Quality Control of Chemical Data Bases

STEPHEN R. HELLER*
Environmental Protection Agency, MIDSD, PM-218, Washington, D.C. 20460

G. W. A. MILNE and R. J. FELDMANN
National Institutes of Health, Bethesda, Maryland 20014

Received June 18, 1976

The problems of quality control of the data in the files that comprise the EPA–NIH Chemical Information System (CIS) are described and discussed. The use of the Chemical Abstracts Registry Numbers (REGN) in these data bases is also described.

INTRODUCTION

Since 1969, a systematic effort has been made by a number of groups collaborating to collect a series of data bases of interest to organic chemists, write programs for searching through such data bases, and disseminate the resulting system as widely as possible via computer networks to the international scientific community. This effort has involved five agencies of the U.S. Government (NIH, EPA, ERDA, NBS and FDA) as well as several European units (BASF and DKFZ in Germany, the ETH in Switzerland, the Netherlands Organization for Chemical Information (NOCI), Sendai University in Japan, and the Mass Spectrometry Data Centre in England). The system that has emerged from this extensive collaboration has been dubbed the EPA–NIH Chemical Information System (CIS).

All the components of the CIS are concerned with either chemical data or bibliographic information on chemicals, primarily organic compounds. The data files range in size from a few hundred entries to over 40 000 entries and include files of mass spectra, carbon-13 NMR spectra, atomic coordinate data, x-ray diffraction data, and x-ray powder diffraction data. The bibliographic files cover primarily the literature of mass spectrometry (60 000 citations) and x-ray crystallography (15 000).

During development of components of this sort, the major preoccupation has been to acquire enough data to ensure that the files are large enough to be representative of a broad range of chemistry. The CIS is now entering a phase in which the quality of the data and the information pertaining to the compounds represented in the files is the weakest aspect of the system. Because of this, there has been an increasing effort during the past two years to devise methods for checking the quality of the files; this is the subject of this communication.

DISCUSSION

Perhaps the most common complaint from users of the CIS has been over the extensive duplication in the data bases. In the mass spectral data base, for example, acetone appeared as an entry no less than ten times. Quite apart from the irritation this causes users, such duplication is costly by the unnecessary use of disk storage and computer resources in searching. Mass spectra are sufficiently reproducible that there is no justification for duplication and the decision has therefore been made to eliminate it.

Elimination, however, implies the ability first to detect duplication and further to identify correctly the best of several entries. Detection of duplicate entries has emerged as a surprisingly difficult problem, and it is this that is discussed here.

Compounds in the data files of the CIS are identified by a name, a molecular formula, and a molecular weight, in addition to a file ID number. Early experiments were therefore carried out to identify duplicate entries by examining coincidences of one or more of these properties. This unfortunately proves not to be feasible because none of these identifiers is sufficiently unique. To use acetone as an example again, it is one of six compounds in the file with molecular formula C3.H6.O and molecular weight 58, and these six compounds are represented in the data base by no less than 30 entries (acetone, 10; propionaldehyde, 7; trimethylene oxide, 6; allyl alcohol, 3; propylene oxide and vinyl methyl ether, 2 each). Identifying duplicates on the basis of molecular formula or molecular weight is not therefore possible, and the problem is further compounded by the laxity in the nomenclature. Of the compounds mentioned above, only propylene oxide and vinyl methyl ether appear with just those names. The others are each identified by more than one trivial name—acetone, propan-2-one, etc.—that has been applied arbitrarily by the spectroscopist who originally measured the spectrum. The outcome of this combination of circumstances is that it is practically impossible to write a computer program to locate duplicate entries in any of the files, and recourse has therefore been made to semiautomated techniques.

In 1964, the Chemical Abstracts Service (CAS) began a systematic registration of all chemical materials. The main identifier that is used in this connection is the CAS Registry Number (REGN), which is an arbitrary number containing nine digits. Each REGN is applied to only one material as the only unique identifier for that material. This system will ultimately supersede the approach that employs the Wiswesser line notation (WLN), because WLN generation is still done manually in most cases and uses an incompletely and non-algorithmically defined set of rules.

During the last two years, the CAS has, on a fixed fee-for-service basis, been registering every compound in the data files of the CIS. This process has also been extended to other EPA files that are not part of the CIS, so as to conform with EPA regulations which require the registration of all file entries.

The CAS is supplied with the name and the molecular formula of each compound. The first step in the registration process is a name match whereby an attempt is made to find this name in the CAS Master Name File which contains about 3.5 million different compounds with about 5 million names or synonyms. If the name match is successful then the REGN and related information can be retrieved from the master file and the registration process is complete. If the name match fails, a structure match is attempted. The chemical structure of the compound may have been provided with the name, but if not, a structure is generated by chemists on the CAS staff. The structure is entered into the CAS computer system using a specially designed "chemical" typewriter, and it is used to search the CAS file of some 3.5 million registered structures.

