# A Framework for the Evaluation of Chemical Structure Databases

Frank Cooke[†] and Helen Schofield*[,‡]

GlaxoSmithKline, 709 Swedeland Road, P.O. Box 1539, King of Prussia, Pennsylvania 19406-0939, and
Chemistry Department, UMIST, P.O. Box 88, Sackville Street, Manchester M60 1QD, U.K.

Received January 12, 2001

Access to desk-top structure and reaction databases through applications such as Chemical Abstracts' SciFinder, MDL's Beilstein CrossFire, and ISIS Reaction Browser has led to changes in information seeking habits of research chemists, the impact of which has implications when database purchasing decisions are made. A semiquantitative assessment is proposed which takes into account key aspects of structure and reaction databases. Assessment criteria are identified which can be weighted according to an organization's information needs. Values are then assigned to criteria for each data source, after which a formula is applied which leads to an indication of the relative value of systems under consideration. The formula takes into account the cost of database products and also the incremental benefit of adding a new system to an existing collection. This work is presented as a generic approach to the evaluation of databases and is not limited in scope to only structure and reactions databases.

## INTRODUCTION

The pharmaceutical industry is a knowledge-driven industry. The chemical knowledge available to this industry can be represented as follows:

$$\text{Info}_{chem} = \Sigma\text{Info}_{chem(proprietary)} + \Sigma\text{Info}_{chem(free)} + \Sigma\text{Info}_{chem(commercial)} + \Sigma\text{Info}_{chem(tacit)}$$

Often the tacit information, $\Sigma\text{Info}_{chem(tacit)}$, i.e., the information in people's heads within the organization, can be the most fruitful but the hardest to tap efficiently. This area of organizational knowledge management will be the subject of another paper. $\Sigma\text{Info}_{chem(free)}$ and $\Sigma\text{Info}_{chem(commercial)}$ are the information sources which contribute to an organization's external learning. Presently, $\Sigma\text{Info}_{chem(free)}$ is largely information available through the Internet. Although the quantity and value of chemical information available through this medium is increasing, it still represents a very small proportion of the total $\text{Info}_{chem}$ in the above formula.[1] This paper therefore focuses on commercially available sources of chemical information, which at present are the most comprehensive sources of external information and knowledge.

Since the early 1980s the shift from information searching by intermediaries to "end-users" has been discussed,[2−4] in terms of the threat to the work patterns of librarians and information scientists, the opportunities for end-users, and also concerns about the efficiency of searches carried out by the newly empowered end-users. This has led to the term "disintermediation", which "relates to the role of the intermediary in acting between information and its end-users. It is the finding of information by an end-user without the need for a third party".[5]

In the late 1970s and early 1980s attempts were made to encourage researchers to undertake their own online searches, with mixed success.[6] In the late 1980s and 1990s a change of emphasis in the availability of commercially produced external bibliographic and patent information occurred as an increasing number of databases became available on CD-ROM. A revolution for chemists occurred when powerful client-server desk-top applications were developed. The first such systems were developed by MDL, who released the Fine Chemicals Directory (a predecessor to the Available Chemicals Directory) in 1983. This was followed by the ORGSYN and Theilheimer reaction databases released as part of the REACCS service in 1985. These products evolved from the pioneering MACCS software used by organizations for managing in-house chemical information since 1979. These were the forerunners of the ISIS system widely deployed today. The full impact of this technology was felt when Beilstein CrossFire[7,8] and SciFinder[9] became available in the mid-1990s. Desk-top systems enable researchers to have access to information directly with no metered charging for online time, searches, or displays.

The end-user searching facilitated by such applications does not remove the need for intermediaries. However, in the period May 1998−May 2000 at SmithKline Beecham there has been an approximate 30% decrease in the number of searches requested through the information analysts group, although during this period the demand for analytical and patent prior art searches has remained constant (see Figure 1). There has been an increase in the number of searches undertaken by end-users, reflecting their increased self-sufficiency, but this has been accompanied by an increase in the complexity of the searches requested of the information professionals. These observations indicate that the increased familiarity with databases by research chemists has led to a greater understanding of the capabilities of the information systems that enables them to formulate more sophisticated queries for the information scientists. The result is an increase

* Corresponding author phone: +44(0)161 200 4468; e-mail: helen.schofield@umist.ac.uk.

[†] GlaxoSmithKline.

[‡] UMIST.

10.1021/ci010360l CCC: $20.00
Published on Web 08/28/2001

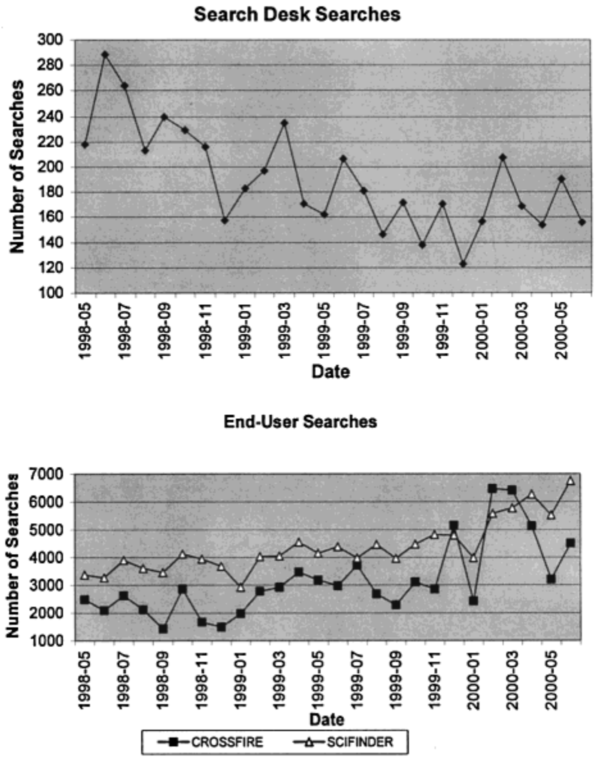


**Figure 1.** Trends in searching activity at SmithKline Beecham. The increase in the number of end-user searches (shown in the lower chart) has been accompanied by a decrease in the overall number of searches (upper chart) received by the information analysts' group and an increase in the complexity of these search requests.

in the time spent on each search request by information scientists.

There are a number of positive aspects to the present shift to end-user searching. Disintermediation leads to decreases in turn-around time for completion of searches and avoids a visit to the search desk and a possible delay in receiving results. End-user chemists can follow their own leads and take full advantage of serendipitous findings enabling them to embark upon tangential research with the potential of new, innovative ideas. This creative process can be inhibited when information scientists search on behalf of users, as most searches involving an intermediary need to be clearly defined and constrained. Additionally, the information scientist is striving to achieve accuracy and precision in the results obtained which tends to prevent serendipity.

However, there is a down side to end-user searching due to the inevitable lower skill level of the end-user compared to that of the highly trained information scientist. End-user searches tend to be less comprehensive and can result in poor recall and relevance of documents. Applications with behind the scenes intelligence, for example a system that uses a thesaurus to back up a natural language query, can help to mitigate these problems. This situation is probably exacerbated in the case of structure-based chemical information systems, which tend to be more complex and require more training than purely text-based systems.

