Datenanalyse von umweltrelevanten Metadatenbanken

Kristina Voigt[1], Gerhard Welzl und Gerda Rediske

Abstract

Metadatenbanken nehmen als Werkzeug zur Unterstützung der Auffindung geeigneter Primärquellen im Umweltschutz eine immer größere Bedeutung ein. Seit einigen Jahren erarbeitet die GSF - Forschungszentrum für Umwelt und Gesundheit drei Metadatenbanken auf dem Gebiet des Umweltschutzes im Zusammenhang mit Chemikalien. Es handelt sich dabei um DADB - Metadatenbank der Online Datenbanken, DACD – Metadatenbank der CD-ROMs, DAIN - Metadatenbank der Internet Ressourcen. Im Vordergrund dieses Beitrages steht die Datenanalyse der Inhalte der Metadatenbanken mittels Methoden aus dem Gebiet der explorativen multivariaten Statistik. Es werden 50 relevante Deskriptoren (Objekte) für die Datenanalyse ausgewählt. Da die drei Metadatenbanken DADB, DACD und DAIN betrachtet werden sollen, liegt eine Datenmatrix 50 x 3 vor. Dabei wird die Datenmatrix derart umsortiert, daß bezüglich der Medienverfügbarkeit "ähnliche" Deskriptoren nebeneinander zu liegen kommen. Durch Minimierung der Summe alle "Abstände" zwischen benachbarten Deskriptoren entsteht eine Datenmatrix mit homogener Struktur. Dabei kommen Algorithmen zur Lösung des sog. Traveling Salesman Problem zur Anwendung. Es wird u.a. deutlich, daß Internet Ressourcen bei den meisten Umwelt-relevanten Deskriptoren geringere Datenverfügbarkeit aufweisen als Online Datenbanken und CD-ROMs.

1 Metadatenbanken für Umweltchemikalien

Die Forschung zu globalen Umweltveränderungen stellt aufgrund der notwendigen Interdisziplinarität besonders hohe Anforderungen an ein nachfrageorientiertes Informationsmanagement. Dabei steht der Aspekt der Unterstützung der Suche nach und dem Zugriff auf umweltforschungsrelevante Datenbestände in einem stark interdisziplinären Umfeld im Vordergrund (BMBF 1997). Es gibt in Deutschland Informationssysteme, die wissenschaftliche Nutzer mit Hilfe von Metadaten (Daten über Da-

[1] GSF - Forschungszentrum für Umwelt und Gesundheit, Institut für Biomathematik und Biometrie, Ingolstädter Landstr. 1, D-85764 Neuherberg
email: kvoigt@gsf.de, Internet: http://www.gsf.de/institute/ibb/voigt

ten) bzw. Datenkatalogen bei der Suche nach potentiell relevanten Datenbeständen unterstützen. Eine Zusammenfassung wichtiger Metadatenbanken auf internationaler und nationaler Ebene gibt Oliver Günther in dem Buch "Environmental Information Systems" (Günther 1998). Dort werden auch die in der GSF-Forschungszentrum für Umwelt und Gesundheit im Auftrag des Bayerischen Staatsministeriums für Landesentwicklung und Umweltfragen entwickelten Metadatenbanken für Umweltchemikalien betrachtet. Es handelt sich um folgende drei Metadatenbanken für Umweltschutz und Chemikalien: DADB - Metadatenbank der Online Datenbanken, DACD - Metadatenbank der CD-ROMs, DAIN - Metadatenbank der Internet Ressourcen.

Auf den Workshops "Hypermedia im Umweltschutz" wurde die Metadatenbank DAIN bereits im Jahr 1998 (Voigt/Benz 1998a) und 1999 (Voigt/Benz 1999a) vorgestellt. Die Metadatenbank der Internet Ressourcen für Umweltchemikalien DAIN ist unter der URL http://dino.wiz.uni-kassel.de/dain verfügbar (Dokumentenstand: 510 am 07.05.99).

Von großem Interesse ist es für die effektive Nutzung der Metadatenbanken nun, welche Deskriptoren (Umwelt-relevante bzw. Chemikalien-relevante Parameter) in welchen Medien bevorzugt anzutreffen sind und ob es Deskriptoren gibt, die überwiegend in einem bestimmten Medium vorkommen. Aus diesem Grund wird eine statistische Auswertung der Inhalte der Metadatenbanken vorgenommen.

