Computer Editing of General Subject Heading Data for *Chemical Abstracts* Volume Indexes†

R. D. NELSON,* W. E. HENSEL, D. N. BARON, and A. J. BEACH

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received March 14, 1975

The *Chemical Abstracts* (CA) General Subject Index includes entries for such subjects as chemical and physical properties, reactions, uses and applications, and classes of chemical substances. The editing of the headings including subdivisional terms for these index entries has been converted from a fully manual operation to a system where the majority of the validation and modification of the input index entries is done by computer programs. Four types of operations are performed on the input index heading data: validation of the term against control files, modification of the term to correspond to the accepted CA index heading, correction of spelling and keyboarding errors (including character fonts) under defined conditions, and addition of certain diagnostic statements to the input index data for review by an editor. Consequently, 96% of the index headings are fully edited by the computer programs and only the 4% receiving diagnostics require additional human editorial review. This contrasts with the previous need for a 100% manual editing of General Subject Index headings for all CA volume and collective indexes. These programs are the first phase in the development of a vocabulary management system for general subjects parallel to the rigid vocabulary control exercised by the Registry System over chemical substance index entries.

Throughout the development of the Chemical Abstracts Service (CAS) chemical information processing system, the control of all information entering the CAS data base has been emphasized. Controlling the terminology and structure of index entries has been particularly important in the past for facilitating the production and use of the printed indexes, and its importance increases as the amount of indexed information grows and as the production of CAS indexes includes more computer based operations. Vocabulary control at CAS is exemplified by Registry System[1] techniques which support control of the chemical substance name vocabulary in CAS data bases through systematic computer interrelationships of the name and structure records. Now, procedures are being developed which will provide machine-checkable control of the General Subject vocabulary segment of the index data base which includes entries on chemical and physical properties, reactions, uses and applications, and classes of chemical substances. The purpose of this paper is to describe the system which has been developed and implemented to provide computer assistance in the control and validation of the index heading data of the CA General Subject Index.[2]

Editing Procedures. An index entry consists of three basic parts: index heading, index entry modification, and CA reference. An index heading is a word or group of words that designates a topic under which all data relating to that subject are indexed. Duplicate headings and duplicate modifications of alphabetically sorted entries are suppressed in the printed index, as illustrated in Figure 1, in which the heading information appears in boldface type. The primary access point to the heading is found in the Concept Heading data element, a unit of information of specific content and format which consists of information found in the first segment (labeled "1") of the heading. The heading may include a subdivision, exemplified in "2" and "5," which is that word or phrase following the comma or within the parentheses. Heading subdivisions are used primarily to organize headings with large numbers of entries into subgroupings of related interest or to differentiate among homographs. These subgroupings[2] may be (a) Qualifiers (as in "2"), e.g., analysis, biological studies, and reactions; (b) Functional Categories (as in "2.1"), e.g., anhydrides, esters, hydrazones, and polymers (see Table III for additional examples); (c) any subdivision established for a particular purpose (as in "2.2"), e.g., azo, blood, electro-, and apparatus (see Tables II and III for additional examples); (d) physiological heading subdivisions (as in "2.3"), e.g., metabolism and neoplasm, which are placed in the Qualifier data element; or (e) Homograph Definitions (as in "5"), e.g., plant, seaweed, and genus (see Table II for additional examples).

The Synonym part of the heading (shown at "6") which appears in italics is not controlled by these computer edit routines.

In the computer readable CA Integrated Subject File (CAISF), which corresponds to the General Subject Index,[3] a heading consists of the Concept Heading data element and, for some entries, a Functional Category, Qualifier, or Homograph Definition data element.

The index entry modification (as in "4") gives additional information about the index heading and appears in lightface type in the printed index. In the computer readable file, this information is found in the Text Modification data element. Not all index entries require a modification.

The abstract reference (shown at "3") appears with each index entry and refers to the appropriate CA abstract. The final character of the reference is a letter check digit which allows computer validation of the numerical part of the reference. Three types of source documents are identified by capitalized code letters preceding the reference: B for book, P for patent, and R for review.

Computer control procedures for the heading data were implemented with the Volume 76 General Subject Index (January–June, 1972). Prior to Volume 76, the heading for each index entry, after initial input to the data base, had to be proofed manually and validated by professional and clerical personnel prior to its appearance in the index. Since the majority of the index headings are correct at input, much of the manual review consisted of identifying those headings within the total data base which contained some type of error such as misspellings, variations in gram-

† Presented before the Division of Chemical Literature, 168th National Meeting of the American Chemical Society, Atlantic City, N.J., Sept 11, 1974.

* To whom correspondence should be addressed.

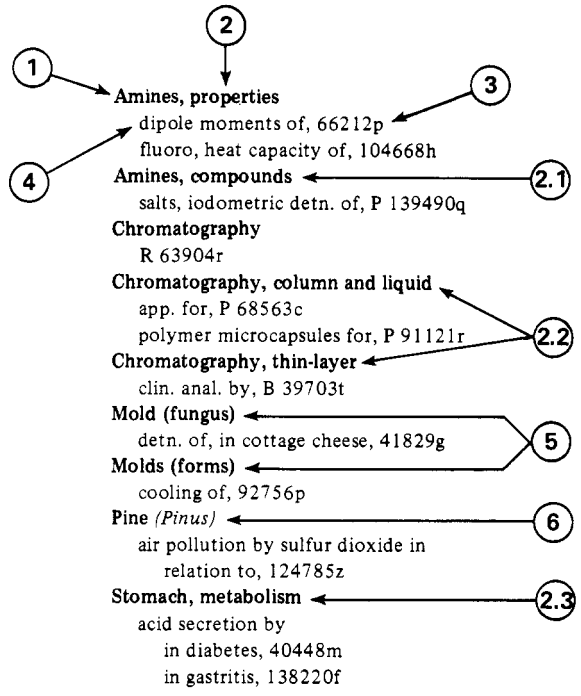

Figure 1. Illustrative key to general subject index entries.

matical endings, incomplete or improper combination of heading terms, heading terms not approved for CA index usage, or the oversight of applying cross references. Computer identification of those headings which do not meet output standards would, at the initial editing stage, aid the technical editor in easily locating problem entries as well as identifying the type of problem.