Since databases are now commonly acquired as subscription packages rather than as pay-as-you-go services, decisions about purchases have greater implications than the selection of an online database for use on a one-off basis for an individual search. Most information professionals intuitively "know" the most appropriate and best quality sources available, but these must be rationalized and justified in terms of the user's ability to use the applications and their business cost/benefit. It is therefore essential to make a thorough evaluation of databases and their potential benefit prior to subscription, and this paper provides a model for evaluation of chemical information services that have recently become available.

## IDENTIFICATION OF PARAMETERS AND DETERMINATION OF FORMULA

Benefit can be defined in qualitative terms as the information and knowledge gained by an individual or organization which can then be applied to accelerate the research process, resulting in time savings and avoidance of repetition of work already published. To determine the benefit likely to be achieved by acquisition of a database or application, criteria for its evaluation must be identified and the product under consideration evaluated against those criteria.

Other authors have already identified some of the major quality criteria for the evaluation of databases. The EQUIP (European survey of quality criteria for the evaluation of databases) program of work included a survey of a large number of users to determine the relative importance of a number of criteria for assessment of online and CD-ROM services.[10] The survey respondents were users from throughout Europe and were mainly trained information professionals. There was variation in the ranking of the importance of the criteria depending on the location of the respondents, their organization type, user type (librarian/information scientist or end-user), and also medium of delivery under consideration (i.e. CD-ROM or online). Overall the most important criteria were deemed to be as follows: coverage (coverage of subject area; comprehensiveness; coverage as described by producer), accessibility/ease of use (the search software; online thesaurus; saved searches), timeliness (update period; currency of material included in each update), and consistency (of records within a database; consistent use of fields; consistent indexing). Although this work is of relevance to the present study, there are some significant differences. The study does not consider the client-server desk-top systems, predominantly for use by end-users, under consideration here, and the criteria were determined with all subject areas in mind rather than considering chemical information systems specifically.

A set of criteria for assessing reaction databases has been highlighted by Zass[11] who states that reaction databases should be judged by their information content and the means of accessing that information. The present authors are in agreement that the coverage of a database is its most important feature. If the reaction, reference, or structure is not contained in the database, no amount of sophisticated search functionality will extract the information.

Comparisons between chemical information systems have been made;[12-14] these have generally been with the view of identifying overlap between sources and assessing their capabilities of dealing with particular types of search request. Although comparisons of coverage form an important part of the present work, earlier studies have not taken the further step of analyzing sources from the viewpoint of their overall benefit to an organization and the incremental value of

**Table 1.** Search Functionality Subelements with Definitions and Explanations

| search type | functionality assessment criteria definitions and notes |
|---|---|
| structure | substructure, using atom lists, generic groups |
| reaction | substructure, reaction conditions, mapping, bond changes, valence changes |
| similarity | allows specification of the degree of "likeness" between two molecules or two reactions |
| property | physical, spectral, etc. and use of numeric searching techniques |
| bibliographic | titles, authors, controlled index terms, abstract keywords, etc. |
| full text | the whole text of the original article is indexed |
| search term expansion | ability to deal with unstructured and natural language text queries, including synonyms, abbreviations, and plurals, often achieved with the assistance of a thesaurus |
| basic index | insulates the user from knowing which data field to search. This is especially important in a complex data structure used infrequently by novice users e.g. basic index in CrossFire |
| context/proximity operators | availability of Boolean AND, OR, NOT operators, and enhancement of search precision through availability of operators which allow terms to be adjacent, present in the same sentence, field, paragraph, etc. |

**Table 2.** Nonsearch Functionality Subelements with Definitions and Notes

| feature | functionality assessment criteria definitions and notes |
|---|---|
| import/export data | support for multiple formats for text files e.g. for management of references using packages such as EndNote, for structure/reaction files e.g. Molfile, SMILES, and standard delimited text files for importing into analysis tools |
| link to best of breed tools | support for major applications such as Excel for data analysis, Word, 3D visualization and analysis tools |
| postprocessing/data analysis | groups together similar types of information based on user requirements allowing analysis of results from within the application, e.g. reaction clustering |
| 3D structure representation | the ability to draw, search for, and display three-dimensional structures or reactions |
| full text linkage | links to the original primary journal article, patent document, etc. |
| search history/reuse | ability to review search statements already performed and to reuse them as needed |

purchasing a new database or application when subscriptions to others are already in place.

To obtain a customizable framework for the estimation of fit and potential impact of new data sources, we propose a general formula that uses three major factors: ease of use, functionality, and coverage.

As discussed above, the content of a database has been identified by other authors and ourselves as being fundamental to the value of a system. To reflect the criticality of the content of a database, coverage has been included as a multiplier in the formula. Thus if the relevant content of a product under evaluation is low, as determined by the needs of an organization, then coverage will have a low value and hence the overall value of the product will be low. Alternate approaches might lead to an artificially high overall value in cases where the functionality and ease of use of a product are good but the coverage is poor. Hence, using appropriate weightings for each of the subelements within the major factors we can estimate this value by the following:

$$\text{value} = \text{benefit/cost per user}$$

and

$$\text{benefit} = [(x\Sigma_i F_i w_i + y\Sigma_i E_i w_i) \times \Sigma_i C_i w_i]/(x + y)$$

where F = functionality, E = ease of use, C = coverage, i = individual subelement, x and y are the weightings for functionality and ease of use, respectively, and w = weighting of individual subelement.

This is not meant to evaluate to a precise number with an absolute meaning. It is proposed as a semiquantitative method of analyzing the potential value to the customer and organization of purchasing additional data sources. The contributing factors and their subelements are outlined next.

**Functionality.** Structure and reaction search functionality is fundamental to a chemistry database, and essential requirements are detailed in Table 1. Increasingly, structure-based information systems are providing additional functionalities enabling the extracted data and information to be further manipulated and analyzed (Table 2). For example, Figure 2 shows a structure retrieved from SciFinder that has been exported into an additional application, WebLab Viewer, launched from within SciFinder; this has enabled the retrieved molecule to be visualized as a space-filling three-dimensional structure.

