Table III. Z_{ij} Values between Real Proteins and Pseudo Ones Produced by Adding Random Amino Acid Sequences

length of seq added[a]	Z_{ij}			
	CCCS[b]	LWLV6[c]	O4RTPB[d]	TVMVGM[e]
0	-26	-25	-24	-31
0.5	-11	-13	-14	-14
1	-9	-10	-11	-8
1.5	-7	-8	-8	-6
2	-6	-6	-6	-4
2.5	-5	-5	-5	-4
3	-4	-5	-4	-4
3.5	-3	-4	-3	-4
4	-3	-4	-3	-4

[a] Expressed as ratios of lengths of random amino acid sequences added to those of real proteins. [b] Cytochrome *c*. The sequence length is 111. [c] ATPase, a chain. The sequence length is 248. [d] Cytochrome P450. The sequence length is 491. [e] Kinase-related transforming protein. The sequence length is 746.

Table IV. Similar Proteins of Z_{ij} Values Less Than -11 to Protein A27776

Z_{ij}	protein nos.	names
-16	SLONA1	protamines (salmines) AI and AII
-16	IRTR59	protamines CII
-16	IRTR1A	protamine (iridine) IA
-14	IRTR2	protamine (iridine) II
-13	IRTR1B	protamine (iridine) IB
-12	IRTR42	protamines pRTP242 and pTP8
-11	IRTRC3	protamines CIII and pRTP43

from the first protein with the following 150 ones in the protein database, and the results are shown in Figure 1.

Apparently the Z_{ij} values decrease linearly with the Needleman-Wunsch scores in a field of Z_{ij} values less than ca. -5 and the scores more than 60, but not in other fields.

The Z_{ij} values between a real protein and pseudo ones produced by adding some lengths of random amino acid sequences to the end of the real one were calculated as shown in Table III. The Z_{ij} values increase with increasing lengths of random amino acid sequences.

These results indicate that the distance Z_{ij} can be used in the search of closely related proteins as well as the Needleman-Wunsch scores. As an example, the results of a search for similar proteins to protein A27776, Pretamin C11-Rainbow trout, are listed in Table IV.

In the next the execution time of the calculation of Z_{ij} was examined.

Random amino acid sequences of lengths 100, 250, 500, 1000, and 3000 were produced, and the time to calculate Z_{ij} between those sequences and proteins in the database was measured. It was found that the time was almost constant at 38 s, irrespective of the lengths. The total execution time needed to calculate Z_{ij} and then rank proteins in increasing order of Z_{ij} was 45-60 s.

In conclusion, the present system gives results similar to the previous ones, but quickly, in the search of closely related proteins, and thus it is found to be useful.

REFERENCES

(1) Needleman, S. D.; Wunsch, C. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48*, 443-453.
(2) Lipman, D. J.; Pearson, W. R. Rapid and Sensitive Protein Similarity Search. *Science* **1985**, *227*, 1435-1441.
(3) Nakayama, S.; Shigezumi, S.; Yoshida, M. Method for Clustering Proteins by Use of All Possible Pairs of Amino Acids as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 72-78.

Requirements for and Challenges Associated with Submission of Machine-Readable Manuscripts†

MARIANNE C. BROGAN‡ and LORRIN R. GARSON*
Publications and Operational Support Divisions, American Chemical Society, 1155 16th Street NW, Washington, D.C. 20036

Received April 10, 1990

Computer-assisted composition, first used for primary journals in the 1960s, became an important component of publishing in the 1970s and the dominant production method of the 1980s. During the last half of this decade, the availability of word-processing tools and the affordability of computer systems have made electronic submission of manuscripts common in many publishing applications and have intensified discussions of the practicality of electronic submission for other applications. Authors and publishers share a common interest but have different perspectives on requirement and expectations. Utilization of electronically submitted material at any stage of the publication process would significantly impact that stage, with peer review, technical editing, and production being the principal targets for change. Challenges exist in the translation of tables, mathematical expressions, chemical structures, and other graphics from a variety of word processors to formats required for composition. Database applications demand the additional requirement of data-element identification.

INTRODUCTION AND BACKGROUND

Sixteen years ago at a meeting of the American Chemical Society (ACS), one of us (L.R.G.) was asked by an ACS member (who had just acquired a word processor): "When can I submit my manuscripts to the ACS on diskette?" The response was, "Well, I'm not sure, but it doesn't seem to be terribly difficult to me." Since 1974, that same question has been posed at every ACS meeting, at many journal editorial advisory board meetings, and on many other occasions.

Approximately 12 years ago, most manuscripts submitted to the ACS were prepared by typewriter. Today, a majority of manuscripts are prepared on a wide variety of word-processing systems. Virtually all scientific and technical publishers use high-end, computer-controlled composition systems. Thus, both authors and publishers produce and process manuscripts in machine-readable forms. Unfortunately, those forms are not compatible, which is a major barrier to practical, cost-

† Presented at the Symposium on Electronic Methods of Document Preparation and Information Exchange, Division of Chemical Information, 198th National Meeting of the American Chemical Society, Miami Beach, FL, Sept 14, 1989.
‡ Journals Department, P.O. Box 3330, Columbus, OH 43210.

