Lessons Learned from Full-Text Journals at OCLC

Thomas B. Hickey

OCLC operates many services and programs for libraries, but the major ones are an online cataloging system, an interlibrary loan system, and a reference service. The reference service (FirstSearch) includes full-text databases as well as databases of abstracts and indexes. We currently serve more than 25,000 libraries, have more than 1,100 full-text with-graphics journals online, and the FirstSearch databases contain more than 250 million records.

We started working on full-text journals by examining new approaches to information display. We felt that there would be a gradual movement from the availability of only metadata electronically to the full text of reference works, journal articles, and finally books. We started investigating and working with Donald Kuth's Metafont and TeX as soon as they were available, even to the point of doing our own ports of the systems. In general, this is the progression we have seen, although the wide availability of journal articles has taken longer than we expected twenty years ago.

In the early 1980s, there were two main problems regarding electronic display of text: fonts and speed. Scientific text especially requires a large number of special characters (glyphs) that were not generally available. In fact, electronic versions of fonts were hardly available at all except for the dot matrix ones embedded in hardware such as displays and printers. Speed problems had several aspects. Communications speed was the worst—we were working with 1200-baud modems typically with very expensive ($1000+) 9600-baud modems gradually becoming available. Printing was restricted to machines that could support a bitmap image of a page and typically took 5 to 10 minutes/page to display. Display of a page on CRTs was faster but still could take ten seconds or more because of the lack of hardware support for special fonts and limited processing speeds.

The research department at OCLC has had two main projects with full-text journals. The first was Graph-Text (Hickey & Calabrese, 1985). We worked with American Chemical Society (ACS) journals obtained through Chemical Abstract Service (CAS). CAS had a very sophisticated system, developed in the 1970s, which stored their journals in structured records in a database at a time when most publishers had no concept of storing any sort of electronic version of their material. We took tapes of their articles in their database format, translated them into TeX, and then into TeX's standard device independent format, DVI, creating new glyphs as needed using Metafont. Since the graphics were not on the tapes, we scanned the original documents, cropped out the graphics, and linked those to the graphic callouts in the articles. Our first system relied totally on metadata for identification and selection. It batched requests for overnight downloading and printing since delivery of an article with its associated fonts and graphics could take up to half an hour over modems.

Before we tested the system in libraries, we decided that PCs had progressed to the point where display of the formatted text and graphics was possible interactively, so we revised and extended the system to do this. To avoid slow transmission speeds, we stored the data on CD-ROMs. To store the data, we devised our own directory structure since there was no standard method of storing files on CDs at the time. These were, of course, single speed CD readers with very slow seek speeds, but they were much faster than running the system remotely over modems. We used Hercules Graphic cards for display on 286 PCs and the first of the Canon laser printers for printing (Hickey & Handley, 1987). The system worked fairly well, but we were never satisfied with CD-ROMs as a delivery mechanism for journal articles. There tend to be too many of them and they tend to be out of date.

Our second major project worked with John Wiley & Son's *Kirk-Othmer Encyclopedia of Chemical Technology* (Hickey, 1988). We obtained this from them in a simple text format designed for loading into an ASCII database system. We translated this into SGML and used the SGML to drive TeX and indexing for the database. For our testing, we concentrated on a single volume of the twenty-four volume encyclopedia and successfully translated it. For display, we used a Wyse 700 display which offered black and white resolutions comparable to those on CRTs today. Unfortunately no window system was available to make display of articles easier, so again we were forced to write software, such as window management and character display, which today is taken for granted. Over the past fifteen to twenty years, we have seen the following progression in the delivery system and format of our data:

Delivery	*Format*
1200-9600 Baud Offline	TeX/DVI

CD-ROM	TeX/DVI
Online, proprietary client	SGML/TeX/DVI
Web Browsers	HTML
	PDF & Image

OCLC's first commercial venture in formatted full text was called Electronic Journals Online (EJO). This was a joint project with the American Association for the Advancement of Science and started with a new electronic journal, *Current Clinical Trials.* We wrote our own client for the display of the articles that were coded in SGML and translated using TeX into DVI. All the fonts used were developed in-house. EJO had several dozen journals mounted at its height, but we found working with the SGML too expensive. Each publisher's DTD and special formatting requirements simply took too much staff time to make the system affordable. When the economics of the system became apparent, OCLC revamped the system into Electronic Collections Online (ECO). ECO now has more than 1,100 journals online and often loads more than fifty new journals each month, a rate much higher than we could ever have reached with the earlier SGML system. Each of these journals is stored in Portable Document Format (PDF), a format devised by Adobe to eliminate many of the portability and rendering problems that their PostScript format has. We do minimal processing on the data to mount metadata about each journal, issue, and article and link it to the PDF files. The whole system is closer to OCLC's FirstSearch database model and is being completely integrated with the FirstSearch system.

