Proceedings of the 1993 International Conference on Refereed Electronic Journals. 1-2 October 1993, Winnipeg (MB): University of Manitoba (1994) (ISBN 0-919932-02-9), 14.1-14.12.
[Anonymous FTP: ftp.cc.umanitoba.ca /e-journal/ICREJ93/brailsfd.wp5 Base font: Times Roman 12pt]

ADOBE'S ACROBAT™ — THE ELECTRONIC JOURNAL CATALYST?

David F. Brailsford

Electronic Publishing Research Group
Department of Computer Science
University of Nottingham
Nottingham, U.K. NG7 2RD
Internet: dfb@cs.nott.ac.uk

Abstract: Adobe's Acrobat software, released in June 1993, is based around a new Portable Document Format (PDF) which offers the possibility of being able to view and exchange electronic documents, independent of the originating software, across a wide variety of supported hardware platforms (PC, Macintosh, Sun UNIX etc.).

The principal features of Acrobat are reviewed and its importance for libraries discussed in the context of experience already gained from the CAJUN project (**C**D-ROM **A**crobat™ **J**ournals **U**sing **N**etworks). This two-year project, funded by two well-known journal publishers, is investigating the use of Acrobat software for the 'electronic' dissemination of journals, on CD-ROM and over networks.

Keywords: Acrobat; PDF; FTP; Gopher; CD-ROM; electronic journals.

1. INTRODUCTION

For some time now publishers, librarians and all users of scholarly publications have been subjected to a series of sermons to the effect that the future of publishing lies with 'electronic documents' in one form or another. We are told that the printed form of a document will become just a small part of a much bigger story and that the process of reading documents will soon be revolutionised by further advances in computer and network technologies. Before too long, the story goes, we may be able to travel on an airplane carrying our book-sized personal computers complete with attached pair of Virtual Reality spectacles. Rather than reading paper-based reports we shall be viewing a series of electronic documents by looking at the correct area of the virtual screen and giving the appropriate command (*"Next Page!", "Zoom In!", "Print It!", "FAX it to Head Office!"* etc.). These commands will be picked up by a tiny microphone built in to the spectacles, and will be recognised in an instant by sophisticated voice-recognition software. It goes without saying that if a document is not immediately available on the book-size personal computer it will be fetched, quite automatically and transparently, over ultra-high bandwidth telecommunications links from the user's home or office computer or from some global database of available documents. The network and database software will *know* where to look.

Whether you regard this vision as a dream or a nightmare, there is little doubt that something of the sort will come about. However, it has never been entirely clear how we can move from our present paper-dominated dissemination of documents towards the Brave New Electronic Future. In certain areas a start has been made: an increasing amount of material is now available over networks and on CD-ROM but the overriding problem at the moment is the lack of *de facto* standards for electronic dissemination of high-quality documents. By 'high quality' I mean that the electronic format should have all of the richness of the printed document in terms of typefaces, diagrams, full-text search and so on — in other words, the electronic pages should be more than just ASCII text or scanned-in page images. To these requirements we could also add that the chosen format should be robust with respect to transmission over networks and capable of being viewed with software available on *all* of the popular computer platforms (IBM PC, Macintosh, UNIX systems etc.).

The next section reviews some of the options currently available for networked distribution of 'electronic' journals.

2. FORMATS FOR ELECTRONIC DOCUMENTS

2.1 The ASCII 'Jail'

The US ASCII character set acts as a lowest common denominator for the transmission of information around the world. Unfortunately, limitations in network hardware and electronic mail software mean that vital information may be lost, or garbled, unless the characters are restricted to the 7-bit subset of ASCII. This restriction is not too onerous for sending computer programs and simple messages, so long as these stick to the unaccented characters of the Western alphabet. ASCII is quite useless for sending diagrams, photographs, oriental characters or any other complex material.

On the other hand ASCII does have the virtue of being searchable for keywords and other text strings. But this is only useful, for typeset material, if the result of that search can be related to a position on the printed journal page. Since ASCII appears as fixed-pitch typewriter-like characters when viewed on screen (and given that very few journals these days are published as typescript) it follows that the ASCII text of an article will generally have completely different line and page breaks to those seen in the printed copy.

For all these reasons John Warnock, the CEO of Adobe Systems Inc., regards ASCII as a 'jail' from which we must break free if we are to transmit complex documents to one another.

