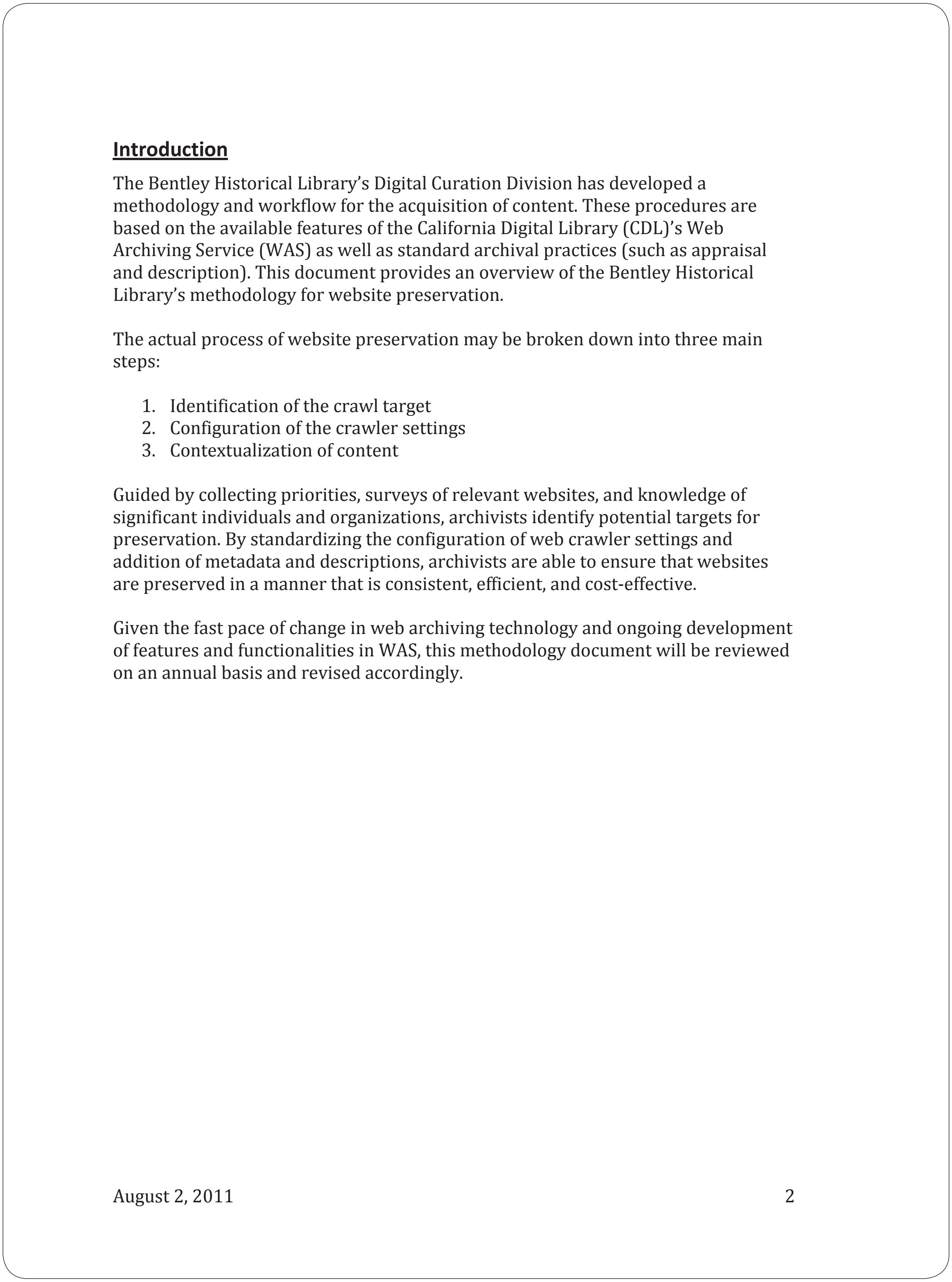
Version 2.0 (August 2, 2011)

Table of Contents

UNIVERSITY OF MICHIGAN

BHL Web Archives: Methodology for the Acquisition of Content

http://bentley.umich.edu/dchome/webarchives/BHL_WebArchives_Methodology.pdf



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.

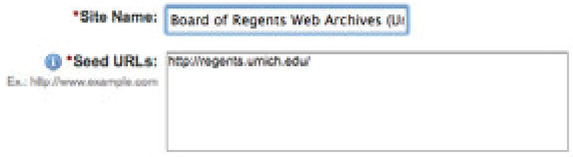
Identification of Content

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).



*Site Name: Board of Regents Web Archives (U

*Seed URLs: http://regents.umich.edu/
Ex.: http://www.example.com

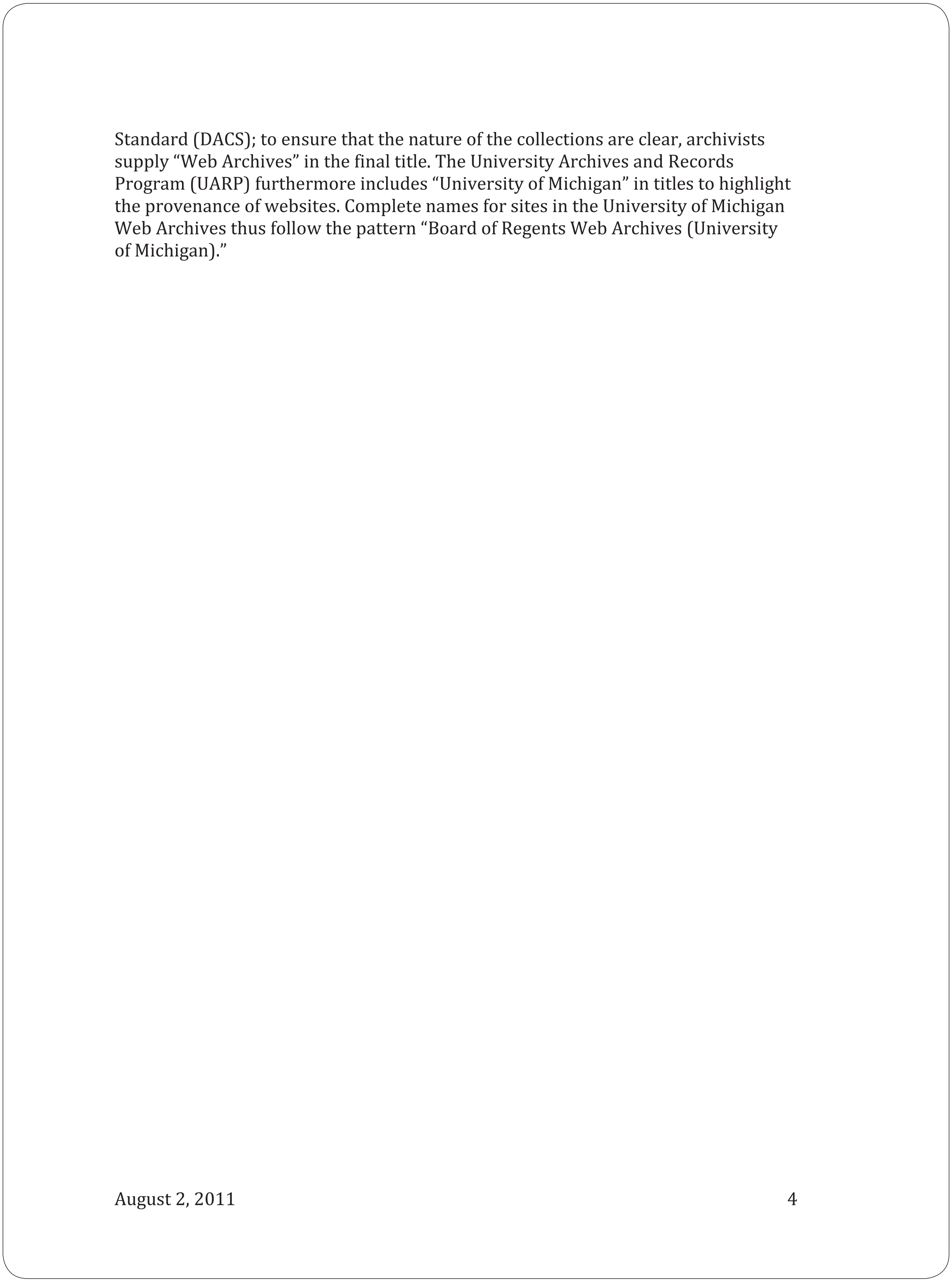
Figure 1

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

UNIVERSITY OF MICHIGAN

BHL Web Archives: Methodology for the Acquisition of Content

http://bentley.umich.edu/dchome/webarchives/BHL_WebArchives_Methodology.pdf

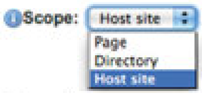


Standard (DACS); to ensure that the nature of the collections are clear, archivists supply “Web Archives” in the final title. The University Archives and Records Program (UARP) furthermore includes “University of Michigan” in titles to highlight the provenance of websites. Complete names for sites in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



Scope: Host site

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.

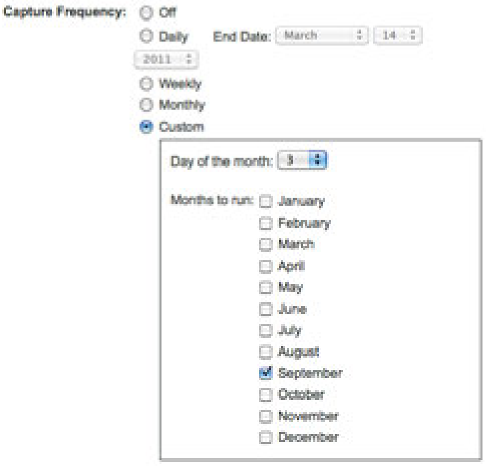
Linked pages: determines whether or not content from other hosts/URLs will be captured; archivists have two options for this setting. If set to "No," the crawler will only archive materials on the seed URL entered by the archivist; if "Yes," the crawler will follow hypertext links one 'hop' to capture 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).



Capture Frequency: ○ Off
○ Daily End Date: March 14 2011
○ Weekly
○ Monthly
● Custom

Day of the month: 3

Months to run: □ January
□ February
□ March
□ April
□ May
□ June
□ July
□ August
☑ September
□ October
□ November
□ December

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.

