OPEN ACCESS

EURASIA Journal of Mathematics Science and Technology Education
ISSN: 1305-8223 (online) 1305-8215 (print)
2017 13(10):6533-6543
DOI: 10.12973/ejmste/78095

Analysis of Ambiguous Information about Chemical Compounds in Online Databases

Martin Rusek
Charles University, Fakulty of Education, Department of Chemistry and Chemistry Education, Prague, CZECH REPUBLIC
Antonín Jančařík
Charles University, Fakulty of Education, Department of Mathematics and Mathematical Education, Prague, CZECH REPUBLIC

Received 20 April 2017 ▪ Revised 21 August 2017 ▪ Accepted 16 September 2017

ABSTRACT
The Wolfram applications enable direct access to several main information sources (databases) and enable the data to be analysed easily. This offers developing information literacy on authentic material in chemistry education. For chemical compound identification: CAS, CID and Beilstein Numbers are used. In this paper, the authors focused on the ambiguousness of carbohydrate identification using these numbers. Altogether, there are 1498 compounds listed under the CAS Number, whereas CID and Beilstein Numbers are not assigned to every compound with CAS. During the analysis, several hundred entries with duplicate records were found, providing ambiguous information about the compounds. The authors focus on ambiguous database records, further analyse the physical properties of two compounds with several assigned identification numbers. This approach offers a suggestion on how to work with this topic in education, offering fruitful educational content.

Keywords: chemical compounds, big data, database search, ambiguities

INTRODUCTION

To prepare students for their integration into the contemporary information society, modern technologies should be integrated into individual school subjects. Separating ICT has been proven ineffective e.g. by the last ICILS (International Computer and Information Literacy Study), where the students with the best scores were reported not to be due to ICT education (Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014). It seems that making ICT a natural part of every school subject would be more effective. When considering the role of technology more broadly in science education Bell, Gess-Newsome, and Luft (2008, p. 4) argue that "much of the value... can be found in its capability to allow students to work with data, to enhance visualization of complex concepts and to facilitate communication and collaboration. The students encounter real data from a particular field which require certain skills in order to be processed. The data then not only serve as natural subject matter, processing skills come naturally too.

Chemical data are not an exception. The question how to find desired data and how to verify their accuracy rises. This skill has become one of the crucial skills of a responsible member of contemporary society. With the massive development of the Internet and the time students spend on it daily, school education needs to focus on using the Internet too.

Correspondence: Martin Rusek, *Charles University, Fakulty of Education, Department of Chemistry and Chemistry Education, Magdaleny Rettigove 4, Prague, Czech Republic.*
martin.rusek@pedf.cuni.cz

Contribution of this paper to the literature

- The paper provides an example of possible use for databases of chemical compounds in education.
- The example is given on the database of organic compounds, which contains several duplicities and also errors. They can serve as a motivational problem for students.

This paper is another contribution to the topic on integrating Mathematics and Science education (see Kim & Aktan, 2014). The authors focus on using data primarily intended for mathematical processing in the Wolfram Alpha internet application, which draws the data from several chemical databases and the Wolfram Mathematica app, designed for mathematical calculations. The authors of this paper chose these applications as they are a suitable tool for further calculations with the obtained data and are also used in education.

Connecting education with mobile technology moves the whole process into another dimension. Besides formal education, informal and non-formal education are distinguished (see Filippoupoliti & Koliopoulos, 2014; Greenhow & Lewin, 2016) as the process of learning enhanced by mobile technology, no longer depending on a school classroom. Nevertheless, even at school students can, almost instantly, gain a view of almost all compounds' basic physical and chemical properties. There is then more time to have students engaged in tasks based on more difficult cognitive processes.

One new educational goal is to teach students to orientate themselves in the maze of information. Its reliability and accuracy as well as the ethics behind the use of the information is very important.

