Comparative Evaluation of the Performance of Online Databases in Answering Toxicology Queries

Lyn Robinson
University College London

Ia McIlwaine
University College London

Philip Copestake
BIBRA International

Christina Anderson
BIBRA International

Abstract

An evaluation of toxicology information resources is reported, comparing commercial online databases and a specialised in-house database. A mixed qualitative/quantitative approach, using ten test queries and detailed failure analysis was used. The main conclusions are: the in-house database is superior to any 'general' database in recall and precision; commercial databases are a useful complement, usually providing unique material; a range of databases should be used for good recall; for the commercial databases, complex search strategies are necessary, using the specific access points of each database; retrieval failures are due primarily to coverage, secondly to poor indexing of specific toxic effect. Detailed discussion is given of indexing policies and search strategies.

Introduction

Toxicology is recognised as a subject area with a particularly rich and diverse set of information resources; see, for example, Robinson (in preparation), Liverman, Ingalls and Fulco (1997), Wexler (1990), Wexler (1988), Kissman and Wexler (1983). This reflects both the ethical and economic importance of provision of accurate and complete information on the toxic, and other harmful effects of chemical substances (and hence the investment in information systems by both governmental agencies and commercial organisations), as well as the development of toxicology as an interdisciplinary science.

Because of this diversity of resources, and the economic and the ethical importance of the subject area, toxicology has been a 'test bed' for a significant number of comparative evaluation studies, comparing information sources and searching techniques. These evaluations have taken widely different forms, and used a variety of methodologies; a detailed review and comparison is given by Robinson (in preparation). In brief, they have included:

- annotated listings, usually highly selective, with independent comment on each resource [e.g. Hargreaves (1980), Dekker (1990), Marra et.al. (1996)]
- qualitative comparisons, providing comparison of sources, and advice on best practice in searching [e.g. Bresnitz, Rest and Miller (1985), Vidgen and Madge (1989), Stafford and Logan (1990), Running (1996)]
- quantitative comparisons, typically of coverage, use, or retrieval parameters [e.g. Montgomery (1973), Cotton and Livesey (1980), Voigt and Bruggermann (1995), Ludl, Schope and Mangelsdorf (1996)
- mixed qualitative/quantitative evaluations - the most elaborate, and usually most useful, form of evaluation [e.g. King (1983), Cassidy and Kostrewski (1986), Bawden and Brock (1982, 1985)]

The study reported here is an example of this last form of evaluation.

One consistent finding of these evaluations has been the continuing importance, for the retrieval of toxicology information, of secondary bibliographic abstracting and indexing services implemented as online databases: examples are CASearch (Chemical Abstracts), Biosis (Biological Abstracts), Medline (Index Medicus) and Embase (Excerpta Medica). These are characterised by a wide coverage of material within their subject area, by detailed indexing, and by provision of several access points for retrieval, including controlled indexing and classification, uncontrolled keyword indexing, and free-text searching of title and abstract. Given that they are general discipline-oriented resources, catering for a wide variety of user interests, they cannot be adapted specifically for the needs of the toxicologist. The ability to access a wide variety of such resources within the common interface of a host system such as those offered by the Dialog Corporation is, however, a particular benefit for toxicology searching, where information from a variety of sources and a variety of perspectives may be needed.

Also among the resources available for toxicology searching are specialised databases, created and operated by research or commercial organisations, and generally referred to as 'in-house' databases. Being toxicology-oriented, these resources can employ specific criteria for coverage, and specific indexing languages and policies, to meet the needs of the toxicologist. They were found to be of value in an earlier study, particularly because of their wide coverage of material which may not be found in other sources (Bawden and Brock 1982). More generally, the ability of in-house databases to provide high levels of recall and precision in retrieval of biomedical material has been demonstrated (Gretz and Thomas 1995, Sodha and van Amelsvoort 1994, Datta 1988).

A comparative evaluation of the strengths and weaknesses of these two kinds of resource will be valuable in two respects, as part of a general study of the communication of toxicology information (Robinson in preparation). First, it may indicate the current value of each resource type, and suggest any changes in the future. Second, it can cast more general light on such factors as coverage, indexing, and retrieval features in any resource.

The general research questions which this study is intended to address are therefore:

- how good are commercial online bibliographic databases at answering toxicology queries ?
- are there any particular problems and weaknesses with them ?
- how do they compare in performance with a specialised in-house database ?
- what is the future for these two kinds of source ?

More specifically, 6 hypotheses are to be tested:

1. That an in-house database, created specifically to deal with toxicology information, would be superior in performance to any commercial database, and usually to a combination of databases, over a range of queries
2. That, despite (1), commercial databases provide a useful complement to an in-house database over a range of queries
3. That a range of commercial databases is always needed to give adequate recall for any toxicology query
4. That searching for toxicology in commercial online databases always requires a relatively complex search strategy to achieve good recall
5. That there are still problems in attaining good recall and precision in commercial databases, associated with the definition of particular toxicology concepts, and with consistency and depth of indexing
6. That the problems noted in (5) may be overcome by the kind of indexing adopted in a specialised in-house toxicology database

This investigation evaluates two things at once. Firstly, the performance of commercial databases, because they are an important resource for toxicology information. Secondly, an example of a specialised, in-house database, because this may perform particularly well as a result of its coverage and indexing. Both of these factors can reflect the specialised purpose of the database, ie. that it is tailored in a way that is not possible in a general database.

If superior performance is offered by specialised databases, then making such files more widely available could improve the cost effective provision of toxicology information. Additionally, lessons could be learned from them, and applied to improve the performance of more general, commercial databases.

The BIBRA TRACE database

This specialised database is maintained by the Information and Advisory Service of BIBRA International, a contract research and consultancy organisation, specialising in toxicology. Its role is to support the consultancy activities of BIBRA in industrial, environmental, food and agricultural toxicology.

The database currently contains approximately 115,000 records, from over 3,000 publications, largely dating from 1987, but with the selective addition of older material. The database is updated daily by input from BIBRA staff, who are themselves experienced toxicological evaluators, largely from the scanning of approximately 100 journals, secondary sources and 'grey literature'. About 200 records per week are added.

Database records comprise a brief bibliographic reference, with a reference to the BIBRA location in which the original document is held. Indexing is by chemical substance – Chemical Abstracts Service Registry Number for specific compounds, and other chemical identifiers (file codes) for particular classes of substance – and a series of codes for toxicological concepts, e.g. acute exposure, study in humans, ecotoxicity. Free text searching of titles is also possible. Emphasis is placed on careful selection of material and on focused indexing. The records are prepared by specialist toxicology staff, who also perform database searches.

Methodology

Methods; general

This study is, in essence, a detailed comparative evaluation of retrieval from a set of databases. The method adopted was the mixed quantitative / qualitative approach, since this is capable of producing the richest, and most meaningful, results (Bawden 1990, Robinson in preparation).

The specific method chosen was that of test queries with detailed failure analysis. This is particularly suitable for this type of investigation, since it yields detailed and reliable information on the overall strengths and weaknesses of the various sources, and on specific points of indexing and searching (Robinson in preparation). In this particular, toxicology, application, it also enables direct comparisons to be made with other studies of toxicology resources carried out nearly 20 years ago (Bawden and Brock 1982,1985) in order to assess changes over time, and with consequent technological advances. A detailed comparison of the methods and results of those studies and the present one is given by Robinson (in preparation).

Methods; queries

A series of 10 test queries was constructed, in order to investigate a broad section of toxicology. Queries were therefore drawn up to be representative of a variety of toxicology facets, and included specific and general chemical substances, and substances without a clear definition in terms of chemical structure. General and specific toxic effects, and factors such as route of administration, species affected, and the environment of the effect, were also included. An over-riding proviso was that the queries had to be sensible and realistic; the kind that could be encountered in practice. Queries were constructed by the author in conjunction with BIBRA Information and Advisory Service staff. Whilst these were not real queries, they were chosen to be typical of the sort of questions encountered in real life. Time period limitations were used when it was necessary to restrict the size of the output to be analysed.

The 10 queries used were:

1. Contact toxicity of lavender oils in humans.
2. Toxicity of sage oils
3. Acute toxicity of angelica
4. Carcinogenicity of chloro-cresols
5. Teratogenicity of formaldehyde, items published in 1995,1996 or 1997
6. Ecotoxicity of ethylene glycol, items published in between 1990 – 1997 inclusive
7. Aquatic toxicity of propoxur
8. Cases of chloracne due to dioxin exposure, data published between 1987 and 1997 inclusive
9. Evidence of increased mortality in humans due to occupational exposure to styrene, items published in 1995, 1996 or 1997
10. Acute effects of inhalation of toluene in humans or other mammalian species, items published in 1995, 1996 or 1997

Methods; details

Searches were carried out using databases on the Dialog system. This host system provided all the databases which were likely to be of use, given that the searches were confined to text-based bibliographic files, i.e. excluding factual/numeric databanks (such as RTECS or DOSE) and files with graphic chemical substructure searching capabilities (such as Chemical Abstracts CAS Online system). Full details of the search process are given by Robinson (in preparation).

Choice of databases
For each query, the set of files to be searched was initially determined from the searcher's prior knowledge and experience, from the Dialog database documentation, and from searches of the Dialindex database, examining all possible files. This set was then refined during a series of initial exploratory searches. In general, the focus was kept to scientific databases: regulatory files, though they would certainly contain relevant information, were excluded from this study, so as to keep the analysis feasible [but see future work section]. The final set of databases chosen was that which would be, in the investigator's judgement, the selection used by an expert searcher seeking high recall. Files were usually only included if they gave either good recall or some unique items; sometimes, however, databases were included in the analysis if this was not the case - even if they gave no useful material at all - if their inclusion exemplified some point of retrieval effectiveness, or lack of it, or if it gave an interesting comparison with other queries.

