N94-33821

Issues for Bringing Digital Libraries into Public Use

David W. Flater
University of Maryland Baltimore County
Baltimore, MD 21228 U.S.A.
Phone: 410-455-3000 Fax: 410-455-3969 davidf@cs.umbc.edu

Yelena Yesha
University of Maryland Baltimore County
Center of Excellence in Space Data and Information Sciences
NASA Goddard Space Flight Center
Greenbelt, MD 20771 U.S.A.
Phone: 410-455-3000 Fax: 410-455-3969 yelena@cesdis1.gsfc.nasa.gov

Introduction

The U.S. Government has recently begun a renewed effort to support the research and development of digital libraries. This happened as part of a larger package intended to hasten the evolution of a new and improved information infrastructure. The project is intended to benefit everyone to some extent, not just the institutions that currently have wide-area network access[1]. The questions facing us now are therefore not of an exclusively technical nature. People hearing the phrase "digital library" are bound to worry about things like the relative availability of romance novels when the public library on the corner decides to go electronic.

Such worries are not entirely unfounded, and we will attempt to address them in this paper. We will also try to answer the questions, What exactly *is* a digital library? How long have they been around? What are they likely to become, and what has that got to do with romance novels, anyway?

Digital Libraries Today

Depending on what sort of digital library is being discussed, one may claim that they already exist or that they could not exist for at least ten years. To some, a digital library is any collection of electronic books. This phrase can be taken loosely to include any kind of structured data, so that existing archives and even databases could be construed as digital libraries. However, in the future, the phrase electronic book is likely to refer to a particular standardized encoding for arbitrary collections of data. Electronic books, or e-books, can encompass quite a lot more than what is currently thought of as a book. Standards already exist for how to encode pictures, sound, and movies for storage, retrieval, and playback on a computer. Furthermore, e-books can be interactive. An e-book can give you all of these elements together in something reminiscent of a video game, it can be the ASCII text of the world's most boring stereo instructions, or it can be any combination, however well orchestrated or poorly pulled-off. Today's personal computers already have the processing and storage capacity to support full-fledged e-books, let alone today's low-cost workstations. What remains for the future, some would say, is to make use of these resources with digital libraries. Digital libraries are destined to be the first true archives of *integrated* multimedia data. Today we have corpora of text, image, and sound files; soon we will have corpora of electronic books which contain all these things in a single organized package.

To others, a digital library is not just any collection of e-books, but a facility which makes use of a limitless number of resources which are accessed through a wide-area network. The focus is not upon individual "books," but on the data repositories which provide them. The digital library as such would be the unification of all the available repositories through a single interface. That interface would naturally be able to cope with whatever kind of data, multimedia or otherwise, is retrieved. While it might seem as if the production of multimedia

e-books would be the greatest challenge, that technology already exists. The intelligent management of networked resources is what remains a difficult topic for research. This research has been going on for some time in an attempt to manage the already existing repositories of data which are public on the Internet. With the new initiative to build a better information infrastructure, it has become even more important that we improve the ways in which we manage, locate, and retrieve information over wide-area networks.

The above two "definitions" of digital libraries are only mutually exclusive in the sense that they both lay claim to the same buzzword. There is nothing really preventing those technologies from being integrated. Individual "digital libraries" will be needed to provide e-books to the world, and we will need some way of providing access to those libraries that allows users to find the repositories they need and to access them without undue trouble. Although the latter topic is more interesting from a research perspective, the former is more important to ordinary people who are more interested in what digital libraries can do for them than in how it is accomplished.

Digital libraries in the Future

It is probably a safe prediction that many serious efforts will be made to build digital libraries. A lot of money will be spent on getting books and other items onto computer media, on the physical media themselves, and on the software and hardware required to read these electronic books. However, after this startup period, very little of the cost of a real life digital library will have to do with the cost of building or maintaining the library itself. The real cost will be licensing fees.

Most proponents of digital libraries want to see them benefit everyday people, not just academicians and engineers. In order for digital libraries to be a popular success, they must contain popular e-books. In order for publishers to provide e-books, they must make a comfortable profit. Therefore, we may predict that if popular entertainment shifts into the realm of digital libraries, it will be after sufficient propagation controls and billing procedures have been put in place to insure that whoever has to pay for the e-books will pay for all those who make illicit copies as well. With sufficient effort, the number of illicit copies could be small indeed. The propagation controls could be as draconian as any which have ever been applied to computer software, and all the more sinister for having had so many years to be perfected. It could be insured that a customer has the use of a data object exactly *once* with a sufficiently encrypted data object and an obfuscated, self-destructing piece of software which is necessary to make use of the data. The software could insist on contacting an authentication server maintained by the publisher before executing, so that copying the software itself would achieve nothing. First-run movies and premium sports broadcasts would probably be handled this way, and the companies providing the copy protection stand to make as much money as their clients.

On a more pedestrian level, publishers will want to sell copies of more ordinary e-books to people and hope that the cost of blank media will be sufficient to discourage truly rampant piracy. When digital phone service becomes widely available in the U.S., a new species of 900 numbers will appear. Small computer companies, always in search of a niche with profit potential, will offer access to well-maintained and well-stocked digital libraries. These libraries will offer reasonably new movies, music, periodicals, and books, similar to today's video, record, and book stores. Customers may pay a specific price for each item they retrieve, a per-kilobyte charge, or both, but the objects retrieved will not be subject to draconian controls and customers will be able to use them as much as they please. A large portion of the proceeds from these operations will be returned to the publishers in the form of licensing fees.

