Scholarly Information

William Y. Arms

A theme of recent EDUCOM conferences has been the merging of technical areas which have traditionally been separate. The same is becoming true of scholarly information, but universities have been slow to react to the need.

The problem is simple. A student writing a paper, a faculty member preparing a course, or a scholar working on a research project begins by assembling information from many sources. These sources can include libraries, museums, photographic archives, commercial services, computer data bases, personal contacts, and private files. The search may be on-campus or world wide. In some fields of study, assembling information can form the major part of a research project; in others it is an essential building block.

Computing has the potential to improve this process, but requires coordination. Otherwise the various areas will continue to develop services that fulfill parts of the need but do not provide the links that would allow scholars access to all the resources of a modern university.

LIBRARIES

In the field of information, the pioneers have been the libraries. Long ago they realized that merely to collect books was of little value to scholars. Librarianship developed as a profession around the disciplines of cataloging and classification, tools used to give information about library collections.

The principles of librarianship are carefully spelled out in documents such as the *Anglo-American Cataloguing Rules*, and library schools have been established to teach these principles to new librarians. Nobody claims that the classification systems or subject headings are perfect, but they are in widespread use and provide a reasonably effective way to find items in a library.

Scholars often require more information than can be found in an orthodox catalog. Secondary information services exist to fill this need. These provide information—titles, keywords, or abstracts—about individual journal articles. Most secondary services are discipline-specific. Some are huge. For instance, *Index Medicus, Chemical Abstracts,* and *Lexis* cover the entire fields of medicine, chemistry, and law respectively. Others are tiny.

When library computing developed in the early 1970s, two major success stories were shared cataloguing and on-line computer searching of secondary information services.

Shared Cataloguing

Cataloguing a book accurately is a skilled and time-consuming task. Since many libraries acquire the same books, it is sensible for libraries to share their records with each other. This is not an easy computing problem. Bibliographic data is extremely subtle, and an effective shared cataloguing system requires an enormous number of terminals to use a very large bibliographic data base. The pioneer in this area was OCLC under the direction of Fred Kilgour. OCLC has been followed by a number of other systems, most notably the Research Libraries Group based at Stanford University.

William Y. Arms is vice provost for computing and planning, Dartmouth College, Hanover, New Hampshire.
Reprinted with permission from the EDUCOM Bulletin, *v.18, no.3/4, Fall/Winter 1983, p.21–23. The EDUCOM offices are located at P.O. Box 364, Carter and Rosedale Rds., Princeton, NJ 08540.*

OCLC was able to build on earlier work by the Library of Congress and the British National Bibliography in establishing an international format for exchanging catalog records between computer systems. This format, known as MARC, is supported by all major cataloguing services. Dartmouth was an early member of both OCLC and the Research Libraries Group. Over the past ten years shared cataloguing has allowed the library to improve the quality of its cataloguing and build up a large machine-readable data base despite the recent budget pressures.

Information Retrieval Services

Large secondary information services produce so much material that searching them becomes a major problem. In this field the computer pioneer was the National Library of Medicine. The library had an early computer system to assemble the numerous items for printing in *Index Medicus*. As a result, the entire text was available on magnetic tape. The earliest medical search system, Medlars I, was a batch processing system which searched these tapes to find articles that matched specified search profiles.

When this concept spread to other disciplines, two requirements emerged. The first was a demand for online searching. The second was "standard procedure" for users. Secondary information services use a wide variety of approaches; indeed, the disciplines they serve are so diverse it is difficult to envisage any single standard satisfying them all. Yet it is important for library staff to be able to use them with a minimum of training.

Several commercial companies provide libraries with on-line searching of secondary information. The first was Lockheed, with the system now known as Dialog, followed by SDC and BRS. These companies acquire data bases from many sources, mount them on-line, and provide a standard search procedure. This is a competitive business and the companies use advanced methods for storing and searching huge data bases, including free text searching.

These two major achievements are now converging. Libraries are beginning to replace local card catalogs with on-line computer systems. These use both the MARC records produced through shared cataloguing and the methods of data base searching developed by the various bibliographic services. At Dartmouth, the Pew Foundation provided funds to load the MARC records developed on OCLC and Research Libraries Group computers onto a duplicate of the BRS search system. This was a convenient way to provide a generally available on-line catalog.

NON-BOOK MATERIALS

The success of library computing has led to extensions in a variety of areas. Some of these are traditionally housed within the university library; examples are maps and manuscripts. Others, such as artifacts and paintings, are likely to be found in the university museum. Some areas, such as films and photographs, have a variety of homes in different universities. Collectively these are sometimes called "non-book materials".

For a number of reasons computing progress has been slower in these areas than in libraries. One reason is that most of the materials are resources for the humanities, usually less well funded than the sciences. In addition scholars in the humanities have been less familiar with computing than their colleagues in the quantitative disciplines. Another difficulty is that library automation has made its contributions in sharing information about items that are held by many libraries; most manuscripts, paintings, and museum objectives are unique. Finally, no widely accepted standards exist for cataloguing and classifying most scholarly materials other than books and journals.

Despite these difficulties, numerous attempts have been made to develop information systems for museums and other non-book materials. Funding has been limited, but still much useful work has been accomplished.

Recently this work has received a champion in the J. Paul Getty Trust. The Trust has the prestige to coordinate many areas and the long-term funding to tackle some of the underlying problems. The Trust has projects in a wide variety of fields. One is to build a computer catalog of the collections of a group of museums and galleries, beginning with paintings. This will include several major national museums

and two universities, Dartmouth and Princeton. Another project is to catalog several enormous photographic archives.