Initial search
Files were searched separately, or together using OneSearch, as convenient; in some cases, the number of hits was so great that separate searching was essential to avoid system space filling up. In all cases, specific searching features of particular files, e.g. CAS Registry Numbers and Biosis Biosystematic Codes were used where appropriate; there was no 'lowest common denominator' searching. Chemical dictionary and general reference files, such as ChemName, ChemSearch and Merck Index, were used as necessary in identifying chemical substance identifiers and other terminology.

Broad classificatory terms, such as Biosis concept codes (Prasek 1998), were not generally included in the search strategies, as it was considered that these would be likely to lead to too much irrelevance; however, it was found that for some searches, the Biosis concept codes relating to toxicology were essential for maximum retrieval from this database. Other possible useful Biosis codes, e.g. 37013 Public Health/Environmental Health/Occupational health, 37013 Public Health/Environmental Health/Air, Water and Soil Pollution, and 35500 Allergy, though possibly useful for optimal recall, were not included as being outside the practice of a searcher under normal conditions.

CA Search sections headings are automatically included in the basic index, and hence would be retrieved if equivalent text terms were put into the search: thus TOXIC? would retrieve any items in the Toxicology section (CA204XXX), while AIR and POLLUT? would find those in Air Pollution and Industrial Hygiene (CA 259005). Similarly, Excerpta Medica EmClass section headings, such as Toxic Substances and Protective Effects (3726060500), are included in the basic index, and hence accessible by text search.

Full details of the search strategies used for each query are given by Robinson (in preparation).

Search refinement
The initial search strategy was refined in two ways. Firstly by inspection of the indexing terms assigned to records which appeared relevant, to see if any additional or alternative terms could be added to the strategy, and also to reveal terms responsible for the retrieval of large

numbers of irrelevant records. The second approach was a failure analysis, carried out by checking for the presence of relevant items in other files by author name, or by title terms other than those used in the initial searches (so that, e.g., mis-spelling of a term would not affect the failure analysis as well as the initial retrieval). Items found in this way were examined to see if the search strategy could be improved, and the process repeated. It took between one and four iterations, according to the nature of the query, to arrive at an optimal search strategy.

Relevance judgements and final failure analyses

Once the optimal search strategy had been resolved, ie. no new relevant records were identified by the refined search, relevance judgements from the final record set were made by the searcher, usually on the basis of title and abstract, sometimes from the rest of the online record, e.g. index terms, and where necessary and feasible from examination of the original document. Although relevance judgements are notoriously subjective and variable, this process at least ensured consistency. No attempt was made to determine the novelty or quality of information provided (though it was noted that in several cases essentially the same paper, to judge from the abstract, had been published in more than one journal). Nor was extra weight given to items, such as BIBRA Toxicity Profiles, substantial review articles, or extracts from compendia, which might in themselves have provided much extra information.

Toxline was considered as a single database for the purposes of summarising recall, so that multiple hits from individual subfiles were counted only once; for assessing precision, however, each hit from each sub-file was counted once in the 'total hits'. Toxline's precision ratings are therefore, arguably, inconsistently poor.

A final failure analysis was carried out on the set of relevant records, and a matrix constructed to show the presence or absence of each record in each database, and where appropriate, the reason for record retrieval failure.

The measurement of recall has particular problems in evaluations of operational systems. It is, of course, dependent on consistent and sensible determination of relevance; Blair (1996), for example, discusses methods of calculating recall so as to minimise uncertainty in recall estimations because of differing relevance judgements. Beyond this, however, is the fundamental problem that assessment of absolute recall requires calculating the ratio of the number of relevant items retrieved to the number of relevant items present in the database. This latter figure can only be determined by examining each and every item present in the database being searched, and judging its relevance; this heroic expedient was in fact used on the test database used for the Cranfield II retrieval experiments, in which every document in a set of 1,400 items was individually checked for relevance against each of 287 test queries (Cleverdon and Keen 1966). This approach is clearly not practical for any operational database, and alternative means are required; several have been proposed, and are discussed in detail by Robinson (in preparation). They include use of test queries for which a comprehensive bibliography already exists, and some form of statistical sampling, but these are unsuitable for this sort of evaluation. The only remaining alternative is the use of *relative* or *apparent*, rather than *absolute*, recall. By this, the evaluator makes no attempt to use the absolute number of relevant items in a database, or in a sample taken from it, but rather uses a surrogate set of relevant items, against which retrieval is judged. In a comparative evaluation, this is likely to comprise the total set of all relevant items retrieved by use of any database or search technique (Bawden 1990 p 79-80, Tague 1981 p 69). Any item present which was not found by any of the search procedures used, especially if failure analysis is employed, is likely to be so inadequately indexed, as to be effectively irretrievable, and hence can be safely ignored for purposes of comparison of databases, search terms, etc. Tague (1981 p 69) argues that 'relative recall seems appropriate in comparative testing' (as here, where comparisons are being made between databases and search techniques, and it is the relative effectiveness of these, rather than the absolute number of items retrieved, which is of interest). She also suggests that this is 'virtually the only possible approach to recall in testing large operational systems'. Indeed, this approach was used in several earlier evaluations of toxicology resources (), and was therefore adopted for these evaluations.

TRACE

The BIBRA TRACE database was searched according to normal best practice, and relevance judged from the title and where, necessary, inspection of the original document and personal knowledge of BIBRA staff. On the rare occasions on which relevance judgements initially differed, these were resolved by discussion of individual cases.

Comments on methods

This methodology is very time-consuming, with several working days typically taken up in the search and analysis process for each query, and a similar time in completing the overall analysis. Nonetheless, it proved very effective in casting light on the issues under investigation. Although only ten queries were used, several hundred records were examined, and many aspects of indexing were revealed.

Results

For each query, two tables of results were compiled, giving quantitative summaries of retrieval for (a) the set of items retrieved from commercial databases, and (b) the set of items retrieved from commercial databases plus the TRACE database. These include the following values:

a)

- total number of relevant records present, whether retrievable or not, in any of the commercial databases searched (n)[1]
- number of relevant records retrieved by each database (r)
- % recall for each database, calculated as the % of relevant records found from n
- number of unique records per database (u)
- total number of records retrieved from each database (t)
- % precision, calculated as the % of relevant records out of the total number retrieved from each database (r/t)
- % recall for each individual database, calculated as the % of relevant records found out of the total number of relevant records present in the database ($recall_{idb}$)

b)

- total number of relevant records found from all databases searched, including TRACE (N)
- number of relevant records retrieved by each database (r)
- % recall for each database, calculated as the % of relevant records found from N
- number of unique records per database including TRACE (U)
- total number of records retrieved from each database (t)
- % precision, calculated as the % of relevant records out of the total number retrieved from each database (r/t)
- % recall for each individual database, calculated as the % of relevant records found out of the total number of relevant records present in the database ($recall_{idb}$)

Precision and recall are quoted rounded to the nearest percent. 'Unique' items are those judged retrievable only from that database by any feasible search strategy, whether or not they are present in others. For each individual database, the recall of items present in that database is also shown, denoted as $recall_{(idb)}$

Cause of failure to retrieve relevant items is then summarised as due to coverage or to failings of indexing for each component of the search (tables c and d). If the failure is due to more than one reason, e.g. lack of adequate indexing for both chemical substance and toxic effect, each reason for failure is counted separately.

Brief comments on search strategies, and on the results of the failure analysis, are given here; full details, including a detailed account of the development of queries, and their modification in the light of the failure analyses, are given by Robinson (in preparation).

[1] The recall figures in the a) tables, for the commercial databases, are quoted against the total set of items present in any database, whether or not they were retrievable from any; those which were not retrievable were identified by the search of the BIBRA database.

Individual Queries

Query 1 - Contact toxicity of lavender oils in humans

The search was taken to include all contact allergies and dermatitis, and skin sensitivity, whether induced occupationally or otherwise. Definition of relevance was generally straightforward, though in some cases it was necessary to consult the full paper to determine whether data relating to lavender oil were included.

A range of databases, including Medline, CA Search, Biosis, Scisearch, Food Science and Technology Abstracts, Pascal, CAB, and Life Science Collection, were investigated. However, all except Toxline and Embase gave only a small number of non-unique items, and are therefore not included in the analysis, with the exception of Medline, which is included for comparison with Embase.

a) For the 18 items present in the three commercial online databases, the performance figures are as follows:

	Embase	**Medline**	**Toxline**
Retrieved	10	2	4
Recall (%)	56	11	22
Unique	8	0	2
Total hits	11	2	8
Precision (%)	91	100	50
Recall (idb)	63	18	31

A combination of Embase and Toxline is sufficient to find all the relevant material retrievable from these files, i.e. excluding the 6 items which cannot be found in any commercial database because of lack of indexing of substance. The performance of Embase is particularly good, comparing favourably with Medline in this case.

b) When the references present only in the TRACE database are added, the performance figures for the 22 references and 4 sources are as follows:

	Embase	**Medline**	**Toxline**	**TRACE**
Retrieved	10	2	4	19
Recall (%)	45	9	18	86
Unique	3	0	1	10
Total hits	11	2	8	23
Precision (%)	91	100	50	83
Recall (idb)	63	18	31	86

c) Causes of failure to retrieve relevant items over the 3 commercial databases are these:

- item not present 11
- substance indexing 24
- toxic effect indexing 0

Failures of substance indexing predominate, due largely to records describing data on many substances, which are not indexed individually, with only general terms being applied. Coverage is also a significant factor, largely due to references being in 'obscure' sources, and in one case to a Medline reference almost certainly destined for Toxline being too recent to have appeared there. The profile of toxicity terms used results in no failures in any database for this reason.

d) Causes of failure to retrieve relevant items over the 4 sources are these:

- item not present 29
- substance indexing 25
- toxic effect indexing 3

'Items not present' in the online databases when TRACE's contribution is included are a mix of older pre-online material and items from largely non-journal sources.