Im folgenden soll die Datenanalyse der Inhalte der drei Metadatenbanken im Vordergrund stehen. Durch eine geeignete Analysemethode soll die Auswertung der Metadatenbanken derart gestaltet werden, daß Bereiche von "ähnlichen" Deskriptoren mit "ähnlicher" Verfügbarkeit in den Medien erkennbar werden.

2 Methoden in der Ökometrie, Biometrie, Chemometrie

Auf dem Gebiet der Chemie und der Umweltwissenschaften haben sich die Begriffe Chemometrie und Ökometrie etabliert. Chemometrie wird verstanden als eine fachliche Disziplin in der Chemie, die statistische oder andere mathematische Methoden verwendet, um ein Maximum an chemischer Information durch die Analyse der chemischen Daten zu liefern. Oftmals wird auch der Begriff "Ökometrie" verwendet, für die Anwendung von chemometrischen Methoden im Bereich der Umweltwissenschaften (Einax 1997). Analog werden unter dem Begriff der Biometrie die vielfältigen Anwendungen in der Mathematik, insbesondere der mathematischen Statistik, in den biologischen und ihren verwandten Wissenschaften z.B. Agrarwissenschaften verstanden (Lorenz 1996). In diesem Beitrag wenden wir mathematisch statistische Methoden auf die Inhalte von umweltrelevante Metadatenbanken an. Dementsprechend könnte man in diesem Zusammenhang von "Database-metrics" sprechen (Voigt/Welzl 1999b).

Um Beziehungen in komplexen Datensätzen aufzudecken, werden Techniken (statistische Methoden und Visualisierungsmethoden) angewandt und weiterentwickelt, die auch als "data mining" bezeichnet werden. Ziel ist es, unbekannte bzw. unerkannte Strukturen aufzudecken.

Ausgangspunkt ist eine Matrix mit den Werten von K Variablen für N Objekte. Ziel der Analyse ist es, eine Matrix mit einer möglichst homogenen Struktur zu erhalten. Dazu werden in der Ausgangsmatrix Permutationen bezüglich der Reihen und Spalten vorgenommen. Eine sogenannte Purity-Funktion ist definiert als Maß für die Homogenität einer Matrix. Mittels der Purity-Funktion wird der Eintrag einer Zelle der Matrix mit benachbarten Einträgen verglichen. Der Wert einer Purity-Funktion ist daher um so größer, je ähnlicher sich benachbarte Objekte und Variable sind. Häufig wird die Optimierung der Purity-Funktion getrennt für Objekte und Variable durchgeführt. Auf der Basis von Abständen zwischen zwei Objekten (Deskriptoren von Inhalten von Umweltdatenbanken) kann dann die Purity-Funktion dadurch maximiert werden, daß diejenige Anordnung mit der minimalen Summe aller paarweisen Abstände gesucht wird. Dieses Optimierungsproblem ist für größere N nicht exakt lösbar; es gibt N!/2 mögliche Anordnungen. Zur Lösung dieses Problems - als Traveling Salesmen Problem bekannt - gibt es aber verschiedene approximative Lösungsalgorithmen, z.B. die Methode des simulated annealing (Lawler 1985). Bei unserer Auswertung kam das in Press et al. beschriebene Programm zur Anwendung (Press 1986). Je nach Fragestellung sind unterschiedliche Abstandsmaße für Objekte sinnvoll, z.B. Summe der Abstandsquadrate oder Manhattan-Abstand. In dieser Analyse wird ein Abstandsmaß basierend auf Rangkorrelation bevorzugt.