The computer editing routines include the matching of index heading data against CAS standards, the standardization of data and indication of data changes through applicable diagnostic messages ("diagnostics"), and the provision of diagnostics that identify the input data which cannot be changed to CAS standards automatically because there are various valid alternatives. (In the latter case, human intervention is required to provide the correct alternative.) Thus, one major problem, that of finding and identifying the error, is solved. The effort saved is significant since about 330,000 General Subject Index entries are prepared per volume (every six months) with the corresponding processing rate of 12,700 index entries per week.

The operations performed by the General Subject Edit routines are summarized as follows:

1. The input index heading terms matching the control file are output as approved heading terms.

2. Changes between the singular and plural forms of a word that involve only the final "s" allow input of either form of the Concept Heading while algorithmically outputting only the approved form. Singular–plural changes have been cited as one of the troublesome problems of any vocabulary handling system.[4]

3. Input of a synonym (or nonapproved term) results in the output of the approved Concept Heading. In effect, this arrangement automatically applies a cross-reference, modifies a heading to its final form for the index, or provides singular–plural changes which involve characters in addition to the letter "s" (i.e., cherries to cherry, alkali to alkalies).

4. The first Roman alphabetic character of the Concept Heading is capitalized if input as a lower case character. Specific exceptions are recognized by the program.

5. The diagnostic APPLY CROSS REFERENCE is issued for those input records belonging under another heading but which the programs are not able to correct. The choice of the proper index heading requires human intervention.

6. A NO MATCH diagnostic is applied to work units (groups of data which pertain to one indexable entity or idea) containing Concept Headings not processed by the above procedures.

7. Since some of the NO MATCH diagnostics are the result of spelling errors, work units carrying this diagnostic are automatically processed through a spelling correction algorithm whereby some Concept Headings are changed to the approved CA index form.

8. Certain Concept Headings, processed according to procedures 1, 2, 3, and 7, are further interrogated to determine whether the index entry requires a subdivision. The improper presence or absence of a defining term triggers the appearance of a diagnostic. In addition, the subdivisional term, if present, is matched against a control file to assure its correctness.

The Control File.[5] The General Subject Edit Control File is a group of permanent computer files used to verify the correctness of all index headings for the CA General Subject Index. These files are rigidly controlled to assure continuity of the index heading data within a volume and among the volumes of a collective index period. In building these files, the position has been taken that any data added to the file must be thoroughly reviewed to meet index standards and to assure compatibility with the CA Index Guide. The addition of uncontrolled data into the control files is avoided partly by the absence of any automatic updating features for the files and by limited access to the files which assures controlled building and maintenance.

Printed forms of the files as well as machine-readable files are available for editorial review and are used in the control and maintenance of the file contents. The format of the printed listings is illustrated in Tables I–III. Six types of information are incorporated into these listings: (1) the computer address for each record, (2) the CA approved primary access point of the index heading (Concept Heading, which has the mnemonic CTH), (3) synonymous terms (Synonym, SYN) and their Concept Heading equivalents, (4) Concept Headings with specific required subdivisions either Qualifier (QLF) or Homograph Definition (HOM), (5) Concept Headings with Qualifier and Functional Category (CAT) subdivisions, or (6) Concept Heading User Data Codes (UDO) which assist specified processing and identification functions.

The first and most important aspect in file building is to verify the validity of all heading terms. The appropriate data elements and computer codes are then assigned to the data. Specialists decide, using the Index Guide, whether a new concept is an acceptable heading or should be a cross-referred term. From this process, a list of Concept Headings with their associated data elements is compiled for input into the computer. The original file (General Subject Edit Control List 1, cf. Table I) containing 13,082 terms was a composite listing of more common Concept Headings which appeared in the Eighth Collective Subject Index and of new headings effective for Volume 76, the first volume of the Ninth Collective Index period (1972–1976). Further additions have been made by including new incoming terms, numerous genus-species and common names of animals and plants from a CAS taxonomy file, and terms validated in the Index Guide. At present the file contains 45,231 records of Concept Headings. Of these, 4000 are nonapproved terms. For example, "Molecular heat" is equated to "Heat capacity" (see Table I). "Molecular heat" is the synonym and appears in the column with the heading CTH or SYN, and "Heat capacity" is the approved Concept Heading and appears in the CTH ASSOCIATED WITH SYN column. Over half the file is composed of taxonomic terms (about 28,500). The vocabulary of the Concept Headings in the CA General Subject Index devoted to chemistry, chemical en-

Table I. Chemical Abstracts Service General Subject Edit Control List 1

BATCH	GRP	UDO	GRP	CTH OR SYN	GRP	CTH ASSOCIATED WITH SYN
4505	195	1E1Q	194	Alkanes		
4544	210	1X	211	Alkanesulfonic acids		
4819	741	9S	742	Alkanna tinctoria		
4820	45	1X	46	Alkanols		
4505			198	Alkaptonuria		
4544	212	1X	213	Alkene oxides		
4505	197	1E1Q	196	Alkenes		
4505	193	1N	192	Alkyd resins		
4544	320	1X	321	Alkyl		
4505			182	Alkylamino groups		
4844			5	Alkyl astatides		
4540			32	Alkylates		
4505	238	2F	183	Alkylating agents		
4505			184	Alkylation		
4505			185	Alkylation catalysts		
4583			158	Alkyl azides		
4505			186	Alkyl bromides		
4505			187	Alkyl chlorides		
4506			17	Amidoximes		
4542			71	Amidrazones	72	Hydrazidines
4544	717	1X	363	Amine oxides		
4506	21	1E1Q	20	Amines		
4844			106	Aminimides		
4506			44	Amino acid metabolic disorder		
4506	23	1Q	22	Amino acids		
4525			56	Arteries	57	Artery
4506			286	Arteriosclerosis		
4525	313	1B	314	Artery		
4544	316	1X	317	Association		
4508	112	1B	113	Brain		
4622	229	2V9S	230	Brake	349	Bracken
4622	256	2H	228	Brakes		
4541	193	1G	15	Bromometry		
4629	419	2V9S	5	Bryophyllum calycinum	6	Kalanchoe pinnata
4629	420	2V9S	7	Bryophyllum crenatum	8	Kalanchoe laxiflora
4640	3	9S	4	Bryum ventricosum		
4544	690	1X	691	B-strain		
4629	424	2V9S	15	Bubalus	16	Water buffalo
4629	425	2V9S	17	Bubalus bubalus	18	Water buffalo
4542			236	Bubble caps	237	Plates and Trays
4508			173	Bubble chambers		
4622	262	2H2V9S	248	Mold		
4526			69	Molding compositions		
4501			57	Molding of plastics	58	Molding of plastics and rubbers
4554			86	Molding of plastics and rubbers		
4501			59	Molding of rubbers	60	Molding of plastics and rubbers
4622	263	2H	249	Molds		
4599	322	2H	201	Mole		
4553			622	Molecular heat	623	Heat capacity
4551			124	Molecular index		

gineering, medical sciences, etc., accounts for the remaining (about 13,000) output CTH terms, e.g., "Hydrolysis", "Distillation apparatus", "Hyperthyroidism".

