Table I. Comparative Performance of Various SDI Systems[3]

					This System	
	SDI-1	SDI-2	SDI-3	SDI-4	Full recall basis	Est. recall basis
U·D,[a] No.	6670	619,550	1,860,000	532,932	23,751	144,229
Hits, %	6.2	2.3	0.9	0.6	8.3	8.2
Trash, %	9.0	1.1	0.5	1.4	5.1	5.0
Miss, %	5.8	36.6	[b]	[b]	1.2	1.2±
Pass, %	79.0	60.0	98.6	98.0	85.4	85.6±
Selective reaction, %	15.0	3.4	1.4	2.0	13.4	13.2
Hit ratio (=P.F.)	0.41	0.68	0.66	0.30	0.62	0.62

[a] U·D is the product of the number of users and the number of documents processed. [b] Data not obtained; lumped with Pass.

was used in deriving the %-miss figure given for the SDI-1 System in Table I of this paper.[2]

In Table I, the data cited in Savage's paper representing four systems labeled SDI-1, SDI-2, SDI-3, and SDI-4 are displayed together with comparable data from our system on a "full-recall basis" (U·D = 23,751) and on an "estimated-recall basis" (U·D = 144,229). The full-recall basis includes only data obtained from the 55 cases described earlier; the estimated-recall basis includes all available data between 6608 and 6712. The estimated value of 1.2± for %-miss in the last column of Table I is inferred from the preceding column and is not based on feedback from randomly selected references. Since the percentages found for the "hits" and the "trash" in these two columns are almost identical, we infer that the percentage of the missed articles will also be substantially the same.

LITERATURE CITED

(1) Cleverdon, C., *Nat. Acad. Sci.* **1**, 687 (1959).

(2) Hensley, C. B., *et'.al.*, *IRE, Trans. Eng. Management* **EM-9**, 55 (1962).

(3) Savage, T. R., *Am. Doc.* **18**, 242 (1967).

Data Retrieval for a Large Organic Synthesis Project

H. J. ACKERMANN, E. H. KOBER, R. E. McARTHUR,
R. E. MAIZELL, and D. A. SHERMER
Olin Mathieson Chemical Corp., 275 Winchester Ave., New Haven, Conn. 06504

Received November 26, 1968

A data processing system for storing and retrieving experimental data from an organic synthesis project is described. The data keypunched from a specially designed notebook are processed in an IBM 1800 computer to provide multifaceted printouts for the project scientists and research management.

The laboratory project referred to in this paper involves the successful development of a new organic synthesis. Some 80 chemists and engineers have participated in various phases of the project.

The project moved along at a good pace during the first few months, and as data began to accumulate, the need for an automated system to store and access the data became clear.

By the time the data processing system became operational, the project had progressed past the initial bench stage and into the development phase. The benefits of the system were almost immediately apparent. Project scientists were relieved of the time-consuming task of tabulating data for weekly oral review meetings and for periodic progress reports. Research management was able to study advances in the progress of the project through rapid access to laboratory data. Consultants were able to supply valuable suggestions with a minimum expenditure of time. Listings of starting materials, catalysts, solvents, and products were updated by computer printout. Patent attorneys were able to utilize the file in search of information and data for patent applications.