0095-2338/90/1630-0271$02.50/0

effective electronic transfer of information between authors and publishers. Because there are opportunities to publish products in nontraditional formats such as CD-ROM, online databases, and videotext, publishers are also increasingly concerned with being able to use data from their composition processes for creating these new products. Again, there is an incompatibility issue, with data formats created for print products and formats required for the new products.[1] Furthermore, if publishers wish to combine data to create new products such as online databases that include information from multiple publishers, another level of incompatibility arises. Publishers use a wide variety of composition systems, and data are stored in very different file formats—even if the same composition system is used![2,3] How can these incompatibilities be addressed and managed?

The concept of a "generalized markup language" (GML) was proposed by Goldfarb, Mosher, and Peterson in 1970 to avoid respecifying markup instructions in documents to accommodate changes in format or output devices.[4] There are basically two types of document markup, generic and specific.

1. Generic markup is primarily concerned with document content, i.e., identification of particular data elements such as authors' names, title of paper, abstract, footnotes, and references. The concept of a generic markup language formed the basis of IBM's Document Composition Facility (DCF), which has been used extensively for over a decade. The ACS, building upon the experience and expertise of its Chemical Abstracts Service Division, has also used this approach since 1975 in publishing journals through a proprietary computer-controlled composition system.
2. Specific markup is primarily concerned with text appearance and page layout: how the printed page will look (fonts, point size, centering, indentations, etc.).

Although both types of markup are important, generic markup is the most crucial for multiple reuse of material and for sophisticated searching and display.

Over the years, the concept of GML has evolved into what has become known as SGML (Standard Generalized Markup Language). In fact, SGML has become an international standard.[5] In itself, SGML is not a tagging scheme: it does not specify particular markup codes for a document. Rather, SGML is a language that provides rules for defining document structures and generic tagging schemes. The two best known SGML applications are (1) the AAP (Association of American Publishers) guidelines and (2) the CALS (Computer-aided Acquisition and Logistics Support) initiative.

The AAP guidelines define three document types: book, article, and serial.[6,7] In 1989, the AAP turned the management of the guidelines over to OCLC-Online Computer Library Center in Dublin, OH. The guidelines are currently being considered for adoption as a NISO (National Information Standards Organization) standard.

The CALS project was initiated by a committee of the Department of Defense (DOD) in February 1987. This initiative led to the publication of a military standard (MIL-M-28001) in February 1988, and DOD is expected to require that all military technical documentation conform to the CALS specification.[8] It is possible that other Federal agencies such as the Food and Drug Administration, Environmental Protection Agency, and all granting agencies may require submission of documents using the CALS specification or some other SGML application.

Progress has been made by several publishers in being able to accept and process machine-readable submissions. A notable example of this is the journal *Tetrahedron Computer Methodology*, edited by W. Todd Wipke and published by Pergamon Press, in which submissions may be as ChemText, Microsoft Word, or flat ASCII files or in hard copy.[9] Since 1982, the *Biophysical Journal* has accepted manuscripts in machine-readable form[10] and has published information about doing so in its Instructions to Authors from 1982 onward. Likewise, the American Physical Society in 1983 published guidelines, have encouraged use of these guidelines, and have accepted manuscripts that follow these guidelines.[11,12] Preparation of manuscripts in machine-readable form has also been described by the publisher of the well-known *Chicago Manual of Style*.[13]

From discussions we have had with several scientific and technical publishers,[14] at present it is not clear that journal manuscripts received in soft-copy form are processed at a cost equal to or less than that for manuscripts received in traditional format. Nevertheless, there is broad interest on the part of many publishers to receive as well as authors in being able to send electronic submissions. The experience of the Publications Division of the ACS in receiving machine-readable manuscripts from authors has been mixed.[15] For the Journals Department, of the almost 11 500 manuscripts published in 1989, fewer than 10 were accompanied by an electronic version, and only the hard copy was used. Such submissions must be handled as special cases and are therefore more costly to process. Consequently, authors have not been encouraged to submit their manuscripts in this manner, even though specifications for such submissions are described in *The ACS Style Guide*[16] and editorials and articles addressing electronic submissions have documented ACS practice.[17-19] For the Books Department of the American Chemical Society, for typeset books almost 95% of all submissions from authors are now in electronic form, and processing these data has been cost effective and has significantly reduced production time.

Over the years it has been suggested that the ACS accept camera-ready copy from authors rather than be concerned with capturing these data in machine-readable form or incurring the expense of typesetting. With the widespread availability of high-end word processors and desktop publishing systems, which provide much improved quality of presentation over typescript or first generation word-processing systems, this notion is more frequently being put forth. In fact, the ACS does use camera-ready material from authors for information included as supplementary material. However, the use of camera-ready material has two serious disadvantages. First, the quality of camera-ready material is generally inferior to that of typeset material: the physical presentation is not at professional standards and the format does not lend itself to the added value of professional technical editing. Second, camera-ready copy would not allow secondary products to be created such as database building as is currently being done in the creation of Chemical Journals Online on STN International, which is a direct byproduct of the typesetting process. Indeed the current generation of word-processing software is seriously deficient for database needs because it produces output which is format or presentation oriented and not content based; that is, identification of critical data (tagging) is not made. In the long term, the ACS is committed to providing information by electronic means, which requires primary data to be in electronic form and appropriately characterized.

