Analysis of the Registry of Toxic Effects of Chemical Substances (RTECS) Files and Conversion of the Data in These Files for Input to the Environmental Chemicals Data and Information Network (ECDIN)

OLE NØRAGER, WILLIAM G. TOWN,* and J. HOWARD PETRIE**

The Joint Research Centre of the European Communities, I-21020 Ispra (VA), Italy

Received August 24, 1977

A data bank for environmental chemicals, ECDIN, is being developed at the Joint Research Centre of the European Communities in cooperation with universities and research institutes in the nine member states as a part of the Environmental Research Programme of the EC. During the pilot phase of the project, data from the Registry of Toxic Effects of Chemical Substances have been incorporated into the data bank. Conversion of the data into ECDIN input format was necessary before inclusion of the toxicity data in ECDIN, and the computer programs used for this format conversion have produced various statistics for the contents of the RTECS files. Analyses of the data in three editions of RTECS are presented.

INTRODUCTION: THE ECDIN DATA BANK

In 1973 the Council of Ministers of the European Communities (EC) decided to include in the Environmental Research Programme of the EC a project for a data bank on environmental chemicals.[1,2] This Environmental Chemicals Data and Information Network (ECDIN) is to be an instrument which would enable all people engaged in environmental management and research to rapidly obtain reliable information on chemical products of environmental significance.

The project is designed as a Community action in which the Commission's own Joint Research Centre (JRC) closely collaborates with competent institutions in the member countries. To ensure this collaboration community funds were made available to permit the sponsorship of research contracts concluded with the institutions for data collection and processing as well as for system development.

In the first years of the project the work within the JRC has been concerned with an analysis of the data considered to be relevant for inclusion in the data bank and with the development of a general scheme for classification of these data. Further, the format for input of data to the bank as well as the format for display of data have been designed, and the necessary computer software has been developed and implemented for the IBM 370/165 computer at the computing center (CETIS) of the Ispra Establishment of the JRC. The SIMAS system[3] for storage and retrieval of data, originally designed for a library of computer programs, has been updated by members of the staff of CETIS to meet the requirements of the ECDIN data bank, and the SIMAS system has been used by ECDIN for demonstrations during the pilot phase of the project.

SIMAS allows the system designer to set up a number of classes each containing objects which may be catalogued. In addition, keywords and searchable identifiers may be assigned to the "objects". In the ECDIN implementation one "class" in which the "objects" were chemical compounds was used as it was not possible to search across classes.

To enable the data stored in SIMAS to become retrievable, it was necessary to develop a thesaurus for ECDIN. The SIMAS system contains procedures for thesaurus construction and maintenance. The thesaurus is organized in a number of broad keyword groups, each of which may contain a number of narrow keyword groups. Each "narrow keyword group" may in turn contain a hierarchy of keywords. One improvement of the SIMAS system introduced for the ECDIN implementation was the ability to associate numerical values with keywords and to search them with operators such as: equals, less than, greater than, less than or equal to, greater than or equal to, between (for ranges), error (to allow for uncertainty in data). Further the units in which the original measurement was made could be input and automatically translated into a standard unit. This facility was also made available to the searcher.

Table I. ECDIN Categories

Category code	Title
IDN	Identification
CSI	Chemical Structure Information
PCP	Physical and Chemical Properties of the Pure Substance
CAD	Chemical Analysis Data and Methods
SPL	Supply, Production and Trade
TPH	Transport, Packing, Handling and Storage
USE	Use and Disposal
DTE	Dispersion and Transformation in the Environment
TOX	Effects of the Chemical on the Environment (including toxicity)
CRR	Control Regulations and Recommendations

The data elements considered to be relevant to ECDIN have been organized into ten categories, the titles of which are given in Table I together with the identification codes assigned to these categories. The ten categories are further divided into fields and, in some categories, into subfields. The structure of the SIMAS system has demanded the organization of the data into logical records, one for each compound. Each logical record may contain data for over 100 data elements, where a data element may be a number such as a melting point or a free text state-of-the-art summary of a field such as mutagenicity. An input format has been designed for ECDIN to allow greater flexibility than is possible in the SIMAS system and to prepare for a future change to a new system.

In the summer of 1976 data for some 3500 compounds, which were selected for the ECDIN pilot phase, were loaded into the SIMAS system at the IBM 370/165 computer in Ispra. This set of data was used, in the autumn of 1976, for a series of demonstrations of the project, and the data bank has since been maintained in experimental operation with increasing amounts of data for selected compounds accessible through on-line retrieval.

During the pilot stage of the project it was decided to include into the data bank in the experimental phase data extracted

* To whom all correspondence should be addressed.

