Available online at www.sciencedirect.com
ScienceDirect
Journal of Chromatography A, 1157 (2007) 414–421
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Development of a database of gas chromatographic retention properties of organic compounds☆

V.I. Babushok, P.J. Linstrom, J.J. Reed, I.G. Zenkevich[1], R.L. Brown, W.G. Mallard, S.E. Stein*

National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

Received 12 February 2007; received in revised form 1 May 2007; accepted 3 May 2007
Available online 18 May 2007

Abstract

A comprehensive database of gas chromatographic retention properties of chemical compounds has been developed using multiple literature sources. The National Institute of Standards and Technology (NIST) database of retention data for non-polar and polar stationary phases currently contains 292,924 data records for 42,888 compounds. The database includes data for Kováts indices, linear indices, Lee indices, retention times and retention volumes. The first release of this database for non-polar stationary phases is available with NIST/US Environmental Protection Agency (EPA)/National Institutes of Health (NIH) Mass Spectral Database (June 2005) and through the internet (NIST Chemistry WebBook). The paper describes the database and the process by which it has been compiled. The format of data presentation and the quality control procedures are described. Data sources of gas chromatographic retention data are also discussed.
Published by Elsevier B.V.

Keywords: Gas chromatography; Retention indices; Kováts indices; Linear retention indices; Non-isothermal retention indices; Database

1. Background

Gas chromatography/mass spectrometry (GC/MS) is a widely used method for the identification of organic compounds in complex mixtures. Identification is typically carried out by matching measured spectra with the spectra in a reference library. Mass spectral reference libraries have been developed over many years. In contrast, there are practically no comprehensive libraries of retention indices for chemical compounds. The reliability of GC/MS identification is substantially increased by the use of both GC and MS identification approaches [1–3]. For example, differences in the structures of branched alkanes and *cis*/*trans* isomers have very little effect on the mass spectra and so make an identification difficult, but using the chromatographic data can often definitively identify these compounds [1,2]. The absence of reliable chromatographic databases has made the combined use of both techniques for identification less widely used.

The fundamental variable tracked in gas chromatography is the retention time. However, since the time is a strong function of experimental conditions, other parameters have been developed that are more independent of the experiment. The retention index (I) is a generally accepted type of data used for the identification of chemical compounds by gas chromatography. The system of retention indices suggested by Kováts [4] is a widely used and recognized system for recording gas chromatographic data for further use in the identification process. The index system suggested by Kováts in 1958 (and its further development—the linear system of indices for temperature-programming conditions [5]), allows the results measured in one laboratory to be used in other laboratories even under different carrier gas flow conditions. The retention index combines two fundamental gas chromatographic properties, the relative retention and the specific retention volume [6]. Kováts suggested a chemical ruler to characterize different chemical compounds on a relative time

☆ Disclaimer: The views and conclusions contained in this paper are those of the authors and should not be interpreted as presenting the official policies or position, either expressed or implied, of the National Institute of Standards and Technology or the US Government unless so designated by other authorized documents. Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof. The US Government has certain rights in copyright of this material.

* Corresponding author. Tel.: +1 301 975 2505.
E-mail address: steve.stein@nist.gov (S.E. Stein).
[1] On leave from the Chemical Research Institute, St. Petersburg State University, St. Petersburg, Russia.

0021-9673/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.chroma.2007.05.044

scale for identification purposes. Using the Kováts' definition, the measure of a chemical compound is its relative time position between neighboring linear paraffin hydrocarbons. This idea has been extended using other sets of compounds for reference markers. The other widely used set of compounds was suggested by Lee et al. [7] and it uses polyaromatic hydrocarbons as marker compounds.

To make use of the retention information for various compounds it is necessary to develop a comprehensive database along the lines of the mass spectral databases. Despite the fact that there has been a great deal of data published, there are currently only a relatively small number of retention data collections available, and many of these are limited to specific sets of compounds (Table 1).

At the early stage of retention data systematization, ASTM efforts on the organization and compilation of GC information were very important (ASTM Committee E-19 on Chromatography) [10]. The results of ASTM activity in the 1960s–1970s were the development of the format for presentation of retention data and the extensive data compilation. Data accumulated in the system in the beginning of 1969 totaled 43,000 data items for approximately 4000 compounds [10]. In spite of these initial efforts, there have been only very limited collections of *I* data.

There have been a number of data collections for specialized areas that include information for certain classes of chemical compounds or for some particular practical applications, e.g., retention data for gibberellins [3], or the well-known collection of retention data for essential oil components [1]. Some collections of retention data are available as published handbooks [1,2,8,10,11,13,14,21–26]. With the tremendous popularity of the Internet, several web-based collections were created [3,9,15–17,27]. Commercial specialized databases with limited scope of data are also available [1,18,19,27,28]. There were several attempts to organize retention index libraries—but to the present day there are practically no work in the area of analysis and evaluation of retention data [29]. Some collections contain reliable and carefully verified retention data, but they have a limited number of chemical compounds.

