Experiments on Access to Digital Libraries: How can Images and Text be Used Together?

Michael Lesk
Bellcore
Morristown, NJ 07960
lesk@bellcore.com

Abstract

Should digital libraries be based on image or text display? Which will serve users better? Experience and experiments show that users can employ either, and that there are technical advantages to each format. Often, material in both formats can be used together, and the long-run trend is probably towards Ascii material, even if reached by a circuitous path via images and OCR during a transition.

1 Introduction

Since Vannevar Bush's original 1945 article, many have dreamed of desktop access to a great library[Bus45]. But of what technology should it be built? Bush originally proposed bar-coded microfilm, an analog storage medium. Today everyone believes that such a library should be digital (including Vannevar Bush himself [Bus67]). But what should the digital information represent? Many like images of pages which look like the current printed journals; others think that we should be storing information in Ascii, with considerable descriptive material along with the text. The image standard format is Tiff with Group IV compression; a complex page fits in about 100K bytes in this form. The text standard format is SGML syntax, with with semantics taken from one of several options, e.g. the AAP (American Association of Publishers) Electronic Manuscript Standard[AAP87] or the Text Encoding Initiative standard.

Among the efforts which deliver images are the ADONIS project,[Ste90] which distributes images of medical journals; the University Microfilms CD-ROMS of IEEE journals; the TULIP project[McK93], involving Elsevier materials science journals and several universities; and the Red Sage project of AT&T, UCSF, and Springer-Verlag[HOS+93], working with biomedical journals. The latter two projects use optical character recognition in order to produce a searchable text. All of these projects involve commercial publishers, and the first two are commercial projects, in which large sets of pages are sold to libraries. The AT&T project is also producing a commercial venture, RightPages,[SOF+92] which is to be announced in early 1995.

Many other projects, of course, are based on ascii text. Commercial vendors have sold full-text of newspapers and legal decisions for more than a decade, but these systems do not involve graphics or illustrations. Systems are now appearing that unite Ascii text with the ability to display pictures or even multimedia. Several commercial ventures of this sort are offered by OCLC, including online journals of Current Clinical Trials (AAAS) and Electronic Letters (IEE); MIT Press is going to offer the Chicago Journal of Theoretical Computer Science in electronic form, and there are also smaller ventures such as Matrix News. Many commercial full-text databases are sold on CD-ROM as well: these include the Oxford English Dictionary, or the Chadwyck-Healey compendium of English poetry.

There are also of course a great many non-commercial projects, including text projects such as Project Gutenberg, and the Oxford Text Archive, vast numbers of bulletin boards including some that are refereed and published by permanent organizations (such

Proceedings of the 20th VLDB Conference
Santiago, Chile, 1994

as the online journals *Postmodern Culture* and *Psycoloquy*). There are also many systems available on Gopher and Mosaic[Val93], including image databases as well as text; for example, the Library of Congress is scanning its exhibits and making the images available.

Many other projects can not be described as either text or image, since they are really numerical data of various sorts. NASA, for example, maintains an enormous database of both earth and space observations; the ICPSR database of social science surveys at the University of Michigan and cooperating universities runs to several terabytes; and there are other large files such as census data, digital mapping, and all sorts of government economic statistics.

1.1 Technical Comparisons

The advantages of Ascii systems includes ease of search, compact storage, minimal network traffic and thus fast response, and more easily read displays. The equipment to access Ascii systems can be cheaper and less powerful since high resolution displays are not essential. The advantages of image systems include a familiar display, simpler software, easy use with any kind of journal, and ease of making printouts.

Image systems consume more computer resources, but are easier to code. To consider the use of resources, look at a single page. In image format, this will be between 50KB and 100KB, depending on the density of print and the size of the paper. In text format, it will be about 2KB to 5KB, plus whatever space is needed to hold the graphical illustrations in what may be some kind of image format. With the American Chemical Society journals, for example, which are perhaps 1/4 illustrative material measuring by square inches of paper, the space required for the figures is two to four times the space for the text (depending on the resolution; the higher estimate is for figures stored at 300 dpi). Much more serious than the bulk of the material is the amount of time required to display it on the screen. For example, a full page bitmap at 300 dpi is about 1 MB; not only will this take perhaps five seconds to move across a typical LAN, but it may take as much or more to get it onto the screen in some kind of viewing window. Text can be displayed much more rapidly. The delays involved in images can be ameliorated by providing lower resolution, since 300 dpi is really too big to read on today's 72 dpi screens anyway, but the basic problem remains.

In coding the advantage is the other way. There is considerable agreement on the format of images, where this is not such agreement on text, so considerable effort gets spent adjusting to the format of each new publisher and perhaps even each new book. Searching systems for text may have to deal with complex material (should users be able to search strings appearing in tables? in equations? in figures?) and may have to deal with local alterations to the search rules (for example, in chemical material, the word "I" may mean iodine more frequently than it is the first person pronoun, and thus should not be on a stop list). The ability to highlight items, adjust the format, and provide other aids, means that text systems tend to become more complex, and thus both harder to program and harder to use.

It may seem that this is silly: the advantage of flexibility in a textual system becomes a disadvantage in implementation, and perhaps as we develop our expertise in human-computer interfacing these problems will disappear, but at present they are still there. Again, one can in principle imitate a image system with an Ascii-based system: one could simulate the typesetting software of the original publication and produce pages for display that look just like the printed journal. Imitating the original journal in this way has in fact been done by OCLC in their Graf-Text project[HC86] and by Peter Goldie with the *Journal of Biological Chemistry*. However, few Ascii-based projects have taken this route, preferring to try to use the flexibility of text based systems to provide adaptive or tailored displays.

Similarly, conversion costs are often less for images. Scanning pages costs only a few cents a page, whereas rekeying or even re-editing (typically to insert meaningful tags on each item) can run up to a few dollars a page. Even scanning an old book, too fragile for the use of a sheet feeder on its pages, costs only about 50 dollars, based on the experiments of the CLASS project[Ken93]. To give an idea how low a cost this really is, both the United Kingdom and France are building new national libraries and in both cases the cost of the building is less than it would have cost to scan the books in it. A description of the problems involved in scanning more complex material is in given by Robinson's recent monograph[Rob93].

For all these reasons, projects which are doing joint preparation of both paper and electronic versions of the same material are often scanning, as are of course projects dealing with conversion of older material (except in the area of literary text). Projects which are preparing entirely new material which is only distributed electronically, on the other hand, normally use some kind of Ascii format.

