Environmental Information Databases

Kristina Voigt

GSF – National Research Center for Environment and Health, Neuherberg, Germany

Abbreviations

BI = bibliographic database; CA = catalogue of chemicals; db = database; DADB = metadatabase of online databases; DACD = metadatabase of CD-ROMs; DAIN = metadatabase of internet resources; FBDB = fact-based database; FT = full-text database; INDB = integrated database; MD = metadatabase; NU = numeric database; RD = research database; RE = reaction database; ST = structural database; TBDB = text-based database; WWW = World Wide Web.

Glossary

Metadatabase
Database which contains descriptive *information* about *knowledge* in a database, including domain assignment, ownership, access restrictions, database model. Other definitions are given in the text.

1 INTRODUCTION

The solution of serious problems in environmental protection, environmental management, and environmental research can be based only on the effective use of comprehensive and reliable information on our environment.[1] This information which is currently being collected in various types of data sources takes the form of biological, physical, chemical, geological meteorological, etc. data describing the state and dynamics of our environment. The availability of information is increasingly important in our society. Information will become one of the most valuable assets in many of our activities. However, it already becomes apparent that due to the proliferation of information, the task to obtain the right information on a specific subject is difficult to solve. This is especially applicable to environmental issues. The key problem in this regard is therefore: where to find the relevant information on environmental questions.

In this article, the focus lies on databases on environmental information of chemical substances. It will be indicated in which primary database to find a particular kind of environmental information. Examples will be given on the quantity of environmental information stored in the databases described.

2 WAYS OF DESCRIBING ENVIRONMENTAL DATABASES

Several ways exist to describe environmental databases. First, databases are categorized by media type. These media types are commonly accepted as online databases, *CD-ROMs*, and Internet resources. Second, databases belonging to a medium are grouped according to their subject into ‘general interest databases’, ‘scientific and technical databases’, and ‘business and regulatory databases’. The third way of classification is the division of environmental databases into database types. Figures 1 and 2 illustrate these various ways of categorizing databases.

The ways mentioned are only the common and obvious ones. Due to the very diverse subject matter of environmental science, there are many other possible ways to classify these databases. For example, environmental databases can be specialized to a specific type of environmental information like ecotoxicity data on chemical substances, or concentration data in environmental media. On the other hand, several databases focus on a specific use of chemical substances, like databases on pesticides or on solvents.[2]

Before these categories are looked upon in depth, the terms environmental data, environmental information, environmental database, and environmental information system should be defined.

2.1 Definition of Environmental Data, Information, Database, and Information System

Environmental data are technical, spatial, and temporal data for the environmental media air, water, and soil. They pertain to questions of waste, noise, dangerous substances, fauna and flora, landscape, nature, and species conservancy. With the help of the analysis and interpretation of those data environmental information can be created.

An environmental database is a particular type of database that stores mainly environmental data. According to environmental informatics experts, a database can be called an ‘environmental database’ if the following three conditions are fulfilled:

1. the majority of data are environmental data
2. a database system is used for the storage of these data
3. the database is established as the basis for environmental uses and inquiries.[3]

Applying this definition to environmental information on chemical substances, the following conditions have to be included:

4. the majority of data are chemical data and information
5. the database is established as the basis for chemical questions and uses.[4]

Environmental Information Systems (EIS), used as a technological–organizational infrastructure to provide

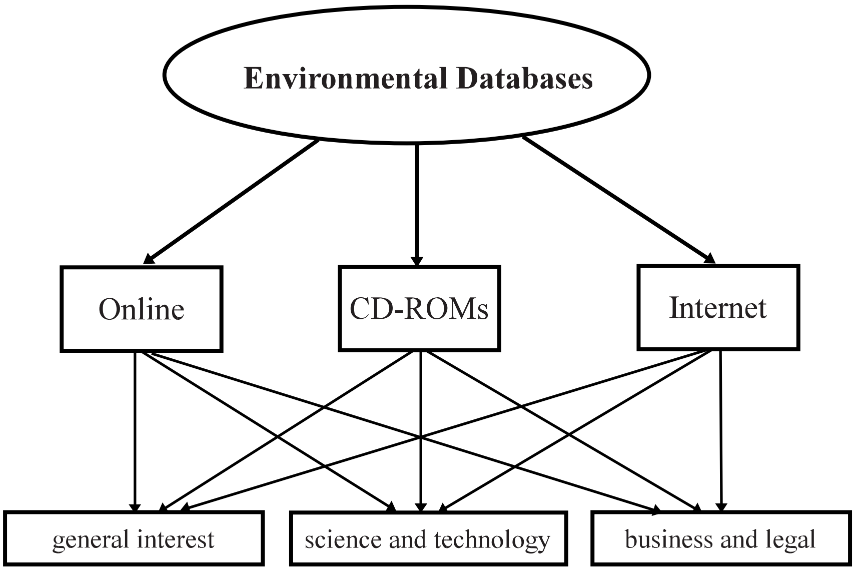


Figure 1 Ways of categorizing environmental databases

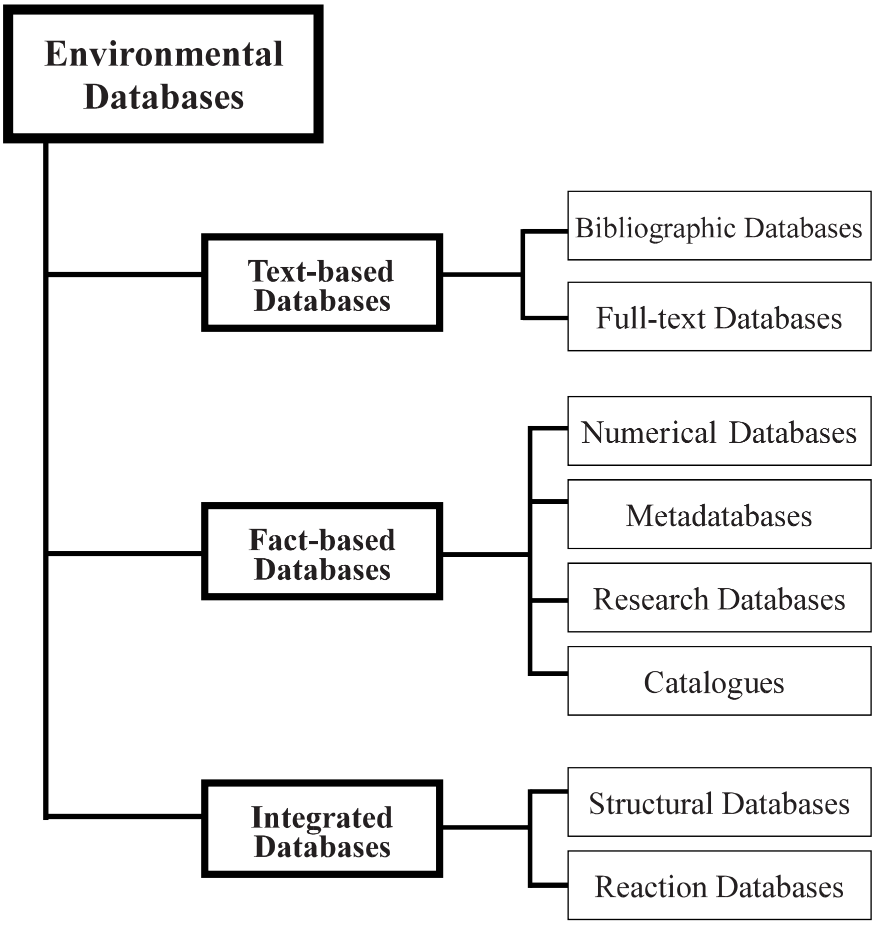


Figure 2 Types of environmental databases

environmental information from special fields in different environmental databases, are often geographically localized. Therefore environmental information systems are sometimes considered to be extended geographical information systems (GIS). However, EIS also hold thematic data (i.e., environmental facts such as measurement values on chemical substance attributes, environmental documents such as text data on literature, research projects, laws, and regulation) or data with temporal reference (e.g., land use alterations of restoration areas, or seasonal fluctuations in dangerous substance measurements). GIS can neither cope with the problem of handling thematic data adequately nor of managing time series data.[5]

2.2 Media for Environmental Databases

As mentioned above, the media online, CD-ROM, and the Internet are used.