The balance of this paper will address details of the ACS experience and describe the challenges facing both authors and publishers in being able to handle machine-readable manuscripts.

QUALITY CONTROL IN JOURNAL PRODUCTION

We begin by examining four questions that are seemingly obvious when considering the processing of machine-readable

submissions from authors for publication in journals.

1. Can keyboarding be eliminated?
2. Can proofreading be eliminated?
3. Will editorial costs be reduced or increased?
4. How will quality be affected?

Can Keyboarding Be Eliminated? In an ideal, simple world, when manuscripts are presented in a consistent fashion with all materials properly coded and/or easily translated, much new keyboarding *could* be eliminated. However, our journals are far from simple, consistency is practiced by a select few, and coding is often directed toward making the manuscript's appearance more attractive instead of toward the requirements for the printed publication or that publication's alternative formats, especially database applications. Coding for journals and books includes matters of boldface and italic, Greek letters and other special character sets, and for the journals also data-element flagging to ensure efficient database production. It is in the area of data-element flagging where books and journals diverge the most. Translation programs exist, as do programs that scan and tag papers, but our experience to date has been that none of these programs work well. If 80% of the material is properly translated and tagged, then one only has to address, via keyboarding, the additional 20%. Unfortunately, that 20%, in a correction mode, is often more expensive to accommodate than rekeying and proofing all material from the hard-copy manuscript.

Can Proofing Be Eliminated? Not with any continued assurance of quality. Whether translation is by program or whether changes are made interactively online, the accuracy of those modifications has to be verified. Our experience with journals indicates verification is most efficiently accomplished in handling the hard-copy form. In book production, both hard- and soft-copy proofing occurs, the preponderance of proofing being done with soft-copy. Most users of word-processing programs will agree that it is much easier to delete or duplicate information from an electronic file than from hard copy, so proofing is necessary to rectify mistakes of translation, deletion, or repetition.

Will Editorial Costs Be Reduced or Increased? Whether it is more efficient to edit online or in hard copy depends to a great extent on the type of material one has to work on and the types of corrections that have to be made. For example, editing online can be more efficient where it is feasible to use global search-and-replace techniques. However, editing online can be less efficient: instead of writing notes such as"space always" or "all minus signs", so that the trivial changes and codings can be done by less highly compensated keyboarding staff, those trivial changes and codings may have to be introduced by editorial staff, thus escalating editorial costs. Conversations with editorial staff of other scientific publishers indicate that 30–50% more time is necessary to ensure equivalent quality in processing journal manuscripts received in machine-readable form.

How Will Quality Be Affected? As publishers, we feel it is vitally important that high quality be preserved. Both journal and book publication, at least in our environment, is viewed as a series of value-added procedures—from peer review and subsequent revision to scientific copy-editing and typesetting; from proofing to correction; from composition to printing and distribution. Each step enhances the quality of the information being disseminated. The technical editing process itself plays a major role in contributing to overall product quality. Most authors and readers are not aware of the extent to which editing occurs. For purposes of reference, a typical manuscript page is double-spaced and consists of approximately 26 lines of text. Light editing may involve fewer than 10 changes on such a page. Most of these edits will be format requests. An average level of editing will be about 20 changes per page, both format and style. Not infrequently heavy editing is required, especially with mathematical and highly technical material, in which case 40 or more changes per page would not be unusual. Quality is not free: it has a price.

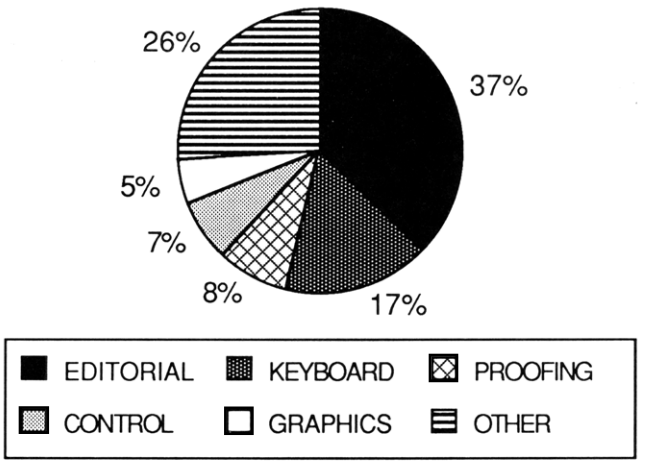

Figure 1. ACS composition and editorial cost breakdowns for journals—1989.

ECONOMIC ISSUES OF JOURNAL PRODUCTION

For some users of libraries a perception may exist that information is free. Librarians who buy books and journals know this perception is false. The laws of supply and demand dictate that librarians will add to their collections, and thus publishers will continue to publish. But costs continue to rise. Fewer copies of higher priced books are purchased, fewer copies of new journal titles are added to collections, and more current titles are dropped. What can publishers do to reverse this pattern? Can publishers control prices?