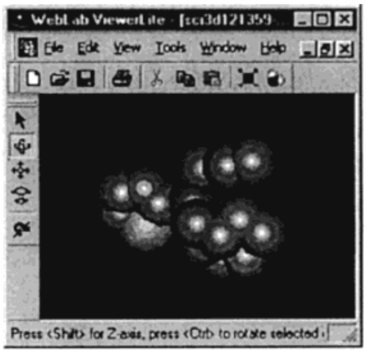
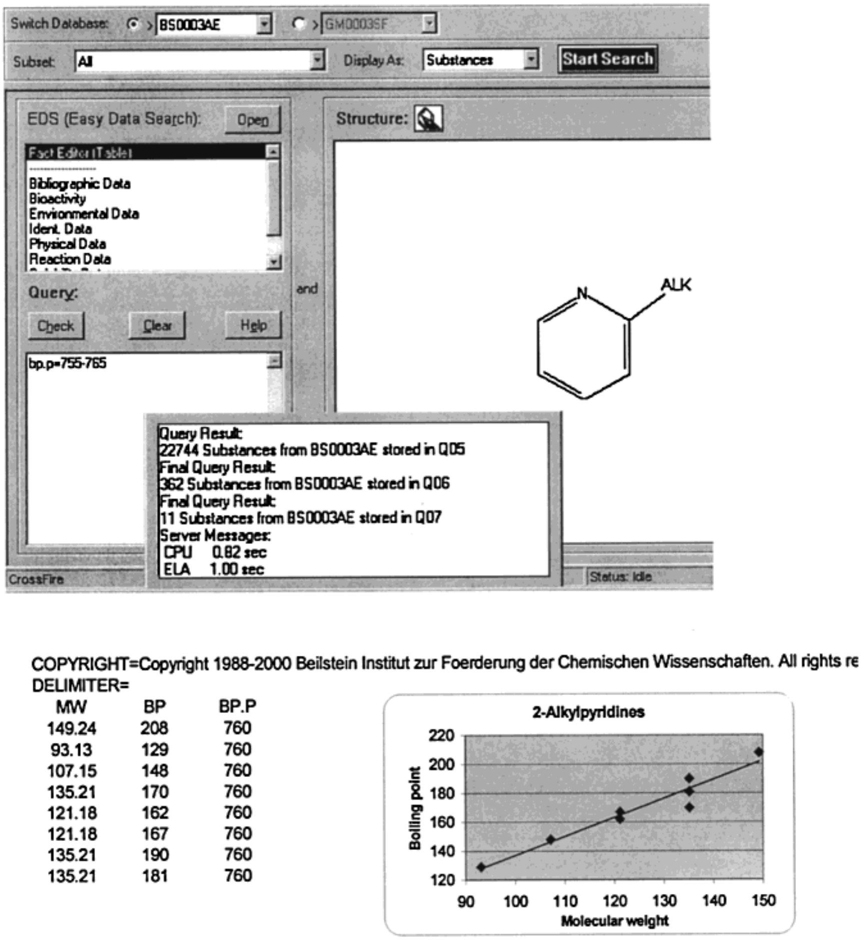

WebLab ViewerLite - [sci3d121359
File Edit View Tools Window Help
Press <Shift> for Z-axis, press <Ctrl> to rotate selected

**Figure 2.** Result of an automated export of a structure from SciFinder directly into an independent application, WebLab Viewer, where it can be further manipulated. (Reproduced with permission of MSI, a subsidiary of Pharmacopeia.)

Many systems allow statistical analysis of the results obtained. SciFinder allows analysis of the bibliographic references, for example by publication year, corporate source, or author that enables trends to be identified. The CrossFire Commander 2000 allows export of data into Microsoft Excel, from which the full functionality of the spreadsheet application can be exploited (see Figure 3). This feature would be evaluated under the criteria of "import/export data" and "links to best of breed tools" in Table 2. A feature of MDL's ISIS Reaction Browser is that the retrieved reactions can be clustered together, enabling rapid review of large answer sets by viewing the first few members of each cluster (see Figure 4); this feature would be evaluated under "postprocessing/data analysis" in Table 2.

A further example of linkage to a "best of breed" application is the "panorama" functionality used in Sci-



bp.p=755-765

Query Result:
22744 Substances from BS0003AE stored in Q05
Final Query Result:
362 Substances from BS0003AE stored in Q06
Final Query Result:
11 Substances from BS0003AE stored in Q07
Server Messages:
CPU 0.82 sec
ELA 1.00 sec

COPYRIGHT=Copyright 1988-2000 Beilstein Institut zur Foerderung der Chemischen Wissenschaften. All rights re
DELIMITER=

| MW | BP | BP.P |
|---|---|---|
| 149.24 | 208 | 760 |
| 93.13 | 129 | 760 |
| 107.15 | 148 | 760 |
| 135.21 | 170 | 760 |
| 121.18 | 162 | 760 |
| 121.18 | 167 | 760 |
| 135.21 | 190 | 760 |
| 135.21 | 181 | 760 |

**Figure 3.** Output of molecular weight and boiling point data from CrossFire Commander 2000, showing retrieval of data restricted to compounds with boiling points measured at atmospheric pressure. Data can be exported directly to Microsoft Excel. (CrossFire is a trademark of MDL Information Systems GmbH. CrossFire screenshot (c) Copyright 2001, MDL Information Systems, GmbH, all rights reserved. CrossFire table is also (c) Copyright 1988−2001 Beilstein Institut zur Foerderung der Chemischen Wissenschaften. All rights reserved. Materials used herein under permission.)

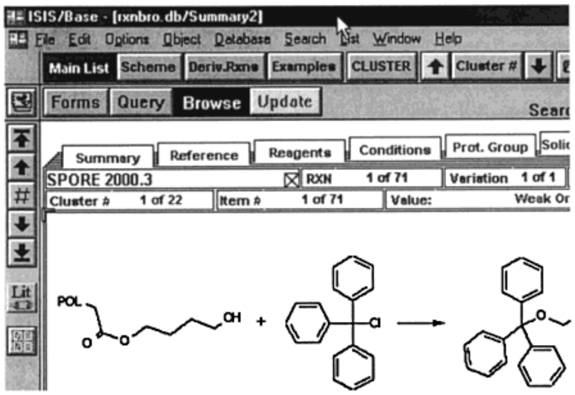
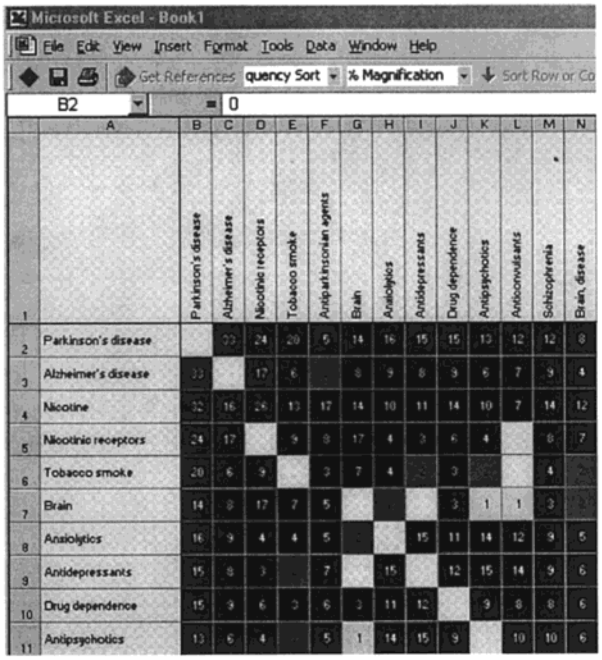