2.2.1 Online Databases and CD-ROMs

The term '*online* database' has been established for databases, which are offered to the public by specialized database providers, called hosts using national and international public data networks or the Internet. CD-ROMs (compact disk read only memory) were introduced in the mid 1980s and are still essentially a read only medium. A single CD-ROM has the capacity to store approximately 650 MB of data.

The world's greatest compilation of publicly available databases is the 'Gale Directory of Databases' (GDD). It comprises commercial online databases, CD-ROMs, diskettes, and magnetic tapes. It was created by merging 'Computer Readable Databases' with the 'Directory of Online Databases' and the 'Directory of Portable Databases'. The current version comprises 9207 databases.[6] The majority of databases (55%) are accessed online and only 23% are available on CD-ROM. This means that 5064 online databases and 2118 CD-ROMs are covered. The GDD treats several major subject classes, such as business, science and technology, health and life science, law, etc. Environmental information can be found in all of these sub-classes although most of the relevant data can be retrieved out of science and technology databases. The exact number of commercial environmental databases is unknown. Estimates made in a previous publication give the approximate number as 2500.[4]

For CD-ROMs several directories exist to help find the relevant CD-ROMs. Standard reference material is published both in printed form and on CD-ROM twice a year. Here the 'CD-ROMs in Print',[7] 'Handbuch lieferbarer CD-ROMs',[8] and the 'CD-ROM Directory'[9] are worth mentioning.

2.2.2 Internet Resources

The medium 'Internet' is a medium created by hundreds of developments that mark our movement toward the new millennium[10] (see ***Internet*** and ***The Internet as a Computational Chemistry Tool***). The number of Internet resources reached the mark of four million in 1995. According to an estimate of the Internet society the number of computers connected to the Internet will be around 120 million in the year 1999.[11] World-wide commercial online services attracted seven million people world-wide in 1995. Internet has served an estimated 20 million people so far. Growth in usage for the commercial services is estimated to be about 40% through 1997. Internet growth is estimated at being roughly twice the rate of the commercial online services.[10]

The question nowadays is not whether to exploit the electronic information superhighway, but how.

2.3 Subject Classes for Environmental Databases

The topic of the classification of environmental databases according to their subject was treated in a special issue of the journal DATABASE. This booklet was published under the title 'Environment Online, the Complete Environmental Series from DATABASE magazine'.[12] This issue includes three different articles which cover the following three different subject classes:

- general interest databases[13]
- scientific and technical databases[14]
- business and regulatory information.[15]

2.4 Classification Types for Environmental Databases

This is the most comprehensive way to categorize environmental databases. In the specialized literature of the information sciences, databases are divided into several types according to their database structure and their information types.[16] On the first level, fact-based, text-based, and integrated databases are distinguished (see Figure 2). The term fact-based databases is not clearly defined. It is commonly accepted that this database type is comprised of facts. However, bibliographic references are often given in factual databases, although they are of secondary importance.[17] The term 'fact' in this context is based on the concept of attributes and characteristics in database theory. Therefore fact-based information is more structured than textual information. In text-based databases the information type 'text' plays the leading role. On the one hand text is modeled or represented, on the other hand this is done by the text itself (abstract). It has to be stressed that most databases are 'mixtures' of text-based and fact-based databases. They are so-called heterogeneous databases.[17] Integrated databases means that they are comprised of a combination of textual, factual, graphical, tabular, etc. information.[16]

The following further classification of environmental databases can be made. An approach has been published by the author.[4] In this current publication only those types of databases which appear in at least two media given in Section 2.2 are included.

2.4.1 Fact-based Databases (FBDB)

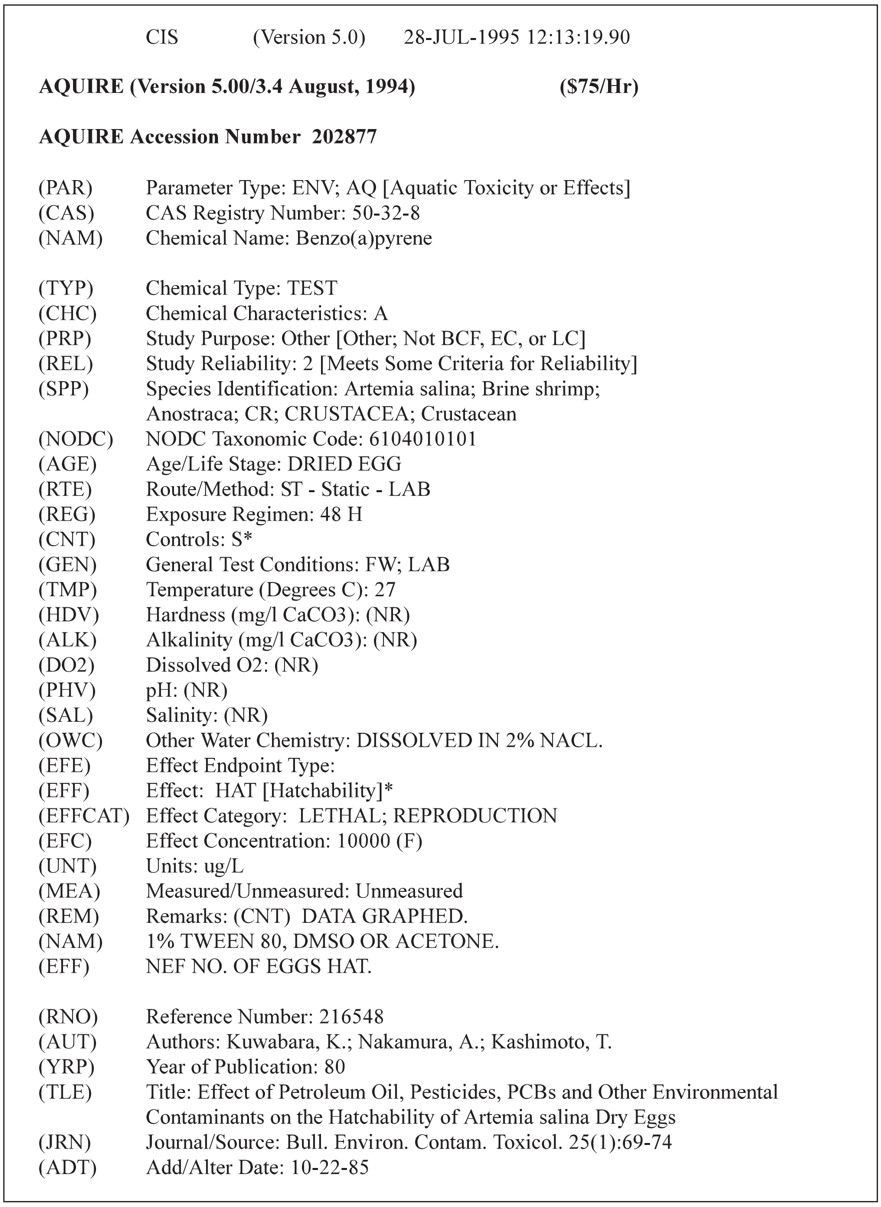
The fact-based databases are divided into *numerical (factual) databases, metadatabases*, research databases, and catalogues of chemical substances. These types will be explained in the following sections.

