by the Russian group do not fulfill those criteria. What are those criteria?

(4) The theoretical chemists are expecting "island of stability" for super-heavy elemnts around element 114. Explain why.

(5) A few experiments have been carried out to locate super-heavy elements in nature. Please describe them.

(6) There was a claim (which turned out later to be wrong) for the synthesis of element 114. Describe the experimental part of the above work.

Figure 2 is an example of a news story that was published in the Science section of *Time* magazine on January 28, 1980. The student has to answer the following questions:

(1) Who was the first one to isolate interferon?

(2) What do you know about the anti-viral and anti-carcinogenicity activity of interferon?

(3) How is interferon isolated today from natural sources? Are those routes protected by patent(s), and if so, who holds the patent(s)?

(4) What are clones?

(5) What the research interests of Dr. Weissmann?

(6) What do you know about the Biogen Co.?

(7) What is the scientific work on which the *Time* article is based?

SYNTHETIC ORGANIC CHEMISTRY LECTURE

The Synthetic Organic Chemistry Literature lecture covers the following topics: locating suppliers of fine chemicals, locating procedures for synthesis of chemical compounds (*Beilstein*, CA, *Organic Syntheses*), designing a synthesis (the use of organic class preparations sources such as *Methodicum Chemicum*, *Houben-Weyl*, *Compendium of Organic Synthetic Methods*, *Formation of C-C Bonds*, *Theilheimers Synthetic Methods of Organic Chemistry*, *Journal of Synthetic Methods*, *Current Chemical Reactions*), information concerning reagents, solvents, and/or techniques (*Reagents for Organic Synthesis*, *Techniques of Chemistry*), and getting an overview (*Organic Reactions*, *The Chemistry of the Functional Group*, the review literature).

USAGE OF NUMERICAL DATABASES

The last part of the program is the usage of a numerical database in the advanced laboratory. The students received a short introductory lecture about data banks in general and numeric data banks in particular. The structure and organization of the NIH/EPA MSSS (Mass Spectral Search System) is discussed. The student learns about some of the database options and their usage [e.g., searching for peaks (PEAK), searching for peaks and molecular weight (PMW), entering a spectrum, searching with a complete spectra.] After running the mass spectra of the unknown, the student elucidates the structure of the compound and then turns to the terminal in order to run a computerized structure elucidation, comparing the results obtained from the two routes.

CONCLUSIONS

Our feelings are that a student that participates in all four parts of the program receives a good basis for the modern approach in chemical information. The ones that take only the compulsory 4-h lecture receive the basic information about the chemical literature. Those students learn some of the aspects of chemical information (the use of numerical databases and synthetic organic literature) if they elect the Advanced Organic Chemistry laboratory.

Quantification, Retrieval, and Automatic Identification of Numeric Data in Organic Chemistry Journals

JOSEPH J. POLLOCK

Chemical Abstracts Service, Columbus, Ohio 43210

Received December 20, 1983

The full text of American Chemical Society (ACS) primary journals may now be searched on-line. An important benefit of this capability is that it makes available a substantial amount of highly current numeric data, which can be used to help identify unknown compounds or retrieve data on known ones. Such searches are readily performed with high recall and precision. The amount of certain types of numeric data is estimated, and an algorithm for identifying them in the journal text is discussed in detail.

INTRODUCTION

Computer searching of the content of chemistry journals has traditionally proceeded via index entries and focused on concepts rather than data. However, computer technology has now advanced to the point where it is possible to search primary journals directly and for data. A particular aspect of this topic is explored here; the quantity, use, and automatic identification of numeric data in the experimental sections of three American Chemical Society (ACS) journals concerned with synthetic organic chemistry. All ACS journals can now be searched on-line in a publicly available file.

The first half of this paper estimates the amount of characterization data (defined below) in three ACS journals and demonstrates how one can use it to identify an unknown compound or retrieve data on a known one. The second half presents an algorithm for identifying a "package" of numeric data in primary journal text comprising a CAS Registry Number (CAS REG) for a substance, its name and/or symbol, and its characterization data together with their measurement conditions.

(1) QUANTIFICATION OF CHARACTERIZATION DATA

When ACS journals were examined for numeric data, it rapidly became clear that the kind most likely to repay further study was "characterization data", the type that chemists

Table I. Documents with Content Indicative of Characterization Data

Property	Search Statement	Documents
MP	MP	6,789
	MP with (DEGREE adj C)	6,385
UV	UV	5,874
	UV same (EXPERIMENTAL ADJ SECTION)	3,352
	UV with NM	2,790
	LAMBDA adj MAX	2,415
BP	BP	2,895
Optical Rotation	ALPHA adj #D	772
Refractive Index	N#D	333
Circular Dichroism	(CD or (CIRCULAR adj DICHROISM) same THETA	265
IR	IR	8,164
	(IR or INFRARED) same (EXPERIMENTAL adj SECTION)	6,776
	IR with CM	5,089
NMR	NMR	12,265
	NRM same (EXPERIMENTAL adj SECTION)	7,851
	(NMR same [EXPERIMENTAL adj SECTION]) and (JOCEAH.CD. or ORGND7.CD.)	3,408
	(13C or (C or CARBON) adj 13) same (EXPERIMENTAL adj SECTION)	2,531
	113CD adj NMR	23
	(23NA or (NA or SODIUM) adj 23) with NMR	40
MS	(MASS adj SPECTRUM) OR MS OR (M ADJ E) OR (M ADJ Z) same (EXPERIMENTAL adj SECTION)	4,279
	(MASS adj SPECTRUM) OR MS) same (EXPERIMENTAL adj SECTION)	4,023
Yield	YIELD$	26,222
	YIELD$ same (EXPERIMENTAL adj SECTION)	7,972

Table II. Documents with Content Indicative of Selected Physical Properties

Property	Search Statement	Documents
Moessbauer Spectrum	(MOSSBAUER or MOESSBAUER) same (EXPERIMENTAL adj SECTION)	78
Lethal Dose	LD	407
Dissociation Constants	PK	856
	PKA	1,634
	PKB	55
Free Energy	DELTA adj G	1,298
Free Energy	DELTA adj F	417
Enthalpy	DELTA adj H	1,690
Entropy	DELTA adj S	1,214
Chromatographic Retention	RF	7,262
	TR	771

Table III. Substances with Characterization Data

		Data/Document	
Property	Substances	Range	Average
Mass Spectra	9,000	0-13	2.4
NMR Spectra	60,000	1-20	8.3
IR Spectra	25,000	0-21	13.6
UV Spectra	13,500	0-11	8.0
Melting Points	36,000	2-21	5.3
Boiling Points	7,000	0-7	2.5

report for a synthesized substance both to link identifying data to the substance name and to support the structure assigned. This is by far the easiest type of data to generate, to identify, to use, and to measure. Typically, it consists of single physical measurements or spectra, and its main utility to others is for substance identification and structure elucidation. There are at least hundreds, and probably thousands, of other types of numeric data, all of which occur much less often than characterization data and many of which are far more complex. The study was also confined to the three ACS journals that carry substantial amounts of characterization data: the *Journal of Organic Chemistry* (JOC), the *Journal of Medicinal Chemistry* (JMC), and *Organometallics* (OM).

