FEATURE ARTICLES

Perspectives on Editorial Operations of Chemical Abstracts Service

RUSSELL J. ROWLETT, JR.†
Caropines, Myrtle Beach, South Carolina 29577

Received December 17, 1984

These are perspectives on the major reorganization and retraining of the scientific editorial staff that accompanied the computerization of Chemical Abstracts Service (CAS) from the mid 1960s through the early 1980s. The corps of volunteer abstractors was replaced over a 15-year period with in-house, full-time, document analysts. Concurrently, the former separate processes of abstracting and indexing were combined in a single intellectual step. New techniques were developed for measurement of editorial productivity, and new groups were made responsible for constructing structural formulas and for naming new complex chemical substances. By 1982, a sufficient organization was in place, operating efficiently, to carry CAS forward into the online information age.

The mid 1960s to early 1980s was a period of major change for the editorial and production operations of Chemical Abstracts Service (CAS). Much has been written[1,2] about the development and installation during this period of computer hardware and software that enabled CAS to survive in face of continuing growth in both the chemical literature and the multidiscipline interests of chemists. Little has been written about the massive reorganization and retraining of the scientific staff that was so necessary to continue to build the quality data base in accord with the life-long standards of CAS. These are perspectives on the human engineering that paralleled the many computer advances. It is the story of professional scientists revamping their organization and work assignments in a period of both rapid growth in the literature and adaptation of new technology in production.

In the last half of the 1960s CAS set out to build an automated information-handling system that would produce printed abstracts and indexes more efficiently and economically while at the same time provide the basis for new approaches to accessing and retrieving information.[3] This dual objective led to two significant results. First, it allowed the editorial and production operations to survive despite continued literature growth and rising costs.[4] Second, it provided the computer-readable data base from which multiple services have been derived,[5] the most important of which may be CAS ONLINE. Without this foresight of 20 years ago, today's online services would not be possible.

PHASE OUT OF VOLUNTEER ABSTRACTORS

From its beginning in 1907 through the mid 1960s, CAS relied on volunteers to produce abstracts for publication in *Chemical Abstracts* (CA). In 1966 the number of volunteers reached a peak of 3292.[6] Obviously, an automated information-handling system, with timely delivery of information and references, could not be built on the basis of unpredictable, even though dedicated, world-wide volunteers. The first, big, people-oriented task was to bring the preparation of abstracts into the Columbus, OH, offices. Here, abstracting could be performed by scientists uniformly, timely, and in high quality. However, this was a formidable challenge. For 60 years the Columbus editorial offices had been concerned primarily with editing abstracts, preparing index entries, editing index entries, and naming chemical substances. There were very few Columbus-based abstractors.

The world-wide volunteers included both language and subject experts in many fields of science. Could such expertise be obtained in Columbus? Could the high-quality subject content provided previously by volunteers be maintained in Columbus? The volunteers were active in research with hands-on experience. Would an office-bound scientist of similar training, but dissociated from on-going research, be able to maintain current knowledge of specific subject fields? These were tough questions for CAS management, but the step had to be taken if a timely information service derived from a computer-readable data base was to be built.

Volunteer abstractors were phased out gradually over a period of about 15 years. In 1984 only 4%[4] of the abstracts were obtained from workers outside of the Columbus offices. (Work performed by the remaining foreign associates is considered a part of the Columbus offices.) These were from difficult-to-read foreign documents. To accomplish this big change, additional scientists had to be employed. The best sources were competent industrial scientists who wanted to work outside of actual laboratories and who were former users of CAS services. The on-board staff had to be trained both in reading and in understanding additional languages. This was accomplished through regular courses of The Ohio State University plus classes organized in the CAS building. A Russian language professor gave regular in-house classes for several years on a full-time basis.

It was discovered early that in-house abstractors and indexers maintain their knowledge of current research accomplishments through their daily close association with the literature of their scientific field. They are reading daily about the cutting edge of new research and technology developments with access to a broad spectrum of all areas of science. In

† R.J.R. is a former Editor and Director of Publications and Services of Chemical Abstracts Service. Present address: 18 Meadow Oak Drive, Caropines, Myrtle Beach, SC 29577.