3 Auswertung der Metadatenbanken bezüglich ihrer Inhalte

3.1 Datenmatrix (50 Objekte, 3 Variable)

Die Struktur und die Inhalte der Metadatenbanken sind ausführlich in mehreren Publikationen beschrieben (Voigt 1998 b,c). Die drei zu analysierenden Metadatenbanken sind in ihren inhaltlichen Datenfeldern einheitlich strukturiert. Von besonderer Bedeutung ist das sog. Deskriptorenfeld, welches den umwelt- und chemierelevanten Inhalt der zu beschreibenden Original Datenbank charakterisiert. Hier sind Themengebiete aus den Bereichen Identifikationsmerkmale von Chemikalien (CAS-Nummer, Strukturformel, Summenformel, Molekulargewicht etc.), Produktions- und Verwendungsdaten, Daten zum Vorkommen von Chemikalien in den Umweltmedien, physikalisch-chemische Eigenschaften, Abbau- und Akkumulationsdaten, ökotoxikologische Daten, toxikologischen Daten, Carcinogenität, Mutagenität, Teratogenität etc., Daten zum Arbeitsschutz, Gesetze und Verordnungen etc. aufgeführt. In Tabelle 1 sind insgesamt 50 Deskriptoren (in englischer Terminologie, die der Metadatenbank DAIN zugrunde liegt) aufgelistet.

Nr.	Deskriptor	Abk.	DADB	DACD	DAIN	DADB %	DACD %	DAIN %
1.	abiotic degradation	ABD	234	131	41	48,8	34,7	10,9
2.	acute oral toxicity	ACT	226	205	141	47,1	54,2	37,6
3.	detection in air	AIR	252	167	41	52,5	44,2	10,9
4.	algae toxicity	ALT	123	122	48	25,6	32,3	12,8
5.	bioaccumulation	BIA	179	130	58	37,3	34,4	15,5
6.	biodegradation	BID	225	138	61	46,9	36,5	16,3
7.	CAS-number index	CAI	188	180	69	39,2	47,6	18,4
8.	chemical name	CAN	453	370	311	94,4	97,9	82,9
9.	carcinogenicity	CAR	223	176	156	46,5	46,6	41,6
10.	CAS-number	CAS	189	181	211	39,4	47,9	56,3
11.	chronic mammalian toxicity	CHT	168	141	118	35,0	37,3	31,5
12.	chemical name index	CNI	452	362	187	94,2	95,8	49,9
13.	daphnia toxicity	DAT	110	111	55	22,9	29,4	14,7
14.	dermal toxicity	DET	104	106	133	21,7	28,0	35,5
15.	distribution coefficients	DIC	161	118	7	33,5	31,2	1,9
16.	detection in drinking water	DRW	237	119	22	49,4	31,5	5,9
17.	eye irritation	EYI	136	114	135	28,3	30,2	36,0
18.	fish toxicity	FIT	171	141	61	35,6	37,3	16,3
19.	flash point	FLP	154	148	123	32,1	39,2	32,8
20.	detection in food	FOO	175	142	23	36,5	37,6	6,1
21.	handling and storage	HAS	218	125	100	45,4	33,1	26,7
22.	inhalative toxicity	INT	95	56	138	19,8	14,8	36,8
23.	melting point	MEP	239	246	154	49,8	65,1	41,1
24.	molecular formula index	MFI	72	59	17	15,0	15,6	4,5
25.	molecular formula	MOF	86	95	104	17,9	25,1	27,7
26.	molecular weight index	MOI	35	20	8	7,3	5,3	2,1
27.	molecular weight	MOW	269	176	61	56,0	46,6	16,3
28.	mutagenicity	MUT	235	180	150	49,0	47,6	40,0
29.	neurotoxicity	NEU	66	111	29	13,8	29,4	7,7
30.	photodegradation	PHD	166	126	46	34,6	33,3	12,3
31.	production volume	PRV	115	43	22	24,0	11,4	5,9
32.	reproductive toxicity	RET	190	139	133	39,6	36,8	35,5
33.	R-S-sentences	RSS	35	59	53	7,3	15,6	14,1
34.	detection in sediments	SED	164	88	13	34,2	23,3	3,5
35.	detection in sewage sludge	SEW	183	126	5	38,1	33,3	1,3
36.	skin irritation	SKI	141	115	138	29,4	30,4	36,8
37.	detection in soil	SOI	169	152	22	35,2	40,2	5,9
38.	structural formula	STF	33	57	29	6,9	15,1	7,7
39.	structural formula index	STI	16	25	3	3,3	6,6	0,8

