OPTICAL STORAGE OF PAGE IMAGES AND PICTORIAL DATA - OPPORTUNITIES AND NEEDED ADVANCES IN INFORMATION RETRIEVAL

William R. Nugent
Jessica R. Harding
Automated Systems Office
Library of Congress

ABSTRACT

We describe two current development projects at the Library of Congress using high-density optical storage, both of which require more advanced and improved computer-based information retrieval methodologies than existing bibliographic retrieval systems. A much greater emphasis will be placed on the information content of the articles rather than on the broad subject categories in general use for computer retrieval citations to book materials. Needed approaches include the linking of selected external reference sources and the extraction of character-encoded index information from the page images by OCR techniques. Our position is that extremely high density storage of articles in a single subject area requires a finer resolution in the retrieval system accessing the material. A single side of a 12-inch digital optical disk, for example, can hold 15 years of the Journal of the ACM.

1.0 THE DIGITAL OPTICAL DISK PROJECT

1.1 Text Materials to be Stored as Digitized Images

The Library of Congress has embarked upon a pilot project using digital optical disk to store the digitized images of one million pages of text over a two-year period and provide rapid access, display and printing of those images to seven work stations located in the three buildings of the Library.

High-use current periodicals will be given initial emphasis and will include journals in science, technology, and business; newspapers in German, Portuguese, Japanese, Thai, French, and Hebrew; government documents such as the Budget of the United States and the Congressional Record; United States Agriculture Decisions and Social Security Rulings from 1960-1975; maps, atlases, and microfilm, selected manuscript collections and music. Currently, these periodicals are in such demand that copies are not always available due to the large number of requests.

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.

At the document preparation station, printed library materials will be entered into the Library Of Congress Information System (LOCIS) database with the document number corresponding to the Library retrieval number. A machine-readable batch index sheet will be generated as the leading page of each identifiable record. Documents are then batched and sent to the input station where they will be scanned and digitized at 300 lines per inch resolution, by a high speed scanner. During this scanning process, the machine-readable batch index sheet is also read, and this indexing information is sent to the Library of Congress indexing data file. The resulting images are temporarily stored on a digital magnetic buffer where the images are inspected for quality and proper sequence. The data on the magnetic disk is then transferred onto digital optical disk. The 12-inch disk to be used has a storage capacity of 10^{10} bits which is equivalent to about 20,000 book pages, 12,500 periodical pages, or typically two to three years of a weekly news magazine.

The Library's central computer system will support user searching via access points in the Library's existing information retrieval system (LOCIS) and will interface with the image storage system to provide access to the text of the articles by means of document control numbers. After the first page is displayed on the CRT, the user will be able to page through the article, using commands similar to existing LOCIS commands which allow similar movement within the bibliographic search files. The user will have the option to command an offline batch printer to print those pages that he wishes to receive in hard copy form.

1.2 Text Access Methods to be Applied

There are three text access methods which will be experimented with in the pilot project. The first two methods represent the two extremes in terms of utility and cost for the majority of the high use periodical materials which will be stored on the system and for which no existing search tools are available on the Library's computer system. The third method is for the small fraction of this material for which the Library's Congressional Research Service has already provided computer

search tools in connection with its Citation file, a data base of citations to current material relating to public policy, among which are citations and brief abstracts of articles in some high-use periodicals.

The first method, which we call the "Bound Volume" approach, is the digital optical disk equivalent to the usual library practice of collecting a year's output of a periodical and binding the issues together into a single physical volume, which volume becomes the item accessed and used by the reader. To use the previous example of the Journal of the ACM, a researcher searching the computer literature would use external reference guides such as H.W. Wilson's Applied Science and Technology Index, literature citations he had noted, personal recollection, or knowledge that a development of interest occurred in a particular year and was reported upon in JACM. In a conventional library he would then request a particular year of JACM, and a bound volume of issues published that year would be delivered to his reader's desk. Upon delivery, he would have available all of the access tools which were a normal part of that bound volume: successive tables of contents with article titles, authors, and corresponding page numbers; possibly an annual index of titles, authors, and in some less likely cases, arrangements by subject area or by permuted title terms, also with issue and page number pointers. In some cases, a succession of illustrated cover pictures with text highlighting feature articles or special topics would also be present and useful.

The digital optical disk equivalent to this approach represents the least cost method in terms of human data preparation work of manual indexing, abstracting or adding information other than that on the printed pages. Only periodical identification, issue and article identification and page number data need be keyed entries in character encoded form.

