Getting the Picture: Observations from the Library of Congress on Providing Online Access to Pictorial Images*

CAROLINE R. ARMS

ABSTRACT

OVER THE LAST FEW YEARS, THE LIBRARY OF CONGRESS (LC) has increasingly created digital reproductions of visual materials to enhance access to its resources. Digitization is now a mainstream activity in the Prints and Photographs Division (P & P) and the Geography and Maps Division (G & M). Both divisions work closely with the National Digital Library Program to make their incomparable resources accessible over the Internet to the general public through the American Memory Web site (http:// memory.loc.gov/). They also use the digital images to serve their more traditional clientele in the reading rooms. Retrieval from a collection of digital images offers special opportunities to apply new technological advances, as illustrated elsewhere in this issue. However, retrieval often takes place in broader contexts. The Print and Photographs Division seeks to enhance access to its international pictorial holdings, whether digitized or not. Within American Memory, the focus is on retrieval by the nonspecialist from a body of materials related to the history and culture of the United States, materials heterogeneous in both original and digital form. A yet broader context is retrieval from the comprehensive collections of the entire Library of Congress. Beyond enabling retrieval, LC is concerned with facilitating use of the materials retrieved, consistent with any associated rights. This article describes selected aspects of LC's practical experience and current practices from digital capture through interactions with users, with an emphasis on the integration of access to pictorial images online with other services and activities at LC.

*This article is exempt from U.S. Copyright.

Caroline R. Arms, Library of Congress, 101 Independence Avenue S. E., Washington, DC 20540-9300

LIBRARY TRENDS, Vol. 48, No. 2, Fall 1999, pp. 379-409

Considering the Challenges of Access and Retrieval

Choices made at the Library of Congress in relation to access, retrieval, and use of its pictorial materials reflect many requirements and desires:

- to serve different audiences from expert researchers to the K-12 and higher educational communities and the lifelong learner;
- to support a variety of uses as appropriate, including citation (in print and online as active hyperlinks), study and comparison, convenient reproduction for classroom or personal use, and high-quality reproduction for publication;
- to facilitate access to digital pictorial resources in conjunction with access to related materials in all forms;
- to find a balance between demand by users for ever more detailed description for resources currently accessible and for access to more of its collections;
- to allow digitization to serve a future role in long-term preservation of materials originally created in many forms;
- to find practical solutions in the absence of well-established standards and contribute to the informed development of standards where necessary; and
- to build systems that can be deployed today with large quantities of images and to enhance services incrementally taking into account economic and organizational realities.

The pictorial collections of the Library of Congress present enormous challenges for both physical and intellectual access. The Prints and Photographs Division (P & P) holds over 13 million images, including photographs (published and unpublished), cartoons, posters, documentary and architectural drawings, and ephemera, such as baseball cards. Most of the images are photographs related to the United States, many acquired in large collections. Photographic prints may have been captioned (usually by writing on the physical artifact) or organized by the photographer or by the institution or individual from whom a collection was acquired. Many images, however, are held only as negatives, which pose special problems for identification, housing, and service and are not always accompanied by individual captions.

Cataloging pictorial items is labor-intensive. P & P estimates that it takes an hour to produce a brief record for an average item for inclusion in the Library of Congress' main catalog. This allows time to handle the item appropriately, note identification numbers, record basic information about the creator, date, physical artifact and reproduction rights, devise a descriptive caption where necessary, assign a few subject headings, and proofread. One hour stretches to three or four if an attempt is made to verify the information accompanying the piece, to describe what a picture is about rather than simply what it is of (consider a political cartoon or a photograph of a notable event), or to provide added contextual information, such as where a picture was subsequently published or biographical notes on the subject of a portrait. This degree of effort can only be justified for a small portion of the P & P holdings—e.g., for fine prints and posters.

Before machine-readable cataloging, access to the Prints and Photographs Division resources was primarily through a card catalog with entries for groups of material and through "browsing" files organized in thematic hierarchies but with no separate item-level description or control. Because of the size of the collections, many items are still only accessible this way. Since 1989, a priority for the Library of Congress has been reducing the backlog of items waiting to be included in public access systems; efforts in P & P since then have focused on physical organization and cataloging for materials that were previously unprocessed or in high demand.

Visual approaches to browsing pictures offer an alternative to detailed cataloging of individual items. Once a picture is retrieved, however, most users need information about the picture in order to cite it, confirm its applicability as evidence or illustration, or determine whether permission is needed to reproduce it. At the very least, the user wants to be able to find a particular picture again (preferably directly rather than by browsing), request a reproduction, or seek permission to reproduce it. In the Prints and Photographs Division reading room, a variety of clues and experts are available to allow identification. Item-level control may be deferred until demand for the particular item is demonstrated. When a user requests a reproduction of an item that has not previously been cataloged, it is assigned a reproduction number and an item-level record is created. Digitization accelerates the need to apply identifiers to the physical items, both for tracking during conversion and quality review and to relate the digital copy to its physical source.

The Prints and Photographs Division regards its digital reproductions, even the high-resolution images being prepared under the current conversion contracts, as surrogates. Based on his experience as chief of the Prints and Photographs Division, Ostrow (1998) reviewed the nature of large historical pictorial collections and how they are traditionally used in reading rooms. He emphasizes the constructive role digital surrogates can play for researchers and in cutting down the need to handle fragile originals. He also discussed shortcomings of digital images as replacements when used for historical documentation. The Library of Congress has not yet used digital reproduction to replace physical originals, even when the originals will soon be unusable. To preserve the information for the longer term, brittle books are currently microfilmed and deteriorating negatives are replaced by high quality photographic copies.

For many uses and users, however, the digital surrogates suffice. Digitization has furthered the objectives of the staff in the Prints and Photographs Division to serve patrons in the reading room better and of the National Digital Library Program (NDLP) to make resources available to a much broader public beyond LC's walls. The remote audiences, however, present new expectations and a wider range of tasks for which pictures are needed; they also lack access to expert assistance. The challenges of serving many audiences and supporting retrieval in many contexts will continue. LC's current technical architecture is based on a modular framework that allows different interfaces to take advantage of the same catalog records and the same digital content. The same digitized picture can be accessed through LC's comprehensive catalog, through American Memory, or through a catalog that is tailored to pictorial resources.

Looking Back

Released for public access over the Internet in early 1998, the Prints and Photographs Online Catalog (PPOC) (http://lcweb.loc.gov/rr/print/catalog.html) is the most recent interface to an increasingly comprehensive catalog to the division's holdings. Where available, records in PPOC are accompanied by digital images. This catalog builds on work which started in 1982 when the division began reproducing selected collections electronically (initially on videodisc) and cataloging the images for LC's Optical Disk Pilot Program, described by Elisabeth Betz Parker (1985). In December 1993, a dedicated workstation with an array of videodisc players and a separate monitor for displaying images was introduced as a public service in the reading room and dubbed the "One-Box." The "One" in One-Box represented the goal of the Prints and Photographs Division to develop a reference gateway that could provide access to all their holdings. In 1996, a digital version (known initially as the Digital One-Box and taking advantage of the capabilities of the World Wide Web) was introduced in the reading room. The Digital One-Box became the Prints and Photographs Online Catalog and was released on the Internet after incremental improvements based on experience with users. By December 1998, PPOC provided access to twenty-five collections covering over 5 million physical items. In the P & P reading room, PPOC provides access to 355,000 digitized images. Over 60 percent of these images are accessible through American Memory; others are out of scope, or access is restricted because of copyright or other reasons, such as privacy. American Memory and PPOC share digital image files and catalog records for the overlapping content and rely on much of the same program code.

The American Memory pilot project in the early 1990s explored the use of digital images on CD-ROM. The Prints and Photographs Division participated actively in the pilot and, in June 1994, the first release of

American Memory on the World Wide Web comprised three collections of photographs. By December 1998, thirteen collections from P & P, representing 220,000 original items, had been released on American Memory.

