

Development of CAOCI and Its Use in ICI Plant Protection Division<sup>†</sup>

S. BARRIE WALKER

Technical Information Section, ICI Plant Protection Division, Jealott's Hill Research Station, Bracknell, Berkshire, United Kingdom

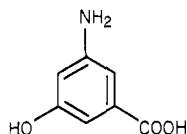
Received December 3, 1981

Research chemists involved in the synthesis of novel organic compounds need chemical intermediates and spend valuable time searching through suppliers' catalogs for useful compounds. The Commercially Available Organic Chemicals Index (CAOCI) data base, an application of WLN, provides the chemist with a means of identifying the commercial availability of a specific organic compound or group of closely related compounds by means of substructure search. The data base has been developed since 1974 when six member companies of the CNA (U.K.) agreed to cooperate in its production.

Chemists involved in the preparation of novel organic compounds for evaluation as agrochemical, pharmaceutical, or other products require starting materials and intermediates in laboratory quantities. It is usually the case that where a chemical can be purchased this is preferable to the chemist spending his valuable time making his own intermediates—in other words, time should be devoted to making speculative compounds for screening rather than the chore of making commercially available compounds.

How does the chemist know whether or not he can buy a specific compound? He undoubtedly will have an assorted collection of chemical suppliers' catalogs that he can look through, but this is a time-consuming and often unrewarding job. Many suppliers' catalogs give only an alphabetic name index. Better ones provide cross-indexing of names, molecular formula indexes, and sometimes chemical-class indexes. Aldrich is the only company to offer a substructure search facility, but of course this covers only their own catalog.

Let us examine the stages that the chemist goes through. Firstly, he has to think up a name for the molecule he requires. This can be a major hurdle.



For instance while the molecule above will have a systematic name, it could be named as an aniline, as a benzoic acid, or as a phenol by the chemist, depending on his use for the chemical. What is the answer to this problem? Well, whatever chemical name you give to a compound, and there are of course examples where many more than three names could be used for one compound, it will have just one Wiswesser Line Notation (WLN):<sup>1</sup>

ZR CQ EVQ

WLN could, then, provide the key to easy access to availability of chemicals.

The idea of using WLN to bring together the same compound from different suppliers' catalogs was discussed within the U.K. Chapter of the Chemical Notation Association—the CNA (U.K.)—at the beginning of the 1970s. Up to that time the main application of WLN had been for the indexing of in-house chemical collections, and using WLN as a tool to bring together information on the same compound was a new departure.

However, by 1972 the production of an index of all commercially available organic compounds was still being discussed

and no progress had been made. During the winter of 1972/73 an index of 20 000 organic chemicals from 10 small catalogs was produced in ICI Plant Protection Division (PPD). This was done in support of an index of suppliers' information which already existed in ICI, but which was limited to data on just 12 U.K. suppliers. The sorted output was taken to a meeting of interested people in the CNA (U.K.), and this small data base became the starting block for a project which began in 1974.

CNA (U.K.) and U.S. members were invited to contribute some form of resource, e.g., coding effort, card punching, or computer time and programming effort to compiling a larger file on the basis that work would be shared and the product made available to all contributors. A number of people volunteered their company's cooperation in the project, and then the crunch came. The group began to work out just how much effort would be needed in an initial phase of the project. We finally ended up with six participating companies.

Imperial Chemical Industries Ltd.

The Wellcome Foundation Ltd.

The Boots Company Ltd.

Pfizer Ltd.

Glaxo Group Ltd. (then including Allen &amp; Hanburys)

Beecham Pharmaceuticals

With some companies being more involved than others, but in reasonable proportion to their size, ICI took on the role of coordinating the work and producing computer programs linked to parts of CROSSBOW<sup>2</sup> software to establish and maintain the file. Coding and card punching effort was provided by the other contributors, and a file of Aldrich data was acquired (by ICI) and added into the file.

