ProRank a Software Tool used for the Evaluation of Environmental Databases

K. Voigt[a], **S. Pudenz**[b], **R. Brueggemann**[c]

[a] *GSF-National Research Centre for Environment and Health, Neuherberg, Germany, kvoigt@gsf.de*

[b] *Criterion Berlin, Germany*

[c] *Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany*

Abstract: On one hand a lot of research has been performed and an enormous amount of data generated concerning protection of human health and the environment with respect to chemical substances. On the other hand great data gaps for environmental chemicals have been detected in the run-up to a new chemicals policy in the EU, known under the name of REACH. In our approach we elucidate the data availability on chemicals applying the Hasse Diagram Technique (HDT) which originates in discrete mathematics as an environmetrical and chemometrical method. The software package used is ProRank software for multi-criteria evaluation and decision support (http://www.prorank.biz). We evaluate 15 environmental and chemical Internet databases which respect to the existence of data on 24 chemicals (12 pharmaceuticals and 12 high production volume (HPV) chemicals) in these resources. The applied methodology reveals the best and the worst databases and conflicts among them, due to different information content. The consequences of the aggregation and weighting of attributes by applying special features of the ProRank program will be demonstrated. The information gap especially for pharmaceuticals entering the environment is demonstrated. Impulse should be given for future research in the generation of new and valuable data.

Keywords: Environmental Databases; High Production Volume Chemicals; Pharmaceuticals; Discrete Mathematics; Hasse Diagram Technique; posets; Software Tool

1. INTRODUCTION

The EU's chemicals policy has renewed the interest on pre-screening and evaluation procedures for so-called existing chemicals. REACH stands for Registration, Evaluation and Authorisation of Chemicals. This new EU regulation will replace 40 existing legal acts and create a single system for all chemical substances. REACH will require manufacturers and importers to gather comprehensive information on properties of their substances produced or imported in volumes over 1 tonne per year and to submit the necessary information to demonstrate their safe use in a registration dossier to the European Chemicals Agency [Europa, 2005]. The regulation does not only encompass industrial chemicals but all chemical classes including e.g. pesticides and pharmaceuticals. Many chemicals have been detected in various environmental media so far including pharmaceuticals. Pharmaceuticals are omnipresent in wastewater world-wide. For several years pharmaceuticals have also been detected in surface water, ground water, drinking water, soil and sediment. Exemplifying the quantities 50,000 drugs were registered in Germany at the end of the nineties for human use, 2,700 of which accounted for 90 % of the total consumption and which contained about 900 different active ingredients. In the UK 3,000 active substances were licensed at the same time [Kümmerer, 2001]. Studies on the availability of data on chemicals including pharmaceuticals were performed in recent studies by the authors [Voigt and Brüggemann, 2005], [Voigt et al., 2006]. The data availability is an urgent prerequisite to scrutinize chemical substances concerning their environmental behavior and effect. The imminent question to answer is whether publicly available databases comprise information on environmental chemicals and in a further step what kind of information is available for the chemicals.

In the present paper we want to scrutinize the data availability for 12 high production volume (HPV) chemicals and 12 pharmaceuticals. The evaluation is performed applying a multi-criteria evaluation and decision support method, named Hasse Diagram Technique (HDT).

2. MULTI-CRITERIA EVALUATION METDOD

2.1 Background of the Hasse Diagram Technique

The basis of the Hasse Diagram technique (named HDT for short) is the assumption that a ranking can be performed while avoiding the use of an ordering index [Halfon and Reggiani, 1986]. For an evaluation of the objects they must be compared. The comparison is done by examining characteristic properties (attributes, descriptors) of these objects. If the evaluation is aimed to assess criteria, then the attributes or (synonyms: descriptors) are thought of as measures, how well a criterion is fulfilled. Attributes are -in the case of the object "x" denoted as q(1,x), q(2,x),...,q(m,x) and often written as a tuple q(x). Often the properties are gathered to a set without reference to actual values realized by the objects. This set of properties is called an information base IB. Often subsets of IB are needed. Consider now two objects x and y, then we say $y \geq x$ (with respect to the m properties of interest) if
$q(i,x) \leq q(i,y)$ for all i =1 ,2,..,m and there is at least one i* , for which $q(i^*,x) < q(i^*,y)$ (because of the demand "for all" this definition is denoted as "generality principle")
If $q(i,x) \geq q(i,y)$ or $q(i,x) \leq q(i,y)$ for all i=1,..,m then the objects x and y are comparable. The mere fact that x is comparable with y (without the information about the orientation) is often denoted as $x \perp y$.
Often however one finds
$q(i,x) < q(i,y)$ for one index set I' and
$q(i,x) > q(i,y)$ for another index set I'' with $I' \cap I'' = \varnothing$.
In that case, the objects x and y are incomparable and one writes: $x \parallel y$. The order relation defined here is known as product order. There are many other ways to define order relations.
The main frame of HDT is therefore (the four-point-program [Brüggemann and Welzl, 2002]:

