BIJLAGE T

SGML and TEX at Elsevier Science Publishers

Jeroen Soutberg
Elsevier Science Publishers
Amsterdam, The Netherlands

Augustus 31, 1990

Contents

Introduction

SGML and TEX are in use at Elsevier as production tools for an increasing variety of products. In the following we describe the prevailing trend in publication technology, and the roles SGML and TEX play in the developments at Elsevier.

ESP

Elsevier Science Publishers is one of the larger scientific publishers in the world. We publish about 600 journals and annually about 600 books in several fields of science. The publications appear under two imprints: Elsevier and North-Holland.

Elsevier

The Elsevier imprint is used for publications in biology/medicine, chemistry, agricultural and earth sciences. Well-known examples are:

- Biochimica & Biophysica Acta,
- Journal of Chromatography,
- Ecosystems of the World (book series).

North-Holland

The North-Holland imprint is used for publications in physics, materials science, engineering & design and mathematics and linguistics. Examples are:

- Nuclear Physics (A, B),
- Surface Science.

Manuscript routing

The publishing world is in transit from a state of established procedures and practices to a state of constant adjustment to the technical and procedural opportunities arising from technological developments. These changes have a profound effect on the routing of manuscripts in the publication process.

We give a simplified description of this transition by splitting it up into three stages:

- The traditional situation;
- The present – transitional – situation;
- The future: database publishing.

Manuscript routing – traditional

Traditionally there have been two ways of producing a printed publication: (1) Preparing a typeset publication from the manuscript submitted by the author, and (2) printing a publication from camera-ready material prepared by the author.

These scenarios yielded quite different print qualities, but correspondingly different costs were involved. These differences made it relatively easy to choose a scenario for a project, depending on its specific requirements.

The traditional scenarios are shown schematically in figure 1.

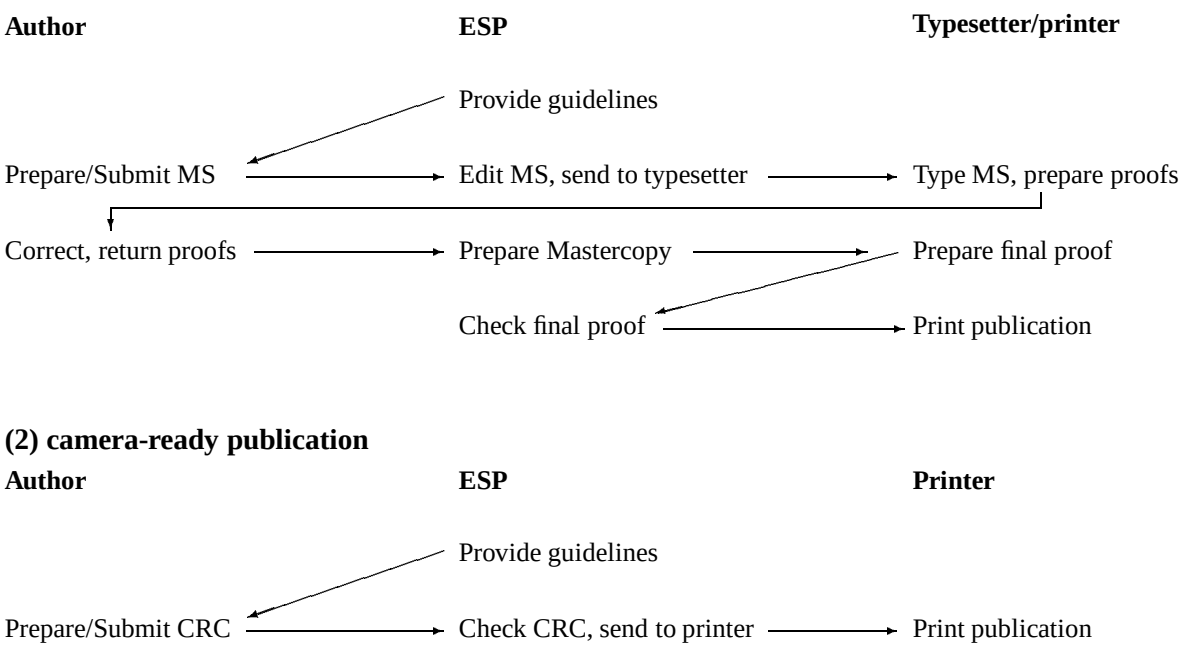

Figure 1: Traditional scenarios for printed publications

Manuscript routing – present

The advent of electronic text processing (in typesetting systems as well as in 'word processors') and the availability of laser printers has upset the clear distinction between the traditional scenarios.

On the one hand, the CRC output an author may produce using a text processing system or a DTP package is much better than in the past, and it may even approach the quality of typeset material. On the other hand both authors of traditional manuscripts and authors of CRC material may want to submit their information in electronic form, in the first case for saving proof-reading time or accelerate publication, in the latter because they do not have a high-quality output device available. In both cases the publisher will have to process an electronic 'manuscript', possibly with media and/or code conversion. Also, the boundary between publisher and typesetter becomes less clear (who does the conversion, who edits the files, . . .).

