RJC

RASĀYAN *J. Chem.*
Vol. 10 | No. 1 |1 - 5 | January - March | 2017
ISSN: 0974-1496 | e-ISSN: 0976-0083 | CODEN: RJCABP
http://www.rasayanjournal.com
http://www.rasayanjournal.co.in

UNVEILING LITERATURE DATA: CHEMICAL, PHYSICAL AND BIOLOGICAL PROPERTIES IDENTIFIER (CPBI)

João P. Leal[1,2,*] and Rui C. Santos[2]

[1]Centro de Ciências e Tecnologias Nucleares (C[2]TN), Instituto Superior Técnico, Universidade de Lisboa, Campus Tecnológico e Nuclear, Estrada Nacional 10, ao km 139.7, 2695-066 Bobadela LRS, Portugal

[2]Centro de Química e Bioquímica, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal

*E-mail : jpleal@ctn.tecnico.ulisboa.pt

ABSTRACT

In this contribution an identifier called CPBI, which makes the access to properties values of chemical compounds easy and reliable, is proposed. It can be included in all published papers without any surcharge of costs to the magazines and authors. A magazine that wants to include this identifier need not enter into agreements with any entity. With this identifier, it will be simple to anyone retrieve published properties values and include them in any database. People can keep using the same data mining program they use previously or even use a search engine (Chrome, IE, and Firefox) to look for properties in the net. This will represent an added value for authors and editors, assuring a reliable way for retrieving properties values published in literature.

Keywords: Chemical Properties, Data mining, Internet Identifier, Physical Properties.

INTRODUCTION

The number of scientific publications as well as the number of new compounds found or synthesized has increased significantly in recent years. According to Chemical Abstracts Service there are currently more than 100 million known substances of which 10 million were added in less than one year, corresponding to about 15,000 new substances in each day.[1,2] The experimental properties determination, however, could not set the same pace. Thus, it is crucial to develop methods to predict chemical, physical and when possible biological properties in a reliable way. In last 40 years, Chemoinformatics became a science in its own right making decisive contributions to the development of chemistry despite the fact that there are still many problems to be solved in the area of chemoinformatics.[3] In any case, the quality of the predicted data relies mainly on the size and quality of existing experimental databases. Therefore, it is of top importance to extract all the information available in the literature. This task is not simple, even if access to the scientific literature is granted, since properties values are published in text, in tables and sometimes even in figures and most of the time in an unstructured way. An attempt to archive and exchange published experimental thermophysical and thermochemical property data as a unique file by article started in 2003. The files are posted online through cooperation between the Thermodynamics Research Center (TRC) at the National Institute of Standards and Technology (NIST) and major journal publishers in the field.[4] The results are presented in an Extensible Markup Language (XML)-based IUPAC standard[5,6] (that is open to public consultation), covers some properties and can be accessed with the programs supplied by the authors or provided by the journals publishers. Instead of a similar approach (create a database), the aim of this contribution is to propose an identifier that can be included in all published papers that, in a structured approach, make the access to properties data easy and reliable. With this identifier it will be simple to anyone, anywhere, retrieve published data and include it in their own databases.

http://dx.doi.org/10.7324/RJC.2017.1011527

The CPBI Identifier

The first step in such an identifier is to define a way for identify the molecule to which the property is assigned. The task of identify molecules in a proper way is an old problem but is in fact solved with the introduction of the IUPAC International Chemical Identifier (InChI). The InChI is a hierarchical, extensible, scalable and understandable unique identifier for chemical substances developed by IUPAC.[7] Since 2010, the development of the standard has been supported by InChI Trust, a not-for-profit member-owned charitable organization, from which IUPAC is a member.[8] The current version is 1.04 and was released in September 2011. It differs from the widely used CAS registry number in three aspects: the format and algorithms are non-proprietary; it can be computed from structural information; and do not have to be assigned by some organization. In addition, with thorough practice, most of the information in an InChI is human readable. Nevertheless, the full InChI turned out to be too lengthy for easy searching, and for that reason InChIKey was developed.