Chemistry-oriented proven data are collected by several databases accessible either online via special portal or via special programmes such as Wolfram Alpha. The databases contain inexhaustible amount of information. In this study, the authors chose the topic of alkanes. When searching for a certain compound by its name, several entries are usually found, identified by various identification numbers. This fact may, not only in education, cause misunderstanding, sometimes even students' demotivation. The authors of this text therefore decided to map the database entries and analyse some discrepancies.

THEORETICAL BACKGROUND

The aim of science education may be seen in the development of scientific literacy, "the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen" (OECD, 2016). It is evident frontal education is unusable to achieve these goals as learner activity is crucial. There are several solutions to promote education in order to develop students' scientific literacy. Among others is the use of complex tasks, experiments (Anonymous, 2015), laboratory works, project-based education or inquiry-based education (Anonymous, 2011).

The potential of information and communication technology (ICT) cannot be ignored (see e.g. Král & Řezníčková, 2013). With enough evidence of its efficacy, teachers seem to be likely to adopt ICT in their education (see Anonymous, 2017). The most developing field – mobile technology – should not be excluded at schools either (Libman & Huang, 2013; Sha, Looi, Chen, & Zhang, 2012). Apart from native functions such as calculator, camera GPS, etc., they offer to run apps which widen the usability of an ordinary smartphone and enable a various range of operations with it. ICT is inherently associated with the Internet. Connectivity is the trend of modern time moving the place of education into a different position. E-learning (more precisely Massive Open Online Courses), instructional videos – separate or under a movement such as the Khan Academy – attract more and more learners. The movement of users on the Internet generates so called big data (Pence & Williams, 2016). They find use not only in marketing, financial engineering or advertising, they may be used in education as well.

Chemistry education does not stand aside in this process. Chemistry as a scientific discipline has developed rapidly in the last three decades. Not only have new elements been discovered and the periodic table of elements filled, the number of discovered compounds grows exponentially. The selection of the core subject matter is therefore more challenging than any time before. In this respect, ICT enables an interesting alternative to

searching for information about various compounds (Lebedeva & Zaitseva, 2014). This is the field the authors of this text focus on in this paper.

Despite IUPAC (International Union of Pure and Applied Chemistry) periodically issues recommendations for chemical compounds' nomenclature, naming especially the more complicated compounds is sometimes ambiguous (trivial names, systematic names, semi-trivial names). The rules for naming more complex compounds are not strict and are being applied differently. For this reason, an identification number is being used for these purposes. There are several such numbers: CID (PubChem Compound Identidication), CAS (Chemical AbstractsServices) or the Beilstein number. In various databases, an overview of a compound's properties, both empirical and structural formulas, boiling point, melting point etc. can be found. This also offers new possibilities for education. The ability to use these databases, which contain links to compounds, seems to be one of the modern aims of chemistry education (cp. Pence and Williams, 2010).

When analysing compounds and their indication in the databases, certain variances can be found. They are the subject of this text. A database entry in schools is offered, for example, via a Wolfram Mathematika program (Abramovich, 2014; Lebedeva & Zaitseva, 2014; Weisstein, 2014). It is a Computer Algebra System which, as well as Wolfram Alpha, allows access to all sorts of databases, including chemical databases. To be precise, 24 significant chemical databases are used in order to get data about chemical compounds.

Except for general information about a particular compound, compounds' formulas as well as molecule models are available. These visual aids can improve the efficacy of education (cp. Eilks, Wittec, & Pietzner, 2009; Anonymous, 2017).

The Mathematika program is being used mostly at technical universities. For primary and secondary school education, applets programmed in this software are used. They are available within the Wolfram demonstration project. This project offers contemporarily more than 10 000 educational materials, out of which more than 800 are focused on chemistry education. All these applets are free to use by both teachers and students via a program called Wolfram CDF Player.

AIMS AND METHODS

The primary impulse which led to a further literature review were data acquired via the Wolfram Alpha program. This Computational knowledge engine can search for data and further perform calculations, see **Figure 1**.