Query 2 - Toxicity of sage oils

There are, in fact, several species of sage plant (Salvia). For this query, oil from any species was considered relevant. In some bibliographic records, it was not clear which form of sage oil was involved. All forms of toxicity to all organisms were included.

Five databases were found to be useful:

- Embase
- Medline
- Toxline
- Chemical Abstracts
- Biosis

A sixth, Life Science Collection, was included in the search process, but provided no useful information; it included one relevant item, but this could not be retrieved because of lack of substance indexing.

a) Performance figures for the set of 24 references present in commercial databases, though not necessarily retrievable from them, is as follows:

Reference	**Biosis**	**Medl**	**Emb**	**Toxl**	**CAS**
Retrieved	3	2	6	9	12
Recall (%)	13	8	25	38	50
Unique	0	0	3	4	7
Total hits	11	7	20	29	23
Precision (%)	27	33	30	32	52
Recall (idb)	33	25	55	75	86

CASearch and Toxline, and to a lesser extent Embase, appear to be the 'best' commercial databases for this search.

To retrieve all 22 references which can be found from commercial databases requires a combination of CASearch, Toxline and Embase. Pairs of databases, for this search, are quite effective, retrieving as follows:

CASearch plus Toxline	17/22
CASearch plus Embase	16/22
Toxline plus Embase	13/22

b) Results including TRACE references (set of 30 retrievable references)

Reference	**Biosis**	**Medl**	**Emb**	**Toxl**	**CAS**	**TRACE**
Retrieved	3	2	6	9	12	22
Recall (%)	10	7	20	30	35	73
Unique incl	0	0	0	1	2	10
Total hits	11	7	20	29	23	26
Precision (%)	27	33	30	32	52	85
Recall (idb)	33	25	55	75	86	100

c) Causes of failure to retrieve relevant items present in any commercial databases are these:

- item not present 100
- substance indexing 25

- toxic effect indexing 2

Coverage is therefore clearly the predominant factor.

d) Causes of failure to retrieve relevant items present over all sources are these:

- item not present 107
- substance indexing 25
- toxic effect indexing 2

Query 3 - Acute toxicity of angelica

'Acute toxicity' was taken to mean any harm resulting from a single dose or short-term exposure. It was assumed to relate either to accidental / occupational / clinical trial exposure in humans, accidental exposure in mammals (domestic or farm animals) or toxicity testing in mammals.

This definition proved somewhat imprecise in practice, as it was not clear whether certain effects, contact dermatitis for example, which are usually though not invariably associated with acute exposure, should be specifically included. It was decoded to search simply for the 'acute' concept, regardless of type of toxicity.

'Angelica' similarly proved a difficult concept to define with precision, since it does not correspond to a single chemical substance, or group of substances. It is commonly taken to refer toextracts from any parts of any of the umbelliferous plants of the genus *angelica*, though other definitions are more exact; the Merck Index, for example, defining angelica as the fruit or root of the plant species *Angelica archangelica*. There are also specific chemical substances, e.g. angelic acid, angelica lactone and imperatoin, which are derived from angelica plant sources. Finally angelica extracts form components of a considerable variety of oriental, specifically Chinese and Japanese, medicines, as well as some Western medicines, e.g. Friar's Balsalm.

Given the difficulty of making a precise definition, the pragmatic choice was to include any substance noted in the primary document as being derived from any angelica species. Full recall would have required an identification of all substances derived from the genus, and an identification of all medicines from whatever tradition, incorporating them, and inclusion of all these terms in a search strategy.

The final set of databases used was:

- Biosis
- CA Search
- Medline
- Toxline
- Embase

CA Search did not give any useful material with the final search strategy adopted, but has been included for comparison.

A wider variety of databases was initially investigated, including Scisearch and International Pharmaceutical Abstracts, but none added any significant amount of material, nor anything unique.
a) For the 29 references present in any commercial online database, though not necessarily retrievable from any, the performance figures were as follows:

	Biosis	**CAS**	**Medl**	**Embase**	**Toxline**
Retrieved	3	0	2	14	5
Recall (%)	10	0	7	48	17
Unique	1	0	0	12	2
Total hits	3	0	3	27	7
Precision (%)	100	n/a	67	52	71
Recall (idb)	23	0	13	56	29

If, for comparison, the performance figures are calculated against the 19 items which are retrievable from any online database, the recall figures become:

	Biosis	CAS	Medl	Embase	Toxline
Recall (%)	16	0	11	74	26

The good performance of Embase is evident in either case, though it requires three databases (Embase, Toxline and Biosis) for complete recall of the retrievable items. It is noteworthy that CA Search does not provide any useful material for this search; although several items are present, they lack substance or toxicity indexing. This unusually poor performance is due to the search lacking any true chemical substance specification, usually CA Search's strong point.

b) For the total 34 references, the performance figures are as follows:

	Biosis	CAS	Medl	Embase	Toxline	TRACE
Retrieved	3	0	2	14	5	19
Recall (%)	9	0	6	41	15	56
Unique	0	0	0	11	1	15
Total hits	3	0	3	27	7	19
Precision (%)	100	n/a	67	52	71	100
Recall (idb)	23	0	13	56	29	100

The TRACE database appears as slightly superior to the best commercial database, Embase, in terms of recall, and considerably superior for precision. A combination of TRACE and Embase would find 31 out of the 34 items.

c) Reasons for failure among the commercial databases are as follows:

- Item not present 44
- No substance indexing 24
- No toxicity indexing 17
- No acute indexing38

The main problem is lack of coverage. This is in part due the 'grey literature' aspect, but also, for this query, particularly to the presence of material from oriental journals covering the medicinal literature of that area. Embase is particularly strong in covering this type of material, and in this search includes a series of 6 papers from the same researchers and published in the same journal (*Japanese Pharmacology and Therapeutics*), which are not included in other services.

The lack of substance indexing is due entirely to the omission of specific substances, rather than to problems with the 'angelica' terminology.

Toxic effect specification was a particular problem in this query, partly due to its complete absence, e.g. in items reviewing all aspects of classes of medicine or cosmetic, and partly due to the difficulties in specifying acute exposure noted earlier.

d) Reasons for failure over all the databases are as follows:

- Item not present 59
- No substance indexing 24
- No toxicity indexing 17
- No acute indexing38

Query 4 - Carcinogenicity of chloro-cresols

Carcinogenicity was defined as being the occurrence, or otherwise, of tumours in humans or any experimental species, or as other forms of data, such as mutagenicity or epidemiology, where specific conclusions on carcinogenicity were drawn by the authors.

Chloro-cresols were defined as compounds containing a single cresol ring, substituted with between 1 and 4 chlorine atoms; this gives a finite number of compounds, and is therefore not a true substructure search, which would be open-ended. This meant that the search could feasibly be carried out in textual bibliographic databases, the structure search being achieved by use of the words chloro and cresol in various combinations, together with Registry Numbers found from chemical dictionary files. Databases allowing chemical substructure searching would have allowed a more direct approach, but would not necessarily have given greater recall.

The initial Dialindex search suggested a set of 5 databases as being potentially useful, and these all gave useful material. They were:

- CA Search
- Toxline
- Medline
- Biosis
- Embase

The Chemname and Chemsearch nomenclature databases were also used to find chemical substance search terms. These were searched both by synonym and by structural features - molecular formula, ring size and composition, etc. - to ensure good coverage. Of the relevant compounds identified, 44 were mentioned at least once in Chemical Abstracts during the relevant time period. In order to select those likely to be of sufficient commercial significance to have been tested for toxicity - and to simulate 'realistic' searcher behaviour - the 15 compounds occurring more than 10 times in Chemical Abstracts were chosen. Had the full set of substances been used, including those not mentioned in the secondary source, the search would have been equivalent to a 'true' substructure search.

a) For the15 references present in any commercial online database, the performance figures were as follows:

	CAS	**Toxline**	**Medl**	**Biosis**	**Embase**
Retrieved	1	11	2	2	3
Recall (%)	7	73	13	13	20
Unique	1	9	0	0	1
Total hits	4	15	2	2	9
Precision (%)	25	73	100	100	33
Recall (idb)	33	92	33	29	43

The particularly good performance of Toxline for this query is notable, though CA Search and Biosis also give unique items.

b) For the total 24 references, the performance figures are as follows:

	CAS	Toxline	Medl	Biosis	Embase	TRACE
Retrieved	1	11	2	2	3	14
Recall (%)	4	46	8	8	12	58
Unique	1	6	0	0	0	10
Total hits	4	15	2	2	9	22
Precision (%)	25	73	100	100	33	64
Recall (idb)	33	92	33	29	43	88

The TRACE database gives best overall recall, whereas Toxline gives greater precision. Overall a combination of the in-house and commercial sources is necessary to give acceptable recall. A combination of Toxline and TRACE finds all except 1 relevant item.

c) The reasons for failure to retrieve, for the 15 references in commercial databases, are:

- Item not present 40
- No substance indexing 16

In this query, specification of toxic effect did not cause any loss of retrieval, in commercial databases.

d) The reasons for failure to retrieve over all the databases are these:

- Item not present 47
- No substance indexing 17
- No toxic effect indexing 2

Query 5 - Teratogenicity of formaldehyde, papers published in 1995, 1996 or 1997.

