Author Name Processing at Chemical Abstracts Service: Name Matching Using Nonunique Bibliographic Identifiers

KATHRYN M. SOUKUP* and SILAS E. HAMMOND

Chemical Abstracts Service, Columbus, Ohio 43210

Received December 9, 1981

For elimination of much of the labor-intensive and repetitious manual editing of author names for the Chemical Abstracts (CA) Volume and Collective Author Indexes, a computer-based system was developed at Chemical Abstracts Service (CAS) which matches author names by using the nonunique bibliographic identifiers available in the literature. The computer system closely emulates the intellectual editing process previously performed and also provides new data edits not possible with the manual system. Called the Author Index Manufacturing System (AIMS), this system compares personal author names recently published in CA with names already present in a master file containing five years of author data. AIMS determines if a recently published name has been previously recorded on the master file. If it has been, AIMS automatically edits (expands) either the incoming name or the file name to contain the most complete information available. This paper describes the design and development of AIMS, the "name upgrading" process, and the impact of AIMS' installation on CA Author Index production.

INTRODUCTION

Could CAS "register" authors as it does chemical substances?

This was one of the questions asked early in the design of the Author Index Manufacturing System (AIMS). To make the CA Author Indexes useful to a researcher using an author name as a reference point when gathering data about a particular chemical substance or concept, CAS has always attempted to collect papers published by the same author under one listing in the CA Volume and Collective Author Indexes. If an author could be uniquely "registered" in some way, no matter how the author's name appeared in a paper, all papers by the author could automatically be collected in one place in the Author Indexes. Such an automated system could not only improve the quality of the indexes but could also save CAS much of the manual effort being expended in identifying all of an author's papers and seeing that the papers were collected at one access point in the Author Indexes.

After investigation, it turned out that, unlike substances, CAS could not uniquely identify (i.e., "register") each author for the documents abstracted by CAS. A lack of available unique author identifiers in the literature was the primary reason registration was not possible. Despite this, CAS did develop a computer system using available nonunique bibliographic identifiers to automate author name processing substantially and to facilitate the collection of all the papers published by an author at a single access point in the Author Indexes.

CAS was not the first organization to investigate the "name matching" process. Work in the literature describes personal name lookup techniques as applied to various data bases.[1-3] The described methodologies compress or concatenate characters to key sets of characters and work best on data bases smaller than CA's author files, which contain over 5.5 million names. Western Airlines uses a surname-match procedure (algorithm) for checking flight reservations.[4] In checking a reservation, if the name is not found on the passenger list for a flight, the algorithm retrieves potential name matches from the passenger list. In this way, certain misspellings of the name can also be retrieved. The encoding technique, which compresses surnames by eliminating vowels and other characters, is not applicable for use in large collections of information (data bases) since it would result in many matches. For example, if CAS used this technique, some author names could have over a million "matches". Other work has also been done on personal name matching. A technical report to the National Bureau of Standards on "Accessing Individual Records from Personal Data Files Using Non-Unique Identifiers" is a survey of name-match methods for use by government agencies.[4] The report looks at the problems caused by the Privacy Act of 1974, which places restrictions on the federal, state, and local governments' use of the social security number as a personal identifier. For some government agencies, compliance with the Privacy Act involved changes in implementation of their retrieval algorithm. The report describes other methodologies applicable to the problem of retrieving individual records by using nonunique identifiers.

For CAS processing, if authors supplied a social security number or some other unique identifier with each paper, the name-match process would be significantly easier. Since CAS has access to only certain bibliographic data—those supplied in the literature which are cited in the CA abstract heading—to use in its author name-match process, the CAS research team looked at methods to automate the name-match process by using the available nonunique bibliographic identifiers. The results of this investigation led to the development of the AIMS computer system.[5]

AUTHOR NAME EDITING AT CAS

Before explaining the name-match process used in AIMS, we will discuss the CAS author name-editing philosophy and process.