2.2 Scanned Pages

A second possibility for creating an 'electronic' journal is simply to scan in the pages, using a document scanner, and to store the pages in bitmap form (e.g TIFF or Group 4 FAX) as a collection of black and white dots. This approach has been extensively used in some first-generation electronic document systems but there are many difficulties. A scanned A4 page at 300 dpi requires almost 1 Mbyte of storage unless compression techniques are used. By dropping the resolution to 150 dpi, as in Group 4 FAX, and applying compression algorithms, such a page can be stored in as little as 2.5 Kbytes but the image quality is poor and colour is out of the question.

Since scanned images cannot be searched for strings of text, it follows that any indexing, searching and database access has to be done on auxiliary files of text which are cross-linked in some way to the images themselves. This approach is taken, for example, in KnowledgeSet Corporation's *Knowledge Retrieval System* and in the *Right Pages* project being conducted by AT&T(Story, 1992, pp.l7-26). The need to keep the text files up-to-date with respect to the page images to which they refer leads to vendor-specific database and viewer software which is tailored to this 'dual representation' approach.

3. PROPRIETARY SYSTEMS

The problems alluded to in the previous subsections have led various vendors to look for an underlying document representation which is more powerful than ASCII text or scanned pages. The *WorldView* system from Interleaf uses an underlying CGM (Computer Graphics Metafile) representation, *Common Ground* from No Hands Corporation uses Macintosh PICT files while the *Replica* system from Farallon Computing uses GDI and Quickdraw primitives from the PC and Macintosh environments respectively. These last two approaches have the advantage that the use of system primitives from the Macintosh and the PC makes it easy to capitalise on the windowing display software already present in the operating system. On the other hand it also means that they are tied to specific hardware platforms in the first instance.

All three of these examples suffer the drawback that the representation used for displaying a document on the screen does not bear any immediate relationship to the one used in printing out the paper form of the document on a laser printer. In this respect the strongest candidate for a low-level representation would be some new development from PostScript, which was first produced by Adobe Systems Inc in 1984 and has established itself very rapidly since that date as the industry-standard page-description language (Adobe, 1990). It is available on a variety of output devices with resolutions ranging from 300 dpi on laser-printers to 2000 dpi (or better) on typesetters.

Late in 1992 Adobe Systems announced a new product called Acrobat which was based around Level 2 PostScript but with many extra features including file compression options, a 7-bit ASCII representation and hypertext links. This interesting development will be discussed more fully in section 5.

4. THE EP-ODD JOURNAL

In 1987 I founded a journal called *Electronic Publishing — Origination, Dissemination and Design* (*EP-odd* for short) and became its Editor-in-Chief. This journal is published by John Wiley Ltd and it appears four times per year (Brailsford, 1989, pp. 482-493). The articles cover topics ranging from hyphenation to hypertext and from typography to SGML tags. The contributors and subscribers are, in the main, computer scientists interested in Electronic Publishing and professional practitioners from the print and publishing industries.

It might be thought that a journal with such a title would be refereed and disseminated electronically from the very outset, but there were a number of predictable difficulties. My US co-editor and all of our editorial board were anxious to create a solid and highly-respected journal. There were severe doubts in the early days (and there still are) as to whether any electronic journal, however scrupulously its papers were refereed, could compete in prestige with the existing traditional journals. Moreover there was the practical problem that the computer hardware and software used by our subscribers was split among three systems (IBM PC, Macintosh and UNIX), so how could we provide viewer software which retained compatibility with these three systems and with the PostScript that we were determined to use for the printed version? There was no immediate solution to the difficulty. The screen version of PostScript (called Display PostScript) was considered for a time but had to be rejected because of unwieldy file sizes and performance problems with early releases of the software. In the years from 1988 to 1992 we made extensive use of electronic mail for exchanging referees' reports, production schedules and so on. In addition the co-editors and production staff exchanged draft papers either in PostScript form or in either of the supported source text formats (LATEX or UNIX *troff*). All we were able to do in planning for a truly electronic journal was to save the source code and PostScript for every published paper and to draw up a ’wish list’ for our ideal electronic format which needed to have the following properties:

(a) It could not be based simply on existing PostScript with no changes. Some of the reasons for this have been given already but another issue is that the PostScript coding of a book or a magazine article is of little use to subscribers. All they could do would be to print out the material they already have in their hands. There is little ’added value’. We wanted a ’distilled’ version of PostScript which would be free from printer-specific details. Existing compression techniques for text, colour images and animated video (LZW, JPEG, MPEG etc.) needed to be in place, so that file sizes could be reduced.