2 Systems

It is possible to combine image and ascii systems in a variety of ways. There are many examples of multimedia systems. Just as a possibility, Figure 1 shows a sample screen from a program built as a demonstration for a project involving a computer archive of the poetry of W. B. Yeats. It combines pictures (in this case one of Yeats himself and one of an illustration in an edition controlled by the Yeats family; it also has sound recordings of two of his poems; and then the Ascii text of a few of the poems in a variety of versions. Part of the menu to select versions of *The Rose Tree* is shown along with two versions of that poem. Mosaic, of course, is full of multimedia pages including animations and sound as well as just text and pictures.

More interesting is the ability to combine image and ascii of exactly the same material. For example, Figure 2 shows a small piece of the British Library general catalog in which a search has been done in Ascii, in the bottom left window, and then the result is also shown as an image of the original catalog page, with its complex and faded typography. A supplementary window on the right shows the entry from the *Dictionary of National Biography* for the author queried. One reason this makes sense is that the format of the printed catalog is familiar to a set of users who have been making intensive use of this particular catalog, in some cases for decades.

2.1 The CORE Project

Recently Cornell has been leading a project which has provided both Ascii and image interfaces to primary chemical journals. This project, CORE, is a collaboration of Cornell and the American Chemical Society, Chemical Abstracts Service, Bellcore, and OCLC. In this project, we scan the journal pages either from paper or microfilm, and we also have access to the database of full text maintained by ACS (and available commercially through STN). Figure 3 shows the data flow in the project. The pages of the printed journal are scanned, and the same pages are retrieved from the database of full text maintained by Chemical Abstracts Service.

Data preparation involves two particularly tricky aspects. One is the conversion of the ACS text database format into SGML. The AAP EMS standard did not anticipate a great many data fields found in the chemical journals, requiring many extensions to the standard, for such items as Chemical Abstracts structured index terms, tables of abbreviations, or four different kinds of in-line graphical items. There are also problems such as dealing with some 3000 character codes, translating the citations in such a way as to permit automatic retrieval of the corresponding items, and matching up the figure callouts with the available graphic items. Such problems are typical of any kind of conversion effort; in fact the ACS material is unusually well marked up as it comes to us, and all twenty journals are basically in the same format (there are minor differences in such areas as reference formats). The most complex conversions affect equations and tables. Several programs are used to display the text, and they must run on a variety of workstations. To avoid placing a burden on these programs, equations are actually turned into bitmaps: they are translated from the ACS format to *eqn* and then processed through Postscript into bitmaps. Tables are set up as Ascii but using fixed-width characters to avoid having to deal with the complexities of different character widths on different kinds of terminal equipment.

A more difficult problem is the identification of graphic items in the page images. The ACS text database contains the tables and equations, but not the figures or chemical schemes, which are pasted in during page makeup. As a result, we attempt to extract these from the page images and then match them with the figure references in the text. Various heuristics have been tried, as described in earlier papers ([Les90] and [Les91]). However, the major problem has been the high accuracy required. At first we thought that identifying too many graphical items would be relatively unimportant, but some articles have a large number of figures and schemes, and an early mistake in the graphics identification results in all the later items being mislabeled, which is very frustrating to the users. Although imperfect, we nevertheless wind up with a database of several hundred thousand pages, available in both formats. This does give the user the option of checking one format against another, in a pinch.

2.2 Image Access

Several different interfaces have been built using this material. One, from Bellcore, is an image based interface, in which the user gets to search the text, but see only the page images. It permits the user either to browse the pages by table-of-contents, or to search them. A sample screen is shown in Figure 4. This screen shows four windows: a greeting window, the table of contents of an issue of the *Journal of Organic Chemistry*, a small-resolution page from the journal, and overlapping that, an enlargement of part of that page. The enlargement is a part of the 300 dpi scanned page; the small-resolution page, at Cornell, is a 100 dpi equivalent reduction from the larger page, with anti-aliasing used to increase readability (this is not visible in the 1-bit-per-pixel reproduction used in this paper). Note that the tables of contents are taken from the Ascii files, not from the images. The user can also search for items, using a standard Boolean type search routine, and the result is a list of titles and authors, which can be clicked on in the usual way to view pages. Each page image provides, in a small scrollbar at the top, the page range covered by the article, with the ability to scroll to any page. Other buttons permit printing the article as well as moving forward and backward in the list of articles displayed.

The user may of course have multiple articles open at once, but is likely to run out of screen space. Resolution limits are a problem today with image systems: a full size 8.5x11 inch page, printed in fine print, is hard to put on the screen in an easily readable form without making it too big to fit. For good legibility one would like 150dpi resolution (we had people complain even when we used 200dpi) and this requires about a screen size of about 1200x1500, which few people have. And the hope of putting pages side by side, as is common with paper, is almost impossible with most of the workstations that might be found today on a scientist's desk.

To save time, we now store two forms of each page: the low resolution pages are computed in advance. Each 300 dpi page is reduced by a factor of 3 in each direction, thus taking 9 bits of the original page into 1 pixel in the reduced page. Two bits of grey scale are saved with each pixel. That is, in the original image between 0 and 9 bits may be black. In the final pixel, the grey level is set between 0 and 3. Tests showed that the most important information to save was the information about light pixels. That is, the mapping from 0-9 to 0-3 goes as follows: 0 bits on becomes grey level 0; 1 bit on becomes grey level 1; 2 or 3 bits on becomes grey level 2; and more than 3 bits on becomes grey level 3, or black. This gives the majority of the improvement in legibility (on a color or grey-scale screen) with a storage cost of only a factor of 2.

These low resolution pages are somewhat smaller (about 1/2 the size) of the full resolution pages. Since we do not have enough magnetic storage to store every page image, Cornell puts the full resolution pages on a jukebox and the low resolution pages for the most recent years on magnetic disk. The precomputation of the low resolution pages not only saves time in displaying the pages for the users, but also increases the number of pages that can be kept on magnetic storage, thus lowering jukebox traffic and the resulting delays from platter conflicts. Caching can be used as another method to reduce jukebox traffic; once the first page of an article is retrieved, we could retrieve the rest of the article and transfer it to magnetic storage. It is not clear yet what the optimum caching strategy is.

2.3 Ascii Access

In addition to the image display system, there are two Ascii based systems available. One is a variant of the Bellcore SuperBook system[ERL+89] and is shown in Figure 5. SuperBook provides a screen with two main windows, one on the left showing the table of contents, and one on the right showing the text. The table of contents is a hierarchy expandable in a 'fisheye' view, in which the brief list of top-level headings can be expanded to show details of subsections in any area. There is a small search window at the bottom left, in this instance used to search for the word "buckyball." The system displays the number of hits in each section of the journals to the left of the corresponding line in the table of contents area. Clicking anywhere in this area moves the display to that section of the text. Note that unlike many hypertext systems, there is a single linear ordering of the text, and thus the user can always keep oriented; it is always clear what is "before" and "after" the page currently being read. Moving forward and backward in the text display can move from one article to another, for example.

