Anderson, C., Copestake, P. & Robinson, L. (2000). A specialist toxicity database (TRACE) is more effective than its larger, commercially available counterparts. Toxicology, 151, pp. 37-43. doi: 10.1016/S0300-483X(00)00264-X

A specialist toxicity database (TRACE) is more effective than its larger, commercially available counterparts

Anderson C.[1], Copestake P.[1], Robinson L.[2]

1. TNO BIBRA International Ltd, Woodmansterne Rd, Carshalton, Surrey, SM5 4DS
2. School of Library, Archive and Information Studies, University College London, Gower St, London WC1E 6BT

ABSTRACT

A comparison of the retrieval precision and recall of a specialist bibliographic toxicity database (TRACE) with that of a wide range of commercial on-line databases indicated that the larger size and resources of the commercial databases did not, for a series of test queries, assure superior retrieval of relevant papers. The specialist database, in which document selection and indexing is undertaken by the same (in this case) expert toxicologists who subsequently use the database in their day-to-day work, achieved markedly better retrieval, using simpler search strategies, than the commercial databases. Small specialist databases may offer a valuable alternative to existing commercial databases.

INTRODUCTION

Two criteria are fundamental to the performance of a bibliographic database - *recall*; the proportion of the relevant documents available which are identified, and *precision*; the proportion of documents identified which prove relevant. In the experience of the Information and Advisory Service at TNO BIBRA International Ltd, which provides expert advice on the toxicological hazards of chemicals and risks resulting from exposure, the commercially available on-line toxicological databases perform poorly when judged by these criteria.

Published studies that have sought to evaluate the ability of selected databases to identify toxicological or medical data generally report rather low values (often of 50% or less) for recall or precision (Gehanno *et al.* 1998; Kleijnen & Knipschild, 1992; Ludl *et al*. 1996; McKibbon *et al.* 1990; Spoor *et al*. 1996). It has been suggested that the performance of a database is related to its focus on a particular field and to the involvement of specialists in the indexing of each paper to identify the chemicals and details of importance in a study (Haramburu *et al*. 1991). A database that fulfills the criteria of specialist involvement and focus (in this case on hazard and risk evaluation) is TNO BIBRA's TRACE (Toxicity Research Association Chemical Effects) database, on the health effects of chemicals.

To compare the relative performances of the most appropriate commercially available databases and a small specialist database (TRACE), in identifying data pertinent to the hazard or risk of chemicals to man and the environment, a representative range of search queries was used to interrogate each.

DATABASES EVALUATED

The specialist database

The Information and Advisory Service at TNO BIBRA International Ltd (formerly BIBRA International) has, for many years, provided information and independent advice, evaluation and comment on the toxicity/ecotoxicity and beneficial effects of chemicals (including foods, natural dietary components, agrochemicals, cosmetics, toiletries, packaging materials and industrial chemicals) to a wide range of industrial sectors and government departments internationally. The TNO BIBRA databank, the product of years of expert selection of relevant papers and reports from key journals, secondary sources and 'grey' literature, is accessed in-house by the TRACE electronic database. Papers entered on TRACE are indexed for study chemical(s) and assigned a few of the two-character keys used to describe the study by the toxicology evaluators who rely on the database for all aspects of their work. TRACE contains around 120,000 records.

The commercial databases

Between three and eight commercially available databases were interrogated for each search query. These were selected from: BIOSIS, CAB, CA Search (CAS), Embase, Healthstar (HlthStar), IAC Prompt (IAC), Life Science Collection (LifeSci), Medline, Occupational Safety and Health (OSH), Pascal, Pollution Abstracts (Pollut), SciSearch, Toxline and Water Resources Abstracts (Water). Indexing of the databases, although rigorous, can not, in general, be said to have been undertaken by those who are also regular database users. Typically such databases contain between 1-12 million citations.

DATA COLLECTION

Ten search queries were compiled to include both specific and general chemical substances, and specific and general aspects of toxicity; details are given in Table 1. The test queries were felt to represent a broad range of queries pertinent to chemical hazard and to be typical of those received by the TNO BIBRA Information and Advisory Service.

The commercial databases most appropriate for each search query were selected by one of us $(LR)^2$, who was also responsible for the design of the search strategies for these databases. All searches were carried out to completion, using extensive failure analysis, further details of which may be found in Robinson *et al.* (2000).