40.	subacute toxicity	SUT	121	89	12	25,2	23,5	3,2
41.	synonyma index	SYI	400	362	258	83,3	95,8	68,8
42.	synonyma	SYN	446	367	305	92,9	97,1	81,3
43.	teratogenicity	TER	190	164	149	39,6	43,4	39,7
44.	TLV-values	TLV	63	79	31	13,1	20,9	8,3
45.	transportation	TRS	171	114	50	35,6	30,2	13,3
46.	use of chemical substance	USE	179	151	166	37,3	39,9	44,3
47.	vapor pressure	VAP	215	159	138	44,8	42,1	36,8
48.	waste disposal	WAD	223	141	95	46,5	37,3	25,3
49.	water solubility	WAS	212	194	42	44,2	51,3	11,2
50.	detection in water	WAT	241	164	35	50,2	43,4	9,3

Tabelle 1
Datenmatrix 50 Deskriptoren und 3 Medien

Folgende Dokumentenstände lagen der Analyse zugrunde: DADB: 480, DACD: 378, DAIN: 475. Es werden sowohl die absoluten Zahlen als auch die prozentualen Anteile in den jeweiligen Metadatenbanken aufgeführt. Die Prozentzahlen bilden die Grundlage der weiteren Datenanalyse.

3.2 Datenanalyse im Umweltschutz

Eine graphische Auftragung der Häufigkeitsverteilung der einzelnen Deskriptoren in der Reihung der Tabelle 1 läßt nur sehr generelle Aussagen zu, die bereits in den Proceedings des 20th Annual National Online Meeting in New York publiziert wurden (Voigt/Welzl 1999b). In den meisten Deskriptoren ist das Internet den beiden "klassischen" Medien online Datenbanken und CD-ROMs unterlegen. Online Datenbanken und CD-ROMs liegen in der Datensituation dicht beieinander und weisen -bis auf wenige Ausnahmen- eine höhere Informationsfülle als Internet Datenbanken aus. Um den Datensatz besser und schärfer zu analysieren, kommen im folgenden Methoden der explorativen multivariaten Statistik zur Anwendung.

Die multivariate ("mehrdimensionale") Statistik ist ein Hilfsmittel zur Lösung von Fragestellungen, bei denen mehrere Variable simultan betrachtet werden sollen. Im Umweltbereich, wo häufig Probleme vorliegen, bei denen große Datensätze mit vielen Objekten und mehreren Variablen ausgewertet werden müssen, kommen also multivariate statistische Methoden oft zur Anwendung. Eine umfassende Beschreibung von statistischen Methoden in Umweltwissenschaften ist im Buch von Stoyan et al. "Umweltstatistik" zu finden (Stoyan 1997). Die Visualisierung von mit Schwermetallen belasteten Regionen in Baden-Württemberg mit Hilfe von explorativen statistischen Verfahren wurden bereits von Welzl et al. durchgeführt (Welzl/ Voigt 1998, 1999).

3.3 Datenanalyse der Inhalte der Metadatenbanken: relative Werte

Im folgenden soll die Datenanalyse der Inhalte der drei Metadatenbanken im Vordergrund stehen. Wie schon unter 3.2. erwähnt, ist die absolute Verfügbarkeit der meisten Deskriptoren im Internet geringer als in online Datenbanken und CD-ROMs. Besonderer Augenmerk sollte bei der Analyse darauf gerichtet werden, wo die Internet Ressourcen ihre inhaltlichen Schwerpunkte legen. Von Interesse ist es daher, die **relative** Verfügbarkeit der einzelnen Deskriptoren innerhalb der einzelnen Medien zu betrachten. Daher wird als Abstand zwischen den Deskriptoren ein Maß basierend auf Rängen ähnlich dem Spearman Rangkorrelationskoeffizienten herangezogen. Ziel ist es, die Datenmatrix derart zu sortieren, daß "ähnliche" Deskriptoren - in Bezug auf ihre relative Verfügbarkeit in den Medien online, CD-ROM und Internet - nebeneinander stehen. Dazu wird der in Kapitel 2 beschriebene Algorithmus angewandt. Mit der Methode des simulated annealing wird die optimale Anordnung der Deskriptoren bestimmt.