METHOD

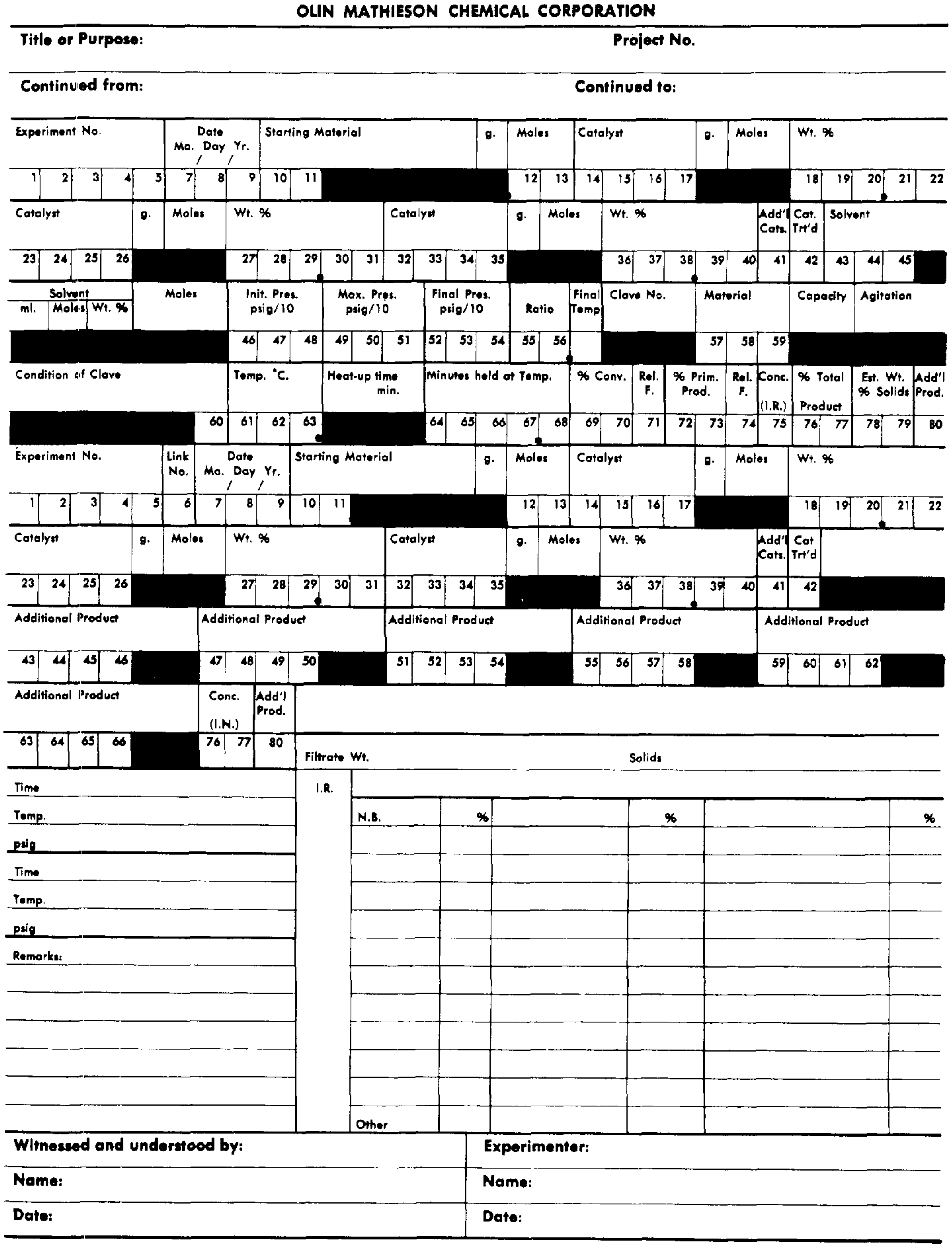
As our first step in developing this data processing system, we conducted group and individual in-depth interviews with the project scientists to pinpoint the kinds and relative importance of data being generated. We were then able to categorize the kinds of information most frequently used as follows: starting materials, catalysts, solvents, reactors, temperatures, time, pressures, product(s) yield, conversion, and by-product(s). Schematically, the data to be filed were generated from the notebook records of the reaction:

$$\boxed{\text{Stg. Mtl. A}} + \boxed{\text{Stg. Mtl. B}} \xrightarrow[\text{Temp. + Press. + Time}]{\text{Cat(s). + Solv(s).}} \boxed{\text{Prod. A}} + \boxed{\text{Prod. B}} + \boxed{\text{Prod. C}}$$

For compact storage of input, it was decided to encode certain kinds of information and to maintain 'concordance tables to identify coded data—e.g., catalysts.