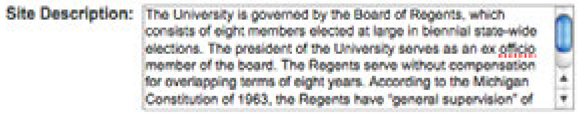
As the foregoing discussion reveals, the accurate and effective configuration of crawl settings must be based on the archivist's appraisal of content and understanding of the target site's structure. The failure to consider these factors may lead to a capture that, on the one hand, is narrowly circumscribed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

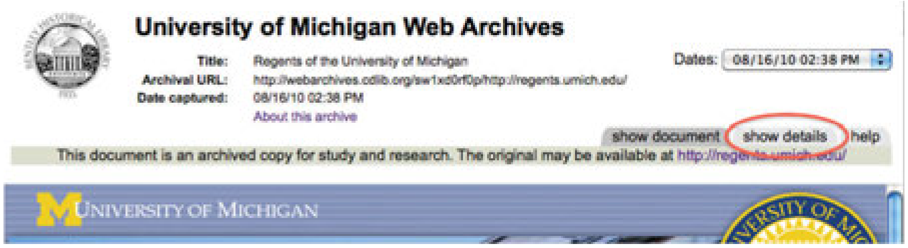
WAS provides a 'Site Description' field so that archivists may contextualize preserved websites with an overview of the creator and/or subject matter (see Figure 4).



Site Description: The University is governed by the Board of Regents, which consists of eight members elected at large in biennial state-wide elections. The president of the University serves as an ex officio member of the board. The Regents serve without compensation for overlapping terms of eight years. According to the Michigan Constitution of 1963, the Regents have "general supervision" of

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).



University of Michigan Web Archives

Title: Regents of the University of Michigan
Archival URL: http://webarchives.cdlib.org/sw1xd0rf0p/http://regents.umich.edu/
Date captured: 08/16/10 02:38 PM
About this archive

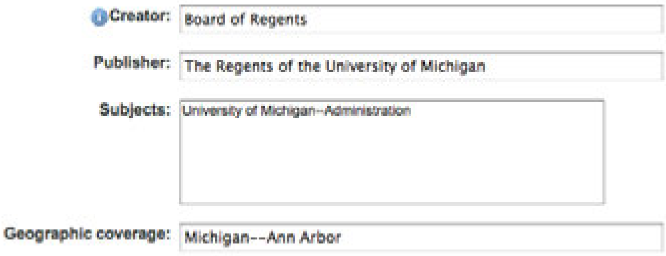
Dates: 08/16/10 02:38 PM

This document is an archived copy for study and research. The original may be available at http://regents.umich.edu/

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).



Creator: Board of Regents

Publisher: The Regents of the University of Michigan

Subjects: University of Michigan--Administration

Geographic coverage: Michigan--Ann Arbor

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).

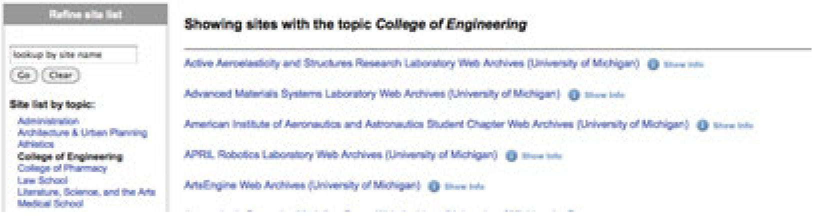
- *Publisher* refers to the entity ultimately responsible for the production and presentation of content. Although the publisher may often be identical to the creator, the Regents of the University of Michigan are recognized as the collective publisher for all sites affiliated with the university. Similar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



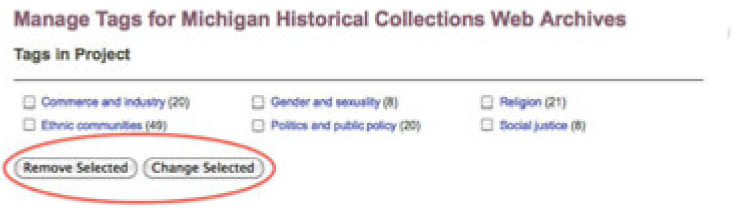
Site list by topic:
- Administration
- Architecture & Urban Planning
- Athletics
- **College of Engineering**
- College of Pharmacy
- Law School
- Literature, Science, and the Arts
- Medical School

Showing sites with the topic *College of Engineering*

- Active Aeroelasticity and Structures Research Laboratory Web Archives (University of Michigan) Show Info
- Advanced Materials Systems Laboratory Web Archives (University of Michigan) Show Info
- American Institute of Aeronautics and Astronautics Student Chapter Web Archives (University of Michigan) Show Info
- APRIL Robotics Laboratory Web Archives (University of Michigan) Show Info
- ArtsEngine Web Archives (University of Michigan) Show Info

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).



Manage Tags for Michigan Historical Collections Web Archives

Tags in Project

□ Commerce and industry (20)
□ Ethnic communities (49)
□ Gender and sexuality (8)
□ Politics and public policy (20)
□ Religion (21)
□ Social justice (8)

Figure 8

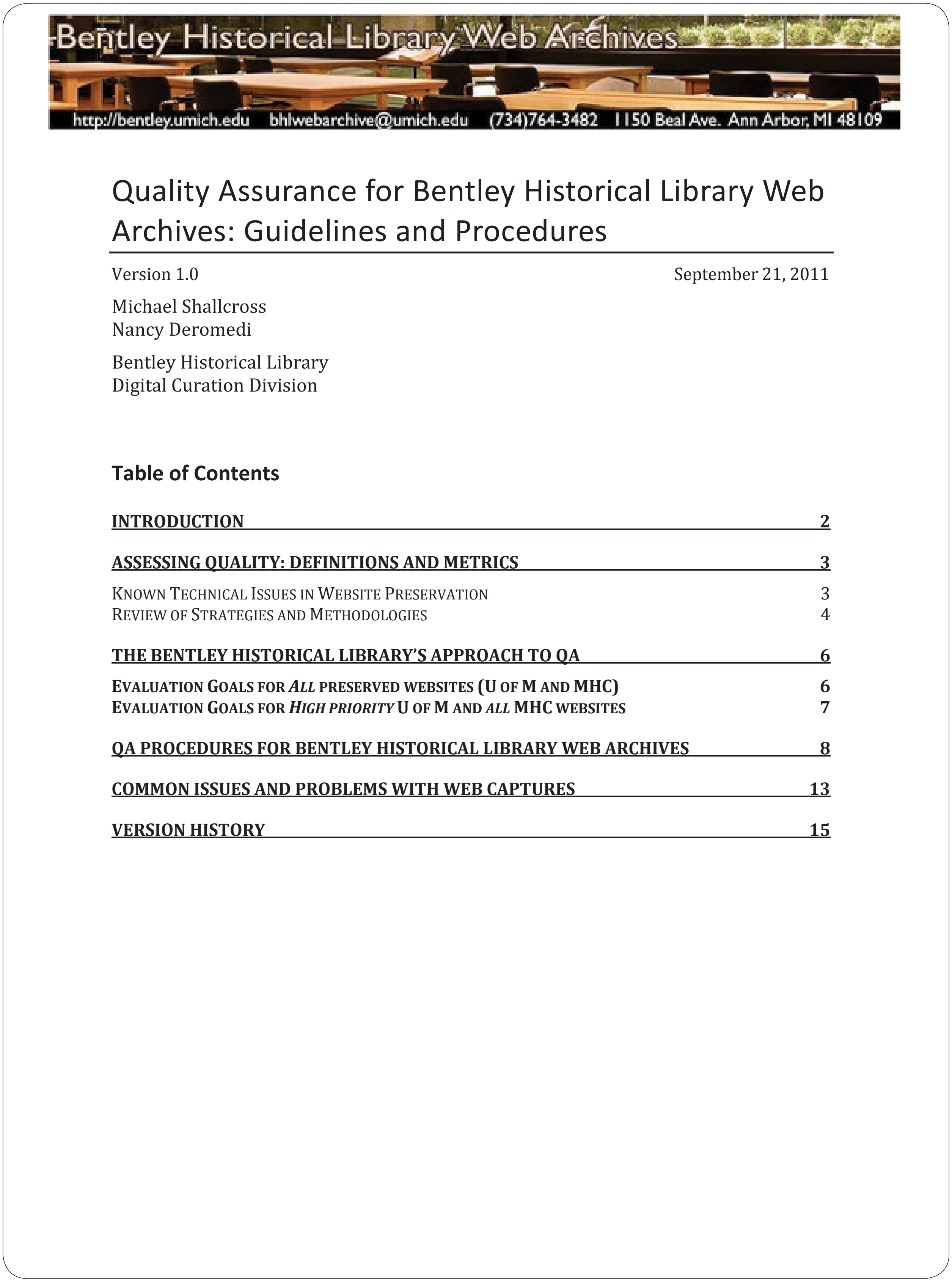
Many sites in the web archives do not have tags because they do not fit into these established categories and tagging is only effective when there are a significant number (i.e. five or more) of related sites. Archivists may, however, add tags to existing archived websites should the need arise.

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.

UNIVERSITY OF MICHIGAN

Quality Assurance for BHL Web Archives

http://bentley.umich.edu/dchome/webarchives/BHL_WebArchives_QAguidelines.pdf



Bentley Historical Library Web Archives

http://bentley.umich.edu bhlwebarchive@umich.edu (734)764-3482 1150 Beal Ave. Ann Arbor, MI 48109