Recognizing that there was a clear need for a comprehensive database of retention index values that could be used along with mass spectrum search methods, NIST began collecting and evaluating the gas chromatographic literature with the aim of developing an evaluated database of retention indices (NIST GCRID) [30]. While the primary goal of such a database was to provide additional tools to reduce the rate of false positive identifications when using mass spectral libraries, the database is expected to be useful in reducing errors in reporting of retention data in the literature. The current database contains 292,924 data records for 42,888 chemical compounds measured on non-polar and polar stationary phases. The collection contains over 7000 sources of gas chromatographic properties including original papers, technical reports, conference proceedings and Internet sources (1958–2006). The first release of the *I* database for non-polar stationary phase is available as a part of NIST/EPA/NIH

Table 1
Published collections of gas chromatographic retention data

Retention data collections	Data scope	Ref.
Adams' collection	Essential oil components; own measurements; DB-5, 1606 compounds.	[1]
Pacakova and Feltl collection	General; literature data; data for various polarity stationary phases	[2]
Gaskin, MacMillan, GC-library of gibberellins	Gibberellins and related compounds; own data; OV-1; available on the WEB	[3]
Sadtler library	General; own measurements; OV-1, SE-54, CW-20M, CW-20M (cross-linked), approximately 2000 compounds	[8]
Flavornet, Cornell University	Flavor and fragrance compounds (fragrance chemistry); own and literature data; OV-101, DB-5, OV-17, CW-20M; 738 compounds; WEB-based collection	[9]
ASTM collection	General; literature data; retention data on various stationary phases of different polarity	[10]
Collection of the DFG commission for clinical-toxicological analysis	Toxicological compounds; own and literature data; dimethylsilicone stationary phase; 4500 compounds	[11]
Jennings and Shibamoto collection	Flavor and fragrance compounds	[12]
Kondjoyan and Berdague collection	Flavor and fragrance compounds	[13]
Bogoslovsky, Anvaer and Vigdergaus collection	General; literature data	[14]
Vinogradov' collection	Essential oil components; literature data; non-polar and polar stationary phases; approximately 2000 compounds, WEB-based collection	[15]
HortResearch collection	Pheromones; literature data; approximately 2000 compounds, WEB-based collection	[16]
The Golm Metabolome Database	Metabolites, own data, WEB-based collection	[17]
VERIFY Database	Chemical weapon compounds; own data, commercial database	[18]
ESO 2000, Database of Essential Oils	Essential oil components, 1800 compounds, commercial database	[19]
Flavor Database, Citrus Research and Educational Center, Florida State University	Flavor compounds; DB-1, DB-5, DB Wax; WEB-based collection	[20]
CRC Handbook Series on Chromatography	Hydrocarbons, drugs, pesticides, amino acids and amines, terpenoids, literature data	[21]

The web references were valid at the moment of paper preparation. There can be some changes in the provided web addresses by developers of the corresponding databases.

Mass Spectral database (June 2005) [31], as well as on the internet (NIST Chemistry WebBook) [32]. This paper will discuss the details of database formation and content as well as document the methods of data quality control.

2. Development of database of retention properties of chemical compounds

The database of retention properties contains four main elements: a bibliographic database, the gas chromatography retention data, a library of chemical compound names and the data entry system. The library of chemical names includes a very large set of common names as well as the systematic names and chemical structures. This allows the literature data to be associated with the correct chemical even when different names for the same chemical compound can be found in literature. The entry system allows the entry of bibliographic information, retention data and the conditions of measurements including specific column, temperature programs and other relevant variables. In addition, it is possible to use the entry system to assign *I* values based on known values for some compounds in the set even if the data is given only in terms of retention time. The data are subsequently processed for quality control to determine the consistency of the retention data.

2.1. Overall description of data included in the database

The database includes several types of retention measurements: isothermal Kováts indices [4], the non-isothermal indices from temperature-programming measurements in accord with the definition of Van den Dool and Kratz [5], and Lee indices [7] (isothermal and non-isothermal)—all measured on both non-polar and polar stationary phases (Table 2). In addition, data on measurements of absolute and relative retention times (t_R) and retention volumes (V_R) are collected in the database as well as data using non-standard reference compounds.

