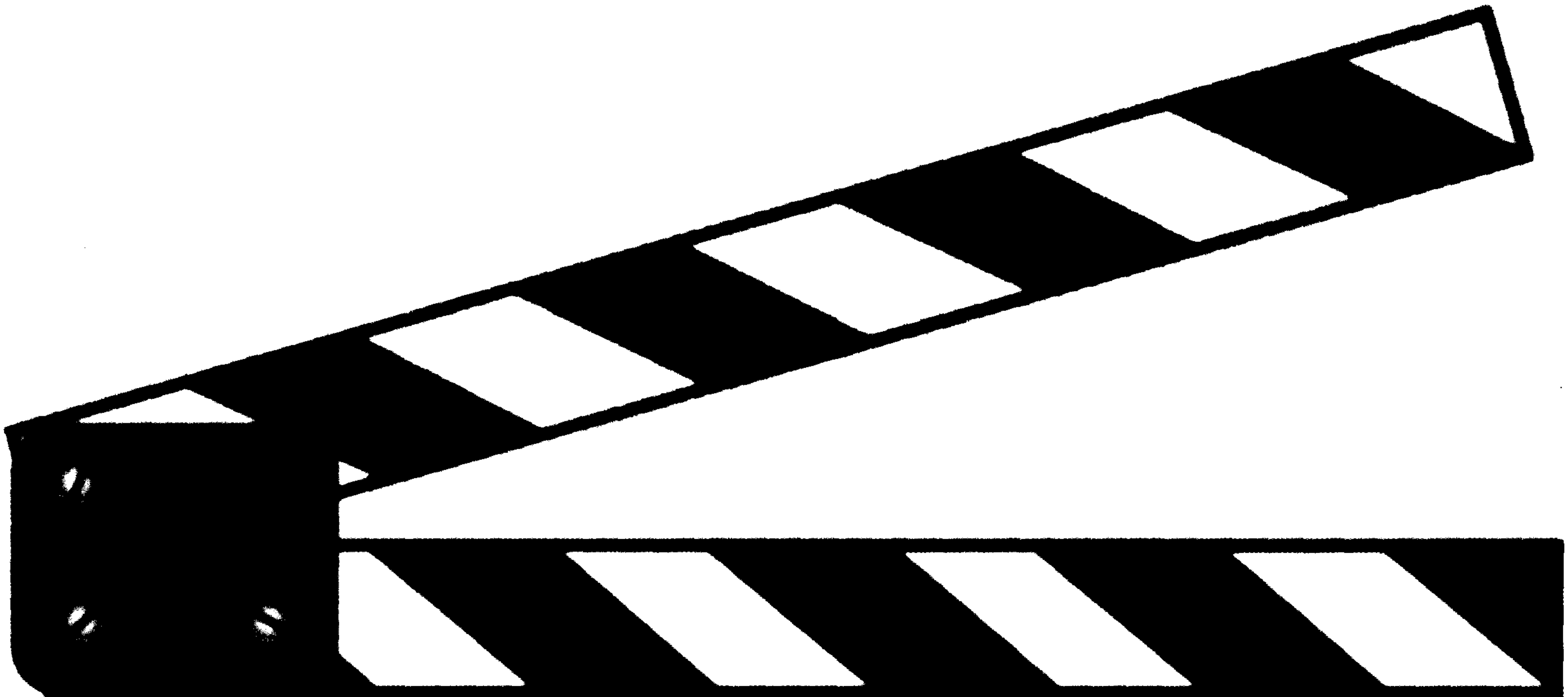
8-1969

Data retrieval in mass spectrometry by an optical coincidence system

Cynthia Helmintoller O'Donohue



Project Name: ODonohue_Cynthia_1967

Date: 7/10	Patron: DTP	Specialist: Jackie

Project Description:
MASTER THESES

Hardware Specs:

DATA RETRIEVAL
IN MASS SPECTROMETRY
BY AN
OPTICAL COINCIDENCE SYSTEM

BY

CYNTHIA HELMINTOLLER O'DONOHUE

A THESIS
SUBMITTED TO THE GRADUATE FACULTY
OF THE UNIVERSITY OF RICHMOND
IN CANDIDACY
FOR THE DEGREE OF
MASTER OF SCIENCE IN CHEMISTRY

APPROVED:
W. Allan Powell
J. Stanton Pierce
F.B. Leftwich
Peter B. Bahler
Richard A. Mateer

AUGUST, 1967

DEDICATION

Dedicated to my husband whose patience, understanding, and encouragement has made this work possible.

ACKNOWLEDGMENT

On completion of this project I wish to thank Dr. W. Allan Powell. His interest and guidance in this problem have been without peer.

I am deeply grateful to Dr. M. Arthur Manzelli for his many valuable suggestions and gracious assistance in this project and to Mr. W. F. Kuhn for his kind assistance in explaining the many problems that confront his group in mass spectra retrieval.

I would like also to express my gratitude to Philip Morris Incorporated for allowing the use of their facilities to carry out this project, and to the managers and director of Philip Morris Research Center for their encouragement.

Finally, I wish to express my appreciation to Mrs. Margaret Opocensky who so kindly consented to type this thesis.

TABLE OF CONTENTS

Page

INTRODUCTION

The object of this work was to develop an information storage and retrieval system for use in mass spectroscopy. The information to be stored was the compound's name, the mass spectrum, and distinctive physical and chemical properties. It was desirable that the system be one that could be used in the mass spectral laboratory.

At the time this project was initiated, no system for mass spectroscopy had been developed with the characteristics of simplicity, rapid retrieval, and practicality of use in the laboratory without elaborate and space consuming equipment. Thus, a system was needed to aid the laboratory personnel in compound identification and prediction of empirical formula and structure.

HISTORICAL

Analytical laboratories from earliest times have had the problem of efficiently utilizing spectral data to identify an unknown compound analyzed by mass spectrometry. In one of the first methods devised, the known compounds were run with the unknown to determine if the data obtained were identical. After the experimental conditions had been standardized, the data obtained from testing known compounds were recorded and stored in a file system. As the number of known compounds increased, the problem arose of how to employ the stored data, as it was no longer practical to compare the data for each known compound to that of the unknowns.

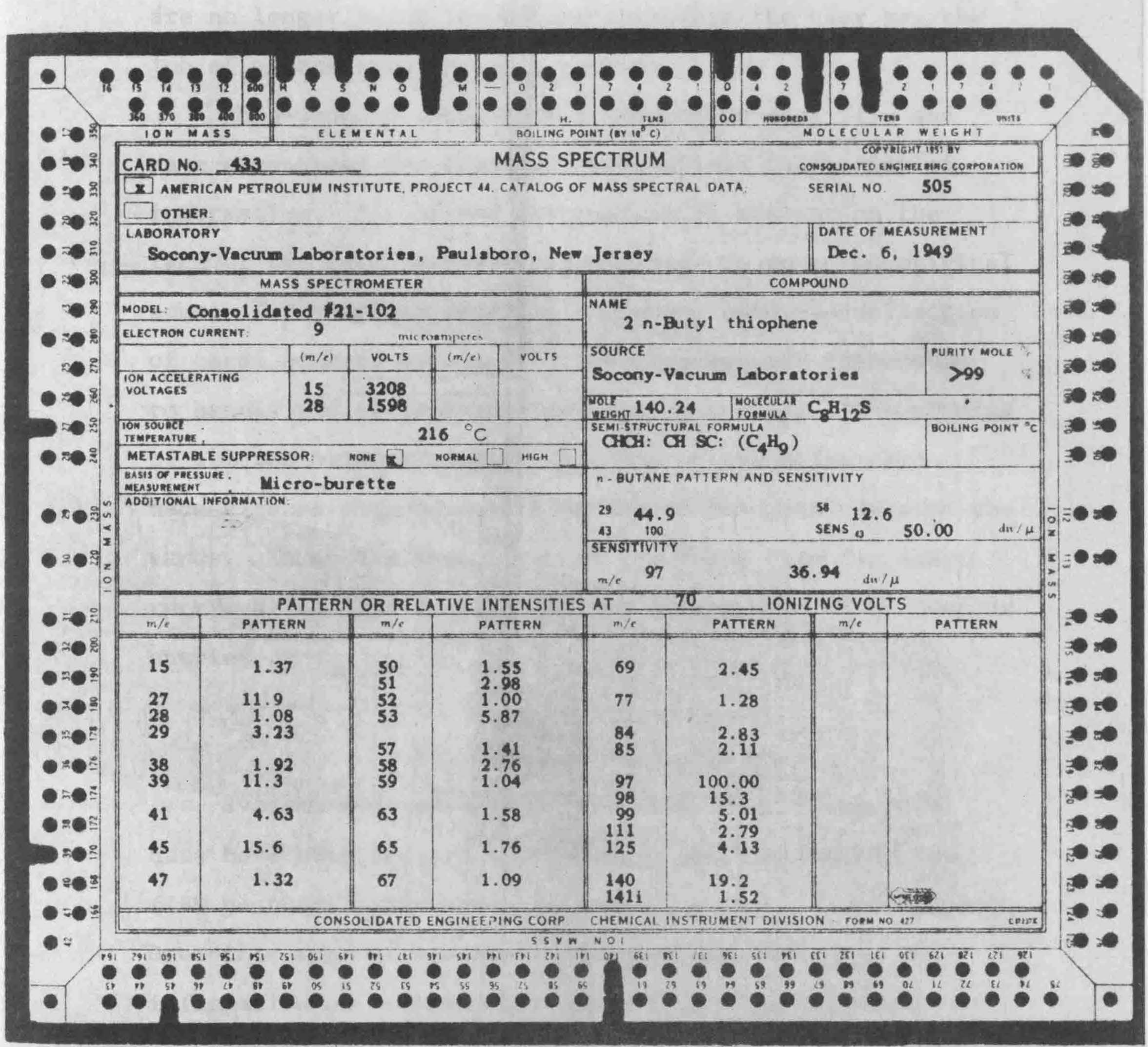
An early method employing mass spectra comparison for compound identification called for the determination of the parent peak, and thus the molecular weight. The available files were then searched by molecular weight; i.e., all compounds with that particular molecular weight were examined to determine if their spectra were similar to that of the unknown. This system was simple, straightforward and not too difficult while the spectral files were small; however, there has been a tremendous growth in published spectra from the academic and industrial research laboratories as the workers determine the mass spectra of their compounds of interest. At the present time there are available the American Petroleum Institute certified spectra, Dow Chemical uncertified spectra, Philip Morris unpublished spectra, and in addition various other compilations of spectra. This involves the examination of 30 or more compounds in at least three different storage sources to track down an unknown. A recent compilation (1) attempted to improve the spectral retrieval by listing the compounds not only by molecular weight but also in order of increasing peak mass for each of the first three strongest peak masses. However, the worker is still required to examine quite a few spectra within one of these regions to find a possible match for the unknown.