Two very large collections are being digitized and released for public access in phases. The first, known as Built in America (http://memory.loc.gov/ammem/hhhtml/hhhome.html), comprises photographs, architectural drawings, and "data" pages of typed textual documentation from the Historic American Buildings Survey and Historic American Engineering Record (HABS/HAER). As of March 1998, HABS/HAER documented 35,000 sites and structures through 363,000 negatives and paper artifacts. The other large collection is entitled America from the Great Depression to World War II (http://memory.loc.gov/ammem/fsowhome.html). It contains approximately 165,000 negatives and transparencies from the Farm Security Administration and the Office of War Information (FSA-OWI).

Today's Snapshot

The National Digital Library Program was established in 1995 as a five-year program. LC management is currently considering how best to build on NDLP's achievements and incorporate digital content more extensively into its collections. Figure 1 is based on a diagram developed to describe the component activities and related systems that provide the infrastructure for producing or incorporating digital content into LC's collections and providing coherent access to those resources. The diagram is based on the experience of staff in the NDLP, in the custodial divisions whose content NDLP has helped digitize and provide access to, and in LC's central technology service organization, which has a small group of programmers building and maintaining the computer applications that support both American Memory and the Prints and Photographs Online Catalog. The framework in this diagram will be used to organize the observations and experiences described in this article.

Making Digital Reproductions

The Prints and Photographs Division has chosen to use expert contractors to prepare the digital reproductions of pictorial materials. The use of contractors allows the Library of Congress to take advantage of special equipment without the need to build in-house facilities to handle a wide variety of physical formats. Contractors are also better able to keep up with the latest technological improvements in hardware and develop specialized software that applies the latest techniques for capturing and processing large quantities of images, since the investment required can be allocated across many projects and customers. In early 1998, a multiyear contract was awarded for the generation of digital images for pictorial materials after evaluating responses to a request for proposals (Library of

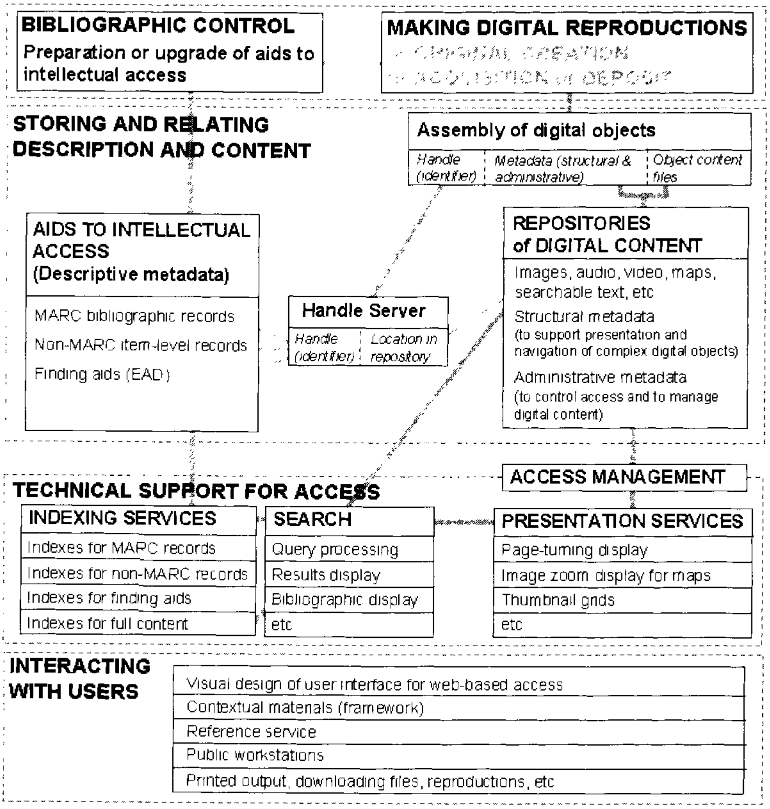
BIBLIOGRAPHIC CONTROL
Preparation or upgrade of aids to intellectual access
MAKING DIGITAL REPRODUCTIONS
or ORIGINAL CREATION
or ACQUISITION or DEPOSIT
STORING AND RELATING DESCRIPTION AND CONTENT
Assembly of digital objects
Handle (identifier)
Metadata (structural & administrative)
Object content files
AIDS TO INTELLECTUAL ACCESS (Descriptive metadata)
MARC bibliographic records
Non-MARC item-level records
Finding aids (EAD)
Handle Server
Handle (identifier)
Location in repository
REPOSITORIES of DIGITAL CONTENT
Images, audio, video, maps, searchable text, etc
Structural metadata
(to support presentation and navigation of complex digital objects)
Administrative metadata
(to control access and to manage digital content)
ACCESS MANAGEMENT
TECHNICAL SUPPORT FOR ACCESS
INDEXING SERVICES
Indexes for MARC records
Indexes for non-MARC records
Indexes for finding aids
Indexes for full content
SEARCH
Query processing
Results display
Bibliographic display
etc
PRESENTATION SERVICES
Page-turning display
Image zoom display for maps
Thumbnail grids
etc
INTERACTING WITH USERS
Visual design of user interface for web-based access
Contextual materials (framework)
Reference service
Public workstations
Printed output, downloading files, reproductions, etc

Figure 1. Infrastructure for Managing Digital Collections and Providing Access.

Congress, 1997, RFP97-9). The contract, awarded to JJT, Inc., covers a variety of original formats, including transmitted-light items (e.g., negatives and transparencies) and reflected-light items (e.g., photographic prints and baseball cards) but excluding oversize items such as architectural drawings. The RFP provides an excellent description of LC's objectives and the criteria considered important in digitizing pictorial materials. Some details, however, were modified during the contract startup and as production began. In line with the requirements to perform capture at LC, the contractor has established a scanning facility in a small room in the P & P division. On-site scanning allows items that cannot leave LC to be scanned directly rather than via photographic intermediates, and reduces manpower needs for shipping and tracking large quantities of material. Although LC used photographic intermediates for early projects, they inevitably introduce some degradation in quality as demonstrated in the RLG Technical Images Test Project (Reilly, 1995).

However much effort is devoted to describing specifications in a proposal or contract, LC has learned that other factors are important to the success of digitization projects. A cooperative working relationship with frequent communication is invaluable, as is the development of mutual trust. Early in this project, the contractor demonstrated a commitment to careful handling that allayed concerns of conservators. As indicated in a recent report from the Image Permanence Institute (Frey, 1998), operator expertise and visual sophistication are needed for successful digitization in an archival environment. After careful and productive experimentation with test batches of materials, LC is able to rely on the scanning contractor's technical judgment on many matters. The staff from the Prints and Photographs Division and the contractor's team have shared objectives, with ambitious goals for quality balanced by a need for productivity.

Quality of Digital Reproductions

The current Prints and Photographs Division practice is to scan most pictorial items at spatial resolutions of 3,000, 4,000, or 5,000 pixels on the long side. The choice depends on various factors, including the size and quality of the source and the visual content or intent of the work. For baseball cards (http://memory.loc.gov/ammem/bbhtml/bbhome.html), which are small, capture was at 3,000 pixels. The negatives in the HABS/HAER collection are being scanned at 5,000 pixels since they are intended to document architectural and engineering details and may be used in the future to support reconstruction or restoration. Scans from copy negatives are usually at 4,000 pixels since the quality of the copies does not warrant higher resolution. Anne Kenney of Cornell University and Lou Sharpe of Picture Elements, Inc. have used a conceptual structure for suggesting levels of resolution for capture of illustrations in nineteenth-century printed books in a study for LC's Preservation Directorate (Kenny et al., 1999). They consider the increasing resolutions needed to capture the essence of a picture for screen display, the detail of its visual content (such as a wisp of hair in a portrait), or the structure of the original artifact (for example, to distinguish different types of engraving). Although P & P does not use these categories explicitly, the aim is usually to capture the visual details rather than to reveal the artifactual structure. The equipment used by the contractor captures color images at 36 bits/pixel; any processing is at 48 bits/pixel to ensure that transformations do not introduce artifacts or lose detail. Due to limitations of current technology and standard formats, the tonal resolution is reduced to 24 bits/pixel for delivery to LC as an uncompressed TIFF image. Black and white photographs are captured at 12 bits/pixel grayscale, processed at 16 bits/pixel, and stored at 8 bits/pixel.