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

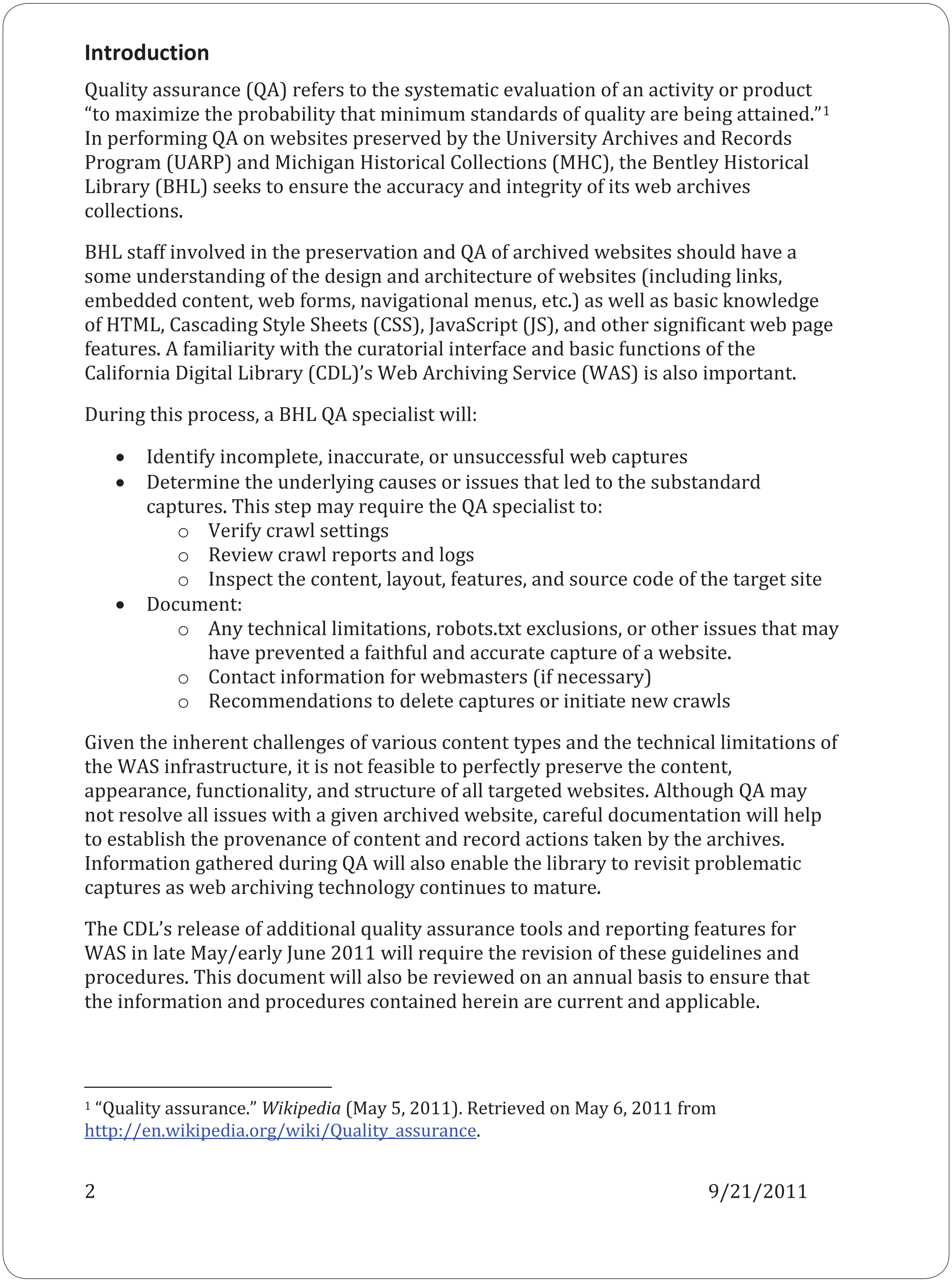
Version 1.0 September 21, 2011

Michael Shallcross
Nancy Deromedi

Bentley Historical Library
Digital Curation Division

Table of Contents

UNIVERSITY OF MICHIGAN
Quality Assurance for BHL Web Archives
http://bentley.umich.edu/dchome/webarchives/BHL_WebArchives_QAguidelines.pdf



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.

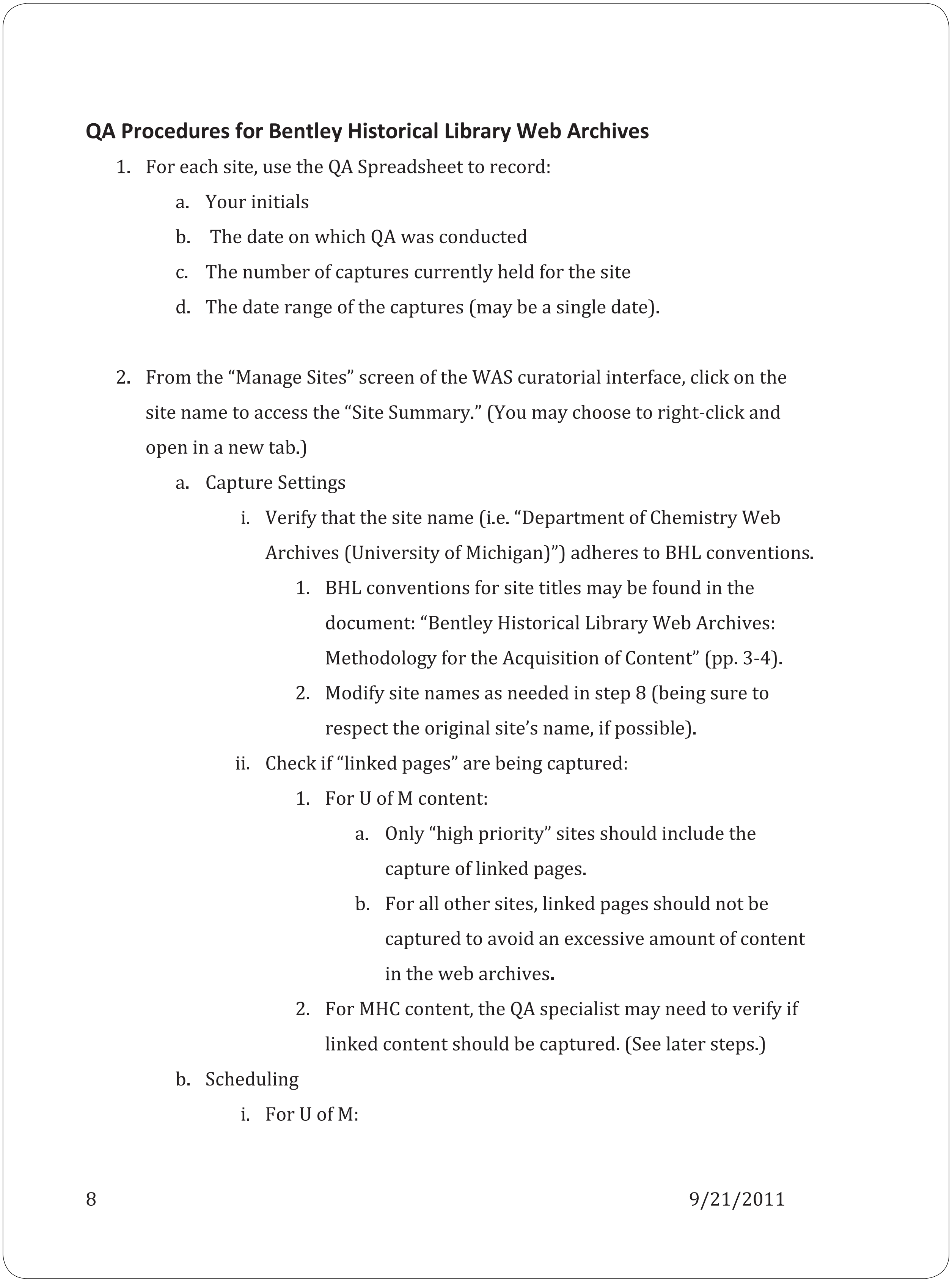
The CDL's release of additional quality assurance tools and reporting features for WAS in late May/early June 2011 will require the revision of these guidelines and procedures. This document will also be reviewed on an annual basis to ensure that the information and procedures contained herein are current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

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

UNIVERSITY OF MICHIGAN

Quality Assurance for BHL Web Archives

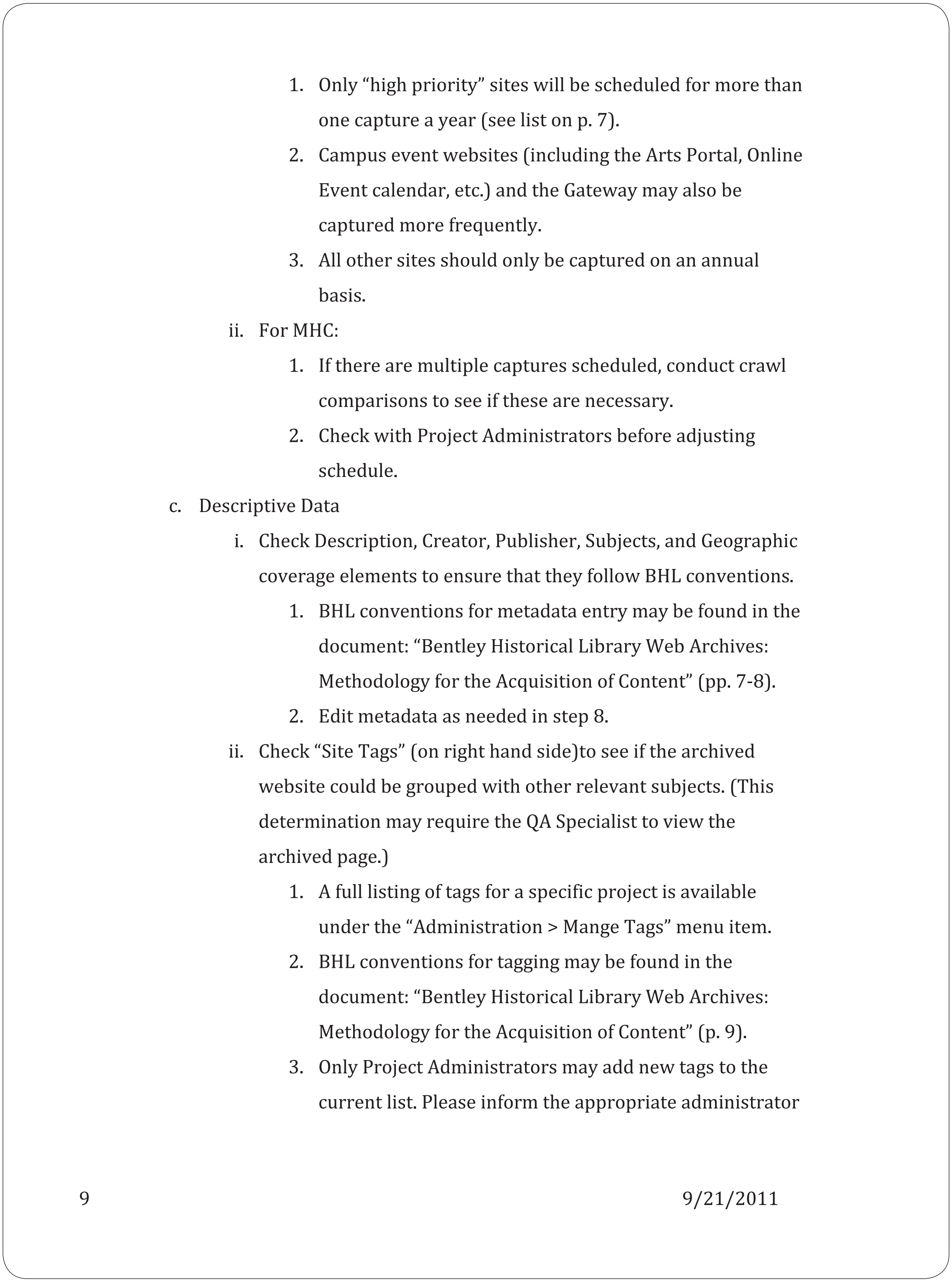
http://bentley.umich.edu/dchome/webarchives/BHL_WebArchives_QAguidelines.pdf