This phase of the work took approximately 18 months, and by mid-1976 the first product was available. By Sept 1976 the members of the group were in possession of hard copy molecular formula and WLN-ordered indexes. At this stage the group decided on the rather unpronounceable name of CAOCI for the project—Commercially Available Organic Chemicals Index.

Let us now return to our hard-pressed bench chemist frantically thumbing through his jumble of catalogs. Reference to CAOCI will quickly tell him whether the chemical is commercially available or not. Having ascertained that any specific compound is available and from which suppliers, the chemist can then refer to the appropriate catalogs to check pack size and price and order accordingly. If the chemist can see that a compound is available from five suppliers, perhaps one of them is his favourite or has a reputation for supplying quickly.

It did not take long before the molecular formula and WLN-ordered indexes were in the hands of the chemists, who quickly accepted the product and wanted more coverage. At

<sup>†</sup> Paper presented at the Symposium of the Chemical Information Division of the American Chemical Society, Las Vegas, Aug 1980.

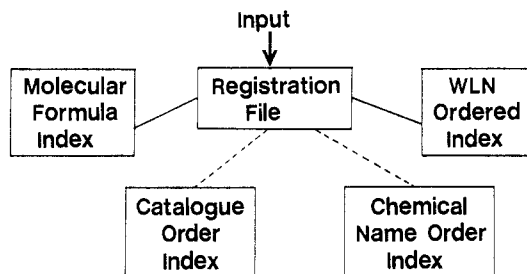


Figure 1.

this stage the index contained about 50 000 records derived from 18 catalogs, giving information on 20 000 different chemicals. The next task of the group, therefore, was to extend the index by including more catalogs, but we were also faced with the daunting task of updating the existing index. Adding in new data proved much easier than updating catalogs already on file. Despite considerable correspondence with suppliers there was some unwillingness on their part to provide updating information, and so coders had to resort to comparing an out-of-date catalog with a new one, marking the catalogs appropriately so that new compounds could be coded and added in.

We then had to decide what to do with deleted items, and it was agreed that we would flag these on the index as "withdrawn" rather than remove them. This is because when updating the catalog information it is not easy to establish if the supplier is the only supplier of the material or just one of many. Where there are other suppliers, then an enquirer will obviously avoid the entry marked "withdrawn" but where the supplier is the only company offering the compound, then our experience in ICI has shown that it may still be possible to obtain some of the material.

Over the next 2 years the group continued to add more catalogs, update existing catalogs, and improve programs written to handle the data. The indexes grew and were produced on microfiche to save space and computer stationery! The participating companies now had in their possession an index which could be used readily in three forms:

The Registration File which is essential in updating the index

The WLN-Ordered File which is useful to information officers or those who speak WLN

The Molecular-Formula-Ordered Index which can be used by the chemist

Let us look at these in order and discuss them in more detail. The main file is the registration file. This is the one used in updating the index and consists of six data elements:

- the catalog name
- the catalog reference number
- the WLN for a compound
- the molecular formula for a compound
- the chemical name
- and in some cases a withdrawn flag

The file is held on the computer by catalog coden and within this by catalog reference number. Virtually the only use of the registration file is for the updating and maintenance of the index. From this master file two further indexes are produced by sorting and splitting the file (Figure 1). The first of these is the molecular formula index (Figure 2) which is further sorted by WLN within the same molecular formula. This is the index which can be used by the chemist to look for specific compounds without the help of an information officer. In Plant Protection Division, we have found it necessary to set up units throughout the research block where chemists have access to the molecular formula index and an up-to-date set of catalogs. A further set in the library means that a chemist searching the literature for ideas can relatively quickly see