The file size for a color image at 5,000 pixels on the long side is around 50 Mb for the archival uncompressed TIFF. The archival grayscale HABS/

HAER images approach 20 Mb. Smaller derivative versions are created for convenient access, retrieval, and use, although the Prints and Photographs Division usually makes the archival versions of images not subject to copyright protection available for downloading. Current practice (which differs a little from the RFP) is to create three smaller versions. A thumbnail, 150 pixels on the long side for almost all items and at 8 bits/pixel, is designed for inclusion in item-level bibliographic displays and grids of thumbnails. This size delivers adequate performance over the Internet and supports rapid visual browsing on screen. Care is taken with the quality of thumbnails since a poor thumbnail may keep a user from looking at a relevant image.

For general use, two service versions are currently created. For convenient display on any screen and rapid downloading over the Internet, a JPEG image at 640 pixels on the long side is generated with moderate compression (usually at a reduction of around 15 to 1 for color and 8 or 10 to 1 for grayscale images). For the user who wishes more detail or a better image for printing or other re-use, a JPEG with lighter compression is created at 1,024 pixels on the long side. When creating the derivatives or during quality review, the contractor selectively applies techniques to reduce the moiré patterns that may be generated when images with regular patterns (such as siding on a house) are reduced in size. The technique most commonly applied is to blur the image at full size, reduce to the desired size, and then sharpen the smaller image.

The tonal quality of images is limited by the quality of the initial scan, which depends on the equipment, its calibration, the judgment of the scanning operator in using its capabilities, and environmental characteristics, such as dust and lighting. The scanning stations designed and installed by the contractor at the Library of Congress incorporate prototype MARC II digital cameras built by Udo and Reimar Lenz of Munich, Germany. The contractor has developed software to control settings, including the camera's height above the glass scanning surface which can be lit from below for transmitted-light materials and above for reflected-light items. No glass is placed directly on top of materials. Curtains and air filters around each station provide control over lighting and dust. Target images that allow objective measurement of scanner performance on spatial resolution and grayscale representation are scanned with each batch of material and stored.

For prints, the objective is to reproduce the tonality of items as they exist. The monitors on scanning and review stations are calibrated carefully, although it is recognized that no monitor can reproduce all tonal qualities of all originals in all viewing conditions. For negatives and original positive transparencies, the objective is to create a positive image using photographic sensitivities that might be expected of a skilled darkroom technician. For the documentary HABS/HAER images, the instructions are to balance tones and capture all the information, avoiding loss of detail in shadows or highlights. Histograms of grayscale tonal values are used as an aid but not to control the process. All manipulations performed on each item, whether during scanning, quality review, or when creating derivatives, are recorded in a database by the contractor.

Archival images from collections scanned recently have been downloaded and reproduced in newspapers. For the February/March 1999 issue of *Civilization* magazine, a publication with higher quality requirements, archival digital images of baseball cards and some photographs were used. Some Prints and Photographs Division staff members argue that the high quality images being created currently could potentially serve as preservation masters in the future. LC has found it expensive to make photographic copies of deteriorating negatives. In addition, findings by the Image Permanence Institute suggest that scans from second-generation photographic materials are noticeably inferior to those from originals (Frey, 1998). The ability of the digital images currently being created from negatives to serve all the purposes served by photographic copies will be evaluated, and the practice of making photographic copies may be discontinued.

Workflow and Productivity

High throughput for scanning requires attention to workflow and modification of systems to avoid bottlenecks. The contractor has achieved an average rate of 375-400 negatives scanned in an eleven-hour day on a single station through a number of noteworthy innovations. One innovation is to plan and test transformations (including cropping and re-alignment) and exact steps for creating derivatives on small images (approximately 1,000 pixels on the long side) and then apply them to the full-size images automatically. A batch of these small images is sent on magnetic tape each night to the company headquarters in Texas, where highly skilled staff review the day's work and plan the transformations, which are recorded for automatic retrieval by the contractor's staff at LC. Another innovation relates to the size of individual files. Initially, operators were waiting while the large image file from one scan was stored before the preview of the next item could be displayed. The contractor modified the computer system to allow these tasks to run in parallel.

A limiting factor on throughput is now the capacity of the dual Pentium II workstations, which will be upgraded by the contractor as faster processors become available and can be tested. Another bottleneck is network bandwidth for loading the archival image files to the Library of Congress' servers. Rather than degrade network and server performance for users, image files are loaded directly to LC's storage system from CD-ROMs created by the contractor. Workstation capacity and network bandwidth have also been limiting factors for the Geography and Map Division (G & M) where large format color maps are scanned by LC staff at 300 dpi and 24 bits/pixel, sometimes creating individual files of over 300 Mb. After each image is reviewed and cropped, a compressed version is created using wavelet compression. This post-processing generates a heavy computing load. Using 166 MHz Pentium stations with 32 Mb of RAM, the production rate was roughly six maps per day. New dual Pentium-II processors with 500 Mb of RAM have generated a sevenfold increase in throughput. In G & M, catalogers can still keep pace with the scanning operation. In the Prints and Photographs Division, however, the scanning capacity far exceeds the capacity for item-level cataloging.

Preparing Aids to Intellectual Access

Describing pictorial materials accurately is time-consuming and expensive. Unlike a book, which usually has a title page on which basic information is recorded, an image does not describe itself. Words are needed to indicate the place or event represented in a photograph, its creator, the names of people portrayed, and when it was taken. Providing effective access to large collections poses a challenge; unfortunately, many of the solutions for access in physical archives do not transfer as readily to the online environment as individual descriptions for each item, particularly when the aim is to provide coherent access to resources of all types.

The Prints and Photographs Division has made extensive use of collection-level and group-level records for cataloging its holdings. If a group of related items is housed in a single container, a single catalog record describing the contents will allow the user in the reading room to request the container and browse through its contents visually. Larger collections have sometimes been described both by collection- and group-level catalog records that point to a paper-finding aid that offers a structured hierarchical listing of the collection. Group-level catalog records for components of a large collection allow integrated general access from within LC's main catalog or the Prints and Photographs Online Catalog and identify the physical location and any paper-finding aid or index that offers more specific access. The finding aid provides a convenient mechanism for exploring the entire collection, once it has been identified as relevant, through the logical organization selected for its physical storage or any alternative indexes provided. As mentioned earlier, many items, particularly those acquired before automation, are accessible only through "self-indexing" filing schemes. For example, P & P has files for each president with subcategories such as cartoons, homes and haunts, and family. Biographical files hold portraits of many individuals, famous and not so famous. An advantage of such an approach is that the labor of preparing individual records for each picture is avoided. A disadvantage is that access is constrained by a single logical arrangement unless copies are made.

The physical filing scheme might seem straightforward to reproduce in a digital environment for visual browsing. However, browsing through a physical folder full of assorted pictures at a table in the reading room is not the same as browsing online. With the physical item in hand, the user will naturally scan the picture, turn it over to look for a caption, and make inferences from the physical nature of the item. In an online environment, some physical clues (such as size) are obscured and skimming through full-size images is awkward. Moreover, users have different expectations; online they expect captions to be legible and presented consistently. Even if a physical item has no individual identification, the user can ask a librarian about it by pointing to it. In an online environment to support remote users, the explicit recording of minimal identification for an image is essential if any appropriate use or reference is to be made. The challenge is to find lightweight approaches to description and organization that support both convenient visual browsing and also search-based retrieval on whatever descriptive information is available.

To date, items digitized from the Prints and Photographs Division's collections have mainly been described in item-level records, mostly in the MARC format, but with widely different depths of detail in the description. The first use of group-level records, described more fully below, has been made for the HABS/HAER collections. The records describe the intellectual expression and the original form of the material and provide a link to the corresponding digital reproductions. Information about the digital files is not recorded in the MARC bibliographic records since these are considered surrogates for reference purposes rather than separate works. One pragmatic reason for using the item-level approach initially is that it was easy to adapt existing search and retrieval tools and use the same records for PPOC, for American Memory, and for the main LC catalog. Another is that staff were confident that they knew how to design and build a first system around item-level records. Recent projects, at the Library of Congress and elsewhere, have begun to show how finding aids and group description can be used to advantage for providing access to pictorial materials.