The ACS has experimented with on-line searching of primary journals for several years. Currently, a file with 4 years of most ACS journals may be searched on-line. This file was searched to determine the amount of each kind of characterization data and to assess its retrievability and searchability. That is, can one retrieve the data for a known compound, and conversely, given various types and amounts of data, can one identify the file compound to which it applies? It is also important to determine the number of substances involved.

Characterization data typically consist of either single physical measurements [e.g., mp (melting point), bp (boiling point), optical rotation, refractive index, or chromatographic retention times] or spectra [e.g., IR (infrared), UV (ultraviolet), NMR (nuclear magnetic resonance), MS (mass), or CD (circular dichroism) spectra]. To assess the utility of the ACS journals in this regard, it is important to know the number of substances characterized and the frequency of each data type.

Most CAS REGs cited in JOC and OM represent new compounds with characterization data, so estimating them gives a reasonable measure of the number of substances characterized. The CAS REGs in both full papers and short communications (which tend to have fewer characterized substances and less numeric data) in an issue of each of these two journals were therefore counted, and the number of substances with numeric data in the experimental sections of a volume of JMC (which did not then include CAS REGs) was similarly noted. Estimating the ratio of papers to short communications in different ways and extrapolating suggest that the number of characterized substances per year is 23 500–25 000 in JOC, 4700–5450 in OM, and 2400 in JMC, approximately 32 000 substances per year for all three journals. When this study was conducted, the file contained 3 years of JOC, 1 of OM, and 7 of JMC, which corresponds to 90 000 characterized substances. By the end of 1983, there will be 1 more year of each journal on-line, which suggests 120 000–130 000 substances, and the total number of documents will have risen from 27 000 to 40 000. Thus, to correspond to the end of 1983, the estimates given below for specific data types should be increased by approximately one-third. The number of substances searchable on the ACS file will then be roughly comparable to most existing numeric databases. However, the file has the significant advantage that several different properties are associated with most substances, which greatly facilitates searching (see Correlative Searching below).

Estimating the number of occurrences of specific types of characterization data is a two-stage process as an on-line search reveals only how many documents contain specific types of data (Tables I and II). Individual documents must then be examined to determine the number of substances involved and give the estimates shown in Table III. The search statements in Tables I and II, which are only meant to be illustrative starting points for on-line searchers, are in a command language that provides the following proximity operators: AND, in the same document; SAME, in the same paragraph; WITH, in the same sentence; ADJ, adjacent or separated only by stopwords and in the specified order. Here, the SAME operator is used with EXPERIMENTAL SECTION to ensure that only paragraphs in the experimental section of the paper are retrieved as these are very likely to represent data rather than theoretical discussions. This leads to some underestimation as numeric data do occur in other sections of papers (e.g., Results) and some journals do not have formally designated experimental sections.

The number of instances of characterization data for individual substances was calculated by examining samples of the papers retrieved. For example, searching for mass spectral terms in the experimental sections retrieved over 4000 documents. Examining 26 of these selected at random from various portions of the retrieval set revealed detailed mass spectra associated with 62 individual substances. The number of spectra per document varied from 0 to 13. Extrapolating to the full file suggests that it contains from 9000 to 9500 mass spectra for individual substances.

A mass spectrum was only counted for the above calculation if it contained a reasonable number of peaks. However, five levels of detail are commonly given in these papers: (1) an exact molecular ion value, e.g., *m/z* Calcd for $C_{14}H_{15}N$: 197.1204; Obsd: 197.1209; (2) existence of the parent ion to confirm that a compound synthesized has the appropriate molecular weight, e.g., *m/e* 282 (M^+); (3) a few peaks; (4) a longer list of peaks with relative intensities, and possibly the parent ion, indicated; (5) as (4) but with mass losses identified, e.g., *m/e* 245 (M + 1), 244 (M), 227 (M − OH), 226 (M − H_2O), 214, 207 (M − OH − HF), 200, 195 (M − H_2O − CH_2OH), 187, 185, 157 (P + C_2H_4), 152, 149, 131 (P + 2H), 130 (P + H), 114. Similar sampling for other properties produced the estimates shown in Table III.

RETRIEVABILITY OF DATA

This section illustrates retrieving numeric data for a specific compound from the ACS primary journal file: first trying to retrieve all data (recall) and then searching for particular data types (precision). One difficulty in selecting a compound for a data-retrieval search is that most of the substances on the file are new and appear in only one paper. Selecting such a compound and then retrieving it would be somewhat less than convincing. Phencyclidine was chosen as the subject compound because it has considerable intrinsic interest. A potential disadvantage is that it has been known for over 2 decades, so one would not expect standard physical data for it to be present in a file covering only the last 3–4 years.

Phencyclidine [1-(1-phenylcyclohexyl)piperidine, PCP, CAS Registry Number 77-10-1] was introduced as a general anaesthetic 2 decades ago and performed splendidly—as an anaesthetic. Unfortunately, it is also a potent psychomimetic, causing intense, bizarre, and long-lasting symptoms including euphoria, suicidal and homicidal urges, and schizophrenia. Ironically, although these side effects caused its rapid discontinuance as a human anaesthetic, they are responsible for its major current use as possibly the most dangerous recreational drug available today, under the name of "angel dust". It is also often sold to unsuspecting users as cocaine, LSD, or THC.

As a first step, all references to phencyclidine were retrieved and examined for numeric data to determine absolute recall figures for later searches. The simple term PHENCYCLIDINE alone retrieved 20 of the 21 documents concerning phencyclidine. The following data types were found [type (times found)]: mp (2); ^{1}H NMR (2); ^{13}C NMR (1), MS (4), molform (1), chromatographic properties (3); ED_{50} (1); free energy (1); chronoamperometric D values (1); membrane binding (1).

