

BUILDING THE PENNSYLVANIA DIGITAL LIBRARY: Guidelines for Topics, Materials, Priorities, Best Practices, Next Steps

1. Introduction

In 2006, the Collection Development Working Group of the Pennsylvania Advisory Committee for Collaborative Digitization (PACCD) began to think about developing a comprehensive, distributed statewide digital collection of materials about Pennsylvania's history, culture, and society. PACCD is a statewide coalition of Pennsylvania-based organizations whose leaders wish to develop a collective, communal approach to the digitization of and access to historically important resources housed in collecting institutions across the Commonwealth. Discussion among members of the PAACD Working Group resulted in the submission of a request to the Office of Commonwealth Libraries for grant support to develop a set of guidelines for the subject matter and content of the digital collection.

Funded for 2007/2008 by monies available through the Library Services and Technology Act (LSTA), the Working Group convened a panel of scholars, teachers, librarians, and archivists who are expert in the history and resources of the Commonwealth. The panel was charged under the grant to develop guidelines for selecting subjects and types of material in order to help libraries, museums, and other cultural heritage organizations, together with the funders that support them, set priorities for digitizing their collections.

The guidelines developed by the panel explain how and why the digitized material will be used by a variety of audiences. They also list desiderata for digitization. Additionally, the guidelines provide a starting point for gathering information on the technical processes and recommended best practices for digitization. It is hoped that following these guidelines' recommended practices and specifications will result in high-quality digital resources to be used by the many communities concerned with Pennsylvania history, culture, and society.

The Working Group hopes to hear during a public review period from all interested parties about the guidelines, recognizing that not everyone will agree with the priorities our group has suggested. In order to keep these guidelines current and flexible and ensure that they meet the changing needs of the range of collecting institutions, we propose that the recommendations made in these guidelines remain in place for three years and then be reviewed and revised.

2. Audiences for Digitized Resources

The Pennsylvania Digital Library content will serve the following audiences:

Middle and high school students and teachers / National History Day. Students will be able to use an online environment that is enticing to them--and teachers will be able to create lesson plans and class assignments--using sources that are unique or rare and cannot be accessed except at the institution holding them. In so doing, they will meet Pennsylvania Standards for history that call for students to develop the skills of explaining,

analyzing, interpreting, synthesizing, and evaluating the many kinds of historical sources. This is especially true in the case of activities relating to National History Day, a program that reaches nearly one million middle and high school students nationwide and engages them in historical research using primary sources.

Local and family historians. Nationwide, about 35 million people use the Internet for family history research, and Pennsylvania libraries and archives report significant use of their collections for this purpose. Increased digitization of historical materials will aid a growing audience interested in connecting with personal and community histories. Since in many cases only portions of a given collection will likely be on line, those portions will serve as a hook to bring researchers to the libraries and archives that host the physical materials.

College students. Pennsylvania ranks fifth in the nation in higher education enrollment and third in the number of higher education institutions. The same motives that animate the Pennsylvania Standards also prompt college and university History programs to ask their students to use archival and other primary sources more innovatively. Having access to a large body of documents and other historical sources online will encourage college students to be more ambitious in undertaking their class projects and will provide them with a wider array of resources than may be found on their own campus.

Scholars and researchers. This audience, numbering in thousands, uses Pennsylvania's historical collections and extends awareness of them by teaching, publishing, and giving presentations. A large array of digitized Pennsylvania materials available on the Web will call attention to the key role Pennsylvania has played in many histories, histories which, in turn, will allow Pennsylvanians to gain a richer sense of the Commonwealth's heritage. Like other user groups, scholars and researchers are also likely to want to travel to Pennsylvania's libraries and archives to view the materials in their original format.

Heritage tourism development. Heritage tourism has proven itself to be a significant economic driver in Pennsylvania. Digitized Pennsylvania materials serve the heritage tourism sector in two ways: by providing the raw material from which tourism professionals develop public programming and by providing history travelers, who prefer to research destinations and topics as part of their trip planning, with additional reasons to visit or to extend their stays.