To circumvent this tiresome and time-consuming chore of manually searching the mass spectral files, various systems have been devised to limit the number of compounds necessary to be examined for each unknown.

I. Edge-Punched Cards

One of the earliest and simplest systems of storage and retrieval for mass spectral data was developed by Zemany in 1950(2). The compounds were coded on edge-punched cards by molecular weight, elements present, and selected prominent peaks. On one side of the card the hundreds, tens, and units holes were reserved for the molecular weight. An additional five or six holes were used to designate the elements present other than C or H. Through the use of deep punches 208 spaces were made available for direct indexing of the peaks. Only the five or six most prominent peaks were coded. For mass 200 and up each space was assigned ten mass units. Each card for a given compound had a copy of the spectrum pasted directly on the card or else the record of the peak heights was transcribed onto the card.

The Consolidated Engineering Corporation, Inc., published a set of edge-punched (McBee) cards containing mass spectra (Figure I). This file provided four fields of selection. One numerical field was used for molecular weight and another for the selection of boiling points. The third field was used for indicating the presence of various elements; i.e., H, X(halogen), S, N, O, C, and M(miscellaneous). The remaining holes on this card were devoted to peak mass. Six to ten of the largest or most distinctive peaks were punched in this field. These cards


ION MASS
ELEMENTAL
BOILING POINT (BY 10° C)
MOLECULAR WEIGHT
HUNDREDS
TENS
UNITS

CARD No. 433 **MASS SPECTRUM**

COPYRIGHT 1951 BY CONSOLIDATED ENGINEERING CORPORATION

☒ AMERICAN PETROLEUM INSTITUTE, PROJECT 44. CATALOG OF MASS SPECTRAL DATA. SERIAL NO. 505

□ OTHER

LABORATORY: Socony-Vacuum Laboratories, Paulsboro, New Jersey

DATE OF MEASUREMENT: Dec. 6, 1949

MASS SPECTROMETER

MODEL: Consolidated #21-102

ELECTRON CURRENT: 9 microamperes

	(m/e)	VOLTS	(m/e)	VOLTS
ION ACCELERATING VOLTAGES	15 28	3208 1598		

ION SOURCE TEMPERATURE: 216 °C

METASTABLE SUPPRESSOR: NONE ☒ NORMAL □ HIGH □

BASIS OF PRESSURE MEASUREMENT: Micro-burette

ADDITIONAL INFORMATION:

COMPOUND

NAME: 2 n-Butyl thiophene

SOURCE: Socony-Vacuum Laboratories

PURITY MOLE %: >99 %

MOLE WEIGHT: 140.24

MOLECULAR FORMULA: $C_8H_{12}S$

SEMI-STRUCTURAL FORMULA: CHCH: CH SC: (C_4H_9)

BOILING POINT °C:

n - BUTANE PATTERN AND SENSITIVITY

29 44.9 58 12.6

43 100 SENS 43 50.00 dv/μ

SENSITIVITY

m/e 97 36.94 dv/μ

PATTERN OR RELATIVE INTENSITIES AT 70 IONIZING VOLTS

m/e	PATTERN	m/e	PATTERN	m/e	PATTERN	m/e	PATTERN
15	1.37	50	1.55	69	2.45		
		51	2.98				
27	11.9	52	1.00	77	1.28		
28	1.08	53	5.87				
29	3.23			84	2.83		
		57	1.41	85	2.11		
38	1.92	58	2.76				
39	11.3	59	1.04	97	100.00		
				98	15.3		
41	4.63	63	1.58	99	5.01		
				111	2.79		
45	15.6	65	1.76	125	4.13		
47	1.32	67	1.09	140	19.2		
				141i	1.52		

CONSOLIDATED ENGINEEPING CORP CHEMICAL INSTRUMENT DIVISION FORM NO 427

Figure I

A McBee Mass Spectrum Card

are no longer being issued and therefore the user has the job of maintaining the file current.

Edge-punched cards form a convenient desk file and have advantages for the small specialized collection of information. All of the information is present on the face of the card, eliminating the need to go to an original source for complete details. However, once the collection of cards numbers over 1,000, they become very cumbersome to handle and present difficulties in sorting. The placing of all the original source material on the McBee card necessitates someone compiling and printing the data on the cards. Thus, the usefulness of the McBee card for easy retrieval is lost in an expanding data system with numerous entries.

II. Hollerith-Type Cards

Systems for retrieval of several types of chemical data have been devised which employ only Hollerith-type (IBM punched) cards and a sorter (3, 4, 5). Eastman Kodak used this type of system for storage and retrieval of information on chemical compounds (3). The information entered on the cards had to be in a format acceptable to a computer. It was not physically possible to enter all the desired information on one card. The system finally involved

17 cards that could be used to completely describe an organic compound according to structure and functional groups. It was felt that by the time the card file became quantitatively unwieldy the information would be on magnetic tape for computer use.

Kuentzel (4) used Hollerith-type cards for infrared absorption and chemical structure data. Each card contained the details of absorption, chemical structure, physical properties, and a reference to the location of the original data for a given compound. To enter this information on a single card a code was employed in which each punch within a column had a designated meaning.

In developing a Hollerith-type card system for mass spectral data, McLafferty and Gohlke (5) concluded that the sorting steps present in most other systems, both edge-punched and Hollerith-type, required too much time. They eliminated the sorting step by using IBM punched-cards to prepare lists for each mass number (m/e) of all compounds having a significant peak at that mass number. To prepare the listing it was necessary to have 22 duplicate sets made of the master deck. The compounds had their ten highest peaks and five additional "peculiar" peaks listed. To identify a compound it was necessary to examine all the compounds listed under the particular m/e chosen.

The American Society for Testing and Materials Committee E-14 set forth its own method which used features from proposed systems (5) and systems already in industrial use. Only 51 of the 80 columns in each card were used so that the individual laboratories could enter any additional information that was essential to its operation. The cards contained the six strongest peaks, molecular weight, serial number, source code, and the chemical structure code (identical with the chemical structure code used in the commercially available Wyandotte-ASTM Infrared punch card system). In addition to the punched cards the American Society for Testing and Materials offered a book index of mass spectral data containing five separate indexes based on molecular weight, most intense peak, second, third, and fourth most intense peaks. The spectral data in this book consisted of the four strongest peaks and their relative intensities with a coded number and name of the compound.

III. Computer Systems

In 1965 Cook et al. (6) reported on a computer control and data processing system. The mass spectrometer was connected to an on-line digital computer which was programmed to control the instrument, accumulate the data, compute and evaluate the mass spectral data with a print out. This system was capable of detecting and storing peak heights at a rate of 200 peaks/second.

With the aid of a digital computer Abrahamsson et al. (7) stored all m/e's and intensity values for each spectrum as well as the name of the compounds, the molecular weight, the source and serial number of the spectrum. By use of an appropriate program the information in the file could be listed or sorted by various items such as molecular weight and chemical name. The computer is so programmed that using the key of the five strongest lines in the spectrum a search of 10,000 spectra takes about five minutes. It is planned to add to the files the elemental composition and a type classification.

Hamming et al. (8) used a computer program in an analytical research mass spectrometry laboratory for several different purposes. By the addition of subroutines to the main program, correlation studies that involved searching complete mass spectra could be performed. Smith et al. (9) have devised several methods of matching spectra and have written the necessary computer programs. The relative efficiencies of these have been compared with one another and with that of Abrahamsson (7). Despite spectral variations due to differences in instruments or sampling, efficient retrieval has been achieved.

Computer forms of storage and retrieval of chemical data require a systematic form of nomenclature. Chemical nomenclature is not suitable as input data because of the many permissible variants of a name; e.g., paranitraniline, p-nitroaniline, and 4-nitroaminobenzene are three correct names for the same compound. Thus a notation for the chemical structure must be used. A notation is simply a means of delineating the geometrical structure of an organic molecule in a linear algebraic form.

Many systems have been devised for specific needs but only two systems will be mentioned here. Silk (10) modified existing systems into a new notation that required only one alphabet and one set of numerals so that the standard typewriter and computer print-out system had an adequate range of symbols. The basis of the system was the use of the first four letters of the alphabet to cipher the fundamental chains and ring systems. Additional symbols were employed to denote a variety of functional groups. Numerals were used to indicate the number of substituents.