In the right window the current page is shown. This text is reformatted on the fly, including therefore the ability to highlight matching search terms, an operation not possible in the image system. It can also be easier to read (since the characters are generated by the usual computer display commands, and formatted for the window size) and faster to display. Footnotes, tables, and figures are indicated by icons in the right margin. The icon for a figure is a small version of the actual figure, recognizable in this case as some kind of structural drawing. Clicking on this icon pops up a larger version of the drawing, as shown in Figure 6. Note, however, that the user must perform an action to see the diagram. This has proven to be a problem, as will be discussed later, and newer versions of SuperBook put the drawings inline so as to call the user's attention more firmly to the graphic. Lack of screen space prevents us from placing each figure on the screen in large size each time it appears.

The other Ascii-based interface available is Scepter, written by OCLC. An example of Scepter is shown in Figure 7. Scepter, unlike SuperBook does not impose an ordering on the articles, each of which is independent. This is sometimes an advantage: in SuperBook users sometimes do not realize when they have moved into a new article, whereas in Scepter they can not make this mistake. At the top is a window showing the results of a search, with the first few titles and authors. The details of the search specification ("nitrobenzene") are in the next window on the left; note that many of the qualifiers, instead of being part of a complex Boolean expression the users would have trouble remembering, are specified in menu choices by clicking. The actual article is shown to the right. In the article window, the left side shows the components of the article including a default list of figure thumbnails, letting the user see quickly what the figures are in the article. Clicking on any of these brings up the full figure, as shown at the bottom left.

Experience with Scepter has shown that users greatly like the menus giving them a choice of parts of the article: and what they prefer are the pictures. They look at the figures and schemes first, followed by the author and title, and then by tables and references; only as a last, desperate measure do they read the text, it seems. Our users are basically graduate students and research staff. They are most often browsing, rather than doing specific searches; and they are not normally subscribers to the journals themselves, so that the electronics are a substitute for a trip to the library, not for their own copies. Perhaps the most heartening statement is that the most frequent complaint of the users is that we should have more journals in the system.

3 Experiments

In an effort to compare the image and text versions of the same material, we ran some systematic experiments at Cornell, using 36 students as experimental subjects in a controlled sitation. Dennis Egan of Bellcore ran these experiments, with two chemistry professors at Cornell serving as consultants to design the questions, and 1000 articles from the *Journal of the American Chemical Society* used for data. The students were divided into three groups: one had the journal issues on paper along with the corresponding paper volumes of *Chemical Abstracts*, one had the image retrieval system, and one had the SuperBook text system.

3.1 Tasks

Five types of tasks were set as exercises, with several examples of each type. The tasks ranged from straightfoward to very difficult. The simplest task is a factual question plus the cite to the article in which the answer is given, e.g. "In the article 'Total Synthesis of Ginkgolide B' by E. J. Corey, M. Kang, M. C. Desai, A. R. Ghosh and I. N. Houpis, *JACS* v. 110, p. 649-651, what is reported as a medically important property of ginkgolide?" This type of problem is called the "citation" task. The "search" task presented the subjects with a similarly precise question, such as "What is the calculated P-O bond distance in hydroxyphosphine?" but did not tell them in which article the answer could be found. A particularly interesting task is the "browsing" task, in which the students are given only one issue of the journal (about 80 articles) and a list of eight fairly concisely named topics (e.g. "bridgehead halides") and asked which topics were mentioned in that issue. The "essay" task was similar to the search task, except that instead of a factual question to be answered the students were given a topic such as "phospholipids" and asked to write a few paragraphs about it. Finally, the hardest task was the "analogous transformation" task, in which the students were shown two structural diagrams and asked to propose a way of transforming the first into the second; neither diagram appears exactly as shown in any of the articles, but a similar transformation is presented somewhere (and the students are not told where).

The browsing task was supplemented with an effort to measure serendipity. Every library user has so many stories about finding things by accident that one sometimes wonders if random retrieval would be an appropriate strategy. To attempt to measure serendipity, we supplemented the browsing task with a second stage. The journal issue was taken away from the students, and they were given a second list of eight topics and asked again which appeared, relying entirely on their memory of whatever they had picked up browsing through the issue.

3.2 Protocol

The total experiment took three sessions for each student, a total of half an hour of training plus six hours of experimental work. Each student used only one of the three possible formats. The results of the exercises were graded by graduate students who did not know which answers were written in which condition, and the time required to do the task was also recorded.

The students used different strategies at times, depending on which format they were using. Paper users, for example, relied more heavily on flipping pages in the journal, or reading the table of contents; the computer users tended to do searches. This was particularly evident in the browsing task. Doing the analogous transformation problems required considerable ingenuity on the part of the students, since we did not provide a substructure search capability. So they tended to generate as many terms as they could that described such structures, and scan the matching articles or paragraphs.

4 Results

The detailed results of these experiments have been presented elsewhere[ELK+91]. The most important conclusion is that searching with computers is enormously easier and more effective than searching on paper. For the search task, as an example, where the students had to spend at least 15 minutes, the students with paper spent nearly twenty minutes but nevertheless gave up more than half the time and scored only 23% right answers. Both computer users, the image users and the SuperBook users, scored over 75% correct answers. The paper form of *Chemical Abstracts* was available to the students using the journals on paper, but they were not sufficiently familiar with it. They did receive half an hour of training, which appears to be inadequate. For example, they had trouble with common names of chemical substances, not knowing how to use the Trivial Name Index in *CA* to find the systematic names used in the indexes.

4.1 Task Performance

Although computer searching had great advantages, reading the articles was just as easy on paper. In the citation task, where the article to be read is stated in the question, all three groups of students scored over 75% and all took about five minutes to read the article. Some problems in the SuperBook interface (e.g. the failure to notice when they had wandered into a different article) in fact caused the students using SuperBook to spend an extra two minutes searching or reading the wrong article, so that it wound up generally slower. Similarly, on the analogous transformation task, the image system was the most effective of a more-or-less unsatisfactory total results (only organic chemistry students are likely to be able to do this at all), and SuperBook was weakest, because students might bypass the right figure and not realize they had found the article with the answer in it.

All three formats (SuperBook, image and paper journals) yielded similar scores for the browsing task, but in different ways. The students with paper journals were more likely to miss a relevant article than to claim to have found something relevant when it wasn't really there. The students with the computer systems, and particularly those with the image interface, made the reverse kind of error: they rated too many subjects as present that weren't there, more rarely failing to find something that was there. What that means is that they simply found lots more stuff. Part of this reflects the difficulty of searching in general, particularly, as mentioned above, using *Chemical Abstracts* without adequate experience or training. Part of it reflects the ease with which computers can drown inexperienced users in material: for example, of undergraduate searches on the University of California online catalog, MELVYL, those that retrieve any titles at all retrieve an average of 400.