Funders. In these PACCD guidelines, subject areas and materials are identified, prioritized, and linked to the needs of Pennsylvania's diverse user communities. With this information at their disposal, both funding agencies and libraries will be better able to plan and set digitization priorities. Having these guidelines will also enhance the possibilities for inter-institutional collaboration as cultural collections repositories work together to assemble critical masses of materials. Without precluding the development or funding of digitization projects which do not fit these guidelines, the guidelines will help to focus the attention of funders and libraries on those materials that would provide the most utility for the groups listed above.

3. Topics and Material Types

With today's technology, a wide variety of materials can be digitized for wider access. We on the Working Group, as well as our panel of expert colleagues and others interviewed about primary historical sources, highlighted many types of materials that they, their colleagues, researchers, family historians, and students use in their work on Pennsylvania's history and current affairs.

We propose a tiered approach to digitization of materials types and topics/periods. In the first tier are four "priority" topics. We advise that collecting institutions and funders give top priority to them, and we prioritize them in order to help focus the community on creating a critical mass of materials. Within material types, we recommend that priority be given to materials with high research value that are also at risk of significant degradation. Recognizing that a much wider range of materials is desirable for digitization, we also suggest "general" topics and materials for second priority consideration.

We encourage institutions and funders to develop partnerships and cooperative projects that will devote resources to these priority materials and topics. Our listing these topics by no means should be understood as discouraging institutions from digitizing other materials based on the strength of their collection in a particular area, a local need, or an upcoming event, such as an anniversary which may focus widespread public interest on a topic. Ideally, these materials would be related to the four themes to which we have given priority, but when an institution finds that it does not have significant holdings related to these four themes, it should meet institutional and constituency needs as seems best.

Topics

We assume that interest in digitizing material will consider its historical significance. All history, like politics, is local, and we recommend that collecting institutions think first about family, neighborhood, municipal, and county-level materials that illustrate themes, periods, or events of national and world interest. Collecting institutions should look for local documentation that will contribute to the study of the broader issues and events we have listed below. We also encourage the inclusion of state birth and death records under access guidelines similar to those followed by states that have chosen to make these records available online.

Pennsylvania was and is remarkably rich in agricultural, mineral, and energy resources, which have drawn successive immigrant populations to pursue occupations in these fields, as well as a range of manufacturing industries in the nineteenth and twentieth centuries. The people, industries, resources, and related transportation and communication systems have both shaped and been shaped by the great conflicts of those centuries, notably the American Civil War, World War I, and World War II. These key subjects in the history of the Commonwealth of Pennsylvania can provide users with both a chronological view of the development of the state and a glimpse into how Pennsylvanians have dealt with progress and change. The interplay of these forces is vital to the understanding of our shared history.

Priority Topics

We recommend, therefore, that among the six topics and periods listed as "General Topics of Interest" immediately below, collecting institutions and funders focus first on the following four aspects or subsets in order to build a critical mass of material:

- The environment and land use
- Population groups from 1850 to the present
- Transportation and industry
- Major nineteenth- and twentieth-century conflicts: American Civil War and World Wars I and II

General Topics of Interest

Having addressed these four priority (sub)topics, collecting institutions and funders should concentrate second on these six more general categories:

1. The Environment: Pennsylvania's geology, flora, fauna, and land use from prehistory to the present.
2. The Economy: from the 1600s to the present, important economic and labor trends including agriculture, commerce, and manufacturing that have informed the growth of the Commonwealth.
3. Politics and Government: from the first elections held in the State to current local, state, and national elections, materials that provide insight into how Pennsylvanians govern themselves and into Pennsylvania's relationships with other states and nations.
4. Society: race, gender, Native American populations, religion and social justice, slavery and the underground railroad and their place in determining the course of Pennsylvania history, along with the effect of battles and wars on the people of Pennsylvania from the earliest settlers through the Revolutionary War, the Civil War, the two World Wars, and the many other conflicts of the nineteenth and twentieth centuries.
5. Technology, Science, and Medicine: especially from 1850 on, technology, innovation, and invention in Pennsylvania and its importance to the nation and the world.
6. Culture: from built spaces to fine arts to pop culture, especially in the twentieth century, Pennsylvanians as cultural trendsetters.