Table IV. Numeric Data Retrieved for Phencyclidine

Property	Statement	Hits/Docs	Hits/Paras
MP	PHENCYCLIDINE same MP	1/4	1/8
Free Energy	PHENCYCLIDINE same (DELTA adj G)	1/1	1/1
Effective Dose	PHENCYCLIDINE same ED	1/3	1/4
1H NMR	PHENCYCLIDINE same (1H adj NMR)	1/4	1/4
13C NMR	PHENCYCLIDINE same (13C adj NMR)	1/5	1/54

Table V. Numeric Data Retrieved for a Specific Compound

MS	248 (M+), 189, 105 (100), 77
1H NMR	(CDCL3) +DELTA+ 3.28-3.70 (2 H, M), 3.87 (3 H, S), 5.07 (1 H, DD), 6.17 (1 H, BR S), 7.33-7.87 (5 H, M)
IR	1760, 1740, 1680 CM-1
MP	121-123 +DEGREE+ C

Note that although only 7 of the 21 documents yielded numeric data, this is not a practical problem as only paragraphs containing the search terms need be examined. Thus, one can rapidly and confidently discard irrelevant documents. It is impressive that a very simple search on a file that goes back only a few years provided a substantial amount of numeric data on a well-known compound chosen at random.

Users experienced in bibliographic searching will not be impressed by the precision (Table IV) obtained in trying to retrieve specific data on phencyclidine. However, this does not have the same significance as it would for bibliographic retrieval (or substructure searching) for two reasons. First, one sees the actual data, not a reference to it. Second, the number of retrievals is usually very small. For example, although the precision for the MP search is only 25%, it took less than 1 min to review the entire retrieval (eight paragraphs) and discover phencyclidine's mp.

Searchability of Data. Experiments show that it is easy to retrieve file compounds with extremely high precision by using only a small part of the numeric data normally available. For example, the data given for N^1-benzoyl-5-(methoxycarbonyl)-2-imidazolidinone in a single paper are shown in Table V.

Suppose that a chemist had prepared or isolated the above compound independently and wished to identify it. How easy is it to retrieve substances in the file compatible with the above numeric data?

The mass spectrum is an important illustration of the recall and precision that may be expected. When only the parent ion (M^+) was used, 159 documents were retrieved, which is too large to be useful; when the largest peak was added, only 15 documents were retrieved, which is manageable; and when all four peaks were searched, only two compounds were retrieved, one of which was a false drop because the 248 was a UV measurement in the same sentence.

The IR results were equally precise even though only three, rather common, peaks are reported. Requiring that the three peaks appear in the same sentence retrieved only two compounds, the other compound being benzyl 7-[(*Z*)-2-(methoxyimino)-2-phenylacetamido]-1-oxa-2-oxocepham-4α-carboxylate, whose quoted spectrum also contained peaks at 3370 and 1790 and which would have been readily distinguished by correlative searching, e.g., using MS or mp data. Similar results were obtained with UV spectra for other compounds.

Specific NMR spectral values cannot be searched as this implementation does not index decimal numbers. However, NMR can readily be retrieved for display by specifying that appropriate text terms appear in experimental paragraphs. The above results are typical; retrieving a specific compound from the file with high precision is generally straightforward with the information normally available to a bench chemist.

CORRELATIVE SEARCHING

The power of "correlative searching" (of matching on two or more independent properties) is even more striking. When the M^+ peak of the imidazolidinone was combined with the mp, only the subject compound was retrieved. For a compound that is a solid at room temperature, a mp is usually the simplest datum to obtain, so one would expect to find a lot of compounds with the same mp. In fact, 6789 documents contained the term mp, 213 had mp adjacent to 121, and 56 included the phrase mp 121–123. Correlative searching is analogous to ANDing together subject terms that are highly posted but rarely point to the same document.

Searches for 15 other randomly selected substances were entirely consistent with the imidazolidinone result; in each case, combining the parent (or largest) ion with the base peak retrieved at most two documents. Even more surprising, it is not usually necessary to include text terms in the profile. Thus, a search for *688 WITH* (*44 ADJ 100*) retrieved only two spectra, the expected mass spectrum and a false drop containing a UV peak! (This is possible because authors often format an entire experimental section as a single sentence with semicolons separating data types.)

In a large file containing only one data type (e.g., IR or MS), one would expect the difficulty of pinpointing a specific compound to increase sharply with file size. For correlative searching, this is not a problem as the characterization properties are essentially independent. Thus, much larger files can be searched simply. Lindsay et al.[1] relate that while there are 14 715 813 possible isomers of *N,N*-dimethyl-1-octadecanamine, the DENDRAL program generates only 1 284 792 structures when the mass spectrum is taken into account and only 1 when both the NMR and the mass spectrum are used.

PROBLEMS IN SEARCHING TEXT FOR NUMERIC DATA

Considering that the present implementation of the ACS journals was not intended to support numeric data searching, it is impressively successful in doing so. The problems encountered in this study arose from four sources: the specific implementation; the structure of the file itself; missing search operators; and difficulties inherent in searching free text.

The most serious deficiency is that decimal numbers are not present in the inverted index and are therefore unsearchable. The most important practical effect of this is that one cannot search specific NMR, optical rotation, or refractive index data as these are almost invariably given as decimal reals. For a similar reason, if a molform ends in a colon, a common style for analytical results, it is also truncated (e.g., $C_{21}H_{18}O_3$: becomes $C_{21}H_{18}O$), leading to both precision and recall errors. It must be emphasized that these are problems with one particular implementation only.

Another implementation problem is that systematic names are segmented at certain characters. For example, *p*-(dimethylamino)phencyclidine is stored as DIMETHYLAMINO and PHENCYCLIDINE and will be retrieved in a search for PHENCYCLIDINE because there is no operator to restrict the search to self-complete terms.

The major problem with the structure of the ACS file is that the CAS REGs are linked to documents, not to individual substances. For example, searching for *77-10-1.RN AND MS* only guarantees that both PHENCYCLIDINE and MS were indexed for the same document, not that the mass spectrum of phencyclidine was given; indeed, it is more likely to be that of another compound in the same document. For this reason, it would be preferable to insert each CAS REG after the name (or symbol) to which it refers. As shown in the identification section (below), this can be accomplished algorithmically.

The search system used has the following proximity operators: AND, in the same document; SAME, in the same paragraph; WITH, in the same sentence; ADJ, adjacent or separated only by stopwords and in the specified order. However, due to the unpredictability of free text, these are not always sufficient for numeric data, for which several additional operators are needed: for range searching, for recognizing numeric strings, for masking in certain types of data, and for an unusual kind of Boolean logic.