In each CA issue, CAS attempts to publish author names in a format as close as possible to the format used in the literature. The names may be incomplete or complete depending on how much information about the name appears in the literature. For example, in the literature an author name may be incomplete, i.e., the first or subsequent name(s) may be missing or present only in the form of initials. In these cases, only initials or perhaps a single forename will appear in the bibliographic data for the abstract in CA. If a complete author name appears in the literature, i.e., the surname and first and subsequent names are present, then the full name will be published in CA.

For the weekly CA issue author index, however, only the author surname, initials, and CA Abstract Number (CAN) are published. The purpose of this index is simply to direct the CA user to the abstract of the author's paper in that issue of CA.

* To whom correspondence should be addressed at Chemical Abstracts Service, P.O. Box 3012, Columbus, OH 43210.

0095-2338/82/1622-0172$01.25/0

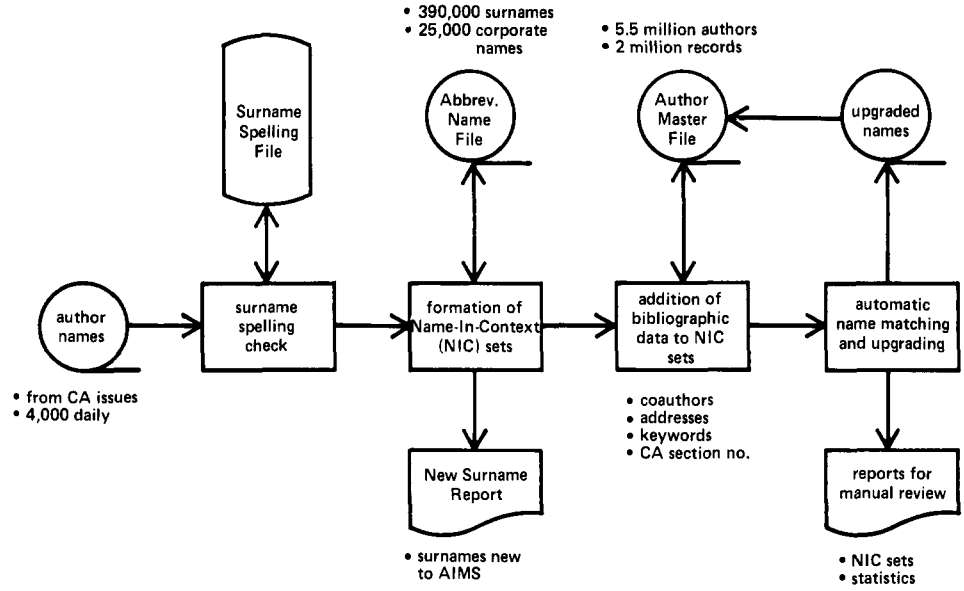

Figure 1. Programs and files processed in AIMS.

In the Volume Author Index, which contains author name data from 26 CA issues, author names are edited and published in as complete a form as possible with the titles of their papers and patents. In this index, the names are expanded to their fullest—even if the authors' full names do not appear in the literature—to differentiate clearly between individuals and to ensure that all papers and patents associated with a specific individual or corporation are indexed at the same point.

CAS has always gone to considerable lengths to obtain the full names of authors for the CA Volume Author Index and the 5-year cumulative CA Collective Author Index so that the citations are as accurate and complete as possible.

As noted, personal author names are recorded initially in the CAS information base in the form in which they appear in the primary literature covered by CA. This means that over a period of time an author's name may be recorded in a variety of forms, e.g.: Brown, J.; Brown, J. A.; Brown, J. Allen; Brown, James; Brown, James A.; Brown, James Allen. All of the above names may not refer to the same person. In the past, the task of determining if a collection of names all referred to the same person and, if so, expanding the initials to full names required extensive manual effort. The task of manually reviewing author names and finding the information necessary to expand those names that were not complete enough to identify the author unambiguously added 3 months to the time required to produce a Volume Author Index and required nearly 2 years to produce a Collective Author Index. In reviewing the author names, the editors assigned to this task used the following data for determining if authors such as J. Brown and James Brown were the same person: author addresses, coauthors, subject of work (titles of the papers). If enough of the data were the same, i.e., a name match had occurred, the editors would expand J. Brown to James Brown. The process of expanding initials to full first names in a personal author name is called "name upgrading".

