Theory into Practice: working with SGML, PDF and LaTeX at Elsevier Science*

Martin Key

Elsevier Science Ltd, England
m.key@elsevier.co.uk

1 The Company

While I do not want to make this article a plug for Elsevier, it is first necessary to put our activities into context. Therefore, for those who do not know us, Elsevier Science is part of the Reed Elsevier Group and, in terms of number of journals, is by far the largest publisher of scientific journals in the world. The original Elsevier Company was Dutch based, but now, through acquisition and merger, is an international company with offices in the Netherlands, UK, USA, Switzerland, Eire and the Far East. We publish well over 1,000 scientific, technical and medical journals covering all sections of academe and business.

2 The move into electronic publishing

Elsevier's major customers are academic and research institutes throughout the world. Traditionally, academic publishing has relied on authors submitting papers via external academic editors who arrange for the necessary peer reviews. Once accepted, papers are sent to Elsevier for copy-editing, typesetting and compilation into issues. As a result we have in the past received paper manuscripts of varying levels of presentation from around the world. Over the last 10 years it has become apparent that most authors use some form of word processing or computer generated text to prepare their papers. To have these papers typeset means rekeying the manuscript and, what is worse, ending up with electronic files produced by many types of typesetting equipment and software with minimal chance to reuse this material at a later date. For some years the Elsevier Group have been looking at ways to avoid rekeying manuscripts whilst at the same time automating the production process, produce proofs more quickly and create electronic files for multiple use in the foreseeable future.

After many surveys, experiments and discussion groups it was clear that Elsevier should work to accepted international generic standards in order to achieve these goals. The major standards agreed on were Standard Generalised Mark-up Language (SGML) for text, Tagged Image File Format (TIFF), Joint Photographic Experts Group (JPEG) and Encapsulated PostScript (EPS) for graphics and PostScript, and the Portable Document Format, (PDF), also known as Acrobat, for pages. Unlike typesetting codes, SGML does not drive any particular application but can be readily converted to numerous formats for typesetting on paper, database applications, CD-ROM and so on. It is therefore an ideal archive medium. TIFF, JPEG and EPS are well documented graphic file formats and are widely supported in terms of external applications. PDF is, perhaps, a risk in that it is the property of a commercial developer (Adobe) but its great flexibility and rapid acceptance by professionals and the academic community, together with the track record of PostScript itself — now a de facto standard — makes its long-term future seem relatively safe. The decision by Adobe to make the Acrobat reader available free-of-charge is another positive sign.

3 The concept of Computer Aided Publishing (CAP)

Once the standards were agreed the process known internally as CAP (Computer Aided Publishing) took clearer shape. There are a number of activities which form part of CAP. These include the following: the converting of manuscripts and artwork into electronic files; structuring of text with SGML; editing on screen; automatic proofing; moving and maintaining files on a network; creating SGML (text) and graphic files; receiving PDF files from our typesetters. In addition, a number of journals receive, and use, papers in LaTeX format which will be discussed later.

4 Practicalities: How we do it

CAP started in Elsevier in January 1994, in both Amsterdam and Oxford, with a limited set of journals. The number of journals has been increasing rapidly and in 1995, as software and hardware stabilises, the number of journals is being increased dramatically.

The first action when receiving a paper, either on paper or disk, is to log the information on to our production tracking system. All the important details are recorded — title, authors, number and type of graphics, whether it is available on disk etc. This record follows the manuscript throughout its production process and is updated at each stage of its progress through the system. Elsevier encourage authors to submit on disk, and the numbers are rising. If it is on disk it is initially converted to our standard CAP format which

*Reprint from the Annals of the UK TeX Users Group **Baskerville**, Volume 5.2, March 1995. Published with permission of both Baskerville editor and author. Presented at the UK TeX Users Group conference 'Portable Documents: Acrobat, SGML, and TeX', on 19 January 1995, London, England.

allows it to be used by our SGML tagging and editing tool — Pandora — which was developed by staff working in Amsterdam. If it is only available on paper it is either OCR (Optical Character Recognition) scanned and then converted into the CAP format or, if the paper is too complex for scanning, it is keyed by off-shore keying agencies. Whatever the route, it arrives at our Pre-Edit Department in the generic CAP format. Simultaneously graphics are scanned — TIFF for line art and JPEG for half-tones — or redrawn and saved as EPS in some instances.