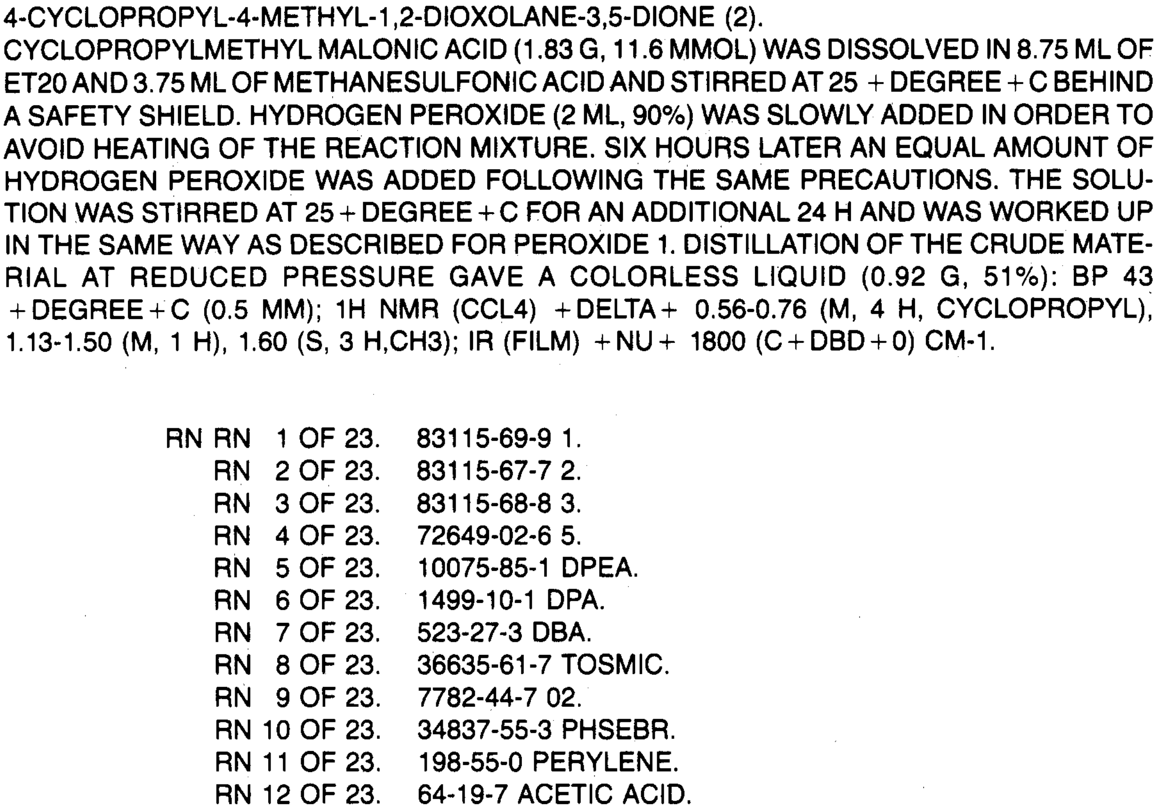
A *range* operator is needed because measurements are usually given to one more significant figure than can be reproduced in a different laboratory. For example, melting points may vary by one or two degrees but are often quoted to the nearest half degree. Similarly, IR spectra, which have a range of over 2000 wavenumbers, are quoted to the nearest integer, while ^{1}H NMR spectra, which have a range of perhaps 10 units, are often measured to the nearest 0.01 unit. Small differences in purity and laboratory conditions can cause measurements to vary by more than this. Thus, one needs to be able to specify a range.

A *numeric* operator is needed because sometimes one wants to specify that a term be a number rather than a specific value; e.g., (MASS *ADJ* SPECTRUM) *WITH number* would retrieve sentences containing numeric data rather than those that simply mention mass spectra. Similarly, a masking operator is needed for data labels that include variable conditions, e.g., ALPHA#D and N#D, in which # is the temperature at which the measurement was taken. Without this capability, one must input all possible variants to ensure complete recall.

An unusual kind of search logic is also needed, which can be informally expressed as "match if field is present, ignore if not". For example, the user may specify an IR peak of 1650. If a substance has an IR field, then the 1650 peak must be present for the record to be a potential retrieval. However, if the record does not have an IR field, it should not be discarded.

Three problems inherent in free-text searching are the great variability of natural language, the difficulty of defining context, and homonymy. All are illustrated by problems in searching for mass spectral data in experimental sections. Consider, for example, the form of the data retrieved in the phencyclidine search: M/E (% OF BASE) 243 (75, M+), 242 (40), 200 (100), 186 (40), 166 (40), 91 (80), 84 (30). One would like to be able to identify the parent ion as this is the most precise way of matching an unknown spectrum. Many mass spectra may contain a 243 peak and thus will be retrieved by (*M ADJ E*) *WITH 243*, but it will be the parent ion in very few of these. However, there is no easy way to specify that 243 be the first peak given. If there is no text between *m/e* and the data, then (*M ADJ E*) *ADJ 243* will suffice. However, various phrases (e.g., relative intensity) may intervene. Similarly, the parent ion may be identified by M+ or PARENT ION, which may be preceded or followed by the relative intensity. In addition, a paragraph in the experimental section is often a single sentence, with different data types separated by semicolons, so a UV peak of 243 (nm) may be retrieved because it is in the same sentence as *m/e*.

Homonymy rears its ugly head with respect to both *m/e* and MS. *m/e* must be expressed by *M ADJ E*, which also retrieves such names as M. E. Stevenson. Similarly, MS is also an abbreviation for millisecond. Other examples abound: CD denotes both CIRCULAR DICHROISM and CADMIUM; ED denotes EFFECTIVE DOSE, EDITOR, and EDI-