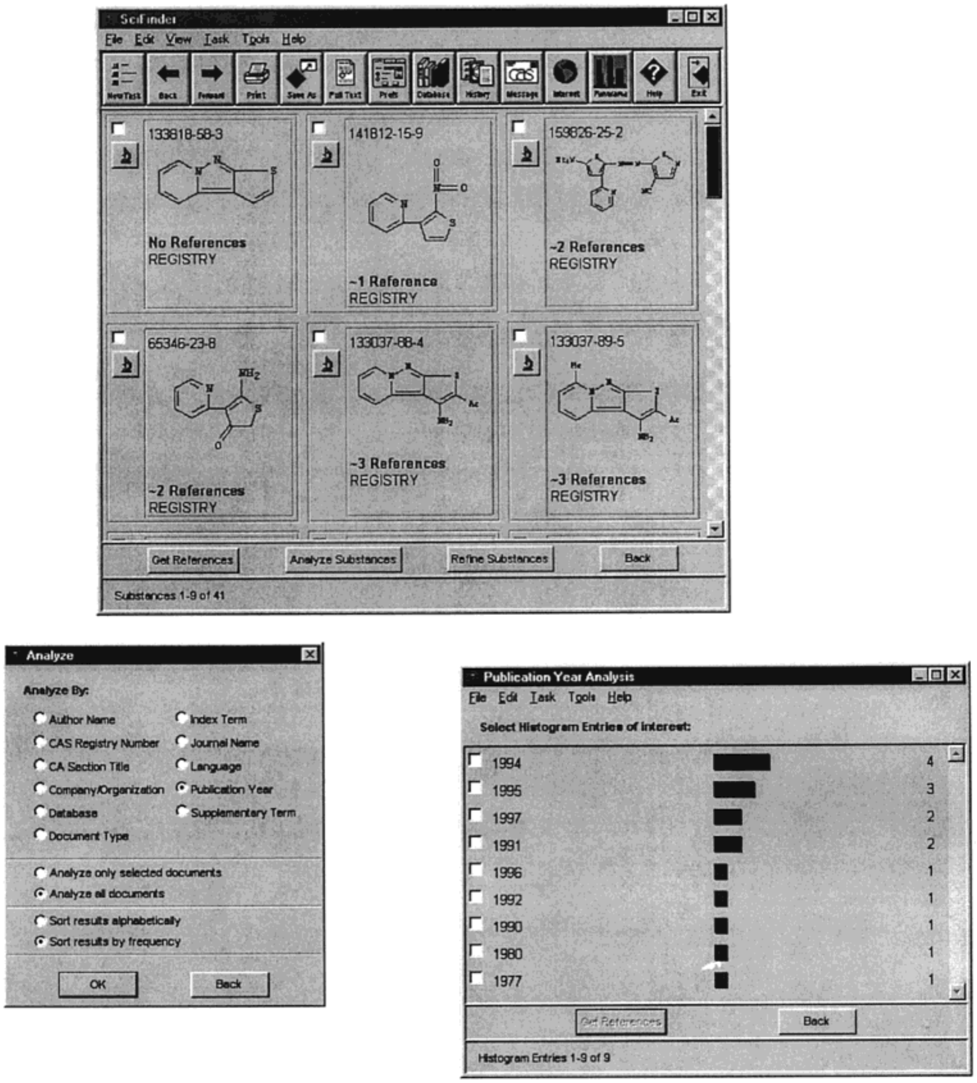
**Figure 4.** The first of 71 reactions from the first of 22 clusters extracted from an ISIS search. Scanning a few reactions from each cluster allows quick assessment of relevance of reactions retrieved. (ISIS is a trademark of MDL Information Systems, Inc. ISIS screenshot (c) Copyright 1991−2001, MDL Information Systems, Inc., all rights reserved, used herein under permission.)

Finder2000 (see Figure 5). Data are exported to Excel in the form of a matrix of related concepts that can be manipulated within Excel, and the individual cells retain their “linkage” through SciFinder2000 to the original bibliographic reference. By clicking on one of the Excel cells a user is taken back to SciFinder, and the relevant bibliographic



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | Parkinson's disease | Alzheimer's disease | Nicotinic receptors | Tobacco smoke | Antiparkinsonian agents | Brain | Anxiolytics | Antidepressants | Drug dependence | Antipsychotics | Anticonvulsants | Schizophrenia | Brain, disease |
| 2 | Parkinson's disease | | 33 | 24 | 20 | 5 | 14 | 16 | 15 | 15 | 13 | 12 | 12 | 8 |
| 3 | Alzheimer's disease | 33 | | 17 | 6 | | 8 | 9 | 8 | 9 | 6 | 7 | 9 | 4 |
| 4 | Nicotine | 32 | 16 | 26 | 13 | 17 | 14 | 10 | 11 | 14 | 10 | 7 | 14 | 12 |
| 5 | Nicotinic receptors | 24 | 17 | | 9 | 8 | 17 | 4 | 3 | 6 | 4 | | 8 | 7 |
| 6 | Tobacco smoke | 20 | 6 | 9 | | 3 | 7 | 4 | | 3 | | | 4 | |
| 7 | Brain | 14 | 8 | 17 | 7 | 5 | | | | 3 | 1 | 1 | 3 | |
| 8 | Anxiolytics | 16 | 9 | 4 | 4 | 5 | | | 15 | 11 | 14 | 12 | 9 | 5 |
| 9 | Antidepressants | 15 | 8 | 3 | | 7 | | 15 | | 12 | 15 | 14 | 9 | 6 |
| 10 | Drug dependence | 15 | 9 | 6 | 3 | 6 | 3 | 11 | 12 | | 9 | 8 | 8 | 6 |
| 11 | Antipsychotics | 13 | 6 | 4 | | 5 | 1 | 14 | 15 | 9 | | 10 | 10 | 6 |

**Figure 5.** Example of the panorama functionality of SciFinder, showing the numbers of references that contain combinations of concepts. (Used with permission of CAS, a division of the American Chemical Society.)


SciFinder
File Edit View Task Tools Help
NewTask Back Forward Print Save As Full Text Prefs Database History Message Internet Panorama Help Exit
Get References
Analyze Substances
Refine Substances
Back
Analyze
OK
Back
Publication Year Analysis
File Edit Task Tools Help
Get References
Back

□ 133818-58-3 No References REGISTRY
□ 141812-15-9 ~1 Reference REGISTRY
□ 159826-25-2 ~2 References REGISTRY
□ 65346-23-8 ~2 References REGISTRY
□ 133037-88-4 ~3 References REGISTRY
□ 133037-89-5 ~3 References REGISTRY

Substances 1-9 of 41

Analyze By:

○ Author Name
○ CAS Registry Number
○ CA Section Title
○ Company/Organization
○ Database
○ Document Type
○ Index Term
○ Journal Name
○ Language
● Publication Year
○ Supplementary Term

○ Analyze only selected documents
● Analyze all documents

○ Sort results alphabetically
● Sort results by frequency

Select Histogram Entries of Interest:

| | | |
|---|---|---|
| □ | 1994 | 4 |
| □ | 1995 | 3 |
| □ | 1997 | 2 |
| □ | 1991 | 2 |
| □ | 1996 | 1 |
| □ | 1992 | 1 |
| □ | 1990 | 1 |
| □ | 1980 | 1 |
| □ | 1977 | 1 |

Histogram Entries 1-9 of 9

**Figure 6.** Analyze function of SciFinder showing structure records retrieved from a search and the relevant bibliographic records subsequently analyzed by publication year. (Used with permission of CAS, a division of the American Chemical Society.)