0095-2338/85/1625-0061$01.50/0

addition, their expertise continues to be expanded through attendance at national symposia and seminars such as Gordon Conferences, ACS short courses, lectures by visiting scientists, and joint seminars with The Ohio State University. A few members of staff who needed specific additional scientific background were encouraged to take regular university courses as a part of their normal work day. Over the 15-year phase out of volunteers, long-time users of CAS services did not complain about any loss in quality of content. There was none.

There were two significant losses in the transfer of abstract production to Columbus. The volunteer abstractors constituted a large, organized group of users who were in periodic contact with the editorial offices. They offered advice on new fields of chemistry and were constantly alert for errors. This assistance has not been replaced. Also, there was a concurrent body of almost 100 Section Editors, some of whom read galley proof of their Section for every CA abstract issue. With the advent of computer composition, there was no more galley proof and no time for such careful review. Section Editors were renamed Section Advisors, but understandably, their interest waned when they had little chance to interact directly in the editing stages. They could comment only after an issue was published. After-the-fact comments are never as stimulating as being a part of the initial action.

The CAS Editorial Advisory Board, organized in 1975, and now discontinued in favor of a larger user panel, provided good advice to the Editor on data base content improvements and transmitted many excellent user comments and suggestions. They also served to answer questions directly from users in their large organizations. The value of this Board and its service were never fully appreciated by the total CAS management. They were a distinct help to the Editor in the difficult transition period.

Volunteer abstractors were phased out with appreciation and recognition for their long years of valuable service. Certificates were awarded to all. Those with 10 and 25 years or more of service were recognized with lapel pins and special certificates. This was one of the most popular gestures ever extended by CAS. Gratitude was expressed by volunteers from around the world. Many volunteers had abstracted for more than 40 years. The former CA Editor E. J. Crane often referred to these as "the iron men of CA".[5] There were 26 chemists who were abstractors or Section Editors for 50 years or more.

ABSTRACT PREPARATION IN COLUMBUS

The problem of absorbing most of the abstracting in Columbus was exacerbated by the continued simultaneous growth of the literature. During this period of approximately 15 years, the number of documents abstracted in a given year doubled. In these years were published more than half of all the abstracted contained in CA since its origin in 1907. Thus, the editorial staff had to handle the ever increasing literature growth plus the abstracting being absorbed from the volunteers. These tasks were accomplished within available budgets as approved by the ACS Committee on CAS and the ACS Board of Directors. In the late 1960s many in ACS–CAS management felt these double tasks could not be accomplished at the same time. An important ingredient of success was the gradual substitution of computer handling for many of the manual chores formerly done by professional scientists. Thus, the latter were freed to spend more time on their professional assignments. Also, the "can do" attitude and commitment of the CAS editorial staff had much to do with the success of this important step.

For its first 60 years CA abstract formats were similar but differed according to their subject content. With over 3000 volunteers, standardization was hardly practical. With a centralized abstracting group, standardization was not only possible but mandatory to simplify and accelerate the process. Accordingly, in the early 1970s CAS adopted the "findings-oriented" abstract approved by the American National Standards Institute.[7] This was not a major change, but it was a substantial improvement in building a data base that now can be searched more consistently.

UNIFIED DOCUMENT ANALYSIS

Until the 1970s CA abstracts were printed and distributed before the corresponding indexing was even started. Abstract identification numbers were established by the printer during page make-up. Such numbers were required for index references. The advent of the computer made possible the use of temporary abstract numbers to which index entries could be referenced. The temporary numbers could be replaced later by the actual published references. Thus, for the first time it was possible to consider combination of the two steps, abstracting and indexing. Such a combination became known as Unified Document Analysis.

This was the most important and most difficult change in professional staff assignments ever undertaken at CAS. Its long-range implications were very significant in savings of both time and costs. Studies revealed that the intellectual effort required to understand a scientific document and prepare both the abstract and the index entries was about equal to the time required to perform one of the tasks separately. The time-consuming step was the intellectual understanding of the document. When abstracting and indexing were performed separately, the intellectual understanding step was performed twice, usually by different staff members.