2.4.1.1 Numeric databases (NU). In numeric or factual databases the user can retrieve the information wanted, e.g., ecotoxicity data of a chemical substance, immediately. In factual databases in science and technology, data are not isolated but are stored in combination with other attributes and characteristics. According to Staud[16] they usually include the following three parts:

- a bibliographic section
- an information section
- a data section.

The following is an example of an online search for ecotoxicity information on the chemical substance benzo(*a*)pyrene as executed in the factual database AQUIRE (aquatic information retrieval system) at the host CIS. AQUIRE is a valuable database in the field of aquatic toxicity of chemical substances. It covers approximately 6000 chemical substances.

Search Example 1 CIS File: AQUIRE Item: 202877 (edited and abbreviated). Search was performed using CAS-Registry Number 50-32-8 as search term.



```
                CIS          (Version 5.0)      28-JUL-1995 12:13:19.90

AQUIRE (Version 5.00/3.4 August, 1994)                    ($75/Hr)

AQUIRE Accession Number  202877

(PAR)      Parameter Type: ENV; AQ [Aquatic Toxicity or Effects]
(CAS)      CAS Registry Number: 50-32-8
(NAM)      Chemical Name: Benzo(a)pyrene

(TYP)      Chemical Type: TEST
(CHC)      Chemical Characteristics: A
(PRP)      Study Purpose: Other [Other; Not BCF, EC, or LC]
(REL)      Study Reliability: 2 [Meets Some Criteria for Reliability]
(SPP)      Species Identification: Artemia salina; Brine shrimp;
           Anostraca; CR; CRUSTACEA; Crustacean
(NODC)     NODC Taxonomic Code: 6104010101
(AGE)      Age/Life Stage: DRIED EGG
(RTE)      Route/Method: ST - Static - LAB
(REG)      Exposure Regimen: 48 H
(CNT)      Controls: S*
(GEN)      General Test Conditions: FW; LAB
(TMP)      Temperature (Degrees C): 27
(HDV)      Hardness (mg/l CaCO3): (NR)
(ALK)      Alkalinity (mg/l CaCO3): (NR)
(DO2)      Dissolved O2: (NR)
(PHV)      pH: (NR)
(SAL)      Salinity: (NR)
(OWC)      Other Water Chemistry: DISSOLVED IN 2% NACL.
(EFE)      Effect Endpoint Type:
(EFF)      Effect:  HAT [Hatchability]*
(EFFCAT)   Effect Category:  LETHAL; REPRODUCTION
(EFC)      Effect Concentration: 10000 (F)
(UNT)      Units: ug/L
(MEA)      Measured/Unmeasured: Unmeasured
(REM)      Remarks: (CNT)  DATA GRAPHED.
(NAM)      1% TWEEN 80, DMSO OR ACETONE.
(EFF)      NEF NO. OF EGGS HAT.

(RNO)      Reference Number: 216548
(AUT)      Authors: Kuwabara, K.; Nakamura, A.; Kashimoto, T.
(YRP)      Year of Publication: 80
(TLE)      Title: Effect of Petroleum Oil, Pesticides, PCBs and Other Environmental
           Contaminants on the Hatchability of Artemia salina Dry Eggs
(JRN)      Journal/Source: Bull. Environ. Contam. Toxicol. 25(1):69-74
(ADT)      Add/Alter Date: 10-22-85
```

The above mentioned three parts ‘information’, ‘data’, and ‘bibliographic’ section are clearly outlined. The data-fields ‘CAS-Number (CAS)’ and ‘Chemical Name (NAM)’ belong to the information section, and the fields ‘Author (AUT)’, ‘Title (TLE)’, ‘Journal (JRN)’, etc. are regarded as bibliographic information. Important chemical identification parameters like structural formula, molecular formula, and molecular weight are not included in this factual environmental database. The main and most important part is the data section providing data on the toxicity of crustaceans including test conditions.

Other examples of numeric databases including environmental information are HSDB (Hazardous Substances Databank) and ECDIN (Environmental Chemicals Data and Information Network) (see ***Chemical Safety Information Databases***). HSDS gives extensive information about 4500 dangerous substances on practically all information types mentioned in Section 3. ECDIN is still available online and on CD-ROM. It comprises data on approximate 100 000 chemicals, but the given datasets on each substance are often not completely filled with actual data. For some time ECDIN has also been available on the Internet free of charge.[18]

2.4.1.2 Metadatabases (MD). Metadatabases are also categorized as fact-based databases. In this type of database, objects – in this case databases – are described with the aid of datasets. A prominent example is the already mentioned metadatabase ‘Gale Directory of Databases’ which is available in printed form, online, and on CD-ROM. In addition, every host has its own metadatabase with a description of each database offered online. Examples are STNGuide, DIALOG Bluesheets etc.

Metadatabases for environmental chemicals are established by the GSF-Research Center for Environment and Health in cooperation with the Bavarian State Ministry for State Development and Environmental Affairs. They will be described in Section 3.

2.4.1.3 Research databases (RD). In research databases or better – research project databases – research projects are presented with the aid of datasets.

A German example for this type of database is UFOR (Umweltforschungs-Datenbank, Environmental Research Database) produced by the German Environment Agency in Berlin. In this database more than 35 000 in-progress and completed research projects as well as approximately 8 000 research institutions are stored.[19] UFOR is offered by Data-Star, FIZ-Technik, GBI, and STN. Another German research database is FORKAT (Forschungsberichte aus Technik und Naturwissenschaft – Research Reports in Science and Technology) sponsored by the German Federal Ministry of Education, Science, Research, and Technology), which is online on STN. This database gives descriptions of approximately 25 000 research projects still in progress. CORDIS (Community Research and Development Information Service) is available on CD-ROM and is offered by the European host ECHO. ENREP (Environmental Research-in-Progress) is also available at ECHO. CRIS/USDA (Current Research Information System) is a database which covers research projects in science and technology in the United States. It is available at DIALOG.

2.4.1.4 Catalogues of chemical substances (CA). In this type of database, the objects presented are chemicals. The specific purpose of these catalogues of chemical substances is to provide information from specific chemical companies on their offerings of chemical substances. They provide not only data on prices, but also in many cases information on identification, physical-chemical, ecotoxicity, and toxicity parameters. Especially on the Internet, several chemical catalogues are available free of charge. Interesting examples are the catalogues Aldrich,[20] Fisher,[21] Fluka,[22] Sigma,[23] and Supelco,[24] just to mention a few.

