Comparative Evaluation of Chemical and Environmental Online and CD-ROM Databases

Kristina Voigt,*,† Johann Gasteiger,‡ and Rainer Brüggemann§

GSF-Forschungszentrum für Umwelt und Gesundheit, Institut für Biomathematik und Biometrie, D-85764 Neuherberg, Germany, Computer-Chemie-Centrum, Institut für Organische Chemie, Universität Erlangen-Nürnberg, D-91052 Erlangen, Germany, and Institut für Gewässerökologie und Binnenfischerei, D-12587 Berlin, Germany

Received July 26, 1999

In a constantly expanding world of chemical and environmental information sources, the need for their evaluation gains more and more importance. This paper presents a comparative evaluation of datasources of online databases and databases on CD-ROM (called CD-ROMs in this paper) in the field of environmental chemicals. The approach is based on research results gained in the years 1996/1997. The authors are aware that changes in the database industry may lead to different results. Before the actual evaluation process can be carried out, two major procedures are necessary, namely, the selection of sets of datasources and the definition of evaluation criteria. In order to perform the difficult task of an evaluation based on several criteria, a general order relation has to be introduced. Methods of partially ordered set theory are applied, and the results are visualized by the technique of Hasse diagrams. On the basis of these evaluation results, the datasources are grouped and then evaluated. It will be shown that there are groups of datasources with quite specific property profiles, and only two groups turn out to be relatively better than the others.

1. INTRODUCTION

Chemistry and the environmental sciences are scientific disciplines with an enormous output of and demand for data. Presently, the chemical literature grows by approximately 500 000 publications per year. Since there is no indication that the increase in information in these fields will slow down in the foreseeable future, we shall have to cope with a growing flood of chemical and environmental information. A scientific approach is urgently needed to deal with this vast amount of information.[1]

The enormous increase in chemical and environmental information implies a rise in online databases, CD-ROMs, and Internet resources in these fields. A clear management strategy is required to handle this variety of datasources:

(1) Metadatabases must be defined.

(2) Datasources must be selected and grouped.

(3) Evaluation criteria must be given.

(4) A ranking procedure based on these criteria must be performed.

This article will explain in detail such a management strategy.

2. ELEMENTS OF THE MANAGEMENT STRATEGY

2.1. Summary of Applied Management Strategy. The evaluation process can be broken down into several steps which are shown in Figure 1. As a first step, representative groups of datasources must be established in order to perform a pragmatic evaluation procedure. These groups are called the *sets of datasources or sets of objects*. In our case, objects such as online databases and CD-ROMs are datasources. Second, evaluation criteria must be defined. These can be general ones, such as the price of the datasource, or those criteria specific to chemistry such as identification parameters for chemicals, or environmental-specific ones, such as environmental information parameters. The scoring is carried out according to a six number scoring system. After the definition of the individual criteria for evaluation and an assignment of values to these contributions, the final evaluation will be performed using the mathematical method known as the Hasse diagram technique. Important datasources will be identified by this procedure.

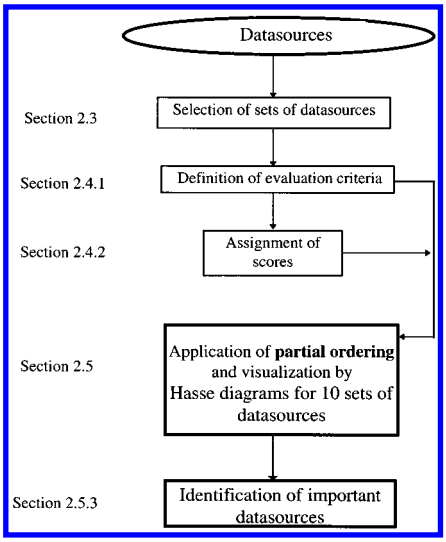

Figure 1. Steps of the procedure for the evaluation of datasources.

* Corresponding author.
† Institut für Biomathematik und Biometrie.
‡ Universität Erlangen-Nürnberg.
§ Institut für Gewässerökologie und Binnenfischerei.