Table II contains a portion of the General Subject Edit Control List 2 listing the Concept Headings (CTH) associated with either a Qualifier (QLF) or a Homograph Definition (HOM). Currently there are 202 CTH–QLF relationships and 241 CTH–HOM relationships on this file. The General Subject Edit Control List 3 (see Table III for a portion of the list) contains 83 Qualifiers which may be associated with Concept Headings in addition to those given in List 2. List 3 also contains the 18 Functional Category terms used with specific Concept Headings in the CA Indexes. These lists contain all the index heading data currently approved for the General Subject Indexes for the Ninth Collective Index period.

The control and classification codes for the associated Concept Headings are listed in the column headed UDO in Tables I–III. Codes beginning with the numeral 1 are associated with diagnostics printed on the index work sheets supplied to the editors to review and correct. Those codes beginning with the numbers 2 or 3 direct data to specified processing operations of the General Subject Index Edit Routine such as checking the validity of CTH–QLF combinations and may also generate a diagnostic. Codes beginning with numbers 4 through 9 are classification codes. Presently only one classification code is used: 9S for the taxonomic classification. The UDO codes now in use are the following.

1B. The CTHs associated with this code are physiological organ and tissue headings requiring a subdivision.[6] This computer code triggers the diagnostic MISSING FRAGMENTATION if a QLF is not present. In cases where a QLF is not to be used at the heading, the appearance of "general" in the Index Entry Note data element will prevent the diagnostic from printing.

1E. This code signifies the existence of heading content notes or assumption notes which appear with a CTH in the Index Guide and triggers the diagnostic SEE INDEX GUIDE NOTE. Heading content notes are statements of index policy regarding the type and range of information to be found at the particular index heading. Assumption notes explain the assumptions made by the indexers which also should be made by the user with regard to the subject described by the particular index heading. Illustrations of notes can be found in the Index Guide at the headings "Al-

Table II. Chemical Abstracts Service General Subject Edit Control List 2

BATCH	GRP	UDO	GRP	CTH	GRP	QLF	GRP	HOM
4694	1	1X9S	2	Absinth			3	plant
4694			4	Absinth			5	liqueur
4695			1	Adjuvants	2	immunological		
4694			6	Alaria			7	fluke
4694			8	Alaria			9	seaweed
4694	13	1X	14	Albedo			15	neutron
4694	10	1X	11	Albedo			12	citrus peel
4694	16	1X	17	Albedo			18	biological phenomenon
4694			19	Albedo			20	reflection phenomenon
4695			352	Albinism	353	plant		
4695			354	Albinism	355	animal		
4695			6	Alkylating agents	7	biological		
4695	3	1X	4	Alkylating agents	5	chemical		
4694			21	Alligator			22	genus
4694			23	Alpaca			24	animal
4694			27	Alsophila			28	moth
4694			25	Alsophila			26	fern
4694			29	Amaranth			30	Amaranthus
4694			31	Ammonite			32	fossil
4695			10	Antimetabolites	11	plant		
4695			8	Antimetabolites	9	animal		
4694			37	Aristotelia			38	plant
4694			35	Aristotelia			36	insect
4695			15	Ascites	16	syndrome		
4695	12	1X	13	Ascites	14	neoplasm		
4694			39	Ash			40	Fraxinus
4694			41	Ashes			42	residues
4694			43	Asterococcus			44	alga
4694	609	1X	45	Asterococcus			46	bacterium
4694			47	Ataxite			48	rock
4694			52	Azurins			53	proteins
4694			54	Bacillus			55	insect
4695	19	1X	20	Baths	21	biological		
4695			22	Baths	23	compositions		
4695			17	Baths	18	apparatus		
4695			24	Batteries	25	primary		
4695			26	Batteries	27	secondary		
4694			58	Benthos			59	sea bottom
4694	585	9S	56	Benthos			57	marine organism
4694			325	Mitella			326	animal genus
4694			327	Mitella			328	plant genus
4695	397	1E	142	Models	143	physical		
4694	593	1E	329	Mold			330	fungus
4694			331	Molds			332	forms
4694	584	9S	333	Mole			334	animal
4694			335	Mole			336	gram molecular weight
4694			337	Mole			338	neoplasm
4694			596	Mummy			597	preserved body
4694	600	1X	598	Mummy			599	organic matter

cohols", "Dyes", "Light", and "Replacement nomenclature".

1G. This computer code triggers the diagnostic GENERAL HEADING. Headings associated with this code are restricted to more general types of information as shown by this example from the Index Guide.

Viscosity

Studies of viscosity itself, or of the viscosity of classes of substances, are indexed at this heading. For studies of viscosity of specific substances, see those specific headings.

1N. A CTH having this code should not occur with a Functional Category or Qualifier. If a Qualifier or a Functional Category is used in association with one of these headings, the diagnostic HEADING NOT CATEGORIZED OR QUALIFIED will appear.

1Q. The CTH having this UDO must be linked to a Qualifier which is associated with a UDO of 2C (Table III), or to a Functional Category. If the CTH is not so linked, the diagnostic MISSING CATEGORIZATION will appear.

1X. The CTH associated with this UDO is cross-referred to an approved index heading. This code triggers the diagnostic APPLY CROSS REFERENCE because editorial review is needed to determine the choice of the proper heading.

2C. The CTH must have a Qualifier containing one of these terms: analysis, biological studies, occurrence, preparation, properties, reactions, and uses and miscellaneous. Some additional subdivisions are coded 2C—"dilute", "ice", "vapor", "vitreous", and "molten"—for reasons discussed in Heading Subdivision Control, the next section of this paper. The 2C code triggers the diagnostic QUALIFIER NON-MATCH if the data in the Qualifier do not match with one of the foregoing terms.

2F. This CTH must have a Qualifier. This code triggers the diagnostic QUALIFIER NON-MATCH due to absence of the Qualifier or non-match of data in the Qualifier with a CTH-QLF combination on the General Subject Edit Control List 2 (Table II).

