Ann. Ist. Super. Sanità, vol. 30, n. 4 (1994), pp. 411-416

On-line information sources on chemical substances

Maria CASTRIOTTA and Adriana DRACOS

Servizio Documentazione, Istituto Superiore di Sanità, Rome, Italy

Summary. - Information technology has brought about changes in the work patterns of researchers and scientists. After some hints on the on-line facilities needed to be connected to the international host computers, an analysis is made of some of the main automated sources available to retrieve information on chemical substances. Special emphasis is given to textual-numeric data banks, first reviewing the main chemical dictionaries, like Registry and Chemline, and then focusing on those sources that offer immediate information in case of emergency. Among the Toxnet files, produced and managed within the US National Library of Medicine Toxicology Information Program, play a very important role in offering publicly available data on toxicology and on hazardous chemicals. Therefore, the Hazardous Substances Data Bank (HSDB) and the Registry of Toxic Effects of Chemical Substances (RTECS) are described for their relevance thereon. Other data banks produced in Europe, like the Environmental Chemicals Data Information Network (ECDIN) and the very specialized Major Hazard Incident Data Service (MHIDAS) are also briefly outlined. To integrate this overview on online information, the attention is then shifted on sources having the characteristic of reference databases: prestigious files covering the international scientific literature, as CA/Chemabs, Toxline/Toxlit, Embase, Medline are introduced. Implications of on-line technology in enhancing information access in the next future are discussed, pointing out the new tools created to meet the information needs of end-users.

Key words: chemicals, bibliographic databases, factual databases, on-line systems.

Riassunto *(Fonti di informazioni su linea per sostanze chimiche)* - Le moderne tecnologie dell'informazione hanno contribuito a modificare il *modus operandi* di ricercatori e scienziati. Dopo alcuni cenni sugli strumenti necessari per collegarsi con i distributori delle basi di dati su linea, si analizzano qui alcune delle maggiori fonti di informazione sulle sostanze chimiche. In particolare, vengono delineate le caratteristiche delle banche dati di tipo testuale-numerico, esaminando dapprima i dizionari chimici maggiori, come il Registry e il Chemline, e quindi mettendo a fuoco in particolare quelle fonti che offrono l'informazione immediata in caso di emergenza. Tra queste, le banche dati che afferiscono alla rete Toxnet, prodotte e gestite nell'ambito del Toxicology Information Program (TIP) dalla statunitense National Library of Medicine, giocano un ruolo molto importante mettendo a disposizione del pubblico i dati tossicologici sui composti chimici. Vengono quindi illustrati, per la loro rilevanza in materia, gli archivi Hazardous Substances Data Bank (HSDB) e Registry of Toxic Effects of Chemical Substances (RTECS). Segue l'esame di alcune banche dati prodotte in Europa, come Environmental Chemicals Data Information Network (ECDIN) e Major Hazard Incident Data Service (MHIDAS), specializzata nel registrare gli incidenti che hanno dato luogo al rilascio di sostanze. Per integrare questa breve rassegna sull'informazione chimica su linea, non possono essere tralasciate quelle fonti che hanno la caratteristica di basi di dati bibliografiche. Vengono quindi presentati archivi prestigiosi riguardanti la letteratura scientifica internazionale, come CA/Chemabs, Toxline/Toxlit, Embase, Medline. In conlusione, si danno brevi cenni sui nuovi strumenti messi a punto nel campo dei supporti elettronici dell'informazione, per le esigenze di un pubblico sempre più vasto.

Parole chiave: composti chimici, basi di dati bibliografiche, basi di dati fattuali, sistemi su linea.

Introduction

Modern society has to live together with an evergrowing quantity of chemicals, consisting in products and by-products of a great variety of industrial activities.

In such a wide market, the prompt availability of any information regarding physical-chemical properties and toxicological data of chemical substances on the market is fundamental, both for governments, who have to regulate and control trade and for the scientific community.

As a matter of fact, information technology has contributed to change the work patterns of researchers with the creation of electronic files, which, if adequately exploited, increase retrieval of data and improve quality of searches.