Isothermal Kováts retention indices are determined by the relationship [4]

$$I_x = 100n + \frac{100[\log(t_x') - \log(t_n')]}{[\log(t_{n+1}') - \log(t_n')]}$$

where t_n' and t_{n+1}' are adjusted retention times of the reference *n*-alkane hydrocarbons eluting immediately before and after compound "X", and t_x is the adjusted retention time of compound "X". Linear indices (non-isothermal indices in accord with the definition of Van den Dool and Kratz [5] from temperature-programming measurements) are defined by the following equation:

$$I_x = 100n + \frac{100(t_x - t_n)}{(t_{n+1} - t_n)},$$

where t_n, t_{n+1} and t_x are net retention times. Lee retention indices are determined by analogy with the linear Kováts indices for the following reference compounds: benzene (assigned index 100), naphthalene (200), phenanthrene (300), chrysene (400) and picene (500). It should be noted that there are relatively minor differences between the retention indices defined for isothermal data (the original Kováts definition) and indices determined in accord with the definition of Van den Dool and Kratz. The Lee indices must be treated independently as are those using any other set of standard compounds.

The collection is limited to gas chromatographic data for chemical compounds measured on widely used non-polar and polar stationary phases. The aim of database development is to provide retention data for comparable, well identified and commercially produced columns, which can be applied in the identification process on widely used stationary phases. The list of stationary phases considered is presented in Table 2. The methylsilicone phase is taken as the standard non-polar phase. In addition, data for slightly polar stationary phases are included. The polyethelene glycol based stationary phases were considered as standard polar phases. Available data demonstrate that retention data for polyethylene glycol stationary phases constitute more than 90% of measurements on different polar phases. Table 3 contains the average values of McReynolds' constants [33] for stationary phases presented in database, which characterize a polarity of stationary phases.

2.2. Sources of gas chromatographic properties. Bibliographic database

Available collections of retention indices are summarized in Table 1. The retention data presented in these collections are of differing quality. Several data collections represent results of authors' measurements and they include results of accurate determinations of retention indices for well-defined conditions [1,8]. Some collections contain the results of measurements of different authors compiled from the literature. It can be seen that several libraries represent the compilation of data for diverse chemical compounds, e.g., [2,8,10]; while most published collections represent compilations for certain chemical classes of compounds or for certain practical applications [1,3,9,11].

Table 2
GC stationary phases presented in the NIST GCRID database

Type of stationary phase	Stationary phase	Trade name
Non-polar	- Dimethylpolysiloxane	OV-101, HP-1, DB-1, SE-30, etc.
	- Methyl silicone, 5% phenyl groups	DB-5, SE-54, HP-5, Ultra-2, etc.
	- Apiezons (L, M, N)	
	- Squalane	
	- Apolan-C87 (24,24-diethyl-19,29-dioctadecylheptatetracontane)	
Polar	Polyethylene glycol based stationary phases (MW > 2000)	Carbowax 20 M, Innowax, CP-Wax 52 CB, etc.

Table 3
McReynolds constants of stationary phases considered in the NIST GCRID database

Stationary phases	McReynold's constants				
	x'	y'	z'	u'	s'
Dimethylpolysiloxane	15 ± 3	56 ± 2	45 ± 1	65 ± 3	43 ± 4
Methyl silicone, 5% phenyl groups	27 ± 7	71 ± 4	64 ± 3	95 ± 5	63 ± 5
Apiezons L, M, N	35 ± 4	22 ± 15	21 ± 7	41 ± 11	49 ± 10
Squalane	0	0	0	0	0
Apolan-C87	21	10	3	12	25
Polyethylene glycol based stationary phases[a]	337 ± 37	580 ± 50	395 ± 45	587 ± 40	558 ± 75

Data are based on publications [2,26,33–35] and were averaged through available data for different phases.
[a] Data included for acid and base modified polyethylene glycol stationary phases.

Our data set covers literature published from 1956 to 2005 and includes both the printed literature and internet sources of retention data. The database contains retention data for various classes of organic compounds including natural products, environmentally important compounds, drugs, food contaminants, etc. The printed literature sources include original journal articles, technical reports, conference proceedings, dissertations and personal contributions. Fig. 1 contains the distribution of sources in the database by a publication year. As can be seen from the figure there has been a steady increase in publications with quantitative retention data in gas chromatography. While every effort has been made to find all of the sources of retention data, it is certain that some have been missed. Note that some of the articles published in 2004–2005 have not yet been reviewed, so there is an apparent decrease in the number of sources in these years. The vast majority of data sources are regular peer-reviewed journal articles (>90%). Internet sources contribute less then 2%.

An archive of sources of gas chromatographic data was created. When possible the electronic version of the paper was used. Currently the collection of sources of gas chromatographic properties of chemical compounds contains more than 7000 sources. The information on the data sources was realized in the form of a bibliographic database with full bibliographic descriptions.