For the CA Ninth Collective indexing period (1972–1976), more than 4.5 million author names were reviewed during production of the ten 6-month CA Volume Author Indexes, and almost 2.5 million of these names were reviewed again for the production of the 10-volume Collective Index. This massive task required more than 75 000 h of human effort over the 5-year period.

AIMS PROCESSING: OVERVIEW

AIMS edits author names using an algorithm that closely emulates the intellectual editing methods that were used to compare incoming names with names already on file and then automatically upgrades author names when possible. The use of AIMS allows human intellectual effort to be more properly directed toward solving problems the computer is unable to handle.

Names coming into AIMS from the CA issues are processed through a series of programs and files that are summarized by the flow chart shown in Figure 1.

The Surname Spelling File contains approximately 390 000 edited personal author surnames and 25 000 edited corporate author names extracted from the most recent 5 years of CA Author Indexes. During daily processing of primary documents to be abstracted by CA, every incoming personal author surname and corporate name is checked by the computer against this file before being published in CA. All surnames and corporate names that match the file are forwarded to the CA issues. If an exact match is not found, the surname or corporate name is printed out for intellectual review and correction, if necessary. This human/machine review and correction process improves the accuracy of the personal surnames and corporate names that appear in the CA issues and also reduces the number of names that must receive intellectual review for the CA Volume Author Indexes. Of the 4000 incoming surnames and corporate names processed daily, less than 7% do not match names already on this file. Of these, less than 1% are the result of input errors. The other 6% are correct as input but are new to the file.

The Abbreviated Name File contains the surname and first two initials for each of the approximately 5.5 million authors in the Author Master File. To begin the name-upgrading process, a newly recorded personal author name is compared with names on the Abbreviated Name File so that a set of potential name matches—called a Name-In-Context (NIC) set—may be compiled. An NIC set contains an average of ten potentially matching author names. When potential matches are obtained, the Abbreviated Name File provides links to the correlative bibliographic information for the author names located in the Author Master File. The computer system then retrieves the related bibliographic information (e.g., coauthors, addresses, etc.) from the Author Master File, adds it to the NIC set, and forwards the data to the Author Name Upgrader program. Surnames new to AIMS are printed for manual review. Each record in the Abbreviated Name File contains all the authors sharing the same surname and initials, e.g., all of the J. A. Brown's are contained in a

single record. The Abbreviated Name File contains 875 000 records, occupies 200 million bytes, and resides on four reels of magnetic tape.

In addition to author names, the Author Master File contains pertinent bibliographic information for the 2.2 million documents abstracted in CA over the past 5 years. For each document, the file contains the author(s) names, the title of the paper, the name of the city where the work was done, the CODEN for the primary publication, the CA section number, and subject-related keywords (assigned by CAS). This information is used to ensure highly accurate matching of author names during the name-upgrading process. New information is added and old is deleted on a regular basis so that the file is both current and stable in size. Each record in the file contains all the information for one document. The Author Master File contains 2.2 million records, occupies 700 million bytes, and resides on 15 reels of magnetic tape.

The Author Name Upgrader program is the heart of AIMS. This program compares the incoming author name and its related bibliographic information with the edited names and related information extracted from the Author Master File. To minimize computer storage space and facilitate processing, the bibliographic data used in name matching is put into hash-coded form. The program then applies a specified set of criteria for determining whether an incoming author is already present on the master file. If a sufficient number of the comparison criteria are met and if the incoming author name is less complete than the master file name, it is upgraded to the more complete form of the name present on the Author Master File. If a match occurs and the incoming name is more complete than the name on file, the file name is upgraded. If the incoming name and related information cannot be matched adequately with a name on file, the incoming name and its potentially related names are printed for intellectual review.