** Franklin Institute GmbH, D-8000 München, Germany.

0095-2338/78/1618-0134$01.00/0 © 1978 American Chemical Society

Table II. Distribution of the Data Classes over the Entries

Data class	Class code	1974	1975	1976
Prime name	A	12639 *1753*	16372 *1850*	21723 *2010*
Notes relating to identification	C	153 *8*	390 *13*	516 *15*
CAS registry number	D	5786 *1753*	6517 *1850*	8216 *2010*
Molecular formula	F	8571 *1739*	12998 *1839*	18488 *1997*
Molecular weight	H	8495 *1731*	12928 *1834*	18151 *1967*
Wiswesser line notation	J	5898 *1679*	6241 *1749*	6883 *1850*
Compound class code	N	4400 *678*	6948 *812*	10306 *905*
Department of Transportation name	P	*b*	*b*	985 *313*
Synonyms	R	7306 *1537*	11620 *1721*	16537 *1874*
Toxicity	T	12599 *1738*	16201 *1803*	20919 *1920*
Aquatic toxicity	U	*b*	441 *280*	444 *284*
Toxicology and cancer reviews	V	*b*	197 *115*	359 *166*
Federal standard	W	581 *354*	604 *362*	1283 *498*
NIOSH criteria document	X	45 *14*	68 *32*	218 *58*
Supplementary toxicity data	Y[a]	307 *144*	567 *235*	665 *284*
File maintenance	Z	4656 *1133*	6898 *1503*	8521 *1733*
No. of rejected records		107	193	142
No. of synonym entries	B	31697	47180	59571

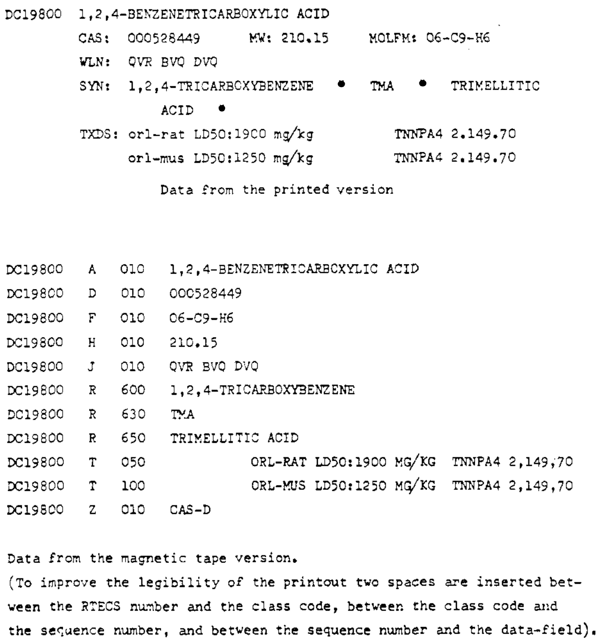
[a] Class code U in the '74 edition. [b] Not present in this edition.

from the Registry of Toxic Effects of Chemical Substances (RTECS), published annually in book form by the National Institute for Occupational Safety and Health (NIOSH).[4,5] The data elements contained in RTECS are all relevant to one of the ten data categories in the ECDIN data bank. The RTECS files which exist in machine-readable form on magnetic tape contain data for a large number of compounds, and it was therefore possible to include data from RTECS into the ECDIN files for a relatively low cost per compound. The processing of the RTECS files was performed automatically, without any detailed manual validation of the data, as the data are not in general thought to be a permanent part of the ECDIN data bank, but rather as a useful tool for testing of the ECDIN input, retrieval, and display techniques.

The following sections in this paper describe the data contained in RTECS and the processing of the RTECS toxicity data in Ispra, and communicate the results of the various analyses of these data resulting from this processing.

REGISTRY OF TOXIC EFFECTS OF CHEMICAL SUBSTANCES

RTECS is published by NIOSH in accordance with "the requirements of Section 20(a)(6) of the Occupational Safety and Health Act of 1970". (All citations in this section are from the introduction to the 1975 edition of RTECS.) Nearly all entries in RTECS correspond to pure compounds, because of difficulties in the unique identification of commercial mixtures, a problem which is also recognized by the ECDIN project. In the introduction to the 1975 edition of RTECS, the number of unique toxic substances is estimated to be 100 000, and a comparison of this number with the number of compounds present in the three latest editions of the Registry in Table II shows that even if the present growth rate is maintained, the final goal of the inclusions of "all toxic substances" in RTECS is still some years ahead.



```
DC19800  1,2,4-BENZENETRICARBOXYLIC ACID
         CAS:  000528449      MW: 210.15     MOLFM: 06-C9-H6
         WLN:  QVR BVQ DVQ
         SYN:  1,2,4-TRICARBOXYBENZENE  •  TMA  •  TRIMELLITIC
                    ACID  •
         TXDS: orl-rat LD50:1900 mg/kg          TNNPA4 2.149.70
               orl-mus LD50:1250 mg/kg          TNNPA4 2.149.70

                    Data from the printed version



DC19800  A  010  1,2,4-BENZENETRICARBOXYLIC ACID
DC19800  D  010  000528449
DC19800  F  010  06-C9-H6
DC19800  H  010  210.15
DC19800  J  010  QVR BVQ DVQ
DC19800  R  600  1,2,4-TRICARBOXYBENZENE
DC19800  R  630  TMA
DC19800  R  650  TRIMELLITIC ACID
DC19800  T  050              ORL-RAT LD50:1900 MG/KG  TNNPA4 2,149,70
DC19800  T  100              ORL-MUS LD50:1250 MG/KG  TNNPA4 2,149,70
DC19800  Z  010  CAS-D

Data from the magnetic tape version.
(To improve the legibility of the printout two spaces are inserted bet-
ween the RTECS number and the class code, between the class code and
the sequence number, and between the sequence number and the data-field).
```

Figure 1. Examples of data from RTECS 75.

