ELSEVIER

Journal of Chromatography A, 974 (2002) 223–230

JOURNAL OF CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

Automated storage of gas chromatography–mass spectrometry data in a relational database to facilitate compound screening and identification

J.A. Staeb[a,*], O.J. Epema[a], P. van Duijn[a], J. Steevens[b], V.A. Klap[a], I.L. Freriks[a]

[a]*Institute for Inland Water Management and Waste Water Treatment, RIZA, P.O. Box 17, 8200 AA Lelystad, The Netherlands*

[b]*TLC Chemistry and Software, Amsterdam, Simonshavenstraat 60, 1107 VC Amsterdam, The Netherlands*

Abstract

This paper describes a database containing mass spectra from gas chromatography–mass spectrometry (GC–MS) measurements as a tool for easy screening for multiple compounds. In this way additional compounds can be reported from the same run together with routine pesticide monitoring with little effort. The relevant analytical data from the GC–MS measurements are transferred automatically to a database. Search algorithms in the database, containing the US EPA and Dutch NEN GC–MS identification criteria as standard settings, are used to identify compounds in the data. Screening of samples analysed in our laboratory show the ubiquitous presence of—up until now in monitoring largely overlooked—compounds in surface waters in The Netherlands. Most frequently found compounds include TAED (complexing agent), 2-methyl quinoline (industrial solvent), atrazin and desethylatrazin (pesticide and degradation product), caffeine (human consumption), surfinol-104 (anti foaming agent), HHCB (Galaxolide) and AHTN (Tonalide; fragrances). The database can also be used to quickly search a large number of datafiles for rare contaminants. This way, some interesting compounds such as pentoxifilin (a pharmaceutical) and Irgarol 1051 (an antifouling compound) were found.

Keywords: Data storage, automated; Database, GC–MS; Identification criteria, GC–MS; Irgarol 1051; Pentoxifilin

1. Introduction

In view of the large number of organic compounds present in environmental samples, the data usually reported by analytical laboratories to their customers are remarkably limited. It is due to quality control considerations that only a few compounds are reported while other compounds that are present in the same chromatographic run are skipped. It is often later that interest in other compounds arises. However, even semi quantitative indication of the presence of those compounds then turns out to be an elaborate task. The acquired GC and GC–MS data of the original analyses are usually not easily accessible for quick later screening. In order to account for changing interests over time and to allow multiple sample screening, another method of data storage is required.

Ouchi described the versatility of modern desk-top database packages for data management in general and their applicability to analytical chemistry in particular [1,2]. Lipinski and Stan reported the use of a database for screening food samples on pesticide residues [3]. The analytical data described by these

*Corresponding author. Tel.: +31-320-298-657; fax: +31-320-298-799.

E-mail address: j.staeb@riza.rws.minvenw.nl (J.A. Staeb).

0021-9673/02/$ – see front matter
PII: S0021-9673(02)01309-2

workers consisted of GC traces obtained by flame ionization detection (FID), FPD (flame photometric detection) and nitrogen–phosphorus detection (NPD). Acquired GC data were automatically loaded into a database program. The major function of the program was to compare sample chromatograms with those of a reference set of pesticides and matrix chromatograms.

For GC–MS spectra KIWA, RIZA and several waterworks developed an expert system called Infospec [4]. The formula is that all members load the database with analytical data (retention indices and mass spectra), while KIWA maintains the quality of the identification and quantitation parameters. Sample results may be stored but peak assignment is fixed after the results are put into the database. It is not possible to change assignments after new compounds are discovered except for predefined ''unidentified'' compounds.

In view of the large number of still unidentified compounds in the environment we decided that it was desirable to retain the exact spectrum in the database in order to allow for renewed identification when new knowledge becomes available. To that end, in the presented approach an algorithm was written that reports all peaks in a chromatogram above a certain threshold to a database. In this paper we present the algorithm and the relational database. Search facilities in the database allow retrospective automatic or manual assignment of compound names to GC–MS peaks using standardized or user defined GC–MS criteria.