**Table 3.** Factors that Contribute to the Ease of Use of a Database or Application

| positive factor | description |
|---|---|
| intuitive/self-guiding | interface functionality is obvious, comprehensive without being confusing and is consistent with user expectations |
| spiral design | application is designed to "lead" the novice user toward the desired end-point in a stepwise process |
| context sensitive help | specific help is available for the application context currently in use |

records pertaining to that cell are retrieved for further analysis.

**Ease of Use.** Factors leading to greater ease of use of a database are listed in Table 3. These increase the value of a database as they can contribute to time savings by end-users.

One such factor is "spiral design", an example being the "analyze" feature in SciFinder. Having arrived at a set of results the user is presented with several intuitive choices one of which is "analyze". By clicking the options provided the user is led, i.e., "spiralled" to a more precise set of results (see Figure 6). Results can thus be organized in accordance with criteria stipulated by the user, such as index term, document type, or supplementary term. These answer sets can then be selected and further analyzed if desired.

**Negative Factors.** The more interfaces within an application or between applications that have to be learned the less likely the application(s) will be used. There is a training component to each major interface within an application. If there is more than one application, then the interfaces should have a similar look and feel and method of operation. We consider the three major interfaces for chemical applications are as follows:

- main (the general interface for communication, text searching, and manipulation of answer sets)
- structure (the structure and reaction building interface)
- display (the structure, reaction, and reference viewing, and navigation interface)

If similar interfaces can be deployed or if interfaces share some common features with other applications, e.g. CrossFire can use the ISIS/Draw application as the structure-drawing interface, then this is a benefit. Structure drawing is a major interface hurdle for the user to overcome. Anecdotal evidence has shown that beyond two applications (i.e. a total of possibly six major interfaces) there are rapidly diminishing returns and beyond three there may be serious negative returns. Hence the need for a minimum number of interfaces between similar applications is very important. As stated above, where different applications may share a common component, e.g. a structure drawing package, this is reflected in a lesser training need in the evaluation spreadsheet.

**Coverage.** As for all subject areas, the coverage of a chemical database is crucial. Coverage encompasses the time period of the literature covered, the up-to-dateness of the information, the types of publications covered (e.g. patents, journal articles, conference proceedings), and the numbers of each of these. The type of information covered can vary, and for chemical information sources could include bibliographic references, structures, reactions, physical and chemical properties.

There is generally overlap between chemical information systems, but usually each has significant unique content. This unique content must be quantified according to an organization's needs. If an organization is not interested in the physical or spectral property content of a database, then its high score in a database has no particular value.

It is the unique incremental information of value to the organization that is key here and it is incumbent on the organization to quantify this incremental information according to its needs before it can be used in the overall equation. Ideally, we would seek the fewest data sources or systems that give us the greatest coverage. However, other variables, such as cost, can affect this approach. If more systems can cover the same information domain for lower cost, then this might be an alternative. However, now we have to consider that additional data sources often lead to additional interfaces each of which have to be learned. This is another negative factor in the overall equation as the preferred scenario for the end-user chemist would be to have a single interface for all chemical information systems, to simplify access to comprehensive coverage. Summing up, there should be minimal overlap between good quality data sources, and the incremental coverage has to be worth the additional cost.

The coverage subelements identified for evaluation are as follows:

- structure (numbers and types)
- reaction (numbers and types)
- properties (numbers and types)
- time period (retrospective)
- currency (up-to-dateness and frequency of updates)
- literature (range of journals; comprehensiveness)
- patents (countries covered and comprehensiveness)

**Cost.** There are always costs associated with the deployment and ongoing support of new applications. These could include the following: updating, monitoring, training, documenting, and troubleshooting as well as the sometimes hidden costs of hardware amortization and specialized support software for the application. These soft costs can be difficult to calculate, but IT/IS departments can generate reasonable estimates to show the full cost of ownership of the new data source. For example Table 4 shows a sample application with all these costs applied. In this example the total expenditure of $316 500 is far beyond the cost of the raw data at $250 000.

**Table 4.** Breakdown of the Costs Associated with an Example Application (AppX)

| | yearly totals | |
|---|---|---|
| AppX budget | hours | total |
| receive and load data | 30 | $3000 |
| test data load | 15 | $1500 |
| release to production | 20 | $2000 |
| train customers | 250 | $25 000 |
| ongoing support | 200 | $20 000 |
| contract negotiation | 25 | $2500 |
| license optimization | 25 | $2500 |
| purchase of data | | $250 000 |
| server cost per year | $10 000 | $10 000 |
| total expenditure | | $316 500 |
| number of searches run | | 53 000 |
| value of searches | | $530 000 |
| ratios | | |
| average cost per user | $1500 | |
| benefit ratio (search value/expenditure) | 1.67 | |
| number of searches per person per year | 251 | |
| labor rate $/hour | $100 | |
| estimated value per search | $10.00 | |
| number of users | 211 | |

This model does allow the ready calculation of meaningful financial ratios that can be used to optimize deployment and support. Also, it can give an estimation of the benefit/cost ratio which is calculated by dividing the value of searches (VoS) by the total expenditure. The VoS is derived by multiplying the number of end-user searches by the value per search. The value per search is estimated by looking at alternate ways the information might be obtained, for example, through an alternate online search or through a professional intermediary. The cost per user is calculated by dividing the total expenditure by the number of users; the worked example in Table 4 thus shows a cost per user of $1500 for the sample application, AppX.

## FORMULA APPLICATION

Each subelement of functionality, ease of use, and coverage is assigned values for the particular data source or application being reviewed. Weight factors for each subelement are assigned according to the needs of the organization and used to create a composite for the major factors. The major factors are combined, again with suitable weight factors to arrive at an overall score for benefit. As an example of the application of the formula, we consulted with information professionals and scientists in our institutions, SmithKline Beecham and UMIST, and assigned weight factors and values for each of the relevant subelements. These were assigned as perceived relevant to our institutions; individual assignments would need to be made in accordance with a particular organization's information needs and search requirements. It should be noted that values and weights may need to be reassigned if a new product comes on to the market, if the functionality, ease of use, or coverage of an existing product changes, or if the information needs of an organization change. Subelements can also be added or removed to reflect such changes.

Once subelements and weightings have been identified and values assigned, the benefit/cost ratio is then calculated using the cost per user values derived in Table 4, e.g. if we were to use this application the cost per user would be $1500.

