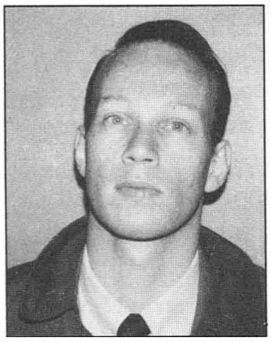
THE UNIVERSITY LICENSING PROGRAM (TULIP):

A LARGE SCALE EXPERIMENT IN BRINGING ELECTRONIC JOURNALS TO THE DESKTOP

Jaco Zijlstra

TULIP, a co-operative project involving Elsevier and nine university libraries in the USA, aims to determine the feasibility of the networked distribution of journals, to understand the economic and practical viability of the method and to study usage patterns. Jaco Zijlstra describes how electronic files for 43 journals in materials engineering are being created and used to achieve the three objectives.



Jaco Zijlstra is Project Manager TULIP, Elsevier Science B.V., 655 Avenue of the Americas, New York, NY 10010
tel: 212-633-3757
fax: 212-633-3975
e-mail: j.zijlstra@elsevier.nl

Starting point

TULIP started early in 1991. At this time a number of smaller experiments had been started or were in the planning stage. Elsevier was participating in a research project at Carnegie Mellon and had been approached by several other universities to join in single-site experiments. These experiments however were mostly rather idiosyncratic and unlikely to lead to generalizable findings.

University systems and library leaders at a number of schools had been talking with Elsevier to find a way to accelerate the development of large scale systems for the distribution in electronic form of traditional journal information — information presently found only in print. Elsevier Science was looking at the same question from the publisher's side and was looking for experience on which to make strategic developmental and investment decisions, whether in search software, document delivery systems, PostScript or SGML database files or network development.

During a few Coalition for Networked Information (CNI) meetings in the spring of 1991, it was agreed that if ten or fifteen universities would commit to the same basic experiment, then a publisher could justify investing in the creation of a major testbed. University participants outlined a project and organized a group of universities on the spot and so TULIP was started.

The universities were invited to submit project proposals, and Elsevier started to establish the technical and organizational framework necessary for such a large project. Ultimately, the TULIP program became operational in January 1993, and nine universities had decided to participate: Carnegie Mellon University; Cornell University; Georgia Institute of Technology; Massachusetts Institute of Technology; University of California (all campuses); University of Michigan; University of Tennessee; University of Washington, and Virginia Polytechnic Institute and State University. The program is planned to run through 1995.

TULIP's objectives

TULIP is a cooperative research project, testing systems for networked delivery and use of journals at the users' desktop. The participants set three objectives at the outset.

i) Technical - to determine the technical feasibility of networked distribution to and across institutions with varying levels of sophistication in their technical infrastructure. "Networked distribution" means sending the information both across the national Internet and over campus networks to the desktops of students and faculty. Elsevier delivers the journal information to participating universities in standard formats. The universities incorporate the information into local prototype or operational systems. A wide variety of delivery alternatives, search and retrieval systems and print-on-demand options will be compared.

ii) Organizational and economic - to understand, through the implementation of prototypes, alternative costing, pricing, subscription and market models that may be "viable" in electronic distribution scenarios; comparing such models with existing print-then-distribute models and understanding the role of campus organizational units under such scenarios. The overall goal is to reduce the unit cost of information delivery and retrieval. "Viable" means economically and functionally acceptable to all parties.

iii) User behaviour - to study reader usage patterns under different distribution (technical, organizational and economic) situations. Improvement in the functionality of the information, whether as to article structure or retrieval tools, will also be considered. Certain data will be collected uniformly at all sites for analysis in the aggregate and for comparison among different systems.

To reach these objectives, each university has as much local autonomy as possible, subject to standard terms and conditions which are outlined in licences signed with each site.

Why materials science?

The participating universities have in common strengths in the physical and engineering sciences. In looking within these disciplines for a target area, we wanted a field in which there is a mix of researchers comfortable with computers and a less sophisticated community. Materials science provided a field in which there was both a sufficiently large corpus of frequently-cited material within one publishing company and interested faculties. It is also a field in which Elsevier has a large core collection of journals without the need to rely on other publishers, something which everyone thought would only further complicate the process.

Technical foundations

Elsevier is providing electronic files for 43 Elsevier and Pergamon journals in materials science and engineering. These files consist of:

- TIFF bit-mapped page images (cover-to-cover, including tables of contents), scanned from the printed page at 300 dpi (600 dpi for certain applications), Group IV fax compression;
- edited and structured ASCII "heads" for each editorial item, including bibliographic citation and article abstract,
  and
- unedited OCR-generated ASCII full text for use in searching, but not for display.

In addition, as TULIP journals become available in SGML and PostScript as Elsevier re-tools its production processes to provide these formats as standard output from its production processes, they will be made available to the participating universities in those formats as well.

Each university receives, without charge during the project, the electronic full-text (bit-mapped and ASCII) for those journals to which it subscribes in paper. They also receive the bibliographic information for all 43 journals and have on-demand access on a pay-per-use basis to those titles to which they do not subscribe. All but one university mount the subscribed-to full text files locally on their own file servers. This means that biweekly Elsevier's contracted host, Engineering Information (Ei), distributes these voluminous files to each university over the Internet. One university retrieves articles over the Internet from Ei on-demand. Both models are important to test for efficiency and cost-effectiveness.