1. Selecting a set of elements of interest which are to be compared, E. The set E is called ground set. This notation expresses that the ground set together with at least one binary relation among the elements of E gets a structure, which can be often represented as digraph as in the case discussed here.
2. Selecting a set of properties, by which the comparison is performed, called the information base IB.
3. Finding a common orientation for all properties; according to the criteria they are assigned.
4. Analysing $x,y \in E$ whether one of the following relations is valid:
    - $x \sim y$ (equivalence, we call the corresponding equivalence relation R, the equality of two tuples q(x), q(y))
    - $x \leq y$ or $x \geq y$ (comparability)
    - $x \parallel y$ (incomparability, there is a "contradiction in the data of x and y")

The relation defined above among all objects is indeed an order relation, because it fulfills the axioms of order, namely

- reflexivity (one can compare each object with itself)
- antisymmetry (if x is preferred to y then the reverse is only true, if the two objects are equal (or equivalent)
- transitivity (if x is better than y, and y is better than z, then x is better than z).

A set E equipped with an order relation $\leq$ is said to be an ordered set (or partially ordered set) or briefly "poset" and is denoted as $(E, \leq)$.

We note: A set E equipped with a partial order is often written as $(E,\leq)$. Because the $\leq$-comparison depends on the selection of the information base (and of the data representation (classified or not, rounded, etc.) we also write (E, IB) to denote this important influence of the IB for any rankings [Brüggemann and Welzl, 2002].

In our applications the ellipses near the top of the page (of the Hasse Diagram) indicate objects that are the "better" objects according to the criteria used to rank them: The objects not "covered" by other objects are called maximal objects. Objects which do not cover other objects are called minimal objects. Equivalent objects (denoted by Kn) are different objects that have the same data with respect to a given set of attributes. Only one representative of the equivalent objects is shown in the Hasse Diagram.

If empirical posets are to be examined, it is important to establish orientation rules, i.e. which value of attributes is considered to contribute to "badness" and which values to "goodness". When evaluating chemical and environmental Internet databases the availability of data in database x is: value 1 means available information, hence

"good", the value 0 means information unavailable, hence "bad".

2. 2 ProRank Software

The method introduced in this paper is based on discrete mathematics. The commercial software is called ProRank - Software for multi-criteria evaluation and decision support and will be applied here. ProRank presents a rather new approach based on partially ordered sets that can be used to avoid the loss of information by merging characterizing properties and thus preserve important elements of the evaluation and decision-making processes [Criterion, 2006].

The background of the applied Hasse Diagram Technique is explained in a variety of different environmental and chemical as well as statistical journals. A rather comprehensive description can be found in [Brüggemann et al., 2001, 2002]. This is the reason why we will only give a very brief and incomplete introduction into this method at this point.

3. EVALUATION of 15 DATABASES by 24 CHEMICALS

For the evaluation of databases by chemical substances a data-matrix consisting of 15 Internet databases and 12 high production volume chemicals as well as 12 pharmaceuticals is set-up.

3.1 Selection of Objects and Attributes

The evaluation was performed in spring 2005. The chosen databases which are listed together with their later used abbreviation are all available on the free Internet. Not only US databases but also European databases, are covered (see Table 1).

Table 1. List of 15 Numerical Databases.