2.4.2 Text-based Databases (TBDB)

The text-based databases are divided into bibliographic and full-text databases. They still completely dominate the fact-based and integrated databases in the online and CD-ROM categories.[6]

2.4.2.1 Bibliographic databases (BI). In contrast to full-text databases, bibliographic databases do not contain the text itself but only the text model. Often they encompass an abstract which is a summary of the information (knowledge) given in the text. To differentiate bibliographic databases from numeric databases one has to say that bibliographic databases describe information on objects and not the objects themselves (as fact-based databases do). The actual data and information on environmental questions and chemical substances can only be found in the primary literature. Most online databases are bibliographic ones.

Examples of major environmental bibliographic online databases are Pollution Abstracts and Enviroline, both offered by a couple of international hosts. Although these two databases are considered by Orton[25] to be the most important ones for searches on environmental questions, Orton advises combining the searches in these databases with other databases to get more rounded results (see ***Chemical Engineering: Databases*** and ***Chemical Safety Information Databases***). For questions on environmental chemicals he proposes searching the Chemicals Safety Newsbase and Chemical Abstracts. The important German database ULIT (Umweltliteratur-Datenbank = Environmental Literature Databank) produced by the German Federal Environment Agency (Umweltbundesamt) is online on Data-Star, FIZ-Technik, GBI and STN. It offers the best coverage of German information sources. Titles are given in German and English. It covers more than 260 000 documents and is explained in detail by Batschi.[19]

2.4.2.2 Full-text databases (FT). Text is considered in online databases as a means based on natural language to describe reality and to represent knowledge. Most online databases are text-oriented (word-oriented). 71% of the databases in the Gale Directory are text-oriented. Full-text databases have grown the most in recent years.[6]

In full-text databases the complete articles can be retrieved. They contain the complete text of a documentation unit. In most cases this text is completed with additional data-fields, e.g., descriptor, thesauri, classification code fields. An example of a full-text database is the CJRSC (Chemical Journals of the Royal Society of Chemistry) which is offered by STN. The document, shown in Search Example 2 on the topic of waste removal, is abbreviated. Apart from bibliographic information (author [AU], title [TI], source [SO], abstract [AB]) the full article with tables and figures can be retrieved. The information found is very comprehensive but also expensive in comparison to bibliographic databases (see ***Chemical Abstracts Service Information System*** and ***Inorganic Chemistry Databases***).

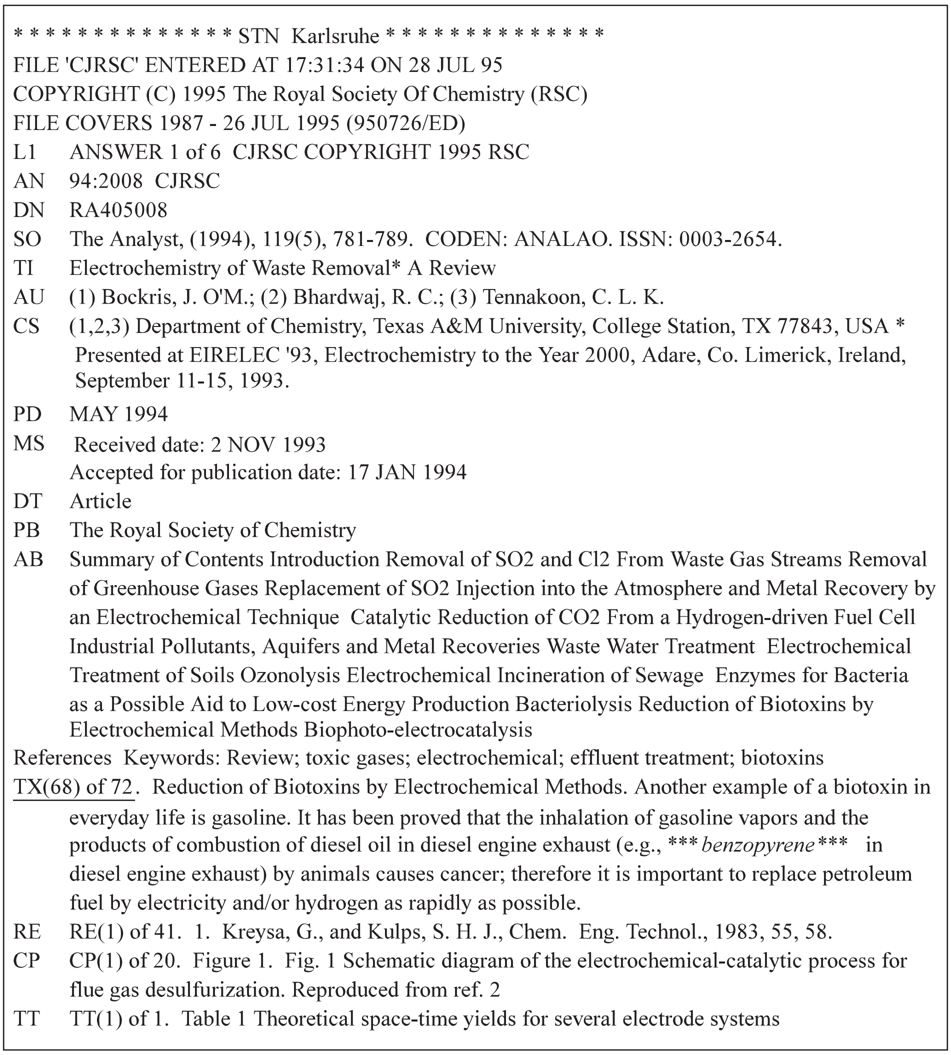
2.4.3 Integrated Databases (INDB)

The integrated databases are divided in this approach into structural and reaction databases. Other types of integrated databases are spectral databases and patent databases. Patent databases have an increasing importance in environmental questions related to technology.[26]

2.4.3.1 Structural databases (ST). Structural databases are databases which comprise chemical structures. The chemical structures are described for the searches in topological form (connection tables). The output results are displayed in graphical form.[17]

A prominent example of a structure database (see ***Structure Databases***) is the REGISTRY File offered by the host STN. This database contains both structure and nomenclature from the Registry System. This file is searched using the Messenger software developed by Chemical Abstracts Service, which provides both identity, full structure, and substructure searching as well as nomenclature searching (see ***Chemical Abstracts Service Information System***). Structure queries are entered as structure diagrams created with text commands or more conveniently drawn on a graphics terminal or on a personal computer with graphics software. CAS supplies a software package called STN Express for use on personal computers. With this software the user can construct a structure query offline, and then go online and transmit the query to STN for searching.[27] Other hosts like, e.g., Orbit.Questel, DIALOG Information Services, and CIS (Chemical Information Systems) offer structural databases as well.