2.3. Database of retention properties

The format of presentation of retention data has been discussed in the literature, e.g., [29]. At the early stage of retention data systematization, ASTM efforts on the organization and compilation of gas chromatographic information led to the formulation of a format for retention data presentation. It was based on a system for compiling published retention data developed by the Gas Chromatography Subcommittee of the Tennessee Eastman Company Analytical Committee [10]. The format for retention data in the NIST database includes many of the same fields used by these earlier works, specifically the data is characterized by the following information:

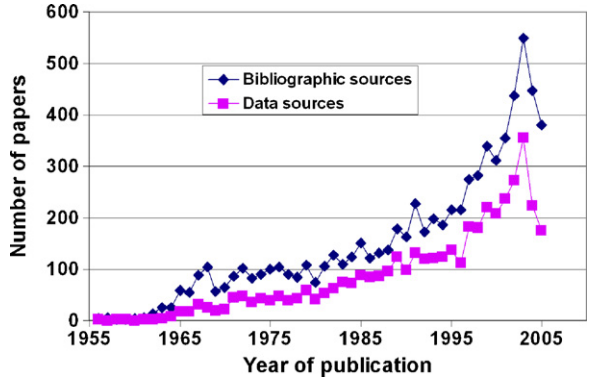

Fig. 1. Number of sources of gas chromatographic data in the NIST GCRID database by year (data for 2004 and 2005 are not complete).

(a) Description of data source (reference to the bibliographic database).
(b) Column specification includes: type of column (capillary or packed), stationary phase specification, column dimensions, stationary phase thickness (capillary column), characterization of solid support (packed column) and carrier gas. A comprehensive list of stationary phases was compiled on the basis of widely used stationary phases and commercially produced columns. It was served as a vocabulary of stationary phases.
(c) Description of temperature conditions include isothermal measurements (temperatures of measurements), linear temperature program (program description) and complex temperature program (program description).
(d) Type of chromatographic retention data (Kováts indices, Lee indices, linear indices, other indices, retention time, retention volume, relative retention).
(e) Description of chemical compound including: the original name of chemical compounds from the data source, the registration number (CAS (Chemical Abstracts Service) Registry number or the unique registration number in the retention database) and the reference to the corresponding structure and synonyms of the chemical name from the library of chemical compound names described below. It was found to be important to keep the original name of chemical compounds as assigned by the author both to avoid possible discrepancies in the defining of chemical compounds from original sources and for quality assurance in the entry system.
(f) Quantitative data of retention properties of chemical compounds. Numerical data on retention properties (retention indices, retention times, retention volumes) are presented in a table form for different temperatures or different tem-

perature programs, and for different columns. Additional fields were used to characterize the data quality (expert estimate of data reliability). Currently the data quality is defined to in one of four categories: reference value; acceptable value; suspicious value; value very likely in error.

2.4. Library of chemical compounds

Both the entry and search programs use a comprehensive library of chemical names and structures. The library contains compound identification information from the NIST/EPA/NIH MS-database [31], NIST Chemistry WebBook [32], and other NIST chemical data projects. The library currently contains the names and structures for approximately 350,000 different chemical compounds. The library of chemical names represents one of the key elements in the database. It allows the search for a chemical compound by name, structure, substructure, molecular formulae and the unique registration number. The detailed description of chemical compound contains the following fields: the list of names (synonyms, trade names, systematic names), structure (represented by a mol-file); molecular formula, molecular weight, IUPAC International Chemical Identifier [36], and a unique registration number of chemical compound. When known, the CAS number is used as the registration number. If the CAS number is not known, a unique registration number is assigned to the chemical compound in the retention database. A new structure is compared to all other structures in the database to ensure each distinct chemical is only entered once.

2.5. Entry program

The data treatment begins with the entry of bibliographic description of the source. Then the data source is analyzed for the presence of retention data and the conditions of the measurements (column, column parameters, temperature conditions, type of retention data). The entry of the conditions of measurements leads to the creation of a template for entry of retention data themselves. The availability of a portable document format (PDF) or hypertext markup language (HTML) data source file allows the automatic preparation of the table of data (name of compound—retention data) in a spreadsheet format through copy-paste operations thus greatly reducing entry errors. After editing the spreadsheet table, its content is copied to the prepared template for this source. Once the data is entered into to the template, the names are automatically searched against the library of chemical compounds discussed above. If any name is not automatically recognized, the interactive process allows the search for this compound under different names or using chemical formulae or structure. If the compound is not found in the library of chemical compounds, a new unique registration number is assigned to it. The data characterizing the new compound (name, structure, registration number) are added to the compound library of the retention database.

3. Data distributions and quality control

A majority of the compounds (53%) in the collection have just one measurement, limiting quality control by comparison of replicate measurements. On the other hand 50% of the measurements belong to only 2.4% of the compounds, all of which have more than 46 replicates. For example, benzene and α-phellandrene are among the most commonly reported compounds. The current database contains 545 data points for benzene and 630 data points for α-phellandrene. Numbers of compounds having different numbers of replicate measurements is given in Fig. 2. Other distributions of this data are described elsewhere [35,37].

