

## The Future of Digital Libraries

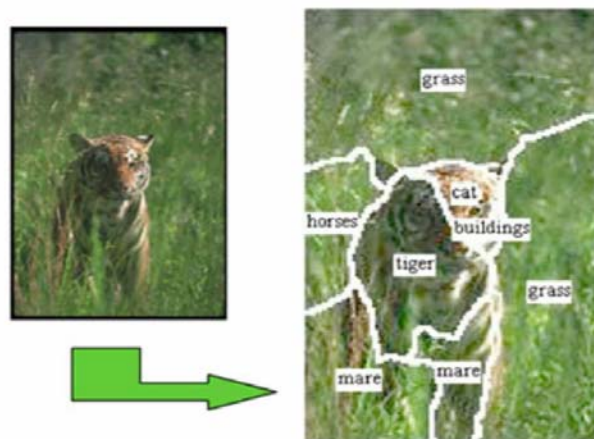
Michael Lesk, Rutgers University

Digital libraries research has produced a product everybody uses: Google. Google came out of the DL group at Stanford University, and is now doing 250 million searches per day. Ask any professor where undergraduates do their research: the answer is overwhelmingly the web, with paper libraries running way behind. Digital libraries increase collaborations with researchers throughout the sciences and beyond: there are projects jointly with scholars of art, music, archeology, engineering, history, astronomy, and so on. The focus on content in DL research produces, almost automatically, multi-disciplinary projects. It enables a great many institutions to participate: almost every university library, for example, has some unique special collection or area of expertise, which can be the base of a research and service opportunity.

Over the last ten years there has been enormous progress. We now know how to convert not just the traditional books, pictures, sounds and video to digital form, but also fossils, buildings, and sculptures. Text searching is now used effectively every day by millions, while research is active on searching 2-D images and sound recordings, both music and voice. The current research frontier in searching and organizing is in 3-D images and in the combinations of techniques needed to search video.

Some examples of research projects in the area are mentioned below, most NSF funded but some funded elsewhere.

- 1 Image searching. Jitendra Malik and David Forsyth at Berkeley are building systems that can learn to attach text labels to images by analyzing shape and color (NSF and other funding). Image searching applications like face recognition are now important for national security purposes, including the analysis of aerial



photographs.

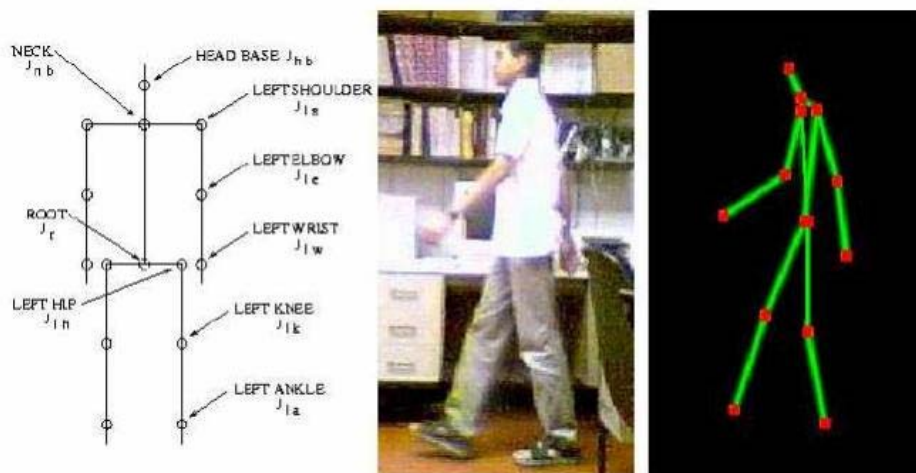
- 1 Modeling historic buildings. For example, Columbia University, using NSF funding, has laser scanned Beauvais Cathedral to get an accurate 3-D model so that structural engineering programs can be used to figure out how to keep it from

falling down again. In the EU, the synagogue at Wiesbaden (destroyed in 1939) has been "virtually" rebuilt. The Mellon Foundation supports an international effort (with US technology) to bring together digitally the images and manuscripts



from the Dunhuang Caves in western China.