(b) The format had to allow for hypertext links and give support for browsing, automatic indexing, page annotations etc. without compromising, in any way, the ability to print the document out to any laser-printer which implemented popular languages such as PostScript, Hewlett-Packard PCL etc. The chosen format ought to be an open system with its specifications in the public domain.

(c) The format needed to be searchable for keywords etc. and be 'close' in some sense to the printable PostScript, thereby giving us resolution independence. We were **not** interested in resolution-dependent bitmap page images (e.g Group 4 FAX) with proprietary search software linking these to some separate ASCII version of the material. However, viewer software had to be easily available to enable our 'electronic' version to be browsed. These viewers needed to be fast and available on a wide variety of platforms. There had to be full support for working in colour.

(d) The area of Electronic Publishing overlaps that of multimedia systems. Looking to the future it was important that any adopted format should have 'placemarkers' for multimedia inserts (sound, animated video etc.) even if the multimedia facilities were not yet fully in place.

(e) In addition to text and image compression options there had to be a '7-bit clean' capability to allow for network dissemination of material via File Transfer Protocols (FTP) (Gien, 1978, pp. 312-319) or electronic mail (e-mail).

The preliminary announcement of Adobe Acrobat, late in 1992, seemed to satisfy just about all of these requirements and a project was set up (later called CAJUN — see section 6) to use this format for disseminating *EP-odd* over networks and on CDROM.

5. WHAT IS ACROBAT?

Acrobat is based on a new Portable Document Format (PDF), developed by Adobe, which remains close to Level 2 PostScript but has a range of compression options available to reduce file sizes (Adobe, 1993). At the moment it is a fixed page format, but there is the possibility of a revisable form at some later stage. There is also a prospect of being able to include video and audio inserts in future releases. Various 'byte accelerator' techniques have been included in the format so that searching for words and phrases is fast.

There is an option in Acrobat for LZW compression on text which gives a compression factor of about 2:1 but bigger gains are obtainable on images, particularly those in colour, where JPEG compression can achieve some spectacular reductions in the region of 10:1.

The novel features in PDF are a set of facilities for 'hot links', 'thumbnail' icons of pages, chapter outlines and page annotations. The first of these — the hot links — provide a gateway into *hypertext* facilities and these can be further developed into hypermedia facilities as appropriate standards emerge. Hypertext links allow for references within a document, and from one document to another, to be 'hotlinked' so that a direct jump can be taken to some destination page. By way of example, a phrase such as '... see reference [22] ...' might be highlighted on the screen in some way. Positioning the mouse on that phrase and clicking might cause a jump to the Bibliography page of the document, where reference number 22 will now be highlighted. A further click on that highlighted reference might cause the book or article in question to be opened up on the screen (provided this material is also on the same CD-ROM or is available in some way from a hard disk or across a network). The wider idea of bringing multimedia material into hyperdocuments would be exemplified by being able to click on the name 'Mozart' to obtain a picture of Mozart, or a digitised sound insert of an aria from *The Magic Flute* or even an animated video insert, in a separate window, of some scenes from the film *Amadeus*.

Turning to the 'chapter outlines' (or 'bookmark') feature we find an ability to create a hypertextual Table of Contents for a book or an article. All sections and subsections can be entered into this hierarchical outline, each entry of which is a link to some predefined view of a fixed page within the document.

The thumbnails for the document pages are miniaturised JPEG-compressed bitmaps of the pages — each one is unique and there is enough detail to be able to recognise a page from the layout of its thumbnail. The thumbnails can be optionally displayed down the left-hand side of the screen and they greatly facilitate fast browsing and random access: one can jump from the page displayed on the screen to any distant page by clicking on the appropriate thumbnail for the destination page. Finally, the page annotations are an electronic version of the 'yellow stickers' that are commonly attached to paper-based memoranda.

All of these annotations and other 'hyperfacilities' are kept separate from the underlying formatted material and they enable any document to be ’personalised’ with the user's own marginal notes and cross-referencing hyperlinks.

PDF has a set of markers for these new hyperfacilities, which can either be added to the PostScript after it has been produced or can be passed down from ’front-end’ text-processing packages into the final PostScript. This is effected by means of a new PostScript procedure called pdfmark.