Search Example 2 STN File: CJRSC Item: RA405008 (edited and abbreviated). Search was performed using the free term `benzopyrene'



```
* * * * * * * * * * * * * * STN  Karlsruhe * * * * * * * * * * * * * *
FILE 'CJRSC' ENTERED AT 17:31:34 ON 28 JUL 95
COPYRIGHT (C) 1995 The Royal Society Of Chemistry (RSC)
FILE COVERS 1987 - 26 JUL 1995 (950726/ED)
L1    ANSWER 1 of 6  CJRSC COPYRIGHT 1995 RSC
AN    94:2008  CJRSC
DN    RA405008
SO    The Analyst, (1994), 119(5), 781-789.  CODEN: ANALAO. ISSN: 0003-2654.
TI    Electrochemistry of Waste Removal* A Review
AU    (1) Bockris, J. O'M.; (2) Bhardwaj, R. C.; (3) Tennakoon, C. L. K.
CS    (1,2,3) Department of Chemistry, Texas A&M University, College Station, TX 77843, USA *
       Presented at EIRELEC '93, Electrochemistry to the Year 2000, Adare, Co. Limerick, Ireland,
       September 11-15, 1993.
PD    MAY 1994
MS     Received date: 2 NOV 1993
      Accepted for publication date: 17 JAN 1994
DT    Article
PB    The Royal Society of Chemistry
AB    Summary of Contents Introduction Removal of SO2 and Cl2 From Waste Gas Streams Removal
      of Greenhouse Gases Replacement of SO2 Injection into the Atmosphere and Metal Recovery by
      an Electrochemical Technique  Catalytic Reduction of CO2 From a Hydrogen-driven Fuel Cell
      Industrial Pollutants, Aquifers and Metal Recoveries Waste Water Treatment  Electrochemical
      Treatment of Soils Ozonolysis Electrochemical Incineration of Sewage  Enzymes for Bacteria
      as a Possible Aid to Low-cost Energy Production Bacteriolysis Reduction of Biotoxins by
      Electrochemical Methods Biophoto-electrocatalysis
References  Keywords: Review; toxic gases; electrochemical; effluent treatment; biotoxins
TX(68) of 72.  Reduction of Biotoxins by Electrochemical Methods. Another example of a biotoxin in
      everyday life is gasoline. It has been proved that the inhalation of gasoline vapors and the
      products of combustion of diesel oil in diesel engine exhaust (e.g., *** benzopyrene ***   in
      diesel engine exhaust) by animals causes cancer; therefore it is important to replace petroleum
      fuel by electricity and/or hydrogen as rapidly as possible.
RE    RE(1) of 41.  1.  Kreysa, G., and Kulps, S. H. J., Chem.  Eng. Technol., 1983, 55, 58.
CP    CP(1) of 20.  Figure 1.  Fig. 1 Schematic diagram of the electrochemical-catalytic process for
      flue gas desulfurization. Reproduced from ref. 2
TT    TT(1) of 1.  Table 1 Theoretical space-time yields for several electrode systems
```

Regarding structural databases on the Internet, the WWW chemical structures database established by the Computer Chemistry Center at the University of Erlangen should be mentioned. Currently this database contains more than 2250 chemical structures collected automatically from the Internet, complete with information about the referring html pages. The search operations supported in this database include full structure search, substructure search, formula-oriented searches, and queries on names, CAS-Numbers, etc.[28,29]

2.4.3.2 Reaction databases (RE). Reaction databases (see ***Reaction Databases***) are databases comprised of information on chemical reactions. These chemical reactions are stored as topological data. They consist of:

- connection tables of the compounds involved
- information on the reaction centres and
- the assignment of the atoms or atomic groups in the reaction.[17]

These chemical reaction databases contain not only reaction-type data, but also in most cases, bibliographic or factual information as well.

This type of integrated database is costly in production and needs special skill and/or special software to retrieve the information wanted. The need and importance for this kind of integrated data will drive forward the production of these databases and the development of retrieval software.

Famous examples of reaction databases are CASREACT, Chemreact, and ChemInformRX, all online on STN. None of the reaction databases covers the whole chemical literature, and furthermore there are time periods that are not included in any of the available databases.[30]

The conversion of the ChemInform weekly abstracting service supporting recent publications dealing with methods of preparation and synthesis in organic, organometallic, and inorganic chemistry is described by Gasteiger.[31,32] The wealth of information on chemical reactions can only be made available in a reaction database. A recent publication[33] describes new developments in this reaction database – especially in the inhouse version for use with MDLs ISIS software – concerning the retrieval and improvement in the evaluation of answer sets. Described is a unique classification feature that clusters retrieved reactions based upon their reaction centers.

A search example in ChemInformRX is given concerning the synthesis, chemical properties, and mutagenicity of 1,6- and 3,6-dinitrobenzo(*a*)pyrenes.

This example – which is edited and abbreviated – shows that reaction databases comply with the definition of integrated databases. The real data of reactions (tables, diagrams) are completed with bibliographic information, AU (author), TI (title), SO (source), AB (abstract). The data-fields RX, etc. (reactions and their diagrams) are factual information. Apart from factual and bibliographic data, this reaction database contains data-fields like KW (keywords) which describe the content of the article.

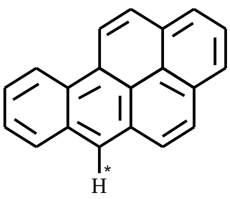
Millions of reactions are available in reaction databases for inhouse or online use. However, the increase in volume has not

Search Example 3 STN File: CHEMINFORMRX Item: RA405008 (edited and abbreviated).
Search was performed using the free-text term benzo(a)pyrene

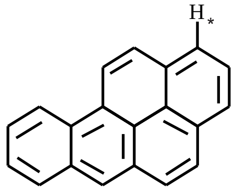
FILE 'CHEMINFORMRX' ENTERED AT 17:46:23 ON 28 JUL 95
COPYRIGHT (C) FIZ-CHEMIE BERLIN
FILE LAST UPDATED: 17 MAY 95 <950517/UP>

L1 ANSWER 1 OF 1 COPYRIGHT 1995 FIZ CHEMIE
AN 9123131 CHEMINFORMRX
TI Synthesis, Chemical Properties, and Mutagenicity of 1,6- and 3,6- Dinitrobenzo(a)pyrenes.
AU FUKUHARA, K.; MIYATA, N.; MATSUI, M.; MATSUI, K.; ISHIDATE, M. JUN.; KAMIYA, S.
CS Natl. Inst. Hygienic Sci., Kamiyoga, Tokyo 158, Japan
SO Chem. Pharm. Bull., 38(11), 3158-3161 (1990), CODEN: CPBTAL ISSN: 0009-2363
LA English
AB Nitration of the **benzo(a)pyrenes** (I) with nitric acid (II) yields the nitrobenzopyrenes (III) - (V) or (III) - (VII) depending on the reaction conditions. The mutagenic activities of all products are studied andcorrelated with their redox potentials.
CC Q1150 other polycyclic compounds
DED 19930426

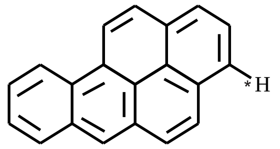
RX(1) OF 2 3 ***I* + 3 II ===> III + IV + V**



I



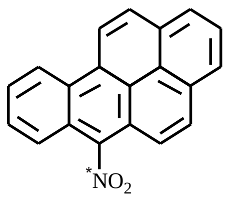
I



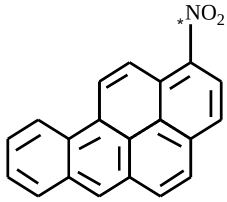
I

$O_2\overset{*}{N}$— OH

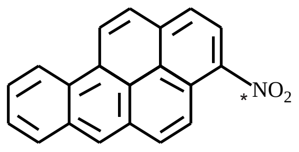
3 II



III
YIELD 44.0%



IV
YIELD 8.0%



V
YIELD 4.0%

RX(1) RCT I, ***157507*** (50-32-8)
II, 136 (7697-37-2)
RGT 4 (108-24-7), Ac2O
PRO III, 157508 (63041-90-7)
IV, 157509 (70021-99-7)
V, 157510 (70021-98-6)
YDS 56.0 %, T 20.0 Cel, TIM 24 hr

KW Arylation
NTE reaction: I (II) -> III + IV + V

necessarily benefited the end user to the extent it might have. Only recently have there been new software developments which are focused on helping end users extract more useful information from large reaction databases. These products were explained by Hayward at the 19th International Online Meeting 1995 in London.[34]

2.4.3.3 Beilstein and CrossFire and NetFire Database. The Beilstein Handbuch der Organischen Chemie contains evaluated numeric and factual data on millions of compounds. The complete Beilstein Online was available at the host STN in 1991. Also the host DIALOG offers the Beilstein database (see ***Factual Databases in Chemistry***).