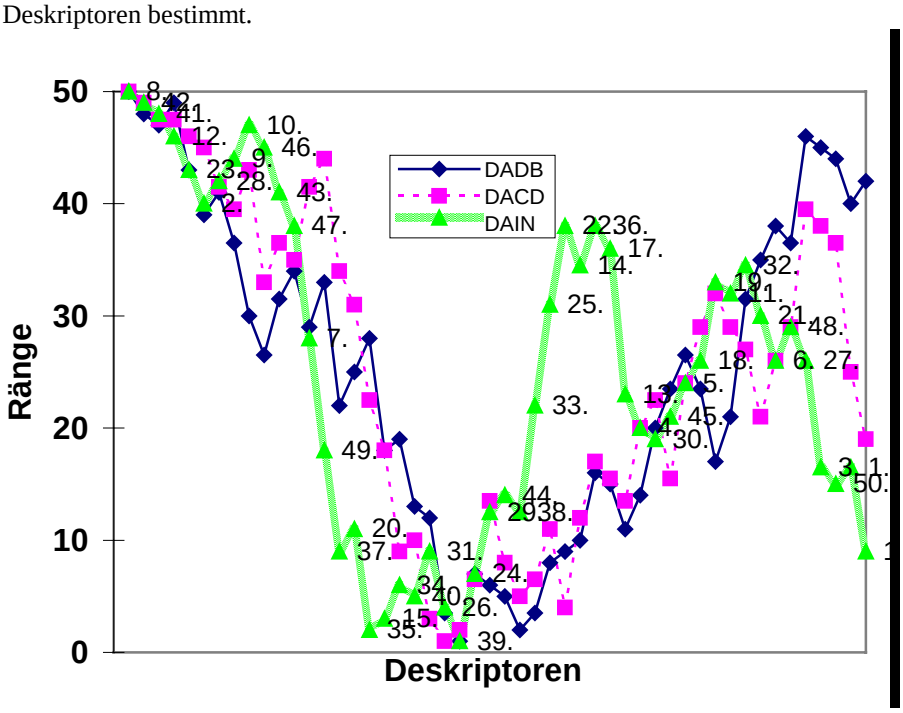

Abbildung 1
Standardisierung der Datenmatrix 50 x 3

Als Ergebnis ergibt sich die in Abbildung 1 dargestellte Anordnung der Deskriptoren. Die Deskriptoren sind dabei durch ihre Rangzahlen bezüglich der einzelnen Medien charakterisiert.

Aus der Abbildung 1 können mehrere Bereiche abgelesen werden, die nochmals in Tabelle 2 aufgeführt werden. Ebenso wird die generelle Datensituation deutlich, nämlich daß die "klassischen" Medien in den meisten Deskriptoren eine bessere Datensituation aufweisen als das Internet. Die einzelnen Bereiche werden im folgenden diskutiert. Die Abbildung wird von links nach rechts analysiert.

Es zeigt sich, daß es Deskriptoren gibt (siehe **Gruppe 1**), die in allen 3 Medien auf hohen Rängen (um den Rangwert 50) vertreten sind. Es handelt sich hierbei um die grundlegenden Identifikationsmerkmale für Chemikalien wie Chemikalienname (CAN Nr. 8), Chemikalien Namenindex (CNI Nr. 12), Synonyma Index (SYI Nr. 41) und Synonyma (SYN Nr. 42). Ohne diese Parameter ist keine Recherche nach Chemikalien in Datenbanken durchzuführen.

Gruppe	Art	Deskriptoren Nr.
1.	IN: h, ON: h, CD: h (~ 50)	8, 12, 41, 42
2.	IN: h, ON: h, CD: h (30-50	2, 9, 10, 23, 28, 43, 46, 47
3.	IN: l (5-25), ON, CD: h (25-45)	1, 3, 7, 16, 27, 37, 48, 50
4.	IN: l (5-10), ON, CD: m (10-20)	15, 20, 31, 34, 35, 37
5.	IN: l, ON: l, CD: l	24, 26, 39, 40
6.	IN: h (>30), ON, CD: l (5-15)	13, 14, 17, 19, 22, 25, 29, 36, 38, 44
7	IN: l (25-30), ON, CD: m (25-30) ON/IN schwankend	5, 6, 18, 19, 21, 30, 32, 45