QA Procedures for Bentley Historical Library Web Archives

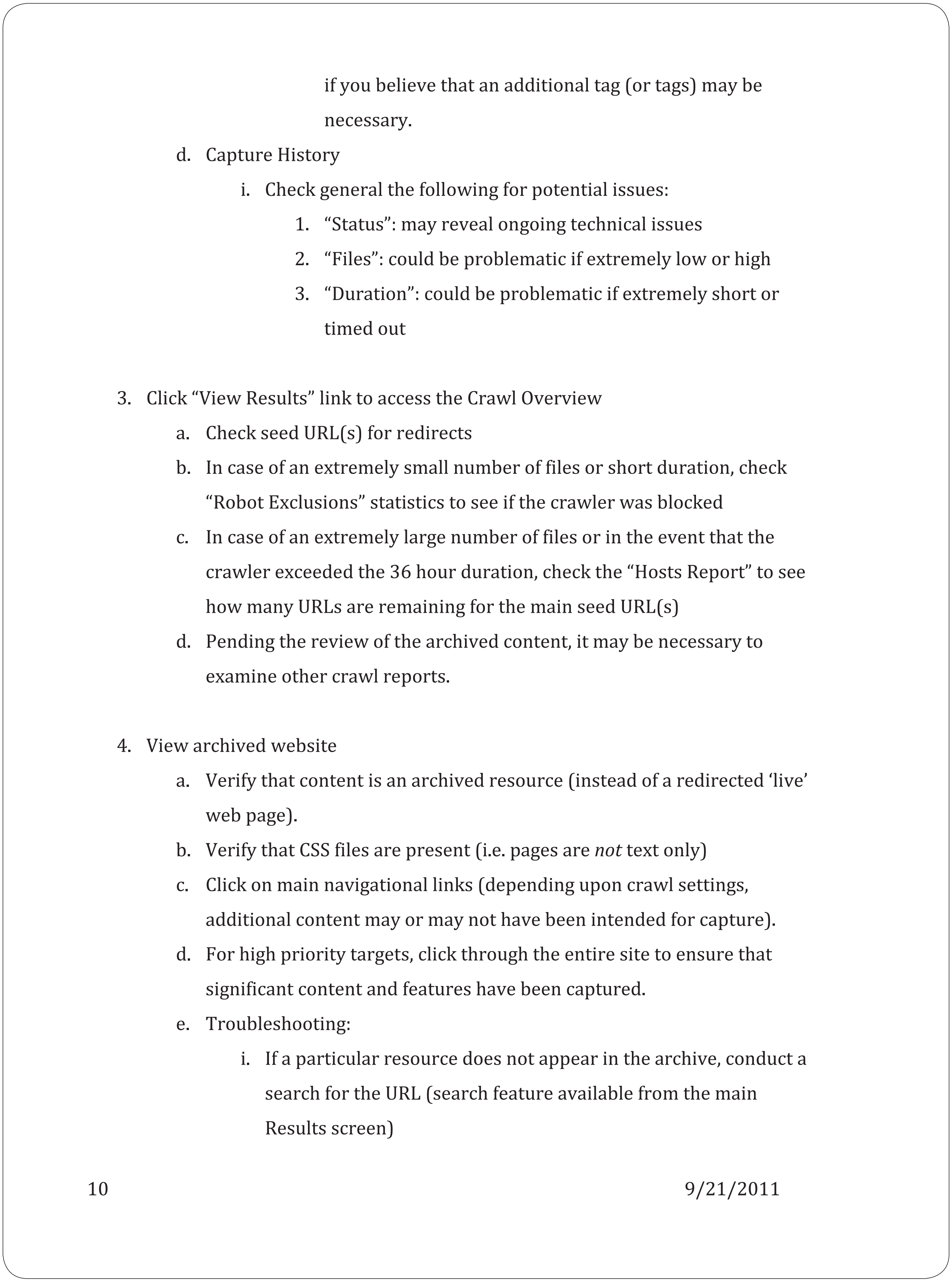
1. For each site, use the QA Spreadsheet to record:
    a. Your initials
    b. The date on which QA was conducted
    c. The number of captures currently held for the site
    d. The date range of the captures (may be a single date).

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



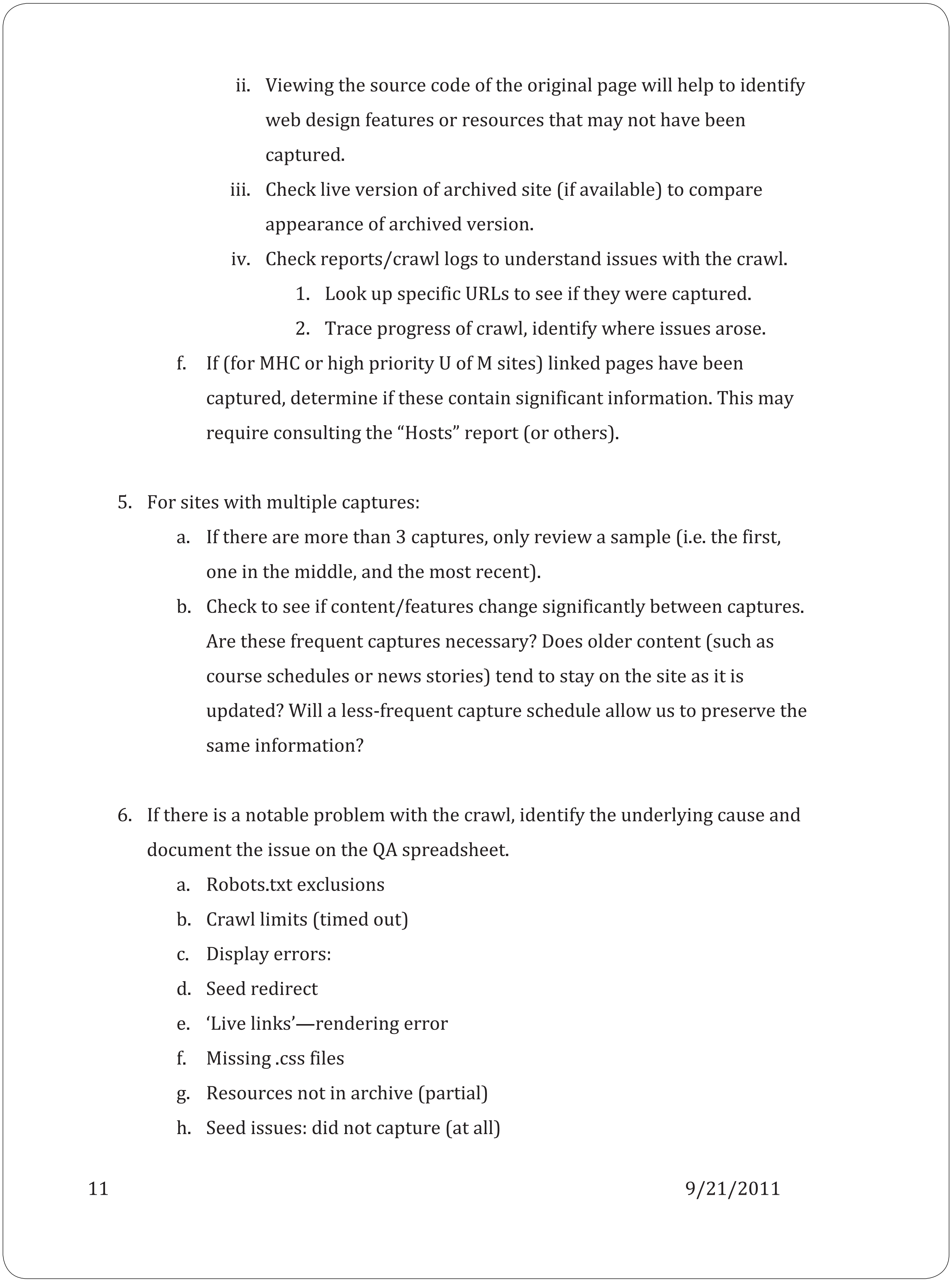
if you believe that an additional tag (or tags) may be necessary.

   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)

UNIVERSITY OF MICHIGAN
Quality Assurance for BHL Web Archives
http://bentley.umich.edu/dchome/webarchives/BHL_WebArchives_QAguidelines.pdf



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:

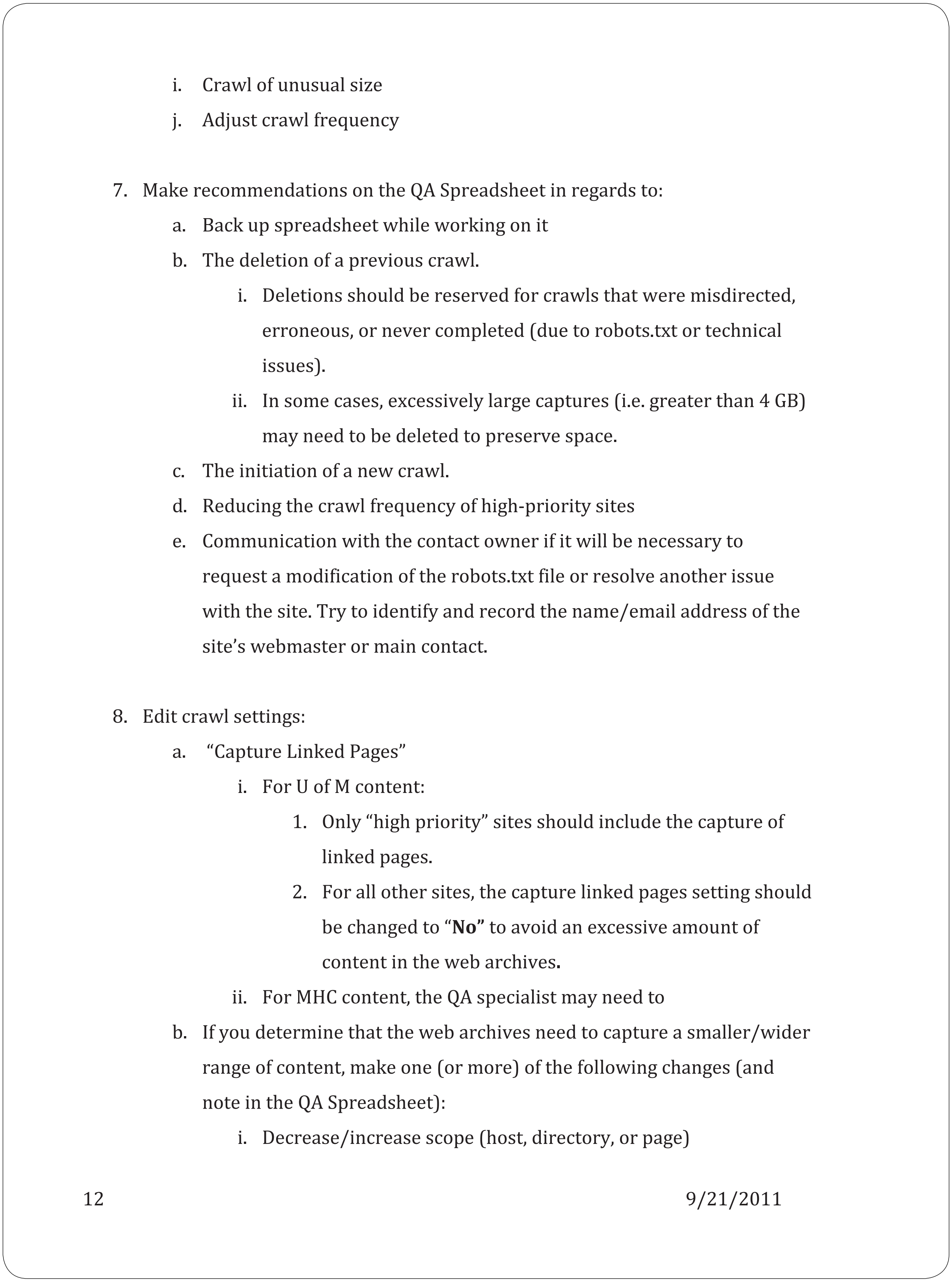
a. If there are more than 3 captures, only review a sample (i.e. the first, one in the middle, and the most recent).