We incorporated into a specially designed notebook (Figure 1) the basic features of the data processing input form to identify easily the areas for keypunching. Space was provided for reference to conventional research notebooks, which continued to be used for detailed records of preparations, compatibility tests, other experiments, and ideas.

Initially, the system took the form of a relatively simple card-oriented data system—and such was the intent. Subsequently, the data storage and retrieval system advanced from an IBM 870 Document Writer, to accounting machine equipment, to an IBM 1130, and, eventually, to an 1800 computer.



OLIN MATHIESON CHEMICAL CORPORATION

Title or Purpose: **Project No.**

Continued from: **Continued to:**

Experiment No.	Date Mo. Day Yr. / /	Starting Material	g.	Moles	Catalyst	g.	Moles	Wt. %
1 2 3 4 5	7 8 9	10 11		12 13	14 15 16 17			18 19 20 21 22

Catalyst	g.	Moles	Wt. %	Catalyst	g.	Moles	Wt. %	Add'l Cats.	Cat. Trt'd	Solvent
23 24 25 26			27 28 29 30 31	32 33 34 35			36 37 38 39 40	41	42	43 44 45

Solvent ml. Moles Wt. %	Moles	Init. Pres. psig/10	Max. Pres. psig/10	Final Pres. psig/10	Ratio	Final Temp	Clave No.	Material	Capacity	Agitation
		46 47 48	49 50 51	52 53 54	55 56			57 58 59		

Condition of Clave	Temp. °C.	Heat-up time min.	Minutes held at Temp.	% Conv.	Rel. F.	% Prim. Prod.	Rel. F.	Conc. (I.R.)	% Total Product	Est. Wt. % Solids	Add'l Prod.
60	61 62 63		64 65 66 67 68	69 70	71	72 73	74	75	76 77	78 79	80

Experiment No.	Link No.	Date Mo. Day Yr. / /	Starting Material	g.	Moles	Catalyst	g.	Moles	Wt. %
1 2 3 4 5	6	7 8 9	10 11		12 13	14 15 16 17			18 19 20 21 22

Catalyst	g.	Moles	Wt. %	Catalyst	g.	Moles	Wt. %	Add'l Cats.	Cat Trt'd
23 24 25 26			27 28 29 30 31	32 33 34 35			36 37 38 39 40	41	42

Additional Product	Additional Product	Additional Product	Additional Product	Additional Product
43 44 45 46	47 48 49 50	51 52 53 54	55 56 57 58	59 60 61 62

Additional Product	Conc. (I.N.)	Add'l Prod.
63 64 65 66	76 77	80

Filtrate Wt. Solids

I.R.

Time							
Temp.		N.B.	%		%		%
psig							
Time							
Temp.							
psig							
Remarks:							
		Other					

Witnessed and understood by:	**Experimenter:**
Name:	**Name:**
Date:	**Date:**

Figure 1. Research notebook–input form for laboratory data storage

The data were to be searchable through multiple routes. These included: mechanical sorting of records; printed listings of the data; collections of previous inquiries and responses; listings of presorted subfiles; printed concordance tables, and indexes for code numbers and names of compounds; and magnetic disk files. Original notebooks and carbon copies of notebook pages bound in experiment number order provide additional back-up and assured quick handling of inquiries from one or more of the various record files. We have found the availability of these multiple approaches useful in answering relatively simple questions which do not require a computer, or if computer time is not immediately available for any reason.

Initially, output was a tabulation distributed to all project investigators, up-dating them on the results of work recently completed. These lists were accompanied by additions to the concordance tables.

Searches of the data files were relatively simple at first. Inquiries were oriented primarily toward fact-finding—e.g., code numbers for compounds, conditions, or results. The records retrieved could be printed out in various ways, as desired by the requester: the total number of hits, experiment numbers only, limited experimental data, total input, and other options.