Tabelle 2
Bereiche von Deskriptoren in der sortierten Datenmatrix[2]

Die **2. Gruppe** umfaßt jene Objekte, die geringere Ränge als die 1. Gruppe aufweisen, jedoch immer noch auf Rängen von 30-50 in allen 3 Medien liegen. Es handelt sich hierbei um die folgenden Deskriptoren: akute orale Toxizität (ACT Nr. 2), Carcinogenität (CAR Nr. 9), CAS-Nummer (CAS Nr. 10), Schmelzpunkt (MEP Nr. 23), Mutagenität (MUT Nr. 28), Teratogenität (TER Nr. 43), Verwendungszweck (USE Nr. 46) und Dampfdruck (VAP Nr. 47). Es handelt sich hierbei um wichtige Parameter aus der Toxikologie sowie um herkömmliche physikalisch-chemische Eigenschaften, die in allen Medien häufig zu finden sind. Ebenso befindet sich der bedeutende Identifikationsparameter für chemische Substanzen, die CAS-Registry-Number, auch in dieser Gruppe. Hier ist anzumerken, daß es bei weitem zur Recher-

[2] Erläuterung der Abkürzungen: IN = Internet, ON = online Datenbanken, CD = CD-ROMs; l = low (gering), M = medium (mittel), h = high (hoch), Zahlenangaben beziehen sich auf die Ränge.

che nach Chemikalien nicht ausreicht, in quasi jeder zweiten Datenbank nur mit der CAS-Nummer recherchieren zu können. Eine geringfügige Verbesserung der CAS-Nummer Situation im Internet gegenüber den beiden anderen Medien kommt in der Graphik ebenfalls zum Ausdruck.

Den **3. Bereich** bilden die Parameter, die eine geringe Belegung im Internet aufweisen, dafür aber Ränge von 25-45 in den beiden klassischen Medien haben. Es handelt sich um: Abiotischer Abbau (ABD Nr. 1), Vorkommen in Luft (AIR Nr. 3), CAS-Nummern Index (CAI Nr. 7), Vorkommen im Trinkwasser (DRW Nr. 16), Molekulargewicht (MOW Nr. 27), Vorkommen im Boden (SOI Nr. 37), Abfallaufkommen (WAD Nr. 48) und Vorkommen in Wasser (WAT Nr. 50) haben. Es handelt sich um die meisten Umweltbelastungsparameter und um einige Identifikationsmerkmale. Der Bereich ist sowohl als dritter Bereich von links als auch als äußerst rechter Bereich auf der Abbildung erkennbar.

Den **4. Bereich** stellt eine Gruppe an Deskriptoren dar, bei denen es eine sehr geringe Internet-Belegung (Ränge 5-10) gibt und die online und CD-ROM Repräsentanz sich im Bereich der Ränge 10-20 bewegt. Es handelt sich um die Objekte Verteilungskoeffizienten (DIC Nr. 15), Vorkommen in Lebensmitteln (FOO, Nr. 20), Produktionsvolumen (PRV Nr. 31), Vorkommen im Sediment (SED Nr. 34), Vorkommen im Klärschlamm (SEW Nr. 35) und Vorkommen im Boden (SOI Nr. 37).

Es schließt sich der **5. Bereich** an, der durch die geringe Belegung in allen 3 Medien repräsentiert wird. Dieser Bereich findet sich in der Mitte im unteren Bereich des Diagramms. Zu dieser Gruppe gehören die Deskriptoren Summelformel Index (MFI Nr. 24), Molekulargewicht Index (MOI Nr. 26) und Strukturformel Index (STI Nr. 39) und subakute Toxizität (SUT Nr. 40). Das bedeutet, daß die wichtigsten Identifikationsmerkmale für Chemikalien in extrem geringem Umfang in umwelt- und chemikalienrelevanten Datenbanken vorhanden sind. Diese Tatsache wurde bereits von Voigt 1997 anhand einer Auswertung mittels verbandstheoretischer Methoden belegt (Voigt 1997).