The Documentation Service where the authors work, having to respond, day by day, to the numerous queries of scientists and researchers of the Istituto Superiore di Sanità, represents a privileged observatory of the questions arising in a scientific environment.

The aim of this paper is to give an overview of the main on-line databases containing information on chemicals.

The on-line background

To get an on-line connection, the user needs a telephone, a computer emulating a terminal, a modem and a contract with a host computer; this is an organization equipped with a mainframe which distributes the on-line files. Once the user has got the connection with the host computer, he/she selects the database he/she is interested in and begins an interactive conversation, based on the query language of the system.

What seems now so simple and speedy is the result of a long effort. The development and use of on-line database services, thanks to new technologies and capabilities that have come together, began in the mid Sixties.

Since then interest in the on-line market has continued to grow, due both to the increasing number of databases produced (from 400 files in 1979-80 to more than 5300 reported in July 1992), and to the implementation of telecommunication networks [1, 2].

An important innovation in this sector is now Internet, defined as a network of networks, the greatest open system existing today. It is used to worldwide connect academic and research institutions, that can in this way avoid telecommunication charges of commercial and private networks. Most host computers are now equipped to offer their on-line services on Internet.

One of the difficulties of on-line search is however the ability to select the appropriate sources for the topic to be searched among the increasing number of databases available. The choice has to be made considering various aspects: some files are characterized by the high number of substances included, others by the high number of parameters considered, and both features are seldom together [3].

On the other hand, not all parameters considered in a database are actually available for every chemical included, whilst their absence does not mean that the information required may not be retrieved elsewhere. Hence it appears that a negative answer from a single file has never to be accepted as final [3].

The first wide distinction regarding on-line files pertains to their structure: they can be classified as source databases or reference databases.

Source databases are generally called data banks for their content of original source data, numeric or textual-numeric; full-text databases are also classified in this group.

Reference databases contain secondary source citations, generally with abstracts to the published literature, and therefore they are also called bibliographic files. In which cases is it better to exploit a data bank or a bibliographic database? If the on-line search has to face with an emergency situation (for instance, a chemical spill from a truck, or a case of substance poisoning), where measures must be taken urgently, data banks constitute the main information sources. On the other hand, if the subject to be investigated needs deep attention and analysis of the most recent literature, bibliographic files should preferably be exploited.

The integration of both kinds of files is often advisable, because of the different sources examined, and also if one considers that even the largest data bank cannot contain all of the substances worldwide identified.

Source databases

The most part of source databases regarding chemicals can be classified as textual-numeric files. They include databases with dictionary or handbook-type data. The main characteristic of this kind of files is that to every chemical identified corresponds only one record. The first problem arising in chemical information retrieval is the proper identification of the substance involved. A fundamental step for the solution of this question was the creation of the Chemical Abstract Service-registry number (CAS-RN). This system derives from CAS's decision in the early 1960's to use the computer to deal with chemical literature, which grew more and more. The system automatically identifies structural diagrams and assigns to each a unique CAS-RN. Between 1965 and the end of 1991 more than 11 million unique structures had been recorded [4]. New structures are being entered at a rate of about 650,000 per year.

The CAS-RN has come now to play a great role by overcoming the ambiguities of chemical nomenclature and uniquely identifying compounds in a wide variety of chemical files [5].

Chemical dictionaries

The development of on-line chemical dictionaries was a consequence of CAS-RN system. Actually the user needed to refer to a file leading from synonyms, tradenames, or standard nomenclature to CAS-RNs and *vice versa*.

One of the first examples of this kind is Chemline. Built in the early 1970's jointly by the US National Library of Medicine (NLM) and CAS, it contains nomenclature and structure information for about 1,3 million chemical substances included in NLM files and in the Toxic Substances Control Act (TSCA) Inventory of the US Environmental Protection Agency (EPA). Searchable fields are CAS-RNs, molecular formulas, formula fragments, ring analysis terms and synonyms.