2H. The CTH associated with this code may require a Homograph. Absence of a Homograph Definition triggers the diagnostic HOMOGRAPH NON-MATCH. The diagnostic also appears if the Homograph Definition associated with the specific CTH does not match that CTH–HOM combination on the General Subject Edit Control List 2.

2V. This code is used for common taxonomic headings

Table III. Chemical Abstracts Service General Subject Edit Control List 3

BATCH	GRP	UDO	GRP	QUALIFIER	GRP	CATEGORY
4696					84	acetals
4696	1	2C	2	analysis		
4696					85	anhydrides
4696					86	anhydrosulfides
4696			15	animal		
4696			16	anterior lobe		
4696			17	anthraquinone		
4696			18	artificial		
4696			19	azo		
4696			20	band structure		
4696					87	base
4696			21	biliary		
4696			22	biological		
4696			23	biological effects		
4696	3	2C	4	biological studies		
4696			24	blast		
4696			25	blood		
4696			26	blood plasma		
4696			27	blood serum		
4696			28	brown		
4696			29	cellular		
4696			30	chemical and physical effects		
4696			31	color		
4696			32	column and liquid		
4696			33	composition		
4696					88	compounds
4696			34	conduction		
4696			35	cyanine		
4696					89	derivatives (general)
4696			37	diesel		
4696	114	2C	36	dilute		
4696			38	disease or disorder		
4696			39	electric		
4696			41	electrochemical		
4696			42	electron		
4696			40	electro-		
4696			43	elucidated		
4696			44	environmental		
4696					90	esters
4696					91	ethers
4696			45	extruded		
4696			46	Fermi		
4696			47	film		
4696			110	flow		
4696			48	fuel		
4696			49	gas		
4696			50	gel		
4696			51	genetic		
4696			52	ground state		
4696					92	hydrazides

and permits the terms to be included in the Index Heading Dictionary so that these common terms may be checked by the spelling correction algorithm.

9S. This CTH is a taxonomic heading and has no associated diagnostic. Because of the large number of infrequently occurring taxonomic terms on file, terms with this code are excluded from the dictionary for the spelling correction algorithm as its effectiveness decreases with increasing dictionary size.

Edit Routines. The General Subject Index headings are validated by using a direct comparison procedure of input heading term against the control file. When the input term matches the file (i.e., the input form is the CA approved form of the heading), the input term is forwarded to the next stage of the program. At this point, about 90% of the input Concept Headings match the control file. Examples of input Concept Headings in approved form are: "Oxidation", "Oxidation catalysts", "Alcohols", "Acid-base equilibrium", "Geochemistry", "Sulfonation", "Pine", "Bacillus fructosus", and "Heat of nitration". Those terms that do not match the file directly are processed through a cross-reference routine in which the input form is changed to the CA approved form of the heading, a singular/plural algorithm, and then through the spelling correction algorithm, details of which are given below. Some of the computer edit processing steps are shown in Figure 2.

In the cross-reference routine, input of a synonym of a heading results in the automatic application of a cross-reference by which the input term is replaced with the CA approved index heading. This application is limited at present to the control of the Concept Heading data element. Examples of this type of change are "Olefins" changed to "Alkenes", "Brake" to "Bracken", "Kinetics of pyrolysis" to "Kinetics of thermal decomposition", and "Ova" to "Egg". Moreover, this application is extended to cover combined headings of the type "Sound and Ultrasound" where related ideas are collected under one primary access point. Some authors do not clearly specify some of the secondary details of a study. For instance, it is not always clear whether a study of sound is in the sonic or ultrasonic region. Therefore, both types of studies are collected under the one combined heading. Thus, either of the input terms "Sound" or "Ultrasound" is changed to the heading "Sound and Ultrasound" through a match with the control file. Studies on acoustics are also listed under and changed by the edit routines to the combined heading "Sound and

Input term	Output term	Diagnostic	Action by Analyst
Arteries	Artery		
Acid	Acids		None needed.
Sound	Sound and Ultrasound		
Ultrasound	Sound and Ultrasound		
Molecular heat	Heat capacity		
Alkyl compounds		APPLY CROSS REFERENCE	Check Index Guide for approved heading and for possible need to change index entry modification or heading subdivisions.
Viscosity	Viscosity	GENERAL HEADING	Check Index Guide for heading content restriction.
Brain	Brain	MISSING FRAGMENTATION	Add Qualifier if necessary.
Amines	Amines	MISSING CATEGORIZATION	Add Qualifier or Functional Category if necessary.
Acids, property	Acids, property	QUALIFIER NON MATCH QLF	Correct the Qualifier to properties.
Cell membrain	Cell membrane	CHECK FOR VALID CHANGE	Verify change.

Figure 2. Computer edit processing.

Ultrasound". Examples of similar changes are the input terms "Aroma", "Odor", or "Odorous substances" changed to "Odor and Odorous substances"; "Phosphate rock" or "Phosphorite" to "Phosphate rock and Phosphorite"; "Molding of plastics" or "Molding of rubbers" to "Molding of plastics and rubbers". About 5.5% of the input terms are changed to the approved CA index headings by this means under current operating conditions.

The operation of a singular–plural algorithm is included in the edit routine to minimize the size of the control file. The algorithm has only the output form of the index heading on file, yet it permits the input of either the singular or plural form. However, the algorithm is limited to headings whose plural form differs from the singular form only by the letter "s". The algorithm resolves the singular–plural difference in single word headings and in the final word of a multiword Concept Heading. Thus, the algorithm changes the input heading "Acid" to "Acids". Examples of other changes are "Acrylic polymer" to "Acrylic polymers" and "Hydroxyl groups" to "Hydroxyl group". The algorithm allows the input of either the singular or plural form of any component of a combined heading with the result that the approved heading is printed out; for example, the input of "Conformation", "Conformations", "Conformer", or "Conformers" results in the output of the index heading "Conformation and Conformers". About 1.9% of the input headings are corrected by the singular–plural algorithm under current operating conditions.

When there are differences other than the letter "s" between the singular and plural forms, e.g., -es and -y vs. -ies, then both the singular and plural forms of the index heading must be included in the control file, and the input form is corrected by associating the less preferred term with the approved CA index term. The computer process used for this operation is the same as that used for automatically applying the cross-references or expansions discussed in the preceding paragraph. The following changes illustrate this control procedure: input of "Appendixes" is changed to "Appendix", "Arteries" to "Artery", "Cacti" or "Cactuses" to "Cactus", and "Knife" to "Knives".