Material Types

In our discussions and interviews we heard about the many kinds of library and archival materials that can be digitized to meet user needs -- published works of local history; maps (especially insurance maps); print and photographic collections; postcard collections; newspapers; diaries, letters, and other personal/family documents; biographical information; business records; arts magazines and architectural sourcebooks; materials in library vertical files including primary resources and press kits; university histories and course catalogs; tax records, wills, death records, court records; city directories and telephone books; records of churches and other local organizations and groups; naturalization records; lists of almshouse residents; Pennsylvania German materials; deeds; Pennsylvania Archives; and census records.

We recommend that significant attention be given to historical sound and moving image documents, most of which will be from the twentieth century. Documentarians, scholars, and students all know the value of being able to locate, use, and incorporate into their work evidence of the sort provided by radio and television programs about local issues and history, industrial films, speeches by politicians, recordings of the dedications of buildings, etc. That historical sound and moving image materials are fragile and are decaying rapidly is amply documented; more worrisome still, most such materials are not even collected at this point into repositories, let alone preserved for future access.

While the urgency attached to the collection and preservation of audiovisual materials is very real because of the threat to their survival, we recognize that digitizing them poses some formidable barriers to entry in terms of expertise, equipment, and cost. Copyright is another issue for many twentieth-century materials, and, once digitized, these materials require major investments in infrastructure for storage, preservation, and ongoing access.

In cases where digitization may not be practical at this time, we urge Pennsylvania libraries to start collecting these materials from the private, commercial, and governmental sources that hold them and to survey and prioritize them as an important first step in a statewide effort to preserve this threatened part of our heritage. Since few institutions have the staff expertise and equipment to digitize these materials in-house, we hope, moreover, that our recommendation of this priority helps to leverage the attention of all parties toward taking a collaborative approach to solving the technical problems of file creation, storage, and ongoing access.

We also suggest that prioritization take into account the funding programs already in place from public and commercial organizations for digitization of such materials as newspapers, diaries, printed works, and census records. We recommend, therefore, that, after they consider digitization of historical sound and moving image materials, collecting institutions focus on kinds of primary sources that are outside the purview of current initiatives for newspapers, diaries, printed works, and census records. Leading examples of a material type of particular interest to many of the audiences in Pennsylvania are vital records, including birth and death records and land records. Preserving the access to these important materials, which were initially published in paper form, is an important goal for digitizing the primary sources on which historical research is based.

With a plan that focuses on high-risk formats and four important topics in the growth of the Commonwealth, Pennsylvania cultural heritage institutions can develop a digital presence that truly represents the history and important contributions of the state.

4. Best Practices and Technical Specifications for Digitization

Libraries, museums, and cultural heritage institutions planning to digitize their collections should follow the widely recommended best practices and technical specifications developed by such organizations as the Office of Commonwealth Libraries, Open Archives Initiative (OAI), Digital Library Federation (DLF), Library of Congress, National Information Standards Organization, and others. By following these standards and practices, institutions will create high-quality images and metadata, which will make their digital resources easily searchable and accessible to the greatest possible audience.

Getting started

The PACCD and its Metadata Working Group have created guidelines for the digitization of collections. These guidelines are available via the Access Pennsylvania Digital Repository site at http://www.accesspadigital.org/pdf/access_pa_dig_rep_guidelines.pdf. The PACCD guidelines are designed to improve the management and preservation of the Commonwealth's digital resources, as well as to ensure preservation of the physical objects upon which the digital resources are based.