In 1980 CAS began to offer access on-line directly to the content of the entire CAS system and since then another important on-line chemical dictionary is available on-line, through the host computer STN: the file Registry.

It contains 12,3 million unique substance records identified by CAS since 1957. A considerable feature of the file Registry is the capability of structure searching by selecting structure fragments from a menu, drawing on a graphic terminal, or typing commands on the keyboard [1]. In this way the user can perform a search starting with a molecular structure to verify whether the substance has already been identified; moreover, the ten most recent documents citing the substance are displayed. It is therefore quite reductive to classify the Registry file simply as a chemical dictionary.

Both Registry and Chemline have the Locator field that indicates, respectively, other STN or NLM files containing information on the substance and indicating regulatory listings where the chemical occurs.

US data banks

When Chemline was first created thanks to a cooperation between NLM and CAS, NLM was already a pioneer in the production of toxicological files. Actually the NLM Toxicology Information Program (TIP) was founded in 1967 because of the increased public awareness about chemical hazards.

TIP functions were above all the creation of complete computer-based files of toxicological information and, further, the realization of every action aimed to render access to these files generally available.

NLM toxicological data banks have become part of an information system called Toxnet. The family of Toxnet files is now quite numerous and it includes also some bibliographic files (Table 1).

HSDB and RTECS are data banks covering general aspects of toxicology; the other files are more specific and subject-oriented within the same field.

For instance, Chemical Carcinogenesis Research Information System (CCRIS) is sponsored by the US National Cancer Institute (NCI) and includes scientifically evaluated data derived from carcinogenicity, mutagenicity, tumor promotion and tumor inhibition studies. It contains over 5000 chemical records.

The Integrated Risk Information System (IRIS) is sponsored by EPA. It contains EPA health risk and regulatory information on 658 chemicals.

GENE-TOX, another EPA file, contains genetic toxicology data on around 3000 chemicals, resulting from expert reviews of the scientific literature.

The Developmental and Reproductive Toxicology (DART) is supported jointly by EPA and the US National Institute of Environmental Health Sciences. It is a bibliographic database containing citations to publications concerning teratology and developmental toxicology. With its backfile ETICBACK, it contains about 68,000 bibliographic references.

In the following attention is focused on Hazardous Substances Data Bank (HSDB) and Registry of Toxic Effects of Chemical Substances (RTECS), because of their relevance.

HSDB produced by NLM, is a factual data bank focusing upon the toxicology of about 4400 potentially hazardous chemicals. HSDB was first launched in 1985 with the name of Toxicology Data Bank (TDB), when it had already been tested for about 15 years and had got a general approval which stimulated NLM to continue and to improve the initiative.

In addition to toxicity data, HSDB carries information in such areas as emergency medical treatment, safety and handling, environmental fate, exposure potential and US-oriented regulatory requirements. The file is organized into 11 categories subdivided into 144 fields. Table 2 lists the fields contained in the category Toxicity/ Biomedical effects (TOXB).

What renders HSDB a very precious information source on chemicals is not only the great quantity of fields included, but also the reliability and quality of the

Table 1. - The family of Toxnet files

Name of database	Type of database/bank
Hazardous Substances Data Bank (HSDB)	factual
Registry of Toxic Effects of Chemical Substances (RTECS)	factual
Chemical Carcinogenesis Research Information System (CCRIS)	factual
Directory of Biotechnology Information Resources (DBIR)	directory
Environmental Mutagen Information Center Back File (EMICBACK)	bibliographic
Environmental Teratology Information Center Back File (ETICBACK)	bibliographic
Toxic Chemical Release Inventory (TRI87, TRI88, TRI89, TRI90)	factual
Integrated Risk Information System (IRIS)	factual
Developmental and Reproductive Toxicology (DART)	bibliographic
The Gene-Tox program of the US Environmental Protection Agency (GENE-TOX)	factual
Toxic Chemical Release Inventory Facts (TRIFACTS)	factual

data. Actually, HSDB undergoes a high level of peer review by a panel of expert toxicologists and other scientists, whose main task is to evaluate how the data were obtained [5].