Another trend in publication is the preparation of secondary information from primary information and the re-use of primary information in new products. It is desirable for these purposes to have the primary information available in electronic form.

Figure 2 shows the schematics for typeset and CRC material with the conversion stage and re-use taken into account. In comparison with figure 1 the distinction between typeset and CRC material is much less clear; theoretically it is now even possible that a typeset and a CRC publication end up following the same routing. Also, these are not the only products anymore.

Manuscript (information) routing – future

Extrapolation of the trends sketched above leads to a situation where information reaches the publisher in many different formats, and is processed into many different end products. The simple distinction between typeset and CRC material then loses its relevance.

It is not feasible to construct bridges ('translation programs') between all of these input and output formats. Apart from the number of programs required (number of input formats $\times$ number of output formats) the need to store information for future re-use would imply storage facilities for the same number of formats.

The ultimate solution to these problems is the concept of 'database publishing': the publisher has all incoming information converted into *one* intermediate format; the information is stored in a database in this format, and all end products are manufactured by making a selection from the database and converting this information into the appropriate format. In this way we have one standard storage format, and the number of translation programs needed reduces to the sum of the number of input and output formats. Figure 3 presents a schematic picture of this concept.

The publisher has now become an 'information broker'

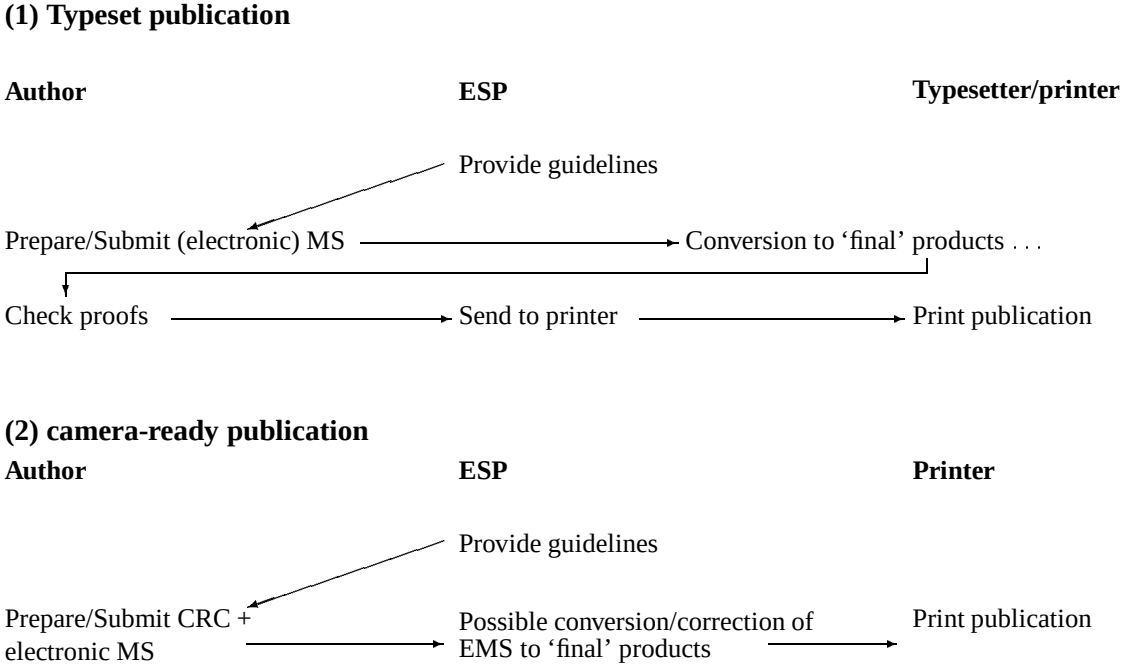

Figure 2: Transitional scenarios for printed publications

with all its valuable material stored in the database. Correspondingly, an author now submits information for the database.

The fact that many different outputs are to be generated from the intermediate format poses strict requirements for this format:

- It should be *presentation-independent* (e.g., the possibilities to identify a heading by its presentation are quite different for printed publications and on-line applications, so the database should contain the heading identified as such; not less, not more).
- It should be *consistent and clearly defined* (e.g., every identification should identify the same, well-defined piece of information).
- It should be *internationally accepted* (to make it possible to exchange information with others).
- It should be *implementable* (e.g., software and hardware available).

The requirements for the intermediate format affect the instructions to authors as well: now the material produced should be in a format which contains sufficient information to generate the intermediate format. This could mean, for instance, that references should be submitted in a special format which can be converted automatically for the database.