Pilot studies[8] were initiated in 1970 on integration of abstracting and indexing from biochemical documents. Throughout these several programs, biochemists, under the leadership of Roger Moody, were more receptive to new ideas and willing to try new procedures. Hence, they were the venue for many editorial pilot studies. Results substantiated the major savings in both time and costs. However, the experience demonstrated that a major reorganization of editorial staff assignments was going to be required.

Almost from CA's origin, the editorial staff was organized by four major tasks: abstracting, although little was done in Columbus prior to the 1970s; abstract editing; indexing; index editing. Within these task areas there were scientists skilled in the four general divisions of chemistry: organic, inorganic–physical, biochemical, and chemical technology. Indexers were expected to construct their own molecular and structural formulas and to name accurately all new substances encountered in their work. Both were time-consuming and critical steps. The naming process was one of the most complex technical operations in the entire editorial process.

Unified document analysis demanded an organization based only on subject content. The four general divisions of chemistry, previously listed, proved adequate. However, three very necessary steps would have to be taken if unified analysis was to be applied to all CA content. (1) A new management tool for measuring analyst production had to be developed. (2) The creation of structural representations of substances to be indexed had to be a separate task performed early in the analysis procedure. (3) The high-skilled task of creating names for new substances to be indexed had to be assigned to a relatively small group of chemists with considerable nomenclature experience.

PRODUCTIVITY MEASUREMENT

The need for a method to monitor productivity was obvious. However, these workers were graduate scientists, professionals. Several members of CAS management and ACS governance

said it could not be done with Ph.D.s. Extensive manual records on each indexer's productivity had been kept by individual supervisors since at least the mid 1960s. These records were compiled according to each supervisor's preference and could not be combined to yield a good overall measure of editorial productivity. Carl Ish, then assistant to the Editor, developed a new technique of productivity measurements leading to a uniform computerized system. In mid 1970 productivity reports on individuals were introduced in the biochemical, organic, and general subject indexing areas.[9] An individual's performance was compared to an average for scientists doing work of similar difficulty in language and subject content. Little resistance was encountered, and the reports were an improvement in individual performance evaluations. They were mandatory for following weekly and monthly production. This emphasis on overall editorial productivity began at CAS almost 12 years before the appointment of the 1982 Baker Task Force on CAS Productivity.[4] While these measurements were initiated with indexers, they were expanded gradually to all groups involved in unified document analysis. They were an excellent management tool during this transition period.

STRUCTURE DRAWING AND SUBSTANCE NAMING

Operation of the CAS Registry System[10,11] has been credited many times with allowing CAS to survive in the face of the more than one million substances that must be identified and indexed each year. The ability of Registry to retrieve names and molecular formulas for all substances indexed since 1965 frees the analyst to concentrate on only the new substances, about 350000 each year. This is a substantial savings in time and effort. However, in Unified Document Analysis substance identification had to take place at the earliest possible moment in the process. In the early 1970s the Registry step required more time than today. It was not possible to wait for the complete document analysis to input new substance identifications. Further, it was not possible to train all analysts to input structures or to generate names for new substances. Structuring conventions necessary for Registry input were too demanding, and the index nomenclature system was far too large and detailed.

Structures were drawn according to procedures revamped and expanded during 1971–1973. These included guidelines for drawing Markush structures, flow diagrams, polymers, and coordination compounds, plus three-dimensional inorganic and organic compounds. A team of lesser trained chemists was assembled to input structures already selected for indexing. The Registry System then retrieved index names for all known substances and returned only new substances for naming. The latter were handled by about 2 dozen highly trained inorganic and organic chemists skilled in CAS name-selection practices. This group was organized and led by Robert White, a former assistant to the Editor. It was assembled from some of the most experienced and best producers within the CAS staff. Its formation accelerated the naming process and eliminated time-consuming discussions that occurred when many indexers were struggling to find the proper name for a new substance.

A significant parallel decision, controversial at the time, was made by the Editor, and it also accelerated the substance naming process and aided Unified Document Analysis. Beginning with the *Ninth Collective Index* period (1972–1976), CAS made greater use of the fully systematic names available through the basic IUPAC nomenclature conventions.[12,13] Without this decision, today's volume of new substances could not be named and processed expeditiously. The decision also supplied more consistent names for similar structures and thus assisted in today's substructure and online searching, and in use of the printed index.