Table 1: Search queries.

1. Contact toxicity of lavender oil in humans.
2. Toxicity of sage oils.
3. Acute toxicity of angelica
4. Carcinogenicity of chlorocresols.
5. Reproductive toxicity of formaldehyde, papers published in 1995, 1996 or 1997.
6. Ecotoxicity of ethylene glycol, papers published 1990-1997.
7. Aquatic toxicity of propoxur.
8. Cases of chloracne due to exposure to dioxins.
9. Evidence of mortality in humans due to occupational exposure to styrene, documents published in 1995, 1996 or 1997.
10. Acute inhalation of toluene, items published in 1995, 1996 or 1997.

Judgments on the relevance of each paper identified were, for consistency, made on the basis of information presented in the bibliographic record alone. The final judgment, the unanimous opinion of all the investigators, was simply based on the probability that a paper

would contribute information useful to a particular query. (Judgments as to the number of documents which are "available" and "relevant" are, of course, highly subjective, but it is necessary to ascribe working values to these parameters for a comparison of performance to be possible. Further details on methodology are available in Robinson (in preparation) and Robinson *et al*.)

Although it was not possible to evaluate the true number of papers, relevant to each query, which had been published, a working value (relevant records; RR) was derived using the total number of unique relevant records identified within TRACE and the commercial databases searched. For each query, recall (R; the proportion of RR identified in any one database) and precision (P; the proportion of records retrieved in a database which were judged to be relevant) were evaluated for each database.

Precision and recall would, in general, be expected to have an inverse relationship to one another; as incremental improvements in recall become progressively harder to achieve, so will precision be reduced. Recognising that both parameters are key to the effective performance of a database, a further composite value, an index of performance (IP), was determined to encapsulate both features in the simplest manner possible. A database that identifies one relevant record for a search where RR is 100 clearly has a precision of 100%, but a recall of only 1%. The inadequacy of such a database would be reflected by an IP value of 1% (where IP=RxP).

In an attempt to compare the "ease of searching" of each database, the number of search terms used for search queries was identified. Although the structure of search terms differs between databases, the use of Boolean algebra to link terms means that it is possible to define 'individual search terms', for the purpose of the present study only, by their adjacency to the Boolean operators *and* and *or*, or to the start or end of a search phrase. (Full details of search strategies are presented in Robinson (in preparation)).

RESULTS

The recall and precision achieved for each search in a particular database are presented in Table 2 (the full data on which this synopsis is based are presented in Robinson *et al*.), and a summary of the performance of the specialist and commercially available databases is shown in Table 3.

For the ten search queries examined, recall ranged from 31% to 92% for the specialist database, TRACE, and from 0% to 63% for the commercial databases; the corresponding means were 64% and 20%, respectively. Of the databases considered in the 'first league' by the evaluators at TNO BIBRA (Biosis, Medline, Embase and Toxline), the recall of the best performer, Toxline, was around half that achieved by the specialist database.

Table 2: Results of comparative searches in a specialist and commercial databases.

SEARCH QUERY	Lavender oil	Sage oil	Angelica	Chloro-cresols	Formalde-hyde	Ethylene glycol	Propoxur	Dioxins	Styrene	Toluene	MEAN

RR		22	30	34	24	37	42	75	38	12	38	
TRACE	R%	86	73	56	58	73	69	31	26	92	79	64
	P%	83	85	100	64	100	97	64	91	18	94	80
Biosis	R%		10	9	8	11	5	12	45	42	11	17
	P%		27	100	100	8	18	69	71	71	36	56
CAB	R%							13				13
	P%							33				33
CAS	R%		35	0	4	8	12	49	3		3	14
	P%		52	n/a	25	50	63	80	100		100	59
Embase	R%	45	20	41	12	8	10		61		16	27
	P%	91	30	52	33	30	67		66		35	51
HlthStar	R%									33		33
	P%									57		57
IAC	R%						2					2
	P%						20					20
LifeSci	R%					5		5		50	5	16
	P%					40		33		100	67	60
Medline	R%	9	7	6	8	14			55	33	13	18
	P%	100	33	67	100	50			66	57	38	64
OSH	R%					5						5
	P%					40						40
Pascal	R%									58		58
	P%									100		100
Pollut	R%							4				4
	P%							38				38
SciSearch	R%					8						8
	P%					25						25
Toxline	R%	18	30	15	46	30	36	31	63	58	16	34
	P%	50	32	71	73	18	37	40	39	35	24	42
Water	R%							4				4
	P%							50				50

Values are rounded to the nearest whole number.