Im **6. Bereich** ist nun eine höhere Internet-Belegung als in den anderen beiden Medien zu beobachten. Das Internet ist dabei mit hohen Rängen (30-40) und die anderen beiden Medien mit Anteilen <20 belegt. Zu dieser Gruppe gehören Daphnientoxizität (DAT Nr. 13), dermale Toxizität (DET Nr. 14), Augenreizung (EYI Nr. 17), inhalative Toxizität (INT Nr. 22), Summenformel (MOF Nr. 25), Neurotoxizität (NEU Nr. 29), Hautreizung (SKI Nr. 36), Strukturformel (STF Nr. 38) und TLV-Werte (TLV Nr. 44). Es handelt sich bei diesen Deskriptoren um einige grundlegende Toxizitätsparameter, die z.B. Bestandteil von Sicherheitsdatenblättern sind. Solche Sicherheitsdatenblätter und andere gesundheitsspezifische Informationen findet man häufig im Internet. Anzumerken ist, daß nach der Auswertung der Rangfolge die Strukturformel im Medium Internet eine deutlich bessere Stellung einnimmt als in den klassischen Medien. Das bedeutet, daß die Möglichkeiten des Internets Graphiken darzustellen hier auch bei der Strukturformel für Chemikalien zum Tragen kommen.

Der **7. Bereich** wird repräsentiert durch Deskriptoren, die eine Internet-Belegung auf Rängen 25-30 aufweisen und deren Ränge bei 20-30 in Online Datenbanken und CD-ROMs liegen. Die beiden klassischen Medien variieren in dieser Gruppe jedoch stark. Das trifft zu auf die Deskriptoren Bioakkumulation (BIA Nr. 5), Bioabbau (BID Nr. 6), Fischtoxizität (FIT Nr. 18), Flammpunkt (FLP Nr. 19), Lagerung (HAS Nr. 21), Photoabbau (PHD Nr. 30), Reproduktionstoxizität (RET Nr. 32) und Transport von Chemikalien (TRS Nr. 45). Es handelt sich hierbei um umweltspezifische Parameter aus der Ökotoxikologie. Es wird deutlich, daß diese bedeutenden Umweltdeskriptoren im Internet in geringem Umfange vorkommen. Die Datenlage in den "konventionellen" Medien ist jedoch ebenfalls alles andere als zufriedenstellend.

Es werden nur die Hauptbereiche hier diskutiert. Einige wenige Objekte können nicht direkt einem Bereich zugeordnet werden.

4 Diskussion der Ergebnisse

Die Hauptinformationsquelle ist heute das Internet. Im Mai 1999 waren 171.25 Million online Nutzer, davon 8.4 Millionen in Deutschland (NUA 1999). Da die Informationsinhalte im Internet stetig weiter ansteigen, wird z.Zt. enormer Aufwand unternommen, die Suchmaschinen auszubauen und zu verbessern, damit das Auffinden der gewünschten und geeigneten Informationen effektiver wird (Notess 1999).

Mit dem Internet als Primärquelle für Daten und Informationen sollten nun auch Ansätze zum Vergleich bzw. zur Evaluierung der Inhalte der Internet Ressourcen angestrebt werden. Bisher sind in dieser Richtung wenige Anstrengungen unternommen worden.

Verfahren aus dem Gebiet der explorativen multivariaten Statistik sind für eine geordnete Darstellung der Objekte ein geeignetes Verfahren. Nur eine solche Ordnung erlaubt es, geeignete Schlüsse aus Datensätzen zu ziehen.

In der vorliegenden Arbeit wird ein Ansatz der vergleichenden Darstellung der umwelt- und chemikalienrelevanten Inhalte von Datenbanken in den drei Medien online, CD-ROM und Internet vorgestellt. Es zeigen sich unterschiedliche Schwerpunktsbildungen in den einzelnen Medien bezüglich der Deskriptoren, auf die im Kapitel 3 im einzelnen eingegangen wurde.

Es gibt Bereiche, in denen Internet-Datenbanken niedrigere Ränge als die klassischen Medien einnehmen. Dies gilt im Besonderen für Parameter über das Vorkommen von Chemikalien in den Umweltmedien.

Bei den chemierelevanten Deskriptoren zeigt es sich, daß alle drei Medien niedrige Ränge belegen. Eine Ausnahme bildet die Strukturformel. Hingegen liegen Internet-Ressourcen rangmäßig höher im Bereich der klassischen Gesundheitsparameter wie z.B. einiger Toxizitätsparameter, Haut- und Augenreizung sowie inhalative und dermale Toxizität.