The total file of data from over 10,000 experiments is being added to by the regular keypunching of input from new laboratory and pilot plant records. In addition to programs written for the storage of and quick access to original data, calculation and statistical evaluation routines have been developed. The programming language has been FORTRAN IV.

Inquiries are usually processed by an information scientist who acts as a coordinator for the data system or by the engineers who wrote many of the search programs.

During the development of the system, the requirements of project scientists and management became more complex. Their involvment in the design and use of the data base stimulated them to make suggestions for the necessary improvements.

DISK SEARCHING

The major inquiries indicated a subfile approach. Thus, data from some 4000 experiments of the greatest interest were stored on an IBM-2315 magnetic disk.

Similarly, programs were written to search the files for specific data, to manipulate them, sort them, perform the particular calculations, evaluate the results, and display them in printed format.

Here the use of FORTRAN permitted easy modification to execute specific search requirements. Thus, it was not necessary to design and program a complex routine to anticipate the many possible varieties of search.

A sort routine for the 1800 is used to manipulate the data on 18 specific keys or reference points. The printout can be duplicated and distributed and is used primarily for the rapid comparison of conversion and yield data for related experiments.

The parameters used in searching involve various forms of logic, including conjunction, disjunction, and negation. There is an added feature of specifying ranges for numerical values. Useful results are obtained by screening the data against certain parameters and listing the data in selected groupings—e.g., yield or conversion values falling within certain ranges.

Accompanying this subfile is an updated concordance list of catalyst descriptors in code number sequence and another list sorted into alphanumeric sequence on the first part of the catalyst descriptors.

Since many of the catalysts were mixtures, and some were impregnated on carriers, the components were not readily obtained through either of the concordance tables. A program was written which effectively selected particular terms and fragments within a catalyst descriptor, by use of a nonprintable character keypunched as a prefix to the entity.

Thus, a catalyst descriptor such as: 5% iridium bromide and 2% indium fluoride on alumina, would be keypunched as:

```
1234  5% # IRIDIUM # BROMIDE + 2% # INDIUM

        # FLUORIDE ON # ALUMINA
```

the # character being used as the indicator for an indexable term or fragment.

These terms or fragments, together with a 54-character portion (single line) of the descriptor, were sorted into alphanumeric sequence and printed in the format of a typical KWIC (key-word-in-context) index. The printed index entries for this catalyst descriptor would be:

KEYWORD INDEX

```
1234   MIDE +2% INDIUM FLUORIDE ON    ALUMINA*      5% IRIDIUM BRO
1234   DE ON ALUMINA*     5% IRIDIUM    BROMIDE + 2% INDIUM FLUORI
1234   IRIDIUM BROMIDE + 2% INDIUM    FLUORIDE ON ALUMINA*      5%
1234   *     5% IRIDIUM BROMIDE + 2%    INDIUM FLUORIDE ON ALUMINA
1234   M FLUORIDE ON ALUMINA*     5%    IRIDIUM BROMIDE + 2% INDIU
```

PERSONNEL, TIME, AND OTHER REQUIREMENTS

To get the system where it is now has required the part-time participation of five technical people, a programmer, and a keypuncher. Over a time span of about two and one half years, we estimate that 175 hours were spent in programming, 650 hours in keypunching, 450 hours in coding, 120 hours in processing on the 1130 and 1800 computers, 300 hours in clerical operations, and 100 hours in supervisory functions, for a total of about 1800 man-hours.

The subfile structure, search strategy, calculation, and display have been, and are still, changing with specific needs of the project.

It is difficult to estimate the kind and minimum size of a research project for which one should consider use of the computer for storage and retrieval of experimental data. There are some minimum requirements. In general, the data need to be quantative and sufficiently voluminous to justify automation. If researchers find themselves spending unreasonable amounts of time in searching for data, then automation must be seriously considered.

As indicated, in our case, we found that a subfile of data from 4000 experiments was more than enough justification for automation.

Preferably, a computer should be available already within the research department on site. Ideally, time on the computer and programming should be either on a no-charge basis or carry minimal rates. In addition, there must be available an experienced programmer to act as an advisor or consultant to the technical people who write their own programs. Also essential are a keypuncher and a computer operator within the research department.