Typical MOL form entry

CBH8025	QVR BSI ORTHO-[METHYLTHIO]-BENZOIC ACID [WITHDRAWN*]	BADER	S38292-2
	QVR DSI PARA-[METHYLTHIO]-BENZOIC ACID. 97%	ALDRICH	14552-1
	QVISR THIOPHENOXACETIC ACID PHENYLMERCAPTOACETIC ACID [PHENYLTHIO] ACETIC ACID THIOPHENOXACETIC ACID PHENYL MERCAPTOACETIC ACID PURE [PHENYLTHIO] ACETIC ACID ALPHA-[PHENYLTHIO] ACETIC ACID	ALDRICH FAIRFLD FLUKA K&K KOCH LT P&B PARISH	TO3300-6 P-2730 U78810 K17405 4596H T11735 1987
	QISVR S-HYDROXYMETHYLTHIOBENZOATE S-HYDROXYMETHYL THIOBENZOATE S-HYDROXYMETHYL THIOBENZOATE	ALFA BADER LANCSTER	S12276 LAN-O523
	SHIR DVQ ALPHA-MERCAPTO-P-TOLUIC ACID	BADER	S50118-2
	T55J BVI EVI 2,5-DIACETYLTHIOPHENE	FAIRFLD	D-1510

Figure 2.

Typical WLN entry

QVR CG DG C7H4-CL202	3, 4-DICHLOROBENZOIC ACID 3, 4-DICHLOROBENZOIC ACID 3, 4-DICHLOROBENZOIC ACID 3, 4-DICHLOROBENZOIC ACID 3, 4-DICHLOROBENZOIC ACID 3, 4-DICHLOROBENZOIC ACID 3, 4-DICHLOROBENZOIC ACID 3, 4-DICHLOROBENZOIC ACID	SCHUCHAR ALDRICH CAMBRIAN EASTMAN FLUKA K&K KOCH LT P&B	820436 14493-2 CAM-DO1976 EO5529 U35320 K 3120 6247H D13790
QVR CG DG FVQ C8H4-CL204	4, 5-DICHLOROPHTHALIC ACID	ALDRICH	17988-4
QVR CG D01 CBH7-CL03	3-CHLORO-P-ANISIC ACID 3-CHLORO-4-METHOXYBENZOIC ACID 3-CHLORO-4-METHOXYBENZOIC ACID 3-CHLORO-4-METHOXY BENZOIC ACID	APIN APIN CAMBRIAN K&K	NO144C NO450C CAM-CO1542 K26458
QVR CG D1 CBH7-CL03	3-CHLORO-4-METHYLBENZOIC ACID	K&K	K 2649

Figure 3.

#### Crossbow Search Levels

1. BIT [CHEMICAL FRAGMENT]
2. WLN STRING
3. ATOM BYATOM

Figure 4.

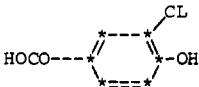
which of the intermediates he has just identified as being required for his new synthesis are available.

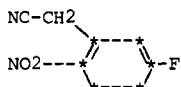
The index can be sorted into WLN order (Figure 3). The main use of this index will be the WLN-speaking Information Officer who can again use it to look for a specific compound but is more likely to use it to look for a class of related compounds. The latter will be brought closely together in the index by having similar WLN's. A WLN KWIC index enables some substructure searching to be done.

There is one other index which can be produced, and this is the chemical-name-ordered index. This needs special programming to cope with the peculiarities of finding the indexing point in a chemical name, i.e., 2,3-dimethoxybenzaldehyde would need to be indexed under "d" and  $\delta$ -valerolactone would need to be indexed under "v". While we have programs in ICI for producing such an index, these were not applied to the file in the period of the project. The main application of such a file is to provide an index which the nonchemist purchasing officer can understand and use.

ICI and some of the other companies have taken the file to a further level. When typical CROSSBOW facilities to search the file at the appropriate level are used, substructure search of the file can be achieved (Figure 4). This may be used to answer search questions from research chemists requiring a range of intermediates, for instance, *m*-halobenzaldehydes, or the biologist requiring a series of analogues of a biological screen lead.

Typical output from a search is shown in Figure 5. This provides the enquirer with the structure and the supplier's name and catalog reference. Where the compound supplied is a salt this is also indicated.