An important method for establishing reliability involves the comparison of replicate values. This first step in the process finds and excludes multiple instances of a single measured value. This is not always evident from the reported data. The distribution of replicate I measurements depends on a number of factors, including different conditions of measurements (temperature conditions, columns, etc.), identification errors and measurement errors (random and systematic). This often leads to a non-random distribution of deviations, often with unexplained outliers [35,37,38]. To illustrate, for the case of dimethylpolysiloxane as stationary phase a recent snapshot of the database shows 80,427 data points representing 9722 species after filtering with data quality procedures noted later. These had an average deviation of 10 I units, a median deviation of 4 I units, but a 99th percentile deviation of 91 units. Note that this distribution is a composite of distributions for individual species, some of which may consist of as few as two data points. As examples, for naphthalene with dimethylpolysiloxane stationary phases (127 data records) range from 1090 to 1300, while the span of I data for α-pinene (857 records) for non-polar phase ranges from 890 to 989 with the most of I values (751 records) in the range from 925 to 945.

In spite of the relatively large number of values measured for some compounds, there are no critical reviews of the data as there are for other chemical measurements. This is due in part to the large spread in I values measured for the same compounds

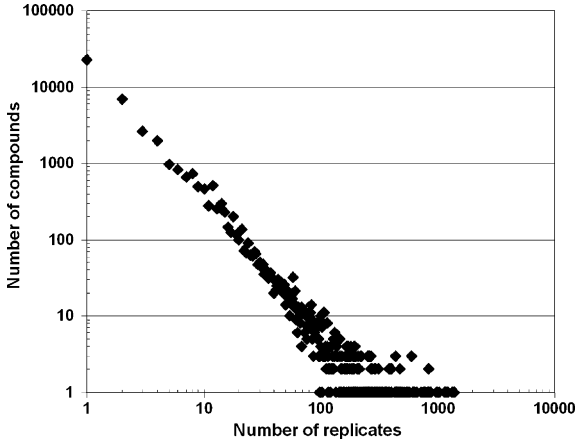

Fig. 2. Distribution of the number of replicate I measurements in accord to the number of compounds.

resulting from both the different conditions of measurements and the effect of some non-controllable experimental conditions (slight variations in the properties of stationary phases, etc.). The lack of measurable parameters that affect the retention indices in predictable ways for these variations hinders data evaluation. Variability of retention data is the main problem in the development of automatic identification procedures that use retention data collections. However, there are clear trends in the data and in many cases the uncertainty in the retention index determination for a single compound is much smaller than the difference between the retention indices of that compound and other compounds which could otherwise be confused with it.

As noted above, multiple measurements of retention indices exist for many chemical compounds. Some compounds exhibit relatively large deviations for *I* determinations performed on equivalent chromatographic columns under similar conditions. There appear to be several causes for the relatively large spread of retention data. One of the causes is incorrect identification of the compound. Where multiple data records exist this can be examined by the construction of histograms (number of measurements—value of retention index) discussed in [35,38]. Histograms showing multiple maxima have been a good indicator of the possible mis-identification of chemical compounds. Other sources of data spread include differing experimental conditions (temperature program, column, etc.), variations in the properties of stationary phase, and the normal experimental errors and data entry errors. Data for chemical compounds with a large spread in their *I* values were analyzed. For this purpose, a standard statistical treatment of retention indices from different sources was conducted with further data analysis. The analysis includes the verification of possible misidentification of chemical compounds [35].

The method described above is very useful when there are multiple records for a single compound and a single phase. However, it is not particularly useful if there are large differences between measured values for a very small number of data for a compound and of no use if there is only a single data point for the compound. Additional quality control of this type of data is provided through the comparison of database values with predicted *I* values. For this purposes the *I* estimations based on a group additivity scheme were used [37]. The increments of different groups of chemical compounds required for these calculations were determined using the database of *I* values for polar and non-polar stationary phases. The possible inaccuracy and errors related to the prediction procedure were taken into account during the comparison process.

In the case of large discrepancies between measured values for the compounds with multiple measurements, and between predicted and measured values, the sources of such discrepancies were analyzed. For the contradictory data, experimental procedures were checked for the use of *n*-alkanes in determination of *I* values, the use of reference compounds, the use of MS identification or other identification procedures, possible problems with derivatization of solutes, thermal stability of compounds under measurement conditions, etc. Additionally, the data review includes the analysis of possible sources in the variability of gas chromatographic retention data (stationary phase, carrier gas, column dimensions, thickness of stationary phase, temperature conditions of measurement, sample size, injection type, column age, etc.). Naturally the data review depends on sufficient documentation of measurements in data sources.