5.1 Acrobat Viewers and the Distiller

Acrobat is currently supported on IBM-compatible PC and Macintosh platforms, with DOS and SUN Solaris/X-windows versions at the *alpha* test stage. Apart from the Distiller and PDFWriter, two versions of Acrobat viewer software are available. The *Reader* provides facilities for browsing existing PDF documents and for printing them out. If the document has already been enhanced with hypertext links these can be followed but they cannot be altered in any way. By contrast the *Exchange* version of the viewer permits a degree of editing with respect to the various ’hyperfeatures’ and it also allows complete pages from other PDF documents to be interleaved with those already present. In what follows we shall use the general phrase ’viewer’ to mean either of the Reader or Exchange versions. Note that PDF does not, at present, allow the underlying formatted text to be altered in any way — it is a fixed-page format.

A program called the Distiller converts PostScript into PDF, carefully transforming any pdfmarks into the corresponding PDF hyperfeatures as distillation proceeds. However, if the text-processing software in use cannot produce pdfmarks directly then they can be added ’by hand’ during the distillation process or from the Acrobat viewers. Figure 1 shows the stages in creating a PDF document for Acrobat use, starting from any DTP or page-makeup software that can produce PostScript.

Interestingly, it is possible to produce PDF directly from word-processors or DTP packages without having to go through a PostScript intermediate stage. This is done by adding an extra PDF printer driver into the PC or Macintosh environments. Figure 2 shows the stages involved.