Name	Abb.
Biocatalysis/Biodegradation Database	BID
Chemicals Information System for Consumer-relevant Substances (CIVS)	CIV
ChemExper Catalog of Chemical Suppliers, Physical Characteristics	CEX
ECOTOX	ECO
Envirofacts	ENV
Environmental Fate Database	EFD
Environmental Health Criteria (EHCs)	EHC
ESIS – European Chemical Substances IS	ESI
EXTOXNET	EXT
GESTIS – Dangerous Substances Db.	GES
HSDB	HSD
International Chemical Safety Cards	ICS
N-Class Database	NCL
Oekopro	OEK
OECD Integrated HPV Database	OIH

Four different types of numerical databases can be distinguished:

- Single databases which cover only one data collection (BID, CIV, GES, HSD, ICS, NCL, OEK)
- Multi-database databases which encompass several databases under the same name and search interface (ECO, ENV, EFD, ESI, EXT)
- Monograph databases which cover extensive reviews on very few chemicals (EHC, OIH)
- Catalogue database (CEX).

The queries were made by CAS-numbers.

Table 2. List of 24 Chemicals.

No.	Name of Chemical	ACR.	Use
1	Bezafibrate	BEZ	PAC
2	Carbamazepine	CAR	PAC
3	Clofibric acid	CLO	PAC
4	Diclofenac	DIC	PAC
5	Diazepam	DAP	PAC
6	Ethinyl Estradiol	EES	PAC
7	Fenofibrate	FEN	PAC
8	Ibuprofen	IBU	PAC
9	Metoprolol	MET	PAC
10	Phenazone	PHE	PAC
11	Roxithomycin	ROX	PAC
12	Sulfamethoxazole	SUL	PAC
13	1-chloro-4-nitrobenzene	CNI	HPV
14	4-nitroaniline	NIA	HPV
15	4-nitrophenol	NIP	HPV
16	Atrazine	ATR	HPV
17	Chlormequat chloride	CMC	HPV
18	Diazinon	DIA	HPV
19	Dimethoate	DIM	HPV
20	Ethofumesate	ETO	HPV
21	Glyphosate	GLY	HPV
22	Isoproturon	ISO	HPV
23	Malathion	MAL	HPV
24	Thiram	TIR	HPV

The selection of the pharmaceuticals was performed in an intensive literature study carried out in the year 2004 [Voigt and Brüggemann, 2005]. Those pharmaceuticals which were detected in several German rivers were chosen as the basis of the test-set [Wiegel et al., 2004]. Concerning the choice of high production volume chemicals we build up on a study made by Lerche et al. [2002]. This test-set has been evaluated regarding different chemicals by Voigt et al. [2006]. In the

following we want to put the emphasis on explaining features of the ProRank software, e.g. the grouping and aggregation procedures.

3.2 Application of ProRank on the Data-Matrix

The data-matrix is constructed representing 15 objects (Internet databases) and 24 attributes (12 pharmaceuticals plus 12 high production volume chemicals). The availability of the chemical n in database x is denoted by 1, the unavailability by 0. In Figure 1 the Hasse Diagram of the complete 15x24 data-matrix is presented.

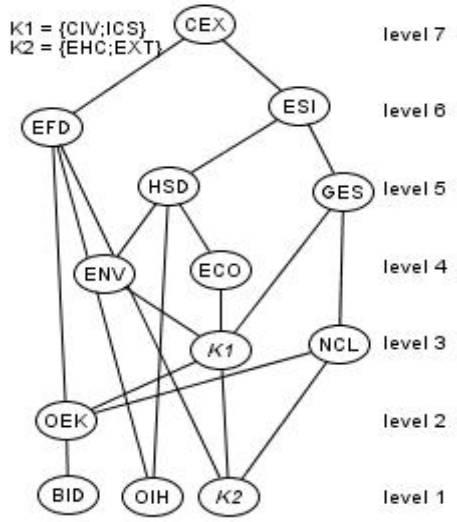

Figure 1. Hasse Diagram of 15 x 24 data-matrix.

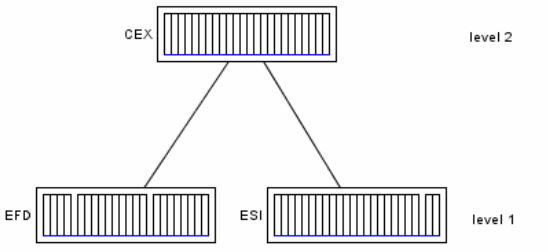

Figure 2. Hasse Bar Diagram of Objects CEX, EFD, ESI. Note the renumbering of the levels.