The time period restriction was made to control the number of potentially relevant references. Although teratogenicity refers specifically to malformations in an organism resulting from exposure to a toxin, the broader concept of reproductive toxicity, and the concept of embryotoxicity, were included in order to avoid losses due to incorrect indexing.

References specifying mixtures or resins were not considered relevant.

Choice of databases
Databases found to be useful were:

- Biosis
- Life Science Collection
- Medline
- Embase
- Toxline
- CA Search
- Occupational Safety and Health
- SciSearch

Of others tried, NTIS, SEDbase and Pollution Abstracts produced no relevant material, Enviroline contained one relevant item but it was no retrievable due to lack of substance indexing, and CAB had one relevant, and retrievable, item, and one irretrievable because of lack of indexing of teratogenic effect.

a) Performance figures for the set of 23 items retrieved from any of the commercial databases were as follows:

	Biosis	**LifeSci**	**Medl**	**Emb**	**Toxl**	**CAS**	**OSH**	**SciS**
Retrieved	4	2	5	3	11	3	2	3
Recall (%)	17	9	22	13	48	13	9	13
Unique	0	0	0	0	4	1	1	0
Total hits	48	5	10	10	61	6	5	12
Precision (%)	8	40	50	30	18	50	40	25
Recall (idb)	25	33	38	30	52	25	50	30

The overall superiority of the Toxline file, particularly in finding unique references, is evident. The best pair of databases is Toxline plus OSH or CAS, finding 12/13 relevant items.

b) Performance figures including the references present only in the TRACE database are as follows:

	Biosis	**LifeSci**	**Medl**	**Emb**	**Toxl**	**CAS**	**OSH**	**SciS**	**TRACE**
Retrieved	4	2	5	3	11	3	2	3	27
Recall (%)	11	5	14	8	30	8	5	8	73
Unique	0	0	0	0	3	1	1	0	24
Total hits	48	5	10	10	61	6	5	12	27
Precision (%)	8	40	50	30	18	50	40	25	100
Recall (idb)	25	33	38	30	52	25	50	30	100

c) Causes of failure to retrieve the 23 items retrievable from any commercial databases are these:

- item not present 91
- substance indexing 41
- toxic effect indexing 38

Coverage is therefore clearly the predominant factor. Failures in substance indexing generally occur with articles reviewing lists of compounds, which are not exhaustively indexed. Failures due to toxicity indexing become increasingly significant as the more 'difficult' material, not found in any database, are included in the analysis.

d) Causes of failure to retrieve items over all the databases are these:

- item not present 101
- substance indexing 41
- toxic effect indexing 38

Query 6 - Ecotoxicity of ethylene glycol, papers published in 1990 - 1997

The time period restriction was chosen to control the number of potentially relevant documents. Ecotoxicity was taken to mean any effect on the general environment, and species within it; excluding toxicity testing and epidemiological studies in humans. Relevance was sometimes difficult to judge because of this rather diffuse concept.

A wide variety of databases appeared at first to provide information relevant to this query. Ultimately 5 were included in the final analyses:

- Biosis
- CASearch
- IAC Prompt
- Embase
- Toxline

Other databases which, though they provided some information, were not included in the final analyses because they provided relatively little additional material, and nothing unique, were CAB, Pollution Abstracts, NTIS and ABI/Inform and (from a different perspective) Magazine Index.

a) For the 23 items present in any of the commercial online databases, of which 4 were considered not retrievable by any sensible search logic, the performance figures are as follows:

	Biosis	**IAC**	**CAS**	**Toxl**	**Embase**
Retrieved	2	1	5	15	4
Recall (%)	9	4	22	65	17
Unique	0	1	1	8	1
Total hits	11	5	8	41	6
Precision (%)	18	20	63	37	67
Recall (idb)	13	100	38	71	33

Toxline gives clearly the best performance of any single database, but there is a considerable spread in retrieval, with four databases needed for complete retrieval. A single unique (and apparently useful) reference is given by IAC Prompt, a business database which might not have been an obvious choice for a search of this kind. Precision is relatively high for CASearch and Embase. (It is worth noting that before the addition of the extra terms from the failure analysis both these databases gave 100% precision; loss of precision was countered by the addition of only one, non-unique, reference in each database.)

b) For the 42 references retrieved from all sources, including the TRACE database, the performance figures are as follows:

	Biosis	**IAC**	**CAS**	**Toxl**	**Embase**	**TRACE**
Retrieved	2	1	5	15	4	29
Recall (%)	5	2	12	36	10	69
Unique	0	1	1	3	1	23
Total hits	11	5	8	41	6	30
Precision (%)	18	20	63	37	67	97
Recall (idb)	13	100	38	71	33	100

c) Causes of failure to retrieve relevant items over all commercial databases are these:

• item not present	46
• substance indexing	18
• toxic effect indexing	4
• environmental indexing	14

While coverage is the predominant factor (particularly in view of the inclusion of the IAC database, which would not be expected to cover scientific material to any extent), substance indexing poses a problem in this case. This is largely due to documents dealing with large numbers of substances, without individual indexing. The difficulties with definition of the 'environmental' aspect have already been noted.

d) Causes of failure to retrieve relevant items over all the databases are these:

• item not present	59
• substance indexing	18
• toxic effect indexing	4
• environmental indexing	14

Query 7 - Aquatic toxicity of propoxur

'Aquatic toxicity' was intended to encompass toxic effects of any kind on aquatic organisms, including fish, and to include both observations in the field and laboratory testing.

7 databases were used in the final analyses:

- CAB
- CA Search
- Water Resources Abstracts
- Pollution Abstracts
- Life Science Collection
- Biosis
- Toxline

a) Performance figures for the 67 relevant items present in the commercial databases, though not necessarily retrievable from any, are as follows:

	CAB	**CAS**	**Water**	**Pollut**	**LifeSci**	**Biosis**	**Toxline**
Retrieved	10	37	3	3	4	9	23
Recall (%)	15	55	4	4	6	13	34
Unique	2	24	1	0	1	2	8
Total hits	30	46	6	8	12	13	58
Precision (%)	33	80	50	38	33	69	40
Recall (idb)	48	73	23	27	44	32	66

The particularly good performance, in terms both of recall and of unique items found, from CA Search is partly due to its (largely exclusive) coverage of a series of papers by a single author (Nishiuchi). Retrieval is otherwise quite diffuse, with the combination of 6 databases required to give complete recall.

b) Performance figures for the set of 75 items, including those retrieved only from TRACE, are as follows:

	CAB	**CAS**	**Water**	**Pollut**	**LifeSci**	**Biosis**	**Toxline**	**TRACE**
Retrieved	10	37	3	3	4	9	23	23
Recall (%)	13	49	4	4	5	12	31	31
Unique	2	20	1	0	1	2	8	16
Total hits	30	46	6	8	12	13	58	36
Precision (%)	33	80	50	38	33	69	40	64
Recall (idb)	48	73	23	27	44	32	66	100

In this case, both recall and precision from the TRACE database are bettered by those from commercial databases. The low recall from TRACE in this instance is a reflection of the dates of coverage, as items are included only from 1987 onwards. This search retrieves a significant body of material from earlier dates. The wide coverage of the commercial databases is seen to be useful in this query.

c) Causes of failure to retrieve relevant items over all commercial databases are these:

- item not present 292
- substance indexing 52
- toxic effect indexing 11
- fish/aquatic indexing 25

Coverage is by far the dominant factor, other retrieval failures being distributed among the three indexing concepts, though lack of substance indexing is more significant than the other two factors.

Failures due to substance indexing are those with documents dealing with tests of large numbers of substances which are not indexed individually, being covered by general terms such as ‘pesticides’. Problems with toxic effect indexing were due to the use of very general terms such as ‘effect’, ‘health’, or ‘quality’. These two problems could be only be overcome by the use of a search strategy of such generality as to be impractical. Difficulties in retrieving the aquatic concept were largely due to the use of very specific terms, such as ‘crayfish’ or ‘tadpoles’, in databases which provided no hierarchical organism indexing; this problem could only have been overcome by the heroic expedient of including the species name of every aquatic species likely to be mentioned.

d) Causes of failure to retrieve relevant items over all the databases are these:

- item not present 344
- substance indexing 52
- toxic effect indexing 11
- fish/aquatic indexing 25

Query 8 - Cases of chloracne due to exposure to dioxins

The query implied exposure of humans, occupational or accidental, and required discussion, comparison or counting of individual cases. General discussions of mechanism, safety precautions etc., and simple statements of chloracne as a result of dioxin exposure, were not regarded as relevant.

A wide selection of databases was initially tested, as Dialindex searches had indicated that they should produce useful material. Life Science Collection, CAB Health, Scisearch, Enviroline and Pollution Abstracts all gave some useful material, but in small quantity, and with nothing unique.