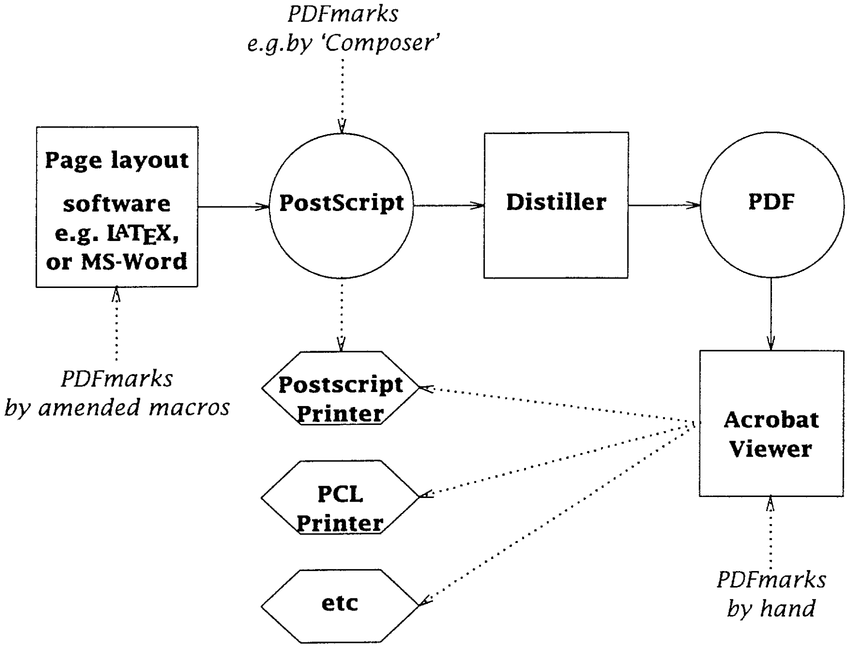

Figure 1: From source code to PDF via PostScript

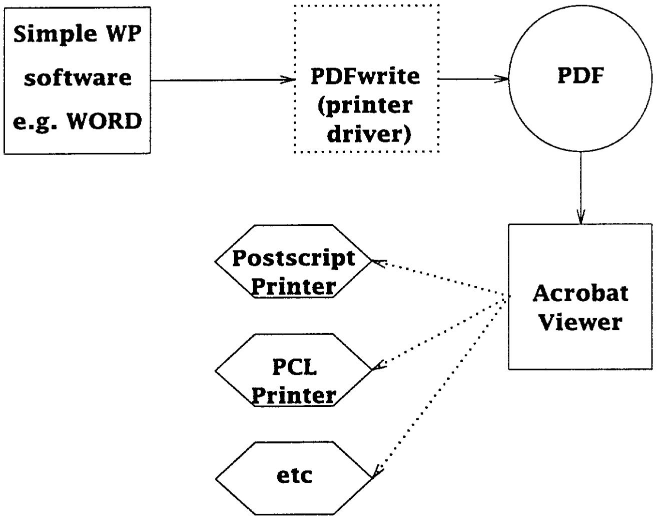

Figure 2: Direct route from application to PDF

6. THE CAJUN PROJECT

Having decided that Acrobat was a very plausible candidate for disseminating high-quality platform-independent journals, funding was obtained from John Wiley Ltd and Chapman & Hall for a two-year project to be carried out in the Electronic Publishing Group at Nottingham. This project is called CAJUN (**C**D-ROM **A**CROBAT™ **J**ournals **U**sing **N**etworks). In addition to the Wiley *EP-odd* journal, Chapman & Hall have contributed material from their journal *Optical and Quantum Electronics* (OQE) and have committed themselves to producing a new journal, *Collaborative Computing* (CC), using PostScript and PDF formats.

The production methods for these journals vary considerably. OQE and CC are produced using Advent Systems' 3B2 software. This package produces PostScript output which distils with no problem but hyperlinks, for the moment, have to be added by hand using the Exchange version of the Acrobat viewer. By contrast, the recommended formatting packages for *EP-odd* have always been LATEX (Lamport, 1986) and UNIX *troff* (Ossanna, 1976, Kernighan, 1982). It is a cause for amazement and incredulity, in some quarters,that computer scientists are happy to work with such lowest common denominator, 'stone-age', software rather than insisting on Word, PageMaker or Quark Express. But the fact is that technical journals *do* have a predictable structure to their content and, unlike fashion magazines, say, they do not require large 2400 dpi full-colour photographs set at bizarre angles inside irregular text layouts! In such circumstances the mathematical and tabular typesetting facilities of LATEX, coupled with facilities for inserting Encapsulated Postscript, are greatly valued.

One of the tasks of the CAJUN project has been to adapt existing *troff* and LATEX macros for the test journals provided by the two sponsors so that, for example, a call-out of a reference can automatically pass down a marker tag, corresponding to a PDF 'hot-link', from this call-out to the Bibliography page. The great advantage of using software such as LATEX and *troff* is that the macro sets can easily be 'got at' for adaptation. Some of the more black-box-like DTP packages will be rather harder to adapt for Acrobat and users will need to have the job done for them. Fortunately, the vendors of the major packages are already working with Adobe to bring this about.

6.1 Dissemination on CD-ROM and Over Networks

The CAJUN project has already created a test CD-ROM consisting of about 40 papers in PDF form. By February of 1994 we hope that to see the first formal release of this product, which at that stage will be up to date with the paper issues of the journal and will have the first 6 volumes in electronic form. A single CD-ROM can hold about 600 Mbytes of information which makes it an excellent archiving medium. The decreasing cost of CD-ROM drives will enable libraries to receive regular updates of 'electronic' journals in this form. This could act as a complement to networked dissemination, or as an alternative to it, if network bandwidth is limited.

Turning to network dissemination, we can identify two categories which we can denote as 'push' and 'pull' methods. Pushing is the sending of information by the publishers to subscribers. Pulling describes the acquisition of documents by journal subscribers who log on to a service provided by the publishers.

Various software tools are available to enable a 'push' service to be implemented and a hierarchy of files to be set up and maintained on the subscriber's computer. The main drawback is that users have to make sure that they possess enough disk space to receive information forced upon them. For this reason we are also experimenting with network information retrieval (NIR) tools to enable users to browse information and download whatever they wish. It is, of course, perfectly possible to use FTP in 'pull' mode to acquire information in this way but another tool which is potentially more flexible is Gopher (Anklesaria, 1993) which allows publishers to run a server which can serve individual files to remote clients.

An experimental 'pull' service is now available on a journal server at Nottingham which is available over the Internet. Six files in PDF form are currently available from the *EP-odd* journal and these will shortly be supplemented by six more papers from OQE and CC. To make use of this service the user needs an Acrobat viewer of some sort, access to the Internet and a copy of either Gopher or FTP. The details for access over Internet are:

Gopher: arrow.cs.nott.ac.uk [128.243.23.11]

FTP: marian.cs.nott.ac.uk [128.243.21.16]
/ep/pub/pdf

6.2 Use of PDF by Referees and Editors

One of the problems that has faced journals that have tried to introduce electronic refereeing is that the ASCII text may convey only a small part of a paper's significance. To go beyond this stage has required, until now, that a journal should use some particular text processing package and that all editors, referees and production staff should possess this same software at the same release level.