4-CYCLOPROPYL-4-METHYL-1,2-DIOXOLANE-3,5-DIONE (2).
CYCLOPROPYLMETHYL MALONIC ACID (1.83 G, 11.6 MMOL) WAS DISSOLVED IN 8.75 ML OF ET20 AND 3.75 ML OF METHANESULFONIC ACID AND STIRRED AT 25 +DEGREE+C BEHIND A SAFETY SHIELD. HYDROGEN PEROXIDE (2 ML, 90%) WAS SLOWLY ADDED IN ORDER TO AVOID HEATING OF THE REACTION MIXTURE. SIX HOURS LATER AN EQUAL AMOUNT OF HYDROGEN PEROXIDE WAS ADDED FOLLOWING THE SAME PRECAUTIONS. THE SOLUTION WAS STIRRED AT 25+DEGREE+C FOR AN ADDITIONAL 24 H AND WAS WORKED UP IN THE SAME WAY AS DESCRIBED FOR PEROXIDE 1. DISTILLATION OF THE CRUDE MATERIAL AT REDUCED PRESSURE GAVE A COLORLESS LIQUID (0.92 G, 51%): BP 43 +DEGREE+C (0.5 MM); 1H NMR (CCL4) +DELTA+ 0.56-0.76 (M, 4 H, CYCLOPROPYL), 1.13-1.50 (M, 1 H), 1.60 (S, 3 H,CH3); IR (FILM) +NU+ 1800 (C+DBD+0) CM-1.

```
RN RN  1 OF 23.   83115-69-9 1.
   RN  2 OF 23.   83115-67-7 2.
   RN  3 OF 23.   83115-68-8 3.
   RN  4 OF 23.   72649-02-6 5.
   RN  5 OF 23.   10075-85-1 DPEA.
   RN  6 OF 23.   1499-10-1 DPA.
   RN  7 OF 23.   523-27-3 DBA.
   RN  8 OF 23.   36635-61-7 TOSMIC.
   RN  9 OF 23.   7782-44-7 02.
   RN 10 OF 23.   34837-55-3 PHSEBR.
   RN 11 OF 23.   198-55-0 PERYLENE.
   RN 12 OF 23.   64-19-7 ACETIC ACID.
```

Figure 1. A simple experimental paragraph and CAS REG table from JOC.

TION; the Greek letter α has numerous referents; and DEGREE may refer to angle or temperature.

In spite of these problems, searching the full text of primary journals is a powerful new tool for chemists. Conversely, searching a numeric database offers certain advantages; e.g., each data type can be stored in its own field, the data can be formatted uniformly and predictably for both search and display, and other data can be associated with each type (e.g., structures or literature references). At the moment, numeric databases and the primary journals are complementary because they contain different sets of substances, the former are usually validated and the latter are much more current. The rest of this paper shows how both currency and searchability may be achieved by using computers to identify numeric data in context in primary journals.

(2) AUTOMATIC IDENTIFICATION OF NUMERIC DATA

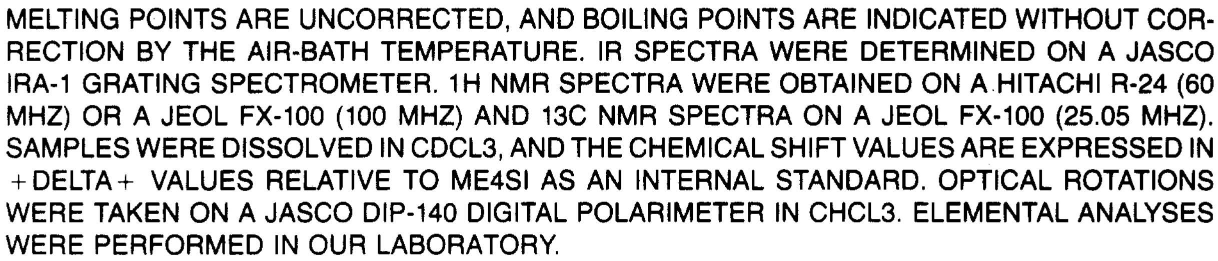
The goal of the work described in this half of the paper was to devise an algorithm for processing experimental paragraphs to identify the subject compound, its CAS REG (if given), and each type of numeric data and to link any measurement conditions given in the experimental section with the appropriate properties. One can view this in terms of retrieving data, structuring the file, or making implicit links explicit. The above information units are automatically associated in the chemist's mind when the paper is read, but the connections are not reflected in the standard file implementation. Linking these units explicitly is a small step toward reflecting the powerful structuring that is partly in the chemist's mind and partly in subtle and complex linguistic forms. The difficulties involved in automating this process may also provide some insight into the cognitive processes of chemists.

The results show that this goal is obtainable. The initial version of the program identified 84% of the subject compounds, 95% of the characterization data, and essentially all of the measurement conditions for a test file and assigned a degree of confidence to its identification of the subject compound. Analysis of the output and trial modifications of the program suggest that the first result can be increased to over 90% and the other two to nearly 100%. It also seems likely that the algorithm could be readily extended to non-ACS journals containing experimental-like paragraphs. Another potential use of the algorithm is to insert CAS REG in experimental paragraphs, which should greatly increase the precision of substance-oriented searches of the primary journals.

The program was written by using SPITBOL,[2] a macrointerpreter for SNOBOL4 (a high-level string-processing language), on the Unix operating system as this allowed rapid development of the algorithm. The theoretical interest of the program lies in the fact that it uses no linguistic theory whatsoever. There is no syntax, no semantics, no inference, and no grammar of any kind, simply observation of surface patterns. One would expect such a strategy to be cost effective but severely limited. Surprisingly, it seems quite adequate for this task.

The test file contained 248 paragraphs from 12 papers published in the three ACS journals. Testing was not confined to any particular kind of experimental paragraph; essentially, the complete experimental sections for 12 papers containing significant amounts or types of numeric data were used. The paragraphs, which were not edited, were used in the restricted character set typical of on-line bibliographic search service implementations. (In only one minor aspect was the restricted character set found to be a disadvantage.) Both the printed and the on-line versions of a paper in JOC or OM contain a table of CAS REGs separate from the text (at the end of the document in the former and in a different field in the latter) that cross-refers the name or symbol selected by the author to denote a substance and its CAS REG. These tables were included in the test file.

The Problem. Consider the typical experimental paragraph and entries from the corresponding CAS REG table shown in Figure 1. The output from the test program for this paragraph is shown in Table VI. Many authors also include



MELTING POINTS ARE UNCORRECTED, AND BOILING POINTS ARE INDICATED WITHOUT CORRECTION BY THE AIR-BATH TEMPERATURE. IR SPECTRA WERE DETERMINED ON A JASCO IRA-1 GRATING SPECTROMETER. 1H NMR SPECTRA WERE OBTAINED ON A HITACHI R-24 (60 MHZ) OR A JEOL FX-100 (100 MHZ) AND 13C NMR SPECTRA ON A JEOL FX-100 (25.05 MHZ). SAMPLES WERE DISSOLVED IN CDCL3, AND THE CHEMICAL SHIFT VALUES ARE EXPRESSED IN +DELTA+ VALUES RELATIVE TO ME4SI AS AN INTERNAL STANDARD. OPTICAL ROTATIONS WERE TAKEN ON A JASCO DIP-140 DIGITAL POLARIMETER IN CHCL3. ELEMENTAL ANALYSES WERE PERFORMED IN OUR LABORATORY.