During the inflationary period of the 1970s, prices for journals were controlled somewhat by changing technology: from letterpress to web offset printing; from hot-metal composition to photocomposition; from composition to camera-ready copy. But those remedies as avenues of savings have been exhausted. What apparently remains is the ability to tap what most authors can now provide: manuscripts in electronic form. A perception exists in this area as well: capturing authors' keystrokes will reduce costs and guarantee quality.

Having explored the quality issue and that of editorial costs in general, we now examine more of the economic issues of journal production, with emphasis on those related to processing machine-readable submissions from authors. Costs associated with publication include composition and editorial, with composition being broken into multiple components. The two most relevant to electronic submissions are keyboarding and proofing, which, combined with editorial, are the ones most subject to change in an electronic environment. Figure 1 is a pie chart showing the percentage breakdown of journal composition and editorial costs in 1989.

Figure 2 gives a breakdown for the costs associated with keyboarding, proofing, and editorial operations for producing ACS journals in 1989. These are the cost centers most likely to be affected by processing machine-readable submissions. Editorial operations constitute the largest expense, 59%, followed by keyboarding at 27% and proofing at 14%. If in processing machine-readable manuscripts the perception of lowering costs for keyboarding and proofing is valid, then publishers must carefully guard against increased editorial expenses. Is it possible to also lower editorial costs? Probably not. Experience in the publishing industry to date is that, if the same level of quality is maintained, editorial costs increase when processing manuscript submissions in machine-readable

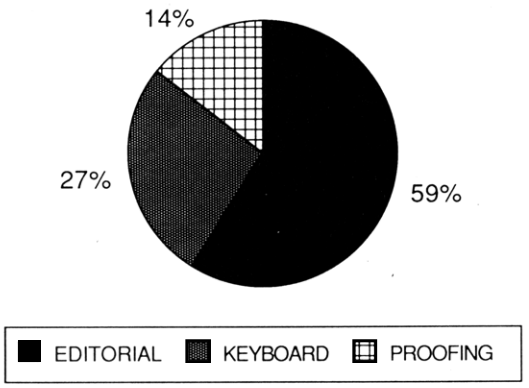

Figure 2. ACS keyboarding, proofing, and editorial costs—1989.

form. For our journal production, the issue is basic: how can we make our electronic capabilities best serve us and enhance the value-added process in a competitive, cost-effective manner?

BOOK PRODUCTION

As already mentioned, our experience in handling machine-readable manuscripts from authors for book production is quite different from that in journal production. At the present time, 95% of the manuscripts received are on diskettes. Of these 95%, 75% can be loaded directly into the computer system used by the Books Department; 25% of the diskettes require conversion by an outside vendor. Most of the remaining 5% of manuscripts are optically scanned to capture the data. Few manuscripts are keyboarded.

The Books Department uses an NCR-Tower, with the UNIX operating system and standard UNIX software tools for processing text files and *troff* and associated macro packages for composition. Editorial staff copy-edit and encode the files online, using the Emacs editor. Widespread use is made of global search-and-replace functions in the editing process. Mathematical equations and tables are composed by using UNIX's *eqn* and *tbl* as preprocessors to *troff*. Chemical structures are prepared on the UNIX system using AT&T's *chem* program and/or ChemDraw on a Macintosh computer. Page makeup is accomplished manually.

Since 1986, when all ACS books were traditionally processed and typeset, turnaround time, the time from which a manuscript is formatted and sent for composition until the bound book is produced, has decreased from 28 weeks to 14 weeks. The editorial process—peer-review through copy-editing and galley correction—has resulted in an estimated increase in effort and time of 20–25%. However, overall editorial/production staff productivity has increased, and although we are not able to give an accurate, quantitative figure for the net effect on staff effort due to soft-copy input, we estimate 30% increase in editorial/production productivity. Management in the Books Department feels that overall quality of the books has improved through adoption of this editorial/composition system and through the receipt of machine-readable manuscripts from authors. Overall cost savings are estimated to be about 10%, but because of the uncertainty regarding the changes in editorial staff effort, this figure is uncertain.

DIFFERENCES BETWEEN JOURNAL AND BOOK PRODUCTION

Of necessity, the operations between the Books Department and the Journals Department are quite different. Staff from the Books Department have preliminary dialogue with authors prior to manuscript submission, leading to greater control in manuscript submissions. The Journals Department does not have such a dialogue and receives manuscripts as accepted by

1. Spacing at end of sentences

   ... for analysis. The precipitate ...

2. Spacing in reference citations

   Smith, L. R. vs Smith, L.R.