In an ideal world, browsing and searching would be supported by item-level descriptions of uniform quality based on perfect information about time, place, and circumstances of creation, using descriptive terms from controlled vocabularies and following common practices for assigning subject terms. Rather than fixed browsing frameworks, groupings could be derived dynamically from the descriptive terms. Computing specialists designing systems often start by assuming that this ideal is easily achievable; to Prints and Photographs Division staff, who are responsible for large archival collections of pictorial materials, it is clearly not. They are nevertheless leaders in promulgating standards of practice since the ideal characteristics serve as goals and some are more achievable than others.

The division is responsible for a manual that supplements the Anglo-American Cataloguing Rules with details appropriate for graphic materials (Betz, 1982), which is included in the *Cataloger's Desktop* CD-ROM (Library of Congress, 1996), and maintains the *Thesaurus for Graphic Materials* (Library of Congress, 1995). Betz (1982) indicates that an individual image often derives its importance from the collection of which it is part, and that full cataloging may not be feasible for all works. For each collection or project within P & P, plans specify the level of granularity (e.g., item-level or group-level) and detail to be employed for cataloging. Factors considered include the research value of the material, its uniqueness, demand (past and potential), practicality, available resources, and how well proposed approaches fit with current systems.

Balancing Quality and Quantity in Practice

Faced with considerable quantities of material to which general access was unavailable, even in the reading room, and the Library of Congress' wish to provide online access for the general public to its collections to the degree possible, the Prints and Photographs Division has found ways to balance the pressures for both full description to support precise retrieval and access to a greater proportion of the holdings.

P & P recognizes different degrees of cataloging quality. Full cataloging, with verification of all information, addition of names to LC's name authority file, preparation of descriptive summaries, and extensive application of subject terms has been used traditionally for collection- and group-level records and for item-level records for certain categories of material, such as fine prints and posters. To prepare a full catalog record for an item takes between three and four hours. Of the collections within American Memory, only the Selected Civil War Photographs (http://memory.loc.gov/ammem/cwphome.html) and the daguerreotypes in America's First Look into the Camera (http://memory.loc.gov/ammem/daghtml/daghome.html) have full cataloging. In contrast, when a user requests a reproduction of an uncataloged item, "minimal" cataloging is performed for the item. Names of people and places are checked against the name authority file in the Library of Congress catalog; prescribed forms are used if found, and conflicts are avoided, but the research necessary to support the addition of a new name to the authority file is rarely performed. A small number of terms are chosen from the Library of Congress *Thesaurus for Graphic Materials (LCTGM)* to characterize the format and genre of the work and provide topical access. Whenever possible, a geographic heading is chosen from Library of Congress Subject Headings (LCSH). For minimal cataloging, no additional research is performed and no attempt is made to describe what the picture is about as opposed to what it is of. Such records typically take two hours to create and review. Reproduction requests for roughly 3,000 new items are received each year;

copy negatives created before this practice started are gradually being cataloged at the minimal level.

For most photographic collections selected for digitization as a whole, the Prints and Photographs Division has cataloged items at a preliminary level using only information from the piece. Captions and dates provided by the photographer are recorded but not verified. Where no specific information is available, a brief descriptive title is devised (and indicated), and an approximate date or date-range is given. Subject terms are applied sparingly, although consistently within a collection. Names are not checked against the authority file. To speed workflow, P & P develops collection-specific automated procedures to complete fields for information that are common to all or many item-level records in a batch. Preliminary cataloging for the photographs of Theodor Horydczak (for the American Memory collection Washington As It Was) took between thirty and forty-five minutes per record.

Large documentary collections, such as the HABS/HAER drawings and photographs that record architectural sites and engineering structures and the FSA-OWI photographs, pose particular problems for access. In making these available online for the public, P & P has explored new approaches to speed cataloging. Collections of negatives are often received by LC in an organization that roughly collocates pictures taken at the same time. For example, about one-third of the FSA-OWI collection was held on strips of 35mm negatives and many of the larger negatives were filed in the batches in which they had first been developed. The agency had provided information on photographers and geographic locations in various forms of documentation given to LC with the collection. For some images, captions and other notes had been recorded on cards. The information was transcribed by Library of Congress staff and merged with boilerplate information to create skeletal records in MARC format. Geographical locations were transcribed in the uncontrolled form used on the cards. Conversion to a standard form was performed automatically without verification. For roughly 35 percent of the images, no details were available beyond a sequenced call number. However, the caption or cataloging for one photograph often applies to or illuminates shots taken before or after it. To give users the clues offered by captions for neighboring images, a feature was added to American Memory and the Prints and Photographs Online Catalog that allowed visual browsing as a grid of thumbnail images sorted by call number.

For the HABS/HAER materials, the challenge was less the shortage of information to support access but more its incompatibility with library practice and formats. HABS and HAER are ongoing programs of the National Park Service, which keeps detailed information on the materials from each architectural site in a relational database. Every few months, batches of material are passed to the Library of Congress to serve and archive. For this project, the Prints and Photographs Division took advantage of the fact that PPOC and American Memory use a general-purpose search engine designed to search several resources simultaneously, not necessarily in the same format. Routines were developed to convert the structure and format of the records in the relational database to a "flat" file of descriptive records more easily searched in combination with traditional bibliographic records. Each "bibliographic" record describes a site or structure. A single identifier provides a logical link to the content of all related documentation that has been digitized; hence the record serves as a group-level record. The documentation for a site may include photographs (with captions listed on separate pages), drawings, photographs, and pages of textual documentation (known as "data pages"). Each form of original content is being digitized independently using procedures for capture and quality review appropriate to the content. Whether the related content is yet digitized or not, records describing the documentation available are included in the database to support access to the physical collection and requests for reproductions. After new batches of image files are prepared and loaded, they are automatically retrievable from the bibliographic display.

Since HABS and HAER are ongoing programs generating new surveys, new copies of the database are retrieved regularly from the National Park Service, transformed, and re-indexed. No attempt is made to add controlled subject headings (although, with encouragement and assistance from the Library of Congress, the Park Service may start to use a controlled set of terms in light of its experience with providing public access). Much information from the original database is treated as notes; topical access is therefore primarily by free text search on the entire record. Automated expansion of query terms to include variants, which is done by default for both the Prints and Photographs Online Catalog and American Memory, is of particular value here.

For any HABS/HAER site, the related documents are treated as groups by original type. Hence, a typical engineering structure has an associated group of black-and-white photographs (with captions listed on separate pages), a group of drawings, and a group of pages of text digitized as page images. For each group, a small file (which LC has nicknamed a "page-turning data set") supports navigation through the group. The file holds sequencing information, links to component image-files, and optional captions. This data set, which can be thought of as "structural metadata" for the complex object representing the group of images, is generated automatically from names of files in a directory. The approach and the naming conventions that support the automatic derivation of the structural metadata were originally developed for American Memory, where it was first used to allow users to "turn" through pages of a short document, such as a theater program, and has since been extended to support page-turning through much longer works. From each page, links to images of higher resolution permit more detailed study or better printing. The HABS/HAER drawings and data pages are presented through the original page-turning interface. For the photographs, the group is presented as a grid (or a sequence of grids) of thumbnails with associated captions. The programming for generating the page-turning data sets, the page-turning interface, and the thumbnail grid displays was done by Library of Congress staff.

The successful use of group-level description supported by thumbnail grids in HABS/HAER may lead to its use in future projects. There may be further opportunities to save cataloging effort by building on existing machine-readable descriptive records. The Prints and Photographs Division takes care to distinguish the levels of cataloging employed in records. Records that support access but do not meet LC's usual quality standards will not be distributed to the bibliographic utilities and other institutions as the basis for copy cataloging.

Preserving and Presenting Context

One characteristic of the Prints and Photographs Online Catalog and American Memory that is not provided in most library catalog systems, which developed primarily as tools for organizing holdings of independent monographs, is the clear delineation of a collection. The importance of a collection may lie in its overall scope and relationships among its parts; preserving the archival integrity of a physical collection is a guiding principle for curators. Online systems for archival pictorial collections should allow the user to wander within the boundaries of a collection with easy access to collection-level information that provides an intellectual context for the entire body of material. The front matter of an archival finding aid serves this purpose. American Memory and PPOC provide this context through an introductory framework of HTML pages, which often also feature exhibit-like presentations. This framework also allows LC to describe how a collection has been cataloged or digitized, in part to help users search more effectively or understand why all items do not have high-resolution reproductions and in part to inform others involved in similar ventures.