CrossFire is the enhanced inhouse version of the Beilstein database. The database consists of about 30 million reports, of which five million are reactions. Hence this database is a reaction database. But apart from that, it covers structural, factual and bibliographic information. This means that CrossFire can also be classified as a structural and most of all as a numeric database. In a recently published article by Lawson,[35] an example is shown, illustrating the use of substance-, reaction-, and document-based search techniques on the same problem. The special power of hyperlinks between reactions, substances, and documents is shown.

A new and valuable literature based database and search system, called NetFire, has recently been made available on the Internet by Beilstein Information Systems. It is the first attempt by Beilstein to provide chemical information outside the main area of extracted and evaluated data. The existing Beilstein system, provided via their CrossFire system, is an in-house system which has received wide acceptance since it was introduced in the mid 1990s. NetFire is a very new and different system, a valuable Internet resource. It is yet another example of the innovation of new and modern electronic products by Beilstein.[36] NetFire provides free access to titles, abstracts, and authors from the organic chemical literature, from 1980 to the present.[37]

3 METADATABASES FOR ENVIRONMENTAL CHEMICALS

It has been demonstrated in Section 2 that the databases on environmental chemicals are very heterogeneous with respect to their types and contents. Knowing these various media, subjects, and types of environmental databases, a comprehensive summary of environmental databases is the next logical step in finding the relevant information on questions concerning environmental chemicals. As mentioned above in Section 2.4.1.2, metadatabases are set up and updated at the GSF-Research Center for Environment and Health. The following three different metadatabases have been established: DADB – Metadatabase of Online Databases, DACD – Metadatabase of CD-ROMs, and DAIN – Metadatabase of Internet Resources. Their current status is given in Table 1.

The set up of the metadatabases and the organization of data-fields was described in several publications.[38–40] Administrative, bibliographic, and content-related fields are given in each metadatabase. In this article only those data-fields treating the contents of the databases will be explained. The content-related data-fields have the same structure in every metadatabase, whereas the bibliographic ones vary according to the different media description structures.

Table 1 Metadatabases for Environmental Chemicals (Status April 1997)

Acronym	Name of Metadatabase	No. of entries
DADB	Metadatabase of Online Databases	470
DACD	Metadatabase of CD-ROMs	397
DAIN	Metadatabase of Internet Resources	96

Content-related fields are: ‘information type’, ‘use of chemicals’, ‘number of chemicals’, and ‘descriptor’. A few data-sources specialize in types of information, e.g., ecotoxicity or identification parameters. On the other hand, there exist special databases for specific uses of chemical substances, e.g., pesticide databases. As there are databases which contain thousands of chemicals and others which only describe hundreds, the field ‘number of chemicals’ is given. For a comprehensive description of the content of a source, a thesaurus containing key words which are of interest to the problem of environmental chemicals, has been developed. These key words can be found in the descriptor field. The thesaurus, which takes into account most of the parameters required by the German Chemicals Act[41] for registering chemical substances, treats not only environmental but also health and workplace exposure aspects. The thesaurus encompasses, for example, the identification of chemicals, data on detection of chemicals in the environment, use of chemicals, economic data, physical–chemical properties, degradation and accumulation data, ecotoxicity, effects on wildlife, toxicity effects on mammals, effects on human organisms, information in relation to the workplace, etc.

A search in these three metadatabases gives indications where to find, that is to say, in which primary data-source(s) to find, the environmental information needed. Concerning online databases, the access to the host(s) in question must be realized in order to search the primary online database(s). CD-ROMs must be purchased. In the case of Internet resources for environmental chemicals these can be searched directly as a link between meta-information and the primary Internet resource is given. The Metadatabase of Internet Resources for Environmental Chemicals is produced in cooperation with the University of Kassel, described by Benz[40] and found under the URL http://dino.wiz.uni-kassel.de/dain.html.[42] As it covers Internet resources which are free of charge, the user of DAIN can link to the primary resource and retrieve the actual data, e.g., biodegradation pathways of chemical substances in the Biocatalysis/Biodegradation database.[43] The other two metadatabases concerning environmental chemical information in online databases and CD-ROMs will soon be available on CD-ROM (for further information please contact the author).

4 EVALUATION OF METADATABASES FOR ENVIRONMENTAL CHEMICALS

Again several methods exist to evaluate the content of these metadatabases. The database types found in each metadatabase can be examined. It is also of extreme importance to know what kind of information is found in each metadatabase representing a medium.

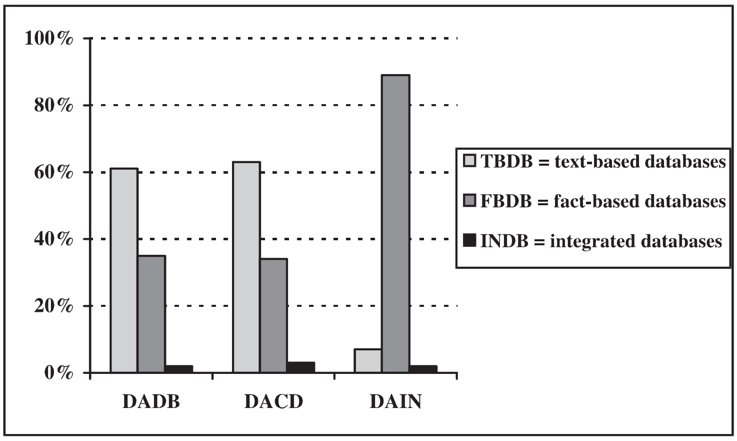


Figure 3 Comparison of three metadatabases concerning database types

4.1 Comparison of Database Types in DADB, DACD, and DAIN

In Figure 3, the percentage is shown for the three discussed types of databases: text-based; fact-based, and integrated databases. As the metadatabases are quite different in size, especially the newest Metadatabase of Internet Resources – DAIN, which still comprises under 100 entries, comparisons are quite vague and only show trends.

In Figure 3, text-based (TBDB), fact-based (FBDB), and integrated databases (INDB) are compared in the three discussed metadatabases. The percentage of each category is shown. The following conclusions can be drawn:

- Text-based databases are still predominant in DADB and DACD. That means that in online databases and CD-ROMs still more bibliographic and full-text databases are found than fact-based and integrated databases. In both DADB and DACD over 60% of the evaluated databases are text-based.
- Fact-based databases have a percentage of over 30% both in DADB and DACD, which means that approximately one third of all online databases and CD-ROMs in the field of environmental chemistry are fact-based.
- Integrated databases are the least represented type of all. Only 2–3% are found in all evaluated metadatabases. Not only in online databases and CD-ROMs but also in Internet resources, is this type very poorly represented.
- Concerning text-based and fact-based databases, the Metadatabase of Internet Resources gives completely different results from DADB and DACD. In DAIN most of the databases are fact-based (89%) and only 8% are text-based. This situation is in extreme contrast to the other two media. For this reason the DAIN results were investigated further.