The Standard InChIKey, sometimes referred to as a hashed InChI, is a fixed length (27 characters) condensed digital representation of the InChI that is not human-understandable. The InChIKey specification was released in September 2007,[9] in order to facilitate web searches for chemical compounds, since these were problematic with the full-length InChI.[10] The Standard InChIKey consist of 14 characters resulting from a hash of the connectivity information of the InChI, followed by a hyphen, followed by 9 characters resulting from a hash of the remaining layers of the InChI, followed by a single character indication of the version of InChI used, another hyphen, followed by single checksum character. It should be noted that, unlike the InChI, the InChIKey is not unique.[11,12] There is a nonzero chance of two different molecules having the same InChIKey, but this chance is extremely small. In a recent publication the collision rate was compared with the theoretical expectations and concluded that InChIKey is adequate for use because considered the size of the involved actual and near future databases the chances of collision are negligible.[13] To determine the InChI and the InChIKey of a molecule, some programs are available, as for example the one developed by InChI Trust[14] or the PubChem website.[15] As an example, for the acetyl salicylic acid (Aspirin) those codes are shown in Figure-1.

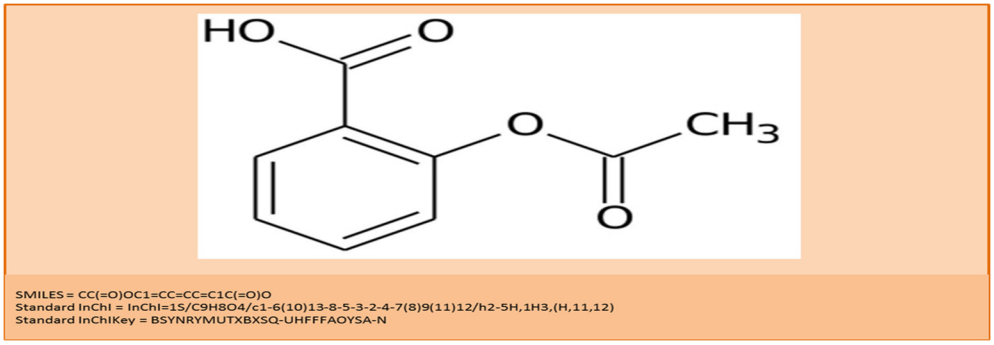

Fig.-1 : SMILE, Standard InChI, and Standard InChIKey for acetyl salicylic acid.

So, by using InChIKey for the identification of the molecule an identifier called "Chemical, Physical and Biological properties Identifier", or CPBI, is proposed with the structure shown in Figure-2.

CPBI = Standard InChIKey/Property1,Attribute1,Attribute2, Attribute3,Attribute4(Value,Error)/Property2, ...

Fig.-2: Structure of the CPBI.

For each property, a four-letter code will be used (e.g., MHFG for molar enthalpy of formation in the gas phase). All possible combinations for a 26 letter alphabet are over 450,000 (not considering numerals), but since for the sake of interpretation and readability the codes should be related with the property they describe, far less useful combinations will be useful. Nevertheless, they will be far enough for the desired properties. Attributes (up to four) for each property are meant to state the conditions where the value is

determined. The first attribute indicates whether the value is an experimental (EXP), computationally calculated (CPC), or a prediction or estimate (EST). The second to fourth ones will state conditions that influence the property (e.g., for solubility the second attribute will be the temperature at which it was measured and the third one the solvent in which the compound was dissolved). In some cases, the second, third or fourth attributes were not necessary. When this happens, their place will be left blank. In fact, any unused attribute can be used as a comment or an additional clarification. For example, orientation can be included as an additional parameter in solid phase thermal conductivity (THCS) for taking into account anisotropy. These flexibility and simplicity are one of the strengths of CPBI. Finally, the identifier will include the value and the associated error of the property between brackets.

For the property units the international system of units (SI) will be preferentially used. So, all temperatures will be in kelvins, enthalpies in joules (or kilojoules), pressure in pascal (and not bar or atm) and so on. This is not a really limitation, since in almost all science journals SI units are already used. This also applies to attributes whenever they are numerical (e.g., temperature values). For the sake of clarity and usability some of the units must not be the canonical SI units. This is the case of molar formation enthalpies (where kilojoules per mole will be used). An exception will be the dipole moment in the gas phase, were Debye will be used as unit (since indicate them in C.m will be not practical). In every case the involved units were defined at the same time the four letter property code is defined (see Table-1, below) so all the process is simple, well defined and transparent.