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)

11 9/21/2011



    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.

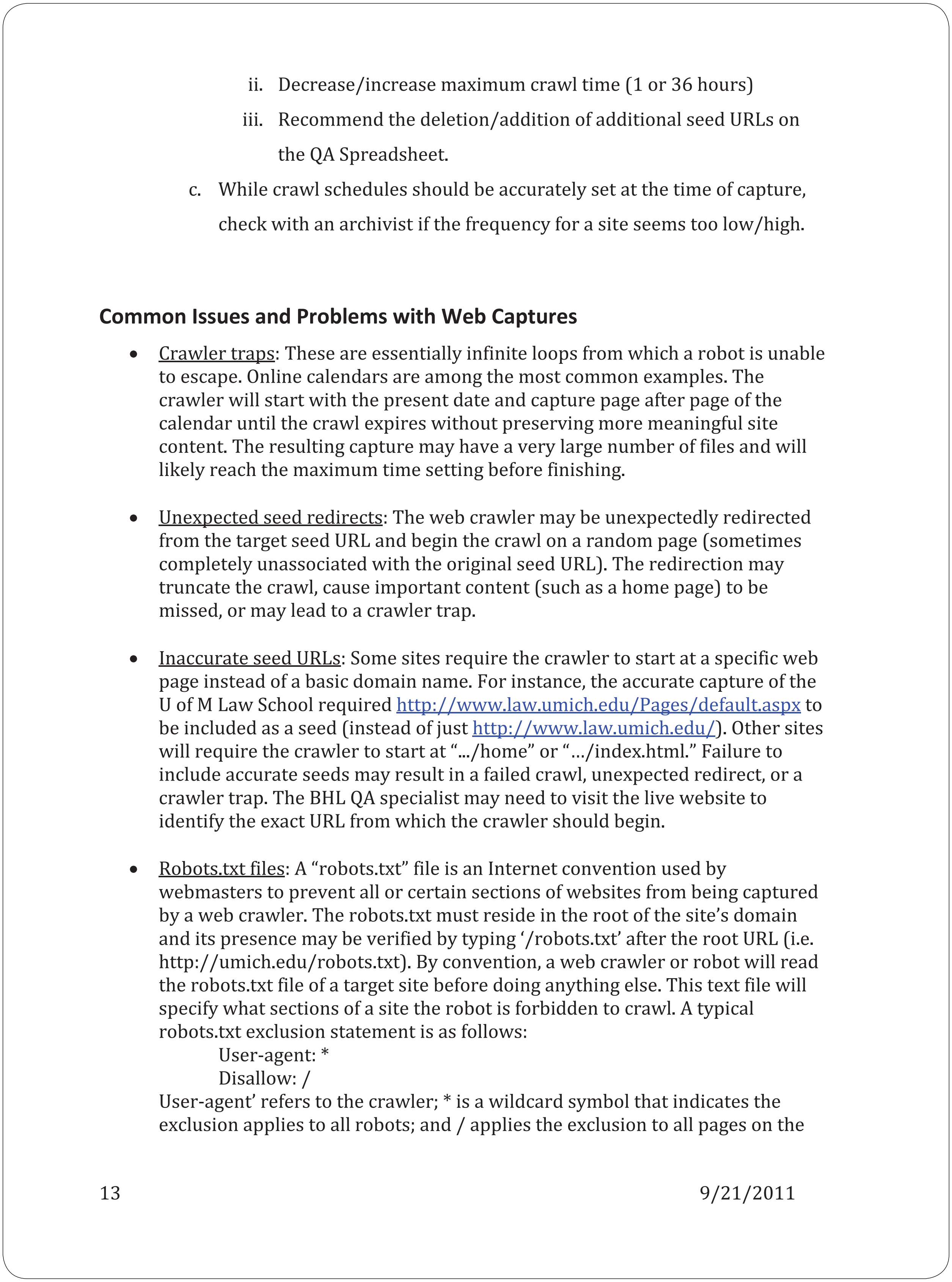
8. Edit crawl settings:
    a. "Capture Linked Pages"
        i. For U of M content:
            1. Only "high priority" sites should include the capture of linked pages.
            2. For all other sites, the capture linked pages setting should be changed to "**No"** to avoid an excessive amount of 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)

12 9/21/2011

UNIVERSITY OF MICHIGAN

Quality Assurance for BHL Web Archives

http://bentley.umich.edu/dchome/webarchives/BHL_WebArchives_QAguidelines.pdf



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.

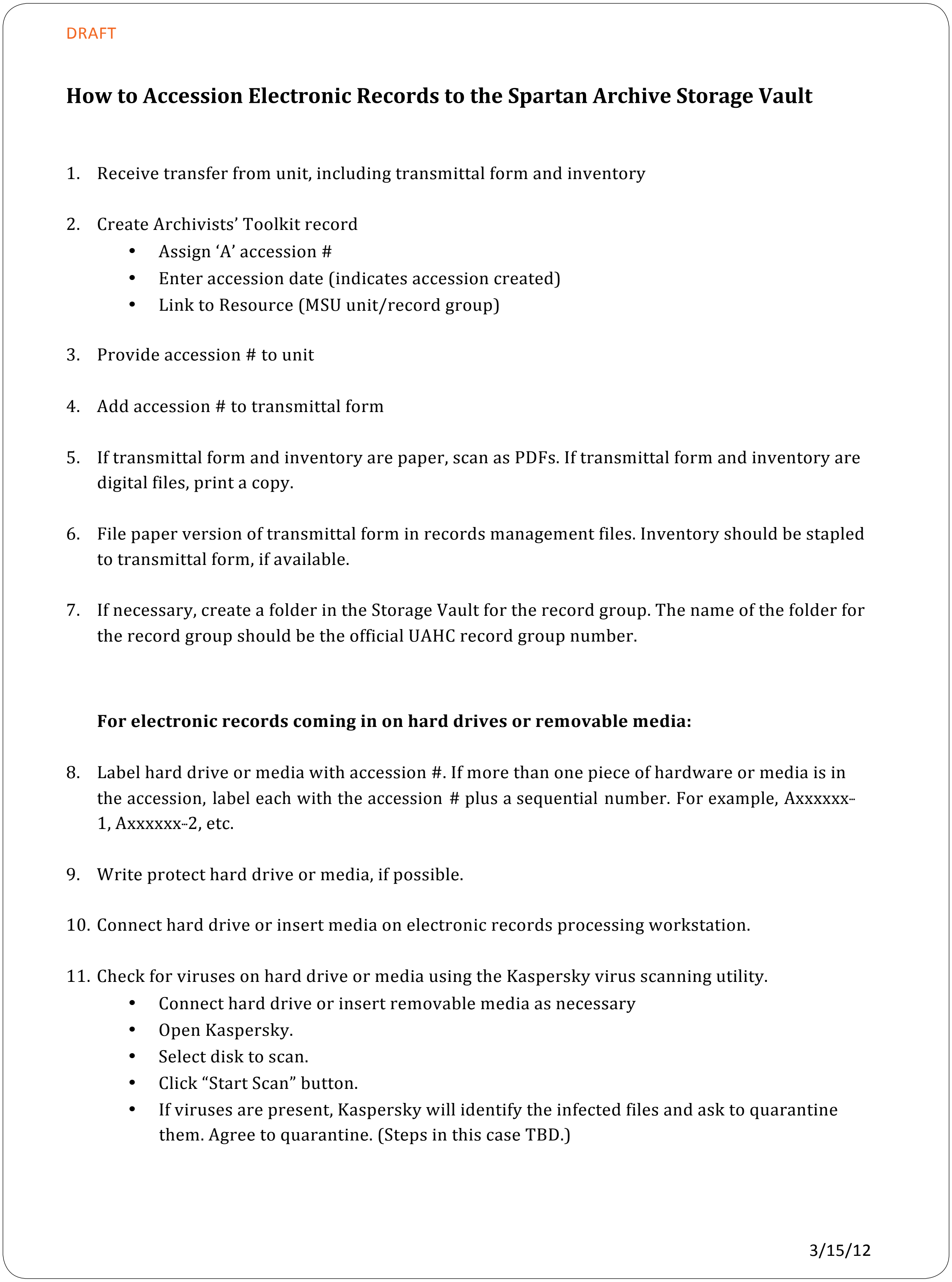
- Unexpected seed redirects: The web crawler may be unexpectedly redirected from the target seed URL and begin the crawl on a random page (sometimes completely unassociated with the original seed URL). The redirection may truncate the crawl, cause important content (such as a home page) to be missed,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