Reported toxicity of individual chemicals is the criteria for inclusion of compounds in the Registry. Toxicity includes for this purpose the ability to produce benign and malignant tumors, to produce death, to produce mutagenic and teratogenic effects, and to produce irritation of the skin and the eyes as well as the ability to produce any other bodily harmful effect. The basis for the preponderance of the entries in RTECS are responses to single or short-term exposures of the toxic substances. Cumulative effects are only considered when these effects are mutagenic, teratogenic, neoplastic, or carcinogenic. Reported toxic effects on human qualify a compound for inclusion in the Registry regardless of the size of the dose and the length of the exposure time. If the reported toxic effect is on an animal and the effect is not one of the four mentioned above, in earlier editions of RTECS the dose had to be below certain limits to be accepted for inclusion in RTECS. These restrictions have been lifted with the publication of the 1976 edition. A table of limiting doses differentiating toxic and nontoxic compounds is comprised in the introduction to the earlier versions of RTECS.

The toxicity data are collected from published literature or from private communications and "no attempts have been made to resolve any question about data that have been published. Of necessity we rely on editing provided by the scientific community before publishing".

The compounds in RTECS are identified by the *Chemical Abstracts* 8th Collective Index name and assigned a RTECS number composed of two letters and five digits. Some entries do not contain a *Chemical Abstracts*' name, but a chemical or descriptive name as published in the source from which the toxic data were derived. These names are accompanied by a definition or a description. Synonyms for the identifying name constitute separate cross reference entries in the Registry, but these entries contain no data except the synonym and a reference to the corresponding prime name. The prime name entries might contain the Chemical Abstracts Service (CAS) Registry Number, molecular weight, molecular fomula, Wiswesser Line Notation, synonyms, toxic data, and other pertinent information. Figure 1 shows an entry in RTECS both as it appears in the printed version of RTECS 75 and as a printout from the magnetic tape version of the same edition.

Table III. Statistics for the Compound Classification in RTECS

Compound classes	1974	1975	1976
Agricultural	1103	1683	1891
	380	*469*	*524*
Carcinogenic	1361	1541	1904
	232	*238*	*302*
Drug	1310	3098	6029
	110	*161*	*188*
Mutagenic	26	30	34
	7	*10*	*12*
Organometallic	230	300	355
	13	*17*	*21*
Teratogenic	101	251	294
	38	*78*	*92*
Psychological effects	658	655	588
	28	*28*	*26*
No. of compounds classified in more than one class	352	529	678
	111	*149*	*200*

Magnetic tape versions of the three latest editions of RTECS have been made available to the ECDIN project through a special cooperative agreement with NIOSH. The magnetic tapes contain exactly the same information as the printed version of RTECS and are organized as fixed length records of 105 characters. Each record in the file is identified by a RTECS number, a class code, and a sequence number which together occupy the first 11 bytes of the record.

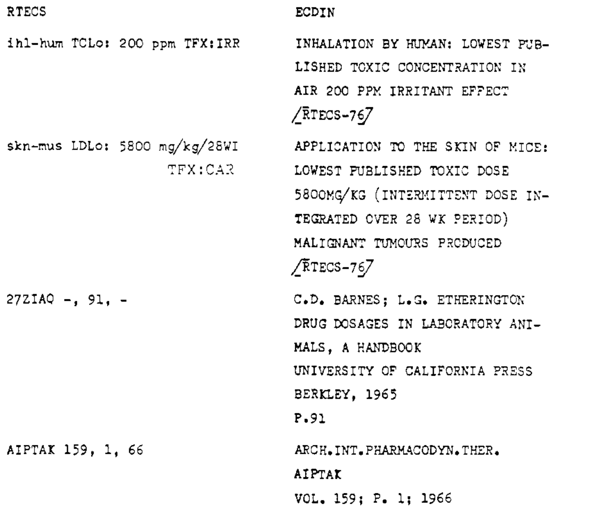
The cross reference entries are contained in records with the class code "B". The first step of the processing of a RTECS tape in Ispra is to copy the tape to a new tape excluding all cross reference records. Also a few other records, such as data records with a RTECS number without any corresponding prime name record, are excluded. The generated tape is used as input to a program, which produces general statistics on the contents of the file. The program examines only the identification number and the class code; it does not validate the class code by any test of the data on the record or perform any other general examination of the data. Table II contains statistics for the distribution of data records by entries.

The records with the class code "N" contain a classification for some compounds into structure, effect, and use classes. Table III contains statistics for the number of compounds in each class derived from these records.

The data extracted from the RTECS files for inclusion in the ECDIN files has until now been restricted to names, synonyms, and toxicity data. Those compounds in the RTECS files, which belong to the set of compounds selected for the pilot phase of the ECDIN project, were identified and extracted automatically via the CAS registry number. Nearly all compounds chosen for the ECDIN data bank have been assigned as CAS number by name and structure matching, and these numbers are present in the ECDIN files. The RTECS files, however, contain CAS numbers for less than half of the compounds. In spite of this paucity of CAS numbers in the RTECS files, no attempt was made to identify compounds occurring in both the ECDIN and the RTECS files by other means, such as names and Wiswesser line notations (WLN).