Figure 2. A simple condition paragraph.

Table VI. Output for Paragraph in Figure 1

```
REG      = 83115-67-7
NAME     = 4-CYCLOPROPYL-4-METHYL-1,2-DIOXOLANE-3,5-DIONE
SYMBOL   = 2
IR DATA  = (FILM) +NU+ 1800 (C+DBD+0) CM-1
1H-NMR   = (CCL4) +DELTA+ 0.56-076 (M, 4 H, CYCLOPROPYL),
DATA       1.13-1.50 (M, 1 H), 1.60 (S, 3 H, CH3)
BP DATA  = 43 +DEGREE+C (0.5 MM)
YIELD    = 0.92 G, 51%
```

a paragraph in the experimental section that gives general information about the characterization data. A simple example is shown in Figure 2.

Note that the condition information may be essential to full specification of the numeric data; e.g., the optical rotation is dependent on the solvent ($CHCl_3$) used. The main problem here is not identifying the different types of condition data but rather distinguishing condition paragraphs from synthesis paragraphs and several other types of general paragraph.

Identifying the Subject Compound. Identifying the subject compound appears trivial and, indeed, is for the example given in Figure 1. However, quite difficult cases do arise and, counterintuitively, it proved the most difficult of the four information units to process.

Four ways of identifying a subject compound in an experimental paragraph are as follows: (a) by position, the subject compound is often given in the heading and, if not, is usually mentioned in the first sentence; (b) by context, the subject compound can often be identified via patterns found immediately before the numeric data; (c) by character set, systematic names normally contain an unusually high proportion of nonalphabetic characters (e.g., hyphens, parentheses, and digits); (d) by roots, most substance names and all systematic nomenclature contain strings denoting chemical entities such as radicals, ring systems, functional groups, and elements. Note that strategy a is necessary (though not always sufficient) because it is the only one capable of identifying the subject compound per se. Also mentioned in the above example are CYCLOPROPYLMETHYL MALONIC ACID, ET20, METHANESULFONIC ACID, and HYDROGEN PEROXIDE. (In fact, HYDROGEN PEROXIDE appears twice, so a frequency method would select it as the most important compound!)

The program uses only position and context, which appears to be sufficient in practice. The character-set strategy was not used because it did not seem to be necessary. However, it can be useful for confirming that a string is plausible as a chemical name and is employed for this purpose in the condition–identification algorithm. The root approach was avoided because the hundreds of substrings required would exceed both the time and storage available.

Use of the Heading of First Sentence. In the simplest and most common case, the name of the subject compound is the heading of the paragraph. However, it may consist of more than one word and/or be followed by a symbol. Also, there may be no heading as such but a sentence containing the name of the subject compound. Some common forms are shown in Table VII.

To a chemist, the above forms are trivially interconvertible. A very powerful artificial intelligence program might also be able to "understand" the equivalence of these forms. The program has no understanding whatsoever; it simply uses the surface patterns. However, it is capable of suggesting (though not confirming) that the two descriptions refer to the same entity and required neither 20 years of education nor thousands of hours of programming. The problem is to define patterns that allow for each case but do not conflict with each other or match inappropriate text strings.

Use of the Substance Symbol. Often, the author uses a symbol in the heading (or first sentence) to denote the subject compound. If this symbol also appears in the CAS REG table, this is a good indication that the correct symbol has been identified. For example, the heading in Figure 1 contains only

Table VII. Contexts of Subject Compounds

Format	Example
WORD	4-CYANOSPIRO[5.2]OCTANE
WORD—PARENTHESIZED SYMBOL	ACETYLTRIMETHYLSILANE (3)
PHRASE	2-METHYL-2-CYCLOPROPYLMALONIC ACID
PHRASE—PARENTHESIZED SYMBOL	+ALPHA+-DEUTERIOVINYL METHYL ETHER (13)
PREPARATION OF compound	PREPARATION OF (+ETA+-C5H5)RE(CO)2D2
PREPARATION OF compound symbol By	PREPARATION OF N-BENZOYL-3-METHYL- 2,3-N-BENZOYL-3-METHYL-2,3- DIDEHYDROHOMOSERINE +GAMMA+-LACTONE (10) BY
SYNTHESIS OF descriptor symbol	SYNTHESIS OF CARBOXIMIDE 10A
compound comma-blank symbol	COCL (PPH3)2(C8H402), 1
compound WAS PREPARED...	2-(ETHOXYMETHYL)-3-METHYL-1,4- NAPHTHOQUINONE WAS PREPARED AS

Table VIII. Data Types Identified

Type	Tag
IR spectrum	IR, infrared
1H NMR spectrum	1H NMR, NMR
13C NMR spectrum	13C NMR
Melting point	MP
Boiling point	BP
UV spectrum	UV
CD spectrum	CD, CIRCULAR DICHROISM
Optical rotation	:+ALPHA+:##D
Mass spectrum	MS, MASS SPECTRUM
RF value	RF, TLC
Molform	pattern
Yield	pattern

a single word and a parenthesized symbol, so the existence of the same symbol in the CAS REG table is powerful confirmation that both the subject substance and the author symbol have been identified. However, experimental paragraphs often contain many strings that fit the format of symbols, and the lower digits are almost invariably present in CAS REG tables.

In the Body of the Paragraph. An experimental paragraph often contains a string or symbol denoting the subject compound immediately before the characterization data. A typical example is ...TO GIVE 45 G (78%) of 3: BP 112..., which has the pattern:

<gave> <yield> OF <symbol> colon-blank <data_tag>

where <> indicates a nonterminal, <gave> is phrases like TO GIVE, GAVE, AFFORDED, and YIELDED, <yield> has the format

<number> blank <unit> blank (<number> %)

and ⟨data_tag⟩ is a label for numeric data (e.g., IR, 13C NMR, MS, or MP), with parentheses indicating optionality and the nonterminal <symbols> being expanded in the obvious manner.

If the symbol in the heading is also found in a pattern of this kind, this is strong confirmation that the program has correctly identified the subject compound and the author symbol. This is especially important when there is no CAS REG table for the paper or the subject compound was not registered in time for publication. The program uses this to assign a degree of confirmation to the identification. Unfortunately, many authors, showing an admirable sense of economy and an understandable failure to anticipate programs of this type, prefer to describe the physical state of the product rather than to identify it redundantly, e.g., gave a colorless liquid (0.92 g, 51%): bp 43 (Figure 6).