2. Experimental

2.1. Samples and analysis

In our laboratory samples of various origins like surface, interstitial, sewage, drinking and groundwaters, sediments and particulate suspended matter were analysed. One standard analytical procedure involves prefiltration of a volume of 0.1–3 l water and subsequent adsorption of organic compounds on a polymer column. After drying the column, adsorbed compounds are eluted with dichloromethane. A recovery standard (1-chlorodecane) is added to the eluent and the sample is analysed by GC–MS. Full scan electron impact (EI) mass spectra are acquired. Data acquisition and processing were performed using an Agilent 6890 GC and 5973 MSD system. The database is built using Microsoft Access.

2.2. Data processing and data transfer

The acquired data are analysed by a technician using a calibration table containing 20–40 target compounds. In the next step an Agilent Chemstation algorithm transfers—next to the identified compounds—all relevant mass spectra into the database (Fig. 1). To this end the algorithm reports all possible interesting compounds present in the chromatogram by integrating all ion chromatograms from m/z 35 up to m/z 500. For each peak the Kovats retention index is calculated and suggestions of the identity are included using the built-in probability based matching algorithm [5] and a mass spectra library. For the Kovats index calculation one has to establish the precise Kovats indices of the calibrated target compounds by comparing them with an alkane mixture on the instrument used. Next, in routine work the algorithm uses the retention times of calibrated target compounds to calculate the Kovats indices of unknown peaks. This way very reliable Kovats indices are obtained even if the column is made shorter or gas flows are not stable over a longer time. On average the chemstation algorithm reports some 800 peaks for each surface water chromatogram. This is comparable to the AMDIS program that reported approximately the same number of peaks for the same test chromatogram. The possibility to import AMDIS results automatically into the database will be ready by the time this paper is published.

2.3. The database

In a relational database information is grouped in tables that can be related via keys. Fig. 2 shows a diagram of the database. The framework consists of three tables of which the composing fields are indicated. The database is divided into two different sections: analytical data and reference data.

The heart of the database is the analytical data section (left part in Fig. 2), composed of the table MSPEAKS and SAMPLEINFO. These tables con-