Table 3: Summary of database performance.

DATABASES	R%: RANGE (MEAN)	P%: RANGE (MEAN)	IP: RANGE (MEAN)	NO. TERMS: RANGE (MEAN)
TRACE	31-92 (64)	18-100 (80)	17-74 (50)	2-6 (4)
Biosis	5-45 (17)	8-100 (56)	0.1-32 (11)	6-26 (15)
CAS	0-49 (14)	0-100 (59)	0-39 (10)	6-26 (15)
Embase	8-61 (27)	30-91 (51)	2-41 (16)	6-25 (14)
LifeSci	5-50 (16)	33-100 (60)	2-50 (14)	6-26 (15)
Medline	6-55 (18)	33-100 (64)	5-36 (11)	6-25 (15)
Toxline	15-63 (34)	18-73 (42)	4-34 (14)	6-26 (15)
CAB; Healthstar; IAC; OSH; Pascal; Pollut; SciSearch; Water	2-58 (16)	20-100 (45)	0.4-58 (11)	6-26 (15)
All Commercial databases searched (mean of means)	0-63 (20)	0-100 (54)	0-58 (13)	6-26 (15)

For the commercial databases searched, the range of values for precision varied from 0% to 100%, with the mean roughly at the mid-point in this spread. The specialist database showed a narrower range of 64% to 100%, with one outlier at 18%; mean precision (including the outlier) was 80%.

A comparison of database performance, taking into account both recall and precision, indicates that for the queries investigated the best performances of the most popular databases had IP values of 32 to 41, compared to 74 for the specialist database. The mean index of performance for all the commercial available databases tested was 13 compared to 50 for the specialist database, TRACE.

Although a crude measure of "ease of searching", the number of search terms used in the case of the specialist database was around one quarter of those required to search the commercial systems.

DISCUSSION

Comprehensive identification and selection of good quality toxicity data is necessary for the reliable compilation of reviews, dossiers and risk assessments. Ideally these data would be available from an accredited (electronic) databank; in practice the suitability and quality of the widely-available databanks is disappointing. The Registry of Toxic Effects of Chemical Substances (RTECS), which contains data on a large number of chemicals, is unfortunately restricted by its data selection policy. Since data included in RTECS are limited to those indicating a positive finding, a "weight of evidence" approach to evaluation is not readily supported by this databank. Another well-resourced US product, the Hazardous Substances Databank (HSDB), favours qualitative statements on hazard over the quantitative insights on dose-response essential to any risk evaluation. EUCLID, the EU databank of accredited toxicity and ecotoxicity data on existing chemicals, presently covers only a small number of chemicals.

In the absence of a reliable (electronic) databank, a comprehensive assessment of toxicological information pertinent to a chemical's possible effects on health depends, in the first instance, on the identification of relevant papers. Only when this identification has been

adequately achieved is the expert evaluation of the data in each paper possible. Until very recently bibliographic databases provided the only effective route to wide interrogation of the published literature. It is possible (although it presently seems unlikely in the short-term) that services available on the Internet may provide alternative or better routes.

Although the data and analysis presented here are limited, they broadly support the findings of other recently published evaluations which have tended to show that commercially available databases may not be particularly effective at identifying specific data. One study, which evaluated the ability of 18 selected databases to identify toxicological data on three test chemicals, found the average precision (for all databases together) ranged from 12% to 33%, for the three searches. For the separate searches on each compound in the databases BIOSIS, Medline, Toxline or Toxlit, precision ranged from 8% to 51% (Ludl *et al*. 1996). A recall study, using a capture-recapture technique (best known to ecologists sampling animal populations) in which a known set of papers is "released" prior to "recapture", only 28% of papers on a specific medical topic was identified in a Medline search, compared to 93% identified using an unspecified "hand search" (Spoor *et al*. 1996). Recall values of 11-63% were reported in a study in which five databases were searched using two toxicology search queries; the method used to derive recall was essentially the same as in the present study (Gehanno *et al.* 1998). Poor precision or recall have been attributed variously to inadequate, incorrect or inconsistent indexing (Dickersin *et al*. 1994; Haramburu *et al*. 1991; Spoor *et al*. 1996; Wilczynski *et al*. 1995).