The three editions of RTECS have been compared with the same ECDIN registry file which contains ECDIN numbers and the corresponding CAS numbers for the compounds selected for the ECDIN pilot phase. This file contains 3749 ECDIN numbers and 3725 CAS numbers. Before the comparison was made, the value of the check-digit in all the CAS numbers in both files were validated and a general check of the format of the CAS numbers was performed. A CAS number was assumed to consist of nine digits, inserting leading zeroes when fewer digits were coded. The check-digit is the last, right-most, character in the number and the value of the check-digit is calculated from the eight other digits. The first digit is multiplied by 8, the second by 7, and so on to the eighth digit which is multiplied by 1. The results of all multiplications are summed, and the result of the summation is divided by 10. The check-digit is given the value of the remainder from this division. The numbers of CAS numbers in the three editions of RTECS rejected after this validation were 46 (1974), 46 (1975), and 61 (1976).

The numbers in italics in the tables are statistics for the subsets of the RTECS files, which contain data for compounds also present in the ECDIN files.



```
RTECS                               ECDIN

ihl-hum TCLo: 200 ppm TFX:IRR       INHALATION BY HUMAN: LOWEST PUB-
                                    LISHED TOXIC CONCENTRATION IN
                                    AIR 200 PPM IRRITANT EFFECT
                                    [RTECS-76]

skn-mus LDLo: 5800 mg/kg/28WI       APPLICATION TO THE SKIN OF MICE:
                      TFX:CAR       LOWEST PUBLISHED TOXIC DOSE
                                    5800MG/KG (INTERMITTENT DOSE IN-
                                    TEGRATED OVER 28 WK PERIOD)
                                    MALIGNANT TUMOURS PRODUCED
                                    [RTECS-76]

27ZIAQ -, 91, -                     C.D. BARNES; L.G. ETHERINGTON
                                    DRUG DOSAGES IN LABORATORY ANI-
                                    MALS, A HANDBOOK
                                    UNIVERSITY OF CALIFORNIA PRESS
                                    BERKLEY, 1965
                                    P.91

AIPTAK 159, 1, 66                   ARCH.INT.PHARMACODYN.THER.
                                    AIPTAK
                                    VOL. 159; P. 1; 1966
```

Figure 2. Toxicity data in RTECS format and in ECDIN input format.

TOXICITY DATA

The RTECS T-type records contain toxicity data from books, journals, reports, and in some cases personal communications to the editors of RTECS. The data are presented as toxic doses or toxic concentrations of the compound for various route-of-administration to various species. The species are mainly mammals and birds, but frogs and toads are also included. The routes of administration and the species are indicated in the T-records using three-letter codes. These codes are chosen as a sort of abbrevation of the full "names" of the species and the routes, but even though the meaning of some codes is obvious, such as "dog" for dogs and "cat" for cats, other codes such as "bdw" for wild birds and "scu" for subcutaneous, are more difficult to understand, and the RTECS files are in general not easily readable without the glossary in the introduction to the printed versions of the Registry.

It was decided that all data retrieved from the ECDIN data bank should be displayed in a directly readable format. This applies equally to data included in ECDIN from RTECS. The SIMAS system only allows display of data in the same format as the data are loaded into the system. Therefore, the necessary expansion of the toxicity data is performed during the conversion of an RTECS file into ECDIN input format.

The format conversion is performed by replacement of each code by a full word or by a string of words. Some examples of data in the RTECS format and in the converted format are shown in Figure 2. The RTECS T-records contain the RTECS number in columns 1–7, the class code in column 8, and a sequence number in columns 9–11. Columns 12–21 are left blank, columns 22–24 contain the route-of-administration codes, and the species codes are in columns 26–28. The dosage code, the dose, and possible specifications of the duration of

Table IV. Statistics for the Distribution of Routes of Administration over the T-Records

Route of administration	1974	1975	1976
Application to the skin	1134	1332	1713
	434	*543*	*654*
Inhalation	1211	1517	1683
	571	*734*	*789*
Intraperitoneal	5838	7423	10274
	813	*1022*	*1149*
Intravenous	2633	4405	5249
	292	*465*	*529*
Oral	7340	9672	12034
	1979	*2594*	*2901*
Subcutaneous	2545	4022	4529
	467	*690*	*778*
Other routes	885	1653	1884
	188	*456*	*517*
No. of rejected codes	24	28	15
	6	*5*	*1*

Table V. Statistics for the Distribution of Species over the T-Records

Species	1974	1975	1976
Cats	279	459	494
	78	*136*	*142*
Dogs	448	778	879
	156	*277*	*295*
Frogs	92	148	162
	31	*48*	*49*
Guinea pigs	630	1001	1143
	222	*376*	*420*
Hamsters	119	172	202
	47	*71*	*87*
Human	574	681	785
	198	*245*	*291*
Mammals (unspec)	27	227	271
	11	*120*	*126*
Mice	9767	13644	18381
	1351	*1773*	*2027*
Rabbits	1590	2377	2794
	498	*748*	*870*
Rats	7671	9774	11323
	1965	*2352*	*2610*
Wild birds	184	189	191
	90	*92*	*94*
Infants, women,	68	130	221
men, children	*24*	*53*	*74*
Others	321	444	520
	164	*284*	*232*
No. of rejected	21	22	0
codes	*2*	*1*	*0*

Table VI. Statistics for the Number of Records with a Specified Toxic Effect in the RTECS Files