A simple numerical code was devised for an IBM punched-card system (11). Information on organic chemicals was stored according to an index number, the BATCH Number, which indicated the structural and atomic features of the chemical. This enabled the computer to be programmed to print a compilation of an easy-to-use structural directory of compounds.

The BATCH Number is assigned according to the following scheme: the B digit represents the nature and number of rings present; the A digit represents the nature and number of atoms present other than carbon, hydrogen, and oxygen; the T digit represents the number of atoms present other than carbon and hydrogen; the C digit represents the number of carbon atoms in the empirical formula; and the H digit represents the number of hydrogen atoms in the empirical formula. Table I gives a few examples of compounds and their BATCH Number.

Table I

BATCH Number and the Corresponding Chemical Compound

BATCH NUMBER	COMPOUND
36144	Thiophene
36148	Tetrahydrothiophene
36156	2-Methylthiophene
36168	2-Ethylthiophene
36168	2,5-Dimethylthiophene
36171	2-Propylthiophene
36243	2-Iodothiophene
36342	2,5-Dibromothiophene

IV. Optical Coincidence Systems

An optical coincidence system is "an information-retrieval system that uses peek-a-boo cards; i.e., cards into which small holes are drilled at the intersections of coordinates (column and row designations) to represent document numbers" (12). One optical coincidence card is assigned to each term. One position on the term card is dedicated to each different numbered compound. Each chemical compound is assigned a number and is given the same dedicated position number on all cards. A hole drilled at that position in a card means the compound has the term assigned to the card. No hole at that position in a card means that the compound does not have the term. Cards representing the terms that describe the question are superimposed over a light source. Light shining through the holes in these cards represents chemical compounds with those terms. The positions of the holes are the numbers of the item.

Schlichter and Wallace (13) developed a system for use in searching infrared spectra. Their system evolved from the Wyandotte-ASTM system which is on IBM punched-cards. The files after conversion required 250 optical coincidence cards for 10,000 compounds as opposed to 10,000 IBM cards, one for each individual compound. The wavelength coding was devised to indicate the types of compounds that would produce a specific type of spectrum rather than a code to separate one compound from another.

Optical coincidence has also been used to identify compounds by their X-ray diffraction powder data (14). Compounds are indexed by the elements present in the material and by their five strongest lines; i.e., the "d" value which is the value of the interplanar spacing of the crystal planes of the substances given in angstroms. Through the use of a negative deck (the absence of elements and the absence of certain lines), it is possible to identify mixtures as well as single compounds.

EVALUATION OF PRIMARY SYSTEMS

I. Operative Systems

Systems of three general types were studied to determine their suitability for further development. These systems were 1) edge-punched cards, 2) computers, and 3) optical coincidence cards. The first two systems were eliminated because of the considerations listed below.

A. Edge-Punched Cards

One of the simplest retrieval systems is a file of edge-punched cards. These cards are used in the conventional data processing mode; i.e., there is one card for each reference. The equipment needed is simple and inexpensive. Besides the cards only a cardholder, a hand punch and a sorting needle are required. The cards form a convenient deck file which is sorted by hand.

The cards can be obtained in sizes ranging from $1\frac{1}{2}$" x $2\frac{1}{2}$" to 8" x $10\frac{1}{2}$" depending upon the needs of the user. Direct, numerical, alphabetical, or any combination of these codes can be used for indexing the cards. In direct coding each hole is assigned a particular meaning, which can be a number, a keyword, or a phrase. The holes are marked by appropriate symbols indicating the code phrases to the user. The entry points of an edge-punched card system are limited to those holes which can be punched around the edges of the card. The perimeter of the card can contain one or more rows of holes. Alphabetical indexing requires a five hole field, numerical indexing requires a four hole field, and direct coding requires a single hole field.

The body of edge-punched cards is left blank so that the user can enter any information that is desirable and necessary in his particular application. This entry on the card can take the form of typing, writing, printing, or pasting a copy of the original data on the card's body. Thus, after sorting, the remaining cards will contain all the information needed by the user, making it unnecessary for him to go to the original source for verification of the reference.

An edge-punched (McBee Keysort) card system published by the Consolidated Engineering Corporation, Inc., for the American Society for Testing and Materials Committee E-14 is already available for mass spectroscopists. The data included in this file are the American Petroleum Institute

Project 44 Mass Spectral Data and spectra from a Consolidated Model 21-103 Mass Spectrometer. All peaks with a relative intensity of 0.46 or greater are included in the mass spectrum, which is printed on the body of the card.

These McBee cards (Figure I) provide four major fields of selection: molecular weight, boiling point, elements present, and peak mass. The molecular weight field (upper right hand corner) utilizes the following units: hundreds, tens, and 1-9, as well as a zero field. When the number to be coded is 1, 2, 4, or 7, the card is punched deep (two or more holes perpendicular to the edge of the card are punched, one hole wide and two holes deep); and when the number requires two punches; e.g., 3, 5, 6, 8, or 9, two shallow punches are used. Each molecular weight is considered to have three digits and if there is one zero present in the group of three, then the zero hole is punched shallow. Two zeros require a deep punch. In Figure I the molecular weight is 140, thus requiring a deep punch for the one in the hundreds field, a deep punch for the four in the tens field and a shallow punch in the zero field.

A similar numerical coding at the top of the card is used for entering the boiling point. The selection of boiling points is by 10°C increments. The negative sign field is punched shallow for minus boiling points and left unpunched for positive boiling points.

The top of the card also contains a field for punching the elemental analysis of the compound. The number of atoms of each element in the compound determines whether the punch is shallow or deep. In this case the H has a shallow punch indicating 1-12 hydrogen atoms, the S has a shallow punch indicating 1-2 sulfur atoms, and the C has a deep punch indicating 5 or more carbon atoms.

The remaining holes in the card are used to denote ion mass. There is a shallow hole for each mass unit from 12 to 100. One deep hole is allotted as follows: one mass unit from 101 to 150; two mass units from 151 to 180; and ten mass units from 181 to 380. Usually only three to eight peaks are punched per spectrum. In Figure I there are six peaks indicated: 27, 39, 45, 97, 98, and 140 m/e's. The person coding the cards determines which peaks are to be selected for punching into the ion mass field.

A serious drawback to this type of system is the loss of sorting ease as the files grow. Once the card holdings increase beyond 1,000, the time required to sort becomes disproportionate to the quantity of data retrieved. When the files become bulky, the user tends to ignore them and will use other means of identification.

Consolidated Engineering Corporation, Inc., has stopped updating the McBee files for industry and therefore the responsibility for keeping the file current falls on the user. The time required for data input for a McBee card is very high. It involves compiling the data, entering it on the card body, checking for accuracy, and coding for the punching. The file's usefulness has become so limited that it does not justify the time and effort of a trained person to keep it current.

The elemental analysis coding does not answer the specific needs of some spectral groups. When a compound is received for mass spectra analysis, it has already been subjected to other means of chemical identification, and the compound's chemical classification is known. There is no way to distinguish between different classes of compounds from the input on the McBee cards; e.g., in the case of oxygen-containing compounds it is not possible to tell if they are alcohols, acids, ethers, esters, etc. (Table II).

Table II

McBee Card Punch Coding For Some Oxygen Compounds

COMPOUND	CLASS	H FIELD	O FIELD	C FIELD
Benzyl alcohol	alcohol	s	s	d
Phenol	phenol	s	s	d
p-Cresol	phenol	s	s	d
Phenyl ether	ether	s	s	d
n-Valeraldehyde	aldehyde	s	s	d
Benzaldehyde	aldehyde	s	s	d
2-Hexanone	ketone	s	s	d
n-Valeric acid	acid	s	s	d
Benzoic acid	acid	s	s	d
Phenyl acetate	ester	s	s	d

s = shallow punch
d = deep punch

Another drawback to the McBee system is the necessity of using deep punching for all ion masses above 100 m/e, as the card size does not provide enough space for all of the required holes (data points). Going to a larger card will not solve this problem. Also a larger card would be extremely awkward to retrieve by manual sorting. When an ion mass greater than 100 m/e is deep punched, an ion mass less than 100 m/e is also punched; e.g., in Figure I the deep punching of the 140 peak has caused the punching of the 60 peak. This would lead to false drop-outs, thus causing the searcher to examine unrelated spectra.

Consequently, an edge-punched card system has already proven unworkable under laboratory conditions due to the proliferation of cards which leads to unsatisfactory sorting conditions, the nonavailability of new cards to maintain the file current, and the deficiency in chemical classification. As these deficiencies can not be readily overcome, the edge-punched cards have not been studied further in the development of an improved mass spectral data storage and retrieval system.

B. Computer-Based Systems

Hollerith cards have been used alone with sorters (3, 4, 5) in storage and retrieval systems. Haefele and Tinker (3) devised a system of storing chemical structures on IBM cards in a format that would be compatible with computer input at a later date. The system that evolved needed up to 17 cards per compound to completely describe the structural configuration of any compound. The various aspects of chemical structure were divided into 17 major groupings. Each of these groupings was assigned a separate card and specific columns in that card were assigned to various functional groups that are part of the major group (Table III).