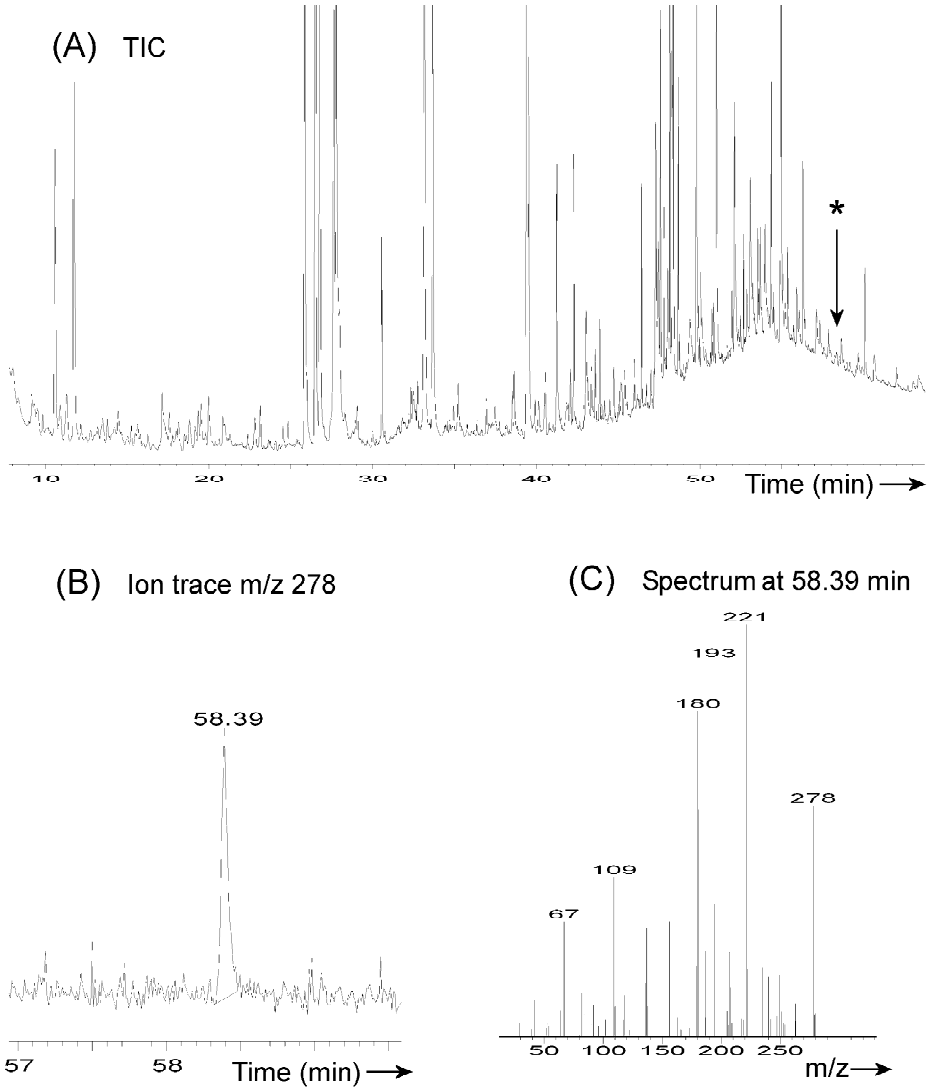

Fig. 1. (A) GC–MS total ion current chromatogram (TIC) of a water sample from the river Rhine. The large number of compounds present form a large hump in the chromatogram. In routine monitoring only 20–70 specific compounds are identified while the other information present in the chromatogram is not used. The data extraction algorithm searches all ion traces from m/z 35 to m/z 500. (B) By searching ion trace 278 (and other traces) a compound is discovered at a retention time of 58.39 min that is not visible at all in the TIC (see arrow with asterisk in A). (C) The background subtracted spectrum at this retention time is automatically transferred to the database and is later identified by the comparison algorithm as the pharmaceutical pentoxifylin.

tain the acquired analytical data, i.e., some 800 peaks for each chromatogram including MS spectrum, Kovats index and sample information. Most fields have self-explanatory names, like SampleName, KovatsIndex, RetTime, Area, Height, MassX and IntensX. The field CompoundName refers to an identified compound or is otherwise assigned as ''unknown''. The field Qcode refers to the reliability of the identification method. Three different codes are distinguished. The highest reliability code ''0'' is given to compounds, which are target compounds identified by the technician using the Agilent chemstation software in the daily laboratory routine. Qcode ''3'' is reserved for unknown compounds.

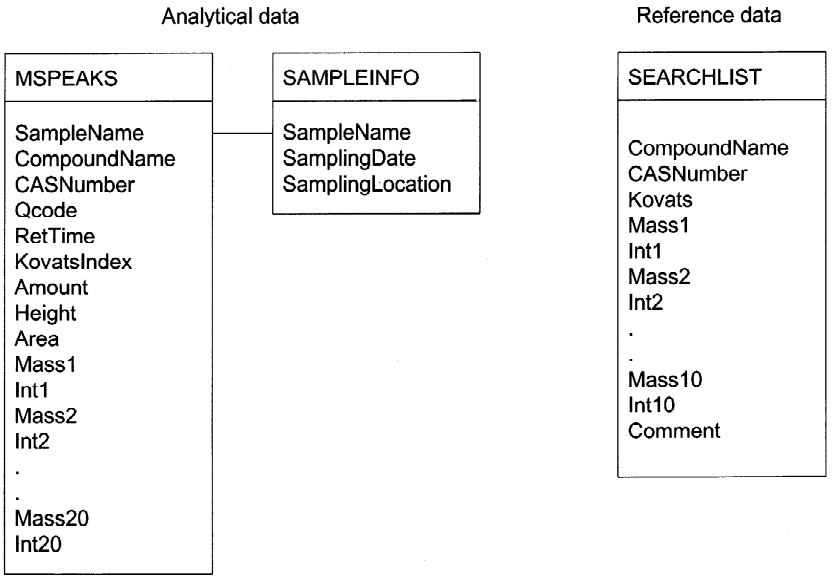

Fig. 2. Diagram of the database framework. A relation between two fields is indicated by a connection line. Table contents are described in the text.