**Table 5.** A Completed Spreadsheet for the Major Applications Discussed in the Text[a]

| wt factors | application | ISIS Rxn Browser (1) | SciFinder (2) | CrossFire (3) | AppX (4) |
|---|---|---|---|---|---|
| | benefit | 6.0 | 25.4 | 11.0 | 29.4 |
| | benefit/cost | 6.0 | 12.7 | 7.3 | 19.6 |
| | **factors** | | | | |
| | coverage | 3.04 | 5.66 | 4.90 | 6.16 |
| | cost per user ($k) | 1.00 | 2.00 | 1.50 | 1.50 |
| 6 | functionality | 3.60 | 4.14 | 4.05 | 6.41 |
| 8 | ease of use | 0.77 | 4.76 | 0.88 | 3.55 |
| | Functionality | | | | |
| 10 | structure query | 4 | 5 | 4 | 5 |
| 10 | reaction query | 4 | 4 | 4 | 5 |
| 2 | properties query | 0 | 1 | 5 | 4 |
| 10 | bibliographic | 3 | 7 | 5 | 7 |
| 8 | full text | 0 | 0 | 0 | 6 |
| 5 | context/proximity operators | 4 | 3 | 6 | 7 |
| 4 | basic index | 3 | 4 | 6 | 8 |
| 4 | import/export data | 5 | 4 | 6 | 7 |
| 4 | link to best of breed tools | 4 | 4 | 4 | 6 |
| 6 | search term expansion | 1 | 6 | 1 | 7 |
| 4 | postprocessing/data analysis | 5 | 4 | 3 | 4 |
| 3 | similarity | 6 | 4 | 1 | 7 |
| 8 | full text linkage | 6 | 5 | 6 | 10 |
| 4 | 3D structures | 6 | 5 | 5 | 5 |
| 3 | search history/reuse | 4 | 1 | 7 | 7 |
| | Ease of Use | | | | |
| 10 | intuitive/self-guiding | 4 | 7 | 4 | 7 |
| 7 | spiral design | 5 | 8 | 5 | 7 |
| 5 | context sensitive help | 6 | 6 | 5 | 5 |
| 8 | interface training (main) | −4 | −2 | −4 | −3 |
| 8 | interface training (structure) | −4 | −3 | −4 | −3 |
| 8 | interface training (display) | −4 | −2 | −3 | −3 |
| | Coverage | | | | |
| 10 | structure | 2 | 6 | 4 | 5 |
| 10 | reaction | 3 | 4 | 6 | 5 |
| 3 | properties | 0 | 3 | 7 | 5 |
| 4 | time period | 4 | 3 | 6 | 6 |
| 8 | currency | 4 | 8 | 4 | 8 |
| 8 | literature | 5 | 7 | 5 | 7 |
| 7 | patents | 2 | 6 | 4 | 7 |

[a] Data are from surveys at SmithKline Beecham and UMIST. AppX is added to illustrate a product to fit a specific niche, i.e. full text. Training needs are shown as negative values in the table for reasons outlined in the text.

Using the formulas

$$\text{benefit} = [(x\Sigma_i F_i w_i + y\Sigma_i E_i w_i) \times \Sigma_i C_i w_i]/(x + y)$$

and

$$\text{value} = \text{benefit/cost per user}$$

the results of such evaluations in our organizations are given in Table 5 for three leading chemical information systems. For example, the “ease of use” of CrossFire, which has a value of 0.88 in Table 5, is determined as follows:

$$(4 \times 10 + 5 \times 7 + 5 \times 5)/(10 + 7 + 5) + (-4 \times 8 - 4 \times 8 - 3 \times 8)/(8 + 8 + 8)$$

Thus it follows that the “value” of CrossFire (7.3) is calculated

$$[(4.05 \times 6 + 0.88 \times 8) \times 4.90]/(14 \times 1.5)$$

If, however, another organization considered that functionality was more important than ease of use, the weightings could be changed to, say, 7 and 5, respectively. The value of CrossFire to that organization would then change to

$$[(4.05 \times 7 + 0.88 \times 5) \times 4.90]/(12 \times 1.5) = 8.9$$

It should be noted that the version of ISIS evaluated here contains only a subset of the databases available through this interface. Also, CrossFire is being evaluated without the available EcoPharm and Gmelin data. In addition, the evaluation of SciFinder does not take account of the recently added references back to 1947. Also note that the “cost per user” numbers used in the tables are only for illustration purposes and do not reflect actual costs.

In reality, additional columns for in-house proprietary databases might be added to Table 5 to give a more thorough picture of the resources available.

Further estimates can be made to assess the overall benefit and benefit/cost for combinations of data sources or applications. Examples are given in Table 6 and Figure 7, which show evaluations for combinations of the systems analyzed individually in Table 5. (For brevity, numbers have been used to represent the applications. These are shown as column headers in Table 5.) Careful consideration to overlap of

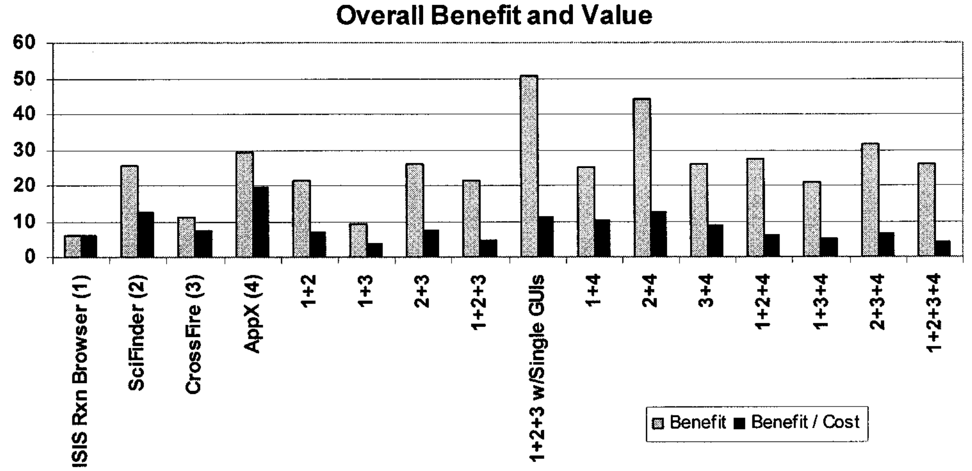


**Figure 7.** A graphical representation of the overall benefit and value of single applications and combinations of applications.