The diagnostic APPLY CROSS REFERENCE is applied to about 0.2% of the input headings that are not CA approved index headings but are included in the control file because of their potential occurrence. Since the headings are not synonyms or near synonyms of approved CA index headings, intellectual review of the input form is needed to determine the approved index heading. Manual reference to the Index Guide[7] is necessary to determine which approved index heading should be used. The term "Association" is cross-referred to one of four possible headings, namely, "Heat of association", "Ions in liquids", "Ions in solids", or "Molecular association". Thus, if the term "Association" is input as the index heading, the edit routines, when referencing the control file, add the APPLY CROSS REFERENCE diagnostic to the index entry work unit. In reviewing the work unit, the editorial analyst determines which of the headings of the cross-reference is the one which most accurately covers the indexable idea and then corrects the index record. The computer record is then corrected in the recycle operation. Input of "Alkanesulfonic acids" illustrates another use of this diagnostic. Because the composite term is distributed between two data elements, there is need of editorial interface to transfer the information to the proper data elements. The portion "Sulfonic acids" belongs in the CTH data element and "alkane" in the index entry modification. The edit procedures are not programmed to perform this type of change because of the variety of output requirements. In this example, the editorial analyst determines the proper data element assignment and also determines which Qualifier, of those coded 2C, is needed in the index entry record.

After the above operations are performed, approximately 2.7% of the input Concept Headings do not match the control file directly. These "no match" terms are routed through the spelling correction algorithm discussed below. The spelling correction algorithm corrects a portion of the "no match" terms equivalent to 0.6% of the initial input. The remaining 2.1% of the input Concept Headings do not match the control file and are assigned a NO MATCH diagnostic. Index entries carrying this diagnostic are reviewed by professional staff to determine the reason for the NO MATCH diagnostic. Three types of situations give rise to this diagnostic. One, the input Concept Heading term is new to CAS and does not occur on the Control File. If the new term is valid, as determined by editorial review, no change is made to the index record, and the new term is added to the file by scheduled, manually initiated update procedures. This procedure provides for new terminology to be added to the control file as it is encountered. The second reason for a NO MATCH diagnostic is misspelled input terms which are not changed by the spelling correction algorithm. These errors must be corrected in the recycle procedure. The third possible reason for the appearance of the diagnostic is the noncorrelation of the input terminology with the terminology on the Control File. This occurs infrequently because of the fairly high degree of controlled input of index entries by the professional staff who apply indexing policies to arrive at approved terminology. The edit routines are based on this type of input and detect the more trivial type of error which, if undetected and uncorrected, would cause misplacement of information in the index.

Heading Subdivision Control. The data in the Heading subdivisions associated with the Concept Heading (CTH) may be present in three different data elements, namely, the Functional Category (CAT), Qualifier (QLF), and Homograph Definition (HOM). These data are computer-edited to validate their correctness with respect to spelling and content and, in certain cases, to assure that the data are correctly associated with the CTH.

Some Concept Headings require a Homograph Definition in order for them to be defined unambiguously. These CTH's are designated by the UDO 2H on Control List 1 (Table I) as illustrated by "Mold" and "Molds". When the 2H code is present, the program validates the CTH, checks for a Homograph Definition in the work unit, and compares the contents of the two data elements against the

Table IV. Index Heading Dictionary

KEY	KEY SET	CONCEPT HEADING	STEM
abal		abalone	
abdo	{	abdomen	abdome
		abdominal diaphragm	abdomi
abla	{	ablation	ablatio
		ablative materials	ablativ
abom	{	abomasum	abom
abor	{	abortion	abor
abra	{	abrading apparatus	abrad
		abramis	abramis
		abramis brama	abramis[b]
		abrasives	abras
abri	{	abrins	abri

CTH–HOM combinations present on Control List 2 (Table II). The CTH–HOM combinations that are permissible are moved on to the next step of the program. The combinations not listed are assigned the diagnostic HOMOGRAPH NON-MATCH so that the index record will be reviewed and corrected by an editorial analyst. For example, the index record having the CTH "Mold" will receive the non-match diagnostic if the Homograph Definition is not present or if the content of the Homograph Definition is anything other than "fungus". These procedures not only check the content of the Concept Heading and the subdivision but validate only the specific combinations as noted in the control file. Thus, the combinations "CTH Mold HOM forms" or "CTH Molds HOM fungus" would get the non-match diagnostic. Since the singular form and the plural form of a Concept Heading may have different meanings, both forms are carried on the control files to avoid having one form changed to the other by the techniques discussed in the section on Edit Routines. Consequently, a high degree of control can be maintained on headings of this type.

A CTH on Control List 1 which has a UDO 2F must be associated with a Qualifier to complete the index heading. After the Concept Heading has been validated against Control List 1, the data in the index work unit are compared with the CTH–QLF combinations on Control List 2. If there is no match, the diagnostic QUALIFIER NON-MATCH is given to the index work unit. If the CTH–QLF does match, the work unit proceeds to the next step on the program. For example, either "chemical" or "biological" can be used as Qualifiers with the CTH "Alkylating agents". Thus, either of the combinations "CTH Alkylating agents QLF biological" or "CTH Alkylating agents QLF chemical" is valid. Any other combination will receive the non-match diagnostic. Any index work unit that has a CTH associated with the 2F code appearing without a Qualifier will receive the same diagnostic. In addition, to illustrate another control function of the edit routine, if "CTH Alkylating agents QLF chemical" is input, the diagnostic APPLY CROSS REFERENCE is added to the work unit due to the association of the 1X code with these terms on Control List 2. The 1X code is used because the concept "chemical alkylating agents" is indexed at the heading "Alkylation" rather than "CTH Alkylating agents QLF chemical". Since the rearrangement of the data to the correct data elements is not done by the program, the addition of the diagnostic alerts the editorial analyst to make the correction.

Another group of Concept Headings requires the presence of data in either the Functional Category or Qualifier data elements. These CTH's are coded 1Q on Control List 1. The permissible contents of the Functional Category are listed in the Index Guide Introduction under the descriptor "chemical functional subdivisions".[6] The "general subject subdivisions"[6] refer to elements permissible in the Qualifier data element. Seven Qualifiers that subdivide headings more specifically in the Chemical Substance Index and in the General Subject Index are coded 2C on Control List 3 (Table III). Other Qualifiers are coded 2C because they describe the type of study being performed on a particular chemical substance. The programs can edit these particular Qualifiers as components of index entries for the General Subject Index or for the Chemical Substance Index. All other Qualifiers in Control List 3, not associated with the descriptor 2C, may be used only with a Concept Heading. CTH–QLF links cannot be defined for machine editing of these remaining Qualifiers because they are not required to co-occur with a Concept Heading; i.e., the Concept Heading may stand alone or be associated with one of these Qualifiers. This means that in this group the Qualifier data, independent of the Concept Heading, are validated by comparison against the Qualifiers on Control List 3. If any of the input Qualifiers do not match the control file, the diagnostic QUALIFIER NON-MATCH is issued.