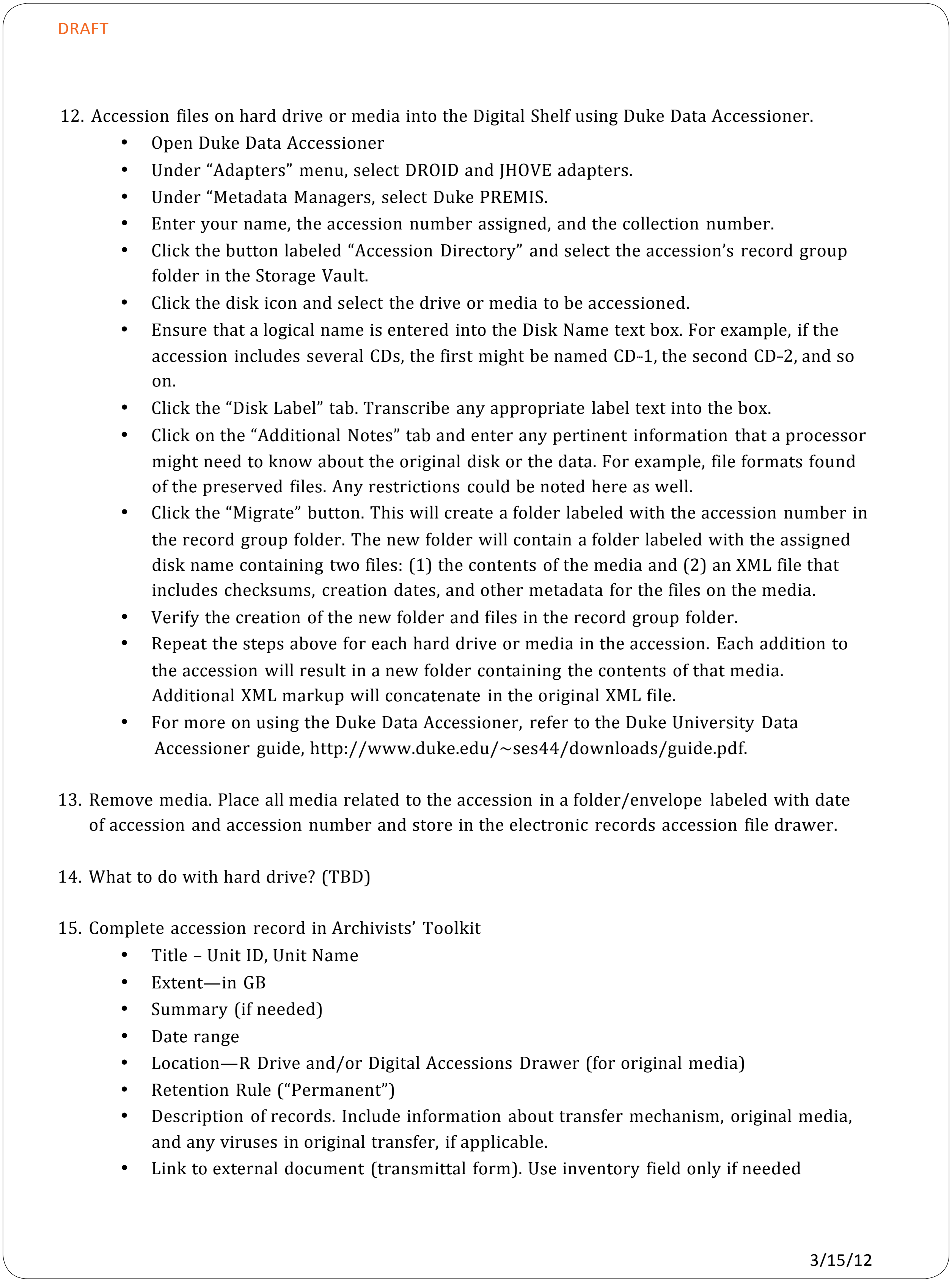
8. Label hard drive or media with accession #. If more than one piece of hardware or media is in the accession, label each with the accession # plus a sequential number. For example, Axxxxxx-1, Axxxxxx-2, etc.

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.

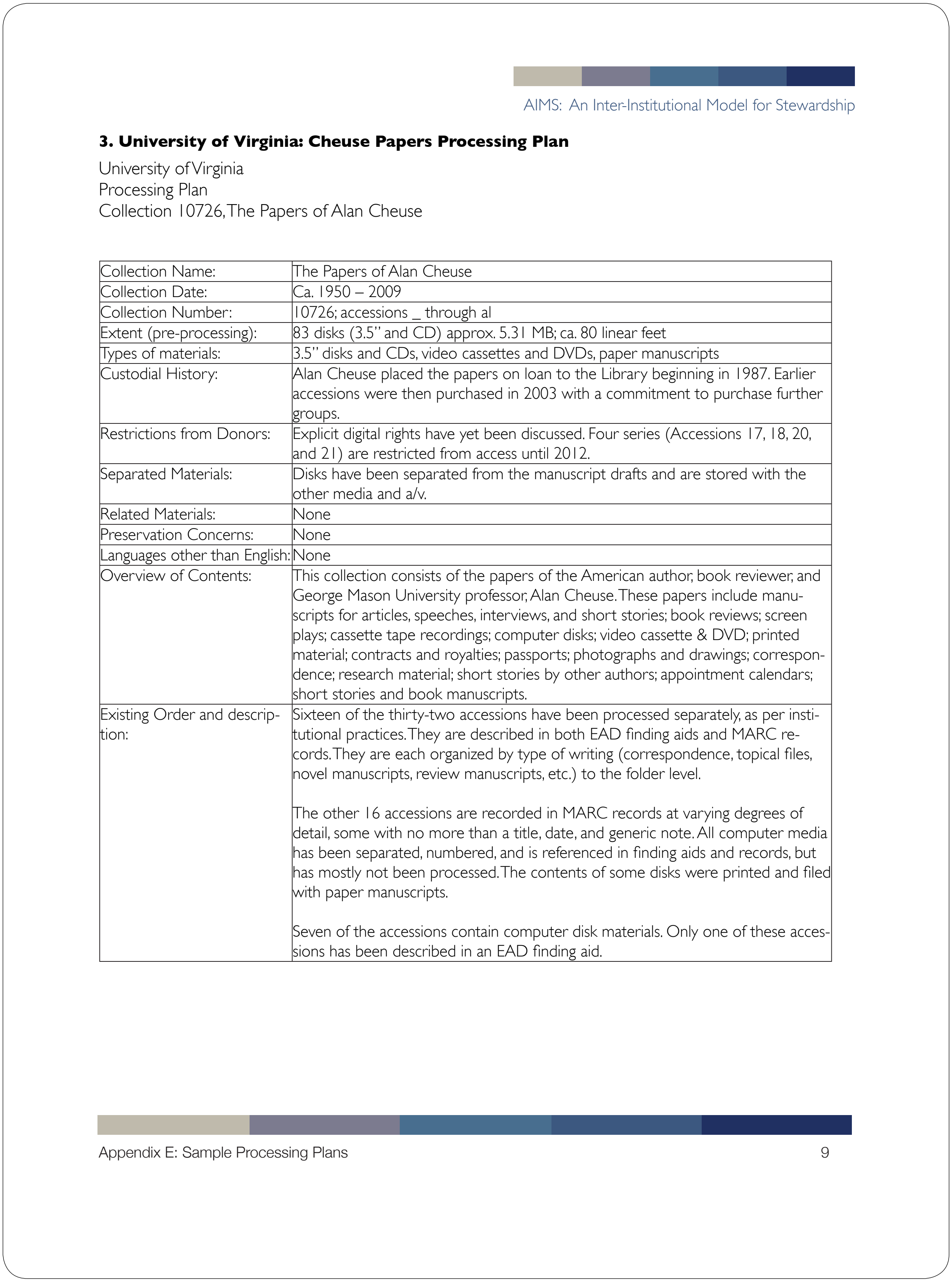
14. What to do with hard drive? (TBD)

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

UNIVERSITY OF VIRGINIA

AIMS Whitepaper. Appendix E: Sample Processing Plans
http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final_appE.pdf

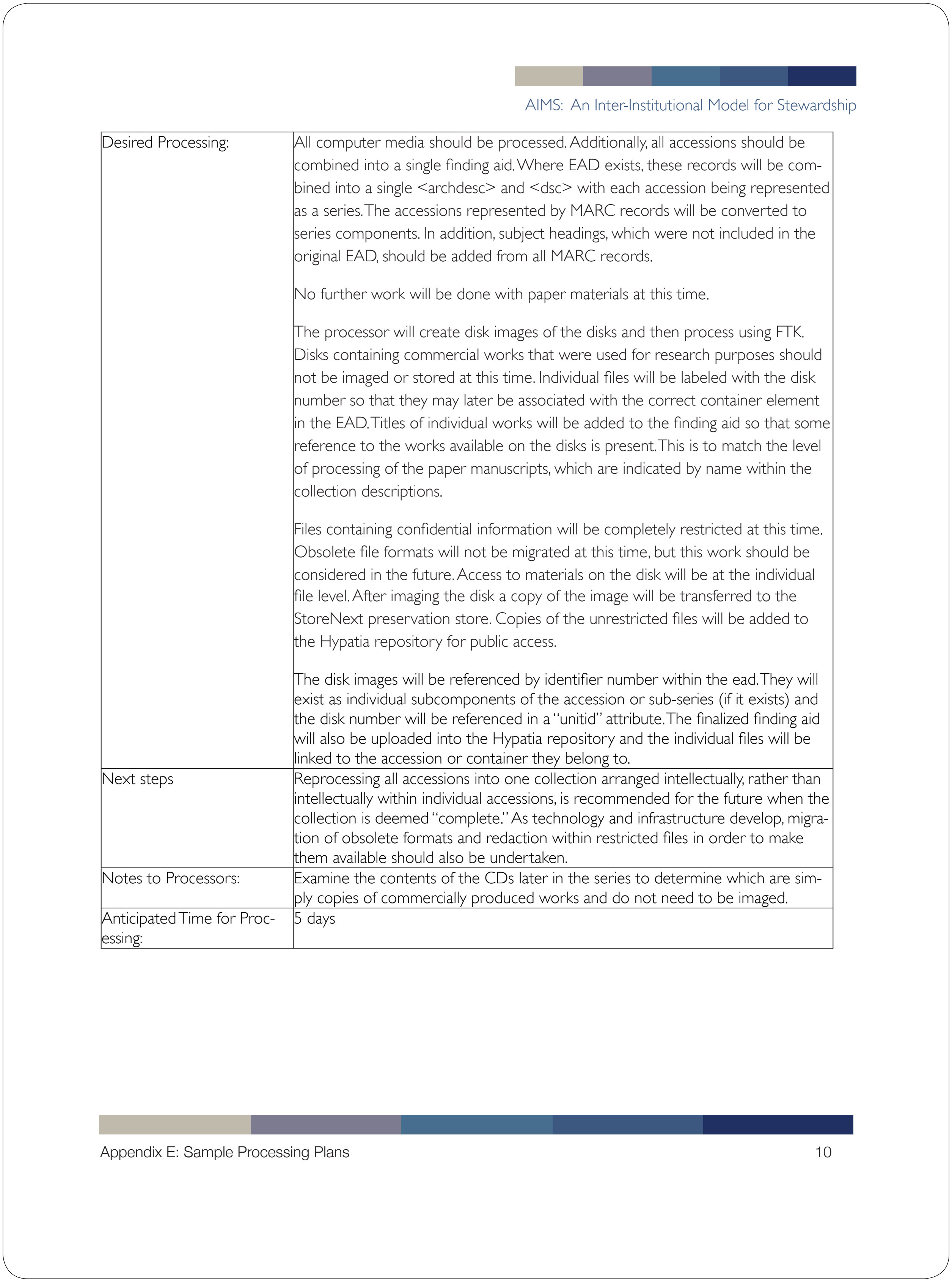


3. University of Virginia: Cheuse Papers Processing Plan

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.