We have been working on bringing electronic documents to users since before it was really feasible and have learned a few lessons along the way:

- *CD-ROM is not the same as telecom.* On the face of it, this sounds obvious, but we were surprised by some of the consequences. We used CD-ROMs to move data to the user instead of setting up modems seeing the CDs as a replacement for the network. We found in focus groups that librarians looked at the CDs much as they did books and journals that they acquired—they owned them. Any sort of per-use charges for items they had physical control over was much less acceptable than per-use charges for items obtained remotely.
- *Math, tables, and layout are 90 percent of the effort.* On a day-to-day basis, these things are what continue to absorb time. Of the three, math is probably the worst if it is a central part of the articles, but tables can become extraordinarily complex and long (e.g., we encountered a fifty-seven page table). In our production databases, we sidestepped most of the layout problems by separating graphics and tables from the text,

but the Graph-Text project tried to match or better the ACS layout and required substantial effort.

- *Production was as expensive as predicted.* One of the main objectives of our early research was to assess how expensive processing the material would be. We went ahead with production plans in spite of the projections of $5 to 15/page and found that, even at higher volumes, these costs were difficult to reduce.
- *SGML helps but not much.* One of our hopes was that getting SGML from the publishers would result in dramatically lower costs. SGML did help. It made it possible to mount journals from publishers. But the math and tables are still there, SGML offers little help in the actual rendering of the text, and our costs to mount SGML journals were too high.
- *TeX is hard to beat.* TeX is not the perfect typesetting system, especially for material you would like to manage in large batches. It is amazing, however, that nearly twenty years later there is still nothing better at what it does.
- *Fonts remain a problem.* The use of PDF insulates us from this problem, but somewhere in the production stream, someone is struggling with yet another character that someone has dreamt up.
- *Publishers were not ready.* Not nearly as true now, but the amount of education and liaison necessary with each publisher was a significant part of EJO's cost structure. Each new publisher was expensive. We could not charge enough to recover our costs for this.
- *Users are harder to change than publishers.* Users in general were not particularly interested in electronic journals until they had experience with the Web and started getting familiar with using information in a networked environment.
- *Proprietary clients are wrong.* There are still people learning this one. Just the barrier of having to install software on each machine that might access a service is too high for a system providing access to scholarly journals. We explored several ways to better integrate our service into the Web before settling on the use of PDF and Adobe Acrobat (Hickey, 1994, 1995).
- *It is the data that matters.* Fancy interfaces are fine, and we spent much time developing, testing, and changing ours, but these are not the most important things. The data are important, and images of pages are fine for this. One feature that is needed is printing, and PDF does a wonderful job at this, even though reading most PDF documents on the screen is difficult at best.
- *End users are less worried about format than publishers.* We suspected this from the start, but publishers have a vested interest in making their material look as unique as possible, and a system that ignores this will

face great resistance from publishers. Our systems did try to maintain these publisher distinctions, but these proved expensive.

Electronic journals now face a whole new set of problems. When we started, we had problems of communications, printing, processing speed, proprietary clients, unavailable fonts, and when to start developing a product. Now the problems are more pricing, licensing, archiving, integration of the digital and the physical, and standards to reduce costs and increase interoperability. It seems that we have accomplished more in the last five years than we did in the ten before that. I hope the same will be true for the next five.

REFERENCES

Hickey, T. B., & Calabrese, A. M. (1985). Electronic document delivery: OCLC's Prototype System. *Library HiTech, 4*(1), 65-71.

Hickey, T. B., & Handley, J. C. (1987). Interactive display of text and graphics on an IBM-PC. In A. H. Helal & J. W. Weiss (Eds.), *Impact of new information technology on international library cooperation* (Proceedings of the Essen Symposium held September 8-11, 1986 in Essen, West Germany) (pp. 137-149). Essen, West Germany: Essen University Library.

Hickey, T. B. (1989). Using SGML and TeX for an interactive chemical encyclopedia. In *National Online Meeting* (Proceedings of the tenth National Online Meeting held May 9-11, 1989 in New York) (pp. 187-195). Medford, NJ: Learned Information, Inc.

Hickey, T. B. (1994). Integrating Guidon with the World Wide Web. *Annual Review of OCLC Research.* Retrieved October 27, 1998 from the World Wide Web: http://www.oclc.org/oclc/research/publications/review94/part1/integuid.htm.

Hickey, T. B. (1995). Guidon Web: Applying Java to scholarly electronic journals. *Annual Review of OCLC Research.* Retrieved October 27, 1998 from the World Wide Web: http://www.oclc.org/oclc/research/publications/review95/part1/guidon.htm.