American Memory and PPOC encourage searching across the entire resource as well as within a chosen collection. To support the easy limiting of retrieval to item-level records from a single collection, collection identifiers (mnemonic codes rather than formal collection titles) are used (recorded in a local field in MARC records). The same coded field supports the automatic export of all records for a collection from LC's main catalog. More than one collection code can be included in a record, allowing the same item to be part of more than one collection. This is needed, for example, because items in LC's collections of daguerreotypes or panoramic photographs may also be part of collections that relate to provenance. In the online environment, virtual collections can be assembled when curatorial or reference staff believe it will be valuable. The Prints and Photographs Division has used this capability for a few small selections made especially for American Memory. Examples include Votes for Women (http://memory.loc.gov/ammem/vfwhtml/vfwhome.html) and Jackie Robinson and other Baseball Highlights (http://memory.loc.gov/ammem/jrhtml/jrhome.html). More extensive use of the capability is made by the Geography & Map Division and the Motion Picture, Broadcasting, and Recorded Sound Division. For example, early motion pictures (many made by Thomas Edison) are presented as a body of material and as thematic collections; a separate virtual collection includes motion pictures made by Edison with his sound recordings.

Forms and Formats for Access Aids

The Prints and Photographs Division chose to use the MARC format for cataloging its pictorial materials primarily for compatibility with LC's main catalog and to allow distribution of records to bibliographic utilities (Zinkham, 1995, p. 48). The vision of a single catalog for all Prints and Photographs Division resources, and a wish to maintain only one copy of any descriptive record were other factors supporting this choice. Almost all records created and maintained by P & P are held in the MARC format, although the records may be derived automatically from a database used for processing the collection, as for the FSA-OWI photographs. Over the past few years, division staff have become skilled in exporting batches of MARC records to a form in which global changes can be made more easily and for converting back to MARC format.

The HABS/HAER records (for which the National Park Service maintains the data) are not converted to the MARC format but to a simpler format easily indexed by the search engine used for American Memory and the Prints and Photographs Online Catalog. This format identifies fields within records by labels in angled brackets (similar to SGML or HTML) and is easy to generate from any database software and to manipulate using simple programs or scripts. To allow coherent searching and presentation of records from both sources, fields have been mapped to MARC, and a common set of index fields is used (e.g., author/creator, title, and subject). Similar non-MARC item-level records are used for several collections in American Memory, particularly for materials that would not normally be cataloged individually in LC's catalog, such as flutes in the Dayton C. Miller collection. Several collections from awardees of the Library of Congress/Ameritech competition will be integrated into American Memory by transforming records from the awardee's database into this format, which is comparable to (and can be mapped to) the Dublin Core elements, with a small set of qualifiers. The Library of Congress expects to adopt an XML (eXtensible Markup Language) representation for Dublin Core descriptive records in the future.

Staff from LC's Manuscript Division and Prints and Photographs Division were involved in the development of the Encoded Archival Description (EAD) standard, a document type definition for the Standard Gener usually added as notes rather than in subject headings or captions. Searches limited to subject fields would miss this information.

Records for some items may nevertheless contain subject terms that have not been checked against an approved thesaurus. Terms may be transcribed from earlier records or suggested by a scholar in conjunction with a special project, such as an exhibit or publication. In the Prints and Photographs Online Catalog, the uncontrolled terms are displayed separately, labeled as Topics rather than Subjects. A third category, Format, is used for terms that describe the form and genre of the original item. However, a search by subject includes all three categories. Initially, a subject search in PPOC was limited to controlled headings until reference staff confirmed that users found the distinction of little value for searching. Within American Memory, the three categories are displayed as a single group.

Users, expert and novice alike, like to look for items in archival collections by date or geographic location. For American Memory, these attributes expose many challenges since consistency across the entire heterogeneous body of materials, although clearly desirable, is hard to achieve. For published works, the date and place of publication, although often easy to ascertain, may bear little relation to the period and location described or represented, which may be of more interest to users. Fixing an unpublished pictorial work precisely in time and space is often infeasible. For most photographs, what matters is where and when the exposure was made. However, if the photographer does not record those details, precision is impossible. For unknown dates, the degree of uncertainty is conventionally represented in ways that are comprehensible to humans once a record has been retrieved—e.g., 189-?, ca. 1892, 1892 or 1893. These conventions, however, create problems for automated retrieval or sorting by date. For some American Memory collections (e.g., George Washington's correspondence), dates are recorded in a standard form so that search results can be sorted chronologically. Effective searching or sorting by date for the Prints and Photographs Division's pictorial materials proves elusive, and PPOC makes no explicit attempt to support it.

For retrieval by geographic location, a more satisfactory solution is available. The Prints and Photographs Division and the Geography and Map Division both make use of a hierarchical nation-state-county-city breakdown (e.g., United States—Illinois—Mercer County—Aledo). Whichever components are known can be recorded. This form has proved useful both for human readers and for generating lists of place-names for browsing and clickable maps that permit retrieval by state. The desirability of having certain elements within catalog records easily parsable by computer programs is reflected both in extensions to the MARC format in recent years and in the discussions surrounding the development of the Dublin Core.

Storing and Relating Description and Content

Identifiers as Links from Description to Digital Content

The Library of Congress maintains its catalog records and finding aids independently of the digital reproductions it makes. It does not expect to embed all descriptive metadata into the content or to manage all descriptive metadata and digital content in one integrated system. This practice follows from the desire for a unified catalog that provides access to information in all forms and facilitates the sharing of cataloging effort among libraries. LC distributes copies of its catalog records to other institutions through its Catalog Distribution Service; some records include links to digital content stored at the Library of Congress. Access to overlapping content in American Memory and PPOC is supported by the same set of catalog records, but this set is separate from LC's main catalog. For the records that overlap with the main LC catalog (currently only a small proportion), copies are made and re-indexed regularly. Now LC has moved its main catalog to the new integrated library management system (ILS). The Prints and Photographs Division hopes to load the rest of its MARC records into the main cataloging system. Copies of the records will still be exported for American Memory and PPOC, for which development will continue in parallel with deployment of the ILS.

Since LC cannot predict how and where its catalog records will be used, the link between a descriptive record and a digital content "object" must be a persistent identifier in a standard format that is globally unique and, unlike Uniform Resource Locators (URLs), will not change if LC moves digital content from one computer to another. An intermediate system is needed to "resolve" the identifier to the correct physical location; when the physical location for an item changes, a record will be updated once in the resolution system rather than in the catalog record and all its copies. LC has installed and begun to use the Handle System® from the Corporation for National Research Initiatives for this purpose. Experimentation with handles as persistent identifiers in catalog records has begun for monographs and maps. The handles are based on the scheme of logical identifiers used for all materials digitized by NDLP.

Each item has a unique two-part logical identifier. As examples, dag.3g05001 is a daguerreotype portrait and musdi.139 is a reproduction of *Powell's Art of Dancing*, a dance instruction manual. Currently, the two parts of the identifier are related to names of directories and file names in the UNIX system hierarchy. In the longer term, the logical identifiers will serve more generally as unique persistent identifiers, however the content is stored. The handle for the dance manual is urn:hdl:loc.music/

musdi.139. A catalog record for this item incorporating the handle is in LC's main catalog and has been distributed to other institutions. The handle resolves to a Web-based presentation of the book, generated dynamically from its digital content, which includes page images and transcribed text. Since only Uniform Resource Locators (URLs) are usable today by most browsers and library catalog systems, the MARC record includes http://hdl.loc.gov/loc.music/musdi.139 as a URL. The use of the proxy server hdl.loc.gov provides an identifier that is usable today but is not entirely independent of physical location, since the proxy service is supplied by a particular computer. Whenever Uniform Resource Names (URNs) are deployed as a standard across the Internet, use of the proxy server can be discontinued.