10.1021/ci9900837 CCC: $19.00
Published on Web 11/23/1999

Table 1. Groups of Datasources[a]

group	name of group	classification scheme			no. of objects
1ON	bibliographic chemical online databases	BI	C	ON	28
1CD	bibliographic chemical CD-ROMs	BI	C	CD	26
2ON	numeric chemical online databases	NU	C	ON	18
2CD	numeric and chemical reaction CD-ROMs	NU	C	CD	19
3ON	bibliographic environmental online databases	BI	E	ON	19
3CD	bibliographic environmental CD-ROMs	BI	E	CD	42
4ON	numeric environmental online databases	NU	E	ON	14
4CD	numeric environmental CD-ROMs	NU	E	CD	20
5ON	environmental chemical online databases	EC		ON	18
5CD	environmental chemical CD-ROMs	EC		CD	25

[a] Abbreviations: CD, CD-ROM; ON, online; BI, bibliographic; NU, numeric; C, chemical; E, environmental; EC, environmental/chemical.

2.2. Metadatabases on Environmental Chemicals. The datasources have been compiled in two metadatabases (databases of datasources), the DADB, Metadatabase of Online Databases, and the DACD, Metadatabase of CD-ROMs. These metadatabases register information on the contents of the primary datasources, the database types (e.g. bibliographic, numeric), and other relevant information such as host, producer, cost, frequency of updating, etc.[2,3] DADB comprises 453 online databases, and DACD 347 sources.

The comparative evaluation of online databases and CD-ROMs reported in this paper is largely based on the contents of these metadatabases.

2.3. Selection of Sets of Datasources. We apply three different schemes to classify datasources in this approach: (1) the kind of medium, online or CD-ROM (ON, CD); (2) the discipline, chemical, environmental, or environmental/chemical datasources (C, E, EC); (3) the type of datasource, bibliographic or numeric (BI, NU). These three classification principles give 12 combinations ($2 \times 3 \times 2$). For technical reasons, the numeric and bibliographic environmental/chemical datasources were treated as a single group. This leaves 10 different groups: five sets of CD-ROMs and five sets of online databases.[4] These groups are called sets of objects in our approach and are listed in Table 1.

2.4. Definition of Evaluation Criteria and Assignment of Scores. 2.4.1. Definition of Evaluation Criteria. Criteria have to be defined to evaluate these 10 groups of datasources listed in Table 1. The chosen evaluation criteria can be distinguished into the following: general criteria such as the size of a datasource and content-specific environmental and/or chemical criteria. A series of evaluation criteria has been chosen to help the user finding the best profile for her/his own purpose. Table 2 lists these evaluation criteria.

The content-specific criteria (see lower section of Table 2) are more important to the group of users considered here (chemists, environmental scientists) than the general evaluation criteria (upper section of Table 2).

Some of the evaluation criteria dealing with chemical information, as well as the scoring system, will be explained in more detail in section 2.4.3.

2.4.2. Assignment of Scores. As a next step, a scoring scheme has to be provided for each of these criteria in order to obtain a numerical value for each criterion.

The comparative evaluation of chemicals in environmental sciences is often called a ranking or scoring system. Such scoring systems were developed in the late 1970s and early 1980s. They were tested and improved, and some of them are still in use today. Examples are the Interagency Testing Committee Scoring System,[5] the Swedish ESTER System,[6] and the German BUA-System.[7] Scoring systems of chemicals are still important in environmental research as recent publications prove.[8,9]

Table 2. Evaluation Criteria for Chemical and Environmental Datasources

acronym	evaluation criterion	type of criterion
SI	size of datasource	general
UP	frequency of updating of datasource	general
CO	cost of one hour online searching or of CD-ROM	general
AV	availability on other media	general
NU	number of chemicals	content-specific
ID	identification parameters for chemical substances	content-specific
IP	information parameters for chemicals (with environmental relevance)	content-specific
US	use of chemical substances	content-specific
QU	quality of online database/CD-ROM	content-specific

Table 3. Scores and Their Meaning

score	meaning	score	meaning
5	excellent	2	acceptable
4	very good	1	poor
3	good	0	insufficient

Scoring systems should generally be regarded as tools for setting priorities among objects, in our case datasources. Following these approaches to a certain extent, we elaborated a scoring system consisting of six levels, each represented by a number, which is explained in Table 3.