3. Spacing in standard abbreviations

   L. R. Smith, Ph.D. vs L. R. Smith, Ph. D.

4. Spacing in chemical names

   bicyclo[2.2.2]octane vs bicyclo[2. 2. 2]octane

5. Spacing in tabular or indented situations

   spacing achieved by use of a tab command

   vs

   spacing achieved by 8 strokes of a space bar

Figure 3. Insertion/deletion of spaces.

editors of 22 research journals. The Books Department processes a relatively small number of large papers, whereas the Journals Department handles a large number of small papers. These factors result in the Books Department enjoying a relatively homogeneous manuscript environment as compared to the Journals Department, which is extremely heterogeneous. In producing the journals, consideration must be given to data-element definition for database building. This factor imposes certain constraints in production. At present, the Books Department is not concerned with database building, but this focus is likely to change in the future.

There are significant quantitative differences between book and journal production as well. In 1989, the Books Department processed data for 14 typeset titles for a total of 200 chapters in 4116 pages. In comparison, the Journals Department published 11 473 papers for 20 journals in 63 000 pages, from 54 different countries. Obviously, the Journals Department must handle machine-readable submissions differently than the Books Department because of the scale of manuscript submissions. Also note that an average manuscript submission results in 20.6 pages for a book in contrast to a journal paper, which is 5.5 pages. This factor also impacts processing methods.

NEED FOR SOFTWARE TOOLS

From our experience, we have identified a number of areas in which we perceive a need for software tools to increase productivity, especially in the area of editorial operations. Among them are tools for

1. insertion/deletion of spaces
2. discrimination for specific meanings among similar characters
3. identification of special characters
4. identification of data elements

Consider these items in some detail.

Insertion/Deletion of Spaces. As shown in the first example in Figure 3, spacing at the ends of sentences presents a challenge. In normally typed manuscripts, the end of a sentence is followed by two spaces. However, in composed material, spacing at the end of sentences is adjusted as part of the hyphenation-justification process. Consequently for the purposes of composition, electronic manuscripts must be processed to replace double spaces at ends of sentences by a single space. Fortunately many modern composition systems, for example Xyvision, reduce double spaces in ASCII input files to a single space and introduce proper spacing as part of the composition process.

-10 °C
vs
−10 °C

Figure 4. Example of need for discrimination for specific meanings among similar characters.

As shown in the second example in Figure 3, in names, initials are separated by a single space. (Contrast that usage with L.R.G.—no spaces being used.) Authors may supply copy with one, two, or no spaces.

In the third example in Figure 3, the abbreviation for doctor of philosophy has no space preceding the "D". Here, too, copy from authors may have extraneous space and corrections are needed.

In the case of the chemical expression bicyclo[2.2.2]octane, no spaces should appear; manuscript copy and associated electronic files often contain spaces, incorrectly.

Finally, the last example given in Figure 3 can be rather subtle. Some word-processing software programs use a tab character, which is a single character, to position text; typists may choose to use spaces for the same purpose. In the example shown, the appearance of the two lines is identical whether a single tab character is in the source text or five spaces. This difficulty is particularly vexing with tabular material. For composition purposes, spacing must be handled consistently, and one cannot mix and match spaces and tabs without creating problems.

Obviously any tool that would aid editorial staff to distinguish among these cases would have to be sophisticated. Such a filter would likely need to be interactive to allow staff to override the software for circumstances that cannot be rule or context based.

Discrimination for Specific Meaning among Similar Characters. Sometimes groups of characters convey special meaning or require special typographic handling. In either case discrimination for specific meanings among similar characters is required. Consider Figure 4, where the first item shows a temperature value as might be submitted by an author. There is a hyphen, followed by two digits (a one and a zero), then a superscripted letter "oh", and finally a capital letter "C". For the purposes of typography, the hyphen is a minus sign and the superscripted "oh" is really a degree symbol. In consideration of database applications it is also important that the representation for such values be consistent. Here, too, software tools are needed to help editorial staff. At this time we do not have adequate experience to either identify clusters of characters with special meaning or the variety of ways in which authors might submit these data.

Identification of Special Characters. The identification of special characters is probably one of the most difficult problems facing publishers in dealing with soft-copy submissions, special characters being those characters not accounted for in the ASCII character set. Each word-processing package codes special characters differently. Moreover, the problem is exacerbated by the use of laser printers because the character printed is dependent upon the font cartridge used in the printer and/or softfonts downloaded from the printer. Thus, any specific character representation in the word-processing file itself may cause the printing of several different characters. Here are some simple examples of the types of special characters that need to be handled.

- Greek characters, α, β, and so forth, appear frequently in chemistry and other scientific and technical literature. Other non-English characters are often used as well. These are illustrated by Mössbauer (German), Ångström (Swedish), Brønsted (Danish), etc. The use of non-English characters is common in ACS journals.

Title	Identified as first block of text
Authors	Separated from title by blank line (individual authors identified by being separated by commas)
Affiliation	Separated from authors by blank line
Abstract	Separated from affiliation by blank line
Paragraphs	Separated from previous material by blank line
Sentences	Identified by period, followed by space, followed by capital and lowercase letters (but consider the problems with specialized abbreviations beginning with lowercase letters)
Section headings	

Figure 5. Selected data elements requiring identification.