- 1 3-D searching. Tom Funkhouser of Princeton (NSF funding again) is doing geometrical methods for 3-D shape searching. If these techniques can be made to work, they are of enormous importance for drug design. Some other remarkable 3-D modeling projects include Jezekeil Ben-Arie's studies of human motion and Tim Rowe's models of the internal structure of fossils, made by taking CT scans of fossils with an industrial high-power scanner. Human motion is now being suggested as a way of identifying people for security applications and the fossil



scans have detected forgeries and been extremely valuable in education.

- 1 The Million Book Project. Raj Reddy with the cooperation of the Government of India expects to scan and put online viewable and searchable versions of one

million books (NSF funding plus major outside matching). Some 20,000 books have been done so far, and another 150,000 are in the works; the rest should be finished over the next 3 years.

- 2 The National Gallery of the Spoken Word. Mark Kornbluh and others, with NSF support, have sound recordings of every President since Grover Cleveland, fifty years of interviews by Studs Terkel, and Supreme Court hearings, among other gems of historical voices.

So why isn't everything that we ever want to read, see or hear online? What remains for research support?

### **a) Data resources**

Support of "data curation" - the care and maintenance needed by the data in digital libraries. Just as research on computer networks had to be supplemented with NSF support of the costs of the actual network service, research on digital libraries produces a need for longer-term support of the data collected, particularly during the time that the library system needs to run duplicate paper and electronic support systems.

### **b) Economics.**

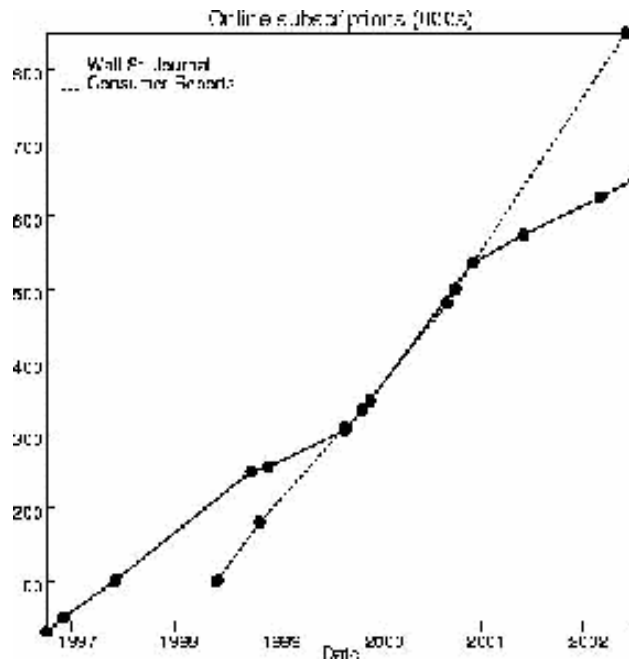
We know how to do large scale digitizations but we're still groping in the dark for how to pay for them. The standard solution so far is site licenses to university libraries. The good part of this is

- ✍ minimal administration: the users don't have to track and pay item by item, and there are only a few big customers for the supplier to deal with
- ✍ fairly sophisticated local support, so the suppliers don't have to worry as much about that
- ✍ retaining the libraries in the role of informations supplier, so that paper and electronic services can be kept in a complementary mode
- ✍ encouragement of libraries to find attractive things and scan them
- ✍ no charges per item, so that readers are not discouraged from using the library.

The bad parts are

- ✍ no access for individuals, e.g. the home genealogists unable to get at JSTOR
- ✍ limitations on use imposed by the license agreements
- ✍ incompatible and different interfaces
- ✍ no new money entering the system either from individuals or university departments, at a time when library budgets are under pressure

Are there any other economic models that might help? We thought the music industry would be in the lead, trying out alternate pricing models; that hasn't happened. Few publications sell successfully to individuals; the Wall St. Journal and Consumer Reports stand out.



## b) Copyright.

Some copyright issues are bundled with economics, but a large part of the problem perceived by the libraries is our inability to get cheap permissions to scan and put online materials that are obviously of no commercial value. The copyright office is considering something to do about "orphaned" materials (out of print but not out of copyright). Some kind of administrative compulsory license "orphaned" books, music and movies would help a great deal. This would be similar to the Harry Fox Agency for "covers" of recorded music, although it is likely that payments would be made to a society rather than to individuals. Note that this would be paying revenue, albeit probably small, to creators that they do not get now.