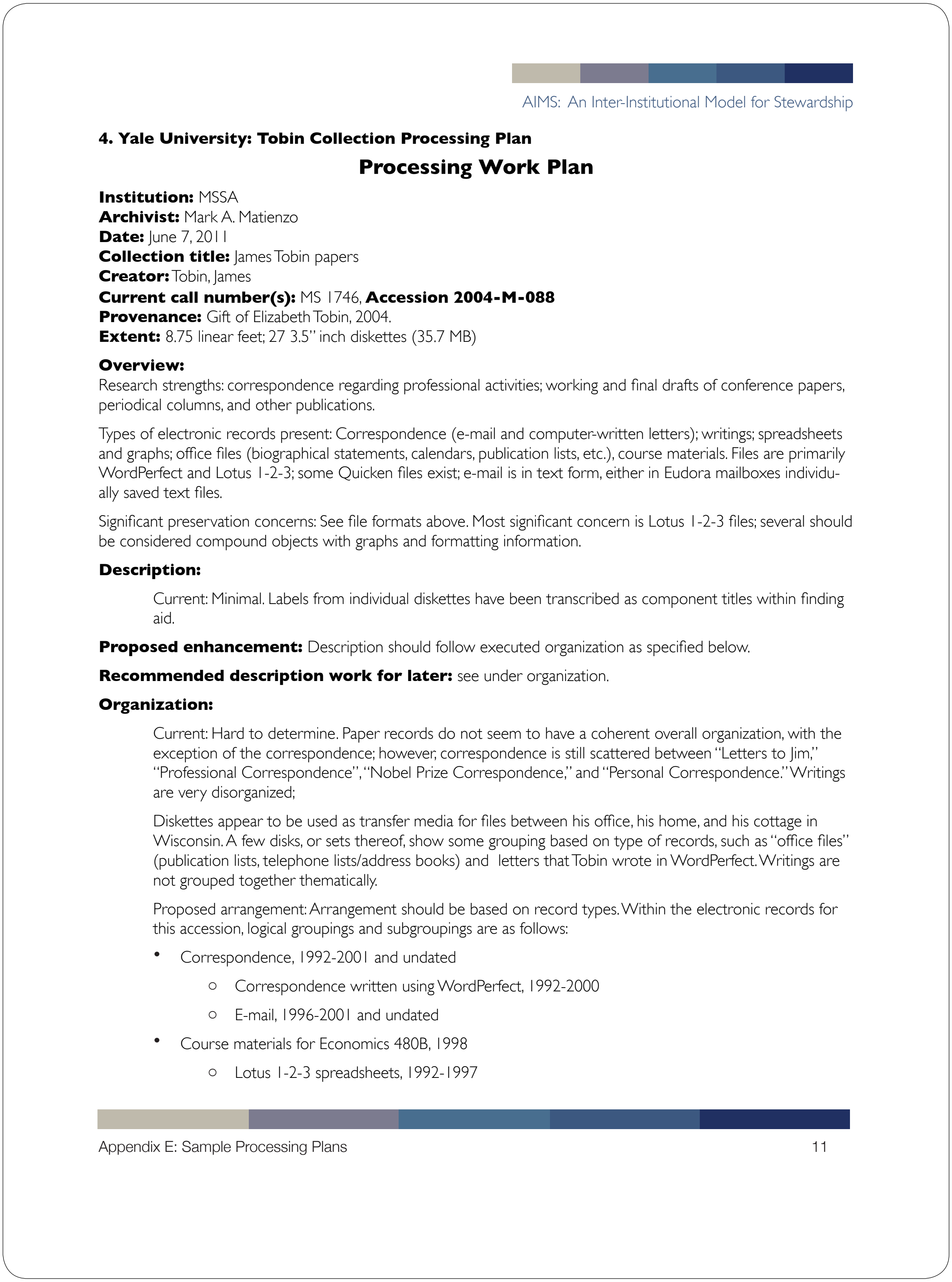
UNIVERSITY OF VIRGINIA
AIMS Whitepaper. Appendix E: Sample Processing Plans
http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final_appE.pdf



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

UNIVERSITY OF VIRGINIA

AIMS Whitepaper. Appendix E: Sample Processing Plans
http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final_appE.pdf



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:

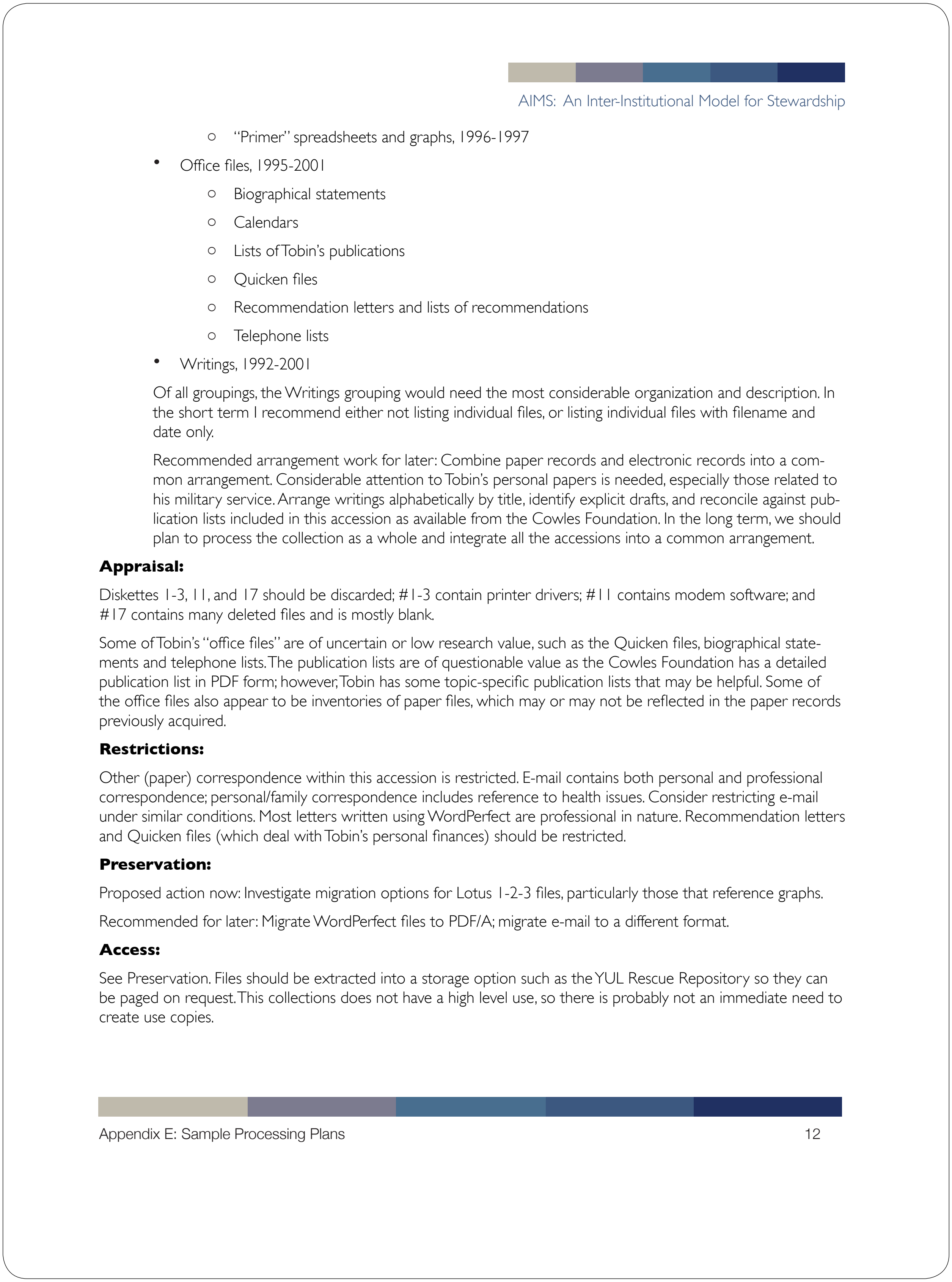
Current: Hard to determine. Paper records do not seem to have a coherent overall organization, with the exception of the correspondence; however, correspondence is still scattered between "Letters to Jim," "Professional Correspondence", "Nobel Prize Correspondence," and "Personal Correspondence." Writings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

UNIVERSITY OF VIRGINIA
AIMS Whitepaper. Appendix E: Sample Processing Plans
http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final_appE.pdf



  - ○ “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.