The diagram (Figure 1) is structured into 7 levels. Only 11 objects are shown explicitly; the two equivalent objects are named K1 {CIV;ICS} and K2 {EHC;EXT}, which are an equivalence classes. In the diagram you find one maximal object CEX and three minimal objects, namely BID, OIH and the equivalent object K2 (CIV;ICS). The partial order should be elucidated by the three objects found in level 7 and 6 which are CEX, EFD and ESI. It is possible to draw a so-called bar Hasse diagram which explains the partial order theory (see Figure 3). The catalogue database CEX (ChemExper Catalog of Chemical Suppliers, Physical Characteristics) is the maximal object and comprises all 24 chemicals. EFD (Environmental Fate Database) has information on 22 chemicals. The chemicals CMC and ISO are missing. This means EFD is worse than CEX. ESI (ESIS-European Chemical Substances Information System) encompasses 23 chemicals, the chemical ROX is missing. This means it is worse than CEX and at the same time incomparable to EFD, applying the generality principle.

3.3 Grouping and Aggregation of Attributes

In decision support it is often of great interest to group and/or aggregate attributes. As mentioned above our test-set comprises different kinds of chemicals, pharmaceuticals and high production volume chemicals (most of them are pesticides). One possibility is to take a look at the sub test-sets which means pharmaceuticals or high production volume chemicals in separate approaches. This analysis shows that the data situation for pharmaceuticals entering the environmental media is extremely poor whereas the data availability for pesticides is considerably better but far away from being satisfactory [Voigt et al., 2006]. In the present approach we demonstrate the grouping of pharmaceuticals on one side and the grouping of pesticides on the other side. In data analysis terms this means reducing the 15x24 data-matrix to a 15x2 data-matrix.

This procedure can easily be carried out by the ProRank program under the feature "grouping and aggregation". In a first step we aggregate all 12 pharmaceuticals to one super-criterion Pharmsuper and all 12 high production volume chemicals to another super-criterion HPVsuper. Both aggregations are calculated by adding without weighting.

Afterwards both super-criteria Pharmsuper and HPVsuper are used for partial order ranking of the databases. The result of this data reduction procedure is given in Figure 3. The left bar represents the HPV chemicals whereas the right bar shows the pharmaceuticals. EFD and OIH are the only databases which comprise more information on pharmaceuticals than on HPV chemicals. The databases NCL, OEK, K2 {EHC, EXT} and BID do not cover information on pharmaceuticals at all.

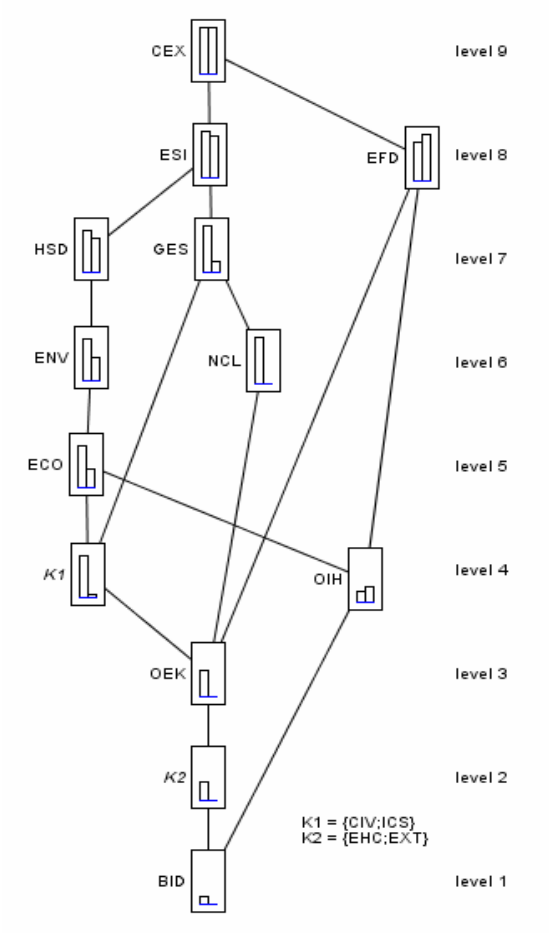

Figure 3. Hasse Bar Diagram of 15 Databases and 2 Super Criteria.