4.2 Further Investigation of Types in DAIN-Metadatabase of Internet Resources

This metadatabase was established in late 1995. It is the result of a cooperation of the University of Kassel with the GSF-Research Center for Environment and Health.[40] The URL for DAIN is http://dino.wiz.uni-kassel.de/dain.html. DAIN focuses on the so-called ‘free’ Internet resources, that is to say those sites which are available without paying any charge. This metadatabase is in the English language and its main advantage in comparison with the two other metadatabases is the possibility of linking to the retrieved databases and hence getting the desired data immediately.

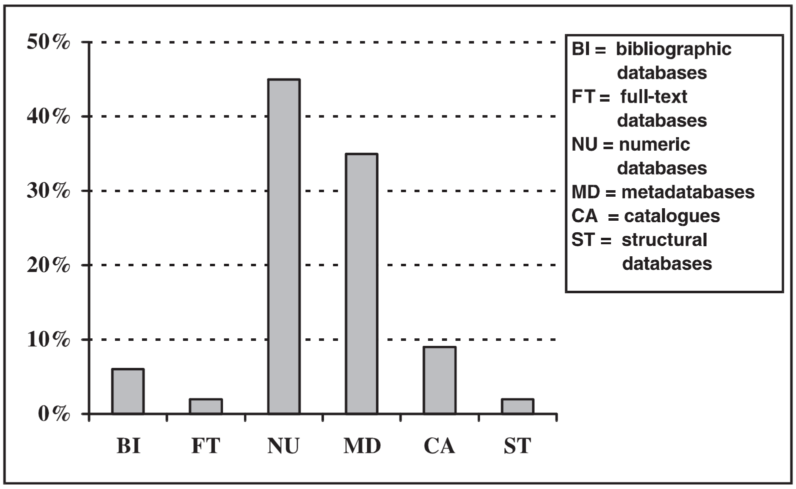


Figure 4 Database types in DAIN

The discussed situation concerning fact-based and text-based resources differs greatly from that of DADB and DACD. We must therefore consider types of databases which are covered in DAIN. In Figure 4 the different types given in Section 2.4 are shown.

Apart from the fact that only very few bibliographic (BI) and full-text (FT) databases are covered in DAIN, the situation in the fact-based area is of significance. Here 45% are numeric databases (NU) and 35% are metadatabases (MD). Chemical catalogues (CA) are also well represented.

Numeric databases (NU) are often collections of Material Safety Data Sheets (MSDS). In addition several factual sites treat pesticides and dangerous substances. For details see the DAIN Metadatabase on the Internet.[42]

In chemical and environmental sciences quite a few approaches exist to gathering the relevant resources. These are lists of web-pages with or without short descriptions of the original site. Heller suggests relevant chemical metadatabases on the WWW[44] and Reichardt gives guidance for environmental Internet resources.[45] The question is, why are there so many resources which refer to other resources on the same subject? Heller gives some possible explanations for that phenomenon. Sites have been put together as a side project or hobby, hence they only have little information, often highly specialized. Information on other resources can be put together in a shorter period of time than setting up factual databases.

The result of the evaluation of DAIN concerning catalogues from chemical suppliers is positive. The Fisher Chemical catalogue[21] and the SIGMA catalogue[23] are significant examples. As already mentioned above, these databases comprise not only information on identification parameters and use of chemical substances but also physical–chemical and toxicity data.

4.3 Parameter Discussion in Three Metadatabases

As mentioned in Section 3, environmental–chemical parameters are used to describe the contents of the databases in the metadatabases. For a comparative evaluation of the content of the sources, the following parameters are investigated: Identification parameters like CAS-Number and structural formula, physical–chemical parameters, e.g., boiling point and n-octanol–water partition coefficient, biodegradation, fish toxicity, acute mammalian toxicity, chronic toxicity, mutagenicity, and health aspects for workers. Only factual databases will be looked at. Table 2 gives the percentual results in all three metadatabases. In DADB 110, DACD 103, and DAIN 43 numeric databases are described.

Again the results of this evaluation have to be seen in context with the different sizes of the metadatabases. Therefore the conclusions drawn out of them can only be very rough ones. The bold figures indicate that the metadatabase in question has received the best result concerning this parameter. In factual online databases the situation with respect to CAS-numbers is the best, followed by CD-ROMs and Internet resources. CD-ROMs dominate the structural formula category of environmental databases. The percentage of 22 is far from satisfactory but a positive trend can be observed. This situation was also described by Warr who already mentioned this trend in 1993.[46] Several CD-ROM producers take the urgent need for structural information into account and incorporate this important identification criterion in their databases. Distribution coefficients like, e.g., n-octanol–water partition coefficients are generally not well represented in databases for environmental chemicals. The situation is particularly unacceptable as such data are urgently needed for modeling the behavior of chemicals in the environment.[47] Data concerning the workplace environment are found on many CD-ROMs. Several CD-ROM databases are explicitly set up for this purpose. Looking at biodegradation, fish toxicity, acute and chronic toxicity, and mutagenicity, the Internet resources outnumber both online and CD-ROM databases. Although this result has to be seen in the light of the fewer datasets in DAIN, it shows a trend in the data situation on environmental chemicals on the Internet. Many data collections which do cover ecotoxicity and toxicity parameters are offered free of charge.

Table 2 Percentage Status of Factual Databases in Three Metadatabases (Status April 1997)

Parameter	DADB	DACD	DAIN
CAS-Number	**82**	70	63
Structural Formula	9	**22**	9
Boiling Point	44	53	**63**
Distribution-Coefficients	**12**	**12**	5
Biodegradation	17	9	**37**
Fish Toxicity	19	12	**30**
Acute Oral Toxicity	31	32	**52**
Chronic Toxicity	16	16	**47**
Mutagenicity	22	20	**58**
Health Aspects for Workers	29	**50**	49

5 CONCLUSIONS AND OUTLOOK

Environmental databases are not only found in every media, that is to say online, CD-ROM, and on the Internet, but are quite varied in their type and contents. The environmentally interested community urgently needs support in finding relevant information and data about environmental chemicals. A first step in helping them is to set up and distribute metadatabases on the subject. However, this can only be a beginning. The three media categories, online, CD-ROM, and Internet, differ considerably in the types of databases offered and their contents.

A further step is to give precise indications of the importance and quality of the databases. This means performing a comparative evaluation of databases with respect to several different evaluation criteria. Approaches using a mathematical method of lattice theory called Hasse algebra are used and a program named Hasse was developed at the GSF. The Hasse-diagram technique is especially useful in the evaluation of environmental objects, such as chemicals, models, and databases.[2]

6 RELATED ARTICLES

Chemical Abstracts Service Information System; Reaction Databases; Structure Databases; Chemical Engineering: Databases; Chemical Safety Information Databases; Factual Databases in Chemistry; Inorganic Chemistry Databases; Internet; The Internet as a Computational Chemistry Tool; Online Databases in Chemistry.