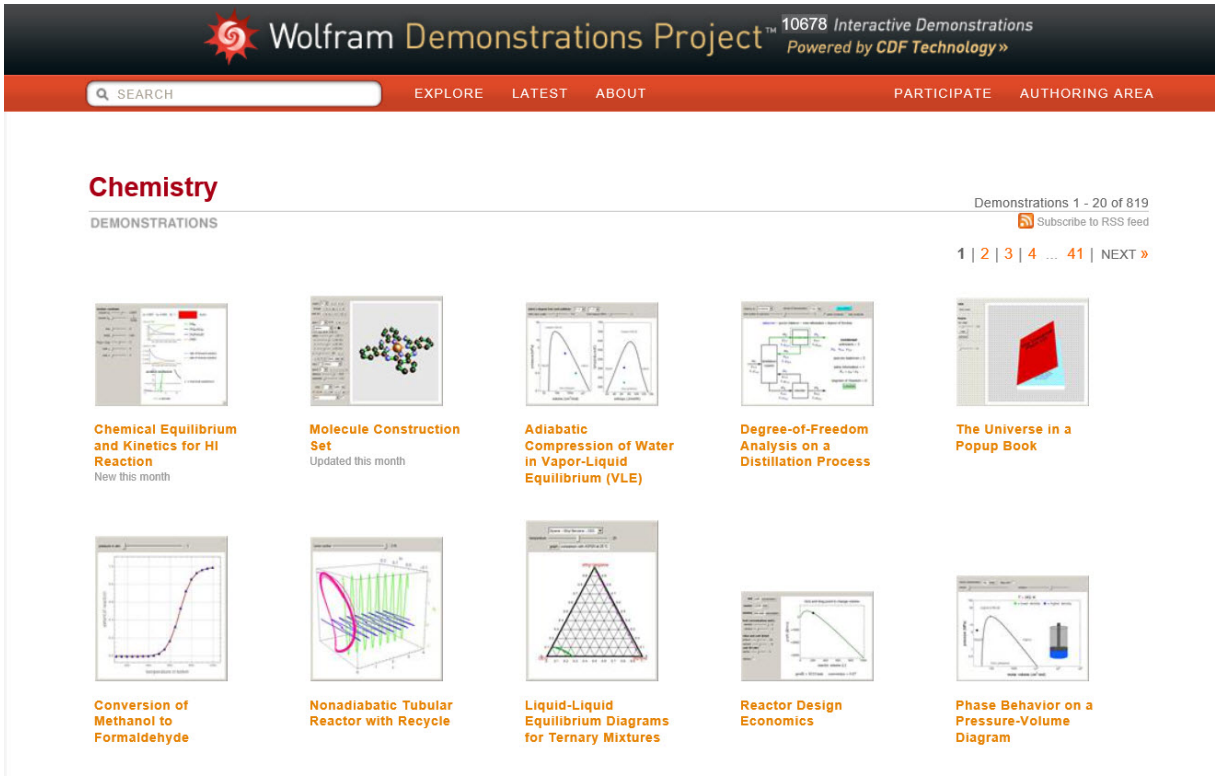

Figure 1. Wolfram Demonstration Project

Because of that it has become a popular tool for Mathematics as well as Science education (Abramovich, 2014; Ersoy & Akbulut, 2014). When working with the program, the authors realised several duplicities. Therefore, they focused on one part of organic chemistry compounds – hydrocarbons. For this group of chemical compounds, the program offers two undistinguishable formulas within isomers. After a closer look, the authors discovered that the reason is the results represent a joint database search. The compounds, which are identical according to their structural formula, were marked with different identifiers. Therefore, they were provided as two separate search results, see **Figure 2**.

Nevertheless, the entries of compounds listed under different identifiers (identification numbers, see above) contained different physical property values. The focus then became to find out such duplicities, and how much the data about "the same" compounds differ. Despite the original idea coming up when working with Wolfram Alpha, software Wolfram Mathematica was used for this research.

Figure 2. Wolfram Alpha – Isomers

METHODOLOGY

The data in Wolfram Mathematica were selected and further processed using the ChemicalData function. This function enables a search for compounds and simultaneously accesses dozens of different features for this compound according to the request (**Figure 3**).