- Unlike typewriter composition and many word processors, typography has several "dashs" of different length. For example, there is the hyphen (-), the em dash (—), the en dash (–), a minus (−), and a single bond. Manuscripts typically contain only the hyphen, which must be editorially transformed into the more appropriate "dash" symbol.
- Chemistry requires the use of many mathematical symbols, only a few of which are shown here: $\leq$ (less than or equal to), $\approx$ (approximately equal to), $\pm$ (plus or minus), $\propto$ (proportional to), $\equiv$ (identically equal to). These symbols per se do not account for the necessity to typeset complex mathematical expressions that contain these mathematical characters.
- There is also a wide variety of other characters, such as • (bullet), □ (square), and © (copyright).

It is a goal of the ACS to be able to disseminate information in a variety of formats and media and to be able to process data in an efficient, cost-effective manner. This goal requires data to be organized in a coherent, preferably consistent, manner. Organizations that take this approach to publishing are called "database publishers". The production of ACS journals requires that streams of text be specifically identified and delineated in what we call data elements. Text in machine-readable form produced by common word processors does not make any accommodation for this requirement. Text is nothing more than a series of characters organized in a manner to give a good appearance when printed. Figure 5 roughly outlines an approach to the identification of data elements as it might be handled by some software package. A detailed analysis and algorithmic approach to this problem are beyond the scope of this paper.[20]

Another approach, of course, is for authors to prepare documents for submission to publishers by using SGML. Several companies are developing software related to the AAP and CALS applications, among them ArborText, Inc. (Ann Arbor, MI), Avalanche Development Co. (Boulder, CO), and SoftQuad, Inc. (Toronto, Canada). For example, SoftQuad's *Author/Editor* program provides a Macintosh-type word processor for entering text within each data element defined by the document-type declaration. The program uses a parser to check for SGML conformance at the time text is being entered by the author.[21] Tools like *Author/Editor* obviate the need for external data-element identification.

SURVEY OF AUTHORS—MACHINE-READABLE MANUSCRIPTS

Obviously if publishers are going to process machine-readable manuscripts submitted by authors, then authors must have the capability and willingness to produce their manuscripts in such a format. In the spring of 1987, we conducted a survey of authors who had published in ACS journals. Previously we had surveyed authors on this topic in 1984 and 1978. In the 1987 survey, questionnaires were sent to 1794

authors, with 45% responses. Of the authors responding, 93% use computers or word processors to prepare their manuscripts. Of this subset of authors, 76% use microcomputers, 13% standalone word processors, 7% minicomputers, and 4% mainframes. The trend over the years has been toward the use of microcomputers, and we believe this trend has continued during the past 2 years. Many authors use several different systems to produce portions of their manuscripts: 25% use graphic packages, 21% use chemical drawing software, and 16% use math packages. Of the responding authors, 90% expressed a willingness to provide their manuscripts in machine-readable form. Thus there is apparently willingness and ability for authors to provide publishers with electronic manuscripts. The Royal Society of Chemistry conducted a similar survey of their authors in 1989 with similar results.[22]

Although authors use many different word-processing packages, only seven packages account for 74% of all those used. WordStar, which was the most commonly used word-processing software in 1987, is clearly not widely employed at the present time. The ACS Books Department now most frequently receives submissions prepared by Microsoft Word and WordPerfect.

Authors employed over 50 different software packages to prepare mathematical equations. The three most commonly used packages are Mac Eqn, by 16% of the respondents using such software; TeX at 10%; and T-3 at 8%.

Authors reported the use of 41 different packages for drawing chemical structures, a surprisingly high number. Undoubtedly, during the past 2 years a number of these packages have fallen into disuse. Our experience is that ChemDraw and ChemIntosh are the most widely used packages at this time, with ChemDraw being the more popular of the two.

Authors reported the use of 81 graphic drawing software tools, an astonishing number. Cricket Graph, MacDraw, and Lotus 1-2-3 were the most commonly used.[23] In the past 2 years there have been many more packages introduced into the marketplace for creating graphic images, especially in connection with desktop publishing; consequently these results are particularly valid today.

Whereas the perceived benefit to publishers of being able to receive electronic manuscripts from authors is reduced costs, what are the benefits to authors, or at least the perceived benefits? When authors were asked this question on the survey, the results showed that 71% expected to do less proofreading and 51% expected reduced costs. However, it is not clear whether the respondents thought costs would be reduced for them or for the publishers. There also seems to be a sense of puzzlement among authors: since the keystrokes have been captured, why cannot publishers use these data? Why go to the added trouble and expense to rekey manuscripts? Only 6% of the respondents expected faster publication by submitting electronic copy.

It remains to be seen what benefits authors would actually receive from electronic submission of journal manuscripts, but our experience with the Books Department indicates general satisfaction among authors with electronic submissions.

CHALLENGES FACING PUBLISHERS

Two central issues exist: (1) containing or reducing production costs and (2) preserving or improving quality. The two issues are not independent. Even though quality can be difficult to quantify, it has a definite cost. There are several secondary issues. There is definitely a need to develop "editorial" software tools for the identification of special characters, data elements, space corrections, and other functions described earlier. There is also a need to develop software for conversion of word-processing files to file formats for composition systems unless software tools for creating SGML documents become widely used by authors—an event we consider unlikely in the next several years. At this time our R&D staff are working on developing software to import files from Microsoft Word and WordPerfect into the Xyvision composition system.