7 REFERENCES

1. N. M. Avouris and B. Page, ‘Environmental Informatics, Methodology and Applications of Environmental Information Processing, Introduction’, eds. N. M. Avouris and B. Page, EURO Courses, Computer and Information Science, Vol. 6, Kluwer, Dordrecht, 1995, ix.
2. K. Voigt and R. Brüggemann, *Toxicology*, 1995, **100**, 225–240.
3. A. Baumewerd-Ahlmann and L. Zink, *Umweltdatenbanken*, eds. B. Page and L. M. Hilty, Umweltinformatik, Informatikmethoden für Umweltschutz und Umweltforschung, 2. Auflage, Oldenbourg, München, 1995, pp. 101–124.
4. K. Voigt and R. Brüggemann, ‘Meta Information System for Environmental Chemicals’, eds. N. M. Avouris and B. Page, Environmental Informatics, EURO Courses, Computer and Information Science, Vol. 6, Kluwer, Dordrecht, 1995, pp. 315–336.
5. B. Page, ‘Database Technologies for Environmental Data Management’, eds. N. M. Avouris and B. Page, Environmental Informatics, EURO Courses, Computer and Information Science, Vol. 6, Kluwer, Dordrecht, 1995, pp. 39–51.
6. M. E. Williams, ‘Highlights of the Online Database Industry and the Internet’, ed. M. E. Williams, Proceedings of the Seventeenth National Online Meeting, Online Databases, Library Systems, The Internet, CD-ROMs, Information Today, Medford, 1996, pp. 1–4.
7. Meckler Managing Information Technology, ‘CD-ROMs in Print 1995: An International Guide to CD-ROMs, CD-I, CDTV, Multimedia & Electronic Book Products, Meckler Managing Information Technology, Esport, 1996.
8. Scientific Consulting Dr. Schulte-Hillen BDU ‘Handbuch der Datenbanken für Naturwissenschaft, Technik, Patente’, 1996, Verlag Hoppenstedt, Darmstadt, 1996.
9. TFPL Publishing, ‘The CD-ROM Directory’, 1996, TFPL Publishing, London, 1996.
10. S. E. Arnold, ‘Publishing on the Internet, A New Medium for a New Millennium’, Infornortics, Calne, 1996, pp. 5–7.
11. Internet Society Network Information Centre, *Byte Magazine*, 1995, July, 69.
12. P. Hane, ‘Environment Online: The Greening of Databases, The Complete Environmental Series from DATABASE Magazine’, Eight Bit Books, Wilton, 1992.
13. F. W. Stoss, *DATABASE*, 1991, **14**, 4, 13–27.
14. P. Gayle Alston, *DATABASE*, 1991, **14**, 5, 34–52.
15. P. Gayle Alston and F. W. Stoss, *DATABASE*, 1992, **14**, 6, 17–35.
16. J. L. Staud, ‘Online Datenbanken, Aufbau, Struktur, Abfragen’, Addison-Wesley, Bonn, 1991, pp. 17–33, 171–173, 220.
17. A. Barth, ‘Datenbanken in den Naturwissenschaften, Eine Einführung in den Umgang mit Online Datenbanken’, VCH, Weinheim, 1992, pp. 191, 404, 411, 413.
18. Joint Research Centre, ECDIN, **http://ulisse.etait.eudra.org/ecdin/ecdin.html**
19. W.-D. Batschi, ‘Umwelt Technologie Aktuell’, 1993, **4**, 354–359.
20. Sigma-Aldrich Corporation. Aldrich Browsable On-line Catalog, **http://www.sigma.sial.com/**
21. Fisher Scientific, The Fisher Chemical Catalogue, **http://www.fisher1.com/index.html**
22. Fluka, FLUKA Browsable On-line Catalog, **http://www.fluka.sial.com/**
23. Sigma, Chemical Company, SIGMA Catalog, **http://www.sigma.sial.com/**
24. Supelco, Supelco Browsable On-line Catalog, **http://www.fluka.sial.com/**
25. D. Orton, ‘Database Reviews: Environmental’, ed. D. Orton, Online Searching in Science and Technology, The British Library, London, 1995, pp. 75–82.
26. U. Georgy, *Cogito*, 1993, **1**, 22–25.
27. G. G. Van der Stouw, ‘Online Searching for Chemical Information’, eds. R. T. Bottle and J. F. B. Rowland, Information Sources in Chemistry, Bowker Saur, London, 1994, pp. 67–103.
28. W. D. Ihlenfeldt, ‘Chemical Structure Search on the World Wide Web’, Learned Information Ltd, 20th International Online Information Meeting, London 3–5 December 1996, Learned Information, Oxford, 1996, pp. 135–142.
29. W. D. Ihlenfeldt, The WWW Chemical Structures Database Homepage and description on the database, **http://schiele.organik.uni-erlangen.de/services/webmol.html**
30. Y. Wolman, ‘Chemical Reaction Databases’, Learned Information Ltd., 19th International Online Information Meeting, London 5–7 December 1995, Learned Information, Oxford, 1995, pp. 157–162.
31. J. Gasteiger and C. Weiske, ‘CHEMINFORM – An Integrated Information System on Chemical Reactions’, Learned Information Ltd., Online Information 89, Learned Information Ltd., Oxford, 1989, pp. 147–154.
32. J. Gasteiger and C. Weiske, *Nachr. Chem. Tech. Lab.*, 1992, **740**, 10, 1114–1120.
33. J. T. Bohlen, ‘ChemInform Electronic Journal and ChemInformRX Reaction database – New Developments’, ed. J. Gasteiger, Software-Entwicklung in der Chemie 10, GDCh-Fachgruppe Chemie-Information-Computer, Gesellschaft Deutscher Chemiker, Frankfurt, 1996, pp. 27–32.
34. J. Hayward, ‘Chemical Reaction Databases: Progress in Reaction Searching’, Learned Information Ltd., 19th International Online Information Meeting, London, 5–7 December 1995, Learned Information, Oxford, 1995, pp. 143–156.
35. A. J. Lawson, ‘Organic Synthesis Pathways in CrossFire’, Learned Information Ltd., 19th International Online Information Meeting, London, 5–7 December 1995, Learned Information, Oxford, 1995, pp. 131–141.
36. S. R. Heller, The Beilstein NetFire System, TrAC – Internet Column, **http://www.elsevier.nl:80/inca/homepage/saa/trac/netfire.htm**
37. Beilstein Information Systems, NetFire Information, the Beilstein Abstracts Database, **http://www.beilstein.com/netfire/netfire.html**.
38. K. Voigt, M. Matthies, and T. Pepping, *Toxicol. Environ. Chem.*, 1993, **40**, 83–93.
39. K. Voigt, *Cogito*, 1995, **6**, 12–15.
40. J. Benz and K. Voigt, ‘Indexing File System for the Set up of Metadatabases in Environmental Sciences on the Internet’, Learned Information Ltd., 19th International Online Information Meeting, London, 5–7 December 1995, Learned Information, Oxford, 1995, pp. 455–466.
41. Chemikaliengesetz – Chem.G.vom 16. September 1980, (BGBl. I S.1718, geändert durch Gesetz vom 25. Juli 1994, BGBl.I S.1703, pp. 1440, 1963, 2705.
42. K. Voigt and J. Benz, DAIN – Metadatabase of Internet Resources for Environmental Chemicals. Homepage and description on the database, **http://dino.wiz.uni-kassel.de/dain.html**
43. University of Minnesota, Biocatalysis/Biodegradation Database, **http://www.dragon.labmed.umn.edu∼lynda/index.html**
44. S. R. Heller, *J. Chem, Inf. Comput. Sci.*, 1996, **36**, 205–213.
45. T. Reichardt, *Environ. Sci. Technol.*, 1996, **30**, 2, 76–81.
46. W. Warr, *DATABASE*, 1993, **2**, 59–62.
47. M. Matthies, *UWSF*, 1991, **3**, 1, 37–41.