The implication of poor precision, particularly for a toxicologically well-studied chemical such as benzene or carbon tetrachloride, may be an unmanageable number of postings, only a small proportion of which, on close examination, will prove relevant. Poor precision therefore incurs (an often considerable amount of) unnecessary time spent reading titles or abstracts to identify papers of interest. The implication of incomplete identification of relevant documents, resulting from poor recall, can of course be potentially more serious.

A typical search in TRACE involves only the relevant Chemical Abstracts Service (CAS) Registry number(s) and a few two-character keys. Most, but not all, of the commercial databases are comprehensively indexed for CAS numbers. While a cautious searcher of the commercial systems would include relevant synonyms in a comprehensive search, difficulties with this strategy, associated with inadequate name indexing or database search restrictions, are by no means uncommon. The power and discrimination of the TRACE keys is matched in the commercial databases only by the use (in most cases) of a much larger number of free text or standard terms, or codes. Appropriate terms or codes are neither standardised nor often easy to identify; in practice considerable expertise is required for an effective search using multiple commercial databases. The keys used in TRACE are intended to be transparent, in that the characters are easily associated with purpose of the key (OR, for example, describes any study in which the route of administration was oral).

The superior performance of the specialist database TRACE may possibly be explained by the involvement of experts both in the selection and indexing of papers. Wide but highly selective scanning of the relevant literature, by those committed to and dependent upon the effectiveness of the specialist database, contributes to its impressive recall. Accurate identification of study chemicals and inclusion of appropriate study keys, by those experienced in the use of study data, ensures a high level of precision. For example, solvent, adjuvant and control chemicals are not indexed but the important, yet sometimes nebulous, details of a study are effectively captured.

Although it is not perhaps surprising that a specialist database (of modest proportions) performs better in terms of precision than its much larger commercial counterparts, the superior recall demonstrated would not have been so confidently predicted. There appears to be increasing evidence, however, that the impressively large commercial databases, once considered key to any comprehensive review of (in this case toxicological) data, may simply be too large, unwieldy and unresponsive to meet the demands of modern information specialists. According to a paper given at the International Health and Safety at Work Conference 1996, "the number of articles retrieved by a search can be unfeasibly large, and many of them are of little value. The latest trend is the appearance of more selective CD-ROM databases which aim at quality rather than quantity of information provided" (Lee, 1996). Unlike TRACE, a specialist database, for which trained toxicologists select only the literature pertinent to their work, the commercial databases try to serve the information needs of a wide spectrum of users. It seems that the varied requirements of different user sectors may be accommodated at the expense of database performance for any individual user. Where user objectives mirror those of the database generator, a small specialist database such as TRACE can offer the best retrieval service.

REFERENCES

Dickersin K. *et al*. (1994). Identifying relevant studies for systematic reviews. Br. Med. J. **309,** 1286.

Gehanno J.-F. *et al.* (1998). Assessment of bibliographic databases performance in information retrieval for occupational and environmental toxicology. Occup. envir. Med. **55**, 652.

Haramburu F. *et al*. (1991). Comment: Manual or computer-based literature searches for adverse drug reactions. DICP, DICP Ann. Pharmacother. **25**, 875.

Kleijnen J. & Knipschild P. (1992). Pharm. Weekbl. sci. Ed. **14**, 316.

Lee R. (1996). Medical and healthcare information sources. The International Safety and Health at Work Conference '96. Conference Papers.

Ludl H. *et al*. (1996). Searching for information on toxicological data of chemical substances in selected bibliographic databases - selection of essential databases for toxicological researches. Chemosphere **32**, 867.

McKibbon K.A. *et al*. (1990). How good are clinical Medline searches? A comparative study of clinical end-user and librarian searches. Comput. biomed. Res. **23**, 583.

Robinson L. Ph.D thesis [in preparation]. University College, London.

Robinson L. *et al.* (2000). Int. J. Inf. Man. [In press]

Spoor P. *et al*. (1996). Use of capture-recapture technique to evaluate the completeness of systematic literature searches. Br. Med. J. **313**, 342.

Wilczynski N.L. *et al*. (1995). Preliminary assessment of the effect of more informative (structured) abstracts on citation retrieval from MEDLINE. MEDINFO '95 Proceedings **8**, 1457.