Since the handles are resolved by the handle server, independently of any particular database or application, these identifiers can be used as links from any document or descriptive record. They support links from American Memory, PPOC, LC's main catalog, and from other catalogs that incorporate Library of Congress records. LC will use handles in finding aids to link to related digital content. Users, from scholars to schoolchildren, can use these handlles to turn citations in online papers into active links. The Prints and Photographs Division expects to start using handles for pictorial items now that LC has migrated its cataloging operations to a new library management system.

What Do Identifiers Identify?

An identifier, such as dag.3g05001, does not identify a single image file but a cluster of files, typically an archival master file with derivative thumbnail and service images. The number and characteristics of the images in this cluster may change over time. New thumbnail images have been generated recently for several older Prints and Photographs Division collections for better quality and more consistent sizing. In some instances, service images of a different size have been generated. Catalog records needed no changes since they contain no technical details for the image files. The metadata that expresses the relationships of the component images to the complex object (sometimes also called a meta-object), and describes the individual files, is held elsewhere. The structural information for pictorial images is currently largely implicit in file names; eventually, it will be recorded explicitly in a repository system designed to support the management of digital collections. Within American Memory or PPOC, the identifier for a picture triggers a dynamic presentation built from the associated structural metadata.

For most pictorial items, the presentation embeds a thumbnail within a bibliographic record; larger versions of the image are available by clicking on the thumbnail or a labeled link. A bibliographic record for a typical item digitized recently links to a single "picture object" with a master image and one or two smaller service images. As mentioned earlier, the identifier in a HABS/HAER record links to a much more complex object, consisting of all image files for all the related photographs, drawings, and pages of text, with structural metadata recorded for each category of images. For the furnaces of the Sloss-Sheffield Steel and Iron Company in Birmingham, Alabama, the originals include 134 black-and-white photographs, 49 textual pages, with 20 drawings and 2 color transparencies still to be digitized. Between these extremes of complexity are the digital objects representing baseball cards; these include images of both front and back, each in several sizes.

No standard digital formats have been developed yet to represent objects as complex and varied as those created by the National Digital Library Program and other programs creating digital reproductions. Within the Library of Congress, an effort is underway to develop a common framework for the structural and administrative metadata needed for the variety of materials being converted to digital form. Early thoughts in this area contributed to the collaborative Making of America II project (http://sunsite.berkeley.edu/MOA2/), which involves five research libraries, all members of the Digital Library Federation, to which the Library of Congress also belongs. In the MOA-II project, XML is being used to represent a flexible hierarchical structure for "archival objects." Programmers from LC at the University of California, Berkeley, have developed a set of software tools to record the structural and administrative metadata, generate the XML files that represent the objects, and allow users to navigate the hierarchy and display an object's components. LC hopes to investigate whether the MOA-II archival object format and related tools might be applied effectively to its content. Meanwhile, in the commercial sector, formats for electronic books are proliferating.

When widely accepted formats exist for objects, corresponding tools or helpers for viewing, navigation, and manipulation become available. It is then possible simply to provide access to the stored objects and rely on users having the necessary tools. Individual JPEG images and PDF documents are examples of formats that can be treated this way. For more complex classes of objects, an institution that wishes to provide convenient remote access to individual works it has digitized can generate Web-accessible presentations of each object and support direct links to the presentations. This is the current approach taken by LC when it assigns handles to digitized books or maps. The handle invokes a dynamically generated presentation, in effect specifying a search for a known item within American Memory. However, the user's view of a book or a baseball card digitized at the Library of Congress may be very different from the display of similar items digitized at another institution. It remains to be seen whether users will find the variety confusing or tolerate it as a minor inconvenience given the benefits of access to a wealth of resources distributed across the Internet. LC believes that the community will benefit from increased standardization or at least commonality of practice. Agreed formats for representing complex objects can support not only more effective interoperability but also a shared strategy for archiving and preserving digital materials for the future.

Storage Media

In 1996, the Library of Congress was faced with rapidly increasing requirements for digital storage as plans were drawn up for digitizing millions of pages of text, hundreds of thousands of pictorial items, thousands of maps and sound recordings, and hundreds of motion pictures. At the time, the cost of magnetic disk storage was $1 per megabyte. A decision was made to install a hierarchical storage management system that would automatically move less-used files to magnetic tape cartridges in a robot-controlled jukebox. Initially, a terabyte of disk storage was combined with 4 terabytes of the cheaper tape capacity (1 terabyte = 1,000,000 megabytes). All files were always accessible, but those on tape took longer to retrieve. For two years, this system proved effective but, as the number of files grew, the perceived performance dropped and backups became increasingly difficult to complete during off-peak hours. By late 1998, the cost of disk storage had dropped to between 20 and 30 percent of its 1996 cost and was expected to fall to 10 cents per megabyte by the year 2000. Use of the hierarchical storage management system has been abandoned for the time being, although LC expects to track the improvements in this class of product closely. In early 1999, LC had 19 terabytes of disk storage attached to its pool of UNIX servers and began developing an enterprise storage network that is independent of its primary data network so that backups place a minimal load on the processors and network that support users' activities. The storage network will support different media types as necessary. For example, a robotic jukebox for CD-ROMs would allow direct loading of image files delivered by scanning contractors without affecting regular network performance.

Providing Technical Support for Online Access

Indexing for Search and Retrieval

American Memory and the Prints and Photographs Online Catalog rely on InQuery (from Sovereign Hill Software, recently acquired by Dataware Technologies), a search engine, developed for indexing free text, that can recognize fields in tagged records or documents. The search system was selected for American Memory as appropriate for indexing the full text of book-length works alongside bibliographic records for multimedia materials. InQuery is not a single program but a flexible set of tools for application developers who retain complete control over visual design. It is the underlying engine for very different applications at LC, including: THOMAS, which provides public access to current legislative information; the Handbook of Latin American Studies, a traditional bibliographic service; and LC's archive of finding aids. The ability to include heterogeneous sources of data in a single search has proven valuable for PPOC, allowing the integration of non-MARC records; integration of finding aids should be feasible in the future.

Systems for indexing free text are significantly different from those designed for indexing relational databases or traditional library catalogs. They can be configured to handle word variants automatically and to ignore punctuation. They are designed primarily to take free text as queries and return a list of documents ranked by "relevance." Each product uses its own formula for relevance, but all give higher weight to documents that contain more of the words in a query, higher still if query words are repeated in the document. Words that occur in many documents in a collection contribute less to the relevance score than words that occur infrequently. Most allow searching for phrases or for words close to each other. Strict Boolean operations, however, are seldom useful for searching free text. Initially, many LC staff familiar with traditional catalogs found the lack of Boolean capabilities in American Memory troubling and failed to notice the benefits of a system that found words anywhere, was more forgiving of inconsistency in the data, and required less exactness in query formulation by users. Lengthy undifferentiated lists of "hits" perturbed those used to systems designed when precise searching and small result sets were essential because of technological limitations. Over time, incremental changes have been made (some of which are described below) to the indexing, the search options, and the presentation of search results to address these concerns. Reference librarians in the Prints and Photographs Division involved in the design and testing of PPOC pressed for enhancements that have benefitted the users of American Memory. Most of these enhancements support more precise searching for users who wish to take advantage of the capabilities. Simultaneously, staff who had complained bitterly about shortcomings began to realize the benefits of the different indexing approach and learned how to use it to advantage. Interestingly, remote users of American Memory, whose basis for comparison is often Internet search engines, are less concerned by long lists of hits, although they do look for ways to search more precisely for what they want, particularly in the full-text materials.

When a user enters a query into the single box of a query form in PPOC or American Memory (ignoring any options), several queries are performed in the background and a combined set of results is presented. First, the query is treated as a phrase. In American Memory, the second search looks for records where all the words occur within a passage of twenty words. Since this distinction is primarily of value when searching long documents, rather than catalog records, it is omitted in PPOC. The next search identifies records that include all the words anywhere using an implicit Boolean AND. The final search is for records that include any of the words using an implicit Boolean OR. Duplicates are removed and the three or four sets are presented as a subdivided list with entries in each grouping ranked by relevance as determined by InQuery. The explicit division of the results list was one of the enhancements introduced after complaints about long lists of hits. Users appreciate clues that help them decide how far down a hits list to bother looking. American Memory search forms now also offer the option to limit results to those that include all words entered or the exact phrase. Originally, the limit for a results set was 5,000. As the size of the resource and the volume of usage increased, the effect on performance of building unnecessarily large sets became a concern. American Memory searches now return no more than 500 records by default, although users can choose to raise the limit back to 5,000.