The PACCD guidelines provide several sources for information on optimum scanning and digitization practices for a variety of material formats—audio, video, text, etc.—recommended by such organizations as the DLF (<http://www.diglib.org/>) and the Collaborative Digitization Program of the Bibliographic Center for Research (BCR) (<http://www.cdpheritage.org>).

Audiovisual resources

In the guidelines above, we designate historical sound and moving image resources as material types for priority collection, preservation, and digitization. While all digitization projects pose challenges, the digitization of audiovisual resources offers formidable barriers to success for cultural heritage organizations.

Detailed and reliable information on audio digitization can be found in a number of sources. These include *Digital Audio Best Practices* (<http://www.bcr.org/cdp/best/digital-audio-bp.pdf>), prepared by the Colorado Digitization Program (CDP) Working Group, and *Recommended Minimum Standards for Preservation Sampling of Moving Image Objects* (http://rucore.libraries.rutgers.edu/collab/ref/dos_avwg_video_obj_standard.pdf), published by RUCORE, the Rutgers (University) Community Repository.

Standardized metadata and access

PACCD guidelines recommend the use of Dublin Core metadata to catalog and describe digital resources and provide detailed specifications on the proper treatment of descriptive fields. Using these guidelines, institutions have created the digital resources that constitute the Access Pennsylvania Digital Repository (<http://www.accesspadr.org>), an effort to create a unified approach to displaying digitized collections using the Office of Commonwealth Libraries' statewide license for the CONTENTdm software. This CONTENTdm implementation is a physical way to make digital resources available, allowing for both storage of resources and their retrieval by scholars.

Collections included in the AccessPA digital repository also meet the standards for interoperability established by the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), details on which are available at <http://www.openarchives.org>. Thus, as part of their adherence to best practices and standards, libraries, museums, and other cultural heritage organizations in Pennsylvania should follow these protocols as well in their digitization efforts.

Pennsylvania Digital Library Record Repository

As a result of following these protocols, records for digital collections about Pennsylvania will be harvestable by and searchable through the Pennsylvania Digital Library (PADL, <http://padl.pitt.edu>). Developed by the University of Pittsburgh's Library System, in cooperation with the PACCD, PADL is a statewide metadata repository for non-subscription, non-commercial digital resources created by Pennsylvania libraries, museums, and other cultural heritage organizations.

Rather than serving as a repository for digital collections themselves, the PADL harvests, indexes, and provides a common search interface for metadata created for Pennsylvania digital collections. Thus, collections reside at their home institutions (or the Access Pennsylvania Digital Repository) while metadata adhering to OAI-PMH standards is gathered regularly and then made searchable through the PADL.

Institutions that want to make their collections searchable through the PADL and whose metadata meet OAI-PMH standards should register with the PADL (<http://padl.pitt.edu/index.php/register>) for harvesting of their metadata. Institutions that are part of the Access Pennsylvania Digital Repository do not need to register with the PADL as their collections' metadata is automatically harvested.

Copyright

Copyright expert Peter B. Hirtle writes in *Digital Preservation and Copyright* (http://fairuse.stanford.edu/commentary_and_analysis/2003_11_hirtle.html) that

Digital preservation and access is all about copying. In copyright law, copying is known as "reproduction," and it's one of the exclusive rights of the copyright owner. The right to publicly display a work is also an exclusive right of the copyright owner, as is the right to make an adaptation, known as a "derivative

work." Our desire to keep digital information around for the future runs smack into the exclusive rights of the copyright owner.

Thus, the process of digitization—transforming a work or object from one form, chiefly analog, into an electronic or digital format—is a form of reproduction. Therefore, national and international copyright laws and institutional copyright policies must be considered when digitizing collections and resources with the aim of sharing them with a larger world.

Many historical materials, particularly those created before the twentieth century, may be in the public domain, outside of copyright, and suitable for digitization. However, copyright may still be held for more recent materials, especially audiovisual resources.