Multivariate explorative statistische Verfahren bieten einfache und effektive Hilfsmittel für die graphische Analyse von Datenmatrizen. Andere mathematische Verfahren aus dem Bereich der Verbandstheorie können die Objekte in Form von sog. Hasse-Diagrammen visualisieren. Mit dieser Methode, die die Ordnungsrelation beinhaltet, können gute und schlechte Objekte (in diesem Fall Deskriptoren) herausgestellt werden. Für den vorliegenden Datensatz von 50 Objekten ist die Methode der Hasse-Diagramm-Technik nicht besonders geeignet, da durch die Darstellung jedes Objektes die Graphik unübersichtlich wird (Voigt/Welzl 1999). Eine Reduktion der Objekte beispielsweise mittels eines Clusterungsverfahrens würde zu einem deutlich besser geeignetem Datensatz für die Anwendung der Hasse-Diagramm-Technik führen.

Demnach wird eine Kombination der explorativen statistischen Methoden mit den Methoden der Hasse-Diagramm-Technik in Zukunft in unseren Arbeiten angestrebt.

Literaturverzeichnis

BMBF (1997): Forschung für die Umwelt, Programm der Bundesregierung, Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie, Bonn

Einax, J., Zwanziger, H., Geiß, S. (1997): Chemometrics in Environmental Analysis, Weinheim

Günther, O. (1998): Environmental Information Systems, Berlin

Lorenz, R. J. (1996): Grundbegriffe der Biometrie, Stuttgart

Lawler, E.L. (1985): The traveling salesman problem, Chichester

Notess, G. (1999): Search Engines Into the Internet Age, in: Online, 23, No. 3, pp. 20-23

NUA (1999): http://www.nua.ie/surveys/how_many_online/index.html, May 1999

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T. (1986): Numerical Recipies, The Art of Scientific Computing, New York/NY

Stoyan, D., Stoyan, H., Jansen, U. (1997): Umweltstatistik, Statistische Verarbeitung und Analyse von Umweltdaten, Stuttgart

Voigt, K. (1997): Erstellung von Metadatenbanken zu Umweltchemikalien und vergleichende Bewertung von Online Datenbanken und CD-ROMs, Aachen

Voigt, K., Benz, J. (1998a): Umweltchemikalien und Umweltmodelle im Internet, in: Riekert, W.-F., Tochtermann K. (Hrsg.): Hypermedia im Umweltschutz, 1. Workshop Ulm 1998, Marburg, S. 63-66

Voigt, K. (1998b): Environmental Information Databases, in: Schleyer, P. v. R. et al. (Hrsg.): The Encyclopedia of Computational Chemistry, Chichester, pp. 941-952

Voigt, K., Brüggemann, R.(1998c): Evaluation Criteria for Environmental and Chemical Databases, in: Online & CD-ROM Review, 22, No. 4, pp. 247-262

Voigt, K., Benz, J. (1999a): Statistik zur Auswertung des Zugriffs auf Umwelt-relevante Metadatenbanken, in: Dade, C., Schulz, B. (Hrsg.): Management von Umweltinformationen in vernetzten Umgebungen, Marburg, S. 80-83

Voigt, K., Welzl, G, Rediske, G. (1999b): Comparison of Environmental Databases: Online, CD-ROM, and Internet, in: Williams, M.E. (Hrsg.): Procedings of the 20th Annual National Online Meeting, New York May 18-20, 1999, Medford/NJ, pp. 487-498

Welzl, G., Voigt, K., Rediske, G. (1998): Visualisation of environmental pollution - Hasse diagram technique and explorative statistical methods, in: Sciences, Berichte des Institut für Gewässerökologie und Binnenfischerei Berlin, Heft 6, Sonderheft I, S. 101-110

Welzl, G., Voigt, K., Rediske, G. (1999): Methodische Fragen aus Sicht der Multivariaten Statistik zur Bewertung von Umweltobjekten, in: Weigert, B. (Hrsg.): Methodische Ansätze in der Umweltbewertung, Schriftenreihe Wasserforschung, Band 5, S. 59-70