The contents of the Functional Category for index entries in the General Subject or the Chemical Substance Indexes are validated against the Categories on Control List 3. The diagnostic CATEGORY NON-MATCH is issued to those Functional Categories which do not match one of the entries on the Control List.

About 1.7% of the Functional Categories associated with General Subject index headings and 0.6% of the Functional Categories with Chemical Substance index headings receive the non-match diagnostic. This represents the detection and identification of about 650 errors per processing of one volume of index data. About 3.7% of the Qualifiers associated with General Subject index headings and 2.7% of the Qualifiers with Chemical Substance index headings receive the non-match diagnostic. This represents the detection and identification of about 5500 errors for both types of index entries per volume.

The program edits about 1000 CTH–HOM combinations per volume. Only a small number of these are found to be in error. Corrections to the data in these data elements are made manually through the recycle operation. Approximately 0.7% of all the input General Subject index entries acquire diagnostics associated with the foregoing subdivision data elements.

Spelling Correction Algorithm.[8] The purpose of the spelling correction module is to correct misspelled Concept Headings based on the dictionary of accepted headings, or to reject it as algorithmically uncorrectable.

The spelling correction algorithm is based on the keyed-stemmed Index Heading Dictionary (Table IV). This is a dictionary of alphabetically arranged accepted Concept Headings. The first four characters of each heading are the heading's key. For example, the key for the Concept Heading "abalone" is "abal". The four-letter key serves as an address or pointer to a group of headings whose first four letters are the same and are grouped in an assigned place in the file.

The Index Heading Dictionary is divided into subsets based on the keys of the headings. These subsets are called "key sets". For example, the following words form a key set since each heading has the same key "abra".

abrading apparatus
abramis
abramis brama
abrasives

The unique stem of a Concept Heading is the longest initial character string which distinguishes the heading from the heading immediately preceding it alphabetically and the heading immediately following it alphabetically. For example, of the three accepted Concept Headings:

abalone
↕
abdomen
↕
abdominal diaphragm

"abdomen" differs from "abalone" at "abd" and differs from "abdominal diaphragm" at "abdome". Thus, the unique stem from "abdomen", the longest initial character string which distinguishes it from the heading immediately before and immediately after it, is "abdome". (See Table IV for further examples of keys, unique stems, and key sets.)

Using the keyed-stemmed Index Heading Dictionary, built from the General Subject Edit Control List 1, the spelling correction module either corrects (i.e., the spelling is changed) or rejects input headings, which do not match a Concept Heading (CTH) on Control List 1. The flowchart for the spelling correction algorithm is given in Figure 3. The spelling correction algorithm considers the first four characters of the input heading as a key to access the key set of accepted keystemmed Concept Headings. If there is no key set having the same key as the input heading, the input heading is rejected as uncorrectable and the next heading to be edited is examined. If a key set does exist, each member of the key set is checked character by character for a match in the corresponding positions in the input heading. If a member of the key set is found whose stem matches that of the input heading, then the input heading is changed to this accepted Concept Heading provided not more than 20% of the characters after the stem of the dictionary Concept Heading differ from that of the input heading in a moving window three characters wide. For example

stem
Concept Heading: abdominal diaphragm
Input Heading: abdominal diafragm

the "f" in "diafragm" is not found in the same character position in "diaphragm" because a "p" exists there. Nor is the "f" found on either side of the "p" in "diaphragm". Thus, the "f" is not found in the three-character window "aph" of "diaphragm". The "r" in "diafragm" is not found in the same character position in "diaphragm" because an "h" exists there. However, there is an "r" in "diaphragm" one character to the right of the "h". Thus, the "r" is found in the three-character window "hra" in "diaphragm". Similarly, the "a", "g", "m" of "diafragm" are all found in the corresponding three-character window in "diaphragm". Only one character out of the thirteen (including the blank) after-stem characters is not found in a three-character window. This is less than 20% of the after-stem length, so "abdominal diafragm" is changed to "abdominal diaphragm".

If more than 20% of the after-stem characters of the Concept Heading differ from the input heading in a three-character window, the input heading is rejected as uncorrectable. There is one exception to this. If the input heading ends in a period, the routine assumes the input heading is an abbreviation. If the stems match, the input heading is changed to the Concept Heading instead of being rejected.

The cases cited above are applied to the input heading if a matching stem is found in the key set. If a matching stem is not found in the key set, the input heading is not summarily rejected as uncorrectable. Two "in-stem" changes are attempted if the input heading is the same character length as the Concept Heading in the key set, and the stem is at least eight characters long. The two changes attempted are:

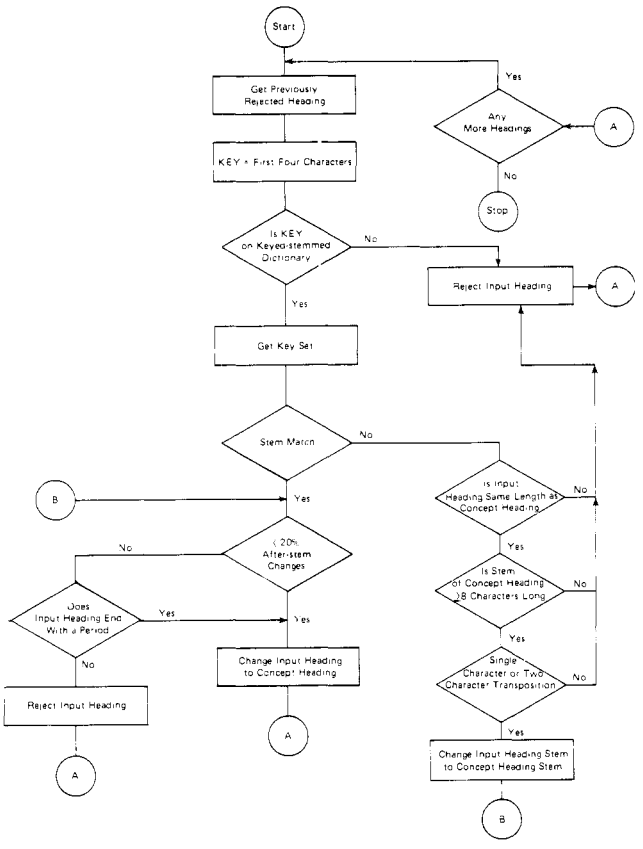

Figure 3. Spelling correction routine flowchart.