To make clear the use of this identifier, some examples are presented in equations (1–3). If someone measured the melting point of acetyl salicylic acid as 136 ± 1 °C it should include the following line:

CPBI = BSYNRYMUTXBXSQ-UHFFFAOYSA-N/MEPT,EXP,1.013E5,(409,1) (1)

and, if estimates the corresponding enthalpy of formation in the solid phase as -815.6 ± 1.4 kJ mol-1, it should write:

CPBI = BSYNRYMUTXBXSQ-UHFFFAOYSA-N/MHFS,EST,298.15,(-815.6,1.4) (2)

Of course, the two lines referring to the same compound can be agglutinated in a single one since the identifier can include various properties for the same compound separated by a slash sign.

CPBI = BSYNRYMUTXBXSQ-UHFFFAOYSA-N/MEPT,EXP,1.013E5,(409,1)/MHFS,EST,298.15,(-815.6,1.4) (3)

A list of proposed codes and attributes for properties is presented in Table-1. A webpage was created in order to maintain a unique list of codes and attributes (http://www.cpbi.ctn.tecnico.ulisboa.pt). Any property not included in those tables (and it should be a huge number) can easily be added as long as new codes can be defined. For that purpose, and to keep the uniqueness of the codes, someone needing a new code should contact the authors and propose the code and attributes as explained in detail in the aforementioned webpage.

CONCLUSION

With this identifier, it will be extremely easy to anyone to retrieve published data and include them in a database. It will be also an asset for the authors and editors that decide to use it, because their data will be more easily accessed. To the publishers it will not force them to have any special agreement with anyone or change anything in their publication structure. The identifier can be included in any place of the paper (e.g. after the keywords or near the acknowledgements). They only have to decide to include it and ask authors to add the identifier in their papers. For the authors they only have to generate the InChIKey code of the measured compounds (as explained in the Introduction) and insert the identifier(s) in their publication.It is not possible to change the past. Therefore, for the past values published in literature, researchers must continue to look for them in the best possible way. But we can assure that in the future,

with only a small effort made by any author, it will be possible to retrieve in an easy and reliable way all the properties values published in literature. This is a huge opportunity we should not disregard.

Table-1: Properties List with corresponding codes and attributes

Property	Property Code	Attribute 2	Attribute 3	Attribute 4	Units
Boiling Point	BOPT	Pressure			K
Melting Point	MEPT	Pressure			K
Triple Point Temperature	TRPT				K
Triple Point Pressure	TRPP				Pa
Refractive Index of gas phase	RFIG	Temperature	Pressure		
Refractive Index of liquid phase	RFIL	Temperature			
Refractive Index of solid phase	RFIS	Temperature	Phase		
Density of gas phase	DENG	Temperature			Kg m^{-3}
Density of liquid phase	DENL	Temperature			Kg m^{-3}
Density of solid phase	DENS	Temperature	Pressure	Phase	Kg m^{-3}
Electrical Resistivity of gas phase	ELRG	Temperature			Ω m
Electrical Resistivity of liquid phase	ELRL	Temperature			Ω m
Electrical Resistivity of solid phase	ELRS	Temperature			Ω m
Surface Tension of liquid	STEL	Temperature			N m^{-1}
Vapor Pressure of liquid	VAPL	Temperature			Pa ≡ N m^{-2}
Second Virial Coefficient	SVCF				dm^3 mol^{-1}
Flash Point	FLPT	Pressure			K
Autoignition Temperature	AIGT	Pressure			K
Thermal Conductivity in gas phase	THCG	Temperature			W m^{-1} K^{-1}
Thermal Conductivity in liquid phase	THCL	Temperature			W m^{-1} K^{-1}
Thermal Conductivity in solid phase	THCS	Temperature	Phase		W m^{-1} K^{-1}
Viscosity in gas phase	VISG	Temperature			N s m^{-2}
Viscosity in liquid phase	VISL	Temperature			N s m^{-2}
Molar Enthalpy of Formation in gas phase	MHFG	Temperature			kJ mol^{-1}
Molar Enthalpy of Formation in liquid phase	MHFL	Temperature			kJ mol^{-1}
Molar Enthalpy of Formation in solid phase	MHFS	Temperature	Phase		kJ mol^{-1}
Molar Enthalpy of Vaporization	MHVP	Temperature			kJ mol^{-1}
Molar Enthalpy of Sublimation	MHSB	Temperature	Phase		kJ mol^{-1}
Molar Gibbs Energy of Formation in gas phase	MGFG	Temperature			kJ mol^{-1}
Molar Gibbs Energy of Formation in liquid phase	MGFL	Temperature			kJ mol^{-1}
Molar Gibbs Energy of Formation in solid phase	MGFS	Temperature	Phase		kJ mol^{-1}
Molar Gibbs Energy of Vaporization	MGVP	Temperature			kJ mol^{-1}
Molar Gibbs Energy of Sublimation	MGSB	Temperature	Phase		kJ mol^{-1}
Molar Entropy in the gas phase	MENG	Temperature			J K^{-1} mol^{-1}
Molar Entropy in the liquid phase	MENL	Temperature			J K^{-1} mol^{-1}
Molar Entropy in the solid phase	MENS	Temperature	Phase		J K^{-1} mol^{-1}
Molar Heat Capacity at constant pressure in gas phase	MCPG	Temperature			J K^{-1} mol^{-1}
Molar Heat Capacity at constant pressure in liquid phase	MCPL	Temperature			J K^{-1} mol^{-1}