2.4.3. Explanation of Two Selected Evaluation Criteria. The evaluation of datasources depends on the selected criteria and on the extent to which these criteria are fulfilled. Criteria can be classified as follows: "ordinal criteria" (the extent of fulfilling these criteria is given by numerical values[10]); "nominal criteria" (the fulfillment of these criteria depends on whether a certain characteristic is present or not[11]). The criterion "number of chemicals found in a datasource" is an example of an ordinal criterion. A database covering a huge number of chemicals is better than another one treating only a few substances. The criterion "identification parameters for chemical substances (e.g. CAS Registry number)" is a nominal one. A datasource in which the search by CAS Registry number is available is better than a datasource where this search is not possible.

Criteria can often be organized in a hierarchy, where a so-called super/subcriterion relation holds. In this case a preselection of a set of criteria is induced which arises from some supercriteria. Details are explained by Brüggemann and Bartel.[12] In the present paper, however, the set of criteria arises from the pragmatic use of datasources, and the importance of each criterion has to be carefully examined. The theoretical background is explained by Brüggemann et al.[13]

The following six characteristics for the criterion "identification parameter for chemical substances" have been selected:

Is the structure searchable?

Is the structural formula given?

Is the molecular weight available?

Table 4. Numeric and Chemical Reaction CD-ROMs (2CD), Their Criteria, and Their Scores[a]

no.	CD-ROM	akk	SI	UP	CO	AV	NU	ID	IP	US	QU
1	CANDATA	CAN	5	3	4	1	4	2	1	4	2
2	CHEM-BANK	CHB	5	3	1	2	4	3	2	2	2
3	ChemReact 10	CR1	5	2	1	1	5	4	0	2	5
4	ChemReact 32	CR3	5	2	0	1	5	4	0	2	5
5	CHEM Source	CHS	3	3	4	1	3	2	4	4	1
6	Current Facts	CUR	5	3	4	0	5	5	1	2	4
7	Dictionary of Drugs	DID	4	2	2	1	4	5	2	3	4
8	Dictionary of Natural Products	DIN	3	2	0	1	3	5	2	3	3
9	Dictionary of Organic Composition	DOC	5	2	0	2	5	5	2	2	5
10	EINECS-Plus	EIN	5	1	2	2	5	2	0	2	4
11	Gefahrgut-CD-ROM	GEF	5	1	0	0	5	2	0	4	3
12	Giftliste	GIF	3	2	5	1	3	2	1	4	2
13	MSDS	MSD	3	3	4	1	3	2	2	4	1
14	NIOSHTIC/DIDS	NIO	5	3	4	1	5	2	2	4	3
15	PEST-BANK	PES	4	3	2	1	4	1	0	3	2
16	RTECS	RTE	5	3	4	2	5	3	2	2	2
17	Sax, Hawley	SAH	4	1	4	1	4	3	1	2	2
18	SIGEDA	SIG	2	2	2	1	1	2	2	5	1
19	SORBE	SOR	3	1	4	1	3	2	1	4	2

[a] Abbreviations for evaluation criteria can be found in Table 2, footnote *a*.

Is the CAS Registry number available?

Is the chemical name given?

Are only synonyma or trades names, etc., available?

All of these characteristics have been aggregated into a one number scheme for the ranking process. Clearly, the most important criterion is whether a molecular structure can be searched for or not. A datasource which comprises only the common names or synonyma obtains a rather low rating. We assigned scores to the six characteristics given above and combined them. The combination containing all six nominal criteria is assigned the highest score, 5, and for the combination CAS Registry number, chemical name, and synonyma the score 1 is given. The complete procedure is detailed by Voigt and Brüggemann.[14]

The evaluation procedure can be carried out by applying the choice of groups of datasources, evaluation criteria, and scores described above.

2.5. Application of Partial Ordering and Visualization by Hasse Diagrams. 2.5.1. Basic Principles. Anyone searching for chemical and environmental information has high interest in the evaluation of datasources. However, a potential user of datasources needs more information than only the quality of the datasource. He/she wants to know the quality of the datasource with respect to his/her own special interests (evaluation criteria). That is the reason why a general quality function is not of much use for the comparative evaluation of datasources. All relevant evaluation criteria given in section 2.4.1 should be examined at the same time. The following question must therefore be answered: How can objects (datasources) be compared if more than one criterion should be considered? In order to answer this question some simple concepts from the theory of partial order will be explained for the evaluation of datasources. It is assumed that each datasource is characterized by a list of values (or scores) corresponding to ordinal and nominal criteria. In such a case, datasource A is better than datasource B if all scores of datasource A are better than those of datasource B. However, not every datasource is always comparable with any other datasource because some scores might be conflicting. In this situation, only a partial order and not a total order (any object can be compared with any other object) can be established. A partial order can be visualized by so-called Hasse diagrams. Applying such Hasse diagrams, significant datasources can readily be recognized. Hasse diagrams visualize so-called maximum and minimum objects. Maximum objects are the "best", whereas minimum objects are the "worst" objects with respect to the evaluation criteria applied. These maximum and minimum objects are usually considered to be of primary interest in the evaluation procedure. A more extensive discussion of the application of Hasse diagrams in the evaluation of databases can be found in other publications.[15,16]