Effect	RTECS code	1974	1975	1976
Carcinogenic	CAR	1012	1190	1427
		286	*317*	*391*
Central nervous system	CNS	117	147	180
		46	*62*	*68*
Irritant	IRR	54	61	75
		40	*45*	*57*
Neoplastic	NEO	871	1019	1325
		195	*218*	*279*
Psychotropic	PSY	230	237	244
		14	*15*	*14*
Teratogenic	TER	170	380	456
		67	*123*	*153*
Other effects	*a*	139	211	280
		56	*82*	*97*
Rejected		37	31	12
		8	*8*	*1*

[a] Allergic reactions, effects on blood clotting mechanism, unspecified effects on blood, effects on blood pressure, corrosive effects, cumulative effects, cardiovascular effects, production of drug dependence, effects on the eye, gastrointestinal effects, effects on the endocrine glands, effects on the muscous membranes, effects on the musculo-skeletal system, mutagenic effects, effects on the peripheral nervous system, effects on the respiratory system, effects on red blood cells, effects on the skin, effects on liver and kidney function, unspecified toxic effects, and effects on the white blood cells.

the exposure and of the toxic effect follow in the sequence cited in columns 30–70. Columns 72–77 contain a code for the source of information, and columns 79–105 a further specification of the source. In the majority of records, the code in columns 72–77 identifies a scientific journal and the data in columns 79–105 specify the publishing year, volume number, and page.

The computer programs, which perform the format conversion, produce at the same time statistics for the toxicity data and detect various types of coding errors. All records, which contain codes not found in the glossary mentioned above, or which contain format errors, are rejected. The number of records rejected owing to errors in the species and route-of-administration codes are included in the statistics in Tables IV and V.

It should be noted that the RTECS file undergoes constant review as a result of both user comments and planned periodic examinations of selected data by NIOSH. The quality of RTECS data (in terms of data field compatibility) has improved significantly since the Registry was first published in 1971 as the Toxic Substances List. As indicated in Tables II, IV, and V, the rejected data records in the 1976 RTECS file represent error rates of 0.6, 0.9, and 0%, respectively.

The dosage code consists of three or four characters. The first two characters indicate whether the dose caused death (LD) or other toxic effects (TD), and whether it was administrated as a lethal concentration (LC) or as a toxic concentration (TC). TD and TC are always followed by Lo, indicating the reported dose (concentration) is the lowest dose (concentration) published or by other means made known to the editors of RTECS. LD and LC can also be followed by Lo, with the same meaning as described above, or by number in the range 1 to 99, which is normally 50. LD50 (LC50) means a calculated dose (concentration) which is expected to cause death of 50% of an entire population of an experimental animal species. The frequencies or occurrence of the different dose codes in the 1976 edition are 971 (LCLo), 404 (LC50), 8119 (LDLo), 23863 (LD50), 253 (TCLo), and 3754 (TDLo); 17 records with toxicity data contain LD or LC followed by a number other than 50.

T-records, which specify a lowest published toxic dose or concentration, should also include a three-letter code specifying the toxic effect. Records containing a lowest published toxic dose or concentration, but invalid effect code, are rejected, and the number of these records is included in the statistics in Table VI. Records containing a lethal dose or concentration and also specifying an effect are also rejected, but these records are not included in Table VI. The programs reject records with a toxic dose given as a concentration in air and those with a toxic concentration given as weight per kilogram body weight. The total number of rejected records is included in the statistics in Table VII.

The assignment of RTECS toxicity data to the fields in the ECDIN input format is performed automatically by the conversion programs. All data from the T-records are placed in the ECDIN category TOX, and within this category allocated to one of the six fields shown in Table VIII according to the following rules:

All records with a lowest published toxic dose or concentration and the effect specified as: (a) carcinogenic effects (CAR in RTECS) or neoplastic effects (NEO in RTECS) are

Table VII. Statistics for the Distribution of T-Records over the RTECS Entries

	1974	1975	1976
No. of entries containing T-records	12599	16214	20925
	1752	*1849*	*2009*
Total no. of T-records	21859	30145	37447
	4849	*6519*	*7324*
No. of rejected records	250	434	366
	195	*120*	*97*
Distribution of the accepted records over the RTECS entries:			
1 record	8589	10807	14385
	697	*560*	*572*
2 records	1872	2447	3276
	378	*368*	*371*
3 records	842	1103	1193
	252	*259*	*285*
4 records	358	526	598
	115	*168*	*185*
5 records	251	362	425
	74	*114*	*134*
>5 records	520	829	939
	202	*330*	*371*

Table VIII. Fields in the ECDIN Category TOX

Field code	Title
MAN	Adverse effects on man
EXP	Experimental studies on animals to assess human toxicity
TLA	Effects on terrestrial animals (excluding data allocated to EXP)
CAR	Carcinogenity
MUT	Mutagenicity
ALL	Allergic and immological reactions

allocated to the ECDIN field CAR; (b) mutagenic effects (MUT in RTECS) are allocated to the ECDIN field MUT; (c) allergenic systemic effects (ALR in RTECS) are allocated to the ECDIN field ALL.

All records not allocated by the rules a, b, and c are allocated according to species: (d) children, human, infants, men, and women are allocated to the ECDIN field MAN; (e) cats, dogs, frogs, guinea pigs, gerbils, hamsters, mammals (unspecified), monkeys, mice, pigs, rats, and rabbits are allocated to the ECDIN field EXP; (f) birds, wild birds, chickens, cattle, ducks, domestic animals, pigeons, quails, squirrels, toads, and turkeys are allocated to the ECDIN field TLA.