The final set of databases used was
- Biosis
- Embase
- CA Search
- Toxline
- Medline

a) For the 36 references present in any commercial online database, though not necessarily retrievable from any, the performance figures were as follows:

	Biosis	**Emb**	**CAS**	**Medl**	**Toxl**
Retrieved	17	23	1	21	24
Recall (%)	47	64	3	58	67
Unique	1	7	1	0	1
Total hits	24	35	1	32	62
Precision (%)	71	66	100	66	39
Recall (idb)	61	77	8	78	83

b) For the total 38 references, the performance figures are as follows:

	Biosis	**Emb**	**CAS**	**Medl**	**Toxl**	**TRACE**
Retrieved	17	23	1	21	24	10
Recall (%)	45	61	3	55	63	26
Unique	1	7	1	0	1	4
Total hits	24	35	1	32	62	11
Precision (%)	71	66	100	66	39	91
Recall (idb)	61	77	8	78	83	34

c) Reasons for failure among the commercial databases are as follows:

- Coverage 64
- Substance indexing 1
- Toxicity indexing 4
- Case indexing 29
- Chloracne indexing 14

CA Search, which provides a single unique item with 100% precision, appears to be very different in retrieval performance from the others. However, causes of failure, expressed as percentages, are very similar, with and without CA Search:

	with CAS	without CAS
Coverage	57	53
Substance indexing	1	1
Toxicity indexing	4	6
Case indexing	26	28
Chloracne indexing	13	11

Coverage is the predominant factor. Problems with ‘case’ indexing and chloracne indexing follow in order. General toxicity and substance indexing pose little problem.

d) Causes for failure to retrieve relevant items over all databases are as follows:

- Coverage 73
- Substance indexing 2
- Toxicity indexing 4
- Case indexing 33
- Chloracne indexing 28

Query 9 - Evidence of mortality in humans due to occupational exposure to styrene, documents published in 1995,1996 or 1997.

An initial exploratory query, encompassing all harmful effects resulting from occupational exposure to styrene, generated too large a number of references, of varied nature, for careful analysis. The query was therefore refined to include the specific concept of evidence of human mortality. Any mention of mortality in populations occupationally exposed to styrene, as well as hard data on mortality rates, was judged relevant.

A wide variety of databases were included in initial searches, as likely to be useful. Finally searching was concentrated on the six shown below. Other databases which produced some useful material, though no unique references, were CA Search, Enviroline, Pollution Abstracts, Environmental Bibliography, Cancerline, Embase and Occupational Safety and Health. The main six resources were:

- Biosis
- Life Science Collection
- Toxline
- Medline
- Healthstar
- Pascal

a) For the 9 references present in the online databases, the performance figures are as follows:

	Biosis	LSC	Medl	HStar	Toxl	Pascal
Retrieved	5	6	4	4	7	7
Recall (%)	56	67	44	44	78	78
Unique	0	0	0	0	0	0
Total hits	7	6	7	7	20	7
Precision (%)	71	100	57	57	35	100
Recall (idb)	56	75	44	50	78	87

Precision figures are high. Toxline precision appears low, as noted earlier, since the same item is retrieved from several sub-files, but only counted as retrieved once. All online files give relatively high recall.

If only the 9 items retrieved from any online database are considered, both Toxline and Pascal provide 100% recall, and Pascal combines this with 100% precision.

b) The TRACE search gave 61 references, of which 11 were judged relevant, compared with 7 relevant items from the commercial databases; 6 of the 7 are in the TRACE set.
Including the TRACE results, the performance figures for the set of 12 items are:

	Biosis	LSC	Medl	HStar	Toxl	Pascal	TRACE
Retrieved	5	6	4	4	7	7	11
Recall (%)	42	50	33	33	58	58	92
Unique	0	0	0	0	0	0	5
Total hits	7	6	7	7	20	7	61
Precision (%)	71	100	57	57	35	100	18
Recall (idb)	56	75	44	50	78	87	92

c) Causes of failure to retrieve relevant items from commercial databases, are these:

• item not present	20
• substance indexing	11
• occupational exposure indexing	7

Unusually, failures in indexing exceed coverage as the reason for lack of retrieval; of course, the findings would be different if the several other databases which produced a very small amount material had been included in the analysis.

Lack of substance indexing and terms for the 'occupational exposure' concept are both causes of failure to retrieve relevant items; the 'mortality' concept presented no problem.

d) Causes of failure to retrieve relevant items over all databases are these:

• item not present	3
• substance indexing	12
• occupational exposure indexing	7

Query 10 - Acute effects of inhalation of toluene, items published in 1995, 1996 or 1997.

It was assumed that this query implied toxicity in humans (accidental exposure or clinical study) or in mammalian species (accidental or experimental exposure).

The difficulties of specifying 'acute toxicity' have been noted above (angelica query). In this case, since the inhalation route was specified, effects such as dermatitis were not under consideration.

Relevance judgements were mostly straightforward, though in some cases there was scope for debate as to what constituted 'acute' toxicity.

Dialindex searches, and initial exploratory searches, suggested that 6 databases might prove useful, and these were all carried through into the final searches.

The databases used were:

- Biosis
- Toxline
- Medline
- Life Science Collection
- Embase
- CA Search

a) For the 28 references present in any commercial online database, though not necessarily retrievable from any, the performance figures were as follows:

	Biosis	**Toxl**	**Medl**	**LSC**	**Emb**	**CAS**
Retrieved	4	6	5	2	6	1
Recall (%)	14	21	18	7	21	4
Unique	0	0	0	1	2	0
Total hits	11	25	13	3	17	1
Precision (%)	36	24	38	67	35	100
Recall (idb)	20	24	23	22	29	5

No database performs well in recall terms, Biosis, Medline, Toxline and Embase being roughly on a par.

Precision is generally poor, though CASearch achieves 100%. Embase and LSc provide unique references.

b) For the total 38 references, the performance figures are as follows:

	Biosis	**Toxl**	**Medl**	**LSC**	**Emb**	**CAS**	**TRACE**
Retrieved	4	6	5	2	6	1	30
Recall (%)	11	16	13	5	16	3	79
Unique	0	0	0	0	2	0	28
Total hits	11	25	13	3	17	1	32
Precision (%)	36	24	38	67	35	100	94
Recall (idb)	20	24	23	22	29	5	100

The TRACE database's 79% recall is far in advance of any commercial database's performance.

c) Reasons for failure among the commercial databases are as follows:

- not present in database 110
- no substance indexing 15
- no toxicity indexing 12
- no inhalation indexing 71
- no 'acute' indexing 72

Coverage is the major reason for failure in retrieval, but the problems with specification of 'inhalation' and 'acute' concepts are clear, and much outweigh those associated with the substance and toxicity components.

d) Reasons for failure over all databases are as follows:

- not present in database 118
- no substance indexing 15
- no toxicity indexing 12
- no inhalation indexing 71
- no 'acute' indexing 72

Summarised results

The results of the individual queries, described in detail above, are now summarised in several ways.

1) First, the databases found to be valuable in at least one of the queries are ranked according to the extent of their usefulness. This gives an indication of the databases likely to be most useful for this kind of search, and for the range of databases which should be considered.

2) Second, for each of the set of six databases used in more than one query, three measures of retrieval performance are summarised across the set of queries.

- Recall is summarised against the items present in each individual database, against those present in the set of commercial databases used, and against the total set including the TRACE database. The first of these measures indicates the ease with which material present in each database can be retrieved; it is essentially a measure of indexing quality. The second and third values measure the contribution of each database to retrieval of the total set of documents, and are therefore measures both of indexing quality and of extent of coverage.

- Precision values are then quoted, as a summary of this measure of each database's capability for focussed searching across the test queries.

All the values above are quoted as percentages, and maximum, minimum and mean values are given. Inspection of the data confirmed that the familiar mean value was as good, or better, a measure of 'average tendency' as the alternative median.

- Uniqueness is then summarised as a measure of the ability of each database to provide items which cannot readily be found from any other source. This is expressed in two ways.

First, simple counts of unique items provided by each database are given. Second, they are expressed as percentages of the total number of unique items for each query: this avoids the summarisation being dominated by the results of queries which happened to produce a large number of unique items. It is arguable that provision of a single unique item in a search in which only two unique items are found overall (counting as 1 item, or 50%) is more significant than five unique items in a search providing twenty such (counting as 5 items, or 25%). The two ways of representing uniqueness allow these factors to be compared.

3) Finally, summaries of failure of retrieval are produced across all commercial databases, across all databases including the TRACE database, and for each individual database. These measures give detailed insight into why relevant documents are not returned when a 'best' search strategy is used.

Summary of valuable commercial databases

'Valuable databases' are those found to provide either a significant amount of relevant material – generally 10% recall, as a minimum – or some unique references.

Because of the way in which identification of databases to be used was carried out – searcher's judgement, based on cross-index searches, examination of documentation, prior knowledge, and initial test searches – it is possible than some potentially useful databases were not included. Indeed, the provision of some useful material by other databases is noted for several queries. However, the process used should identify all the most useful databases, and give a reliable picture of the general pattern of database usefulness.

The databases found valuable in this set of test queries are these;

Database	**No of queries**
Toxline	10
Biosis	9
Embase	8
Medline	8
CA Search	7
Life Science Collection	4
Occupational Safety and Health	1
Scisearch	1
IAC	1
CAB Abstracts	1
Water Resources Abstracts	1
Pollution Abstracts	1
HealthStar	1
Pascal	1

This shows a 'core' of Toxline with the four major disciplinary services, useful for a majority of these queries. Life Science Collection, which includes Toxicology Abstracts, occupies an intermediate position. A tail of 8 databases each contribute to one query. It is likely that a slightly different set of test queries would add more databases to this 'tail'.