The processing of machine-readable manuscripts may potentially increase manuscript control problems. How will production staff know the manuscript on the diskette is the latest revision accepted by reviewers and the editors? This problem exists with submission of paper copy as well, but it may be a greater problem with electronic submissions.

If substantial numbers of electronic manuscripts are processed, it is inevitable that the roles of keyboarders and editors will change. Of course, this involves requirements for different skills, and training becomes an issue. Production itself would also undergo changes, illustrated by the following example: Assuming filters and other software aids are developed to help with processing of machine-readable manuscripts, at what point in production, and by whom, would these tools be employed? In some cases they would be used prior to copy-editing, so editors could offset any damage introduced. In other cases they would be used after copy-editing, so the editor need concentrate only on text and not on typesetting commands. But in the latter case, the material subject to those filters will have to be proofread to ascertain the validity of the application.

This paper has not addressed the issue of processing nontext data, i.e., tabular information, mathematical expressions, line drawings, and photographs or halftones. Each of these items has its own challenges.

In conclusion, while we have made good progress in some areas for receiving machine-readable manuscripts, notably in book production, there is much work yet to be done.

REFERENCES AND NOTES

(1) Martinsen, D. P.; Love, R. A.; Garson, L. R. Multiple Use of Primary Full-Text Information—A Publisher's Perspective. *Online Rev.* **1989**, *13*, 121–133.

(2) Martinsen, D. P.; Love, R. A.; Garson, L. R. Handling Manuscripts in a Multi Word-Processing Environment. *Abstracts of Papers*, 198th National Meeting of the American Chemical Society, Miami Beach, FL, Sept 14, 1989; American Chemical Society, Washington, DC, 1989; CINF 42 (paper presented at the Symposium on Electronic Methods of Document Preparation and Information Exchange).

(3) Look, H. E., Ed. *Electronic Publishing—A Snapshot of the Early 1980's*; Learned Information: Princeton, NJ, 1983.

(4) Goldfarb, C. F.; Mosher, E. J.; Peterson, T. I. An Online System for Integrated Text Processing. *Proc. Am. Soc. Inf. Sci.* **1970**, 147–150.

(5) *Information Processing—Text and Office Systems—Standard Generalized Markup Language (SGML)*, ISO 8879-1986; International Standards Organization: Geneva, Switzerland, 1986.

(6) Bryan, M. *SGML. An Author's Guide to the Standard Generalized Markup Language*; Addison-Wesley: New York, 1988.

(7) *Association of American Publishers Electronic Manuscript Series*, Version 2.0; (1) Author's Guide, (2) Reference Manual on Electronic Manuscript Preparation and Markup, (3) Markup of Tabular Material, (4) Markup of Mathematical Formulas; Online Computer Library Center: Dublin, OH, 1989.

(8) Keiser, B., Ed. *EPSIG (Electronic Publishing Special Interest Group) Newsletter*, March 1989.

(9) Wipke, W. Todd. The First Scientific Journal Published on Disks—Tetrahedron Computer Methodology. *Abstracts of Papers*, 198th National Meeting of the American Chemical Society, Miami Beach, FL, Sept 14, 1989; American Chemical Society, Washington, DC, 1989; CINF 40 (paper presented at the Symposium on Electronic Methods of Document Preparation and Information Exchange).

(10) Parsegian, V. A. Editor's Forward. *Biophys. J.* **1982**, *37* (1).

(11) Physical Review Style and Notation Guide and Physical Review Input Guide for Author-Prepared Compuscripts. *Bull. Am. Phys. Soc.* **1983**, July, A1–18, B1–B57.

(12) "A limited number of author-prepared articles were accepted for APS publications: ... as keystrokes prepared on the author's UNIX system and submitted on magnetic tape; or as keystrokes prepared for input to the UNIX system, submitted on an MS-DOS-formatted disk." From AIP in 1988: An Annual Report. *Phys. Today* **1989**, June, 47–58. In 1988 over 50 000 pages were published by the American Physical Society, on a UNIX-based system; 46 manuscripts, ca. 0.5%, using *troff* were submitted. In 1989, the number of *troff* papers had declined to 41, and 17 TeX papers were used. In 1990, the number of *troff* papers