If a disk having 15 years of the Journal of the ACM were available, its primary point of access would be the journal name and year. This information would be the primary method of selecting a particular disk and confining the page-images accessed to those corresponding to one year of the 15-year collection of the example. Once this had been established, the reader could be given a selection menu to view the annual index if present, successive table of contents, successive covers, successive first pages of articles, unconstrained successive page-images, or a page-image corresponding to a particular page number of the original document. Minimum extra information is added and other than some minor conveniences of direct page access and the avoidance of handling the generally large and heavy bound volumes, the user would be performing the same operations which he would apply to a physical volume of bound issues. But while a significant savings in library shelf space is obtained, this approach offers

no essential advantage or disadvantage to the user of such information over that of using the physical bound volume.

The second method, which we call the "index and abstract" approach, requires a costly intellectual-labor-intensive effort whereby a person familiar with the subject areas covered by the periodical reads the articles and enters index terms and a brief abstract into the computer, and as in the previous case, keys it to the periodical to which it refers, and additionally to the appropriate page numbers of the article. While this produces a quality access tool to the articles on the digital optical disk, its cost exceeds that of scanning and storing the page images of the article.

The third approach is used when a pre-existing search tool exists on the central computer system, as may be the case for some recent articles on public policy such as those indexed and abstracted by LC's Congressional Research Service and searchable on the Library's central computer. In this case, the minimum-required information of the "bound volume" approach is created and a simple flag is placed in the display screen to indicate that a particular article is available on digital optical disk and this flag is made to point to a corresponding disk address.

1.3 Problems With Existing Text Access Methods and Approaches to Improvement

Most library-oriented information retrieval systems emphasize access to books rather than articles in periodicals and tend to follow conventions of cataloging based on long-established traditions of physical card catalogs. Such traditions tend to limit subject access since each additional subject term used required an additional catalog card.

With an entirely electronic system, there are several new approaches of promise which will be examined during the course of this pilot project. One is the linking of the character-encoded reference guides to periodical literature, now becoming available from publishers, to the corresponding printed guides. A second method is computer-based automatic indexing methods which would process the full text of the digitized page images after performing optical character recognition of the page images. The provision of the full text of articles rather than merely bibliographic citations to them is a major improvement by itself, but the automatic extraction of index and abstract information from this same full text is the most needed future improvement.

2.0 THE ANALOG VIDEODISC PROJECT

2.1 Pictorial Materials to be Stored as Video Images

In addition to storing digitally scanned images of printed pages, another aspect

of the pilot project is to convert more than 40,000 images from paper print motion pictures, motion picture stills, prints, photographs and posters, and audio tapes to five laser videodiscs. An additional non-image component of this project is the recording of musical concerts on digital and laser-based audio Compact Discs. Some of the items under consideration to be filmed and incorporated into the videodisc are:

- o 4,000 Cabinet of American Illustration drawings, cartoons, cover designs, sketches for posters and illustrations for magazines, novels and children's books that were executed chiefly between 1880 and 1910 representing approximately 200 artists.

- o 4,000 ink, watercolor, gouache, pencil and crayon by American artists of political and humorous cartoons published in American newspapers and magazines from 1860 to the present.

- o 600 glass lantern slides and approximately 250 black and white photo prints from nitrate negatives.

- o Almost 30,000 glass negatives of the Detroit Publishing Company of views and scenes of the United States, copies of negatives, reproductions and paintings.

- o 58 large format copyright deposit color photoprints made by adding color lithographically to black and white negatives.

- o 5 albums of Detroit Publishing Company mounted photoprints of views of the United States and two foreign countries.

- o 1,687 Farm Security Administration and Office of War Information color slides and transparencies of the United States and Puerto Rico capturing Depression-era America, ca. 1939-1943, which are presently in cold storage to preserve them.

- o 2,250 photoprints of the Middle East presented to the Library by Abdul-Hamid II of views and scenes of the Ottoman Empire, ca. 1893.

- o 2,500 architectural drawings in varied media of the Architectural Drawings Collection.

- o 4,000 political posters from 1965 to 1980, many in Chinese arranged by country.

- o Two television videotape productions.

- o 91,000 motion picture stills.

While preservation is the primary objective of this part of the pilot project, the videodiscs of the non-text collections will provide simultaneous advantages in on-line reference retrieval allowing researchers to survey a large quantity of material without the frequent handling necessitated by manual searches. Because of the fragility of some of the items, such as large posters, glass slides, nitrate film and photographic negatives, they are not now generally accessible to researchers.