As organizations plan and carry out digitization efforts, resources such as the OCLC Digitization & Preservation Online Resource Center's *Copyright Online Resource Kit* and the U.S. government-sponsored Section 108 Study Group's "Background Papers" website (<http://www.section108.gov/papers.html>) can assist with making informed decisions about copyright law and common practices.

Accessible design

Providing rich scholarly content to a wide audience is one of the main goals of digitization. However, that content may be unusable to many scholars unless content creators pay attention to technical compatibility, ease of use, and issues in accessible design. For those organizations that design their own websites and portals to provide gateways to digital collections, it is important to follow good practices in website design. Designs that are clean, uncluttered, easy to search, and compatible with different web browsers are preferred as they make collections accessible to the greatest number of researchers working at all levels of interest. Another important consideration for portal design is accessibility to persons with different abilities and disabilities, including visual, auditory, intellectual/cognitive, and mobility impairment.

The Worldwide Web Consortium (W3C, <http://www.w3.org/>) is the main international standards organization for the worldwide web. One project of the W3C is the Web Accessibility Initiative (WAI, <http://www.w3.org/WAI/>), which creates Web Content Accessibility Guidelines (WCAG) in order to make webpages and web technologies accessible to all. Additionally, Section 508 of the U.S. Rehabilitation Act (29 U.S.C. 794d) (<http://www.section508.gov>) requires U.S. government agencies to give disabled employees and members of the public access to information that is comparable to the access available to others. While there are currently no requirements for accessible website or portal design under the Americans with Disabilities Act (ADA), design that is accessible to as many scholars as possible should guide website development.

Equipment, training, and funding

Although the rewards to scholarship and cultural heritage can be immense, digitizing collections is no small undertaking. In addition to standards and guidelines for digitizing

and cataloging collections, equipment must be purchased, staff trained, and funding secured to pay for the work to be done.

Through its digitization and preservation program (http://www.palinet.org/digitization_preservation_services.aspx), PALINET, the Mid-Atlantic regional library network, provides opportunities for training and continuing education, information on vendors who sell digitization equipment or offer digitization services, and possible sources for funding. Other organizations, such as the Pennsylvania Historical and Museum Commission (PHMC), offer similar direction and information for museums, historical societies, and cultural heritage organizations.

National, regional, and state professional associations for librarians, archivists, and museum specialists provide opportunities for networking and learning from colleagues who have developed digital collections or operate digitization programs. Nationally, these include

- American Association for State and Local History (AASLH) (<http://www.aaslh.org>)
- American Library Association (ALA) (<http://www.ala.org>)
- Association of American Museums (AAM) (<http://www.aam-us.org/>)
- Association of College and Research Libraries (ACRL) (<http://www.acrl.org>)
- Digital Library Federation (DLF) (<http://www.diglib.org/>)
- Northeast Document Conservation Center (<http://www.nedcc.org/home.php>)
- Society of American Archivists (SAA) (<http://www.archivists.org/>)

Regionally and locally, these include

- Pennsylvania Academic Library Consortium, Inc. (PALCI) (<http://www.palci.org>)
- Pennsylvania Library Association (PaLA) (<http://www.palibraries.org>)
- Pennsylvania Federation of Museums and Historical Organizations (PFMHO) (<http://pamuseums.org>)
- Philadelphia Area Consortium of Special Collections Libraries (PACSCL) (<http://www.pacscl.org/>)
- Mid-Atlantic Regional Archives Conference (MARAC) (<http://www.lib.umd.edu/MARAC/>)

5. Looking Toward the Future

In thinking about the many aspects of collection access and especially about the environment for digitization, PACCD sees several related steps that might be taken to promote cooperation in the development of a digital collection about Pennsylvania. Many workflows, organizations, funding streams, and kinds of expertise come into play in collection digitization, and the PACCD represents a strong initial effort at interconsortial cooperation to build the technical, collection development, and administrative network a successful, coherent digital collection for Pennsylvania will require.