Most of the experimental paragraphs examined were amenable to the above techniques (see Results). The commonest exception was the paragraph dealing with more than one substance.

Identifying the Characterization Data. The types of numeric data currently identified are shown in Table VIII, together with the tags that normally label them and that would, of course, be equally useful for searching (# denotes a digit, while "pattern" indicates that context is used because there is no explicit tag).

The various types of numeric data are identified by (1) locating a data tag and (2) accepting all characters until another data tag or the end of the sentence is reached. This very simple method is effective because authors in all the journals examined are strikingly consistent in the way in which they express characterization data. Normally, the author expresses a series of properties in the format

TAG DATA TERMINATOR TAG DATA
TERMINATOR...TAG DATA END_OF_SENTENCE

where the terminator is usually a semicolon–blank, but occasionally a comma–blank.

Identifying Condition Paragraphs and Measurement Conditions. Two problems must be solved to identify measurement conditions successfully: (1) distinguish between condition and noncondition paragraphs and (2) identify statements about each kind of measurement. Curiously, the first problem is harder to solve than the second. Thus, it is not enough to devise an algorithm that processes a file of condition paragraphs successfully because in practice the condition-identification algorithm must be a subroutine that operates only when appropriate, usually once per document. The problem is complicated both because synthesis and condition paragraphs have much vocabulary in common (e.g., use the same labels for NMR, IR, and mass spectra) and because there are several types of general (nonsynthesis) paragraphs (e.g., those describing workup or synthetic procedures that apply to subsequent synthesis paragraphs). Moreover, two or more general paragraphs may be combined and describe reagents, reaction conditions, apparatus, and instrumentation. The following characteristics are used to decide if a paragraph is a condition paragraph.

(A) Sequence. The condition paragraph is often the first paragraph of the experimental section. Even if it is not, it is very unlikely to occur after the first synthesis paragraph because the instrumental conditions discussed will apply to all the compounds characterized. Thus, being other than the first paragraph is mildly disconfirming, while occurring after a synthetic paragraph is strongly so. (This is not necessarily true of other types of general paragraph; e.g., general information about X-ray crystallography or biological activity may be sandwiched between a block of synthesis paragraphs and the paragraphs to which they apply.)

(B) Uniqueness. There is normally only one paragraph concerning the measurement of characterization data. Thus, if a "reasonable" number of condition types are identified in a paragraph, this is moderately convincing evidence that subsequent paragraphs may be ignored. (This is not nearly so certain for paragraphs describing reaction conditions as there may be more than one type of synthesis involved, leading to the pattern: reaction paragraph, synthesis_1, reaction paragrah, synthesis_2, ...)

(C) Heading/First Sentence. If a paragraph has a heading, this is usually decisive whether it is a general heading (e.g., Materials) or a subject compound. The division between heading and first sentence is blurred and the program distinguishes only on the basis of length. If the first sentence contains less than five words, the number of characters indicative of systematic nomenclature is counted in an attempt to identify this and so rule out the paragraph. This only works for headings, not whole sentences, as instrument names often contain numerics (though a more elaborate test would recognize the difference easily enough). Also, if the heading contains a symbol, this is strongly disconfirming.

(D) Style of Paragraph. Condition paragraphs generally use the plural form while synthesis paragraphs strongly favor the singular. Simplistic though it may seem, merely counting WAS and WERE gives a reasonable indication of whether a paragraph is one or the other. Similarly, SPECTRA often occurs in condition paragraphs, but is rarely used in describing the synthesis of a single compound.

(E) Patterns in Paragraph. If a paragraph contains one pattern characteristic of a yield, or several of a quantity, it is very likely to be concerned with synthesis rather than characterization conditions.

Table IX. Occurrence of Data Types and Results

Journal	Subs.	REG	Names	IR	MP	NMR	MS	UV	OR	BP	CH	CD	MF	YD
JOC	5	4	4	4	4	1	3,2	3	5	—	1	3	3	—
JOC	23,1	19,1	18,1	4	12	18,1	4	—	—	—	7	—	6	10,5
JOC	23,5	23,5	23,5	24	9	34	—	—	27	17	—	—	22	25,3
JOC	9	7	9	9	3	9	2	—	—	5	—	—	9	8
JOC	4,7	4,7	5,6	—	7,1	1	1	1,1	—	—	1	—	9	3,5
JOC	23,5	23,4	—	3	19	40	27,1	—	—	2	—	—	16	26,1
OM	11,1	11,1	11,1	9	8	8,1	—	—	—	—	—	—	3,4	7,3
OM	8	—	7	5	—	7	6,1	—	—	6,1	—	—	5	6
OM	8	6	7	2	5	5	5	—	—	—	—	—	7	5,1
JMC	6,1	—	6,1	2	7	2	5	—	—	—	—	—	5	6
JMC	11,9	—	10,6	1	18	5	2	—	30	—	4	—	16	17,2
JMC	9	—	9	2	2	5	5	—	—	—	2,2	—	5	5
Hit	140	97	109	65	94	135	60	4	62	30	15	3	106	118
Miss	29	18	20	—	1	1	4	1	—	1	2	—	4	20
%	83	84	84	100	99	99	94	80	100	97	88	100	96	86

(F) Vocabulary. Some terms are strongly associated with condition paragraphs (e.g., instrument names such as VARIAN or PERKIN-ELMER), while others (e.g., ADDED, WORKED, GAVE, DROPWISE) are highly indicative of synthesis paragraphs. Even though some authors combine general reaction and instrumental conditions in the same paragraph, a preponderance of one type of vocabulary is significant.

RESULTS

The results of processing the experimental paragraphs in the test file are shown in Table IX, where NMR refers to all types of NMR, CH refers to all types of chromatography, YD refers to yield, and the other headings have their usual meanings. The figure before the comma is the number of instances identified by the program (the hits), while the figure after it records the number of failures (misses) and is omitted if it is 0.

As Table IX shows, the program correctly identified 83% of the subject compounds and 95% of the characterization data. There were few subject compound misidentifications. The identification performance can very probably be improved to over 90%, though some cases are almost intractable in principle. Quoting a single figure is deceptive as most of the identification failures came from three papers. The 95% success rate for the identification of characterization data is easier to increase; only trivial modifications are necessary for the program to be able to recognize virtually all characterization data. It is also interesting that the number of properties per substance (excluding yield) was 3.2, which has favorable implications for correlative searching.