This form of Zipfian distribution is to be expected, as being common to most bibliometric situations (Zipf 1935, Brookes 1968). Its clear emergence here gives confidence that the results of this study are reliable[2].

[2] Zipf originated his law as part of a general study of language in the 1930s. In 1934, Bradford and Lancaster Jones at the Science Museum Library in London drew attention to a bibliometric distribution which was later recognised as a specific example of Zipf's distribution: hence the term 'Bradford-Zipf' distribution is often used.

Summary of recall, precision and uniqueness, across all databases

	Toxline	Biosis	Embas	Medl	CAS	LSC	TRACE
Number of queries	10	9	8	8	7	4	10
Recall (individual database)							
Maximum	92	61	77	78	86	75	100
Minimum	24	13	29	13	5	22	86
Mean	60	32	48	27	38	44	97
Recall (commercial databases)							
Maximum	78	56	64	58	55	67	
Minimum	17	9	13	7	3	6	
Mean	46	21	33	23	22	22	
Recall (all databases, including TRACE)							
Maximum	63	45	61	55	49	50	92
Minimum	15	5	8	6	3	5	26
Mean	34	17	27	18	16	16	64
Precision							
Maximum	73[1]	100	91	100	100	100	100
Minimum	18[1]	8	30	33	25	33	18
Mean	50[1]	56	51	64	67	60	80
Uniqueness without TRACE (counts)							
Maximum	9	2	12	0	24	1	
Minimum	0	0	0	0	0	0	
Mean	4	1	4	0	5	1	
Uniqueness without TRACE (percentages)							
Maximum	82	10	80	0	63	33	
Minimum	0	0	0	0	0	0	
Mean	32	2	42	0	23	9	
Uniqueness with TRACE (counts)							
Maximum	8	2	11	0	20	1	24
Minimum	0	0	0	0	0	0	4
Mean	2	0	3	0	4	0	15
Uniqueness with TRACE (percentages)							
Maximum	35	7	50	0	40	2	100
Minimum	0	0	0	0	0	0	29
Mean	10	1	15	0	11	0	68

1 Note the treatment of Toxline precision described above.

Recall
For the commercial databases, recall falls into a fairly clear pattern, with Toxline giving superior performance, Embase second, and the others roughly equal. No single database gives total recall, which can only be achieved by a combination of databases, as discussed in results for each individual query.

When the items present only in the TRACE database are also considered, the recall values for all databases fall, as TRACE adds a significant number of unique records. Toxline and Embase both give the best performance from the commercial databases, but TRACE database achieves a higher recall than both of these. Even so, TRACE's 64% average recall is considerably less than 100%, emphasising the need to use commercial resources as a complement to the in-house, specialist database. The TRACE database recalls more of the records present in the commercial databases, than vice versa. The limited recall demonstrated by the TRACE database is perhaps a reflection of its specific coverage and time period (ie. 1987 onwards).

It is noteworthy that, for three of the databases (Biosis, Medline and CA Search) only about one third of the relevant items present in the database could be retrieved using the 'best practice' search strategies proposed here. The TRACE database achieves the highest average for recall of relevant records actually present in any one database. This could be anticipated as a result of the highly structured, expert and limited indexing system employed.

When recall of items present in at least one commercial system is measured over all queries, no single database gives half of the possible recall; Toxline, with 46%, performs best.

One clear lesson from these results is that, with the exception of one query for which Toxline alone or Embase alone was sufficient for complete recall, it was necessary to use combinations of databases to get good recall. As noted in the descriptions of the individual test queries, combinations of 3 or 4 databases were commonly required. When pairs of databases were used, the combination of Toxline with either CA Search or Embase, depending on the query, generally proved most valuable.

Precision
Precision shows a different pattern, with the best commercial performers here (CAS, Medline and Life Science Collection) being among the poorer for recall. Embase and Toxline, which are the only databases to give recall of above 20% of the total set including TRACE's unique references, give below 60% precision. This illustration of the inverse relation of recall and precision confirms other studies (Cleverdon, 1972); though it should be noted that the differences in precision are not very great, and would be lessened if the duplicate items from Toxline sub-files were discounted. The high average precision for the TRACE database is notable, but so is the range of values achieved, which show a minimum precision of 18%. This indicates that the database may not perform well for queries which contain concepts which cannot be mapped exactly onto the indexing codes.

Uniqueness
Medline produces no unique records, although this could be expected as all Medline toxicology records are also contained in Toxline. The TRACE database provides the highest percentage of unique records, probably because of its specific coverage of grey literature such as government and international organisational material. Within the commercial databases, both the uniqueness measures show essentially two groups of database: one set (Toxline, Embase, and CA Search) which contribute a significant extent of uniqueness, and the remainder which do not. Nonetheless, these latter do, in some instances, provide unique records, and so must be searched if complete recall is required, as must databases located further from the 'core' set identified above. The percentage uniqueness measure, which attempts to standardise the effect of uniqueness regardless of the size of the set of relevant

material, suggests that the order of success in the provision of uniqueness is Embase > Toxline > CASearch.

Ranking of databases
It is tempting to use these results to provide a ranking of the success of the commercial databases in answering these test queries, though the limited extent of the evaluation should dictate that this be done with caution as to the general applicability of results.

With that caveat, the rankings which emerge are:

Local recall (success of the database in providing material within its particular scope)
Toxline > Embase, Life Science Collection > others

General recall (success of the database in retrieving a wide range of material)
Toxline > Embase > others

Uniqueness (success of the database in providing material which other sources cannot)
Embase > Toxline > CA Search > others

Precision (success of the database in providing focussed access to relevant material)
CA Search > Medline > Life Sciences > others

Summaries of failure of retrieval

Across commercial databases

The reasons for failures in retrieval were consolidated and summarised for those references present in any commercial online database. This gave an indication of the relative importance of the general difficulties of toxicology searching in any online database.

Failures were categorised under four headings:

• Item not present	(coverage)
• No substance identification	(substance indexing)
• No general toxicity indication	(subject indexing)
• No specific type of toxicity	(subject indexing)

The last of these consolidates the specification of specific types of toxic effect (e.g. acute toxicity), route of administration (e.g. inhalation) and scope of toxic effect (e.g. ecotoxicity).

Over the 10 queries, and expressed as percentages, the relative importance of the four factors were:

• Coverage	60
• Specific toxicity concept	20
• Substance	15
• General toxicity	5

The dominant effect of coverage, as noted for several of the individual queries, is clear. This is particularly notable, since this analysis refers to items present in at least one of the databases searched.

Across all databases

Reasons for failure in retrieval across all databases, including TRACE, are summarised below, demonstrating little change from the previous table:

• Coverage	61
• Specific toxicity concept	20
• Substance	14
• General toxicity	5

Summary of failure of retrieval in each database
This set of figures shows the relative importance of different types of retrieval failure in each of the six databases used for more than one of the test queries and the TRACE database. This gives an indication of the relative weaknesses of each of the databases.

The four categories of failure factors used in the analysis above – coverage, substance indexing, general toxicity indexing, and specific toxic effect indexing – were also used here.

The failures of each type for each of the six main commercial and the TRACE databases are initially summed over all ten queries.

	Emb	Toxl	Medl	CAS	Biosis	LSC	TRACE
Coverage	80	81	82	106	121	106	121
Substance	29	38	35	23	58	8	4
Toxicity (general)	9	17	13	11	14	6	5
Toxicity (specific)	40	56	44	69	54	10	18
Total	158	192	174	209	247	130	148

They are then divided by the total number of failures for that database, and expressed as percentages.

	Emb	Toxl	Medl	CAS	Biosis	LSC	TRACE
Coverage	51	42	47	51	49	81	82
Substance	18	20	20	11	23	6	3
Toxicity (general)	6	9	7	5	6	5	3
Toxicity (specific)	25	29	25	33	22	8	12

Perhaps the most surprising finding here is how similar the five major databases are in this respect; Life Sciences Collection seeming rather different. The TRACE database fails for the same reasons that the Life Sciences Collection does. This can be anticipated, as they are both smaller, more specialised databases. It should be noted however, that the above table indicates reasons for failure, and not overall database performance. Overall performance can be seen in the previous table (page 64).

Coverage failures are the major factor for all databases; for the five main sources, these lie in a small range between 42% (Toxline) to 51% (Embase and CA Search), while the coverage factor for Life Science Collection is an overwhelming 81%, reflecting the smaller size of this database.

Substance indexing again lies within a small range. CA Search, not surprisingly performs best here with only 11% of failures due to this factor, while the four big biomedical sources lie between 18% and 23%.

Similarly, specific toxicity indexing failures for these four sources lie in a small range between 22% and 29%; CA Search performs relatively poorly on these factors at 33%, reflecting its relative lack of detailed indexing of biomedical concepts.

For both substance and specific toxicity, Life Science Collection performs relatively well, reflecting the preponderant role of coverage in this database's performance.

General toxicity specification causes less than 10% of failures for all databases.

These summary figures support the general picture revealed in the previous failure analysis, indicating that coverage is the predominant difficulty, rather than any particular limitation in indexing. Choice of any particular database, among the five major ones used here, does not seem critical. Of course, this is based on the results of carefully designed and refined searches, taking advantage of the most appropriate searching facilities of each database. It does not necessarily reflect what might happen if simpler searches, using only 'obvious' terms, were carried out.