Table III

Information Coded on IBM Cards

CARD NUMBER	INFORMATION PRESENT
1	Molecular formula
2	Number and type of ring structure
3, 4, 5, 6	Size and number of heterocyclic rings
7	Carbon atoms
8	Oxygen functions outside of rings
9-16	Other functional groups, such as nitrogen, sulfur
17	Patterson ring number and ion codes

As an entire column in a card was reserved for one particular functional group, its presence in the compound was indicated by a punch anywhere within the column. The authors had considered using a positional punch to show how the functional group was connected in the molecule but as this involved linear notation with its attendant rules and rigid regulations, this idea was discarded. Eventually a simple punch code was devised showing only the most important connections; e.g., 1) connection to a carbon in a ring; 2) connection to a carbon not in a ring; 3) connection to a halogen, etc.

The appropriate structure cards were punched with the accession number of the organic compound. The inverted system of filing was used; e.g., to find a compound, such as an aminophenol, the cards containing amino functions and the cards containing phenol functions are pulled and sorted in the proper columns for the amino group and the phenol group. The two sets of remaining cards that contain these functions are matched with a collator for an accession number in common. When matching accession numbers are found, then the compound with that number should be an aminophenol.

Kuentzel (4) devised a system which required one card per compound to store the data necessary for sorting on infrared absorptions and chemical structure. To obtain this result he devised a code for the punching of these cards. The columns on the card were assigned to the following fields: columns 1-28, coding of wavelength numbers; columns 29-31, reserved for future use; columns 32-57, chemical classification and structure; columns 58-62, the number of oxygen, carbon, nitrogen, and/or sulfur atoms present; columns 63-65, melting point; columns 66-70, reserved for the use of the individual laboratories; columns 71-79, source of documentation (could be used to give a journal, volume, page, and year). The storage of all this information on one card was possible through the use of a rigid code. Each punch within a column field had a discrete meaning; e.g., a y overpunch in column 1 indicated that the wavelength was in reciprocal centimeters.

A punched-card filing system for mass spectral data was developed at Dow Chemical Company (5) to assist in the identification of unknowns which were usually present in a mixture. It was found that Zemany's edge-punched cards (2) and Kuentzel's Hollerith-type cards (4) did not serve their needs. This led to a tabulation method which involved prepared lists for each mass number (m/e) of all compounds having a significant peak at that mass.

To prepare the lists of compounds the data was punched into a standard IBM punched-card (Figure II). The following data was punched into the card as indicated: arbitrarily assigned serial number of the compound, columns 1-4; molecular weight, columns 5-7; boiling point in centigrade, hundreds and tens figures only, columns 8 and 9; an x punch in column 8 signifies a negative number and in column 9, a melting point; base peak in m/e, columns 10-12; the second through the tenth highest peaks, columns 13-39; five so-called"peculiar" peaks not in the ten highest, columns 40-54; peaks at a fractional m/e recorded in order of magnitude of the next lower whole mass number, columns 55-76; number of chlorine atoms, column 77; number of bromine atoms, column 78; and the number of the deck, columns 79 and 80. The deck number was vital as the system

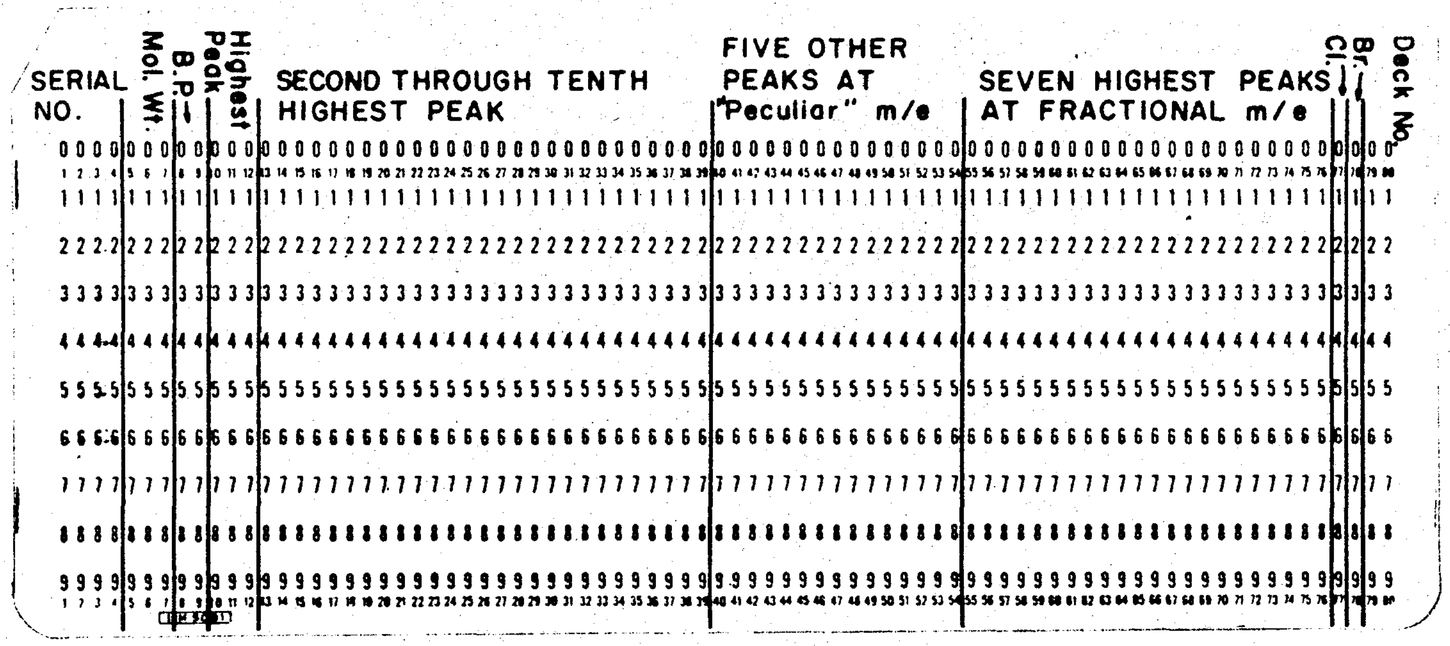

SERIAL NO.
Mol. Wt.
B.P.
Highest Peak
SECOND THROUGH TENTH HIGHEST PEAK
FIVE OTHER PEAKS AT "Peculiar" m/e
SEVEN HIGHEST PEAKS AT FRACTIONAL m/e
Cl.
Br.
Deck No.

Figure II

Dow Chemical IBM Punched-Card System

required 22 duplicates of the master deck with the duplicate sets employing different colors.

By means of an IBM sorter the first of these decks was arranged in order of mass number of the strongest peak (columns 10-12). Deck number two was sorted on the second highest peak in columns 13-15, and so on for the succeeding peaks. Cards that had no corresponding peaks in a particular field were discarded from that deck during the sort on that field.

An IBM collator sorted the cards in order under each mass number. That is, all cards in deck one with their strongest peak at 50 m/e were followed by the cards in deck two which had their second highest peak at mass 50, and so forth. Thus, all cards filed under a particular peak would contain all the compounds that had that peak in their ten highest and five additional peaks.

To identify a compound with a particular peak in a mixture, the cards filed under that m/e were examined. Comparison of these filed spectral cards could be made without any machine sorting unless a certain boiling point, additional peak, etc., was required. These cards were used to prepare a listing so that the cards did not have to be handled each time a search was made.

The Dow workers felt that the advantages of their system were; 1) the elimination of a hand or machine sorting operation; 2) the standard spectra were arranged in order of magnitude of the peak; 3) each mass unit up to 1000 was exactly designated; and 4) fifteen significant peaks plus seven fractional mass peaks could be included without omitting peaks due to an arbitrary relative ion abundance limit.

In 1958 the Sub IV Task Group within the American Society for Testing and Materials E-14 Committee proposed a system which represented an attempt to make the data handling all inclusive on one card (15). The following types of data were coded on the card: prominent specific masses; parent and base peaks; instrument conditions; molecular geometry; functional groups; molecular formula; boiling point; literature reference; and serial number. To code all this information on one card, a rigid set of rules for coding was required. These rules assigned a specific designation to each punch within each particular column. This system was not accepted and a new task group was assigned to develop another system. The final system combined and modified the various features found in McLafferty and Gohlke's system (5), in the original American

Society for Testing and Materials proposed system, and in several private industrial systems. This system is available as a book index or on punched cards from the American Society for Testing and Materials.