(1) two-character transposition changes or (2) one-character exact position changes. If the input heading stem differs from the Concept Heading stem by only one character in the same position (no three-character window is used) or if it differs only in a two-letter transposition, then the input heading stem is changed to the Concept Heading stem. From this point on, the input heading is processed as if the stems originally matched. For example

stem
Concept Heading: abramis brama (8 character stem)
Input Heading 1: abraims brama
Input Heading 2: abranis brama

The stem of Input Heading 1 differs from the stem of the Concept Heading in the two character transposition "im" for "mi". The input heading stem is then changed from "abraimsb" (b = blank) to "abramisb" and processing continues. The stem of Input Heading 2 differs from the stem of the Concept Heading by one character; "n" is used in the same character position as the "m" in the Concept Heading. Again, the input heading is changed to the Concept Heading and processing continues.

In summary, the spelling correction module makes (1) no changes to the key (the first four characters of the Concept Heading), (2) exact position changes and two-character transposition changes to the stem of the heading, (3) after-stem changes to characters only if the total number of characters involved in the change does not constitute more than 20% of the after-stem length, and (4) abbreviation expansions.

Examples of changes made by the algorithm are: "amin/acids" to "Amino acids", "Infrared spctra" to "Infrared spectra", "Aromatic compds." to "Aromatic compounds", "Rhizoctonia salani" to "Rhizoctonia solani", "Magneotstriction" to "Magnetostriction", "Blood group substances"

to "Blood-group substances", "Cell membrain" to "Cell membrane", "Vetcch" to "Vetch", "Prolamines" to "Prolamins", "Red spidermite" to "Red spider mite", and "Fungiside" to "Fungicides". Since the spelling changes are made on a statistical basis and by comparison of the input term with terms from a dictionary, a slight probability exists that a correct input term can be changed to an inappropriate term. Improper changes that could occur are: "Microtus nivalis" changed to "Microtus arvalis" and "Kinetics of protenation" to "Kinetics of bromination". In the first case, since "Microtus nivalis", a correct form, is not in the control file and consequently not in the keyed-stemmed Index Heading Dictionary, the input heading was changed to "Microtus arvalis" as this heading was in the dictionary. To prevent this same operation from occurring again, the correct term is added to the control file. In the second instance, the program in searching for the incorrectly spelled "Kinetics of protenation" on the alphabetically sorted keyed-stemmed Index Heading Dictionary encountered "Kinetics of bromination" first and thus changed the heading to this term. The program never reached the intended heading "Kinetics of protonation" in the dictionary and as a result an improper change was made. Before sending this through the recycle operation, the analyst will change bromination to the proper term with the correct spelling. A safeguard is provided by printing both the input form of the heading and the output form along with the diagnostic CHECK FOR VALID CHANGE in the index work unit copy so the editorial analyst can determine if the output heading is correct. If the change made by the algorithm is incorrect, the original heading in its correct form is reinput during the recycle. However, very few recycles have to be made to the data since less than 2% of the changes made by the algorithm are incorrect. Under present operating procedures, 21.0% of the Concept Headings edited by the algorithm are changed to the correct heading form and the remaining 79% are reviewed by editorial analysts because of the presence of the NO MATCH diagnostic.

Benefits. The major benefit of these edits is controlling, with minimum human effort, the accuracy of the General Subject Index headings with respect to content and CAS editorial policy. Since the edit programs locate and identify the type of error by diagnostics, the editorial analyst needs only to correct the headings that have been flagged instead of having to examine all entries, locate those in error, and then make appropriate corrections as under the former system. The reduction in editorial effort is shown by the following statistics illustrated in Figure 4: (1) 89.9% of the input Concept Headings currently match the control file, (2) 1.9% of the input Concept Headings are automatically corrected for singular–plural problems, and (3) 5.5% of the input headings are corrected by application of cross-references, automatic expansions to the full headings, and special singular–plural corrections. Only the 2.1% of the total input CTH's which do not match the control file and the 0.6% of the headings changed by the spelling correction algorithm need to be reviewed by an editorial analyst after these edit procedures. The automatic corrections also aid the editorial analysts in the preparation of index entries since the analysts no longer need to be concerned with the format details of multiword headings or the singular or plural form of the index heading. The manual effort for editing the General Subject index headings has been reduced over 97%.

Current volumes of the General Subject Index contain about 330,000 index entries, of which approximately 24,000 of the input Concept Headings are automatically corrected, 2,000 are changed by the spelling correction algorithm and about 7,000 receive the NO MATCH diagnostic. The diagnostics, APPLY CROSS REFERENCE, HEADING NOT CATEGORIZED OR QUALIFIED, MISSING CATEGORIZATION and MISSING FRAGMENTATION, appear

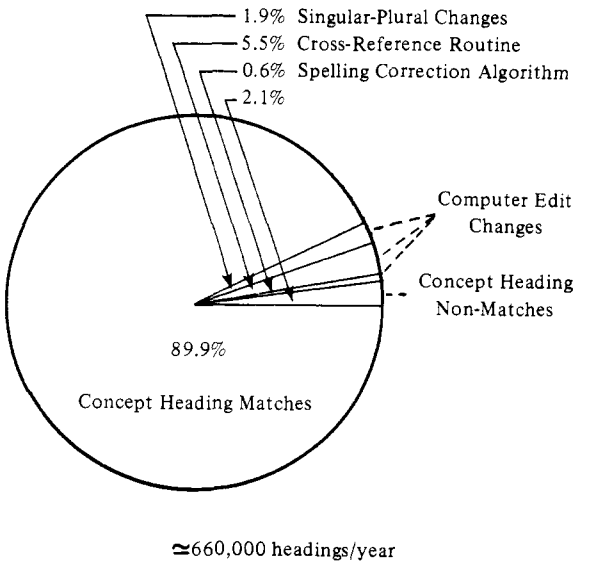

Figure 4. Concept heading control.

with about 0.7% of the input index entries after the Concept Headings have been processed against Control List 1. As a result, about 2,300 entries in the current General Subject Index had to be recycled for corrections of these additional diagnostics. Additionally, 0.7% of the entries received the diagnostics QUALIFIER NON-MATCH, CATEGORY NON-MATCH, or HOMOGRAPH NON-MATCH. Just 4.1% of the input index work units acquire diagnostics which have to be reviewed by an editorial analyst. Thus, the edits on the primary access point, the Concept Heading, along with those on the subdivisions assure control of most of the General Subject Index headings. Only entries with an error need to be recycled for correction. Since the data passing through the recycle operation are not again processed through the edit routines, some errors do appear in the final index. However, the number of such heading errors is small; only about 100 errors of the types discussed in this paper appear per Index volume.