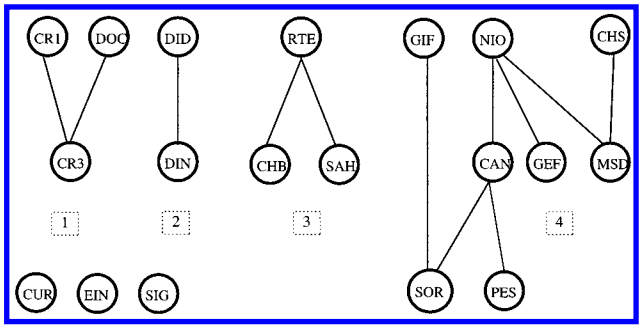

Figure 2. Hasse diagram for the 19 numeric and chemical reaction CD-ROMs of Table 4.

2.5.2. Hasse Diagram for 19 Numeric Chemical CD-ROMs. Of the 10 groups of datasources (see Table 1), we take the group of 19 numeric chemical CD-ROMs (2CD) as an example. This set of datasources also contains reaction databases. Reaction databases are constantly enlarged and improved in size, scope, and user-friendliness[17] and should therefore be considered in this approach together with the numeric datasources. The set of objects, i.e., CD-ROMs, are given in Table 4 together with the scores for their nine evaluation criteria.

Figure 2 shows a Hasse diagram where all nine criteria given in Table 4 are considered simultaneously. The construction of Hasse diagrams is explained in detail by Brüggemann and Halfon.[18]

The Hasse diagram of Figure 2 consists of four isolated diagrams which are called *nontrivial hierarchies* (or, in graph theoretical terms, four components of the graph): the central Hasse diagram (4) on the right hand side, two diagrams (1, 3) that contain only three objects, and one diagram which consists of only two objects (2). It is demonstrated that a specific datastructure is responsible for that phenomenon.[18]

The Hasse diagram shows seven maximum CD-ROMs, objects which have no upper but only lower neighbors (CR1, DOC, DID, RTE, GIF, NIO, CHS). So-called "*isolated objects*" such as CUR, EIN, and SIG can either be regarded as maximum or as minimum objects. The maximum CD-ROMs are better compared to those CD-ROMs connected by direct lines in a downward direction, e.g. ChemReactl0 (CR1) is ranked higher than ChemReact32 (CR3). In order to explain Figure 2 in more detail, NIO (NIOSHTIC/DIDS) is taken as an example. It has five other CD-ROMs which obtained a lower score. In other words, NIO has the following successors: CAN, GEF, MSD, SOR, and PES. (For identification of the CD-ROMs, see Table 4.) *Minimum* objects

are those CD-ROMs for which no CD-ROMs of lower score exist within their diagram. In this evaluation approach minimum objects are as follows: CHB, CR3, DIN, GEF, MSD, PES, SAH, and SOR. As mentioned above, CUR, EIN, and SIG are isolated objects, CD-ROMs which cannot be compared with any other of the remaining objects. CUR, for example, contains chemical structures, numeric data, and bibliographic information for approximately 300 000 chemicals. However this CD-ROM covers "only" the period of the preceding 12 months and is only available on CD-ROM. CUR contains only a few environmental information parameters.[19] It follows that this CD-ROM is superior to other objects in many criteria, whereas inferior in other parameters. This underlines its isolated position in our ranking scheme.

2.5.3. Evaluation Results for 10 Sets of Datasources and Identification of Important Datasources. The Hasse diagram for the numeric CD-ROMs given and discussed in section 2.5.2 shows which CD-ROMs are important. These are the maximum, minimum, and isolated objects already discussed. Each group of datasources named in Table I will now similarly be elaborated.