DOCUMENT ANALYST TRAINING

With a productivity measurement technique in place and the two technical steps of structure drawing and substance naming removed from the chores of regular analysts, CAS was ready to tackle the job of training abstractors and indexers to be complete document analysts. Many human problems occurred. Some abstractors did not have broad enough scientific backgrounds to do satisfactory indexing. They had to be given more training. Some indexers had difficulty in learning to abstract after spending years at only the one task. The work of all beginning document analysts had to be checked carefully to maintain the CAS content quality. Hence, the conversion of former abstractors and indexers to unified document analysts was a slow and painstaking program. The goal to eliminate the duplicate intellectual efforts was well worth the long hard work.

A production development plan for full conversion to unified document analysis was formulated in 1972 and implementation begun.[14] Small groups of abstractors or indexers in each subject area were trained, and their work was carefully monitored. The number in any group and the lengths of training and review depended on the languages of original documents and the difficulty of the subject content. Some learned in a few months; others required more than 1 year. The entire conversion took about 4 years. By 1976 100% of the CAS abstracts and index entries were produced by unified analysis.[15] In addition to the savings in time and costs, the unified process provided index entries in computer-readable format at the same time the abstracts appeared in print and on computer tape. This is a fundamental requirement for a timely online service and is a feature of CAS ONLINE.

In the mid 1970s the revised and simplified CAS Editorial Operations consisted of six major units: four chemical subject areas, Biochemistry, Chemical Technology, Organic Chemistry, and Physical-Inorganic-Analytical Chemistry; plus the two support functions, Chemical Name Generation and Registry and Index Services.

ONLINE EDITING

A concurrent computer development during this time period was online editing for almost all CAS input.[16] The availability of this technique to review simultaneously abstracts and corresponding index entries was another time and money saver. It has been estimated that it alone eliminated over 5.5 million pieces of paper formerly needed for editing abstracts and index entries in the old system. More importantly, this was another significant plus for the analyst. It was a great boost to the analyst's morale. No longer was it necessary to make editing changes in red pencil for unknown keyboarders to input, never really knowing whether the desired changes were ever made, or made correctly. It was never possible to recycle such changes for additional review. Online editing made it possible for the analyst to make the actual changes and see them on the screen. The job was finished, and the analyst took pride in being able to see the final product just as it was stored in the data base and as it would appear in print.

When it was suggested that analysts were going to edit online, some said Ph.D. scientists were not typists and would never willingly work many hours in front of a terminal. A quick study revealed that about 70% of the analysts could type, and many expressed joy in doing so. While they were not skilled at keyboard operation, they were familiar with the techniques and not opposed to trying to learn editing online. Also, it was discovered early that long hours at the terminal were neither necessary nor desirable. About a 2-h stint was adequate and caused little or no discomfort.

Another important perspective should be emphasized. The foresight of the ACS Board of Directors in the late 1950s and

early 1960s in providing an outstanding CAS building facility has been a great aid in attracting and retaining good personnel. Individual offices are provided for most scientists, maintained beautifully with individual controls for light and temperature. Over 50 acres of landscaped ground surround the current two buildings. There is ample, easily accessed parking. Such facilities have been a factor in holding editorial staff turnover to only a few percentage points annually. This is a great asset because initial CAS training for a new Ph.D. scientist requires 12-18 months, a significant investment.

A final perspective is obvious. The described major organizational changes have provided an efficient editorial operation through which CAS can continue to supply timely, complete, high-quality chemical information into the future. The excellently trained and dedicated staff is CAS's greatest asset. It assures the integrity of the data base for the years to come.