Discussion of results

These results demonstrate that the TRACE in-house database performs better on average than any single commercial database used in the test queries. Average percentage recall and precision values for the TRACE database are 64% and 80% respectively, compared to the 34% achieved by Toxline as second best performance for recall, and 67% second best precision percentage achieved by CAS.

The TRACE database provides a high percentage of unique records to the toxicology queries used in this analysis; an average of 68%. For each of the test queries used, no combination of the commercial databases used was able to retrieve all of the items located by the TRACE database. It therefore seems safe to conclude that the TRACE database will usually also perform better than a combination of commercial databases.

The recall and precision figures are averages taken over all ten test queries however, and it is important to note that the TRACE database does not always give the best performance for an individual query, indicating the need to supplement searches of the in-house database with additional searches of commercial databases. Recall from the TRACE database ranged between 26% and 92%, emphasising that successful recall is dependant on the toxicology query matching TRACE's specific coverage. Recall did not reach 100% for any query. This is a further indication of the need to supplement in-house searches with results from commercial databases if high recall is required. Unique references were found in commercial databases for nine of the ten test queries.

In recalling items present in any individual database, TRACE again demonstrated a superior performance in comparison with the commercial databases, 97% compared to the next best of 60%, achieved by Toxline. This may be expected as a result of the limited indexing terms and codes applied to the TRACE database. If an article has been entered and indexed, then the application of a limited number of codes, eg. substance plus toxic effect, should recall the record. In contrast, the commercial systems demonstrated at best an average of only 60% recall of items that were actually present, indicating a deficiency in the record indexing. In these cases, even complex search strategies were unable to retrieve the records.

For average precision, TRACE performs better than the commercial databases, but achieves a minimum of 18% as well as its maximum of 100%, illustrating that the limited indexing can have limitations with regard to some specific toxicology concepts.

All of the queries could be answered to a greater or lesser extent by commercial databases. These databases are useful in answering toxicology queries, and as described above, form an essential complement to the use of the specialist in-house database used at BIBRA. There are many commercial databases available, and although this study was restricted to those offered by the Dialog Corporation, it is clear that those providing at least one useful item are many and varied. The databases chosen for each of the queries were selected in a manner of 'best practice', ie. those that would most likely be chosen in real life by an expert searcher. Although it is therefore possible that other databases would also contribute relevant and possible unique material, it can be seen from the analysis of the test queries that a core of databases can be identified as those likely to be most useful for toxicology queries. These include Toxline, Biosis, Embase, Medline, CASearch and Life Science Collection. These databases can then be ranked in terms of average recall, precision and number of unique items contributed. Here it can be seen that Toxline and Embase offer the best average recall (46%, 33% respectively) and number of unique items (32%,42% respectively) whilst the highest precision is obtained by CASearch (67%). The relatively low recall achieved even by Toxline emphasises again the need to search a range of commercial databases in order to increase recall. Precision is inversely proportional to recall, linking those databases with high recall, eg. Toxline, with lower precision. Again, this may illustrate a difficulty in the indexing of specific toxicity concepts.

While these rankings may be a good indication of general usefulness, it must be stressed that they are constructed from a limited number of test queries and that in any real life query different results may be obtained.

Further insight into the performance of the databases is given by the failure analysis, which examines reasons for retrieval failure in detail. Essentially, failure to retrieve items relevant to a query is due to a combination of factors which may be categorised as coverage and indexing. When reasons for failure are summarised across the commercial databases, and across all databases including the TRACE database, it is clear that coverage is the main factor, as 60% and 61% of retrieval failures respectively are due to the item not being present in the database. Retrieval failure summaries for individual databases also confirm coverage as the highest reason for failure.

Failure due to substance indexing is apparent in all of the commercial databases, although it is low in CASearch, 11%, as may be expected, and in the smaller Life Science Collection (6%). Failure due to lack of substance indexing is even less apparent (3%) in the TRACE database, emphasising its in-depth, if limited, indexing policy of substance and toxic effect. Failure due to general toxicity indexing is less than 10% in all databases. More problematic in every database is the failure to retrieve due to lack of specific toxicity indexing. This could be expected; as the definition of a toxicology concept increases in complexity, consistent and in-depth terminology is harder to apply. Over all databases, including TRACE, 20% of retrieval failures are due to specific toxicity concept indexing.

It is clear therefore, that retrieval of toxicology information could be improved by enhancing the coverage of any one database. However it is also unlikely that any single database is ever going to include records of all items relevant to such a multidisciplinary subject as toxicology. The most comprehensive approach taken so far in this respect can be seen with the Toxline database, which gathers together several relevant databases, and provides an option to search through all of them at once.

Despite the comparatively good recall demonstrated by the TRACE database, it is evident that even a specialised toxicology source is not comprehensive. If the aim is to retrieve the maximum number of relevant items from the literature, then a variety of databases must be searched. However, if the query falls within the coverage of a more specialised one, then this can be a sensible, cost effective, single option.

The Toxline database has addressed coverage by providing a single interface to a variety of toxicology databases, datasets and subfiles; a plausible extrapolation of this principle could be to provide a single, possibly graphical/web based interface to databases containing material relevant to toxicology, thereby resolving, to a greater or lesser extent, the issue of coverage.

Issues of indexing are somewhat harder to resolve. There are advantages and disadvantages to both the limited indexing approach taken by specialist in-house databases, such as TRACE, and the multi-level access point approach taken by the large, general, commercial databases.

The former system supports a high level of precision, but only when the concept required can be mapped directly onto the codes used in the indexing. This type of database seems ideal when the searcher has a clearly defined category of query, which does not require the recall of every item from the literature.
While the limited, highly structured indexing exemplified by the TRACE database does favour a high precision overall, it is clear that this is not applicable to all queries; specifically, for

query 7, where the concept of 'aquatic' could only be addressed by the environmental aspects code, and query 9, where the concept of 'mortality' could not be expressed.

The latter approach clearly varies between databases, which is itself a problem for the searcher. Substance indexing varies in depth, and in the terminology used eg. Registry Numbers, different synonyms and preferred terms. A variety of general toxicity descriptors and headings exists, augmented by a larger number of specific toxicity concept descriptors, codes and subject headings, all of which can be used in addition to free text searching of title and abstract. With the plethora of access points, retrieval from the larger commercial databases can be complex and lack precision. Incomplete indexing can mean that even complex strategies do not always retrieve relevant records which are present in the database. However, it must be remembered that with the exception of Toxline, the commercial databases used are not specifically designed to answer toxicology queries.

The matching of the analysis of document content by an indexer, resulting in a set of index terms, and the analogous analysis of an information need by a searcher, resulting in a set of query terms, is a major issue for effective information retrieval (see, for example, Lancaster 1998, chapter 2). While indexers will generally strive to assign terms describing what the document is 'about', their opinion may not coincide with 'aboutness' as perceived by a searcher (Swift et. al. 1978). Or, as Olafsen and Vokac (1983) put it:

'The indexer has to make guesses at what questions the future user of the system will put. Regardless of how cleverly the guesswork is construed, they are still guesses, while the user approaches the system with his own concrete question, and his associations may be different from those of the indexer'.

One possible way to improve the retrieval of toxicology information from bibliographic databases could be to apply a standard structure to the indexing of items, thereby offering a standard framework from which to construct searches to retrieve items.

The present results can be considered with reference to the three-layer model proposed by Robinson (in preparation). The model breaks down a toxicology statement, such as a title or a query, into primary, secondary and tertiary layers, so that the primary layer describes the substance(s), the secondary layer the general toxic effect (toxicity, side effect, poisoning, carcinogenicity, teratagenicity etc.), and the tertiary layer the way in which the substance is encountered eg. as an acute dose, as an ecotoxic effect etc. It is possible to add specificity to the general toxic effects of the secondary layer, by inserting a sub-layer to contain specific concepts such as resulting effect in man or animals.

Failure to retrieve relevant records which are, in fact, present in a database is seen to be due to incomplete indexing of either substance (1°), general toxic effects (2°), or most importantly, specific (3°) toxic effects. Few items failed to be retrieved due to incomplete indexing of general toxic effects, even though the terminology used to represent these concepts varies between databases, eg. Biosis concept codes. There were failures due to incomplete substance indexing, however, and more precise and relevant substance indexing, such as that employed by the TRACE database, could clearly improve recall from the commercial databases. Often, in studies where a variety of compounds is tested, only the generic name or substance category is indexed, making relevant information impossible to find. Most failures occurred due to incomplete indexing of tertiary concepts however. Terms such as 'acute', 'mortality' or 'aquatic' toxicity are revealed to be frequently not indexed, or described in a variety of ways. The inconsistent and varied approach to 3° concept indexing does have advantages, in that free-text terms can be applied to enrich item description, but at the same time, this requires complex search strategies to ensure item retrieval. In many cases, even complex strategies fail to retrieve all relevant items.

It is possible that no complete solution exists to the problem of indexing toxicology literature, but it may be advantageous to consider each item with reference to the three layer model, so that each aspect is indexed in some way. This would alleviate the problems encountered due to incomplete indexing of substance, general or specific toxic effect. Problems of consistency could be addressed by applying a particular indexing scheme, or thesaurus at any one layer. For example, MeSH headings could be used at the 2° layer. This approach would offer the searcher some guidelines as to how the item had been described, and subsequently improve retrieval. Overall, both searching and indexing could be facilitated.