Finally, Qcode ''DB'' is used for compounds identified via the database (hence DB). Currently, MSPEAKS contains over one million records (=peaks) from over 1300 samples.

The reference data section (right part in Fig. 2) contains a table of currently more than a thousand known compounds. Unlike a normal mass spectral library this table includes Kovats retention indices and is restricted to compounds that are known to occur in the environment.

2.4. Screening using the database

A search algorithm in the database allows the comparison of all (identified and not yet identified) peaks in the analytical data section with the reference data (expert knowledge). In this way the database is able to identify compounds. If the tolerances used for identification are small the reliability of a compound identified by the database can almost equal that of a compound identified by a technician.

A few standard sets of criteria exist that describe the masses that should be present at specified intensities and the allowed Kovats index deviation in order to identify the compound (Table 1). The US Environmental Protection Agency (EPA) criteria [6] require five m/z values at intensities $\pm 50\%$ of the calibrated intensity and use a retention criterion of ± 15 s. The Dutch Institute for Normalisation (NEN) has published a set of criteria requiring three m/z values at intensities within a narrower specified window ($\pm 10\% + 0.1 * \mathrm{Int}_{cal}$) and require a retention deviation window of 0.2% for positive identification [7]. Fig. 3 shows the search parameter screen which allows the user to perform any required search.

After a search is performed it is possible to change ''unknown'' compounds to ''identified'' compounds automatically or manually (see Fig. 4). Next to our

Table 1
Overview of US EPA and Dutch NEN GC–MS criteria and the set of criteria used in the current report

Criterion	EPA identification	NEN identification	NEN indication	Present paper
Number of masses	5	3	3 (above decision limit)	3
Intensity window	0.5*intensity	0.1*intensity+10%	Must be present	20%
Kovats window	15 s	0.2%	1.0%	1.0%