The IBM punched-card (Figure III) contains spectral and chemical structure data. The spectral data consist of the six strongest peaks and their intensities relative to a base peak intensity of 100 units. The mass number of the strongest (base) peak is punched in columns 1-3, the mass number of the second strongest peak is in columns 4-6, the intensity of this peak relative to the strongest is in columns 7-8. Columns 9-28 contain the third through the sixth peaks and their relative intensities. The molecular weight is in columns 29-31. Columns 32-57 are assigned to chemical classification which uses the American Society for Testing and Materials E-13 code, which is employed for the Wyandotte-ASTM infrared spectral coding system. In this code every punch within a column has a designated meaning; e.g., 0 punch in column 40 indicates a methyl group; 1 punch in column 40, ethyl group; 2 punch in column 40, n-propyl group, etc. The E-13 code is also used for the melting and boiling points in columns 63-65. Columns 58-62 can be used for the indication of the number of atoms of carbon, nitrogen, oxygen, and sulfur

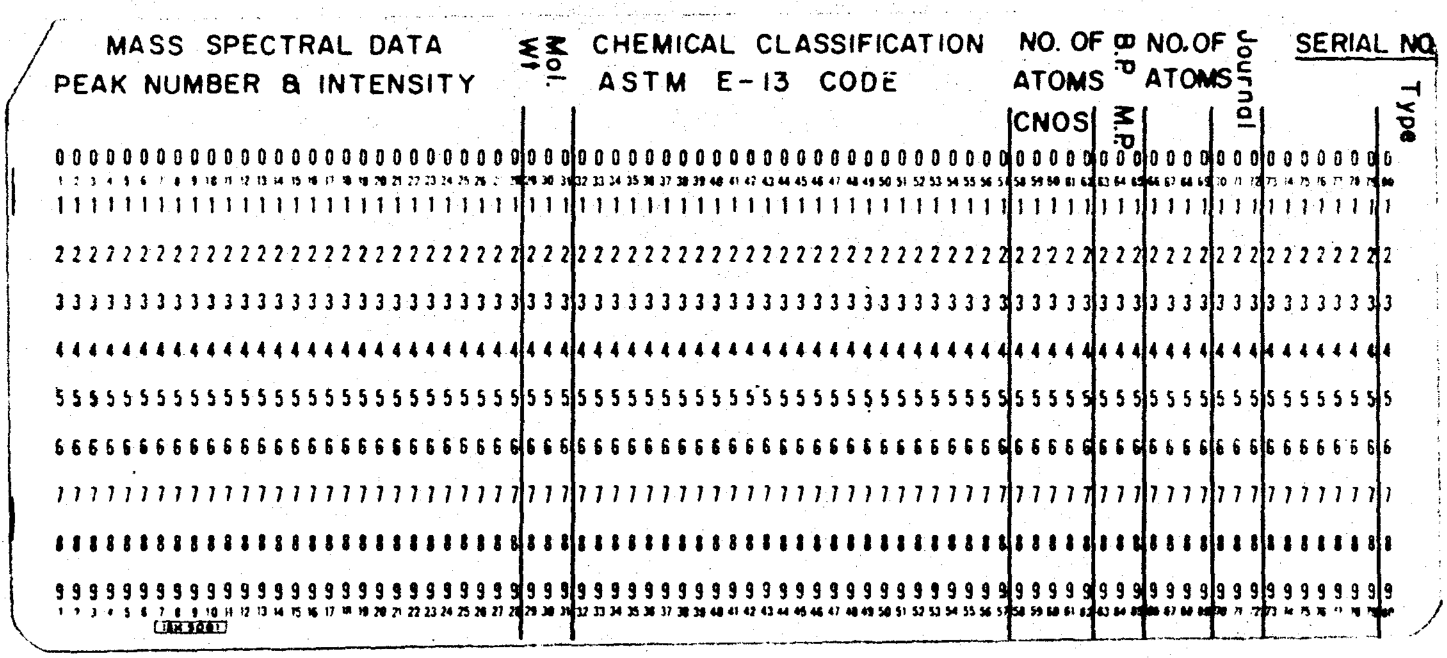


Figure III

ASTM E-14 Punched-Card System

present while columns 66-69 are used for fluorine, bromine, chlorine, etc. An identification code, consisting of a serial number, source of original data, and type of card, is punched into columns 73-80. The serial number is in columns 73-78. Column 79 indicates whether the source is the American Petroleum Institute files, user's file, uncertified, or literature, through the location of the punch within the column. A 12, 5 punch in column 80 indicates that the card contains mass spectral data, as opposed to a 12, 1 in this column which would indicate infrared data, a 12, 2 punch for X-ray data, etc.

Thus, two systems have been developed for use in mass spectrometry that involve Hollerith cards without the need for a computer. However, there are faults in both systems. While the system of McLafferty and Gohlke (5) was simple and straightforward, one big disadvantage was the preparation of the punched-cards. The necessity for 22 duplicate sets of the master deck involved excess time and cards. Also, once a particular mass number was chosen as an entry point to the file system, one still has to compare each one of the known spectra to that of the unknown. There was no provision for the coding of the relative intensities of the major peaks and, therefore, the original source still had to be examined to eliminate many compounds. Chemical classification was not included except for the presence of bromine and chlorine. All of these are felt to be serious drawbacks by some mass spectral personnel.

The American Society for Testing and Materials system overcomes some of the disadvantages of that of McLafferty and Gohlke (5) but introduces some of its own. In an attempt to include the relative intensities and still use just one card per compound, only the six strongest peaks are punched. Frequently, the first six major peaks alone will not separate compounds. Chemical classification is included, but this involves a rigid set of rules for coding and punching which would limit the American Society for Testing and Materials system's usefulness in the laboratory. A code designating each punch within a column with a set meaning is necessary when a system uses cards with a limited number of columns. The only way to obtain additional columns for punching is to add more cards per compound. The American Society for Testing and Materials method is actually set up for laboratory personnel that have ready access to an IBM sorter. The tabulation in book form that is available includes just the spectral data and thus has only a limited usefulness.

It can be seen that these systems based on Hollerith cards have built-in limitations. The objectionable features given above justified the elimination of Hollerith systems from further development.

To date one of the fastest methods for matching and retrieving mass spectral data is with the use of a computer. Cook (6), as noted earlier, connected two thermal emission mass spectrometers to an on-line digital computer, which was programmed to control the instruments and print out and/or punch on paper tape the results of the computations.

Abrahamsson et al. (7) use a digital computer to store the following information for each spectrum: all m/e's with their intensity values, name of the sample, molecular weight, source and serial number of the spectrum, temperature of the ion source and inlet system, electron energy, and an instrument identification number. The computer can be programmed so that any information in the spectra can be listed and sorted by any item; i.e., by molecular weight, chemical name, etc.

A search for spectra to match an unknown can be performed at the full reading speed of the magnetic tape unit, which takes about five minutes to search 10,000 spectra. The computer automatically writes out the name of the compound, its molecular weight, and an index figure which indicates the degree of similarity when an identical or closely similar spectrum is found in a search. With an attached x,y-recorder the computer can draw the mass spectrum in the usual line diagram form.

Hamming et al. (8), Smith et al. (9), and Pettersson and Ryhage (16) have all developed systems to evaluate and search mass spectral data through the use of computers. Smith et al. compared the relative efficiences of the various methods devised by him and others. He found that efficient retrieval had been achieved despite any spectral variations due to any differences that existed in the instrument type or the sampling method that was used.

It can not be denied that at present computers probably are the fastest and most efficient way to identify compounds by spectra matching, but there are inherent drawbacks to this type of system. Computers and computer time are expensive. The system of Cook's (6) costs $90,000, excluding the mass spectrometers. This cost factor is prohibitive to the majority of potential users. Even where computers are available, the speed of retrieval is often overshadowed by the factor of acquiring time on the computer. Most firms schedule computer work to be done in batches. Thus, the mass spectral group might be allotted computer time only every two weeks or even less frequently. Therefore, completing a search in five minutes is not a time-saving system if one has to wait two weeks to be able to do this. Also there is no standard system of chemical notation in use. A recent survey in Europe (17) of 15 organizations revealed that a total of 19 different chemical notation systems had been devised, and that there was no single standard system in use.

Consequently, for the above reasons; i.e., chemical notation need, cost, and time factor, it was decided not to attempt to use computers to solve the mass spectral data storage and retrieval problem.

C. Optical Coincidence Cards

Most of the systems discussed up to now use the conventional mode of data processing; i.e., there is only one reference or compound per card/or cards. Optical coincidence systems are based on an inverted data processing mode; that is, each record unit handles a single characteristic and the identities of the information items are coded on the appropriate record unit.

Coordinate indexing is normally used to code optical coincidence cards (18). In coordinate indexing each item of information to be filed or stored is given an unique serial number. The items can be listed at random as long as the assigned serial numbers are consecutive. Each of the terms used for indexing or coding the data is assigned a record unit, or a characteristic card. Each of these cards contains a fixed number and pattern of assigned but unpunched hole positions in the interior of the card. Each of the holes represents a serial number, starting with 0000 and continuing consecutively up to the maximum that the card can contain. If the term represented by the card applies to an information item, then the item's assigned number would be punched in the position indicated on the card. Thus, a term card will contain only the numbers of the information items which are described by the term of that card.

To retrieve an item stored in this type of system, one chooses the characteristics that best describe the item. The cards; i.e., card X for term X; card Y for term Y; card Z for term Z, are removed from the card holder and are superimposed on a reading box which contains a light source. Wherever the holes coincide; i.e., allow the light beam to pass through cards X, Y, Z (Figure IV), the reference number indicated by that particular hole will have all three of the desired characteristics. The number is obtained by reading the two digit coordinates of the x- and y-axis. This type of system is also known as "peek-a-boo".

One of the best known commercial versions of optical coincidence cards is the Termatrex system of Jonker Business Machines, Inc. The cards have a 9" x 9" coding field which can accomodate 10,000 holes or data points. The available equipment from the firm covers a wide range of input-output devices. For input the drilling equipment can be a simple manually operated drill that uses a drilling template or can be an automatic drill that accepts data from IBM punched-cards, punched paper tape, or magnetic tape.

For output there is a simple card reader that uses a movable scale to read the illuminated holes. There is also available an automatic scanning device that will print the serial numbers of the illuminated holes or that will perform statistical studies on the file cards.

The Termatrex cards are available in units of 100 and in ten different colors. Across the top of each card there are 100 positions numbered from 00 to 99. Within the deck there is a projecting edge tab (Radex tab) at each number position. Each numbered position within a color deck can be coded to represent the characteristic of that card. With these tabs and color coded top edges the Termatrex cards can be filed randomly and yet still be easily located during a search.