Analyzing all five object sets for online databases (ION−50N), we found as maximum objects not only the comprehensive and commonly known large and well-established online databases BEI (Beilstein Database), CAS (Chemical Abstracts), CAB (CAB Abstracts), and RTE (Registry of Toxic Effects of Chemical Substances) but also databases less known in chemistry and environmental sciences, such as CBNB (Chemical Business News Database), CSN (Chemical Safety Newsbase), HSD (Hazardous Substances Databank), MER (The Merck Index Online), MSD (Material Safety Data Sheets from Occupational Health Service), and NTI (National Technical Information Service). On the other hand, it is rather surprising that the popular environmental databases ULI (Umweltliteraturdatenbank), UFO (Umweltforschungsdatenbank), and POL (Pollution Abstracts) are not maximum objects in the evaluation process. The reason for this can be explained by their lack of chemically relevant parameters, such as identification parameters (ID), and their low number of chemicals (NU).

Analyzing the five object sets for CD-ROMs (ICD−5CD), the situation is as follows: The CD-ROM versions of large online databases, e.g., BIOSIS (BIOSIS Previews CD-ROM), RTECS (Registry of Toxic Effects of Chemical Substances), and TOX (Toxicology Literature Online CD-ROM), are maximum objects in our Hasse diagrams. The two thematic excerpts of the Chemical Abstracts Database CAH (CASurveyor Hazardous Materials) and CAV (CASurveyor Pollution Control and the Environment) also receive positive evaluations. The same good evaluation is valid for DID (Dictionary of Drugs on CD-ROM), DOC (Dictionary of Organic Compounds on CD-ROM), NIO (NIOSHTIC/DIDS), and PON (Pollution and Toxicology Database on CD-ROM), which are not available online. Striking is the fact that the German CD-ROMs like GEF (Gefahrgut CD-ROM), GIF (Giftliste, krebserregende, gesundheitsschädliche und reizende Stoffe), and SIG (Siemens Gefahrstoff-Datenbank CD-ROM) are in low ranking positions. One exception is CDR (CD-Römpp), which is an isolated object.

Table 5. Distribution of Evaluation Criteria for Numeric Chemical CD-ROMs (2CD)[a]

	n_{ij}								
score	SI	UP	CO	AV	NU	ID	IP	US	QU
0	0	0	4	2	0	0	5	0	0
1	0	4	2	13	1	1	5	0	3
2	1	7	4	4	0	9	8	8	7
3	5	8	0	0	5	3	0	3	3
4	3	0	8	0	5	2	1	7	3
5	10	0	1	0	8	4	0	1	3
MV	4.16	2.21	2.42	1.11	4.00	2.95	1.32	3.05	2.79
ME	5	2	2	1	4	2	1	3	2
STDEV	1.015	0.787	1.805	0.567	1.100	1.311	1.057	1.026	1.357

[a] Evaluation criteria SI, UP, etc., are explained in Table 1. i = criterion, j = enumerates the score values, MV = mean value, ME = median, and STDEV = standard deviation.

3. ANALYSIS OF EVALUATION CRITERIA

It can be taken from Table 4 that the various evaluation criteria have different significance. The Hasse diagram allows one to derive how these criteria cause the objects to be positioned. Characteristic positions are those of maximum, minimum, and isolated objects. Other striking features are chains (like the sequence PES < CAN < NIO) or antichains (like CR3, DIN, CHB, SAH). The structure of a Hasse diagram depends not only on the set of objects (group of datasources) and on the number of evaluation criteria but also on the evaluation criteria themselves. Regarding these evaluation criteria, the distribution of scores is of great relevance. Therefore, mean values (MVs), medians (MEs), and standard deviations (STDEVs) were calculated for the 10 groups of datasources given in Table 1.

The group of CD-ROMs already discussed (2CD see Table 1) is chosen as an example to interpret these statistical parameters MV, ME, and STDEV. Table 5 gives the number n_{ij} of realizations of the ith criterion within the set of datasources of the jth score value and the parameters of the descriptive statistics of the 19 CD-ROMs with nine evaluation criteria on six possible scores. The aim is to figure out which criterion is a well-distributed one, which is not, and which holds a middle position. It is supposed that if the standard deviation has a high value in comparison with other evaluation criteria, then this specific criterion, i, can be regarded as well-distributed. In analogy to this, low values are regarded as poorly distributed.