A disadvantage of the image system is that it can not highlight search terms within an article. Thus, students who chose to do the browsing task by searching for the terms in the topics and then reading the articles found (most of them), had to read through the article on the screen. This is more tedious than the SuperBook users, who can simply click to get to each consecutive hit. The SuperBook users are thus able to scan through more articles, albeit looking at less of each article.

There was no significant difference between paper and computer users in the supplementary browsing task. In all cases, the students were about 5/8 accurate on the decisions made based only on past memory of the journal issue (as opposed to about 80% accurate looking at the journal, and an expected accuracy of 1/2 guessing at random). There is thus an expectation that serendipity will work as well in computer libraries as in paper libraries.

4.2 Time and Preferences

For some of the tasks, the time to be spent was specified. For example, the students writing essays were asked to spend about 30 minutes on each, and they generally kept to that instruction. Where the time was allowed to vary, it generally varied in the same way with accuracy: those who were doing the problems rapidly were doing them correctly as well. The students also ranged from undergraduate chemistry majors to graduate students; one surprising result was that there was no effect of the extra years of chemical education on the results (Cornell is a high-quality, selective department at both undergraduate and graduate levels). There was a great scatter among individual students, as is common in retrieval tests.

The students were also clearly learning the systems from session to session of the experiment. Particularly on the browsing task, they improved in speed during the six hours of work with the journals. The subjects using electronics improved their speed more than those using paper, but even the paper users got somewhat faster during the three different days during which they performed the experiment. On the citation task, by contrast, there was relatively little improvement in performance.

It is clear that the students, and chemists in general, prefer looking at diagrams and figures to reading the text or articles. Examination of the figures in the article is key in doing many of the tasks chemists often do in a library, particularly the organic chemists working with structural diagrams. Observing people browsing in an ordinary library displays the same phenomenon people flip through journals at a rate which prevents them from reading much text; they are recognizing perhaps a few titles and judging pictures. In fact, their own choice would be to have articles summarized not by the titles and abstracts, but by the author names and small pictures of the figures, sort of a comic book version of the article. This is why Scepter, for example, always provides some of the thumbnails along with the title on the first article display.

Thus, methods for improving access to the graphic components of articles will be welcomed by the users, and extension of journal articles to multimedia, e.g.

animated 3-D drawings of molecular structure, would improve their utility to readers. This can already be observed in the market for computer software for chemists and biochemists, and the importance of reasoning based on stereochemical configurations in much chemical research. The importance of graphics extends beyond chemistry, of course, but these results might need to be viewed cautiously in any application to an area such as literary analysis, where the text might be more important.

5 Conclusions

From our results, it is clear that chemists can make effective use of either kind of interface. On what basis should one choose one format or another? Among the possible justifications are ease of system building, security, and flexibility.

5.1 System Construction

It is easier, as mentioned before, to put together an image system. Each journal and page is treated identically, and little special purpose code need be written. This also means the systems have fewer features, and in fact some are viewed mostly as fast ways of producing photocopies. They also require more screen resolution and network capacity than most desktop workstations normally have, and so image systems are often likely to appear as CD-ROM based systems on dedicated computers.

Ascii systems create problems in asking publishers to format texts more carefully. Many publishers today use a variety of composition techniques and do not carefully identify or tag much of the information in their books or journals. This limits the ability of a text system to reformat, but it can be quite expensive to convert to an SGML syntax. Some are looking at alternatives, of which Adobe's Acrobat seems most attractive today (but is less flexible in terms of reformatting than the SGML-based standards).

Image based systems are greatly limited in the kinds of displays they can present. All the ideas of hypertext and links, for example, are difficult to implement in a page image system. Until recently, hypertext systems were open to the challenge that it was difficult to create hypertext links and it was not clear how many of them would ever be made. SuperBook, for example, opted for a strategy in the primary mechanisms for moving from one place to another in a collection of documents would be using a table of contents or searching for author words. Hypermedia authoring was sufficiently more complex than writing straight text as to raise doubts about its future popularity. However, the rise of MOSAIC on the net has demonstrated that given the audience, people will in fact create quite large volumes of hypermedia material. It will also make people accustomed to the idea that more ambitious formats than merely imitation of a page are possible in a digital library.

5.2 Security

One of the other major issues constraining digital libraries is an attempt to avoid theft. The software publishing industry is subject to a great deal of illegal copying, with estimates of the fraction of copies of software that are pirated ranging from 35% in the United States to 86% in Italy and 99% in Thailand. The total cost of software piracy is estimated at $12 billion per year. It is not clear yet how bad the situation might be with digital libraries. The large database vendors, for example, gain some protection from the fact that they dole out information in fairly small quantities, and the chance that someone who steals a small piece of information will find somebody else who wants that precise bit of data is small. A publisher who provides whole books, on the other hand, has no such protection.

Note that this is not a problem of network security in the normal sense; the danger is not really eavesdropping, as much as the redistribution of material starting with someone who had legitimately bought one copy (or the right to view the material at least once). Thus, encrypting the transaction with the legitimate user, although certainly useful, does not answer the problem. We need a way to frustrate people who download information and then redistribute it without permission.

Image systems are fundamentally harder to steal from than text systems. Typically only a part of a page is visible, and stealing the downloaded stream to re-use it later is somewhat of a pain because of the need to reassemble the entire page. Furthermore, there is additional redundancy in an image library system (that's why it takes so much more disk space) and thus there is an opportunity to try to use that redundancy for protection. Some recent papers suggest hiding codes in the images, with the intent of being able to trace back illegal copies to the original source, so that legal steps can be taken against abusers. For example, Matsui and Tanaka[MT94] propose embedding a code in high frequency noise in a picture, which would not be easily apparent to the viewers. Unfortunately, it could be removed by almost any kind of low-pass filter. Brassil and others[BLMO94] have suggested manipulating the spaces between lines or words in a page image to encode a label; this is harder to remove and even less apparent to the user.

Whether these techniques can be built into an administratively practical system for preventing theft from digital libraries is not clear, and may depend more on the local ethics (e.g. at a university) and the general development of the industry. For example, today universities may buy some databases via a site license giving all students an unlimited right to use them, and other databases via a license which regulates each use. Students will resent the second kind of database, and express a feeling that since they think the second publisher "should" have offered the same kind of unlimited use terms, they should act as if that had happened. Unless and until a standard form of license and access for machine-readable data becomes accepted, this kind of conflict is likely to persist.