Digital libraries owned by educational institutions and research companies, containing specialized scientific and professional e-books, will probably be more open to the public than such libraries are today, but few of these institutions will miss the opportunity to charge for access. In fact, they may be forced to charge just to cover the licensing fees for the e-books, which, like the subscription rates for today's scientific journals, will be high in order to compensate for their relatively small circulation. In the present time, it is already the case that

institutions are charging for access to databases which merely catalog the titles, authors, and abstracts of current periodical literature. A researcher who does not have access to such a database is at a noticeable disadvantage when trying to find related work. The free access databases cannot afford to be as current or as complete as the pay-databases. In the future, the situation will be the same with digital libraries of a technical nature, and those without access will be a step behind in their research.

Public libraries will presumably retain their traditional, underfunded role after they become digital and concentrate on whatever they can afford to provide -- maybe romance novels, maybe shareware e-books. Most government funding will go to educational institutions and to public libraries, and that funding which goes to public libraries will be a pittance compared to what the public at large will spend on premium entertainment through private companies.

It may sadden some to believe that quality news and entertainment will be just as expensive in the future as they are today due to licensing fees, but let us consider the alternative. Suppose that public digital libraries, funded by the government, were the only game in town. Let us even suppose that they were *well* funded and could afford to house whatever they chose. Having spent federal funds on the deployment of a digital library, the government will set standards on what sorts of e-books are worth procuring. We may boldly predict that this will include mostly educational material, but with a large section of romance novels. Without independent digital libraries, e-book publishers will have no incentive to produce e-books other than tutorials and romance novels. Romance novels would effectively become the only officially sanctioned form of entertainment.

Let us give thanks, then, for the independent retailers and for the commercialized publishers that bring us so many choices in what to read. Indeed, e-books could take commercialism to previously inconceivable levels of shallowness as advertisers cash in on the element of future shock. Imagine reading a newspaper in standard black print on a white background, turning the page, and being assaulted by a full color, moving, speaking advertisement with CD-quality Surround Sound [1] blasting subliminal messages from seven different directions. Had enough? Not yet: it could be interactive. An advertisement could *argue* with you until it gets your credit card number. If there is any vestige of human compassion left in corporate society, it will be possible to dismiss these solicitations, or at least to iconify them until they give up and go away.

Help Wanted

Bringing digital libraries to life will require us to make significant progress in the following areas:

- Integration of multimedia data. An extensible standard must be developed for the seamless integration of multimedia data objects into e-books. Hawking proprietary formats will only lessen the chances of long-term success.

- Licensing and copyright issues. While the digital library medium must be open to the public, the contents of e-books must be protected in law and in practice from unauthorized use and duplication. Researchers are still thinking of new problems in this area as they seek to solve the old ones[2,3].

- Wide-area networking. While our desktop computers have enormous capacity for processing and storing information, our wide-area networks have limited bandwidth and accessibility. Public use of digital libraries will require much greater bandwidth. Upgrading to faster network technology will help, but we must also learn to make more efficient use of network resources lest we again use up all the available bandwidth. Alibi[4] is a software system now in development to help us achieve this goal.

[1] Dolby Surround is a registered trademark of Dolby Laboratories.

- Resource discovery. Users must be able to locate the digital libraries which would be useful to them. A number of systems are already in place to support resource discovery through browsing or through cataloging of network resources. Alibi uses a more flexible method for finding resources which may prove to be more useful in the long run. Knowbots[1] are another strategy which is being investigated for resource discovery and information retrieval in digital libraries.

- Information retrieval. How are we to keep track of all our e-books and find the ones that users want? After years of research into ways of retrieving textual documents, we are still not able to automate the process of accurately cataloging arbitrary documents for later retrieval. Research into image retrieval is relatively young, and sound retrieval may be unresearched. Presumably publishers of e-books would be willing to provide catalog entries for them if we could agree on a format for the catalog as well. This is a complicated problem in itself, since the catalog must contain enough data to be useful but not enough to overwhelm the search. There is great potential for innovation in the development of algorithms for retrieving e-books.

Conclusion

In much the same way that the field of artificial intelligence produced a cult which fervently believed that computers would soon think like human beings, the existence of electronic books has resurrected the paperless society as a utopian vision to some, an apocalyptic horror to others[5]. In this essay we have attempted to provide realistic notions of what digital libraries are likely to become *if* they are a popular success. E-books are capable of subsuming most of the media we use today and have the potential for added functionality by being interactive. The environmental impact of having millions more computers will be offset to some degree, perhaps even exceeded, by the fact that televisions, stereos, VCRs, CD players, newspapers, magazines, and books will become part of the computer system or be made redundant. On the whole, large-scale use of digital libraries is likely to be a winning proposition.

Whether or not this comes to pass depends on the directions taken by today's researchers and software developers. By involving the public, the effort being put into digital libraries can be leveraged into something which is big enough to make a real change for the better. If digital libraries remain the exclusive property of government, universities, and large research firms, then large parts of the world will remain without digital libraries for years to come, just as they have remained without digital phone service for far too long. If software companies try to scuttle the project by patenting crucial algorithms and using proprietary data formats, all of us will suffer. Let us reverse the errors of the past and create a truly open digital library system.

References

[1] R. E. Kahn and V. G. Cerf. The Digital Library Project, Volume 1: The World of Knowbots (DRAFT). March 1988.

[2] J. R. Garrett and J. S. Alen. Toward a Copyright Management System for Digital Libraries. 1991.

[3] Workshop on the Protection of Intellectual Property Rights in a Digital Library System. May 1989.

[4] D. W. Flater and Y. Yesha. An Efficient Management of Read-Only Data in a Distributed Information System. International Journal of Intelligent and Cooperative Information Systems, Special issue on Information and Knowledge Management. To appear, 1993.

[5] Discussions seen on comp.infosystems and related newsgroups, June 1993.