For the Prints and Photographs Online Catalog, the ability for users to search by creator, title, subject, and various numeric identifiers, such as call number, was considered essential. The heterogeneity of material searched within American Memory has discouraged the explicit use of field qualification, although it has been used implicitly in browsable lists of subjects (within individual collections) and in bibliographic displays where the user can click on a subject heading or name to search for other records with the same heading. Since the summer of 1998, the option to search by creator/author, title, and subject has been introduced for selected collections within American Memory. Care has been taken to include the same MARC fields in these options as are included in the corresponding options in the main LC catalog. For searching across the heterogeneous American Memory collections, the general search is still the only option provided. As the number of collections and items in American Memory has grown, the simple choice between searching all collections or within a single collection became inadequate. The ability to limit a search to collections that are primarily pictorial (or textual, cartographic, etc.) was introduced in 1997. In January 1999, American Memory introduced a new feature that permits users to pick any set of collections to search.

Under the covers, configuration options have also been changed. Full-text retrieval systems often perform automatic stemming while indexing since this results in much smaller indexes and hence faster retrieval on average. However, since a stemmed index precludes searching for exact word forms, it was determined very early during testing to be unacceptable for searching for titles in bibliographic records. By default, user queries are expanded to include word variants; users can choose to match words exactly. Another technique to reduce index size is to ignore common words, known as "stopwords." After perplexed librarians could not retrieve some titles, the initial stopword list was pruned to match the minimal list used for LC's main catalog (containing only articles, conjunctions, and the most common prepositions—sixteen words in all). In making both these decisions, LC has judged precise retrieval more important than efficiency, given that the effect on performance was not obvious and the costs of computing power and disk space continue to fall.

Access Management

Some of the items digitized by the Prints and Photographs Division are subject to copyright protection or special terms of gift that prevent the Library of Congress from making them freely accessible over the Internet. As part of a prototype repository, a model has been developed for managing access for authenticated users under terms that could, when required, be specified for individual objects. However, no plans are in place for immediate implementation of such a scheme. In the meantime, access to certain collections is limited by Internet IP address to use within LC. Remote users can search and view all the records in PPOC but will not be able to retrieve some images that would be accessible if they were in the reading room.

Interacting with Users: Interface Design

Although both are accessed via Web-browsers, built with the same toolkit, and relying on much of the same code and the same indexes for overlapping content, the visual designs of PPOC and American Memory are strikingly different. Each reflects the instincts of the initial design team and their interpretations of an intended user community's expectations, balanced by technical considerations and modified by incremental changes made in response to feedback. PPOC is plain, with extensive textual explanations and a look reminiscent of the popular text-only catalog displays used previously in the reading room. Initially, reference staff had pushed for a page-oriented display with Next Page buttons, concerned that a scroll bar would be difficult for users to manipulate; this had to be abandoned after it proved too difficult to calculate where page-breaks would come when images were of different sizes. Some of the professional picture researchers who have used the reading room for many years have been slow to adapt to new technology; for most users, however, the mouse is either familiar or soon mastered. American Memory uses color, space, and icons more freely, expecting its primary users to know how to scroll and click through Web pages. PPOC gives priority on displays to information about accessing originals and ordering reproductions, whereas the focus for American Memory is on the online content. However, the underlying functionality of the two search systems is the same since the organization of the descriptive records and digital content and the search capabilities are identical. Each system is constructed of four layered components as shown in Figure 2,

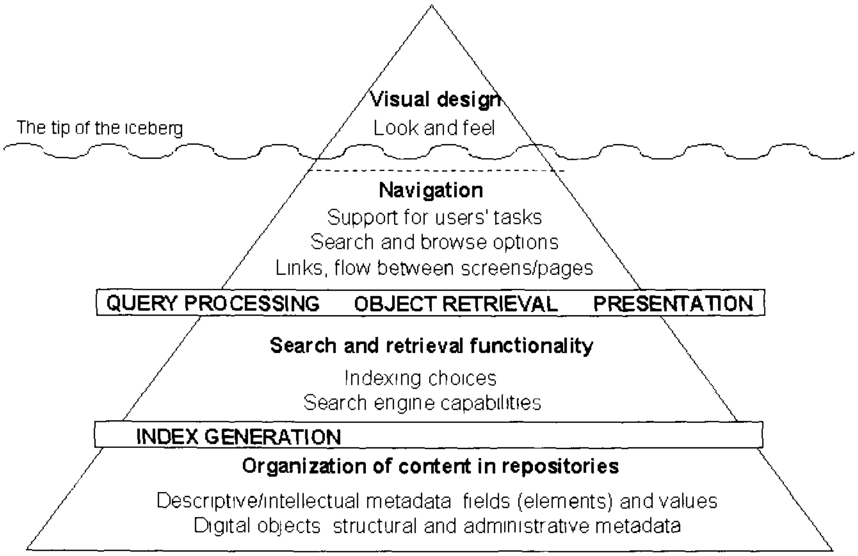
The tip of the iceberg
Visual design
Look and feel
Navigation
Support for users' tasks
Search and browse options
Links, flow between screens/pages
QUERY PROCESSING
OBJECT RETRIEVAL
PRESENTATION
Search and retrieval functionality
Indexing choices
Search engine capabilities
INDEX GENERATION
Organization of content in repositories
Descriptive/intellectual metadata fields (elements) and values
Digital objects structural and administrative metadata

Figure 2. Search Interfaces Rely on Underlying Content and Indexing Capabilities.

a model that describes how many of LC's search and retrieval applications are constructed. They share the three lower layers, differing only at the visual design level.

As development of the two interfaces has continued, staff in the Prints and Photographs Division and the National Digital Library Programs have learned that there are advantages to increased commonality where feasible, if only because the single programming team is able to make enhancements more quickly. Although the bibliographic displays are different, the thumbnail grids and page-turning displays are generated by the same code. Ideas from either community feed the iterative development of both interfaces.

The public mission of the Library of Congress and the requirements of the Americans with Disabilities Act mean that both PPOC and American Memory make less use of icons and graphics than many of today's Web pages. Fast loading of pages is essential to reach the widest possible audience including schools and homes with slow network connections and older computers and software. Pages must be constructed in a way that is usable by assistive software for the handicapped. Rather than create text-only pages as alternatives, LC has chosen to be conservative about adopting technology that may make its collections less accessible. As an example, LC does not use frames in its page-turning presentation.

Facilitating Use of the Images

Two factors control re-use of images: whether a copy is available in a form suitable for the intended use, and whether re-use is permitted by copyright law. Roughly one-third of the electronic mail that comes into the reference desk of the Prints and Photographs Division relates to rights and reproductions. For the American Memory Help Desk, the proportion is much less—around 5 percent. As an agency of the U.S. Government, the Library of Congress claims no rights in the digital versions of the materials it converts. However, rights relating to the original materials pertain also to derivative works, including digital reproductions. Since many of the photographs in the P & P collections were not published and may have been created for hire, ascertaining the copyright status and the identity of possible rights holders for individual items is usually infeasible. For public access, LC has focused on converting materials produced by the U.S. Government, those likely to be out of copyright by virtue of their date of creation, or collections where a single organization or individual appears to hold copyright and commercial interest is unlikely. However, by making items accessible through American Memory, the Library of Congress does not (and is not legally empowered to) warrant them to be in the public domain. When contact has been made with copyright holders, they have usually been happy to allow educational and research use, and LC records these permissions on catalog records and on the Web pages introducing collections. Other forms of right may also apply, including privacy for the individuals represented in photographs. Whether a particular re-use is permissible under the "fair use" doctrine or violates rights is not clearly defined by law or regulation. U.S. law only specifies factors to be considered by a court adjudicating a case brought by a rightsholder. Since LC claims no rights itself and is prohibited from offering legal advice, requests for permission to re-use cannot simply be answered "Yes"—to the frustration of librarians and users alike. However, LC makes every effort to describe its understanding of the factual background and known rights for each collection in order to assist users.