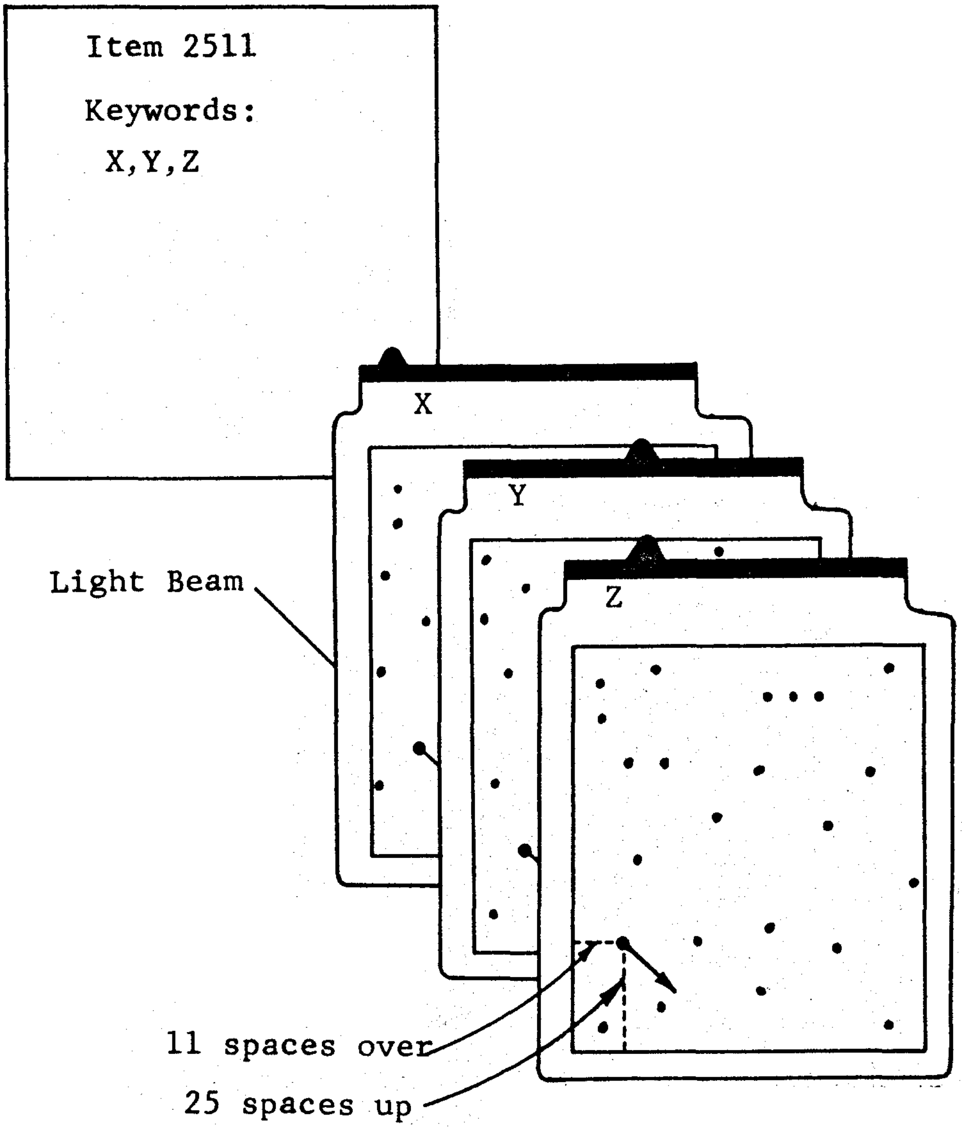


Figure IV

Optical Coincidence System

For files containing more than 10,000 items, Jonker has developed a microfilm "peek-a-boo" system, the Minimatrex, which can contain 100,000 items on each filmstrip. Each filmstrip holds five or ten separate frames with each frame a photographic reduction of a standard Termatrex card. Up to 12 of these strips can be superimposed in a special viewer for a single search.

Users of such systems can search their files in two to four minutes. This time requirement compares favorably to computer searching and is less expensive. Also Termatrex has the advantage of being available for immediate searching at any time. The cards themselves occupy little space. During a search a user is free to modify or change the terms being used as is needed. Termatrex systems seem to have all the features that are desired by mass spectral groups; i.e., desk file, availability, simplicity of use, flexibility, and speed.

Upon reviewing the advantages and disadvantages of all the systems investigated, it was decided to design a system based on the optical coincidence technique for storage and retrieval of mass spectral data.

II. Termatrex Systems

Three working systems using Termatrex cards will be discussed. The first system is in the field of gas chromatography.

When the weekly abstracts from the literature on gas chromatography published by the Preston Technical Abstracts Company in form of edge-notched cards reached a total of 10,000 cards, there was an obvious need for an improved data retrieval system. Jonker and Preston working together devised a Termatrex system which simplified the retrieval of the gas chromatography literature. The resultant system uses 80 Termatrex cards with each card representing one item of an expanded classification system. This index covers the field of gas chromatography through December, 1964.

The index classification employs six major subject headings: general, apparatus and techniques, source of information (language), number of carbon atoms in the compound, type of compound, and stationary phases. The stationary phases group employs 23 cards which can be used singly or in combination to describe up to 299 adsorbents and stationary phases.

The system is used as follows: In analyzing a mixture of five carbon atom alcohols on a Carbowax column to confirm the order of elution, the following cards would be superimposed for optical coincidence reading:

Card for "alcohols"
Card for "five carbon atoms"
Card for "Carbowax"

By the use of a language card the references are limited to those in one particular language, such as English.

Jonker also produces the second optical coincidence system, one originally devised by Matthews (14) for the retrieval of X-ray diffraction powder data. Prior to the Matthews' system, the X-ray powder data had been stored on edge-punched cards and then IBM punched-cards, but neither technique had proven satisfactory. The standard deck of cards now in use consists of 103 cards: three single cards showing the presence of minerals, alloys, and hydrates; 50 cards for elements present; and 50 cards for the "d" value, which is the value given in angstroms of the interplanar spacing of the crystal planes of the substances. These values range from 1.5 to 10 Å, and this range is divided into 50 segments that give an approximately equal distribution of substances within each division. The five strongest lines are used for indexing most of the material.

A supplementary set of cards is available. In this set the range of "d" values included in the first set of cards is shifted by half so that a descriptor card is always available on which values above and below the measured "d" value are included; i.e., for cards with ranges of 1.700-1.799 Å and 1.800-1.899 Å in the first set, the second set would have cards with the ranges of 1.750-1.849 Å and 1.850-1.949 Å.

The third system was devised by Schlichter and Wallace (13) who were seeking a way to quickly retrieve infrared spectral data. They had been using the Wyandotte-ASTM system on IBM punched-cards, but as the file grew to 50,000 cards, the time required for sorting became too long and costly. The developed Termatrex system requires only 250 cards for an input of 10,000 compounds.

The system uses four colored decks of cards: 1) white for general items such as alkyl groupings, olefin groupings, and structure classification; 2) red for functional codes; 3) green for fluoro-carbon functional group codes; and 4) brown for a semi-empirical formula code and a modified wavelength code. The developers of the system retained about 80% of the original ASTM coding used in the Wyandotte-ASTM IBM cards with the addition of a few new codes required by their particular work.

Most of the 100 white cards are used to indicate the elements, types of unsaturation, the structure, types of ring substitutions, miscellaneous and heterocyclic groupings, alkyl groupings, and olefin groupings. The red deck is used to code the functional group units involving oxygen, nitrogen, or sulfur with or without a single carbon atom. Codes have been added for acid halide C=0, P=0, metal C=0, NF, and SF. A set of codes has been developed for a fluorocarbons project and is encoded on a green deck. Approximately 50% of the brown cards are used for a semi-empirical formula code. There are cards for each individual atom up to 20 carbon, 6 nitrogen, 10 oxygen, and 6 sulfur atoms.

The new feature in their system is the modified wavelength code. The wavelength coding of the American Society for Testing and Materials was designed to separate compounds from one another. The developers' work involves new compounds for which there are no previous existing spectra, and the major interest is in determining the types of compounds that could produce a specific type of spectrum. The functional group region is divided into seven large areas between 2.7 to 7.6 μ. The fingerprint area is divided into 0.2 μ intervals up to 11.0 μ and in increments of 0.5 μ thereafter. Each wavelength code has its own characteristic card in the brown deck. If a band is located within ±0.02 μ of a coding unit division, it is coded on the cards falling on either side of the coding division. Bands are coded in order of decreasing strength until 10-15 code units have been used.

Thus, everything is on a relative basis rather than on an absolute basis.

Schlichter and Wallace can identify a compound in their deck of 3,000 compounds in an average time of two to three minutes. This is much faster than sorting the IBM cards of the Wyandotte-ASTM method, and now all of the searches can be done in the laboratory. One of the disadvantages is the high data input time; i.e., 16 man-hours per 50 compounds. However it also takes the same length of time to code data for the Wyandotte-ASTM system.

At present Jonker is working with the American Society for Testing and Materials to develop a method for indexing NMR spectra. A trial index of this data which codes 664 spectra on 266 cards has been issued.

The American Society for Testing and Materials is attempting to develop a system for coding the American Petroleum Institute spectra on Termatrex cards. W. B. Askew of DuPont (19) has been involved in this work but at present a functioning system has not been published. The chemical classification that has been suggested for this system is similar to the one used by Schlichter and Wallace (13). One other proposed feature in the American Society for Testing and Materials' tentative system is the coding of each of the five strongest peaks on an individual Termatrex card.

DEVELOPMENTAL

I. Method Development

The objective of this study was to devise a mass spectral data storage and retrieval system which would satisfy the following requirements for a working laboratory system: quick access, relative simplicity, and the inclusion of definitive parameters.