The text is then tagged using Pandora. The Document Type Definition (DTD) used is the Elsevier DTD (which Elsevier has made publicly available subject to certain conditions) which is fairly complex covering not only text but also tables and mathematics.

After coding and parsing, the text is loaded onto the network server, together with the graphics, using an in-house developed Document Management System which monitors, names and controls the files. As one article can produce more than 20 files, with an average issue of a journal containing 10 articles, the number of files can quickly mount making such management essential. Once the files are on the server, they can be retrieved by the Production Editor who will then edit the article for style, spelling, grammar, etcetera and add any additional tags necessary. Graphic files are also checked at this stage to ensure that the correct graphics are linked to the relevant caption. The file is then parsed again to check its validity. Author proofs can then be produced and, once they are received back from the authors and corrections made, the final SGML and graphic files are exported to the typesetter for making up the final pages.

We expect typesetters to retain the validity of the SGML files when producing the pages, and this is strictly monitored. Due to the complexity of the DTD and the relevant inexperience of most typesetters in using precoded SGML files, we have to work with our typesetters quite closely, answering specific queries and offering advice where necessary. However, we do not expect to develop the systems for the typesetters — that is their responsibility. The final, additional requirement we demand from our typesetters is that they supply each individual article, and other elements of the issue, in PDF format. This means that they must have a PostScript setter in order to create these files.

5 TeX and LaTeX

In some disciplines TeX and LaTeX are used extensively by authors and, not unnaturally, they would like to submit their articles in this format. Experience has shown that this can be hard work for the Publisher. In some cases, hacking in to such a file to find out how the author's carefully developed macros have been used can be very time-consuming and, in some cases, can take considerably longer than having the paper professionally typeset. However, whenever possible, we will try and use submitted LaTeX files and, to a lesser extent, plain TeX files. However, Elsevier encourage authors to use the Elsevier style file which produce a pre-print type output. This style is then replaced with the journal-specific style file which makes the Publisher's task considerably easier. The Elsevier style files, together with the instruction manual, are available from the three CTAN sites or direct from Elsevier.

LaTeX has a number of advantages. Pages in camera ready format can be produced readily in-house without recourse to a typesetter, and PDF files can also be generated from the dvi files. Recently, the Production Methods Group at Elsevier Science Ltd has further developed the 'dvihps' converter and LaTeX macros from the HyperTeX project, to fully retain the hypertext links available in the LaTeX file, as well as generating automatic 'bookmarks' or contents list, directly into the PDF file. In order to meet the full CAP requirements previously mentioned, there is one final part of the equation to be completed — a LaTeX to SGML conversion. Due to the complexity of the Elsevier DTD this is not a simple task but work is currently taking place to see how far down this road it is possible to go.

6 Practical Problems

As with most technical developments there are always problems to be addressed. In the case of CAP they have been surprisingly few. The major problem experienced at an early stage was the lack of SGML editors which could cope with the Elsevier DTD, particularly in the area of tables and mathematics. This problem has been largely resolved by the development of Pandora, a tool which has far exceeded its initial specification as a package which would enable compuscripts to be handled by typesetters. The second problem was one of logistics — how do you train Production Editors to work with SGML on screen editing whilst simultaneously producing journal issues? As previously mentioned, there is also the increased demand we place on typesetters, many of whom have had limited experience of handling complete journals in SGML. Finally, as Production Editors began to use the DTD in earnest, additional requirements are discovered which means that the DTD must be further developed. As a result, the DTD has become a moving target with more complex requirements being asked for almost daily.

7 The Future

Some people may ask why we are putting ourselves through so much pain. Is it worth it? The market is demanding electronic products in addition to, and sometimes instead of, the traditional paper ones. For those publishers who have tried to use typesetters' tapes for such products, the answer is clear. The availability of generic coded data which can be manipulated in multifarious ways is clearly the route to take. In addition to meeting the demands of our market, we are also satisfying the demands of our producers — the authors — who create 'electronic' versions of their articles and who naturally expect that we, the Publishers, should be able to use them. Finally, the Production process itself is being streamlined allowing for more efficient and faster production times.