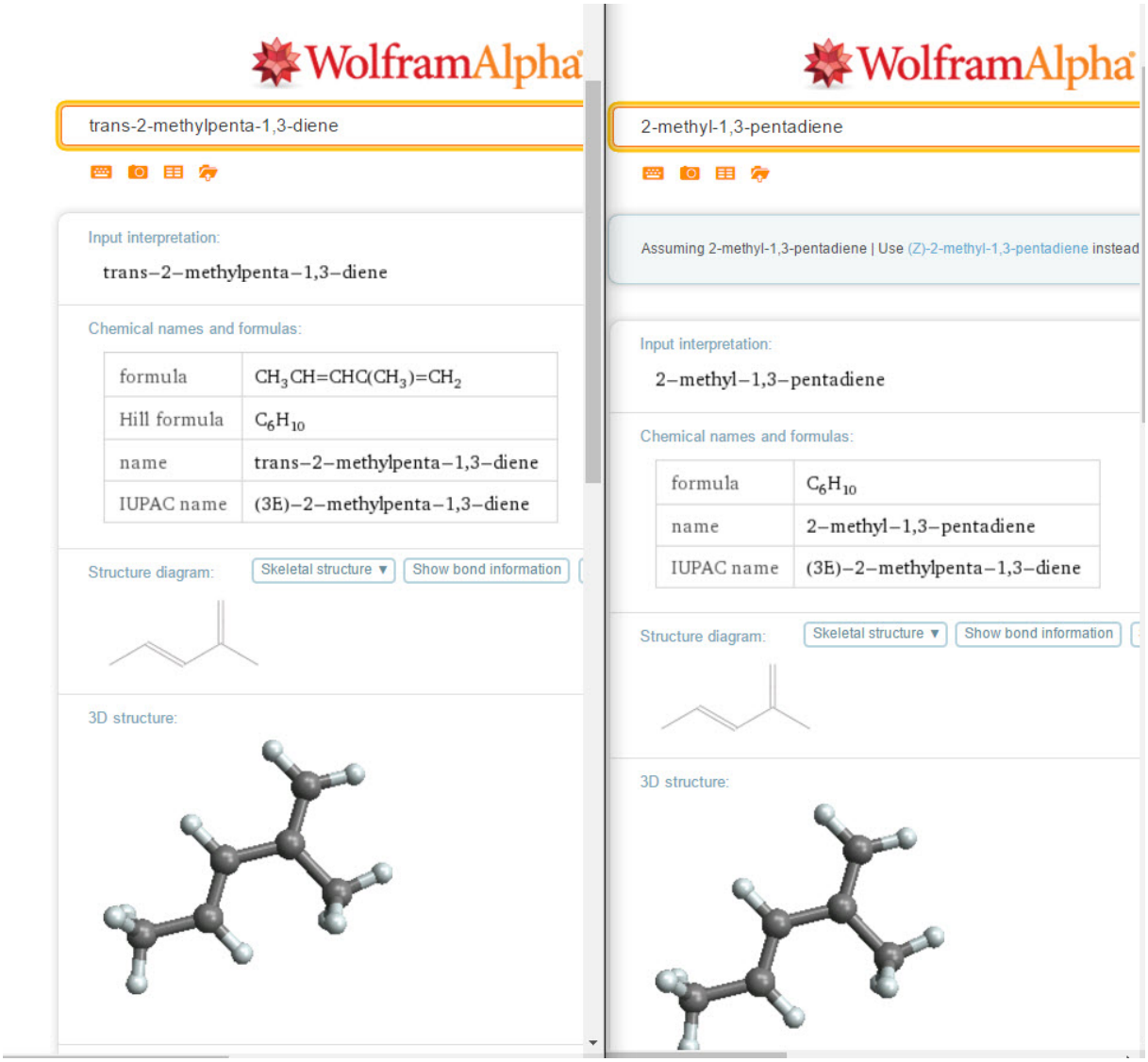

Figure 3. Wolfram Alpha – results

As seen in **Figure 3** above, even though the compounds are identical, the NFPA labels (National Fire Protection Association) are different, which is clearly a mistake. Also, the flash point as well as other physical properties differ (not displayed in **Figure 3**).