We therefore want to list here the many dimensions of and possible next steps for the overall digitization project in Pennsylvania.

1. As noted above, the panel recommends that the guidelines document and its priorities be reviewed every three years in order 1) to keep it fresh by aligning the topics and material types with current interests, and, perhaps more important, 2) to create and maintain an ongoing community of interest in the digital collection about Pennsylvania. The Bureau of Library Development might look to the PACCD for this task. In the immediate future, we or members of the panel of experts can take this document to professional association meetings in order to seek comment on the guidelines and, more generally, discuss the digitization of Pennsylvania materials. Since it is frequently the case that pockets of activity will develop and remain isolated from each other, the PACCD hopes that the website being developed by PALINET and partners can become a community-building hub for collecting information about digitization activities in the Commonwealth.

2. In many cases, statewide digitization projects have employed a survey to assess the amount and kind of material available about the state. In 1998 Pennsylvania participated in a study funded by NHPRC that inventoried document repositories through what is now the Council of State Archivists (CoSA). The report is posted on the CoSA web site at <http://www.statearchivists.org/reports/index.htm>. Since that inventory was taken, collecting institutions and consortia have created many digital collections and are planning more. Possible parameters for a survey could be questions about:

- *Digital projects already in existence*
- *Cataloged but undigitized collections*
- *Cataloged collections the holder has no plans to digitize*
- *Uncataloged collections that are worth digitizing*

3. Given the wonderful incentive offered by the HSLC/AccessPA CONTENTdm implementation, relevant parties should take the opportunity to develop the user community for it by offering workshops on digitization and by encouraging the CONTENTdm user group being formed under the auspices of PALINET.

4. Relevant parties should promote participation in and then work with the results of the Connecting to Collections grant to the Conservation Center for Art and Historic Artifacts from the Institute of Museum & Library Services (<http://www.ims.gov/collections/>); the grant will provide data that could serve as the basis for digitization of materials for reasons of preservation as well as access.

5. Libraries in Pennsylvania should take the opportunity offered by PALINET's Sloan Foundation mass-digitization grant to digitize printed materials about Pennsylvania.

6. Pennsylvania should build on the findings of the current PACSCL hidden collections grant from The Andrew W. Mellon Foundation (<http://www.pacsclsurvey.org/>) to select materials for digitization. In this same vein, PACCD should encourage partnerships in

Pennsylvania to respond to the invitation from the Council on Library and Information Resources for hidden collections grant proposals in the next five years.

7. Pennsylvania should develop infrastructure that will help small institutions contribute materials for digitization through partnerships with larger institutions or through provision to them of expertise and digitization services by a central agency.

8. PACCD, in partnership with Commonwealth agencies, groups representing collecting institutions, and representatives from all levels of the education community, might hold a statewide invitational meeting on digitization based on the guidelines document. Such a meeting should take care to bring small institutions into this discussion.

9. Given the well-documented challenges to collecting, digitizing, serving, storing, and preserving historic sound and moving image materials, PACCD, in partnership with Commonwealth agencies and groups representing collecting institutions, might convene a statewide meeting to discuss how to address this priority of the collection guidelines.

10. PACCD and Commonwealth agencies through their grants should encourage organizations to make records for digital collections harvestable into an OAI registry.

11. Pennsylvania collecting institutions should be encouraged to put their finding aids on the Web so that researchers can explore collections more easily and thereby suggest priorities for digitization.

In this document, the PACCD has laid out not only a set of guidelines for topics and materials to be digitized but a mechanism for their ongoing review and revision; we have listed considerations for and links to information about technical and planning matters and have identified related initiatives and next steps. The members of PACCD are hopeful that, once these guidelines have been through the review and comment process, they will provide a solid foundation on which Pennsylvania can build a digital collection and a reputation for leadership in statewide collaboration.

(Revised May 2008)