REFERENCES AND NOTES

(1) Wigington, R. L. "Computer Architecture for Editorial Processing Within an Integrated Publishing Organization". *J. Res. Commun. Stud.* **1979/1980**, *2*, 25-38.
(2) Seybold, J. W. "Data Base and Journal Publishing at the American Chemical Society". *Seybold Rep. Publ. Syst.* **1982**, *11* (24), 3-16.
(3) "CAS Today", 1980 ed.; Chemical Abstracts Service: Columbus, OH, 1980; p 11.
(4) Platau, G. O.; Metanomski, W. V. "Productivity and Its Measurement at Chemical Abstracts Service". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 8-11.
(5) "The First 75 Years of CAS". *CAS Rep.* **1982**, *No. 12*, 3-9.
(6) Baker, D. B.; Horiszny, J. W.; Metanomski, W. V. "History of Abstracting at Chemical Abstracts Service". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 193-201.
(7) Weil, B. H. "Standards for Writing Abstracts". *J. Am. Soc. Inf. Sci.* **1970**, *21*, 351-357.
(8) Platau, G. O. "Annual Report, Assignment and Abstracting Unit of Publications Division, 1970"; Chemical Abstracts Service: Columbus, OH, 1971.
(9) "Annual Report, CAS Editorial Operations Division, 1970"; Chemical Abstracts Service: Columbus, OH, 1971.
(10) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Service Registry System. I. General Design". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111-121.
(11) Weisgerber, D. W. "Finding Chemical Compounds by CAS Registry Numbers". *Ind. Res./Dev.* **1981**, *23* (5), 156-160.
(12) Rowlett, R. J., Jr.; Tate, F. A. "A Computer-Based System for Handling Chemical Nomenclature and Structural Representations". *J. Chem. Doc.* **1972**, *12*, 125-128.
(13) Donaldson, N.; Powell, W. H.; Rowlett, R. J., Jr.; White, R. W.; Yorka, K. V. "*Chemical Abstracts* Index Names for Chemical Substances in the Ninth Collective Period (1972-1976)". *J. Chem. Doc.* **1974**, *14*, 3-15.
(14) "Annual Report, CAS Editorial Processing Division, 1972"; Chemical Abstracts Service: Columbus, OH, 1973.
(15) "Annual Report, CAS Editorial Processing Division, 1976"; Chemical Abstracts Service: Columbus, OH, 1977.
(16) Weisgerber, D. W. "Applications of Technology to CAS Data Base Production". *Inf. Serv. Use* **1984**, *4*, 317-325.

ARTICLES

Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications

RAYMOND E. CARHART,*† DENNIS H. SMITH,† and R. VENKATARAGHAVAN

Lederle Laboratories, Pearl River, New York 10965

Received March 30, 1984

A simple type of substructure called an atom pair is defined in terms of the atomic environments of, and shortest path separations between, *all* pairs of atoms in the topological representation of a chemical structure. An algorithm is presented for computing atom pairs from such a representation. Two applications of atom pairs to structure-activity problems are described. In the first, a measure of similarity between compounds is defined, and the use of this measure in probing large databases of structures is discussed. In the second, a heuristic technique called trend vector analysis is described. The trend vector summarizes the correlation, within a set of structures, of the occurrence of atom pairs of different types with measured biological activity. These correlations can be used to estimate the biological activity of new compounds. A comparison of trend vector analysis with discriminant plane analysis is presented for one series of compounds.

INTRODUCTION

There currently are several computational methods for exploring relationships between chemical structures and measured properties, especially biological effects, of compounds. Blankley[1] has recently reviewed many of these methods. These techniques share a common problem; how does one express an irregular object like a chemical structure in a regular form that allows the quantitative comparing and contrasting of those structures? Most approaches rely upon *molecular descriptors*, which are numerical values representing selected features of the compounds. Each structure is represented as a list, or vector, of such numerical descriptors and thus may be thought of as a point in a high-dimensional space, with coordinates equal to (or related to) the corresponding descriptor values. With this change of representation, the problem of relating structure to biological activity becomes one of relating position (in the high-dimensional space) with activity. A variety of powerful mathematical techniques (e.g., pattern recognition,[2-4] multiple linear regression,[5] SIMCA[6]) is available for treating such problems.

Many different types of descriptors have been presented in the literature. Substituent parameters such as the Hammet σ[7] (electron-withdrawing power) and the Hansch π[8] (lipo-

†Present address: IntelliCorp, Menlo Park, CA 94025.

0095-2338/85/1625-0064$01.50/0