5.3 Flexibility

Of course, the major long-term advantage of text systems is going to be their flexibility and adaptability. They permit the user to capture the material being displayed for re-use. For example, some of our Cornell chemists are interested in going through the data base to collect certain kinds of (e.g.) spectroscopic measurements and accumulating a private database of such numbers). Text systems also permit one to vary the display format, perhaps to provide larger characters for those with less acute eyesight, or just to arrange the material in a way the user finds more attractive. It is possible to highlight, or to signal hypertext links, or to present paragraphs in isolation or in a different order, or make many other manipulations of the text. One can run together all the titles, or citations, or whatever.

Among the ways in which SuperBook, for example, makes use of the ability to dynamically manipulate Ascii text and adjust the display to the particular user needs are:

(a) The table of contents, which can be presented at any level of detail: the user, by clicking on it, can display subheadings or sub-subheadings in one area of a document and only main headings in another area of the document.

(b) Text pages always begin at a paragraph break, and if a search was done the paragraph at the top of the displayed page contains a match to the search, thus positioning the result of the search conveniently for the user. Unimportant as this may seem, there are experimental results showing that people seeking the answers to a question can find the right page in a printed book and still miss the answer because it is not prominent enough on the page.

(c) Words can be highlighted, the lines are always the right length for the window size the user has provided, and the typography can be adjusted for readability (e.g. the size of superscripts and subscripts can be enlarged compared to normal printing to compensate for the lower resolution of display screens).

(d) Above each text page are given the headings or titles of the document structure leading down to the page being shown, thus helping the user maintain orientation in the document.

Some of these presentation improvements can be done in image systems if optical character recognition has been performed on the images. The RightPages system mentioned before, for example, has the ability to highlight matching items, using OCR results to identify the location in an image which was found by the user query. So long as OCR accuracy is too low to permit displaying the OCR itself to the user, however, it is going to be difficult to do many of the more complex reformatting tasks.

5.4 Summary

We find in the end a conundrum. The image systems are quite popular today, and yet in the long run it seems that text systems are bound to win out. Most of the disadvantages we found for text systems in our experiments can be overcome: inline thumbnails make the graphics more accessible, flashing or audio cues can be used to keep the users informed as to their location in the documents, and with time the wide variety of interface techniques that now confuses the users will shake out and some fairly standard paradigms will dominate. But right now new publishing ventures are sometimes still choosing to go with images rather than text.

Perhaps the answer is in transition. One can imagine starting out by scanning, moving to a system in which scanning is supplemented by optical character recognition, and then to one in which it is replaced by information delivered from the publishers. Among the technology that would make this practical would be better OCR, techniques for deducing document structure from the page layout, and greater adoption of standards for document keying and storage. Image systems will become easier as screen resolutions improve and network speeds grow, but it is not likely that the more fundamental limitations can be overcome. They derive from the single fixed document format, taken from the printed page, which can not be adapted to the user or the particular situation (query or document). Thus, in the long run, we will move to text-based systems, with the major hurdles probably not being technological, but questions of piracy and copyright.

5.5 Acknowledgements

The experimental work described here was carried out by Dennis Egan, Dan Ketchum, and Carol Lochbaum. The work of Stuart Weibel at OCLC, Jan Olsen, Howard Curtis, and Rich Entlich at Cornell, Lorrin

Garson at ACS, and Lorraine Normore and Jim Lundeen at CAS has also been essential for the CORE project. My thanks also to Prof. Robin Alston of University College London and the British Library for his help with some of the examples.

5.6 References

References

[AAP87] AAP. *Standard for electronic manuscript preparation and markup, version 2.0.* Association of American Publishers, August 1987.

[BLMO94] J. Brassil, S. Low, N. Maxemchuk, and L. O'Gorman. Marking of document images with codewords to deter illicit dissemination. *Proc. Rutgers Digital Library Workshop (to appear)*, 1994.

[Bus45] V. Bush. As we may think. *Atlantic Monthly*, 176(1):101–108, July 1945.

[Bus67] V. Bush. *Science Is Not Enough.* William Morrow, 1967.

[ELK+91] D. E. Egan, M. E. Lesk, R. D. Ketchum, C. C. Lochbaum, J. R. Remde, M. Littman, and T. K. Landauer. Hypertext for the electronic library? CORE sample results. *Proc. Hypertext '91*, Dec. 1991.

[ERL+89] D. E. Egan, J. R. Remde, T. K. Landauer, C. C. Lochbaum, and L. M. Gomez. Behavioral evaluation and analysis of a hypertext browser. *Proc. CHI '89, Human Factors in Computing Systems*, pages 205–210, 1989.

[HC86] Thomas B. Hickey and Andrew M. Calabrese. Electro document delivery: Oclc's prototype system. *Library Hi Tech*, 4(1):65–71, 1986.

[HOS+93] M. M. Hoffman, L. O'Gorman, G. A. Story, J. Q. Arnold, and N. H. Macdonald. The RightPages service: an image-based electronic library. *J. Amer. Soc. for Inf. Science*, 44:446–452, 1993.

[Ken93] A. R. Kenney. Digital-to-microfilm conversion: an interim preservation solution. *Library Resources and Technical Services*, 37(4):380–401, October 1993.

[Les90] M. E. Lesk. Full text retrieval with graphics. *Bridging the Communication Gap (NATO AGARD Conference Preprint No. 487)*, pages 5-1 to 5-7, 1990.

[Les91] M. E. Lesk. The CORE electronic chemistry library. *Proc. 14th ACM SIGIR Conference*, pages 93–112, October 1991.

[McK93] C. McKnight. Electronic journals-past, present. . .and future? *ASLIB Proc.*, 45:7–10, 1993.

[MT94] K. Matsui and K. Tanaka. Video-steganography: How to secretly embed a signature in a picture. *Technological Strategies for Protecting Intellectual Property in the Networked Multimedia Environment*, 1(1):187–206, Jan. 1994.

[Rob93] Peter Robinson. *The Digitization of Primary Text Sources.* Office for Humanities Communication, Oxford University Computing Services, 1993.

[SOF+92] Guy A. Story, Lawrence O'Gorman, David Fox, Louise Levy Schaper, and H. V. Jagadish. The RightPages image-based electronic library for alerting and browsing. *Computer*, 25(9):18–19, 23–25, Sept. 1992.

[Ste90] Barrie T. Stern. Adonis-a vision of the future. In G. P. Cornish and A. Gallico, editors, *Interlending and Document Supply*, pages 23–33. British Library, Boston Spa, 1990.

[Val93] Edward J. Valauskas. Ncsa mosaic on the macintosh. *Online*, 17(5):99–101, Sept. 1993.