Another feature of AIMS is its automatic flagging of potential cross-references. Author surnames for which cross-references may need to be created are printed on the Cross-Reference Worksheet. Typically the potential cross-references are compound surnames. The Author Index editors review the worksheet and determine which names need to have cross-references created for them. The cross-references created for the Volume Author Index are automatically formatted in the index with the phrase "See also". For example, in the case of the compound surname Garcia-Rodriguez, the Author Index will have the cross-reference "See also Garcia-Rodriguez" listed under the surname Rodriguez.

Because personal author names published in Slavic language literature are frequently recorded in abbreviated form (surname and initials) these names are not processed by the Author Name Upgrader. The presence of USSR in the country data element or Russian, Latvian, Lithuanian, Estonian, or Bulgarian in the language data element in the record for the document prevents the author name from going through the upgrade process.

Corporate author names are also prevented from being upgraded.

Four times during a 6-month CA volume period magnetic tapes used to produce the CA issues are merged and input to AIMS for selection of the author names and their related bibliographic data.

With each merge, the author names are processed in an AIMS production run. During each production run, AIMS upgrades author names when possible and prints author name data needing manual review. The system also generates statistical reports on the number of names processed and the results of the Author Name Upgrader program. A final production run is made at the end of each volume to format the author names for printing in the Volume Author Index.

Table I. Data Used for Author Name Matching

location of work (city)	CA section number
coauthor(s)	first name
keywords	initial(s)
CODEN	

Table II. Table of Name-Match Conditions

(second initial) + (location of work) + (one keyword)
(second initial) + (three or more keywords)
(first name) + (location of work) + (one keyword)
(first name) + (three or more keywords)
(location of work) + (three or more keywords)
(one coauthor) + (one keyword)
(one coauthor) + (second initial)
(one coauthor) + (location of work)
(one coauthor) + (first name)
(one coauthor) + (CODEN)
(one coauthor) + (CA section number)
two or more coauthors
(second initial) + (third initial) + (location of work)
(second initial) + (third initial) + (one keyword)

Production of each Volume Author Index via AIMS requires about 20 CPU hours on an IBM 370/168 computer.

AUTHOR NAME UPGRADER

Development. The computer process called the "Author Name Upgrader" is based on the logic and operations performed in the manual author name-editing process. In the manual process, Author Index editors determine that two names such as J. A. Brown and James A. Brown belong to the same person by finding similarities in the bibliographic data related to the two names. When enough similarities are present, the names are considered to match. The manual process is based on factual data, experience, and even intuition. In the design of the computer process, what would be considered "enough" data similarities to effect an accurate match?

To answer this question, the research team first looked at the data available for name matching. The bibliographic data available are those items supplied in the literature and identified by CA in the heading for each abstract. In addition, other data assigned by CA (keywords, CA section number) are available for name matching. Table I lists the types of data that may be used in author name matching.

After listing the data that could be used for author name matching and reviewing the manual editing process, the research team developed a set of data combinations which must be present for a match to occur. Each combination of data is called a name-match condition. Table II lists the set of name-match conditions which are applied when two authors share a similar surname and first initial to determine if the authors are the same person.

Author names matching on any one of the conditions listed in Table II are assumed to be the same person. For example, two authors having the same surname and first initial match by sharing the same second initial, location, and one CA keyword. Since past experience revealed that a "coauthor match" was one of the strongest indicators of a probable author match, it is featured prominently in the table of name-match conditions.

Matching authors are then upgraded by the computer if one author name contains more complete information than the other.

The name-match conditions were tested on a sample of data to see if they produced valid author name matches while also not incorrectly matching authors. Since they did, the Author Name Upgrader was developed by incorporating these conditions.