In the past volumes, Subject Index galleys were prepared primarily to edit the context and content of the heading data. Since these edits are now performed by computer, the preparation and editing of General Subject Index galleys has been discontinued with attendant savings. These procedures also provide a shorter time lapse between the appearance of the edited General Subject Index and the last issue of the corresponding volume period.

The computer costs for performing heading data edits for an index volume, which amount to approximately 1¼ hr of IBM 370/168 time per volume, are significantly less than the costs of producing and editing the index galleys for one volume of the General Subject Index under the former manual editing system. Thus, not only are there manpower savings by utilizing the editing system but computer and material savings as well.

ACKNOWLEDGMENT

CAS is pleased to acknowledge the financial support from the National Science Foundation (Contract C656) in the development of the edit routines and their necessary control files discussed in the above report.

LITERATURE CITED

(1) Morgan, H. L., "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service," *J. Chem. Doc.*, **5**, 107–13 (1965); Rowlett, R. J., Jr., and Weis-

gerber, D. W., "Handling Commercial Product Names at Chemical Abstracts Service," *J. Chem. Doc.*, **14**, 92–5 (1974).
(2) See CA Volume 76 Index Guide (1972), Section II, paragraphs 9–10, pp 5I–9I, and Section III, pp 13I–19I for a discussion of the CA General Subject Index.
(3) Data element statistics for Volume 71 are presented by Zipperer, W. C., Stearns, R. E., Jr., and Park, M. L., "The Integrated Subject File. I. Data Base Characteristics," *J. Chem. Doc.*, **13**, 92–8 (1973).
(4) Lancaster, F. W., "Vocabulary Control for Information Retrieval," Information Resources Press, Washington, D. C., 1972, Chapter 18, p 167.
(5) The Control File is a component of the Chemical Abstracts Service's User Aid package, which provides assistance to the chemist and information scientist in the form of 13 Search Aids to be used in accessing the various CAS products and files. These Search Aids are described in "Chemical Abstracts Service Search Aids for the 9th Collective Index Period (1972–1976)" (International Standard Book Number: 8412-0198-6, Library of Congress Number: 74-80986), June 1974, which is available by contacting the Marketing Department of Chemical Abstracts Service.
(6) See CA Volume 76 Index Guide (1972), Section II, paragraph 10, p 8I.
(7) Reference 6, Section I, paragraph 6, p 2I.
(8) Other programs for detecting and correcting spelling errors are described by (a) Alberga, C. N., "String Similarity and Misspellings," *Commun. ACM*, **10**, 302–13 (1967); (b) Blair, C. R., "A Program for Correcting Spelling Errors," *Inf. Control*, **3**, 60–7 (1960); (c) Damerau, F. J., "A Technique for Computer Detection and Correction of Spelling Errors," *Commun. ACM*, **7**, 171–6 (1964); (d) Davidson, L., "Retrieval of Misspelled Names in an Airline Passenger Record System," *Commun. ACM*, **5**, 169–71 (1962).

Polymer Nomenclature, Classification, and Retrieval in the Du Pont Central Report Index†

JOHN L. SCHULTZ

Central Report Index, Information Systems Department, E. I. du Pont de Nemours and Company, Inc., Wilmington, Delaware 19898

Received March 14, 1975

A comprehensive nomenclature system for polymers has been devised in which carbon–carbon addition polymers and polymers of unknown structure are named from their starting monomers; polyamides, polyesters, and polyurethanes from prescribed monomers; and all other polymers of known structure from their structural repeating units (SRU's). Rules are given for aftertreated, graft, and chain-extended polymers. Polymers are classified, for generic retrieval, according to features obvious from their names: each polymer is posted in the Descriptor File under registry-number descriptors corresponding to the registry numbers of its monomers (including artificial monomers derived from SRU's), and a small number of class descriptors are used for overall features of the polymer. Correct registration of polymers is aided by lookup and matching of names and molecular formulas. The molecular formula of a polymer is generated by computer as the sum of the molecular formulas of its monomers. Name and molecular formula lookup are verified by a computer check on the uniqueness of descriptor combinations. Families of polymers are retrieved by searching the Descriptor File. The distinctive feature of the search system is the retrieval of polymer structural details through monomer structural details. A family of monomers can be retrieved by searching the Descriptor and/or Chemical Topology Files for any specified features. The registry numbers so retrieved are translated into the corresponding registry-number-descriptors and automatically resubmitted to the Descriptor File to retrieve the polymers posted under them.

INTRODUCTION

The Du Pont Central Report Index and its handling of chemical information have been described in earlier papers.[3,4] Briefly, the Central Report Index indexes and retrieves the information contained in Du Pont proprietary documents. These functions are carried on by information chemists. Documents are indexed under chemical terms and general terms. Chemical terms denote individual chemical compounds and related concepts, incuding polymers; general terms denote all other concepts. Each general term is the name of a concept, e.g., OXIDATION; each chemical term is a seven-digit alphanumeric registry number assigned sequentially to a compound when it first enters the system. The registry number ties together a series of files, used for identification, document referencing, and classification of compounds. This is shown schematically in Figure 1; for details see an earlier paper.[4] All these files, except the Molecular Formula Card File, are computerized on the IBM 360 and are therefore interlinked to perform a variety of storage and retrieval functions. Currently the files cover 73,000 documents with 116,000 chemical terms: 94,000 for nonpolymers and 22,000 for polymers. These 22,000 polymers have been encountered in the indexing of such major technologies as elastomers, films, plastics, and synthetic fibers over the last 25 years.

DISTINGUISHING AMONG "DIFFERENT" POLYMERS

A polymer is not an individual chemical compound but rather a collection of compounds differing in such chemical properties as molecular weight, linearity, and sequence of structural repeating units (SRU's). This raises the problem of how and in what detail to distinguish between one poly-

† Presented before the Division of Chemical Literature, 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 8, 1975.