Examples of how to read Table 5 might be useful: Selecting the criterion, frequency of updating (i = UP), and score 3 (j = 3), Table 5 shows that this score 3 is given eight times (n_{ij} = 8). In a similar manner it can be demonstrated that the score 4 (j = 4) is given zero times (n_{ij} = 0) for the criterion availability on other media (i = AV).

The evaluation criterion, quality of CD-ROM (QU), can be regarded as positive (which means it has a higher significance than the other criteria) because the STDEV of QU is high. The criteria, cost of the CD-ROM (CO), and identification parameter for chemicals (ID) are also evaluated rather positively. It is a common misconception that databases on CD-ROM are only available at great expense. In our approach it is shown that the criterion, cost of the CD-ROMs, takes a high (positive) position regarding the distribution of scores.

Table 6. Comparative Evaluation of 10 Groups of Datasources[a]

	ED							
group	SI	UP	CO	AV	NU	ID	IP	QU
1ON	1	0	1	1	2	0	1	0
1CD	1	0	1	1	1	0	2	0
2ON	2	1	2	0	2	0	1	2
2CD	1	0	2	0	1	2	1	2
3ON	1	0	1	1	2	0	1	0
3CD	2	1	2	2	2	0	2	0
4ON	2	1	2	0	2	0	1	2
4CD	2	0	2	0	2	1	2	2
5ON	1	0	1	1	2	0	0	0
5CD	1	0	2	1	1	0	2	1

[a] EC = evaluation criterion; 1ON, 1CD = explanations given in Table 1; 2 = good distribution of criterion, 1 = medium, and 0 = bad.

ID has the following span of scores {1, 2, 3, 4, 5}. None of the databases obtains the score 0. This implies that all the CD-ROMs considered here have at least, apart from the chemical name and synonyma, a CAS Registry number. Many CD-ROMs offer the facility to search for molecular structure. It has to be mentioned here that only this described group (2CD) out of the chosen 10 groups of chemical and environmental datasources has a positive situation regarding identification parameters as mentioned above. All the other groups (1ON−5ON, and 1CD, 3CD−5CD) considered here are clearly inferior.

The criteria UP (frequency of updating of datasource) and AV (availability on other media) have to be regarded as poor with respect to the distribution of scores expressed by the low standard deviation. The criterion UP has, for example, the range of scores {1, 2, 3}. The CD-ROMs chosen for this group do only exist as a CD-ROM version and are not available online or on the Internet. That is the reason why the data situation concerning the criterion AV (availability on other media) has also to be regarded as a poor one.

The criteria US (use of chemicals), SI (size of the datasource), NU (number of chemical substances), and IP (information parameters for chemical substances) take a middle position in this evaluation scheme.

To sum up this section, one can state that CO, ID, and QU have a good, UP and AV a bad and SI, NU, IP, and US a middle position regarding the distribution of scores for this set of objects. A bad evaluation also means that the criterion itself and the assignment of scores should be examined. In a further step the positions explained above lead to the assignments of the following numbers: 2 for a high, 1 for a middle, and 0 for a low position.

4. COMPARATIVE EVALUATION OF 10 GROUPS OF DATA SOURCES

As a next step the 10 groups of datasources (see Table 6) will be compared with the help of the already introduced Hasse diagram technique. This approach examines the distribution of the evaluation criteria. As demonstrated in section 3, three possibilities are considered, namely, good (high position), medium (middle position), and bad (low position). Following the scoring ideas, we assigned number 2 (or score 2) for a good distribution, 1 a medium, and 0 a bad one.

We only took into account eight criteria (SI, UP, CO, AV, NU, ID, IP, QU) for the purpose of comparing all 10 sets of objects (1ON−5ON, 1CD−5CD). The criterion US was omitted because it was not available for all groups of datasources. For further explanations see Voigt.[4]

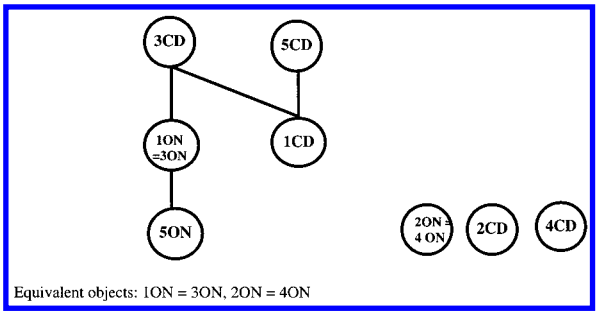

Figure 3. Hasse diagram for 10 groups and 8 evaluation criteria.