Users who visit the Prints and Photographs reading room are often seeking pictures to use in publications. Until recently, only photographic reproductions could provide publication quality copies. LC's Photoduplication Service offers photographic reproductions for a fee set to cover costs. Prints are made from copy negatives or interpositives; when a reproduction request is received for an item for which no copy exists, one is made. Remote access has generated a growing volume of interest in reproductions for personal use—e.g., when family members are recognized. The archivist of the Institute of Regional Studies at North Dakota State University reports a steady stream of requests for reproductions of photographs from the institute's collections in American Memory (The Northern Great Plains:1880-1920). Requests come not only from those with ties to the area but from people who simply like the images of pioneer life. Users with appropriate equipment and fast enough network connections can download the images directly. However, the Prints and Photographs Division reference librarians find that few users are yet comfortable downloading large files; most still prefer to wait for a traditional photographic reproduction. In the future, the Photoduplication Service may use high-quality scans from original negatives (as made for the HABS/HAER collections) as the source for prints of photographic quality.

REFERENCE SERVICES

Reference staff in the Prints and Photographs reading room have seen online access to images as reference surrogates as a goal for many years. They have pressed for convenient one-stop access for both staff and patrons to the division's holdings. Compared with earlier catalog systems, the current Prints and Photographs Online Catalog has advantages beyond a growing volume of content and the ability to serve remote users. As mentioned earlier, free text search capabilities compensate for less than perfect descriptive data. Simultaneous access to images from any workstation with a Web browser creates new possibilities for productivity by both staff and patrons. Recently, many thousands of uncataloged copy negatives were added to PPOC; skeletal records have a reproduction number and a link to corresponding digital image files but no title, creator, or subject headings. Since, following LC's recommendation, reproduction numbers are often cited in publications, staff and patrons can benefit from the ability to retrieve and view images even if they have not been cataloged.

Since PPOC has been accessible over the Internet, the volume of electronic mail reference questions to P & P has doubled from thirty-five messages per month to seventy. Many are basic questions to which answers are already available online, including those seeking permission to use the images or to obtain reproductions. A direct link has been added from each bibliographic display in PPOC to a page that cites the reproduction number and provides detailed information on the services provided by the Photoduplication Service and general information regarding copyright and other potential restrictions on use. The American Memory Help Desk provides standard responses to common questions in a list of Frequently Asked Questions.

Demand for reference service in the P & P reading room has always been high. Remote online access to images, through American Memory and PPOC, seems neither to have stimulated use of the physical collection nor to have reduced the number of users visiting the reading room. Reference staff, however, have noticed that the number of free-lance professional picture researchers using the collection has increased, suggesting that remote access to images over the Internet has either stimulated growth in this niche service industry or allowed individual researchers to discover and explore more sources for pictorial material for their customers. Another effect of online access to images is the need for increased communication among divisions within LC and with other institutions about the availability of digital reproductions for online viewing or use. Effort has been needed to ensure that electronic mail questions on pictorial materials (or any other specialized topic) are dealt with consistently, whether sent to the American Memory Help Desk, LC's main Web service, or directly to a division.

CONCLUSION

The challenge of providing effective online access to visual materials goes far beyond the process of digitization. Indeed, the Library of Congress' experience suggests that by using consultants and contractors expert in the physics inherent in scanners and output devices, the mathematics of image transformations, and schemes for color management, and with the engineering skill to build systems that control them, it is possible to develop procedures that provide high quality images in large volumes. Harder to resolve are the underlying problems of organizing and describing visual materials cost-effectively (whether digitized or not) in ways that help users find the information, illustration, or evidence that they need. Practices developed for the published printed literature need considerable adaptation. Experience with the large FSA-OWI and HABS/HAER collections suggests directions that online browsing, treatment of pictures in groups, and free text retrieval can reduce the need for full cataloging of individual items.

Distinctions between the human interface presented by digital images on a computer monitor and by a folder of prints on a large table provide another aspect of the challenge. Enhancements in interfaces will be inevitable but gradual. Grids of thumbnails provide one approach for side-by-side comparison and rapid browsing through groups of images. Technological advances in processors, networks, and monitors will handle images faster and better and provide more options. More generally, approaches to interface design will improve through better understanding of how people interact with visual materials. At the same time, human ingenuity will suggest new ways to cope with or take advantage of characteristics of the online environment. Consider the use of microfilm: few users are enthusiastic about the medium, but experienced researchers become skilled at scrolling rapidly through a reel focusing on distinctive patterns (such as "target" pages) to indicate when to stop. The modular design of the architecture that supports American Memory and the Prints and Photographs Online Catalog will permit incremental enhancements in response to changes in the technical environment and the expectations of users.

An important objective for the Library of Congress is enhanced access to its comprehensive collections. LC has already made digital reproductions of hundreds of thousands of photographs. However, these constitute only a tiny fraction of its pictorial resources. LC seeks cost-effective solutions to provide integrated access to resources in many forms of expression, whether digitized or not. LC staff most closely involved with developing the current practices for providing online access to pictorial material are hesitant to call them best practices, preferring to consider them as appropriate practices given the state of technology and the Library of Congress' institutional objectives and constraints.

REFERENCES

Betz, E. W. (1982). *Graphic materials: Rules for describing original items and historical collections.* Washington, DC: Library of Congress. Retrieved September 3, 1999 from the World Wide Web: http://www.TLCdelivers.com/tlc/crs/grph0199.htm.

Betz Parker, E. W. (1985). The Library of Congress non-print optical disk pilot program. *Information Technology and Libraries, 4*(4), 289-299.

Flynn, M., & Zinkham, H. (1995). The MARC format and electronic reference image: Experience from the Library of Congress Prints and Photographs Division. *Visual Resources, 11*(1), 47-70.

Frey, F. S., & Reilly, J. M. (1998). *Digital imaging for photographic collections: Foundations for technical standards.* Unpublished report by the Image Permanence Institute, Rochester Institute of Technology. Final Report to the Office of Preservation, National Endowment for the Humanities (NEH GRANT PS-21084-95).

Kenney, A. R.; Shapiro, L. H.; with Berger, B.; Crowhurst, R.; Ott, D. M.; & Quirk, A. (1999). *Illustrated book study: Digital conversion requirements of printed illustrations.* Retrieved October 27, 1999 from the World Wide Web: http://www.library.cornell.edu/preservation/ill_bk_cover.htm.

Library of Congress. (1995a). *Thesaurus for graphic materials: I. Subject terms.* Retrieved September 3, 1999 from the World Wide Web: http://lcweb.loc.gov/rr/print/tgm1.

Library of Congress. (1995b). *Thesaurus for graphic materials: II. Genre and physical characteristic terms.* Retrieved September 3, 1999 from the World Wide Web: http://lcweb.loc.gov/rr/print/tgm2.

Library of Congress. (1997). *Request for proposals for digital images of pictorial materials.* Retrieved September 3, 1999 from the World Wide Web: http://memory.loc.gov/ammem/prpsal9/coverpag.html.

Library of Congress. (1996). *Cataloger's desktop* [CD-ROM, updated annually]. Washington, DC: Library of Congress Cataloging Distribution Service.

Ostrow, S. E. (1998). *Digitizing historical pictorial collections for the Internet.* Council on Library and Information Resources. Retrieved September 3, 1999 from the World Wide Web: http://www.clir.org/pubs/reports/ostrow/pub71.html.

Reilly, J. M. (1995). Technical choices in digital imaging: The technical images test project in review. In P. A. McClung (Ed.), *RLG Digital Image Access Project* (Proceedings from an RLG symposium held March 31-April 1, 1995, Palo Alto, CA) (pp. 85-93). Mountain View, CA: Research Libraries Group.

Acronyms

DLF	Digital Library Federation
FSA	Farm Security Administration
G & M	Geography and Maps Division, Library of Congress
HABS	Historic American Buildings Survey
HAER	Historic American Engineering Record
LC	Library of Congress
NDLP	National Digital Library Program, Library of Congress
OWI	Office of War Information
P & P	Prints and Photographs Division, Library of Congress
PPOC	Prints and Photographs Online Catalog (pronounced pea-pock)
URL	Uniform Resource Locator
URN	Uniform Resource Name
XML	eXtensible Markup Language