Most essential is the backing of research management to provide the funds and encouragement to use the computer, and the willingness of R & D personnel from all phases of the project to get fully involved.

PLANS

The development of the system may require modification of the original input format to accomodate entry of quantitative analytical data or other data from pre-pilot plant or pilot plant operations. More satisfactory input of data may be obtained in the future by use of mark-sensed cards. Also, it may be desirable to consider those features of the IBM 1800 which make possible on-line input if desired. For rapid access to original notebook data, it may be desirable to microfilm these notebooks in a form suitable for automated cartridge readers.

Introduction to Symposium on Training Chemists in the Use of Chemical Literature*

GERALD JAHODA
The Florida State University, Tallahassee, Fla. 32306

Received December 27, 1968

Chemical literature courses can be deadly dull for both the student and the teacher. They are likely to be deadly dull if they consist of a seemingly endless recitation of book and journal titles, with indication of editorial scope, name changes, organization of material, frequency of issue, indexes, etc. Reasons why some courses are still being taught in this way are discussed at this symposium. Suffice it to say at this point that there are better ways. The objective of the symposium is to point out that there are new information systems and information services of interest to the chemist and that there are new ways of acquainting students with these tools.

It seems to me that one of the functions of a course in chemical literature is to instill good information gathering and use habits in our students. There is probably little disagreement with this opinion. The trouble with it is that it begs the question unless we define what we mean by good information gathering and use habits. And this is very difficult to do. We have reasons to believe that our own information habits, the ones we have been taught or have taught ourselves over the years, may no longer be equal to the task. New techniques that are being proposed and tried out appear promising. We have not as yet used these new and largely machine-based methods long enough to be more sanguine about them. They have to be tested and then further refined, just as new methods that are introduced in the laboratory. One such test of machine-searchable indexes prepared by Chemical Abstracts Service is now being conducted in this country and in Great Britain. In Great Britain, chemistry students in their final year of the Ph.D. program are provided with an individualized current awareness service. Keywords that characterize the student's subject interests are matched by a computer against keywords that characterize the contents of newly received documents. Either *Chemical Titles* or *Chemical Biological Activities* is searched, and potentially relevant documents are directed to the student's attention.[1] This current awareness service accomplishes two useful things. It introduces the student to a new information service and it provides a test bed for the service. I hope these tests will be successful, and that as a consequence machine-searchable current awareness services can be offered to a larger number of students.

What conclusions can we draw from the Symposium papers? Herner's approach in his one-day course for working scientists and engineers appears equally valid for chemistry students. The topic is presented by Herner on a problem-solving rather than on a "duty" basis. Techniques for the organization of information, for example, are not presented as something that the student needs to memorize as an academic exercise, but as techniques intended to help him in locating documents needed in his work and in organizing his own document collection. The Martin and Robison survey tells us that a sizeable percentage of the reporting schools have dropped their formal chemical literature courses between 1960 and 1967. The apparent lack of enthusiasm on the part of both faculty and students for such courses is not likely to reverse this downward trend during the next few years. Hopefully, these courses will be replaced by more effective techniques for teaching students what they need to know as users (and eventually producers) of chemical literature. Three techniques which are new for teaching this skill are discussed at the Symposium. They are a "packaged" audio-visual course, machine-searched indexes as teaching aids, and video-taped lectures by experts in the field. The judicious use of these and other new techniques as exemplified by computer-aided instruction may well provide the short term solution for this problem. For the long term solution, more needs to be known about what should be taught.

LITERATURE CITED

(1) Somerfield, G. A., "Students' Chemical Information Project," *Chem. Britain* 4 (2), 71-3 (February 1968).

*Presented before the Symposium on Training Chemists in the Use of the Chemical Literature, Division of Chemical Literature, 156th Meeting, ACS, Atlantic City, N. J., Sept. 8-13, 1968.