**Table 6.** Total Values of Combinations of Applications (as Numbered in Table 5)[a]

| | | combinations | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| wt factors | application | 1+2 | 1+3 | 2+3 | 1+2+3 | 1+2+3 single GUIs | 1+4 | 2+4 | 3+4 | 1+2+4 | 1+3+4 | 2+3+4 | 1+2+3+4 |
| | benefit | 21.6 | 9.1 | 26.1 | 21.5 | 50.8 | 25.3 | 44.3 | 26.0 | 27.4 | 21.1 | 31.7 | 26.1 |
| | benefit/cost | 7.2 | 3.7 | 7.4 | 4.8 | 11.3 | 10.1 | 12.7 | 8.7 | 6.1 | 5.3 | 6.3 | 4.4 |
| | **factors** | | | | | | | | | | | | |
| | coverage | 6.34 | 5.70 | 7.72 | 8.74 | 8.74 | 6.60 | 7.86 | 7.46 | 8.12 | 7.46 | 8.76 | 9.80 |
| | cost per user ($k) | 3.00 | 2.50 | 3.50 | 4.50 | 4.50 | 2.50 | 3.50 | 3.00 | 4.50 | 4.00 | 5.00 | 6.00 |
| 6 | functionality | 5.61 | 5.38 | 5.98 | 6.96 | 7.21 | 7.48 | 7.84 | 7.84 | 8.33 | 8.35 | 8.46 | 9.35 |
| 8 | ease of use | 1.76 | −1.23 | 1.42 | −0.91 | 4.77 | 1.11 | 3.98 | 0.21 | −0.33 | −1.32 | −0.02 | −2.35 |
| | Functionality | | | | | | | | | | | | |
| 10 | structure query | 7 | 7 | 7 | 9 | 9 | 8 | 8 | 8 | 9 | 9 | 9 | 10 |
| 10 | reaction query | 7 | 7 | 7 | 9 | 9 | 8 | 8 | 8 | 9 | 9 | 9 | 10 |
| 2 | properties query | 1 | 5 | 6 | 6 | 7 | 4 | 5 | 8 | 4 | 8 | 9 | 9 |
| 10 | bibliographic | 7 | 4 | 8 | 8 | 8 | 7 | 9 | 8 | 9 | 8 | 9 | 10 |
| 8 | full text | 0 | 0 | 0 | 0 | 0 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| 5 | context/proximity operators | 4 | 6 | 7 | 7 | 8 | 7 | 7 | 8 | 8 | 9 | 9 | 10 |
| 4 | basic index | 5 | 7 | 8 | 8 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 10 |
| 4 | import/export data | 6 | 7 | 6 | 7 | 7 | 8 | 7 | 9 | 8 | 9 | 8 | 10 |
| 4 | link to best of breed tools | 8 | 8 | 6 | 8 | 8 | 9 | 9 | 9 | 9 | 9 | 9 | 10 |
| 6 | search term expansion | 6 | 5 | 6 | 6 | 7 | 7 | 9 | 7 | 9 | 7 | 9 | 9 |
| 4 | postprocessing/data analysis | 7 | 1 | 5 | 7 | 7 | 6 | 5 | 5 | 7 | 7 | 6 | 8 |
| 3 | similarity | 7 | 6 | 4 | 7 | 7 | 8 | 8 | 7 | 8 | 9 | 8 | 10 |
| 8 | full text linkage | 6 | 6 | 6 | 7 | 8 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| 4 | 3D structures | 6 | 7 | 6 | 7 | 7 | 6 | 6 | 6 | 7 | 8 | 7 | 9 |
| 3 | search history/reuse | 4 | 7 | 7 | 8 | 8 | 7 | 7 | 9 | 8 | 8 | 8 | 9 |
| | Ease of Use | | | | | | | | | | | | |
| 10 | intuitive/self-guiding | 7 | 4 | 7 | 7 | 8 | 7 | 9 | 7 | 9 | 9 | 9 | 9 |
| 7 | spiral design | 8 | 5 | 8 | 8 | 8 | 7 | 9 | 7 | 8 | 7 | 9 | 9 |
| 5 | context sensitive help | 6 | 6 | 6 | 6 | 7 | 6 | 6 | 5 | 6 | 6 | 6 | 6 |
| 8 | interface training (main) | −5 | −7 | −6 | −9 | −3 | −5 | −4 | −7 | −8 | −11 | −8 | −12 |
| 8 | interface training (structure) | −6 | −4 | −6 | −6 | −3 | −6 | −5 | −6 | −8 | −6 | −8 | −8 |
| 8 | interface training (display) | −5 | −7 | −5 | −9 | −3 | −6 | −4 | −6 | −9 | −10 | −9 | −12 |
| | Coverage | | | | | | | | | | | | |
| 10 | structure | 7 | 6 | 8 | 9 | 9 | 6 | 8 | 7 | 8 | 7 | 9 | 10 |
| 10 | reaction | 5 | 6 | 7 | 8 | 8 | 5 | 6 | 7 | 7 | 7 | 7 | 9 |
| 3 | properties | 1 | 7 | 8 | 8 | 8 | 5 | 6 | 8 | 7 | 8 | 8 | 10 |
| 4 | time period | 6 | 7 | 7 | 9 | 9 | 7 | 7 | 8 | 7 | 8 | 8 | 10 |
| 8 | currency | 8 | 5 | 8 | 9 | 9 | 8 | 9 | 8 | 9 | 8 | 9 | 10 |
| 8 | literature | 8 | 6 | 8 | 9 | 9 | 8 | 9 | 8 | 9 | 8 | 10 | 10 |
| 7 | patents | 6 | 4 | 8 | 9 | 9 | 7 | 9 | 7 | 9 | 7 | 10 | 10 |

[a] Training needs are shown as negative values in the table for reasons outlined in the text.

coverage and functionality must be made at this stage again taking into account the information needs of the organization.

To illustrate the likely savings on training and the potential increased search efficiency of users if a common graphical user interface were to exist for applications 1, 2, and 3, a column has been included in Table 6 to demonstrate this advantage. In making this evaluation, we have assumed that the full functionality of each application would be retained, which in practice is unlikely to be the case.

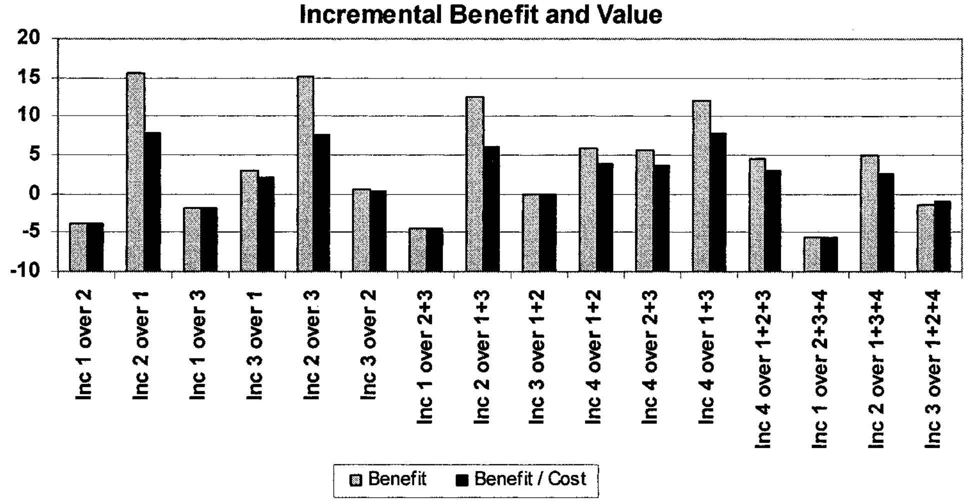


**Figure 8.** A graphical representation of the incremental benefit and value of additional applications. Application numbers are as defined in Figure 7.