The retrieval of data in the SIMAS system via an inverted file requires for the generation of this file the assignment of keywords to the data before the data are loaded into the SIMAS system. The thesaurus is constructed in advance of data loading, and the system assigns four character identifiers to each keyword, which must be used for input of the keyword. During the conversion of format of the toxicity data to ECDIN input format, an automatic indexing is performed for some groups of data allocated to the ECDIN fields MAN and EXP. All records allocated to these fields and containing a lethal dose for oral administration to any species are assigned the keyword AAFL, those records containing a lethal dose for application to the skin are assigned the keywords AAIE, and those which contain a lethal concentration for inhalation are given the keyword AAFV. In these two fields the programs also assign keywords to records containing a lowest published toxic dose or concentration and with the route of administration specified as either oral, inhalation, application to the eye, or application to the skin.

The SIMAS keywords can be associated with numerical values which are searchable. Rather than use the actual doses, the keywords are given a toxicity grading in the range 1 to 5, where 1 means very toxic and 5 less toxic. With this facility we are able to retrieve compounds having a particular grade of toxicity or a range of toxicity grades such as medium to very high toxicity (grades 1, 2, and 3). For doses the gradings are calculated on the basis of the weight of the dose per kilogram body weight; 1 is for toxic doses less than 1 mg/kg, 2 for doses between 1 and 10 mg/kg, and so on until 5 for doses greater than 1000 mg/kg. The gradings for concentrations are calculated from the concentrations expressed in parts per million (ppm). Concentration less than 2 ppm are given the grading 1, from 2 to 20 ppm the grading 2, continuing to 5 for concentrations greater than 2000 ppm.

Table IX. Statistics for the Indexing of Data on Lethal Doses and Concentrations

	1974	1975	1976
No. of records allocated to the ECDIN fields EXP and MAN	19416	27285	34102
	4123	*5708*	*6394*
No. of records indexed as:			
Single lethal dose, oral administration, AAFL	6217	8121	10192
	1664	*2113*	*2347*
Single lethal dose, application to the skin, AAIE	826	989	1325
	344	*444*	*549*
Lethal concentration, inhalation, AAFO	965	1196	1344
	420	*540*	*581*
Statistics for the grading of these keywords:			
Grade 1	68	100	103
	17	*31*	*32*
Grade 2	326	464	529
	114	*169*	*189*
Grade 3	1288	1744	1867
	387	*528*	*560*
Grade 4	3440	4310	5392
	990	*1200*	*1297*
Grade 5	2887	3689	4971
	921	*1170*	*1399*

The bibliographic references in the T-records constitute a special problem for the format conversion. The codes in columns 72–77 of a T-record refer to a table of bibliographic data in the printed version of RTECS. The codes for journals are selected as the CODEN's,[6] whenever these are available. For those journals where no CODEN has been assigned or for which the CODEN was not known by the editors of RTECS and also for books, reports, etc., the codes contain an asterisk as the sixth character instead of a CODEN check character. The codes used in RTECS for personal communications contain lower case letters. Whenever columns 79–105 contain volume–page–year specifications, the format used is volume, page, year. For some sources this format is inadequate and many T-records contain, in column 79–105, report numbers or other data in a free format.

In the ECDIN input files the bibliographic data have a standard format which is independent of the field or category. To convert the data from RTECS to this format, the lists of bibliographic information from the printed versions of RTECS have been coded in ECDIN input format, partly manually, partly automatically. The automatic part of the coding contains a validation of the check-digit in the CODEN's. When the check-digit is accepted, the CODEN is placed in the ECDIN subfield for CODEN's. If the check-digit is rejected, no CODEN is included for the journal concerned, but the code is still accepted for the identification of the source in a conversion run. Table X contains statistics for the frequency of reference to the most cited journal, books, etc. When the reference code is identified in a T-record, the programs search the rest of the record for volume–page–year information. If the data are found in the described format, they are included in the converted file. Only for reports from the National Technical Information Service, Springfield, Va., is the report number in columns 79–105 accepted for inclusion in the converted file; for all other sources, free format data in these columns are ignored.