Both these areas require subject indexes of visual objects such as paintings, vases, and architectural sites. This topic, known as iconography, is extremely complex with no established standards, yet is essential for success in these disciplines. Many of the finest collections are in Europe, which adds the complications of foreign languages and latent chauvinism.

DATA ARCHIVES

The discussion so far has been of computer systems that provide information about traditional scholarly materials such as books or paintings. In other fields, the information is more closely linked to the computer. Data archives were an early case.

Perhaps the best example of a data archive is the U.S. Census; in fact, the Hollerith punched card was originally developed to tabulate census data. More recent censuses have released raw data on magnetic tape. This data is invaluable for studies in several social sciences, but extracting information from hundreds of reels of tape is so tedious that for the most recent census each state has set up a dissemination bureau and several universities have provided their own services. The cost of such service is so great that even universities the size of Harvard and MIT have found it cheaper to work together.

Several universities, most notably in Michigan, have centers whose task is to gather data archives and make them available for research. Project Impress at Dartmouth College, developed during the early 1970s, was a data base system for teaching students how to analyze such data archives, a large number of which are stored on-line. The value of Impress lies in the combination of data archives and good quality search software.

COMMERCIAL DATA BASES

Some academic disciplines use data bases from the commercial sector. These are varied both in quality and scope and have two types of origin. Some, such as the news services, began life as information services used internally by an organization which realized that outsiders would pay for access. Others, such as the services giving information to financial investors, are aimed at specific groups of professionals. By academic standards all these services are extremely expensive.

An interesting experiment in this area is The Source. This commercial company licenses a range of commercial data bases and mounts them on its own time-shared computers. A more or less standard user interface is provided so subscribers can teach a variety of information with minimal training. The Source, in its present form, is of marginal use to scholars, but in five years time such services may mature into more usable form.

COMPUTING INFORMATION

For many years librarians have been asking computing specialists for assistance. Unfortunately, assistance has not been forthcoming. The computing systems of our universities have become enormous collections of poorly indexed tools and resources. In the days that computing was restricted to a few specialists this was not important. When computer users were concentrated into terminal clusters, with many users sitting side by side, word of mouth was still an effective way of disseminating information. Now that computing has become widely distributed across campus, some better way is needed for scholars to learn of the riches at their fingertips.

Dartmouth, as the first university to place emphasis on universal computing, developed a set of indexes that were suitable for a single large time-sharing system. These include an enormous collection of files which can be read either with system commands or from within programs. In addition there are indexes to library programs and publications. Although few universities can rival the completeness of information available at Dartmouth, the system is still far from perfect. One problem is that many of the most useful programs are unknown to central staff. They are in departmental libraries or even in personal catalogs. Another problem is the variety of computer systems. A

user of Dartmouth College Time Sharing may be unaware of a program that runs under the UNIX System.* Finally, computing is always changing. As services are introduced or withdrawn, keeping information up-to-date is a perpetual problem.

INTEGRATION

The word integration is much used in computing, but rarely defined, and even more rarely achieved. Each supplier of scholarly information has a different vision of how to integrate specific areas.

For example, libraries want to integrate their internal data processing, their services to scholars, and their links to other libraries. The aim is for a single description of each item to be used by all library systems.

A scholar has a different set of objectives. A faculty member or student using a library catalog through an on-line terminal is not interested in how smoothly that catalog fits with other data processing carried out by the library. However, after finding a reference in the catalog, the scholar demands follow-up services such as being able to copy the reference into a personal bibliography or word processor. At Dartmouth this problem has been partially solved by the fortunate accident of having a catalog system that runs under the UNIX system. UNIX is primarily an academic operating system and works well with other computers used for teaching and research.

The scholarly information system of the future will have the university providing central coordination of a variety of independent suppliers of information. These suppliers can be large or small, on-campus or off-campus. Since many of these suppliers will not be under the direct control of the university, providing smooth access to them all is not easy. Key aspects of this information system will be:

Quality Control

The university must identify major sources of information and ensure that information provided is accurate and current.

Terminals

A major assumption of computer planning for universities is that within a few years almost all scholars will have a small computer on their desks. One use of such a small computer is as a terminal to larger computers functioning as major information sources. The university must standardize a small number of different types of personal computers.

Communications

Computer planning for universities assumes the existence of campus networks, but Dartmouth is one of the few to have such a network in place. Any terminal or personal computer connected to the network has equal access to all computers on the network and is also able to make off-campus connections using services such as Telenet.

Currently, almost all information services are designed around low-speed serial communications. The future is likely to require much higher capacities, either digital or video, so images or complete documents can be transmitted.

User Interface

Since each information source is likely to have a different user interface, the only way to provide integrated service to the scholar is for the personal computer to translate procedures used by the various sources into some homogeneous user interface.

Today most information services assume that the service is being used directly by a human, either a scholar or a supporting professional. In the future, the user is more likely to be another computer. This requires agreement on application protocols.

Long-term Planning

New technology and new sources of information are going to become available

*UNIX is a trademark of Bell Laboratories.

continuously throughout the next decade. The university must watch these developments, anticipate some, and consciously decide to ignore others.

Each of these areas require standards. One of the most valuable services a university can provide is an acceptable set of standards for computing and information. The difficulty is finding a balance between overstandardization, which restricts flexibility, and the chaos that results when there are no standards.

CONCLUSION

Scholarly information is too big a topic for universities to ignore. Moreover it has so many ramifications that leaving its planning to the library, or worse still the computer center, is unlikely to provide good balance. The only sensible solution is a coordinated plan in which many parts of the university work toward the common goal of providing faculty and students with the information they need for study and research.