The resulting Hasse diagram is shown in Figure 3.

This Hasse diagram comprises three levels, defined by the longest chain, namely nontrivial equivalence classes. 5ON ≤ 1ON ≤ 3CD. The groups 1ON/3ON and 2ON/4ON are so-called nontrivial equivalence classes. These are objects with identical scores for the given set of criteria. The groups 2CD and 4CD and the equivalence class 2ON/4ON are *isolated objects* (see section 2.5.2). These are numeric chemical CD-ROMs (2CD), numeric environmental CD-ROMs (4CD), numeric chemical online databases (2ON), and numeric environmental online databases (4ON). These groups are not comparable with each other if all criteria are considered at the same time. The isolated objects are, for example, scored highly in the criteria, CO (cost of 1 h of online searching in online databases or cost of the CD-ROM), and in QU (quality of datasource). On the other hand, for these isolated objects the criterion AV, availability on other media, is bad.

Maximum objects are 3CD (bibliographic environmental CD-ROMs) and 5CD (environmental chemical CD-ROMs). These two groups seem to be most appropriate for covering our subject "chemistry and environment" and for providing information on these fields of interest. The group 3CD comprises 42 bibliographic environmental CD-ROMs. It received the best score 2 five times out of a possible eight times. That means that five criteria, namely, SI, CO, AV, NU, and IP, are well-defined according to our approach. This group 3CD received the score 0 for the criteria ID and QU, whereas UP is in the middle position (score = 1).

Minimum objects are 5ON (environmental chemical online databases) and 1CD (bibliographic chemical CD-ROMs), which turn out to be the less suitable groups for this type of evaluation approach.

5. CONCLUSIONS AND PROSPECTS

The presented evaluation approach of datasources for chemistry and environmental sciences gives an indication of good and bad datasources and also of the importance of different evaluation criteria. Furthermore, the Hasse diagram of Figure 3 indicates a vague trend: CD-ROMs in the field of chemistry and environmental sciences are slightly better than online databases.

Of utmost importance is the incorporation of Internet resources in this kind of evaluation procedure. First steps in this direction have already been made in establishing another metadatabase for environmental chemicals, DAIN, Metada-

tabase of Internet Resources for Environmental Chemicals, which can be found under the following URL: http://dino.wiz.uni-kassel.de/dain/.[20] This metadatabase is being elaborated in cooperation with the University of Kassel. A first approach for expanding the evaluation process to Internet resources has already been presented.[21] Furthermore, different sets of objects, e.g., the chemically extremely relevant reaction databases or databases which focus on specific chemicals, such as pesticides, should be regarded as separate sets of datasources.

Different evaluation criteria could be taken into account as well as a different scoring scheme. All these variations, however, depend to a large extent on expert knowledge and expert judgment. There is an urgent need to formalize expert knowledge for example by more sophisticated elements of the theory of partial ordering up to formal concept analysis.[22]

The presented approach for the evaluation of datasources as well as of groups of datasources in the field of environmental chemicals can be regarded as a step from information management (setting-up of metadatabases) to knowledge management (evaluation of the contents of these metadatabases). In the future, this direction will be followed and extended by applying partial order set theory and concepts of multivariate statistics and comparing the results.

ACKNOWLEDGMENT

We wish to thank Walter Gruhn, Andrew Fyson, and Hannelore Guth for fruitful discussions concerning the style of this paper. This work was partly supported by the Bayerisches Staatsministerium für Landesentwicklung und Umweltfragen (Bavarian State Ministry for State Development and Environmental Affairs) in Germany.

Supporting Information Available: Figures showing Hasse diagrams for all 10 groups of datasources listed in Table 1. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Luckenbach, R. Past Perfect, Present Perfect, Future Perfect—Quality Assessment and Quality Control Mechanisms at Beilstein. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 923−929.

(2) Voigt, K.; Matthies, M.; Pepping, T. Information System for Environmental Chemicals: Comparison and Evaluation of Metadatabase of Datasources. *Toxicol. Environ. Chem.* **1993**, *40*, 83−92.

(3) Voigt, K.; Brüggemann, R. Toxicology Databases in the Metadatabank of Online Databases. *Toxicology* **1995**, *100*, 225−240.