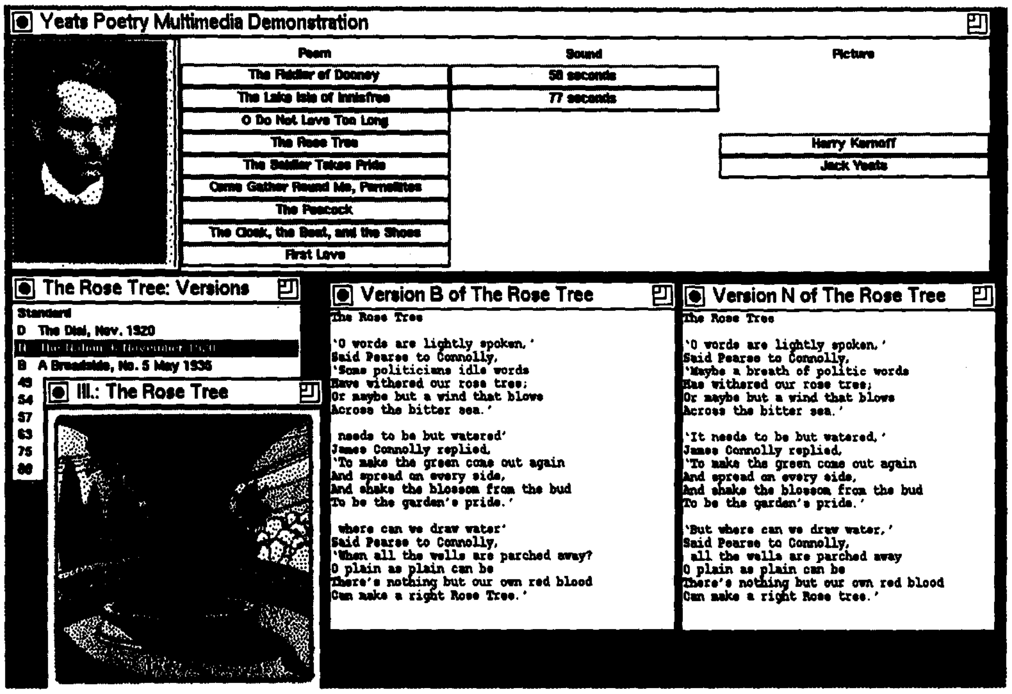

Figure 1. Yeats multimedia poetry screen.

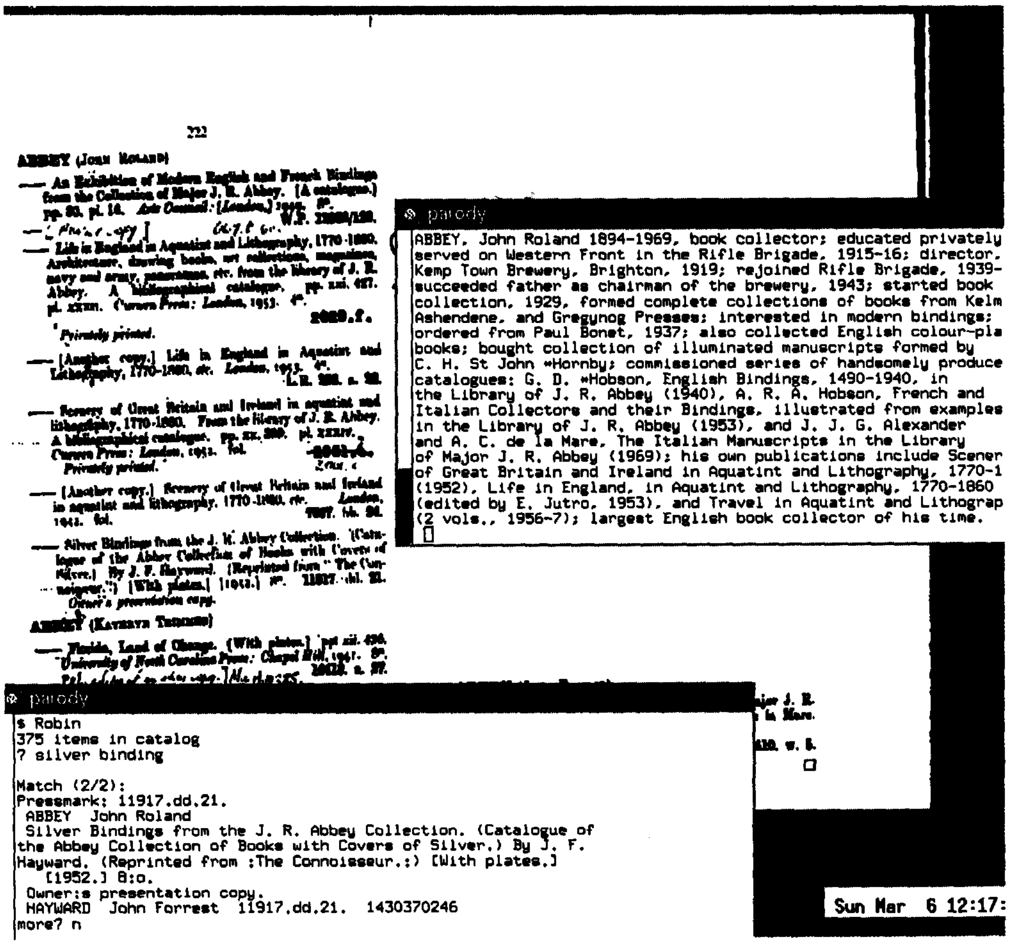

Figure 2. Sample of British Library General Catalog.