Molar Heat Capacity at constant pressure in solid phase	MCPS	Temperature			$J\ K^{-1}\ mol^{-1}$
Polarizability	POLY				$A^2 \cdot s^4 \cdot kg^{-1}$
Molar Ionization Potential in gas phase	MIPG				$kJ\ mol^{-1}$
Molar Electron Affinity in gas phase	MEAG				$kJ\ mol^{-1}$
Dipole Moment in gas phase	DIPM				Debye
Molar Magnetic Susceptibility	MMGS				$m^3\ mol^{-1}$
Lower Flammability Limit	LFLT	Temperature	Pressure		
Standard Reduction Potentials	STRP	Temperature			V
Solubility in mole fraction	SOLM	Temperature	Solvent		
Median Lethal Dose	LD50	Time			$kg\ kg^{-1}$

REFERENCES

1. Chemical Abstracts Service, http://www.cas.org/
2. CAS Statistical Summary 1907-1997, Chemical Abstracts Service, Columbus: Ohio; http://www.shinwon.co.kr/cas/ASSETS/casstats.pdf
3. J. Gasteiger, *SAR and QSAR in Environmental Research,* **25**, 443 (2014)
4. ThermoML Representation of Published Experimental Data; http://trc.nist.gov/ThermoML.html
5. M. Frenkel, R.D. Chiroco, V. Diky, Q. Dong, K.N. Marsh, J.H. Dymond, W.A. Wakeham, S.E. Stein, E. Königsberger, A.R.H. Goodwin, *Pure Appl. Chem.,* **78**, 541 (2006)
6. M. Frenkel, R. D. Chirico, V. Diky, P.L. Brown, J.H. Dymond, R.N. Goldberg, A. R.H. Goodwin, H. Heerklotz, E. Königsberger, J.E. Ladbury, K.N. Marsh, D.P. Remeta, S.E. Stein, W.A. Wakeham, P.A. Williams, *Pure Appl. Chem.,* **83**, 1937 (2011)\
7. The IUPAC International Chemical Identifier (InChI); http://www.iupac.org/home/publications/e-resources/inchi.html
8. InChI Trust – InChI: open-source chemical structure representation algorithm; http://www.inchi-trust.org/
9. Release of InChl Version 1.02 beta; Introducing InChlKey. Chem. Int. 29 (2007). http://old.iupac.org/publications/ci/2007/2906/iw5_inchikey.html
10. The IUPAC InChIKey; https://www.youtube.com/watch?v=UxSNOtv8Rjw
11. E. L. Willighagen, InChIKey collision case (17 September 2011); http://chem-bla-ics.blogspot.nl/2011/09/inchikey-collision-diy-copypastables.html
12. A. Williams, An InChIkey Collision is Discovered and NOT Based on Stereochemistry (1 September 2011); http://www.chemconnector.com/2011/09/01/an-inchikey-collision-is-discovered-and-not-based-on-stereochemistry
13. I. Pletnev, A. Erin, A. McNaught, K. Blinov, D. Tchekhovskoi, S. Heller, *J. Cheminformatics*, **4**, 39 (2012)
14. InChI Software Downloads (Executables and Documentation); http://www.inchi-trust.org/downloads/
15. Pubchem – Compounds identification; https://pubchem.ncbi.nlm.nih.gov/

[RJC-1527/2016]