The ability to send a formatted paper over electronic mail in PDF form gets around this problem completely. The only piece of software that referees require is an Acrobat viewer and a draft version of a paper can then be circulated, regardless of the front-end software used, provided that the paper can somehow be processed into PostScript and distilled. At the later stages of production, once a paper has been revised and accepted, PDF becomes even more valuable. Proof copies of the paper can be e-mailed from production staff to editors, and vice versa, to check that 'widows' and 'orphans' have been removed and that figures and tables are correctly placed. Better still, the 'read only' nature of PDF prevents authors and editors from disturbing the page layout that the production department has decided on. Final revisions to the document can be effected by affixing 'yellow stickers' to various pages, with suggestions as to how the layout might be improved, and then returning the entire electronic document to the production staff for further revision.

7. CONCLUSIONS

Acrobat rates as one of the most significant happenings *ever* in the field of Electronic Publishing. Its potential importance outstrips that of PostScript and yet the attraction of Acrobat as some form of 'electronic' standard lies in its very closeness to PostScript. This means that hard copy is easy to generate and one can distill existing PostScript archives.

Acrobat marks a turning point for publishers and librarians. In the next ten years the publishing business will change out of all recognition with 'electronic' versions of books and journals becoming ever more important. Libraries need to continue their efforts in adapting their organisations to become thoroughly imbued with computing and networking know-how. This is not to say that paper documents **will** go away — it's doubtful if this will ever occur. But increasingly a paper document will come to be seen as a useful two-dimensional snapshot, at a given point in time, of a much richer electronic document. Journal subscriptions may soon be taken out for the electronic version of a journal (perhaps with some kind of campus-wide site licence to permit copies of articles) with the material itself being accessed from some local or remote journal-server machine. The printed version may come to be seen as just a glossy adjunct that appears several weeks, or months, later. Certainly among the academics I have canvassed the view seems clear: printed journals will be needed to browse through in the coffee room and to confer legitimacy on that journal (for the immediate future at any rate). But the electronic version is what will be needed for research and for building up much-needed articles into an electronic reference library.

The CAJUN project has given valuable insights into the use of PDF for journal dissemination. There is no doubt that PDF, even in its current form, goes a long way towards meeting our criteria for a portable format which encompasses both the concrete and the abstract features needed for electronic documents. Indeed, the fascination of PDF for computer scientists is that it establishes *data structures* and a *programming language* for describing documents with the Acrobat viewers being *virtual computing machines* for inspecting them. This idea of documents as computer programs may sound rather terrifying to publishers and librarians still struggling to cope with computerisation of existing, paper-based, publications. But all of us — publishers, librarians, computing companies and computer scientists — need to collaborate on the next big step forward in electronic publishing which is to bridge the gap between paper documents and electronic documents in a manner which is independent of any particular hardware or operating system. I believe that Acrobat can help us do all of these things and those of us involved in the CAJUN project will be doing our best to get PDF and Acrobat established as usable and useful standards within the publishing industry.

REFERENCES

Adobe Systems Incorporated, *PostScript Language Reference Manual,* Addison-Wesley, Reading, Massachusetts (December 1990). Second edition.

Adobe Systems Incorporated, *Portable Document Format Reference Manual,* Addison-Wesley, Reading, Massachusetts (June 1993).

Anklesaria F., M. McCahill, P. Lindner, D. Johnson, and D. Torrey, *F.Y.I on the Internet Gopher Protocol*, Memorandum - University of Minnesota, March 1993.

Brailsford, D. F. and R. J. Beach, “Electronic Publishing —a Journal and its Production”, *Computer Journal* **32**(6), pp. 482-493 (December 1989).

Gien, M., “A File Transfer Protocol (FTP)”, *Computer Networks* **2**, pp. 312-319 (September 1978).

Kernighan, Brian W., “A Typesetter Independent TROFF”, *Computing Science Technical Report No. 97*, Bell Laboratories, Murray Hill, New Jersey 07974 (March 1982).

Lamport, Leslie, *LATEX: A Document Preparation System*, Addison-Wesley (1986).

Ossanna, Joseph F., “NROFF / TROFF User's Manual”, *Computing Science Technical Report No. 54*, Bell Laboratories, Murray Hill, New Jersey 07974 (llth October, 1976).

Story, Guy, Lawrence O'Gorman, David Fox, Louise Levy Schaper, and H.V. Jagadish, “The RightPages Image-based Library for Alerting and Browsing”, *IEEE Computer*, pp. 17-26 (1992).