It can be concluded that the data situation on the pharmaceuticals is poorer than on the HPV chemicals. Within this respect it is logical to weight either the pharmaceuticals higher than the HPV chemicals or vice versa. This can easily be performed by the weighting procedure provided by the ProRank program. In a first step we weight the sum of the pharmaceuticals 0.75 and the HPV chemicals by 0.25. The procedure as well as the resulting Hasse diagram is given in Figure 4. In the upper part of Figure 4 the "grouping and aggregation" dialog is shown displaying the super-criteria Pharmsuper and HPVsuper, the weights used and the resulting criterion labelled as Weighted 1. The two other dialog boxes represent the table (15x1 data-matrix) and the Hasse diagram corresponding to the criterion Weighted 1. This diagram shows a linear order. It may however be the case that one is more interested in the availability of information on HPV chemicals than on pharmaceuticals. Hence we also performed the weighting data analysis in the other direction. Instead of presenting and comparing both linear orders each as a result of different weighting procedures, both criteria Weighted 1 and Weighted 2 will be considered simultaneously.

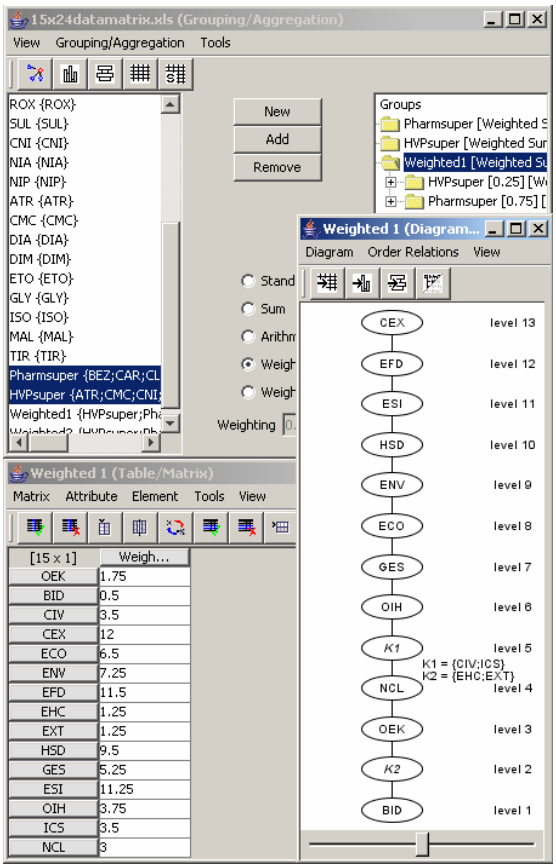

Figure 4. Weighting procedure for Test-sets of Pharmaceuticals and HPV Chemicals in ProRank.

This means that the databases are to be partially ordered by the criteria Weighted 1 and Weighted 2. The result of this procedure is presented in Figure 5. Levels consisting of only one database mean that they are independent from the weighting procedure, both pharmaceuticals and HPV chemicals lead to the same rank. For example it is demonstrated that the weighting selected here has no influence on the positions of BID, K2 (EHC and EXT), OEK, ENV, HSD and CEX. It is also striking that the maximal and the minimal objects CEX and BID are relatively stable against weighting. CEX is the database that contains the largest number of chemicals independent from emphasize (weight) laid on either pharmaceuticals or HPV chemicals side. Incomparable databases, in contrast, indicate a conflict: emphasis on either pharmaceuticals or high production volume chemicals leads to different ranks of a database. For example in level eight EFD and ESI are incomparable. This means that the weighting procedure on the pharmaceuticals side leads to a different position than the weighting procedure on HPV chemicals. The more objects we find in a level, the greater the conflicts due to weighting

and the information content pertaining to HPV chemicals and pharmaceuticals are.