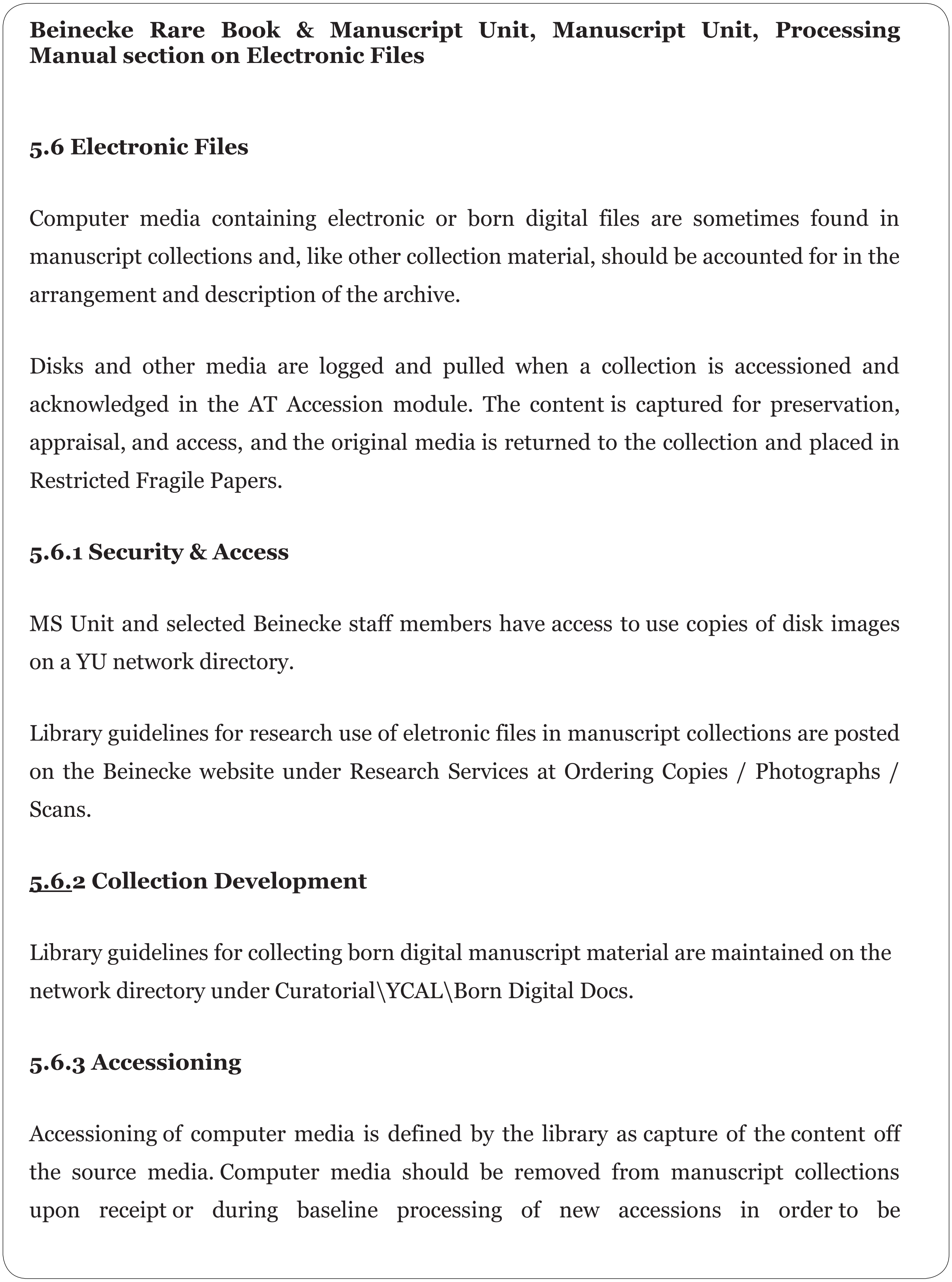
Preservation:

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]

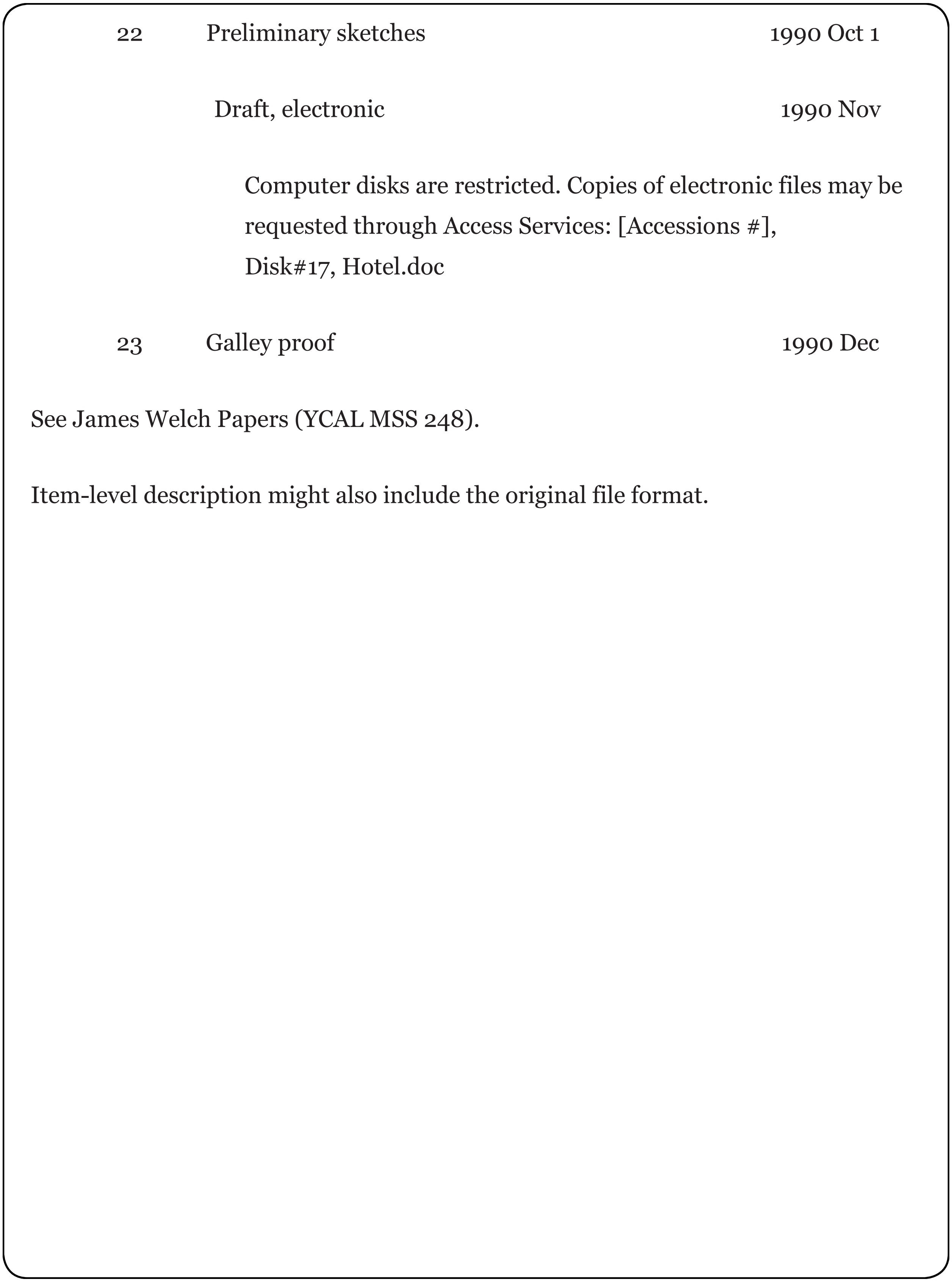
For example:

Series I. Writings

PLAYS

“Hotel Christobel”

4	21	Research notes	1990



22 Preliminary sketches 1990 Oct 1

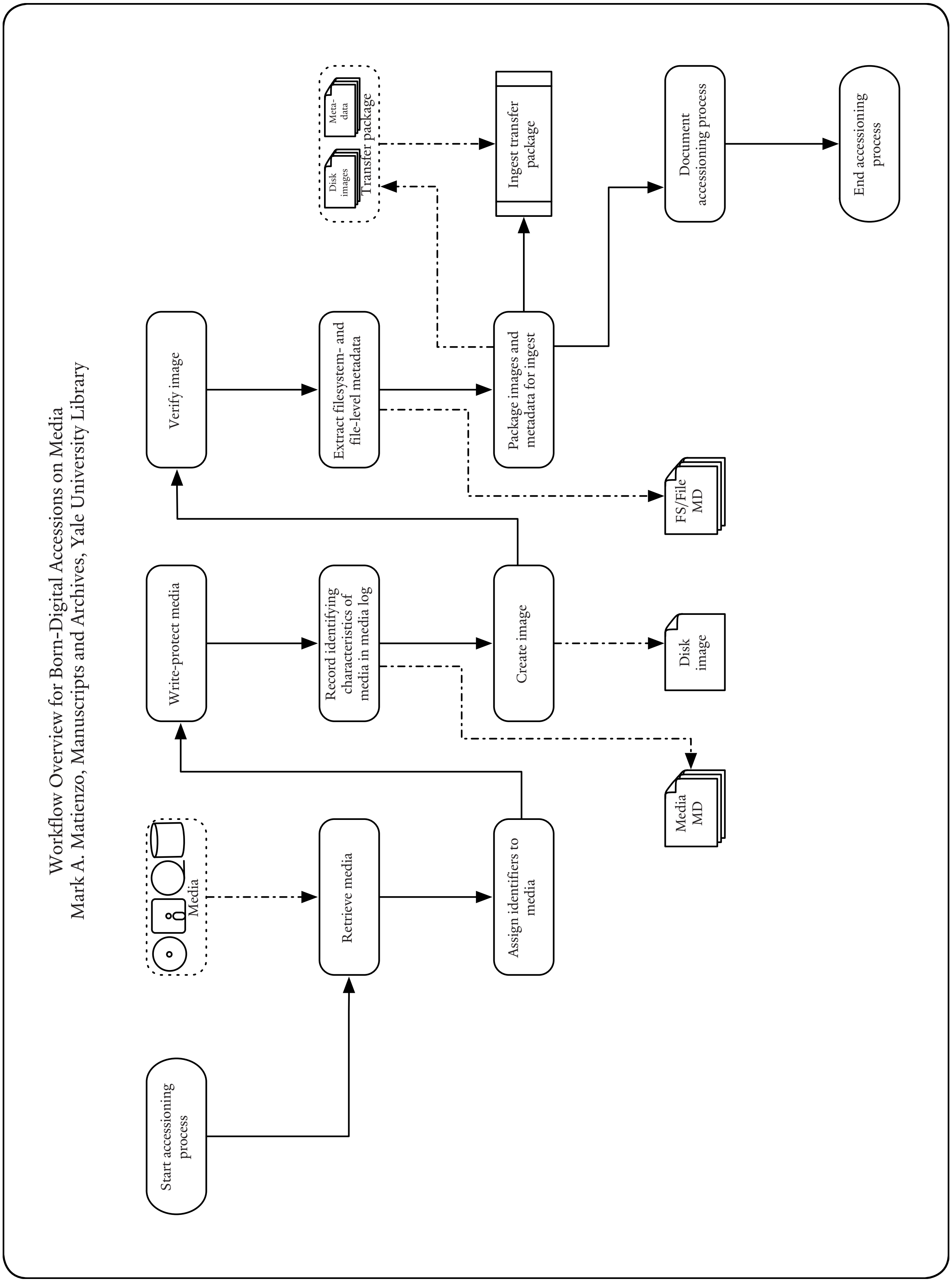
Draft, electronic 1990 Nov

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
Media MD
Create image
Disk image
Verify image
Extract filesystem- and file-level metadata
FS/File MD
Package images and metadata for ingest
Disk images
Meta-data
Transfer package
Ingest transfer package
Document accessioning process
End accessioning process