Since there is a higher probability of similar surname and initial combinations for frequently occurring surnames (e.g.,

Table III

computer upgrades	12 721
manual upgrades	11 134
upgrades made manually and by the computer which were the same	9 046
upgrades made manually and by the computer but which were not the same	811
manual upgrades not made by the computer	1 277
computer upgrades not made manually	2 864
problem upgrades made by the computer	183

Smith), additional name-matching data were required for these surnames before computer matching and upgrading would take place. AIMS applies the more stringent name-match conditions to surnames found in a table of "common surnames", which is part of the Author Name Upgrader program. If the authors have a common surname, it must match on one of the conditions listed in Table II plus one additional piece of data from the following list: coauthor, three keywords, location of work. For example, under the condition

(second initial) + (location of work) + (one keyword)

the following match conditions are valid for common surnames:

(second initial) + (location of work) + (one keyword) + (one coauthor)

(second initial) + (location of work) + (one keyword) + (three keywords)

and so on for all of the name-match conditions. Common names will also match on the following set of special conditions:

(first name) + (second initial) + (location of work) + (one keyword)

(first name) + (second initial) + (one coauthor)

(first name) + (second initial) + (three or more keywords)

Common surnames are expected to have full first and middle names after upgrading. If they do not, they are printed for review. Manual effort is still expended in corresponding with some of these authors to request their full first and middle names.

For testing of the name-match conditions on a large scale, the two major master files in the system, the Author Master File and the Abbreviated Name File, were built by using the data for approximately 1.4 million documents and 4 million authors.

Testing. In checking the name-match conditions on a large scale, a series of tests was set up to compare the computer upgrades to those done manually by the Author Index editors. In each phase of testing, a selected volume of data was processed by the Author Index editors and also by the computer system. The results were compared and modifications made to the computer system as needed. One typical test consisted of data from eight CA issues and gave the results shown in Table III.

An analysis of test results showed that of 11 134 manual upgrades, the computer was able to duplicate 9046 exactly (81%). Upgrades made manually which the computer could not duplicate were those generated from detected spelling errors, direct correspondence with the authors, and human intuition. On the other hand, owing to the computer's superior ability to retain and compare information, the computer was able to generate 2864 upgrades which the human editors could not. However, not all of the computer upgrades were valid. Several problem areas were identified.

It was found that about 1.4% of the computer upgrades were in error. In cases where the incoming name potentially matched two or more names on file, AIMS sometimes chose the wrong name to match. For example, husband and wife coauthors having the same first initials, using only those initials, and working in the same field of chemistry were sometimes upgraded incorrectly by AIMS. Because of these problem upgrades, a special report was created which listed authors who had two or more equally likely, but differing, upgrades. In production, names on this report are not upgraded and are routinely printed for intellectual review.

In the test runs, the results of computer upgrading were not reviewed by the human editors. In production runs, some computer upgrades and all upgrade problems are printed for review. On the basis of this review, the Author Index editors generate keyboard updates to the system. A single keyboard update may result in many computer upgrades. The capability to recycle upgrades is another advantage AIMS has over manual editing.

Results. The objectives of AIMS were to improve the quality and currency of the Author Indexes and to reduce manufacturing costs by eliminating most of the preproduction manual editing.

AIMS improved the quality of the Author Indexes by increasing the accuracy of author name upgrading. The increased accuracy results in less scattering of equivalent author names in the Author Index.

AIMS reduced the time required to produce a Volume Author Index by 2 months. A significant reduction in labor costs and processing time will come in 1982 at the end of the current collective indexing period when the Tenth Collective Author Index containing over 6 million author names (24 000 pages) will be produced with no additional editing time. It previously took 2 years to edit author names for a Collective Author Index.

Comparisons of AIMS editing with the former manual editing process have shown that the use of AIMS produced a reduction of three staff members (from eleven to eight Author Index editors). As a result of system installation, AIMS automated the processing for 90% of the author names for each CA volume. The remaining 10% of the author names are printed for manual review. The names printed are ones that cannot be name matched or ones which name match two or more differing author names and therefore require manual review and resolution. Common surnames not having full first and middle names are also printed for review.

For a typical CA volume, approximately 550 000 author names are processed.

Statistics from production runs show the following profile of the author names being processed: 65% of the names match the master file as input, 6% of the surnames are new, 22% of the names are upgraded by the computer, and 10% of the names are printed for manual review, giving a total of 103% due to overlap in statistics. (Some upgraded names are also printed for review, e.g., common surnames not having full first and middle names even after upgrading.)