LG 28 JUL 1980    COMMERCIALY AVAILABLE CHEMICAL		014861
MOLECULAR FORMULA C7H5-ClO3		
WLN QVR DQ CG		
-----		
CATALOGUE	REF NO	COMPOUND NAME AND DETAILS
ALDRICH	CO4460-5	3-CHLORO-4-HYDROXYBENZOIC ACID HEMIHYDRATE
APIN	NO430C	3-CHLORO-4-HYDROXYBENZOIC ACID
CAMBRIAN	CAM-CO153O	3-CHLORO-4-HYDROXYBENZOIC ACID
K & K	K2179	3-CHLORO-4-HYDROXYBENZOIC ACID
KOCH LT	1122H	3-CHLORO-4-HYDROXYBENZOIC ACID PURISS
P & B	Cl301O	3-CHLORO-4-HYDROXYBENZOIC ACID

LG 28 JUL 1980    COMMERCIALY AVAILABLE CHEMICAL		018914
MOLECULAR FORMULA C8H5FN2O2		
WLN WNR DF B1CN		
-----		
CATALOGUE	REF NO	COMPOUND NAME AND DETAILS
ALDRICH	11544-4	5-FLUORO-2-NITROPHENYLACETONITRILE, 99% (WITHDRAWN)
P & B	FO375O	5-FLUORO-2-NITROPHENYLACETONITRILE
PARISH	1798	5-FLUORO-2-NITROPHENYLACETONITRILE (WITHDRAWN)

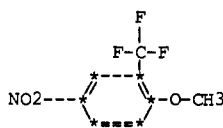
LG 28 JUL 1980    COMMERCIALY AVAILABLE CHEMICAL		070278
MOLECULAR FORMULA C8H6F3NO3		
WLN WNR D01 CXFFF		
-----		
CATALOGUE	REF NO	COMPOUND NAME AND DETAILS
FRINTON	FR-1151	2-METHOXY-5-NITROBENZOTRIFLUORIDE
K & K	K12459	5-NITRO-2-METHOXYBENZOTRIFLUORIDE

Figure 5.

During the 5-year existence of the project, administration was handled by a group of three members of the CNA (U.K.); our chairman, Tony Faulkner of Pfizer, Sandra Ward then of Wellcome but now with Glaxo, who acted as our Secretary, and myself as data and systems coordinator. Regular meetings were held with representatives from each of the member organizations, at which decisions were taken on the progress of the work. Efforts were also made for some consistency between the groups involved in coding. As time went by, and the project was shown to be viable, with large numbers of committed users in each organization, the managements of the representative companies decided that this was an optimum time to hand the file over to an external organization for maintenance, marketing, and development. After much discussion during 1979, the file was handed over to Fraser Williams (Scientific Systems)<sup>3</sup> in Nov of that year and is now available to other organizations. At handover we had expanded the index to contain approximately 120 000 references to 42 000 available chemicals from 48 suppliers.

#### ACKNOWLEDGMENT

While it would be impossible to acknowledge the assistance of all those who helped in the production and updating of the data base, I would particularly like to thank Chris Easton of ICI Plant Protection Division who wrote the update system used for most of the project and who established the links to CROSSBOW software for providing the search and structure print routines described in this paper. I would also like to thank all those people within the cooperating companies whose enthusiasm helped to carry the project through.

#### REFERENCES AND NOTES

- (1) Smith, E. G.; Baker, P. A. "The Wiswesser Line-Formula Chemical Notation", 3rd ed.; Chemical Information Management Inc.: Cherry Hill, NJ, 1976 (\$18 including postage).
- (2) Eakin, D. R. In "Chemical Information Systems"; Ash, J. E., Hyde, E., Eds.; Ellis Horwood Ltd.: Chichester; Chapter 14.
- (3) Fraser Williams (Scientific Systems) Ltd., Glendower House, London Road South, Poynton, Cheshire SK12 1NJ, England.