(4) Voigt, K. *Erstellung von Metadatenbanken zu Umweltchemikalien und vergleichende Bewertung von Online Datenbanken und CD-ROMs* http://vermeer.organik.unierlangen.de/dissertationen/data/dissertation/Kristina_Voigt/html/und Shaker Verlag: Aachen, 1997; p 1-209.

(5) Hushon, J. M.; Kornreich, M. R. In *Hazard Assessment of Chemicals, Current Developments;* Saxena, J., Ed.; Academic Press: Orlando, FL, 1984; pp 63−107.

(6) Landner, L. In *Chemicals in the Aquatic Environment;* Landner, L., Ed.; Springer-Verlag: Heidelberg, 1989; pp 59−72.

(7) *GDCh, Gesellschaft Deutscher Chemiker, Beratergremium für umweltrelevante Altstoffe*; *BUA-Stoffberichte* Hirzel Wissenschaftliche Verlagsgesellschaft: Stuttgart, 1995; pp 1−5.

(8) Swanson, M. B.; Davis, G. A.; Kincaid, L. E.; Schultz, T. W.; Bartness, J. E.; Jones, J. E.; Georges, E. L. A Screening Method for Ranking and Scoring Chemicals by Potential Human and Environmental Impacts. *Environ. Toxicol. Chem.* **1997**, *16*, 372−383.

(9) Russom, C. L.; Bradbury, S. P.; Carlson, A. R. Use of Knowledge Base and QSARs to Estimate the Relative Ecological Risk of Agrochemicals: A Problem Formulation Exercise. *SAR QSAR Environ. Res.* **1995**, *4*, 83−95.

(10) Belke, W.; Graichen, D.; Starruss, M. *Nichtmetrische Klassifizierung von Informationen, Theorie und Anwendung;* Akademie-Verlag: Berlin, 1979; pp 1−6.

(11) Bock, H. H. *Automatische Klassifikation;* Vandenhoeck & Ruprecht: Göttingen, Germany, 1974.

(12) Brüggemann, R.; Bartel, H.-G. A Theoretical Concept To Rank Environmentally Significant Chemicals. *J. Chem. Inf. Comput. Sci.* **1999a**, *39*, 211−217.

(13) Brüggemann, R.; Bücherl, C.; Pudenz, S.; Steinberg, C. Application of the Concept of Partial Order on Comparative Evaluation of Environmental Chemicals. *Acta Hydrochim. Hydrobiol.* **1999b**, *27*, 170−178.

(14) Voigt, K.; Brüggemann, R. Evaluation Criteria for Environmental and Chemical Databases. *Online CD-ROM Rev.* **1998**, *22*, 247−262.

(15) Brüggemann, R.; Voigt, K. An Evaluation of Online Databases by Methods of Lattice Theory. *Chemosphere* **1995**, *31*, 3585−3594.

(16) Brüggemann, R.; Voigt, K.; Steinberg, C. E. W. Application of Formal Concept Analysis to Evaluate Environmental Databases. *Chemosphere* **1997**, *35*, 479−486.

(17) Hayward, J. In *Online Information 95, 19th International Online Information Meeting, Proceedings;* Raitt, D. I., Jeapes, B., Eds.; Learned Information Ltd.: Oxford, U.K., 1995; pp 143−155.

(18) Brüggemann, R.; Halfon, E. Comparative Analysis of Nearshore Contaminated Sites in Lake Ontario: Ranking for Environmental Hazard. *J. Environ. Sci. Health* **1997**, *A32*, 277−292.

(19) Warr, W. A. New Chemical Databases on CD-ROM. *Database* **1993** (Feb), 59−67.

(20) Benz, J.; Voigt, K. In *Online Information 95, 19th International Online Information Meeting, Proceedings;* Raitt, D.I ., Jeapes, B., Eds.; Learned Information Ltd.: Oxford, U.K., 1995; pp 455−466.

(21) Voigt, K.; Benz, J.; Brüggemann, R. In *Online Information 96, 20th International Online Information Meeting, Proceedings;* Raitt, D. I.; Jeapes, B., Eds.; Learned Information Ltd.: Oxford, U.K., 1996; pp 151−160.

(22) Brüggemann, R.; Voigt, K.; Steinberg, C. F. W. Application of Formal Concept Analysis to Evaluate Environmental Databases. *Chemosphere* **1997,** *35*, 479−486.

CI9900837