Such an approach could be combined with the concept of adding together separate databases to improve recall. Many toxicology queries do not fall neatly into a single specialist domain, but encompass several. A unified approach to indexing could offer a consistent interface from which to search any toxicology dataset. This could perhaps be instantiated as a standard, web-based interface to any toxicology data.

The results obtained can be compared with those obtained by Bawden and Brock (1982,1985). This study confirms their findings that:

- Commercial online bibliographic databases are an essential part of any toxicity searching
- It is essential to use several complementary databases for good recall
- 'Peripheral' databases can give useful and unique items
- Complex search strategies are usually necessary, including portfolios of terms
- It is necessary to use the particular access points provided by each database
- An in-house file has particular value in covering material not present in commercial databases
- There are problems in specifying exact toxic effects, routes of administration etc.
- There are difficulties associated with searching for chemical substructures and classes of substance in these databases

Aspects which have changed include:

- The lessening of problems with the specification of a chemical substance – possibly due to changed indexing policies and the greater use of Chemical Abstracts Service Registry Numbers
- Clearer specification of general toxicity, causing few retrieval failures
- The most useful commercial databases found by Bawden and Brock were CASearch and Toxline, with the biomedical databases providing a useful complement to searching. In this study, Toxline and Embase appear to be the most useful pair of such databases for answering toxicology queries. This may reflect changes in both coverage and indexing policies.

Conclusions

With reference to the six original hypotheses, the following conclusions can be drawn from this investigation:

1) The TRACE specialist in-house database demonstrates better average recall and precision than any one of the commercial databases used to answer the test queries. For individual queries, however, TRACE's recall and precision figures were not always superior to each of the commercial databases. In no case was 100% recall achieved by the TRACE database. In all ten of the test queries, the TRACE database contributed unique records. On average 68% of unique material was provided by the TRACE database. This indicates that for most toxicology queries, its performance will not be matched even by combinations of databases.

2) Commercial databases are a useful complement to the TRACE in-house database in providing additional material in answer to the toxicology queries used here. All of the databases contributed some unique material, with the exception of Medline, the toxicology section of which is included in Toxline.

3) For eight of the ten queries, a range of commercial databases was needed to give adequate recall. For the test queries used here, Toxline offered the best average recall of the commercial databases, but even this was only 46%, indicating a need to use additional databases to improve search results. When retrieval failure was analysed for items present in at least one of the commercial databases, 60% of failures were due to coverage, again emphasising the need to use more than one database.

4) In contrast to the relatively straightforward search statements needed to query the TRACE database, search strategies for the commercial databases were in most cases complex and time consuming to create. Lack of standardised indexing terms in the commercial databases requires a number of synonyms, subject headings and alternative terms to be entered in order to ensure good retrieval.

5) The main reason (60%) for limited recall in commercial databases is coverage; i.e. items are not present in the databases. Failure to retrieve due to missing or incomplete indexing occurs in 40% of cases, however, half of which are due to lack of specific toxicity concept indexing. This means that 1 in 5 failures to retrieve relevant information is due to lack of specific toxicity concept indexing. Retrieval failure due to depth of indexing occurs in instances where, for example, a large number of compounds is tested, but not indexed individually, and where indexing terms utilised do not adequately describe the concept being sought e.g. Acute toxicity, ecotoxicity, case studies, occupational exposure and inhalation.

In some cases, exemplified by query 5, teratogenicity of formaldehyde, precision was reduced as a result of imprecise indexing, so that more general terms had to be included to avoid loss of recall, but at the expense of precision.

6) Only 18% of retrieval failures from the TRACE database were due to indexing and so it would seem that the adoption of more precise and structured indexing could help reduce retrieval failures in commercial databases.

As may be anticipated, on average, a specialised toxicology database performs better in answering toxicology queries than any single or combination of the more general, commercial databases. No database currently offers 100% recall and precision however, and when examined in detail, the reasons for this can be identified as coverage and indexing failures. It is therefore recommended that toxicology searches are carried out over as wide a range of databases as possible, to ensure high recall, although for the commercial databases, this may be at the loss of precision.

Problems of coverage could be overcome by adding together databases likely to contain references to toxicology literature. Problems encountered by the varied and often incomplete approach to indexing toxicology items could be addressed by applying a standard model to every toxicology statement, so that each aspect of a statement is indexed in a standard way, to the same level, and using standard terminology. This level of indexing is to some extent employed by specialist databases, but naturally involves a high level of effort in its construction. Additionally, as yet no standard vocabulary exists for describing toxicology statements.

The ease with which web-based technology allows access to distributed information suggests that a single, graphical interface could be designed which would allow access to any networked toxicology dataset, once standard indexing had been applied.

A more general parallel concept is that of the Z39.50 protocol, which is designed to standardise search and retrieval from any database, most specifically, bibliographic databases.

References

D Bawden (1990), User-oriented evaluation of information systems and services, Gower, Aldershot

D Bawden and A Brock (1982) Chemical toxicology searching: a collaborative evaluation, comparing information resources and searching techniques, Journal of Information Science, 5, 3-18

D Bawden and A Brock (1985) Chemical Toxicology Searching: a comparative study of online databases, Journal of Chemical Information and Computer Sciences, 25, 31- 35

DC Blair (1996), STAIRS Redux: thoughts on the STAIRS evaluation, ten years after, Journal of the American Society for Information Science, 47(1), 4-22

EA Bresnitz, KM Rest and N Miller (1985), Clinical industrial toxicology. An approach to information retrieval, Annals of internal Medicine, 103(6), 967-972 [plus erratum 104(2), 290]

BC Brookes (1968), The derivation and application of the Bradford-Zipf distribution, Journal of Documentation, 24(4), 247-259

SI Cassidy and BJ Kostrewski (1986), An evaluation of information resources in household product poisoning, Journal of Information Science, 12, 143-157

CW Cleverdon (1972) On the inverse relationship of recall and precision, Journal of Documentation, 28 (3), 195 – 201

CW Cleverdon and M Keen (1966), Factors determining the performance of indexing systems, College of Aeronautics, Cranfield

L Cotton and JB Livesey (1980), Toxicity data on TOXLINE and RTECS, Proceedings of the 4th International Online Information Meeting, Oxford, Learned Information

VK Datta (1988), Coverage of literature on mycotoxins by computer databases and comparison of this coverage with the TDRI in-house index facility, Quarterly Bulletin of the International Association of Agricultural Librarians and Documentalists, 33(2), 61-78

WJ Dekker (1990), Toxicology information series. IV. Information resources for chemical emergency response, Fundamental and Applied Toxicology, 15(2), 213-220

M Gretz and M Thomas (1995), The in-house database - nicety or necessity, Drug information Journal, 29, 163-169

JM Hargreaves (1980), Literature sources in toxicology: a survey, BLRDD Report 5542, London, British Library

J King (1983), Searching international databases: a comparative evaluation, Library and Information Research Report No 3, London, British Library

HM Kissman and P Wexler (1983), Toxicological information, Annual Review of Information Science and Technology, 18, 185-230

FW Lancaster (1998), Indexing and abstracting in theory and practice (2nd edn.), Library Association Publishing, London

CT Liverman, CE Ingalls and CE Fulco (1997), Toxicology and environmental health information resources for health professionals, Bulletin of the Medical Libraries Association, 85(4), 423-425

H Ludl, K Schope and I Mangelsdorf (1996), Searching for information on toxicological data of chemical substances in selected bibliographic databases, Chemosphere, 32(5), 867-880

CA Marra et. al. (1996), Drug and poison information resources on the Internet. Part 2: identification and evaluation, Pharmacotherapy, 16(5), 806-818

RR Montgomery (1973), An indexing coverage study of toxicological literature, Journal of Chemical Documentation, 13(1), 41-44

T Olafsen and L Vokac (1983), Authors' reply, Journal of the American Society for Information Science, 34, 294

MA Prasek (1998), BIOSIS Previews: the power of codes, Medical Reference Services Quarterly, 17 (2), 43 – 49

L Robinson, Toxicology knowledge and information; the impact of new information and communication technologies. PhD thesis in preparation, University College London

WE Running (1996), A comparison of MICROMEDEX CCIS and CCOHS CCINFOdisc, Medical Reference Services Quarterly, 15(2), 33-51

RJ Sodha and T van Amelsvoort (1994), Multi-database searches in biomedicine, Journal of information Science, 20(2), 139-141

DT Stafford and BK Logan (1990), Toxicological information series. V. Information resources useful in forensic toxicology, Fundamental and Applied Toxicology, 15, 411-419

DE Swift et al. (1978), 'Aboutness' as a strategy for retrieval in the social sciences, Aslib Proceedings, 30, 182-187

JM Tague (1981), The pragmatics of information retrieval experimentation, in Information retrieval experiment, K Sparck Jones (ed.), Butterworth, London, 59-102

GA Vigden and B Madge (1989), Computerized information services in toxicology and their use in a poisons centre, Health Libraries Review, 6, 150-167

K Voigt and R Bruggermann (1995), Toxicology databases in the metadatabank of online databases, Toxicology, 100, 225-240

P Wexler (1990), The framework of toxicology information, Toxicology, 60, 67-98

P Wexler (1988), Information resources in toxicology (2nd edn.), Elsevier, New York

GK Zipf (1935), The psycho-biology of language, Houghton-Mifflin, Boston, MA.
(The relevant chapter on the form and behaviour of words is reprinted in The origins of information science, AJ Meadows (ed.), Taylor Graham, London, 1987. Pp 124-141.)