The RTECS is produced by the US National Institute for Occupational Safety and Health (NIOSH). It contains toxic effects data on over 122,000 chemicals. Both acute and chronic effects are covered, including data on skin/eye irritation, carcinogenicity, mutagenicity and reproductive effects.

The six fields making part of the category TOXB are reported in Table 3.

A typical toxicity field includes route, species, study type, dose, effect and reference.

RTECS has also a good coverage on chemical substance regulations, containing data on threshold limit values, NIOSH recommendations, standards and regulations.

Table 2. - Fields included in the category TOXB of HSDB file

Acronym	Field
TOXS	Toxicity summary
TXHR	Toxic hazard rating
EMT	Emergency medical treatment
EMLS	Life support
EMCE	Clinical effects
EMLAB	Laboratory
EMTR	Treatment overview
EMTOX	Range of toxicity
ANTR	Antidote and emergency treatment
MEDS	Medical surveillance
TOXS	Toxicity excerpts
HTOX	Human toxicity excerpts
NTOX	Non-human toxicity excerpts
TOXV	Toxicity values
HTXV	Human toxicity values
NTXV	Non-human toxicity values
ETXV	Ecotoxicity values
NTP	National toxicology program studies
IARC	IARC summary and evaluation
TCAT	TSCA test submissions
POPL	Populations at special risk
PHMK	Pharmacokinetics
ADE	Absorption, distribution and excretion
METB	Metabolism/metabolites
BHL	Biological half-life
ACTN	Mechanism of action
INTC	Interactions

European data banks

The sources examined until now are produced in the United States and for this reason they sometimes lack European information, particulary legislation.

A file that has to be mentioned for its European coverage is doubtless the Environmental Chemicals Data and Information Network (ECDIN).

ECDIN is produced by the Commission of the European Union Joint Research Centre at Ispra. It contains chemical identification data on approximately 130,000 compounds, generally used in the European countries, produced in large quantities or representing an actual or potential risk [3]. The file is structured into 62 parameters, grouped into 9 categories, regarding toxicity, environmental impact and safety data.

Not all records have complete information: acute toxicity data are available for approximately 20,000 substances; data regarding environmental impact and safety are available for less than 2000 chemicals [3].

The sources considered are handbooks, monographs, journals and research reports. Some portions of the International Register of Potentially Toxic Chemicals (IRPTC) are included in ECDIN as a result of a cooperation between the Commission of the European Union and the United Nations Environment Programme.

ECDIN is distributed on-line by DIMDI in Cologne, Germany, and can be searched in a user-friendly language.

A British specialized data bank of particular interest for chemical search is the Major Hazard Incident Data Service (MHIDAS). It contains references to incidents involving hazardous materials, reporting date and place of the incident, material name and quantity of substance released, brief description of the event. Its coverage dates back to 1964.

It is an invaluable source for information on chemical accidents. Its quality is guaranteed by the Safety Reliability Directorate of the United Kingdom Atomic Energy Authority. It contains more than 4000 records and is hosted on-line by ESA at Frascati (Rome).

Reference databases

What has been examined so far are some of the most important textual-numeric databases, each one with its particular features, not only as to the topics covered, but also as to their structure. Some of them are not simply factual data banks, because they also contain bibliographic references.

Nevertheless, in order to have a complete overview of the published literature on a chemical, it is necessary to exploit reference databases.

Among them Chemabs/CA ranks the first for what concerns chemistry. Its producer is the above mentioned CAS in Columbus, Ohio, and its on-line coverage dates back to 1967. It contains citations to the worldwide literature in organic, analytical, physical, applied, macromolecular and biochemical chemistry and chemical engineering. CA has a good coverage for the European East countries literature. Several on-line services distribute CA on-line.

A composite file specifically designed for toxicological information is Toxline/Toxlit. It is one of the first products of the NLM-TIP, created with the precise scope of collecting, in a unique database, the toxicological references from numerous different files. Therefore, relevant subsets from Biosis, Medline, EPA, CIS and CA are included.