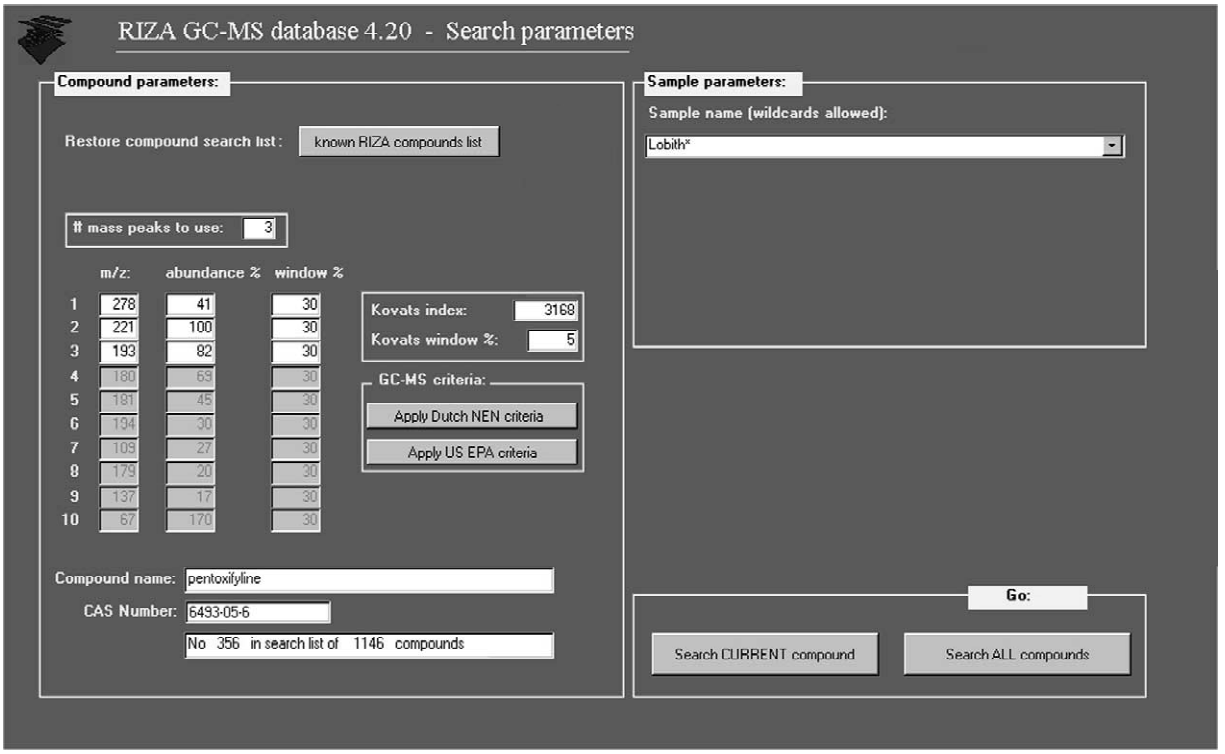

Fig. 3. Screenshot of the ''search parameters'' screen. On the left side the mass spectrum of a compound of interest (up to 10 m/z values and intensities) can be typed in with tolerance windows for the intensities. In the Kovats index field the Kovats index and tolerance window can be given. Next a search on one specific sample or all samples in the database can be started by pressing the button ''search CURRENT compound''. Alternatively a list of >1000 known compounds can be loaded (''known RIZA compound list'') and one or all compounds can be selected. Next a search may be started for all known compounds (''search ALL compounds'').

own spectra and Kovats indices the reference data section contains literature spectra and Kovats indices. Of course data acquired on our own instruments is the most reliable for the evaluation of the samples measured on the same instrument, but literature data are helpful for compounds we have not measured ourselves.

3. Results and discussion

3.1. Identification criteria for compounds

It was found that when using strict criteria such as the EPA or NEN criteria (Table 1) many compounds were missed. This is in line with expectation as these criteria are developed to provide a high certainty of the presence of the compounds (and thus to avoid false positive results). When using strict criteria, as a consequence some compounds—that are in fact present—are omitted (so-called false negative results).

For screening purposes it is better to apply less strict criteria in order not to miss important contaminants, but at the same time one has to accept a larger number of false positive results. The most important reasons for false negative results were:

(1) Contaminated spectra due to coelutions. The algorithm has a built-in background subtraction, but this fails at very low concentrations. When coelutions at certain m/z values or at high concentrations are present the spectrum is not pure enough to be recognized. A good deconvolution algorithm in the future might help to bring better spectra even at low concentrations.