Basic to any storage and retrieval system is the coding of information to be stored. In this case, the coding of the compound's spectrum in a Termatrex system would involve numbering the spectra consecutively. For retrieval purposes there should be a numerical listing giving the compound's name and the source in which the original spectrum could be found. The source could be an original article in a journal, any file or compilation used in the laboratory, or any spectra that are used from a personal file.

After numerous meetings with the mass spectral group at Philip Morris Research Center, certain data characteristics were selected to be included in the system. These characteristics will be discussed in the following order: chemical classification, molecular weight, boiling point, base peak, and other selected peaks with their relative intensities. The specific requirements of each of these characteristics in a Termatrex system and the coding of the characteristics on Termatrex cards will be discussed.

The major groupings to be included in the chemical classification are hydrocarbons, nitrogen compounds, nonmetallic element present, natural products, oxygen compounds, sulphur compounds, metallic-organic compounds, and other compounds. Each of the major groupings in the chemical classification is allotted 20 cards, which allows 19 subheadings. In no major group have all the cards been assigned to specific subheadings. This permits the system to have new subheadings added when the need arises.

A compound's serial number would be coded into the appropriate major group card. If the compound belongs to a subgroup under that major heading, then the compound's number would also be coded in that subgroup card. This creates a system in which a compound can be retrieved by knowing either the general or the specific chemical class. The specific class, when known, would serve to effectively limit the number of unknown compound possibilities.

The majority of the compounds received for mass spectral analysis do not have molecular weights over 500. The molecular weight range of 500 to 1000 can be incorporated into the system with greater increments between the division points. At first, it was proposed to have two sets of molecular weight term cards to provide overlapping. To do this would have required almost 100 Termatrex cards. The proposal was found undesirable, since the entire determination would be incorrect if the searcher miscalculates more than one mass unit in selecting the parent peak (= molecular weight). Finally digital coding was chosen for molecular weight, which would require a working deck of 30 Termatrex cards. There are sets of 10 cards for the hundreds unit, 10 cards for the tens units, and 10 cards for the ones unit. For a molecular weight of 136, the one, the three, and the six cards would be chosen from the hundreds, tens, and ones units, respectively. In those cases where the parent peak can not be determined accurately it is possible to examine all candidate compounds by using only the hundreds and the tens cards.

The mass spectral group felt that the boiling point code need only cover the range from 0 to 300°C. It was proposed to use increments of 20° and two sets of cards. Cards 65-78 would cover the boiling points of 0.0 to

280.0°C in increments of 20°. Card 79 would be used for degrees above 280.1°C. The second set would be coded as follows: card 80, 0.0-10.0°C; cards 81-94, 10.1 to 290.0°C in increments of 20°; card 95, boiling points over 290.1°C. With these two sets of overlapping range cards it would always be possible to choose a card that had values above and below the boiling point figure.

The first system proposed for peak coding was to code each peak on its own individual Termatrex card, and the base peaks would have been listed on celluloid overlays which would have covered a range of peaks. This idea was rejected since a large number of compounds have base peaks in definitive areas of mass numbers and the system did not provide enough resolution in those areas.

Another proposed system that would have greatly compounded the total number of cards would have been for each base peak to have its own set of cards. That is, all compounds with a base peak of 43 would be coded on one set of cards. Each set of cards would consist of a card for each individual peak exhibited by the spectrum of any compound with a base peak of 43.

None of the above proposed systems had taken into consideration the relative intensities of the coded peaks.

A system was proposed in which the stored mass spectra would have been presented in graphic form on celluloid overlays. The coordinates would have been the mass number of the peak on the abscissa and intensities relative to the base peak on the ordinate. This proposed system was eliminated from further consideration because the celluloid overlays could not be used constructively in conjunction with the Termatrex cards containing the other parameters.

Finally a system using peak mass number and relative intensity was evolved. In this system each peak number would be coded on its individual card. The peak intensity would be used as a limiting factor to decrease the number of compound possibilities. The intensities would be coded on colored celluloid transparent Termatrex cards. The system would employ five sets of transparencies, one for each of the following ranges of intensities: 0-16%, 15-27%, 25-53%, 50-79%, and 75-100%. The overlap of the intensity ranges allows for a 5% variation in the intensity value. Based on the daily laboratory run of n-butane, the usual deviation is 4.5% or less for the individual intensities. There would be six transparencies in each of the above sets; i.e., one each for the following peak mass number ranges: 1-39, 40-49, 50-59, 60-79, 80-99, and 100-120 m/e's. Only the peaks within the intensity range coded on a set of transparencies would show optical coincidence when the transparency was superimposed over the peak term cards. The coinciding holes of compounds with the same peak mass number but whose peaks were of different intensities would appear as colored light beams.

Following consultation with the Jonker people, a technique known as positive-negative input was evolved. By the use of this technique the relative intensity can be associated with its individual peak. This technique uses the system given in the preceding paragraph for the peak mass number and relative intensity with the slight modification given below.

Each compound, instead of being assigned a single number, is given a family of numbers. In this case each compound is identified by a group of five consecutive number codes; for example, 0000, 0001, 0002, 0003, and 0004. Each one of the five number codes is assigned to a peak. The peaks which are numbered are the five strongest peaks of 120 m/e or lower. When the compound is coded on a characteristic card, all five number codes are drilled. The individual numbers are used only on the positive-negative cards. The positive-negative cards are opaque Termatrex cards, and there are six sets, one for each of the following peak mass number ranges: 1-39, 40-49, 50-59, 60-79, 80-99, and 100-120 m/e's. There are five cards in each of the above sets, one each for the following intensity ranges: 0-16%, 15-27%, 25-53%, 50-79%, and 75-100%. On these cards all the positive data for the peak mass and intensity ranges are drilled as well as the negative data outside that peak mass range; e.g., on the card, 40-49 m/e and 15-27%, all numbers identifying peaks below 40 m/e and above 49 m/e would be drilled as negative data. For peaks of 40-49 m/e, only those numbers identifying peaks in that range which have an intensity of 15-27% would be drilled as positive data. That is, number 0010 identifying peak mass 41 with an intensity of 20% would be drilled as positive data, while 0020 assigned to another compound's peak mass 41 with an intensity of 70% would not be drilled.

II. Method Developed

The optical coincidence system devised for storing and retrieving mass spectral data consists of approximately 410 Termatrex cards. The system uses 110 cards to code the compounds' parameters, and 300 cards to code the peaks present in the spectra. The spectra which are entered into the file are arbitrarily assigned five consecutive serial numbers. A numerical listing of the serial numbers gives the compound's name and the spectrum's original source.

Both the blue and the yellow edge-coded decks (Table IV) are used for the chemical classification. The numbers shown in the table refer to the individual blue or yellow card assigned to the specific characteristic. The major groups begin with numbers 1, 20, 40, 60, and 80 in the blue deck and numbers 1, 20, and 40 in the yellow deck. The orange edge-coded deck (Table IV) contains the boiling points, and is also used for the digital coding of the molecular weights. The green edge-coded deck contains peaks of mass numbers 0-99 m/e with each peak assigned in numerical order to cards 0-99. The white edge-coded deck contains peaks of mass number 100-199 m/e with each peak assigned in numerical order to the cards (the last two digits of the peak mass number are identical with the card's two digits). In the black edge-coded deck each card represents two mass units and is used for peaks of 200-399 m/e; i.e., card 00, 200-201 m/e; card 01, 202-203 m/e; card 02, 204-205 m/e, etc. All fractional peak masses are coded at the lower numerical value; e.g., a peak of 42.5 m/e would be coded as a peak of mass 42. Only the ten strongest peaks in each spectrum

Table IV

Classification of Cards in the Blue, Yellow, and Orange Decks

Blue

No	Class	No	Class	No	Class	No	Class
1	Hydrocarbon	20	N Compd	40	Non-Met.	60	Natural prod.
2	Aliph Satd	21	Amides	41	B	61	Alkaloid
3	Aliph Unsatd	22	Amines	42	Br	62	Amino Acid
4	Alicyclic	23	Nitriles	43	Cl	63	Carbohydrate
5	Aromatic	24	Nitro	44	F	64	Lipid
6	Arom Fused Ring	25	Hetero N	45	I	65	Terpene
		26	Hetero N & other hetero	46	P		
				47	Hetero not O,N,S		

Blue

No	Class
80	Others
81	Inorg.
82	Polymer

Yellow

No	Class	No	Class
1	O compd	20	S Compd
2	Acid	21	Sulphones
3	Alcohol	22	Thiocyanate
4	Aldehyde	23	Thiol
5	Ether	24	Hetero
6	Ketone		
7	Phenol		
8	Hetero		
9	Anhydride		
10	Ester	40	Met-org.
11	Lactone		

Table IV
(continued)

Orange

No	Boiling Point	No	Boiling Point
65	0.0- 20.0	80	0.0- 10.0
66	20.1- 40.0	81	10.1- 30.0
67	40.1- 60.0	82	30.1- 50.0
68	60.1- 80.0	83	50.1- 70.0
69	80.1-100.0	84	70.1- 90.0
70	100.1-120.0	85	90.1-110.0
71	120.1-140.0	86	110.1-130.0
72	140.1-160.0	87	130.1-150.0
73	160.1-180.0	88	150.1-170.0
74	180.1-200.0	89	170.1-190.0
75	200.1-220.0	90	190.1-210.0
76	220.1-240.0	91	210.1-230.0
77	240.1-260.0	92	230.1-250.0
78	260.1-280.0	93	250.1-270.0
79	over 280.1	94	270.1-290.0
		95	over 290.1

are coded. The purple edge-coded deck contains the 30 positive-negative input cards used for the correlation of relative intensity and the peak mass number. The ranges of 1-39, 40-49, 50-59, 60-79, 80-99, and 100-120 m/e are on cards 1-5, 11-15, 31-35, 41-45, 51-55, and 71-75, respectively. These cards cover the relative intensities of 1-16%, 15-27%, 25-53%, 50-79%, and 75-100% for each peak mass range. Only the five strongest peaks of 120 m/e or lower are coded on these cards.