Table X. Statistics for the Frequency of Citation of the Most Cited References

Title	1974	1975	1976
Am. Ind. Hyg. Assoc. J.	829	893	1021
	262	*296*	*287*
Arch. Int. Pharmacodyn. Ther.	669	832	889
	64	*73*	*85*
Am. Med. Assoc. Arch. Ind. Health	213	263	282
	128	*102*	*113*
Am. Med. Assoc. Arch. Ind. Hyg.	371	404	461
Occup. Med.	*187*	*215*	*221*
Arzneim.-Forsch.	246	328	358
	17	*33*	*32*
Biochem. Pharmacol.	292	300	295
	36	*34*	*34*
Br. J. Cancer	235	266	284
	85	*89*	*91*
Chemical Biological Coordination	809	809	805
Center[a]	*93*	*92*	*96*
Cancer Res.	412	466	619
	66	*78*	*103*
Chicago University Radiation	733	594	219
Laboratory	*94*	*69*	*28*
Food Cosmet. Toxicol.	236	311	394
	69	*75*	*85*
Federation of American Societies	210	220	226
for Experimental Biology[b]	*57*	*65*	*67*
Guide Chem. Used Crop Prot.	61	300	289
	17	*125*	*130*
Hyg. Sanitation (English translation	177	253	252
of *Gigiena Synitariya*)	*99*	*149*	*152*
J. Ind. Hyg. Toxicol.	418	482	517
	270	*303*	*309*
J. Med. Chem.	134	170	3560
	0	*1*	*8*
J. Natl. Cancer Inst.	345	451	532
	91	*104*	*123*
J. Pharmacol. Exp. Therap.	1962	3306	3630
	272	*404*	*434*
National Academy of Science	392	384	377
National Research Council[c]	*48*	*45*	*49*
National Defense Research	487	511	464
Committee	*34*	*39*	*38*
National Technical Information	896	1438	1857
Service, Springfield, Va.	*92*	*185*	*313*
Pesticide Chemicals Official	130	233	223
Compendium[d]	*49*	*134*	*125*
Proc. Soc. Exp. Biol. Med.	242	293	322
	60	*71*	*76*
Toxicol. Appl. Pharmacol.	1785	2280	2358
	529	*654*	*680*
Union Carbide Data Sheet	150	145	516
	17	*16*	*135*
World Rev. Pest Control	214	190	178
	143	*136*	*133*
The Merck Index[e]	594	606	600
	119	*120*	*120*
Psychotropic Drugs and Related	1109	1097	1087
Compounds[f]	*37*	*37*	*37*
Sbornik Vysledku Toxikologickeho	328	322	321
Vysentreni Latek A Pripravku[g]	*62*	*62*	*71*
Chemistry of Pesticides[h]	0	327	325
	0	*192*	*196*

[a] Chemical Biological Coordination Center, Summary Biological Tests, National Research Council, Washington D.C. [b] Federation Proceedings, Federation of American Societies for Experimental Biology. [c] National Academy of Science, National Research Council, Chemical Biological Coordination Center, Review. [d] Pesticide Chemicals Official Compendium. Association of the American Pesticide Control Officials, Inc., Topeka, Kans., 1966. [e] P. G. Stecher, et al., "The Merck Index", An Encyclopedia of Chemicals and Drugs, Merck & Co., Inc., Rahway, N.J., 1968. [f] E. Usdin and D. H. Efron, "Psychotropic Drugs and Related Compounds", 2nd ed, Washington, D.C., 1972. [g] Josef V. Marhold, Sbornik Vysledku Toxikologickeho Vysentri Latrek A Pripravku, Praha, 1972. [h] N. N. Melnikov, "Chemistry of Pesticides", Springer-Verlag, New York, N.Y., 1971.

Table XI. Statistics for the Year of Publication of the Sources for RTECS Based on the Information in the Volume-Page-Year Field of the T-Records

Year of publication	1974	1975	1976
Earlier than 1900	55	90	99
	34	*51*	*58*
1900–1949	3685	5727	6218
	997	*1434*	*1520*
1950–1959	4117	5614	5963
	771	*975*	*1035*
1960–1969	7820	9790	11186
	1679	*2093*	*2202*
1970	600	950	1208
	211	*252*	*268*
1971	650	1119	1516
	151	*372*	*377*
1972	2121	2374	2874
	321	*378*	*416*
1973	277	821	1739
	58	*231*	*278*
1974	56	433	1274
	15	*71*	*104*
1975		71	727
		8	*57*
1976			155
			9

Table XI contains statistics relating to the year of publication. These statistics are based on the information in columns 79–105; the year of publication given for books and reports in the list of bibliographic information is not included.

FURTHER ANALYSES OF THE RTECS DATA

The storage of data on chemical structures, physical-chemical properties, and biological activities of chemical compounds in ECDIN makes the data bank an obvious source for future studies of structure–activity relationships. We are ourselves interested in prediction of the toxicity of compounds in the ECDIN files from other data present in the data bank, and we are continuing the analysis of the RTECS data with a study of the relations between the topological structures and the toxicity of compounds in RTECS using techniques similar to those of Adamson and Bawsen.[7] To facilitate the automatic handling of the toxicity data we have produced a file in which each of the RTECS T-records is represented as a vector.

The interface between other chemical structure systems and the ECDIN data bank is facilitated by the storage of several types of structure representations for each compound, including WLN, CAS connection tables, DARC codes, bond electron matrices, and structure diagrams. The ICI CROSSBOW system[8] has been implemented in Ispra as a part of the ECDIN project to assist with the management of WLN's, and to generate structure diagrams from the WLN's.

The RTECS files contain WLN's for less than half of the compounds present, and many of these WLN's were incompatible with the CROSSBOW system. WLN's have therefore been coded for all compounds with a known structure in the 1975 edition of RTECS. ECDIN files now contain WLN's for some 14 000 compounds from RTECS 75, and the analysis of these WLN's and of the relations between the structural data and the toxicity data is in progress.

For the purpose of substructure search, the CROSSBOW system allows rapid screening of the WLN files to be performed, using a bit screen generated directly from the WLN and stores in the CROSSBOW master file together with the WLN. The bits in the screen are set or not set depending on the presence or absence of 148 defined structural fragments. The fragments are in two classes: those which are closely related to WLN symbols or short sequences of WLN symbols, and those which describe general features of the ring systems. We have made an extension of the fragment generation program and use the fragment-vectors output from this program as structure input in the correlation study. If we succeed in establishing a model which allows reasonably good prediction of the toxicity of compounds with known structures, we would have a tool which could be used not only to estimate toxicity data but for validation of toxicity data in the data bank. In particular, gross transcriptions errors in LD50 data or mistakes with dose units could be detected in this manner.