(2) Deviations in Kovats index. Although a lot of

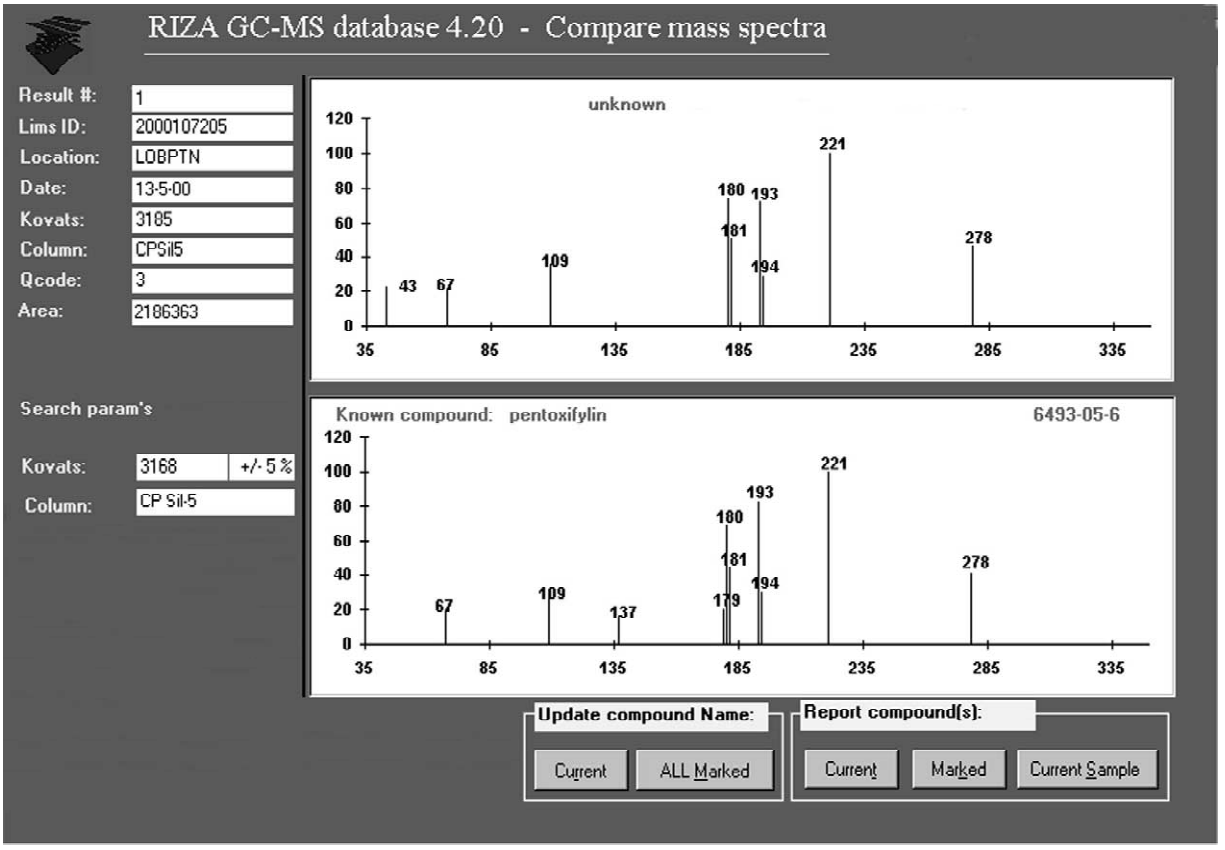

Fig. 4. Screenshot of the ''compare mass spectra'' screen. After a search has been carried out, identified peaks can be inspected visually and compound names can be assigned to the peaks. The information on the upper half of the screen applies to the sample and the unidentified mass spectrum found. The lower half shows the suggested identification.

effort was put into obtaining precise Kovats indices, in everyday routine deviations above 0.2% do occur. However if one realizes the necessity of good retention time calibration and stability this may be improved a good deal.

The most important reasons for false positive results were:

(1) Non-specific spectra. Several compounds with rather non-specific mass spectra (peaks at low m/z values) were reported by the database, but after manual inspection their presence was found to be doubtful. Examples were: dihydromyrcenol (m/z 59, Kovats 1060), benzonitril (m/z 103 and 76, Kovats 949) and L-(−)-menthol (m/z 81,71 and 95, Kovats 1160).

(2) Low concentrations. In some cases peaks with very low intensity were reported. Visual inspection revealed that the peaks had very low intensity and noisy spectra. Therefore the identifications were deemed as unreliable. At present the database has no option to estimate concentrations. In a future version this will be included so it will be easier to remove results at low concentrations.

For the above reasons in this study less strict criteria were used (see Table 1, last column). Next, the database results were checked manually. Indeed several compounds that were reported regularly by the database had to be deleted, as the results were not considered reliable, see above.

3.2. Screening a large dataset of samples

For a large set of daily samples comprising the years 1998–2000 a search was performed for the

>1000 known compounds that were measured on instruments in our laboratory. The 22 compounds most frequently found in the samples were checked manually and are presented in Table 2. In fact 174 compounds were found but only the top list is presented here. Monitoring programs generally are targeted at pesticides and would only show the presence of two of the 22 compounds shown here. This clearly shows that automated screening of regular monitoring samples reveals a lot of extra compounds present.