had declined further, with TeX rising to 140 (late March). The American Physical Society accepts only their version of TeX (RevTeX): their costs of converting other versions make rekeying the paper their production method of choice (personal communication, March 19, 1990).
(13) *Chicago Guide to Preparing Electronic Manuscripts for Authors and Publishers*; The University of Chicago Press: Chicago, IL, 1987.
(14) On Sept 13, 1989, an informal meeting was held in Miami Beach, FL, attended by representative from the ACS, The Royal Society of Chemistry, John Wiley & Sons, Elsevier Scientific Publishers, and Science Typographers, Inc. The purpose of the meeting was to share experiences in direct author submissions in electronic form. All present agreed it is difficult to support electronic submissions for journals, though it has proven cost effective in some cases for book submissions. Also, all agreed that data content specification as well as the text needed to be captured by the publisher and that authors needed to be enticed to do this. Most agreed that storing (and thus ideally capturing) documents with the content-specific information marked was essential for subsequent use of the data, whether in electronic or in reformatted print product.
(15) From 1976 to 1988, 18 papers have been published by using the electronic version of the manuscript, 11 in *Anal. Chem.*, 3 in *Environ. Sci. Technol.*, 2 in *J. Org. Chem.*, 1 in *J. Am. Chem. Soc.*, and 1 in *Inorg. Chem.* Additional papers were available on diskette, but based upon the experience with those published and the complexity of the material, subsequent work was done from the hard copy.
(16) Brogan, M. C. Manuscript Submissions in Machine-Readable Form. In *The ACS Style Guide*; Dodd, J. S., Ed.; American Chemical Society: Washington, DC, 1986; Chapter 5, pp 149-157.
(17) Brogan, M. Analytical Chemistry—A New Approach. *Anal. Chem.* **1977**, *49*, 557A.
(18) Brogan, M. Electronic Manuscripts—One Step Closer. *Anal. Chem.* **1984**, *56*, 784A.
(19) Warner, M. Electronic Publishing in Analytical Chemistry. *Anal. Chem.* **1987**, *59*, 1021A.
(20) Many publishers do not currently include generic coding in their composition process for data-element identification. In processing non-ACS composition data for file building for Chemical Journals Online (CJO), composition data files must be handled as described in this article. Programs for doing these operations were developed by D. P. Martinsen.
(21) *SoftQuad Author/Editor User's Manual*, Version 1.1; SoftQuad, Inc., 1989.
(22) Blackmore, J. (Royal Society of Chemistry, Information Services, Thomas Graham House, Science Park, Milton Road, Cambridge, CB4 4WF, U.K.). Personal communication, Aug 11, 1989.
(23) It is not apparent how the spreadsheet program Lotus 1-2-3 is used to draw chemical structures, but authors did report relatively high use of this package for that purpose.

Computer-Assisted Structure Generation from a Gross Formula. 3. Alleviation of the Combinatorial Problem†

IVAN P. BANGOV

Laboratory of Mathematical Chemistry and Chemical Informatics, Institute of Mathematics, Bulgarian Academy of Sciences, Building 8, Sofia 1113, Bulgaria

Received May 22, 1989

The problem of generation of an exorbitant number of combinatorial operations in the process of isomer enumeration is discussed. The origins of the duplicated structures (isomorphic in the graph-theoretical sense) are examined. A novel approach leading to a substantial reduction of the redundant combinatorial operations is described. Two interrelated schemes: Hierarchical Saturation with Equivalent Saturating Valences (HSESV) and Hierarchical Selection of Saturation Sites (HSSS) are developed and their efficiency is illustrated. Various ways of employment of the available structural and spectral information for alleviation of the combinatorial problem are discussed.

INTRODUCTION

Structure-elucidation systems are designed to produce one or several plausible answeres from a limited amount of structural information. This involves the generation of different optional structures, a process which requires structure-generation programs (generators). The most severe problem in the development of such programs is the generation of an enormous number of combinatorial operations for all but the smallest molecules. Most of them result in either chemically inconsistent or redundant structures.

Many approaches to avoid the redundancy have been discussed in literature. Thus, a linear-notation algorithm based on canons of precedence that order the branches of each tree-like acyclic structure was developed by Lederberg et al.[2] However, a comparison of each newly generated structure with the canonical representations of the structures previously generated is impractical. More advanced is the scheme based on the sequential execution of the "partition" and "labeling" steps exploited in the cyclic structure generator of DENDRAL.[3] Here duplication is avoided in the step labeling by taking into account the topological symmetry.[4]

An elegant mechanism using the concept of connectivity stack[5-7] was devised in CHEMICS. The formation of the canonical (greatest connectivity stack) structure requires an exhaustive generation of all segment permutations. Thus since the segments (primary, secondary, and tertiary components[5]) are of small size, this is a computer-intensive procedure in the cases of larger molecules. Some new developments aiming at reduction of the isomorphism checks have been recently reported,[7] but this still remains a major problem for the CHEMICS generator. Molecular fragments (pieces of the molecular structure with known connectivity between their atoms) are also employed within this scheme, but they are initially degraded into secondary and tertiary components and then the latter are used in the construction process. The existence or the absence of a given fragment is perceived by a substructure search algorithm[7] (an additional time-consuming procedure) applied to each generated structure.

A similar approach to the generation of nonisomorphic graphs has been discussed by Faradjiev.[8] The method implies selection of a graph from a set of isomorphic graphs, which is considered canonical, and the program further generates only canonical representations, using a predicate for canonicity. The maximal adjacency matrix was chosen as the canonicity predicate. Obviously, this predicate corresponds to the connectivity stack in CHEMICS. In the same way, the isomorphism checks are carried out by generating all the permutations, and the fragments can be handled only by a substructure perception algorithm.

† For Part 2 of this series see ref 1.

0095-2338/90/1630-0277$02.50/0