**Table 7.** Incremental Value of Deploying Additional Applications

| | incremental values | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| application | inc 1 over 2 | inc 2 over 1 | inc 1 over 3 | inc 3 over 1 | inc 2 over 3 | inc 3 over 2 | inc 1 over 2+3 | inc 2 over 1+3 | inc 3 over 1+2 | inc 4 over 1+2 | inc 4 over 2+3 | inc 4 over 1+3 | inc 4 over 1+2+3 | inc 1 over 2+3+4 | inc 2 over 1+3+4 | inc 3 over 1+2+4 |
| benefit | −3.8 | 15.6 | −1.8 | 3.1 | 15.1 | 0.6 | −4.5 | 12.4 | −0.1 | 5.8 | 5.6 | 11.9 | 4.6 | −5.6 | 5.0 | −1.3 |
| benefit/cost | −3.8 | 7.8 | −1.8 | 2.1 | 7.5 | 0.4 | −4.5 | 6.2 | 0.0 | 3.9 | 3.7 | 8.0 | 3.1 | −5.6 | 2.5 | −0.9 |

From the previous calculations, the incremental value of acquiring another data source over existing ones can be estimated. This is simply calculated by subtracting the value of an application or combination of applications already purchased from the value for the new combination including the new application under consideration. Again, these numbers only serve to indicate the relative value of a purchase decision. Sample results of this analysis are presented in Table 7 and Figure 8. It should be noted here that values are relative rather than absolute and that negative values merely suggest that a specific combination could represent poor value for money for an organization.

Sometimes additional data sources are purchased for very specific needs. The addition of a hypothetical new application under consideration for purchase, AppX, illustrates this point (see Tables 4 and 5). In this case, let us suppose that it is decided that an organization requires a product that is strong in full text retrieval and AppX meets that criterion. In this case, another combination of data sources might prove more cost-effective. For example, the combination 4 + 2 + 3 appears stronger than 4 + 1 + 2. There will be retraining costs that must be accounted for and are not shown in our model.

The example of the incremental value of AppX over an existing suite consisting of applications 1 and 2 shows only a small positive value. This is primarily due to the fact that several additional interfaces have to be learned, and the incremental coverage and cost cannot balance this out. In this instance it would appear that the purchase of AppX is not prudent.

Disregarding the AppX data, our model indicates that for our organizations the combination of SciFinder and CrossFire is a particularly good one. This combination offers good overall benefit at a reasonable benefit/cost ratio, although it must always be borne in mind that the cost of applications may vary between organizations as for some products local deals are negotiated with suppliers, which might lead to differences in the assessment. As stated above, the overall benefit of different combinations may vary according to the needs of the organization, and this will be reflected in the values assigned against each criterion. Modification of the individual contributing factors or the weight factors for an organization's or individual's specific needs might well produce a different conclusion. This ready customization is a reflection of the power of this type of modeling approach.

Also noteworthy are the large incremental benefits that can be realized, e.g. purchasing 2 when only 1 exists or purchasing 2 when only 3 exists. These applications are good complements. However, purchasing 1 when 3 exists already appears not to be beneficial. The graphs show this quite clearly.

## CONCLUSIONS

The new generation of database applications under consideration in this study have only existed in their present form for around 5 years and have only become widely implemented during the last 3 to 4 years. Their use by the end-user community is still growing and will continue to do so as applications become more user-friendly and their functionality increases. So, their full impact may not yet have been realized. Ultimately users will define the value of systems by their usage patterns, although the efficiency and effectiveness of use is not being measured here.

It is the responsibility of information specialists to understand and evaluate the content, scope, and coverage of potential new resources and hence identify the most suitable, cost-effective, and relevant products to purchase for their organization. We believe this study to be the first to systematically lay down evaluation criteria for desktop chemical structure and reaction databases, with a view to aiding purchasing decisions. Some previous studies, although based on sound methodology, go into so much depth that they become unmanageable in this context or are not relevant to the type of system under consideration at present. Most importantly, the present study provides a working model that an organization can use to measure the incremental value to be gained by adding a new database to an established collection.

The proposed formula is flexible and can be customized by organizations, depending on their individual information needs or circumstances, both in terms of the criteria included and the weightings applied. The model should be reapplied when a new database is under consideration for purchase, with the assessment criteria set according to the content or requirements of the new product. Additionally, if cancellations are needed, for example when information budgets are cut or when an organization's research direction changes, then the formula can be used to indicate possible candidates for cancellation. The formula should also be applied when a change occurs to an existing database or application; for example, recently there was a substantial increase to the coverage of the Beilstein CrossFire database to enhance its coverage of pharmacological and ecological data, accompanied by an increase in the cost of the database. Subscribers have the option to take out a subscription to this new information, and it would be necessary to examine the likely additional value to the organization. Similarly if new functionality is added to an existing application, this may make another application redundant so the formula should be reapplied. This also applies if a new interface becomes available which makes the application easier to use by the end-user, such as is anticipated with the CrossFire Web Client which should greatly increase the ease of use of CrossFire.

A key theme throughout is the benefit/cost ratio which can be determined by comparing with typical online costs. Changes in these costs, or other methods of obtaining the same information, should also be continually monitored.

As usage of chemical information systems continues to move toward applications designed for end-users, the costs of such systems continue to increase, and more databases become available for end-users, the need for a semiquantitative analysis of this type will become increasingly important.

Although this paper has described the application of a formula to the evaluation of chemical systems, this approach is equally valid for other scientific systems e.g. biomedical databases and applications.

## REFERENCES AND NOTES

(1) Schofield, H. Chemists lead the way. *Chem. Ind.* **2000**, 810−812.

(2) Fisher, J.; Bjorner, S. Enabling online end-user searching: an expanding role for librarians. *Special Libraries* **1994**, *85*, 281−291.

(3) Oldroyd, B. K. Study of strategies used in online searching 5: differences between the experienced and inexperienced searcher. *Online Rev.* **1984**, *8*, 233−244.

(4) Warr, W. A.; Haygarth Jackson, A. R. End user searching of CAS ONLINE. Results of a cooperative experiments between Imperial Chemical Industries and Chemical Abstracts Service. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 68−72.

(5) Fourie, I. Should we take disintermediation seriously? *The Electronic Library* **1999**, *17*, 9−16.

(6) Buntrock, R. E.; Valicenti, A. K. End-users and chemical information. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 203−207.

(7) Lawson, A. J. CrossFire*plus*Reactions. In *The Beilstein system: strategies for effective searching*; Heller, S. R., Ed.; American Chemical Society: Washington, DC, 1998; pp 73−97.

(8) Meehan, P.; Schofield, H. CrossFire: A structural revolution for chemists. *Online Inf. Rev.* **2001**, *15*, in press.

(9) Ridley, D. D. Strategies for chemical reaction searching in SciFinder. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1077−1084.

(10) Wilson, T. D. EQUIP: a European survey of quality criteria for the evaluation of databases. *J. Inf. Sci.* **1998**, *24*, 345−357.

(11) Zass, E. Using the Beilstein reaction database in an academic environment. In *The Beilstein system: strategies for effective searching*; Heller, S. R., Ed.; American Chemical Society: Washington, DC, 1998; pp 99−131.

(12) Parkar, F. A.; Parkin, D. Comparison of Beilstein CrossFire*plus*Reactions and the selective reaction databases under ISIS. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 281−288.

(13) Voigt, K.; Brüggemann, R. Evaluation criteria for environmental and chemical databases. *Online CD-ROM Rev.* **1998**, *22*, 247−261.

(14) Voigt, K.; Gasteiger, J.; Brüggemann, R. Comparative evaluation of chemical and environmental online and CD-ROM databases. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 44−49.

CI010360L