Implementation at universities

In 1992 the universities prepared plans as to how they expected to implement the program on their campuses. These plans are diverse, including single-sites, multiple campuses within one institution (where the files will be mounted on one server for all), a co-operative arrangement between two institutions and the possibility of testing regional network distribution to a much larger group of schools. Access tools and distribution systems on campus also include a wide range of alternatives, from high resolution images sent directly to desktop workstations to DocuTech print-on-demand of individual articles and of locally sold subscriptions.

TULIP production

There are several publishing houses in different countries producing the TULIP journals, with significant differences in editing processes, typesetting and printing, as a result of which each journal title has its own layout, its own size and its own typefaces. At present Elsevier Science is consolidating all these different production methods to streamline the output into one electronic format, which is then the basic material for paper as well as 'real' electronic versions of the journals.

As an intermediate step for TULIP, however, we use the paper version of the journals to produce scanned images as the electronic form of the journals. In our scanning office TULIP data is created by the following procedures.

The scanning office receives the journal issues under a special priority subscription arrangement.

After logging in the journal issue, the spine is cut off and the pages are fed into a double-sided, high-volume image scanner. As the scanner does not take into account differences in size between journals , the full size page images are electronically trimmed to the real size of the original page. The page image files are compressed to approximately 8% of their original size for storage.

The page image files are fed into an Optical Character Recognition (OCR) process, in which blocks of black and white dots are interpreted into characters, words, sentences and paragraphs of text.

The top part of the resulting OCRed text is checked and spelling mistakes are corrected, to produce bibliographic information. Title, authors, keywords, abstract, etc. are assigned to specific fields. This is followed by a check on completeness and consistency.

Every two weeks the material is put together into so called 'full datasets'. These datasets contain all page images of the journal issues, the raw or unedited ASCII full text, and a master index to the page images, which includes the bibliographic data.

On average, the size of a page image file is 1 megabyte (Mb) uncompressed, and compressed (CCITT fax group IV) between 70 and 80 kilobytes (Kb). The raw ASCII is about 4 Kb per page, and the bibliographic data of an article is around 1.5 Kb. The length of articles varies quite a bit, ranging from long reviews to short editorials but the average article is about seven pages long.

A typical dataset would contain some 500 articles, 3,500 pages and take up 280 Mb of disk space. The total storage space required for the 1992 and 1993 TULIP material is close to 20 Gb.

Datasets are dispatched to the distributor in the United States, Engineering Information's Article Express office in Westbury, NY, where a number of functions are handled.

- Each new dataset is loaded on magnetic disk. At any given time, there are around 10 datasets loaded for immediate distribution. Older datasets are stored on CD-ROMs.
- For each of the TULIP universities a customized index file is generated from the original master index. A customized index includes all the bibliographic material from journals to which the university subscribes, be it a paper or an electronic subscription, including pointers to the image files. For material to which the university does not subscribe, the bibliographic material is included without the pointers to the image files.
- Again for each university an FTP script is generated. The Internet facility File Transfer Protocol, or FTP, is used to move the files from Engineering Information to the

universities. The scripts automate the sending of the TULIP datasets, including checks on delivery of all files. The FTP script ensures that, for one issue at a time, the image files are being sent according to the university's subscription profile. After each issue the accuracy of the transmission is checked and the next transmission started. At the end of the transmission of all images from a dataset, the complete index, including the bibliographic data, is transmitted.

- The universities "pull" the most recent datasets when they are ready to receive them. That is to say, they initiate the execution of the FTP scripts described above by logging in on Ei's FTP server and "kick off" sending of a particular dataset. As stated above, the average size of a dataset is around 300 Mb. Transferring these large amounts of data across the Internet takes between three and five hours per site. Engineering Information is linked to the Internet by a T1 data link with speeds of 1.5 Megabit/second.
- The datasets are put on a machine and directory designated by the university. In most cases this is a temporary storage facility. From this temporary storage, the files are transferred to the information systems in which TULIP data are kept for use on the campuses, making place for delivery of the next batch.

Engineering Information also runs an article delivery facility of the TULIP titles, so that universities can order articles from journal titles for which they do not hold a subscription. This facility is driven by structured electronic mail messages, allowing for, but not requiring, an automated process on the university side.

Research of user behaviour

Using information technology to bring information to the user's desktop in order to improve the accessibility of scientific information is a major goal of TULIP for Elsevier, as well as for the participating universities. What we do not know as yet, however, is which combination of elements (technical, organizational and economic) makes for a successful implementation. 'Successful' means that users like the system and above all actually use it for their research and study. An important part of the TULIP experiment, therefore, is to gather and subsequently analyze information on the user behavior under different technical, organizational and economic conditions.

Some of the TULIP partners have built in major statistical usage tracking applications to do detailed behavior studies in the coming years. In addition to these local research projects, Elsevier is doing qualitative market research using questionnaires, interviews and the like, as well as quantitative research by gathering usage logs from the TULIP partners. The privacy and anonymity of the user will, of course, be rigorously protected at all levels.

The results of these various research approaches will be combined to produce overall reports which will be part of the final TULIP report to be completed in the spring of 1996.