III. Functioning of the System

A. Coding of a Compound and its Spectrum

The compound, 2,3-dimethylhexane found in the trial set, has been chosen to illustrate the technique of coding as presented in Table V. This compound has been assigned the numbers 0005-0009. This number code, 0005-0009, would be listed in the thesaurus identifying 2,3-dimethylhexane and would give the original source as the American Petroleum Institute certified files. The numbers, 0005-0009, would be drilled in the cards indicated in Table V.

Table V

Coding of 2,3-Dimethylhexane

DECK	CARD NUMBER	DESCRIPTOR	CODE NUMBERS
Blue	1	Hydrocarbon	0005-0009
	2	Aliphatic satd.	0005-0009
Orange	1	Mol. Wt. hundred	0005-0009
	11	Mol. Wt. ten	0005-0009
	24	Mol. Wt. one	0005-0009
		(M.W. 114)	
	70	Boiling point	0005-0009
	86	113-116	0005-0009
Green	27 43	Ten highest	0005-0009
	29 55	peaks	in all ten
	39 57		cards
	41 70		
	42 71		
Purple	15	43 m/e 100%	0005
	44	70 58	0006
	43	71 46	0007
	13	41 28	0008
	2	29 19	0009

B. Retrieval of a Compound

Table VI presents the procedure used in the identification of an unknown compound. When all the cards in Table VI have been superimposed, there will be among the remaining points of optical coincidence code numbers 0185-9, which in the thesaurus will identify hexanoic acid. Its spectrum, obtained from Stanolind Oil and Gas Company, is in the Consolidated Engineering Corporation McBee File.

Table VI

Compound Identification

KNOWN DATA	CARDS TO SUPERIMPOSE
Acid	Yellow 2
116 mol. wt.	Orange 1, 11, & 26
205° b.p.	Orange 75 or 90
Ten highest peaks	Green 27, 29, 39, 41, 45 42, 43, 55, 60, 73

If too many holes still showed optical coincidence, the following purple cards with the limiting intensities would be used.

CARD NO.	MASS RANGE	INTENSITY
45	60 m/e	100%
43	73	42
3	27	36
13	41	33
12 & 13	43	27

IV. Discussion

The preceding system provides the mass spectral personnel with a practical and economical laboratory storage and retrieval system for mass spectral data. This system incorporates all of the characteristics that the Philip Morris Research Center personnel felt were necessary for their own specific needs. Through the employment of the Termatrex method, the stored data is on a minimum number of input cards and requires very simple retrieval equipment. The file of data is kept current by expanding the data input on the body of the existing cards rather than by the addition of new cards to the system. Thus, each updating does not require additional laboratory space for storage nor does it increase the time for searching the data.

As was found by Schlichter and Wallace (13), one drawback to this type of storage and retrieval system is the relatively high data input time. In the experimental working deck the time required to code three to five spectra was approximately an hour. However, regardless of the system used, if all the Termatrex features were present in another system, then the input time for coding would be of the same magnitude.

The developed system offers a significant reduction in the time required to locate spectra in the identification process for unknown compounds. Preliminary studies have shown that it requires approximately five minutes to group the cards containing the coded characteristics of the unknown compound and to view the superimposed Termatrex cards. The present system of searching several individual sources, such as the American Petroleum Institute files and the Dow Chemical uncertified files, requires a minimum of 15 minutes to locate the spectra of possible compounds. Thus, the time saved searching for six spectra in a day would be one hour. Retrieval speed is a very important feature of this system.

The association of the peaks with their relative intensities to delineate the selection of a compound by its spectrum is an innovation. Others have tried to devise a system which included this feature, but no one previously had succeeded in doing this. Of 1500 systems produced or drilled for industry by the Jonker Business Machine Corporation, this is the first one to include a subrelationship; i.e., the intensities are associated to their individual peaks which in turn are related back to the compounds.

Experimental trials have shown this system to be an acceptable, working system. Full-scale employment might suggests slight modifications to further improve the retrieval of data.

SUMMARY

A rapid, simple, and economical laboratory method has been developed for the storage and retrieval of mass spectral data by the use of optical coincidence cards.

A new feature of this optical coincidence system is the correlation of the relative intensity of the peak to the peak mass number. The spectra of compounds are identified by associating the peak mass number and the relative intensity with any of the following characteristics: chemical classification, molecular weight, and boiling point. Initial results obtained with this system show it to be an efficient, time-saving means of identifying compounds by their mass spectra.

GLOSSARY

Collator - A device used to collate or merge sets or decks of cards or other units into a sequence.

Digital computer - A computer that operates by using numbers to express all the quantities and variables of a problem.

Drop, false - The documents spuriously identified as pertinent by an information-retrieval system, but which do not satisfy the search requirements, due to causes such as improper coding, punching spurious or wrong combinations of holes, or improper use of terminology.

Edge-punched card - A card of fixed size into which information may be recorded or stored by punching holes along one or more edges.

Field - A specified area of a record used for a particular category of data.

Hollerith - A widely used system of encoding alpha-numeric information onto cards; Hollerith cards has become synonymous with punch cards.

Magnetic tape - An external storage device in the form of a ferrous oxide coating on a reel of metallic or plastic tape upon which bits may be recorded magnetically as a means of retaining data.

Optical coincidence cards - (Batten) An information-retrieval system that uses peek-a-boo cards; i.e., cards into which small holes are drilled at the intersections of coordinates (columns and row designations) to represent document numbers.

Program - A set of instructions or steps that tells the computer exactly how to handle a complete problem.

Punch - Hole resulting from pressing a sharp edged tool through the card material.

Shallow - The portion of the card removed is between the hole and the edge, and a notch is formed.

Deep - There are two rows of holes around the edge of the card and the notch extends from the edge to the second row of holes.

BIBLIOGRAPHY

1. Cornu, A. and Massot, R., "Compilation of Mass Spectral Data," Heyden and Son Ltd., 1966.
2. Zemany, P. D., Anal. Chem., 22, 920 (1950).
3. Haefele, Carl R. and Tinker, John F., J. Chem. Doc., 4 (2), 112 (1964).
4. Kuentzel, L. E., Anal. Chem., 23, 1413 (1951).
5. McLafferty, Fred W. and Gohlke, Roland S., ASTM E-14 Committee Meeting on Mass Spectrometry, May 28, 1954, New Orleans.
6. Cook, H. D., Baker, F. B., and Hudgens, J. E., 13th Annual Conference on Mass Spectrometry and Allied Topics, May 1965, St. Louis.
7. Abrahamsson, S., Haggstrom, G., and Stenhagen, E., 14th Annual Conference on Mass Spectrometry and Allied Topics, May 1966, Dallas. Also in Biochem. J., 92, 2P (1964).
8. Hamming, M., Wright, W., Gartside, H., Ford, H., and Haley, J., 15th Annual Conference on Mass Spectrometry and Allied Topics, May 1967, Denver. Abstract only.

9. Smith, I. C., Kelly, W., Brickstock, A., and Ridley, R. G., ibid. Abstract only.

10. Silk, John A., J. Chem. Doc., 3 (4), 189 (1963).

11. Barnard, A. J., Jr., Kleppinger, C. T., and Wiswesser, W. J., J. Chem. Doc., 6 (1), 41 (1966).

12. Sippl, Charles J., "Computer Dictionary and Handbook," Bobbs-Merrill Co., Inc., 1966.

13. Schlichter, Naomi E. and Wallace, Ellen, Appl. Spectro., 17 (4), 98 (1963).

14. Matthews, F. W., J. Chem. Doc., 3 (4), 213 (1963).

15. American Society for Testing and Materials Committee E-14 on Mass Spectral Data.

16. Pettersson, Barbro and Ryhage, Ragnar, Anal. Chem., 39 (7), 790 (1967).

17. Frear, Donald E. H., "Survey of European Non-Conventional Chemical Notation Systems," National Academy of Sciences Publication 1278, Washington, D. C., 1965.

18. Bourne, Charles P., "Methods of Information Handling," p. 109, John Wiley and Sons, Inc., 1963.

19. Askew, W. B., private communication, May 23, 1967.

AUTOBIOGRAPHY

I, Cynthia Helmintoller O'Donohue, was born on October 3, 1936, in Washington, D.C. I attended elementary school in Richmond and Portsmouth, Virginia. I graduated from John Marshall High School in Richmond in 1954. I graduated from Randolph-Macon Woman's College in Lynchburg, Virginia in 1957 with a B.A. in Chemistry. In college I received the Una Burton Scholarship which is awarded to the outstanding junior student in chemistry and the Merck Award which is given to the outstanding senior student in chemistry. Upon graduation I was employed by American Tobacco Company until 1961 when I accepted a position with the Medical College of Virginia in biochemical research. In 1963-4 I was enrolled in graduate school at the University of Denver in Denver, Colorado, where I was elected to membership in Iota Sigma Pi, national honor society for women in chemistry. Since July 1965 I have been employed by Philip Morris Incorporated as an associate scientist.