To characterize the retention data a set of quality criteria and corresponding data quality grades were developed. This is currently used for internal purposes of NIST GCRID database development. The following grades of data quality for retention indices are used:

(A) Acceptable data.
  (1) No known problems with data.
  (2) Good agreement with the predicted value.
(B) Suspicious data. The data may exhibit one or more of the following characteristics:
  (1) An uncertain compound assignment indicated.
  (2) The paper contains a description of experimental problems.
  (3) The experimental procedure was insufficiently described.
  (4) Authors used uncommon compounds or samples.
  (5) The presented results are at odds with reference values presented.
  (6) The results of measurements show insufficient agreement between data from different sources for the same compound.
  (7) Comparison with the predicted value of the retention index shows a large discrepancy.
(C) Data very likely in error. The data may be characterized by the following characteristics:
  (1) Data reported deviate considerably from the majority of data from other sources.
  (2) The chemical compound represents an unknown derivative.
  (3) Problems with the experimental measurements (several peaks, overlapping peaks, incomplete peak separation, tails, incomplete derivatization, etc.) as noted in the data source.
  (4) The decomposition of compound was observed.
  (5) Uncertain or erroneous compound assignment.
  (6) The data presented were determined by an unacceptable procedure.

4. The first release of NIST database of gas chromatographic properties

Currently the NIST GCRID contains 292,924 data records (*I*, t_R and V_R values) for 42,888 distinct chemical compounds taken from the original sources published during 1958–2005. The database contains approximately 210,000 data records of experimental retention data for non-polar stationary phases (Table 4). The rest of data represents data obtained for the polar phase (about 73,000 data points for 10,000 compounds). The dataset for polar phase overlaps approximately on 80–85% with dataset for unpolar phase in the presentation of different compounds. Table 5 contains the data distribution according to the type of retention data in the database. As can be seen, most of the

Table 4
Number of retention data measurements performed on different stationary phases

Stationary phase	Number of compounds	Number of data records
Dimethylpolysiloxane	29,942	11,4729
Methyl silicone, 5% phenyl groups	14,636	78,455
Apiezons	3,447	7,030
Squalane	2,311	10,530
Apolan-C87	478	992
Polyethylene glycol based stationary phases	10,025	73,226

Table 5
Types of retention data in the NIST GCRID database

Type of data	Number of compounds	Number of data records
Kováts indices	13,643	67,008
Linear indices (temperature-programming data)	26,110	169,135
Isothermal Lee indices	106	203
Non-isothermal Lee indices	1,322	4,349

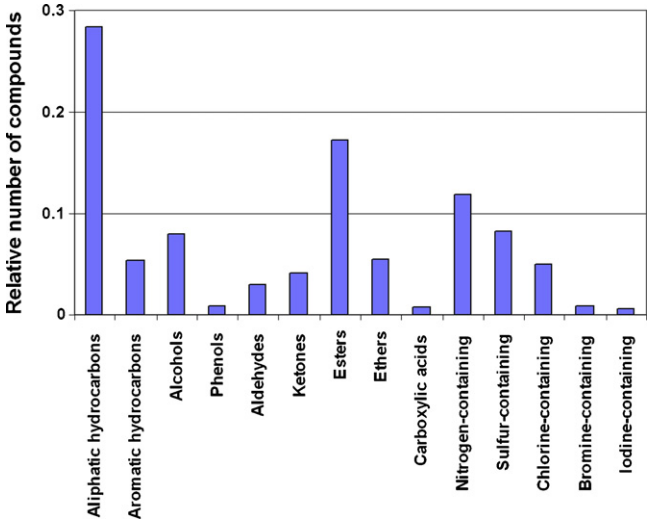

Fig. 3. Relative number of compounds according to compound type in the database.

data represent the measurements of linear indices (temperature programming). Fig. 3 displays a rough picture of the types of compounds in the database. The data presented are based on a partial representative dataset for non-polar stationary phase used as a training set for deriving of group contributions to retention indices [37]. It was assumed that including of overlapping classes does not modify this distribution (number of compounds in overlapping classes constitutes approximately 50%).

The data distribution according the quality grades is shown in Table 6. The June 2005 release of the database is available [31,32]. It includes retention data for 27,948 compounds (125,610 data records) on non-polar and slightly polar phases accumulated before 2005 with "acceptable" grade of data quality. The data included are the isothermal and non-isothermal measurements of Kováts and Lee retention indices. Below a short description of the features of realizations is presented.

The NIST/EPA/NIH MS database [31] includes *I* estimations made with the group additivity procedure [37] in addition to the experimental results of retention index measurements. The estimates use the same group additivity algorithms developed for quality control [37] using the groups that had been identified as useful in estimating boiling points [39]. In the case of the MS database there are mass spectra for over 163,000 chemicals, but only 14,680 of them have experimental *I* data from the new database. The estimation tool allows for a far larger subset of the MS database to be assigned retention indices. The availability of retention indices can increase the identification reliability made by matching mass spectra in the library. The retention indices can also be used as a stand-alone database to assist in confirming the compound identity.