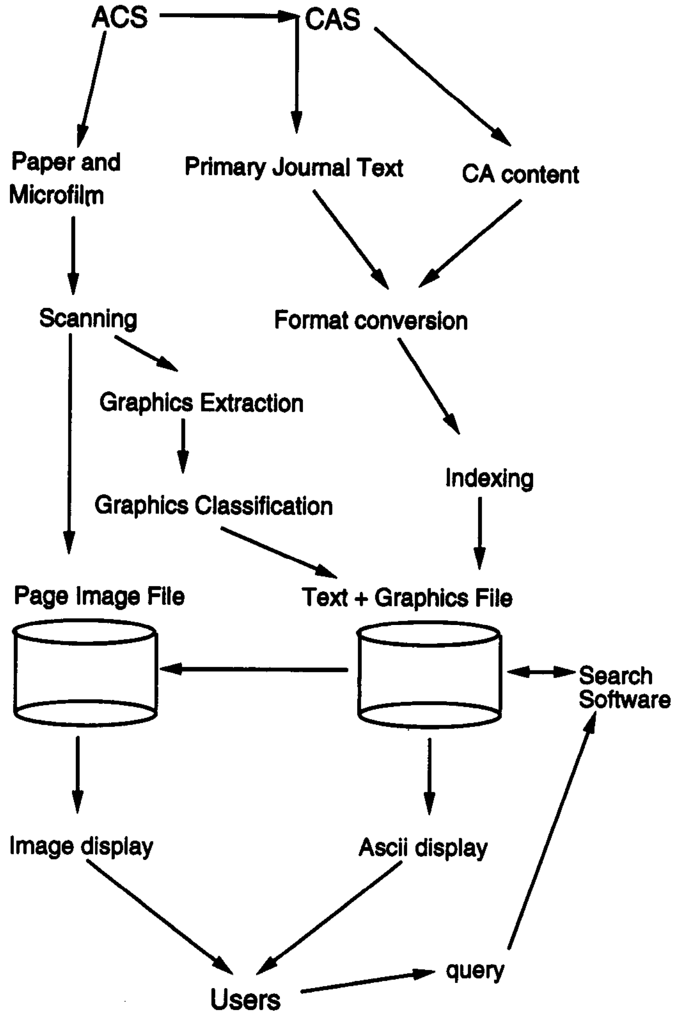

Figure 3. Data flow in CORE project

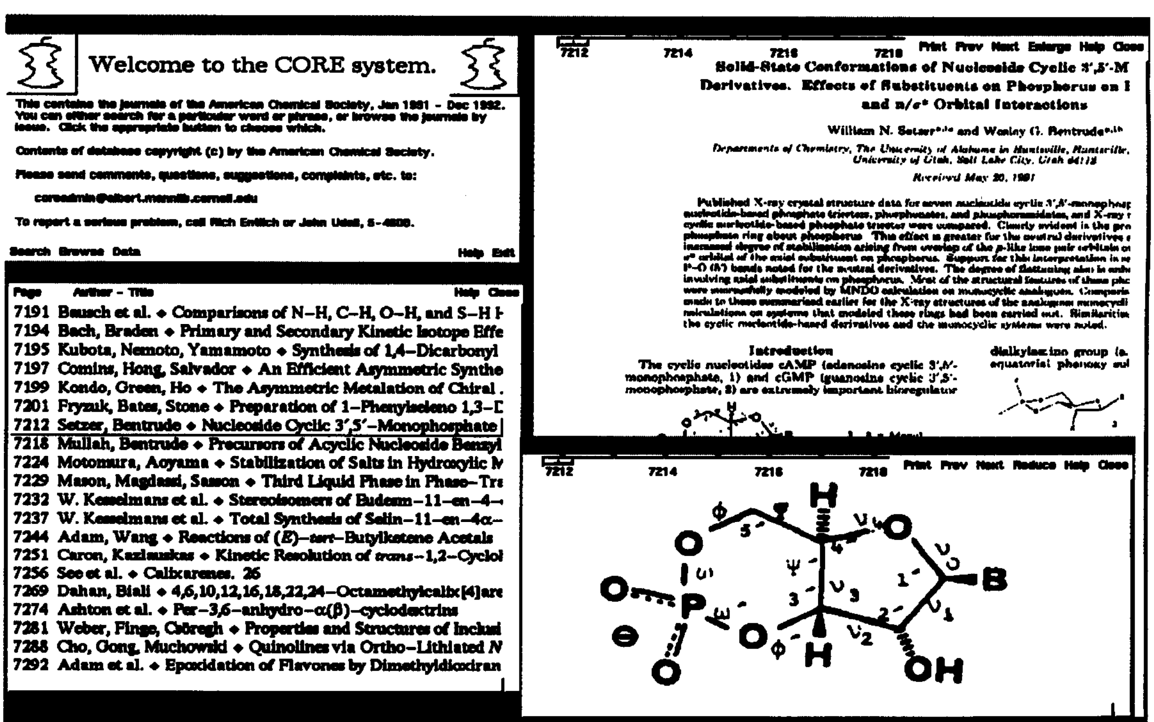

Figure 4. Image display of browsing chemical journals.