The identification program is based on a very simple model of an experimental paragraph: one subject compound per paragraph; self-complete paragraphs (except for elemental analyses, which often constitute a secondary paragraph); a name and/or symbol heading or a first sentence that follows one of a small set of simple patterns; properties consisting of tags followed by data; and data referring only to the subject compound.

If a paragraph does not conform to this model, the program usually fails to some extent, though many of the problems are tractable in principle. However, most experimental paragraphs do conform, although some are notably more obscure than others. For most, one could teach a nonchemist to identify the subject compound; for others, a chemist would be needed; and one or two would mislead most chemists.

Of the 12 papers in the test file, 11 had typical measurement condition paragraphs and one was a special case. The file used to test the condition–identification algorithm consisted of all the paragraphs up to and including the first synthesis paragraph or the last of the noncondition general paragraphs and contained 30 paragraphs: 12 condition, 9 general, and 9 synthetic. It was intended to simulate processing of complete experimental sections, for which this algorithm would be a subroutine active until either a condition or synthesis paragraph had been conclusively identified and then becoming dormant for the rest of the current document. The goal is thus (1) to select condition paragraphs, (2) to reject all other paragraphs, and (3) to identify measurement conditions.

All 11 typical condition paragraphs were identified. The special case, which consisted entirely of an extensive description of chemiluminescence measurement, was not. All nine synthesis paragraphs were rejected, but three general (nonmeasurement) paragraphs were accepted. However, the false acceptances had no practical effect as no conditions were present to be identified. The program recognized that they were not condition paragraphs, because of their lack of measurement conditions, and continued searching. A false acceptance cannot occur when a condition paragraph is the first experimental paragraph.

The occurrence of conditions in the test file is shown in Table X, where C-NMR denotes ^{13}C NMR, P-NMR denotes ^{31}P NMR, FS denotes fluorescence spectra, OR denotes optical rotation, CH denotes various types of chromatography, and the other acronyms have the usual meanings. Identification was essentially complete, but a few qualifying points are necessary.

Most condition statements occupy a single, complete sentence. Complications arise when a measurement condition is expressed in more than one sentence or, conversely, when two data types are dealt with in the same sentence. As an example of the first case, consider the following output from the program:

```
1H-NMR CONTEXT = 1H NMR SPECTRA WERE RECORDED ON A VARIAN EM-360 OR A JEOLCO MH-90 SPECTROMETER.
G-NMR CONTEXT = CHEMICAL SHIFTS ARE REPORTED IN +DELTA+ UNITS (PARTS PER MILLION
```

Table X. Occurrence of Data Types in Condition Paragraphs

Paper	NRM	C-NMR	P-NMR	MS	IR	UV	EPR	FS	OR	MP	BP	TLC	CH
1	X			X	X				X	X		X	
2	X	X			X				X	X	X		
3	X					X				X		X	X
4	X			X					X				X
5	X			X	X	X	X	X		X			X
6	X	X		X	X					X			
7	X		X		X					X			X
8	X	X		X	X					X			
9	X			X	X								X
10	X			X	X					X		X	
11	X			X	X				X	X			
Total	11	3	1	8	9	2	1	1	4	9	1	3	5

RELATIVE TO TETRAMETHYLSILANE) IN THE INDICATED SOLVENT.
G-NMR CONTEXT = COUPLING CONSTANTS ARE REPORTED IN HERTZ.

The first sentence is clearly labeled as a 1H-NMR context, but the other two can, from their own content, only be associated with NMR in general. Since only one type of NMR is mentioned in the paragraph, this presents no problem. Similarly, if such statements are sandwiched between two types of NMR, one can confidently assign them to the first. However, if two types of NMR and conditions such as these are mentioned in consecutive sentences, the scope of the conditions might not be clear. This raises the general problem of how much text should be included in the condition statement, but problems of this kind seem to occur mainly with chromatography and not to be a serious problem in practice.

The converse problem is that certain data types are more likely than others to be textually associated (e.g., melting and boiling points, varieties of NMR) for semantic reasons. The third sentence illustrates a rarer case, where the connection is syntactic; two statements are conjoined because each is brief:
MELTING POINTS ARE UNCORRECTED, AND BOILING POINTS ARE INDICATED WITHOUT CORRECTION BY THE AIR-BATH TEMPERATURE.
1H NMR SPECTRA WERE OBTAINED ON A HITACHI R-24 (60 MHZ). OR A JEOL FX-100 (100 MHZ) AND 13C NMR SPECTRA ON A JEOL FX-100 (25.05 MHZ).
MASS SPECTRA WERE OBTAINED ON A VG-7070F SPECTROMETER AND IR SPECTRA ON A PERKIN-ELMER 177 SPECTROMETER.

The test program simply links the whole sentence to each data type it concerns. Splitting the conjoined sentences, which is more desirable in principle, requires a degree of linguistic sophistication quite beyond the current program. Note, for example, that the verb phrase WERE OBTAINED is elided from the second kernel of the last sentence.

FUTURE WORK

This study is capable of extension along several dimensions. Only a dozen of the thousands of data types that occur in chemistry journals were studied. The three journals used contain most of the characterization data present in all the ACS primary journals but supply only a small percentage of the documents for *Chemical Abstracts*. It seems very likely that many other chemistry serials are available in machine-readable form via computer-assisted composition. A cursory examination of the *Journal of the Chemical Society Perkin Transactions 1* suggests that the identification algorithm could be readily adapted to it. Also, the identification algorithm applies only to experimental paragraphs because these are relatively well-defined and highly structured. Identifying similar data in the bodies of papers would be dramatically more difficult. Whether data could be identified in tables or figures would depend crucially on the file structure of the text. Inserting CAS REG in the body of the text would probably be easier than identifying data. Currently, the data are simply identified, but it might be desirable to massage them advantageously in property-specific ways.

CONCLUSIONS

Now that searching primary journals on-line is feasible, a major source of highly current numeric data has been opened up to chemists, which is already of a comparable size to and complements conventional numeric databases and is growing rapidly. High recall and precision can readily be achieved because of the power of correlative searching.

REFERENCES AND NOTES

(1) Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. "Applications of Artificial Intelligence for Organic Chemistry. The DENDRAL Project"; McGraw-Hill: New York, 1980.

(2) Dewar Information Systems Corporation "MacroSPITBOL Version 3.5 Program Reference Manual"; Dewar Information Systems Corp.: Oak Park, IL, 1980.