The Internet version of the *I* database [32] includes all of the experimental *I* data in the NIST-EPA-NIH mass-spectra database version. However, the web release does not contain the estimated data. The database is available through the NIST Chemistry WebBook (http://webbook.nist.gov/chemistry). The search options include searching retention indices by compound name, CAS registration number, molecular formula, author of *I* data source, structure, and molecular weight. Some options of the use of the "wild card" or pattern search for the chemical name search are available. In addition, exact and substructure searching of the data is possible in the WebBook version.

5. Summary

(1) A large collection of gas phase chromatographic retention data containing over 7000 data sources was created. The database contains over 292,924 records with retention data for over 42,888 chemical compounds for non-polar (mostly methylsilicone) and polar (polyethylene glycol) stationary phases.

Table 6
I data characterization

	Number of data records	Suspicious data records	Data very likely in error
Non-polar phase, June 2005	125,610	1672	4334
Polar phase	67,425	1007	3413
	Data sources	Sources with more than 30 % of suspicious records	
I data sources, June 2005	3276	121	

(2) A database for digital storage of retention properties of chemical compounds and to aid in the identification process was developed. The record structure (format) of retention data presentation and the corresponding vocabularies were developed.
(3) In addition to the numeric data for the retention information, the database included an extensive bibliographic database of retention data sources, and a chemical identification compound library, which includes the names of compounds, the structure, and unique registration numbers for compounds.
(4) Procedures for initial control of retention data quality were developed. These include: analysis of *I* value distributions for each compound, the comparison of predicted retention indices with database values, and the control of chemical compound structures and the corresponding assignment of retention indices.
(5) The first release of the database of retention properties is packaged with the NIST/EPA/NIH MS database (release of June 2005) and is also available through the internet (NIST Chemistry WebBook).