3.3. Searching for a single compound: Irgarol 1051

One rationale behind the development of the database was to be able to search for specific compounds in ''old'' analytical data in a time-efficient mode. Recently the question arose if the pesticide Irgarol 1051, used as anti fouling agent on ship hulls since the late 1980s, had been present in surface water over the last few years. The mass spectrum was available, but not the Kovats index. Using only the mass spectrum of five masses all sample data were searched and indeed five peaks matched. All the peaks had the same Kovats index and visual examination of the spectra revealed that the five masses were indeed present in the right ratio. In some spectra, however, other masses appeared simultaneously demonstrating co-elution of another compound.

Despite the strong mass spectral evidence for the presence of Irgarol 1051 in the samples, no retention time confirmation was obtained. For confirmation the pure compound was purchased, added to one of our standard mixtures and analysed. After this, both the chromatographic and mass spectral characteristics of Irgarol 1051 were known and the compound was added to the list of standard compounds.

3.4. Searching for a single compound: pentoxifilin

In one sample from the river Rhine the drug pentoxifilin was found. A search with the database in

Table 2
The 22 most frequently found compounds by the database in samples from the river Meuse at Eijsden from the period 1/1/1998–31/12/2000

Number of hits	CAS	Compound name
925	10543-57-4	TAED
915	91-63-4	Quinoline, 2-methyl-
886	1222-05-5	HHCB (Galaxolide)
771	1912-24-9	Atrazin
629	6190-65-4	Desethylatrazin
546	58-08-2	Caffeine
544	3622-84-2	Benzenesulfonamide, *N*-butyl-
431	260-94-6	Acridine
387	119-61-9	Benzophenone
364	126-86-3	Surfinol-104
315	25265-77-4	2,2,4-Trimethyl-1,3-pentanediol monoisobutyrate
308	108-75-8	Pyridine, 2,4,6-trimethyl-
306	34590-94-8	Dipropyleenglycol monomethylether
299	1506-02-1	AHTN (Tonalide)
243	6781-42-6	Ethanone, 1,1′-(1,3-phenylene)bis-
211	620-14-4	*m*-Ethyltoluene
210	122-34-9	Simazin
210	5131-66-8	2-Propanol, 1-butoxy-
193	115-96-8	Tri(2-chloroethyl) phosphate
188	1125-21-9	2,6,6-Trimethyl-2-cyclohexene-1,4-dione
187	591-22-0	3,5-Lutidine
173	134-62-3	Diethyltoluamide

The criteria used were: three masses with a 20% intensity window and a Kovats window of 1%. False positive results found with these settings were removed after manual inspection (see text).

all the available samples from the Rhine showed its until now unnoticed presence in two other instances (Fig. 4). However, in all >1000 samples from the river Meuse the compound was not found once. This proves that the compound is present in the river Rhine, but not in the river Meuse.

4. Conclusion

The GC–MS database is to our knowledge the only system that allows retrospective searching of large numbers of GC–MS peaks that were not identified at the time of analysis. Results of screening large datasets may be confirmed on a few random samples by manual identification, but one should realize that such large sets of data cannot be screened completely by hand in a reasonable time.

The advantage of the simple database interface is that no knowledge of the database software is required to conduct a search for a specific compound. For other searches users can use the possibilities of Microsoft Access itself. For instance, instead of searching for a specific compound with a known Kovats index and mass spectrum, one could also choose to screen the database for peaks with intensities exceeding a certain threshold.

References

[1] G.I. Ouchi, LC·GC 17 (1999) 1098.
[2] G.I. Ouchi, LC·GC 17 (1999) 924.
[3] J. Lipinski, H.-J. Stan, J. Chromatogr. 441 (1988) 213.
[4] I. Bobeldijk, J. van Leerdam, E. Doomernik, O. Epema, Th. Noij, in: E. Gelpi (Ed.), Advances in Mass Spectrometry, Vol. 15, Wiley, London, 2001, p. 929.
[5] G.M. Pesyna, R. Venkataraghavan, H.E. Dayringer, F.W. McLafferty, Anal. Chem. 48 (1976) 1362.
[6] EPA Method 1624, 1989.
[7] NEN, Identification Criteria for GC–MS (www.nen.nl).