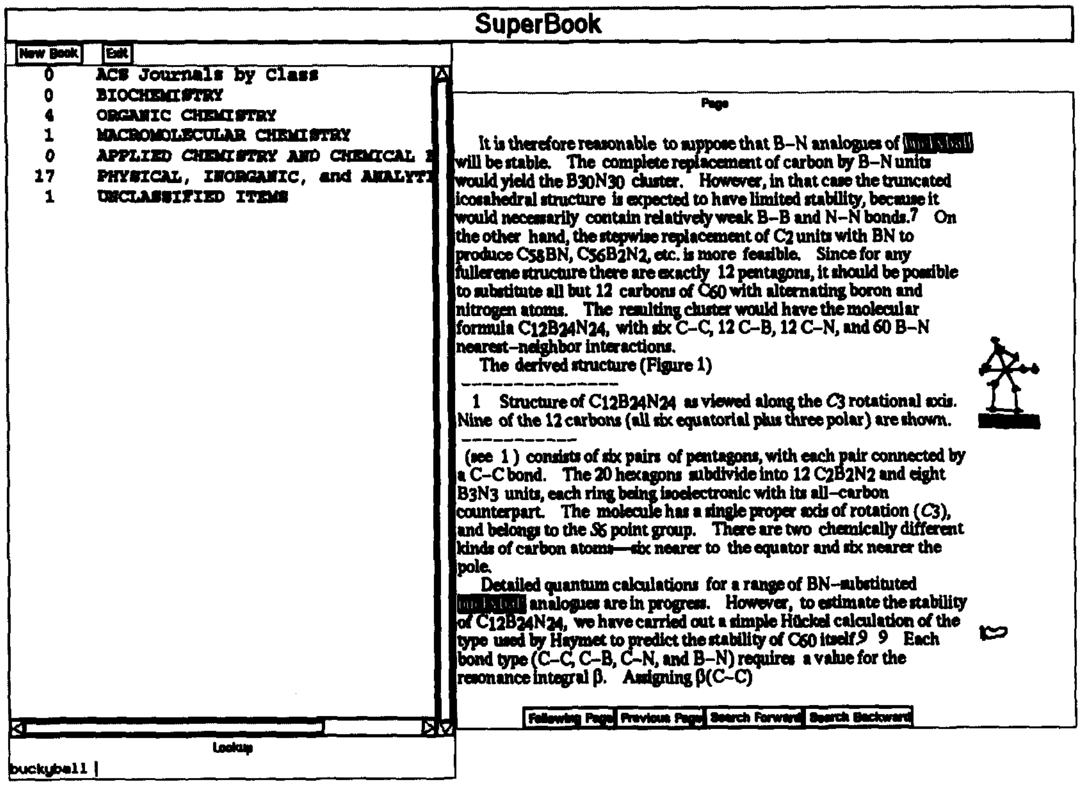

Figure 5. SuperBook screen with contents and text

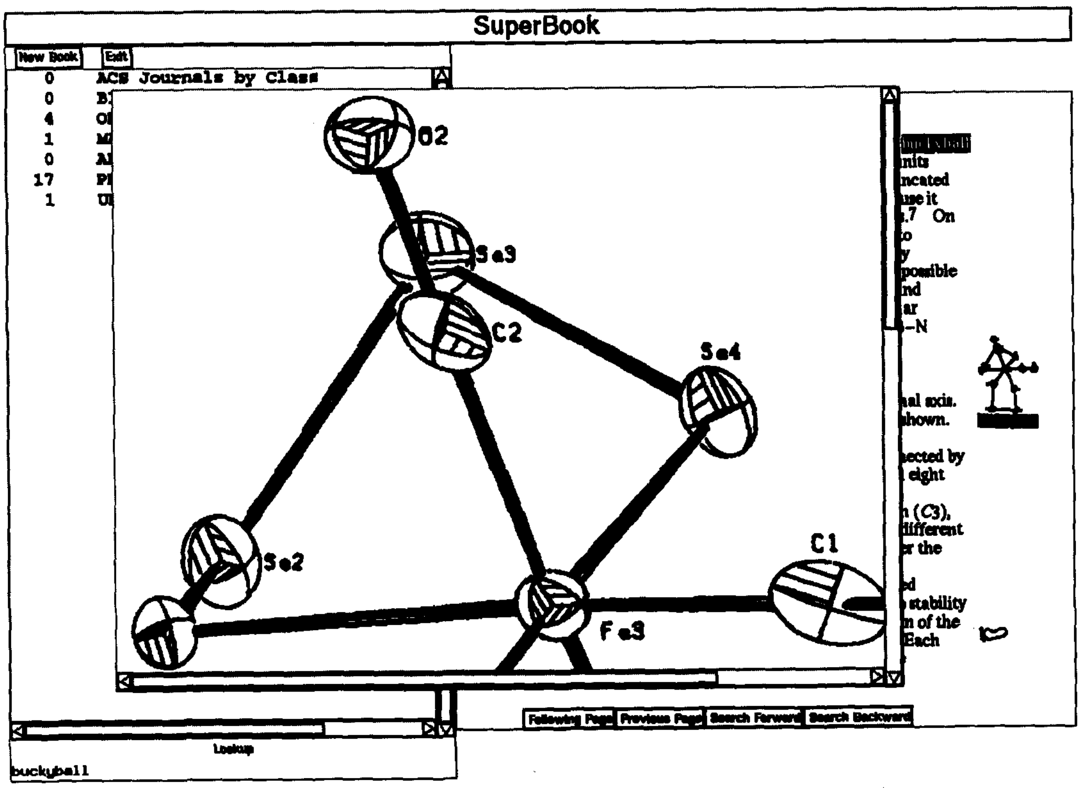

Figure 6. SuperBook screen with graphic enlarged

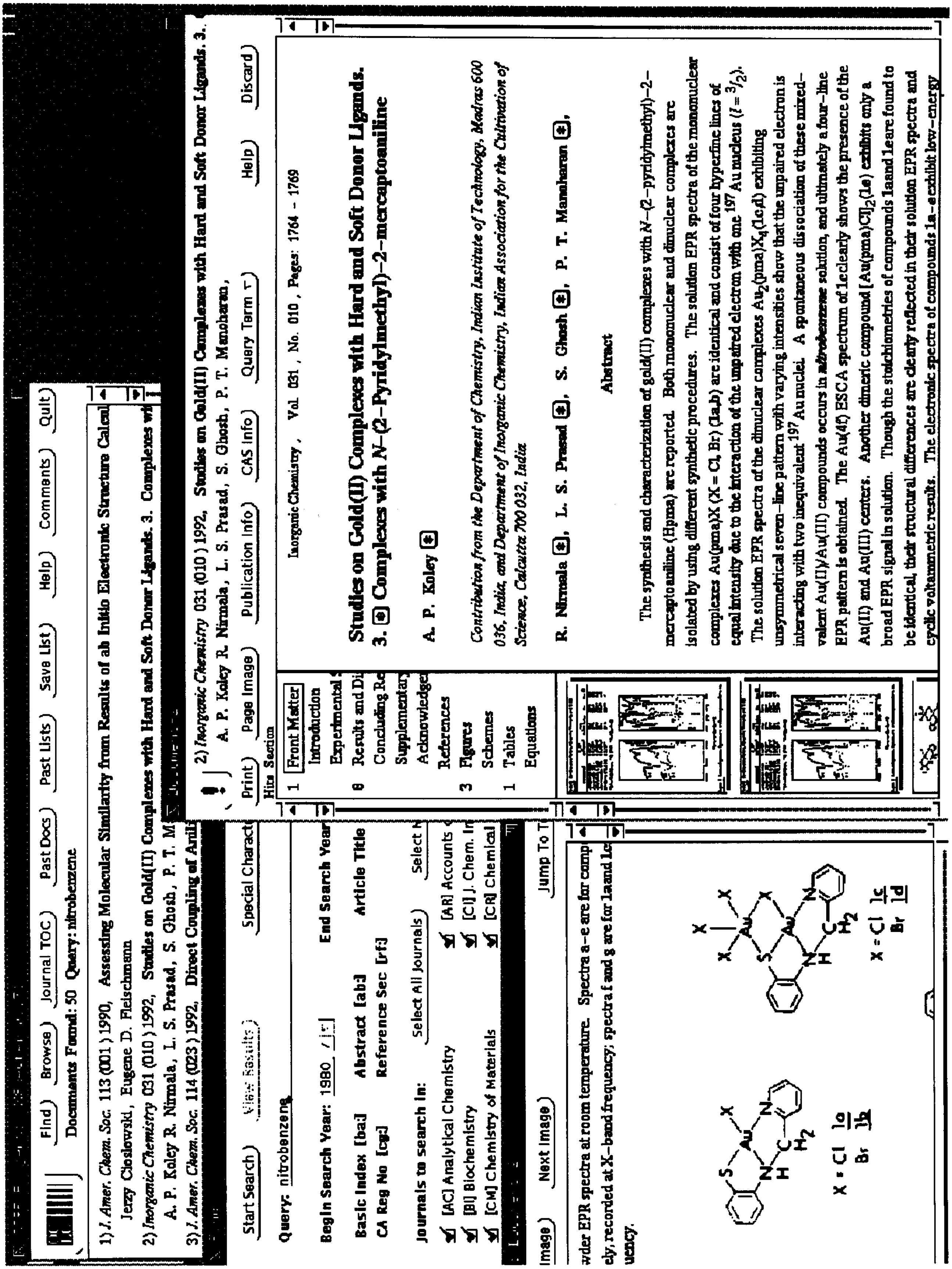

Figure 7. Scepter screen display