For initial input, the percentage of author names matching the master file and the percentage of names new to the file have held relatively constant since the installation of AIMS. Currently of the 35% of the names not matching the file as initially input, 22% are being upgraded to match the file.

Since its installation, the performance of the system has steadily improved. The percentage of input names upgraded by the computer has risen from an initial 14% to the current 22%. This was expected, since the more information that is available on the master file, the more likely upgrades will be made on input names. As previously mentioned, AIMS will also upgrade "old" names—those already on the master file—if

the newly input names contain more complete information. Approximately 38 000 "old" names are currently upgraded per CA volume. The number of "old" upgrades has been decreasing as information on the master file becomes more complete.

Thus, the use of AIMS with its Author Name Upgrader has eliminated the repetitious manual editing for about 90% of the author names processed by CAS.

For others designing a name-matching system, it should be noted that the process described in this paper works better on large data bases where many nonunique bibliographic identifiers are available for matching, i.e., AIMS succeeds because of the nature and size of the data base. Also, the success of AIMS may be attributed in part to the fact that AIMS works with a specialized field (chemistry) and that all of the authors in the data base are already working in a common field.

REFERENCES AND NOTES

(1) Bookstein, Abraham "A Hybrid Access Method for Bibliographic Records". *J. Lib. Autom.* **1974**, *7* (2), 97–104.
(2) Fokker, Dirk W.; Lynch, Michael F. "Application of the Variety-Generator Approach to Searches of Personal Names in Bibliographic Data Bases. Part I. Microstructure of Personal Authors' Names". *J. Lib. Autom.* **1974**, *7* (2), 105–18.
(3) Fokker, Dirk W.; Lynch, Michael F. "Application to the Variety-Generator Approach to Searches of Personal Names in Bibliographic Data Bases. Part II. Optimization of Key-Sets, and Evaluation of Their Retrieval Efficiency". *J. Lib. Autom.* **1974**, *7* (2), 201–13.
(4) Moore, Gwendolyn, B.; Kuhns, John L.; Trefftzs, Jeffrey L.; Montgomery, Christine A. "Accessing Individual Records from Personal Data Files Using Non-Unique Identifiers"; National Bureau of Standards: Washington, DC, 1977; p 195.
(5) Sharpe, Richard B.; Fox, John B.; Hammond, Silas E. "The Information Age in Perspective: Proceedings of the ASIS Annual Meeting"; Knowledge Industries: New York, 1978; Vol. 15, pp 303–5.

NEWS AND NOTES

CAS ONLINE DOCUMENT ORDERING

The ability to place an order for original documents online is now available through CAS ONLINE, the chemical substance search and display system from Chemical Abstracts Service.

The searcher orders original journal articles, patents, or other documents of interest through the CAS Document Delivery Service by typing ORDER while logged-in to CAS ONLINE. Documents may be identified through their abstract numbers in *Chemical Abstracts* or through bibliographic information. The articles can be sent automatically to a stored address.

The cost is $12 for a document of up to 50 pages and $24 for documents of 51 pages of more. CAS will mail the article within 24 h in most cases.

While any article monitored by CAS can be requested, searchers will generally use the service to request documents indentified through a search of the 5.7 million substances in CAS ONLINE. Answers retrieved through that system include full bibliographic information on the 10 most recent references about the substances of interest.

CAS has also moved to encourage practice and refinement of search strategy on CAS ONLINE by offering three services without search charges. All searches of the demonstration file of about 50 000 substances, all sample searches followed by a full-file search, and answers requested in a format that provides only the index name and structure diagram are now offered at the connect-hour fee of $30 per hour, without additional charges.

BASIC'S NEW ADDRESS

Dr. H. R. Schenk (SANDOZ) was appointed Executive Manager of BASIC on July 1, 1982; Mr. H. Kniess (CIBA-GEIGY) and Dr. H. P. Scherrer (ROCHE) will act as Deputy Managers. The new address for BASIC is:

BASIC
Dr. H. R. Schenk, Executive Manager
P.O. Box 4043
CH-4002 Basel, Switzerland
Telephone: (061) 22 44 13
Telex: 65765