References

[1] R.P. Adams, Identification of Essential Oil Components by Gas Chromatography/Quadruple Mass Spectrometry, third ed., Allured Publ., Carol Stream, IL, 2001.
[2] V. Pacakova, L. Feltl, Chromatographic Retention Indices—An Aid to Identification of Organic Compounds, Ellis Horwood, Chichester, 1992.
[3] P. Gaskin, J. MacMillan, GS–MS of Gibberellins and Related Compounds: Methodology and a Library of Spectra, University of Bristol, Bristol, 1991, http://www.plant-hormones.info/gibberellins.htm.
[4] E. Kováts, Helv. Chim. Acta 41 (1958) 1915.
[5] H. van den Dool, P.D. Kratz, J. Chromatogr. 11 (1963) 463.
[6] M.V. Budahegyi, E.R. Lombosi, T.S. Lombosi, S.Y. Meszaros, Sz. Nyiredy, G. Tarjan, I. Timar, J.M. Takacs, J. Chromatogr. 271 (1983) 213.
[7] M.L. Lee, D.L. Vassilaros, C.M. White, M. Novotny, Anal. Chem. 51 (1979) 768.
[8] The Sadtler Standard Gas Chromatography Retention Index Library, vols. 1–4, Sadtler Division, Bio-Rad Laboratories, Philadelphia, PA, 1986.
[9] T.E. Acree, H. Arn, Flavornet, Cornell University, NY, 2004, http://www.flavornet.org.
[10] O.E. Schupp, III, J.S. Lewis (Eds.), Gas Chromatographic Data Compilation. Supplement 1, AMD 25A S1, American Society for Testing and Materials (ASTM), Philadelphia, PA, 1971.
[11] Gas Chromatographic Retention Indices of Toxicologically Relevant Substances on Packed or Capillary Columns with dimethylsilicone Stationary Phases. Report XVIII of the DFG Commission for Clinical-Toxicological Analysis, VCH, Weinheim, third ed., 1992.
[12] W.G. Jennings, T. Shibamoto, Quantative Analysis of Flavor and Fragrance Volatiles by Glass Capillary GC, Academic Press, New York, 1980.
[13] N. Kondjoyan, J.L. Berdague, A Compilation of Relative Retention Indices for the Analysis of Aromatic Compounds, Laboratorie Flavour, Theix, France, 1996.
[14] Yu.N. Bogoslovsky, B.N. Anvaer, M.S. Vigdergaus, Chromatographic Constants in Gas Chromatography—Hydrocarbons and O-Containing Compounds, Standards Publ. House, Moscow, 1978 (in Russian).
[15] B. Vinogradov, Retention Indices of Essential Oil Components, 2004, http://viness.narod.ru/, http://viness.narod.ru/ret_ind.htm (in Russian).
[16] A. El-Sayed, Pherobase, HortResearch, Lincoln, New Zealand, 2005, http://www.pherobase.com, http://www.pherobase.com/database/kovats/kovats-index.php.
[17] The Golm Metabolome Database, Mass Spectra & Retention Time Index Libraries, 2005, http://csbdb.mpimp-golm.mpg.de.
[18] VERIFY–Reference Database for Chemical Disarmament, http://www.helsinki.fi/verifin/.
[19] The Complete Database of Essential Oils, BACIS, ESO 2000, 2006, http://www.leffingwell.com/baciseso.htm.
[20] Flavor Database, Citrus Research and Educational Center, Florida State University, 2002, http://www.crec.ifas.ufl.edu/crec_websites/Rouseff/index.htm#.
[21] CRC Handbook of Chromatography (series of separate volumes: Pesticides, Hydrocarbons, Drugs, Amino Acids and Amines, Steroids, Carbohydrates, Terpenoids), CRC Press, Boca Raton, FL, 1980–1992.
[22] K. Pfleger, H. Maurer, A. Weber, Mass Spectral and GC Data of Drugs, Poisons, and Their Metabolites, Parts I–III, VCH, Weinheim, 2000.
[23] M. Rychlic, P. Schieberle, W. Grosch, Compilation of Odor Thresholds, Odor Qualities and Retention Indices of Key Food Odorants, Institute of Food Chemistry, Technical University of Munich, Munich, 1998.
[24] A.F. Shlyakhov, Gas Chromatography in Organic Geochemistry, Moscow, 1984 (in Russian).
[25] S.A. Rang, A.R. Orav, K.R. Kuningas, A.E. Meister, T.V. Strense, O.G. Eisen, Gas–Chromatographic Characteristics of Unsaturated Hydrocarbons, Academy of Sciences of Estonia, Tallinn, 1988.
[26] W.O. McReynolds, Gas Chromatographic Retention Data, fifth printing, Preston, Niles, IL, 1987.
[27] D.H. Hochmuth, D. Joulain, W.A. Konig, Massfinder Software and Data Bank, University of Hamburg, Hamburg, 2004, http://scientific-consulting.net/Terpenoids.pdf.
[28] Pro-ezGC-Thermodynamic Retention Index Database, Analytical Innovation, Beavercreek, OH, 1999, http://www.aii-usa.com/proezgc.htm.
[29] G. Tarjan, Sz. Nyiredy, M. Gyor, E.R. Lombosi, T.S. Lombosi, M.V. Budahegyi, S.Y. Meszaros, J.M. Takacs, J. Chromatogr. 472 (1989) 1.
[30] J.K. Klassen, S.E. Stein, I.G. Zenkevich, presented at the 23rd International Symposium on Capillary Chromatography, Riva del Garda, June 2000, Abstracts, Rep. A18 (CD-ROM).
[31] P. Ausloos, C.L. Clifton, S.G. Lias, A.I. Mikaya, S.E. Stein, D.V. Tchekhovskoi, O.D. Sparkman, V. Zaikin, D. Zhu, J. Am. Soc. Mass Spectrom. 10 (1999) 10, 287–299 (NIST Standard Reference Database 1A. NIST/EPA/NIH Mass Spectral Library with Search Program. Data Version: NIST 05, 2005).
[32] P.J. Linstrom, W.G. Mallard, NIST Chemistry WebBook, NIST Standard Reference Database Number 69, National Institute of Standards and Technology, Gaithersburg, MD, June 2005, http://webbook.nist.gov/chemistry.
[33] W.O. McReynolds, J. Chromatogr. Sci. 8 (1970) 685.
[34] The Retention Index System in Gas Chromatography: McReynolds Constants, Supelco Bulletin 880, Sigma–Aldrich, Bellefonte, PA, 1997.
[35] I.G. Zenkevich, V.I. Babushok, P.J. Linstrom, S.E. Stein, Presented at the 28th International Symposium on Capillary Chromatography, Las Vegas, NV, May 2005, Abstracts (CD-ROM).
[36] S.E. Stein, S.R. Heller, D.V. Tchekhovskoi, Open Standards for Chemical Information—The IUPAC Chemical Identifier and Data Dictionary Projects, Abstracts, Am. Chem. Soc. 226: U304 082-CINF, Part 1, 2003.
[37] S.E. Stein, V.I. Babushok, R.L. Brown, P.J. Linstrom, J. Chem. Inf. Mod., doi:10.1021/CI600548y, in press.
[38] V.I. Babushok, P.J. Linstrom, R.L. Brown, I.G. Zenkevich, S.E. Stein, Presented at the 29th International Symposium on Capillary Chromatography, Riva del Garda, June 2006, Abstracts, Rep. A16 (CD-ROM).
[39] S.E. Stein, R.L. Brown, J. Chem. Inf. Sci. 34 (1994) 581.