Role of SGML – the intermediate format

Over the last few years SGML has emerged as the primary candidate for storing information in the manner described in the previous section. It goes a long way to meeting the requirements for the intermediate format introduced there. In a different order than above the requirements are met as follows:

- SGML is internationally accepted, an ISO standard.
- SGML is implemented in processing software like Softquad, Writer Station, et c.
- To a large extent, SGML is consistent and clearly defined; over the last year we have found some problems which we dealt with by creating preferred versions of ambiguous solutions for use within Elsevier.
- SGML is meant to be presentation independent. In the last few years a discussion has been going on within Elsevier about the question how far one should go in this. For example, the simple equation $s = v_0 t + \frac{1}{2} a t^2$ cannot be presented in presentation-independent form unless the entire formula is spelled out. Even then it will be difficult to make the factor $\frac{1}{2}$ presentation-independent. The same problem arises in all instances where *for reasons of efficiency* a special presentation was introduced for identification purposes. Our present projects adhere to presentation independence down to the level of the *structure* of the information.

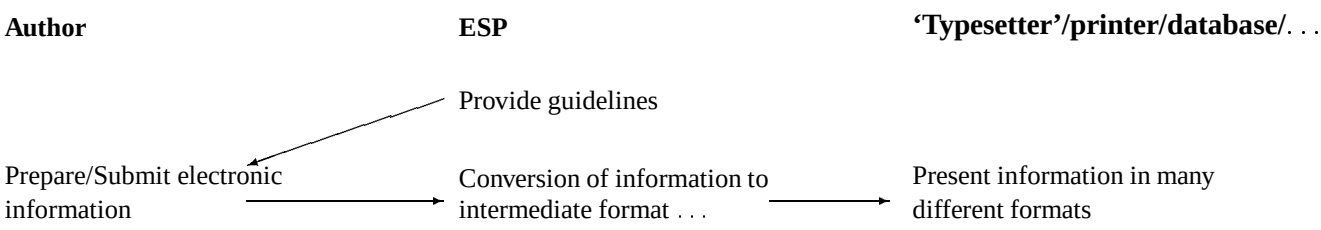

Figure 3: Future scenario for information processing

Role of TEX– Input format, output format

At Elsevier, TEX is used mainly for two rather different classes of applications:

- Where SGML is used as the intermediate format we use TEX as (one of the) output facilities. This permits us to generate, among others, high-quality paper output of proof material formatted in such a way that the information loaded into a database can be validated without looking at SGML codes.
- TEX is also used as the combined input–output format in projects where compuscripts from authors are processed in TEX all the way through. The conversion of TEX to other formats – which needs considerable manual correction – is thus avoided. Also, high-quality output can be generated locally.

Projects

As examples of the different classes of applications mentioned in the previous section we describe two ESP projects representing these classes

CAPCAS

The abbreviation, CAPCAS, stands for Computer-Aided Production of Current-Awareness Services. Although the system can do more or less what its name implies, its status has changed with developments. It is now intended to be a first step towards database publishing.

The scope of CAPCAS is at present limited to the storage of the Heads of journal articles in a database. (By Head we mean the bibliographical information and the abstract at the beginning of an article). Article tails and bodies will be dealt with later. As explained above, the information in the database is stored in SGML format.

Input is provided by our typesetters. This is at present the easiest input route, because a vast majority of articles are still being submitted on paper and the information becomes first available electronically after key-in by the typesetter. The typesetter prepares both the files for the typesetting equipment and the SGML files for CAPCAS from the key-in.

Output is generated for in-house and external applications: in-house for validation of information loaded into the database, contents lists and indexes, externally for secondary services.

- The validation of information loaded into the database concerns mainly the correct identification of fields in the Head, especially those which do not show up in the printed article (e.g., city and country are not typographically distinguished from the organization name in the affiliation). After loading the Head of an article into the database a so-called 'load report' is produced using TEX, whereby all the SGML tags generate unique layout features. As a result, the fields may be identified by simply looking at the layout instead of checking the positions of tags.
- Contents lists and indexes are generated from the database in SGML format. One of the present options is to produce output with the help of TEX, making use of the typical TEX 'programming' possibilities for handling information.
- As yet, no products are being delivered to external customers: we only just started loading the database. Also, before any products may be delivered the exact specification has to be established in consultation with the customer.

Future developments will be the incorporation of the other constituents of articles: tails (i.e. reference/note section), and (ultimately) the bodies (main text, including formulae, tables, figures).

We will start working on the tails fairly soon (probably in the coming year); the body, by sheer size and complexity, will not be tackled for some time to come.

LATEX compuscript processing

Similarly to other publishers, there are several Elsevier journals where most manuscripts received are coded in TEX or LATEX. A project has been set up to create an article routing in which the electronic version of the manuscript is processed, keeping within TEX, and the output is produced with TEX as well; proofs are printed on a laser printer while the final output is produced in high resolution on the in-house typesetting machine. Special software makes it possible to generate output using 'Times' fonts instead of cm fonts.