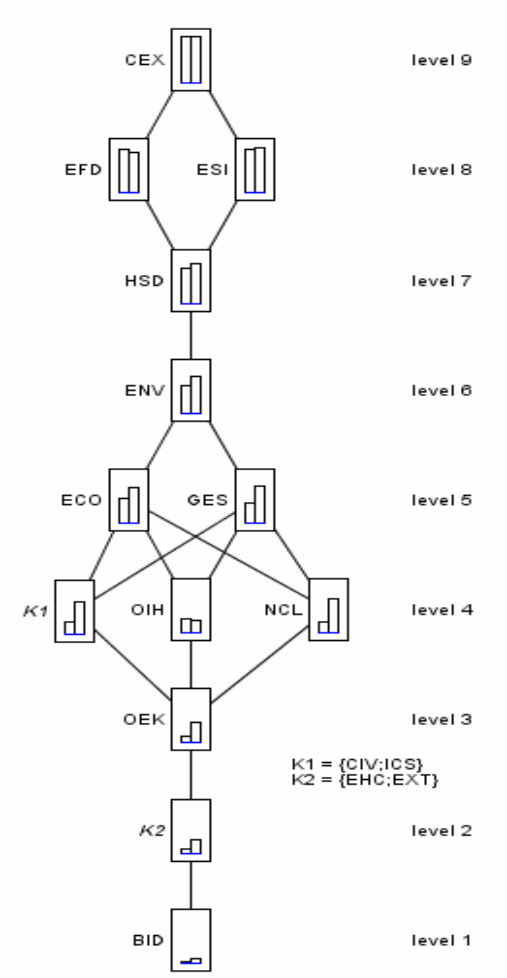

Figure 5. Pharmaceuticals weighted higher (left bar), HPV chemicals weighted higher (right bar).

5. CONCLUSIONS and OUTLOOK

To conclude, one can state that the data situation on the 24 chemicals evaluated in this approach is not satisfactory at all and must be improved in order to protect humans and the environment. Concerning the data situation on pharmaceuticals we discovered that in five databases no information on these substances exists at all. The data situation in the rest of the evaluated databases was very poor, with the exception of CEX - Catalog of Chemical Suppliers, Physical Characteristics. This is a catalogue of chemical substances which provides only data on some basic physical-chemical properties. A special position was found for the database OIH in the performed aggregation and weighting procedure.
For future steps concerning the data availability of chemicals, ways must be found to extract or generate data in order to avoid intensive testing.

Concerning the data-analysis approach we are of the opinion that further evaluation studies aiming at the testing of the information quality of environmental and chemical databases should be performed. In this respect we will evaluate the content of the databases in the direction of environmental fate and pathways and ecotoxicity parameters. We will continue our research into that direction using discrete mathematical methods like the Hasse Diagram Technique as demonstrated in this paper and comparing this method with multi-variate statistical methods.

6. REFERENCES

Brüggemann, R., R. Halfon, G. Welzl, K. Voigt, C. Steinberg, Applying the Concept of Partially Ordered Sets on the Ranking of Near-Shore Sediments by a Battery of Tests, *J. Chem. Inf. Comp. Sci.*, *41*, 918-925, 2001.

Brüggemann R., G. Welzl, Order Theory Meets Statistics in: Voigt, K, G. Welzl (eds.), Order Theoretical Tools in Environmental Sciences. Shaker-Verlag, 9-40, Aachen, 2002.

Criterion, ProRank Software, http://www.prorank.biz, 2006

Europa Rapid Press Releases: http://europa.eu.int/rapid/pressReleasesAction.do?reference=IP/05/1583&format=HTML&aged=0&language=EN&guiLanguage=en, 2005.

Halfon, E., M.G. Reggiani,, On Ranking Chemicals for Environmental Hazard, *Environ. Sci. Technol*., *20*, 1173-1179, 1986.

Kümmerer, K., Drugs in the Environment: Emission of Drugs, Diagnostic Acids and Disinfectants into Wastewater by Hospitals in Relation to Other Sources, *Chemosphere*, *45*, 957-969, 2001.

Lerche, D., E. van de Plassche, A. Schwegler, F. Balk, Selecting Chemical Substances for the UN-ECE POP Protocol, *Chemosphere*, *47*, 617-630, 2002.

Voigt, K., R. Brüggemann, Water Contamination with Pharmaceuticals, *MATCH Commun. Math. Comput. Chem*., *54*, 671-689, 2005.

Voigt, K., R. Brüggemann, S. Pudenz, Information Quality of Environmental and Chemical Databases Exemplified by High Production Volume Chemicals and Pharmaceuticals, *Online Information Review*, *30*, 1, 8-23, 2006.

Wiegel, S., A. Aulinger, R. Brockmeyer, H. Harms, J. Löffler, H. Reinicke, R. Schmidt, B. Stachel,, W. von Tümpling, A. Wanke, Pharmaceuticals in the River Elbe and Its Tributaries, *Chemosphere*, *57*, 107-126, 2004.