2.2 Image Access Methods to be Applied

The Library will investigate the use of the disk in conjunction with a computer-generated captioning method. In this system, the caption identifying the image would be typed into a terminal connected to a microcomputer and stored on a "Mini-Wini", a small magnetic disk using the fixed (e.g., non-removable Winchester disk technology and storing up to 20 megabytes of data.)

Although a data base using an enhanced Library of Congress cataloging format (MARC) is desired it is intended that a separate and abbreviated data base will be established. The user will consult a printed index, determine the pictorial items he wishes to view, key in the pertinent title, creator and date information, and the images of the selected items will appear on the screen for viewing.

One advantage of this configuration of establishing a data base separate from the videodisc is that the information can be updated or changed at any time. Descriptions of the items displayed will be electronically overlaid on the videodisc images.

2.3 Problems With Existing Image Access Methods and Approaches to Improvement

Libraries, by their nature, have tended to emphasize book materials and the cataloging practices common to book materials for all objects in their collections. Just as a chapter gains its importance by being part of a book, so too a particular photograph or a particular work of art has tended to be subjugated to the collection of which it is a part, in terms of cataloging or indexing. Also following book practice, the form, size, and medium of pictorial items has tended to dominate description of content. Effective image retrieval, however, requires identification of and access by the things in the picture. In defense of these historical practices, however, it has not been cost-feasible and is not now cost-feasible to apply the same effort level to the cataloging and indexing of a single pictorial item that is applied to a book of much larger content.

One near term improvement would involve the use of free form descriptions of content, much like those in the catalogs of commercial art galleries and art auction houses, with computer-based content analysis of the full text descriptions. In the longer term, however, it appears that image recognition methods based upon artificial intelligence and scene analysis will become necessary to distinguish the items in a two-dimensional image and their salient characteristics. This is a very broad and pressing problem which is not at all particular to libraries. There are now increasing volumes of digitized images being beamed down

from satellites which must still be manually examined, classified and indexed by human photo-interpreters. We believe that the automated processing of image information is the greatest challenge of the next decade.

3.0 A RETRIEVAL SCENARIO FOR THE LATE 1980's

An important beginning is being made by the Library of Congress' pilot project which we believe is the precursor of a major change in information retrieval systems. By placing primary information items themselves on computer-accessible media, rather than merely references to these information items, a central information retrieval system is potentially made feasible which is both comprehensive and medium-independent. A future researcher should be able to operate a single multi-function terminal and retrieve all items relevant to his topic of interest, independently of the medium that houses these items. Retrieved items would include the full text of relevant monographs, reports and journal articles; relevant pictorial material including photographs, works of art, maps, and motion picture and television documentaries; and the sound of related music, lectures and speeches.

In earlier years, the lack of effective information storage devices was the barrier to such comprehensive information retrieval systems. Today the storage devices exist and the software systems to orchestrate these multiple devices are the only missing links. It is one object of the Library of Congress' pilot project in optical storage to create some of the early links needed.

4.0 REFERENCES

1. Betz, Elizabeth W., Graphic Materials: Rules for Describing Original Items and Historical Collections, Library of Congress, Washington, D.C., 1982, 155p.
2. Betz, Elizabeth W., Subject Headings Used in the Library of Congress Prints and Photographs Division, (Preliminary Edition), Library of Congress, Washington, D.C. April 1980, 479p.
3. Goldstein, Charles M., "Optical Disk Technology and Information," Science, February 12, 1982, pp. 862-869.
4. Nugent, William R., "Application of Digital Optical Disks in Library Preservation and Reference," Proceedings, 1983 National Computer Conference, Anaheim, California, May 16-19, 1983, AFIPS Press, pp. 771-775.

5. Nugent, William R., "Optical Disk Technology," Minutes of the One Hundredth Meeting Association of Research Libraries, Scottsdale, Arizona, May 1972, pp. 62-83.
6. Nugent, William R. and Harding, Jessica R., "Advances in Videodisc Technology for Public Access to Non-Print Collections" to be published Proceedings, First Annual Conference of American Library Association Library and Information Technology Association LITA), September 17-21, 1983.
7. Mole, Dennis, "The Videodisc as a Pilot Project of the Public Archives of Canada," Videodisc/Videotex, Volume 1, Number 3, Summer 1981, pp. 154-161.
8. "Museum of Fine Arts, Boston, Selections From the Permanent Collection," Videodisc Projects Directory, Videodisc/Videotex, Volume 2, Number 1, Winter 1982, pp. 57-58.