Table I. Statistics of CAS Registration

FILE	# COMPOUNDS INPUT	# NAMES MATCHED	% NAME MATCH	# STRUCTURES MATCHED	# STRUCTURES REGISTERED	# NOT REGISTRABLE
MASS SPEC.	49160	17097	34.7	31418	7279	645
XRAY	11148	2527	22.7	8621	1096	0
CNMR	2824	1887	66.8	909	63	28
REGISTERED PESTICIDES	2608	2028	77.7	221	147	359
OIL & HAZARDOUS MATERIALS	915	758	82.8	116	98	41
PESTICIDE ANALYTICAL STDS.	389	344	88.4	45	2	0
DRINKING WATER	309	213	68.9	85	10	1

When the entry is found in this file, the REGN, name, and molecular formula for the compound are retrieved, but if the structure is not in the file, a new REGN is assigned, and the chemical is then registered by CAS and a Ninth Collective Index name is generated. If it is not possible to generate a structure from the information initially provided, consultation with the contracting agency ensues, and if a structure still cannot be defined, the entry is put aside and does not receive a CAS REGN. At present CAS is devising a method whereby these "unregistrable" chemicals (e.g., asphalt) can be assigned a REGN; however, implementation is still a number of months away.

At present, the compounds in seven data bases have been completely registered by CAS, and those in several other files are in the process of registration. The CIS files that have been completed are the MSDC–EPA–NIH Mass Spectral Data Base,[1] the EPA–NIH Carbon-13 Nuclear Magnetic Resonance Data Base,[2] and the Cambridge Crystal Structure file.[3] The EPA files that have been registered are the Pesticide file, the Oil and Hazardous Materials file, the file of Toxic Substances in Water, and the Pesticide Reporting file. Table I gives details of these files, including the number of compounds in each one, the number of successes in each step of the matching process and the number of "materials" that were not assigned CAS REGN, either because no structure could be drawn or because an undefined mixture of isomers was involved.

Once a complete file has been passed through the registration process, the next steps in the duplication removal are fairly clear. The file must first be sorted on the REGN to identify duplicates. Then the quality of each data set for a given REGN must be established. In the case of the mass spectral data base, this is done by means of a program developed by McLafferty and co-workers, which calculates a "quality index" for each spectrum. Finally, all but the best of a series of spectra associated with the same REGN are discarded. Work is now in progress on this phase of the process and is continuing without any notable difficulty.

ECONOMICS

The cost of registration is variable because the amount of work in each case is to some extent unpredictable. With the files discussed above, the cost of obtaining a REGN and the connection table for each compound has averaged about $3.00. This is a high figure, but one which, for the reasons enumerated below, is countenanced. By careful planning of the workload, it has been found that a stream of compound names moving at the rate of about 2000 per month can handled comfortably by CAS.

CONCLUSIONS

The benefits that accrue from the CAS registration process as it is described here are varied, and depend to some extent upon one's perspectives. From the point of view of the CIS system managers, there are some immediate gains, viz., the REGN and a reliable connection table which can be used in the CIS substructure programs. Perhaps the most important benefits are in the longer term. The reliability of the entire CIS, or at least of its data, is enhanced by the addition of the REGN's, and the value of the interfile link that is provided by the REGN's is basic to the structure of the CIS.

From the point of view of the user, the most obvious change is that duplicates are removed from the files. The more important advantages are that compounds with ill-defined or undefinable structures are no longer proffered to him as "answers" to his searches and that the answers which are suggested can be linked to the standard chemical literature through the REGN.

ACKNOWLEDGMENT

One of us (S.R.H.) wishes to thank M. Yaguda, M. Springer, and W. Greenstreet for their support of this project.

REFERENCES AND NOTES

(1) R. S. Heller, G. W. A. Milne, R. J. Feldmann, and S. R. Heller, *J. Chem. Inf. Comp. Sci.*, **16**, 176 (1976).
(2) B. A. Jezl and D. Dalrymple, *Anal. Chem.*, **47**, 203, (1975).
(3) O. Kennard, D. G. Watson, and W. G. Town, *J. Chem. Doc.*, **12**, 14 (1972).

A Flexible Interactive Graphics System for Searching Atom Connectivity Matrices

BO E. H. SAXBERG, DANIEL S. BLOM, and B. R. KOWALSKI*
Department of Chemistry, University of Washington, Seattle, Washington 98195

Received April 14, 1976

An interactive screen-generating structure search system with graphic input and display capabilities is described. The system allows great flexibility in searching atom connectivity matrices for interactively defined substructures or complete molecular structures. Host computer programs are written in UCI Lisp and Fortran, and are accessed via an intelligent graphics terminal. Results on a small collection of molecules are presented and potential applications are discussed.

Large molecular data bases have become necessary in many areas of chemistry. Computerized spectral analysis, for example, requires that a large file of molecules and their spectra be kept for reference. Pharmaceutical companies and other industries[1,2] dealing with molecular design need to keep files of molecules that have been studied in the past, as well as those undergoing current research. Large molecular data bases are also necessary for pattern recognition applications,[3] e.g., using