Review article

Author-printer harmony with SGML

J. D. Painter

Ever since computers came to be applied to composition there has been a case for generalized markup. The fact that we could generate a text on one machine and then expect it to be processed on any of a large number of others has posed authors, editors and publishers with a dilemma that it has taken the best part of twenty years—perhaps longer—to resolve. Two dilemmas, in fact. Not only how to mark it up so as to be sure that its form shall exactly reflect its content; but who shall mark it up—people up- or down-stream of the publisher? The author knows the content but not the typographical niceties of presentation. The compositor knows Bodoni from a Cicero, or Monotype from Autologic, but not what the text means. It has taken a long time to agree a resolution.

I recall having my attention drawn to something called generalized markup in the early eighties when we at Cimtech were investigating, on behalf of the British Library's Research & Development Department, ways of interfacing word processors with photocomposition equipment. I believe the first intimation came from a piece in IBM's Journal of Research & Development by someone unknown to us then, Charles Goldfarb. He proved to be one of the prime movers of the Standard which at long last has seen the light of day as BS 6868—Standard generalized markup language (SGML) for text and office systems (ISO 8879).* I say one of the prime movers. In 1983 there appeared a document whose purpose was, to use its own words, to enable the 'international printing and publishing industries (to undertake trials) during the process of review and adoption of it by ISO'. The publication's title was 'Document Markup Metalanguage—GENCODE™', and it was published by Graphic Communications Association (GCA)—an affiliate of Printing Industries of America, Inc. It was the work of a dedicated few (of whom Charles Goldfarb was one) who had been striving to this end since William Tunicliffe, a founder member of GCA, had introduced the concept of generic coding in 1967.

That it should have taken 16 years to move from an idea to a draft Standard (twenty years to a published one) is a measure of its authority. The two words we find most frequently in commentaries on work-in-progress have been 'rigorous' and 'unambiguous'. The authors of GCA set out to establish, in the words of Tunicliffe, 'a syntactical legitimacy that would provide a solid, logical foundation for a lasting, logical implementation'. Syntactical legitimacy is what we now have.

Every bibliophile understands markup. The change from eight-point to ten-point; from roman to italic; from medium to bold; from one typeface or measure or leading to another—all these aspects of presentation need commands in some vocabulary or other. When machines, especially electronic machines, are used for composition and the human contribution is minimized, such commands must be precise and unambiguous. Hitherto, each manufacturer of a composition system has contrived his own vocabulary. If you wanted to drive a Monotype, you needed to know the Monotype's codes. If you did not know whose machine was to be driven (as was often the case) you did not know what codes to use, and your markup would have to be translated by someone who did.

Since it is authors, or perhaps their editors, who understand best the structure of texts, Tunicliffe's notion was that there should be some means of identifying structural elements such that their presentation could be controlled, regardless of the environment in which the presentation was to be effected. We all knew that a formal document was an organized hierarchy of parts. We knew that a book comprised prelims, body matter and back matter. We knew that body matter comprised chapters and that chapters comprised headings and paragraphs and maybe figures and notes. We knew that figures comprised image data and captions. No-one knew it better than the author. He may not have known, or indeed have cared, that figure captions appear two points smaller than the body matter, or that they would one day be typeset on Snodgrass Ltd's machine, but he did know that they were figure captions. Why could we not have some way of stating, in the text, the fact that 'this is a figure caption, regardless of how figure captions are to be presented and regardless of whose devices shall be used to make this text viewable'? BS 6868 says we can. 'Markup' says the Standard 'is text that is added to the data of a document in order to convey information about it.' Standard Generalized Markup comprises four quite
separate kinds of information—tags, entity references, declarations, and processing instructions. Of all these, perhaps the most fundamental, if not the most important, are the tags, which are used to identify the separate elements which go to make up the logical construct we choose to call a document. The document type is irrelevant. It might be a memo, a press release, a maintenance manual, a journal, or a whole book. Because it is generalized, SGML provides us with an unambiguous way of identifying elements by using tags which can be translated into processing instructions. So, you might say, given this document type, a title shall be set thus; the author’s name thus; a chapter heading thus; and so forth for every identifiable element into which any conceivable document may be decomposed. Other types comprise other elements which are presented in other ways on other machines, but generalized markup copes equally with them all.

So the person who best grasps the structure of a text—its author—is enabled to describe that structure in a way which would be capable—should the Standard be widely adopted, as must be hoped—of being translated into processing instructions, quite regardless of the device which enables the processing and of the coding schemes associated with that device. Thus the difficulties of device dependence and of proprietary coding schemes are overcome. Of equal importance, these days, is the fact that SGML enables the easier construction of retrieval schemes. Even before ISO had adopted GCA’s draft, the Association of American Publishers had begun an investigation of generic coding as a consequence of a proposal from its College Division to devise author guidelines for the preparations of manuscripts on floppy disk. The journal Electronic Publishing Business reviewed its activities in volume 4, part 8. In that issue one contributor, Jo Gargemi of Rocappi, encapsulated the problem in these words: ‘traditionally publishers have rushed to format information on a page and (so) have...fenced the information in. Later, when they have tried to database the information, they have had to strip out all the formatting that they put in when they paginated it. The AAP approaches the publishing process in reverse; it says database first, paginate second.’ It was with a similar principle in mind that the first true implementation of generic coding in the late sixties was made by Chemical Abstracts, whose concerns were as much with ease of retrieval as with typographical presentation.

BS 6868 is not easy to read and harder still to remember. It is a complex document, full of abstractions which are, in the interests of absolute generality, couched in an arcane vocabulary all of its own. Still, it is an important topic and well worth the effort required to understand it. To that end, help is at hand. Joan Smith, one of SGM’s earliest advocates in this country, has written two guides for The British Library—one for authors and another for editors and publishers. These guidelines—BNB Research Fund Reports, numbers 26 & 27—can be obtained from The British Library, reference ISBN 0 7123 31115 and 31123.

J. D. Painter is Head of Education, Cimtech, the National Centre for Information Media and Technology, Hatfield Polytechnic.

---

**Contracted indexes**

The Society of Authors and the Writers’ Guild of Great Britain have, as mentioned in our last issue (16(1), 54) reached an agreement with several publishers—the BBC, Bloomsbury, Century Hutchinson, Faber, Headline—on minimum terms and conditions in respect of original literary works.

The agreement covers: the typescript and its delivery; warranty and indemnity; and copyright fees and index. In respect of the index, the agreement states:

If in the opinion of the Author and [Bloomsbury] an index is required, but the Author does not wish to undertake the task, [Bloomsbury] shall engage a competent indexer to do so and the costs shall be shared equally between the Author and [Bloomsbury], the Author’s share being deducted from the advance or royalties.

Several types of books are excluded:

1. Illustrated works being either:
   (a) books in which the proportion of space taken up with illustrations is 40% or more; or
   (b) specialist works on the visual arts in which the proportion of space taken up with illustrations is 25% or more.

2. Books involving three or more participants in royalties.

3. Textbooks written for the educational market (as distinct from general books of an academic or instructional nature).


In fact, a 50:50 division of the cost of indexing. Two cheers for winning half our loaf!