CONCLUSIONS

The inclusion of data from RTECS in the ECDIN data bank has made part of the Registry accessible for on-line retrieval. The object of the exercise was not merely to create an "on-line" version of the Registry, but rather to use RTECS as a tool in the development of the ECDIN data bank. Nevertheless, our experiment has shown that is is possible to handle the data from the magnetic tape versions of RTECS automatically and to convert them into a form suitable for input into a data retrieval system without any manual editing. The toxicity data have thus been made available in searchable form with ECDIN, to facilitate their combination using Boolean logic with production estimates, use patterns, and other data elements in the ECDIN data bank.

ACKNOWLEDGMENT

We wish to thank the editors of RTECS and NIOSH for making the magnetic tape versions of RTECS available to us, and to acknowledge the grants from the Commission of European Communities and from the Danish Natural Science Research Council to Ole Nørager, which have made this work possible. We also wish to acknowledge the advice given by John C. Gage (consultant to the Commission of the European Communities) on the conversion of the RTECS codes into free text and the assignment of data to ECDIN fields, and to thank Marinus de Groot of the Joint Research Centre for assistance with the translation of the bibliographic codes. We also wish to thank Philippe Bourdeau, Friedrich Geiss, and Heinrich Ott for their encouragement and advice.

REFERENCES AND NOTES

(1) F. Geiss and Ph. Bourdeau, "ECDIN, an EC Data Bank for Environmental Chemicals", *Environ. Qual. Saf.*, **5**, 15–24 (1976).

(2) M. Boni, F. Geiss, J. H. Petrie, and W. G. Town, "The Development of a Data Network on Chemicals and Their Effects on the Environment. The Environmental Chemicals Data and Information Network (ECDIN) of the European Communities", Proceedings from the EURIM II Conference, 23–25 March 1976, Aslib, London, 1977, pp 145–147.

(3) G. Gaggero, C. Lunghi, and C. Mongini-Tamagnini, paper presented at the International Computing Symposium, Venice, April 12–14, 1972.

(4) The Toxic Substances List, 1974 edition.

(5) The Registry of Toxic Effects of Chemical Substances, 1975 and 1976 editions, The National Institute for Occupational Safety and Health, Rockville, Md.

(6) CODEN for periodical titles, American Society for Testing and Materials, Philadelphia, Pa., 1966.

(7) G. W. Adamson and D. Bawden, "A Method for Structure-Activity Correlation Using Wiswesser Line Notation", *J. Chem. Inf. Comput. Sci.*, **15**, 55–58 (1975).

(8) D. R. Eakin, E. Hyde, and G. Palmer, "The Use of Computers with Chemical Structural Information: The ICI CROSSBOW System", *Pestic. Sci.*, **5**, 319–326 (1974).

PAGODE[†]: The Computer-Based Chemical Information System of CLIN MIDY Research Center

SAMUEL BERDUGO,* JEAN BOITARD, JEAN PAUL GERVOIS, ANNE MARIE SEGRETAIN, and ODILE PIETREMENT

Centre de Recherches CLIN MIDY, 34082 Montpellier Cedex, France

Received May 9, 1977

In collaboration with ARDIC, CLIN MIDY Research Laboratories have implemented an in-house database using the DARC topological coding system. This paper describes the general organization of the PAGODE system and its capabilities.

This paper describes CLIN MIDY'S chemical information system. Chemical information must be completed by biological information as is usually done in pharmaceutical firms. But creating a numerical comparative biological database is not an easy task. However, chemical information exists in a standard form ready for computer processing. So our first step was to store the in-house chemical structures for computer handling. The in-house documentation problem requires a more accurate coding system than the huge international chemical system of documentation, and it has quite a different purpose, since, in getting a precise description of any sequence of atoms in a series of molecules, structure–activity relationships use sophisticated coding systems. So the DARC system was adopted. Professor Dubois and his team of scientists in Paris created the DARC[1] system whose diffusion in the chemical industry is promoted by ARDIC[4] (Association pour la Recherche et le Dêveloppement de l'Informatique Chimique; Research and Development of Chemical Data Processing Association) which helped us with the general organization of the CM database and with the coding of the compounds.

The whole project was divided into several steps. The first step was to create the chemical database and to test it in batch processing. This article deals with the problems encountered during the first phase. According to our first results, the DARC system meets our initial requirements.

CODING SYSTEM

Choice of the Coding System. Before choosing the DARC system, we compared other chemical systems of coding commonly used in the pharmaceutical industry, mainly WLN and Ringcode used by DERWENT. Our requirements concerning a coding system were as follows: (1) a one-to-one correspondence between structure and code; (2) the possibility of constructing search keys or screens from the code, which is, of course, a computer requirement and also necessary if one wants to create an inverted file and to get a quick answer from the computer; (3) a description of the molecular topology (structure–activity relationships require the capability of describing any structure, including the sequence of atoms

[†] Programme Automatique de Gestion et d'Organisation de Donnêes.

0095-2338/78/1618-0140$01.00/0 © 1978 American Chemical Society