Table 3. - Fields included in the category TOXB of RTECS file

Acronym	Field
MTSU	Mutagenicity studies [Test system] [Species/route/cell type] [Dose] [Reference]
CTSU	Carcinogenicity studies [Route] [Species] [Study type] [Dose] [Effect] [Reference]
SSTU	Skin and eye irritation studies [Route] [Species] [Dose] [Effect] [Reference]
GSTU	General toxicity studies [Route] [Species] [Study type] [Dose] [Effect] [Reference]
RSTU	Reproductive studies [Route] [Species] [Study Type] [Dose] [Effect] [Reference]
MDSTU	Multiple dose studies [Route] [Species] [Study Type] [Dose] [Effect] [Reference]

With its backfile, it contains approximately 3,7 million citations, dating back to 1965, to the worldwide literature in all areas of toxicology, including chemicals and pharmaceuticals, pesticides, environmental pollutants, mutagens and teratogens. It is hosted by many on-line services.

To complete this overview on bibliographic sources, there are two major biomedical databases that have to be mentioned: Embase and Medline. They are generally considered as complementary to perform exhaustive searches in biomedicine and both of them are in many cases very pertinent for retrieving information regarding chemical, pharmaceutical and drug toxicology.

Embase (Excerpta Medica) is produced by Elsevier Science Publishers in Amsterdam and contains citations from over 3500 journals representing the worldwide literature on human medicine and some areas of biological sciences. In detail, it covers among many other topics, drugs, adverse reactions, occupational health, industrial medicine, pharmacology and toxicology. Searches on Embase are performed through the controlled vocabulary Emtree. It is hosted by a lot of on-line services; some portions of the file are also distributed separately.

Medline is the oldest and most famous bibliographic on-line file on biomedicine. Produced by NLM, it contains references to articles from about 3500 journals published in the US and about 70 other countries. It is hosted by a great deal of on-line services all over the world.

Medline indexers and searchers use the thesaurus MeSH, that represents one of the most important and updated tools in the biomedical information systems. The number of chemical substances considered in Medline is about 30,000.

Conclusions

Most of the on-line files presented in this overwiew are now produced also on a new support, the optical disk, the so-called compact disk-read only memory (CD-ROM). For instance, Medline is produced now in more than ten versions on CD-ROM.

This different approach is having a great success, above all for end-users, because CD-ROMs are searchable in a user-friendly language. To exploit a CD-ROM, the user does not have to connect on-line, but he/she needs a CD-ROM player on his PC. It is not possible to make an *a priori* choice between on-line and CD-ROMs: it depends on the specific situation. Generally speaking, on-line systems are more profitable if the information needed is not confined to specific subjects or disciplines and if the information must be updated and be as complete as possible. On the other hand, CD-ROMs become more valuable if the number of searches on a single file justifies the cost of purchase.

During the last years, the electronic information market has been developing new capabilities above all to meet end-user needs. The most recent innovation is the

creation of on-line full-text databases with the inclusion of graphs and images, which could not be transmitted through the existing telecommunication systems.

Today, with the transmission through optical fibers networks, this problem has been overcome.

Therefore, the electronic tools will soon routinely allow researches and scientists to reach information and get full documents within their laboratories, through worldwide connections.

What was envisaged as utopia and science fiction is now quickly being realized.

Submitted on invitation.
Accepted on 22 June 1994.

REFERENCES

1. *Directory of on-line databases*. 1990. 11 (3). Cuadra/Elsevier, New York.

2. *Directory of on-line databases*. 1992. 13 (2). Cuadra/Gale, Detroit, London.

3. GALANTE, M. 1993. Basi di dati a supporto della compilazione delle schede di sicurezza prodotto. *Informatica e documentazione* 19: 47-57.

4. *Chemical abstracts index guide*. 1992. Appendix II. American Chemical Society, Columbus (OH).

5. WISSMAN, H.M. & WEXLER, P. 1983. Toxicological information. *Annu. Rev. Inf. Sci. Technol.* 18: 185-230.