We also need some understanding of how to deal with items that are not labeled with any producer or date and whose copyright status and permissions can be extremely complex. These form a different category of "orphaned" works but are also generally of no commercial value. Again, some kind of compulsory license system with fees paid to an authors/composers society would be best.

There is a danger of complete loss of material which is produced but kept under the control of the publisher. Often it is insisted that such material has not been "published" (since that would invoke a requirement to deposit a copy with the Library of Congress) and also that it has not been "sold" (since that would give the purchaser the traditional rights of first sale, such as the right to sell the copy onwards to some used book dealer). Instead the publisher says there is only a "license" to access the material from some publisher website or to decrypt some kind of protected medium. If the publisher goes bankrupt, or just decides that the number of licenses sold no longer justifies maintaining the operation, the material can disappear. If it was protected by strong cryptography, there is no practical way to get at it without the cooperation of the publisher; merely getting permission will not be good enough. We should have a copyright law that requires legal deposit of a clear-text version of such

material.

### **c) Scientific data.**

We'd like to have people use data as easily as they now use text. Only ten years ago it was generally believed that people had to be experts to do full text searching; library schools gave semester-length courses in it and we thought competent searching meant taking such a course. Now we have everyone using Google without any training at all, and getting results they consider satisfactory. What will it take to make database courses unnecessary?

There have been several research projects in data visualization, but not enough about interfaces. There are serious problems trying to understand the user's knowledge. After all, is anybody other than a professional chemist going to want direct access to a database of infrared spectra? But just how much computing and chemistry should be required for the interface? We don't know yet.

There are data questions that ordinary people ask all the time ("list all motels in zip code 94123 under \$100/night with vacancies next Sat-Wed"). Some of the problems answering such questions are tangled with economics; when I wrote the preceding question I didn't really want an extra clause "whose management has paid a fee to this website").

Among scientific areas where digital data libraries are already making a huge difference are molecular biology (where the Protein and Genome Data Banks have enabled a major shift from wet-lab experiments to data lookup), astronomy (thanks to the Sloan Digital Sky Survey and the National Virtual Observatory), and earth sciences (the IRIS seismic data consortium is one of many applications of geosciences data, with earth observations pouring down by the terabyte from NASA's satellites).

### **d) Computational humanities.**

Working with humanities scholars is a major effort in the digital libraries research program, along with scientific and medical collaborations. It has been one of the most encouraging areas both for extended collaborations and new technologies. In how many other areas of computer science do practitioners work with literary scholars, architects, historians, and art critics? One of the best examples of the impact of digital libraries on education is Greg Crane's "Perseus" project of ancient Greek literature and culture, and a remarkable mixing of animation, theatre history, and computing is the "Virtual Vaudeville" project of the University of Georgia.

New technologies get developed out of the needs found when humanities data are analyzed. For example, new image processing techniques have been found in the University of Kentucky efforts to make more legible damaged manuscripts (such as the Beowulf manuscript, damaged in an 18th century fire). New clustering and display methods were found by Bruce Schatz and Hsinchun Chen working with the medical literature.

Humanities computing can also introduce us to entirely new problems, such as the

need for multi-lingual or multi-cultural collaborations. An example of new interface needs is the work on the International Children's Digital Library at the University of Maryland, where screen designs are being built to help children who can not read select picture books to flip through.

#### **e) Interface techniques.**

Surprisingly, most digital library interfaces and Web search engines still act as if the typical user didn't have a graphics terminal, but only some kind of "glass teletype". Why can't we have better interfaces, making use of graphical displays? Some systems (look at Jim Gray's "Terraserver", or Ben Shneiderman's work) are doing work in this area, but not enough. As an example of an interesting interface to a very complex



scientific database, look at the SkyServer from [fnal.gov](http://fnal.gov).

#### **f) Software libraries.**

In a wry example of "the shoemaker's children go barefoot" there are relatively few digital libraries of computer software itself ((Netlib is one succesful example). Why not? What kind of organizational techniques or searching techniques will work for software? Since we can parse and understand the semantics of code, why can't we do better than for English at enabling search and retrieval of programs?

This has been only a brief overview of the extent of both progress and prospects in digital libraries. There are many other important problems and opportunities, but we have to have some focus.