The selection of the compounds was based on the grounds of their feasibility, limited only to hydrocarbons. Every hydrocarbon was analysed according to the reference of its record in one of the 3 basic databases (using CAS Number, CID Number and Beilstein Number). Further, its three physical properties significant in chemistry, boiling point, flash point and melting point were recorded. The data acquired this way were saved into a table in which the duplicities were analysed.

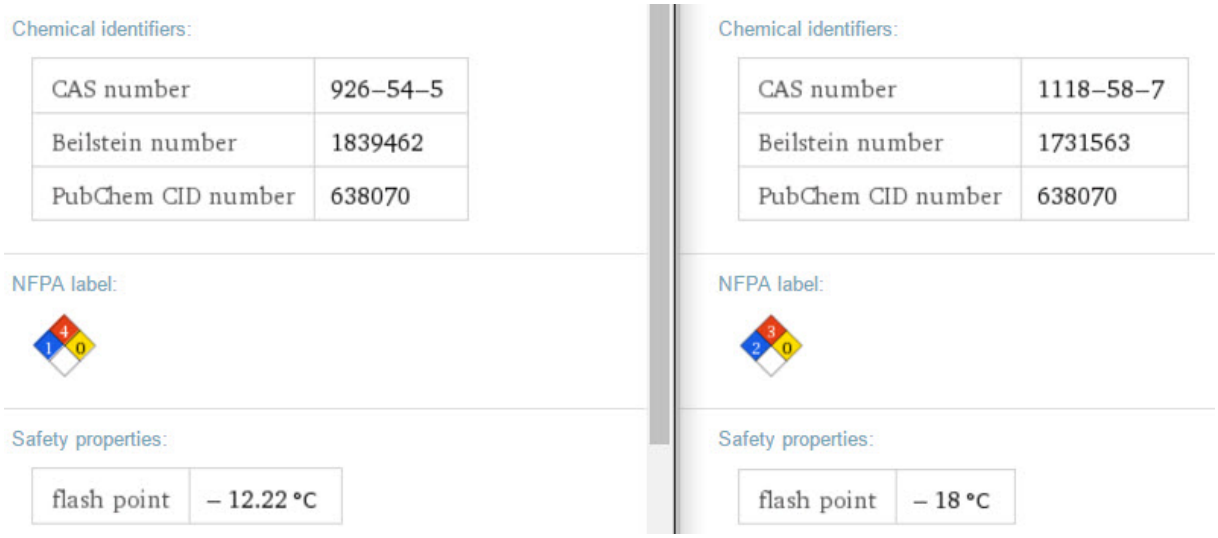

Figure 4. Wolfram Alpha – results

Table 1. Number of compounds found in the databases

Identificator	CAS	CID	Beilstein
Number of items	1498	1449	1088

Table 2. Numbers of duplicities

Identificator	CAS	CID	Beilstein
Number of duplicates	0	96	6

Table 3. Occurrence of particular duplicities

Items of one CID	Occurrence
6	1
5	1
4	1
3	13
2	80

RESULTS AND DISCUSSION

Overall 1499 compounds were identified by the Mathematika program. Every compound was identified by a CAS Number, some of the records did not contain complete data. Concretely, 410 were missing a Beilstein Number and 49 of them a CID Number. The final results for the hydrocarbons found are shown in **Table 1**.

Numbers of Duplicities

Based on the data analysis, the authors of this text conclude CAS is a number describing every compound separately. On the contrary, one CID number may indicate various compounds, see **Table 2**).

While the Beilstein number indicates only six duplicities, every one including two compounds, the CID number includes more compounds (see **Table 3**). This may naturally cause students' confusion when searching for a compound's properties. However, there is also a motivation aspect to this, as students usually like to disclose discrepancies.

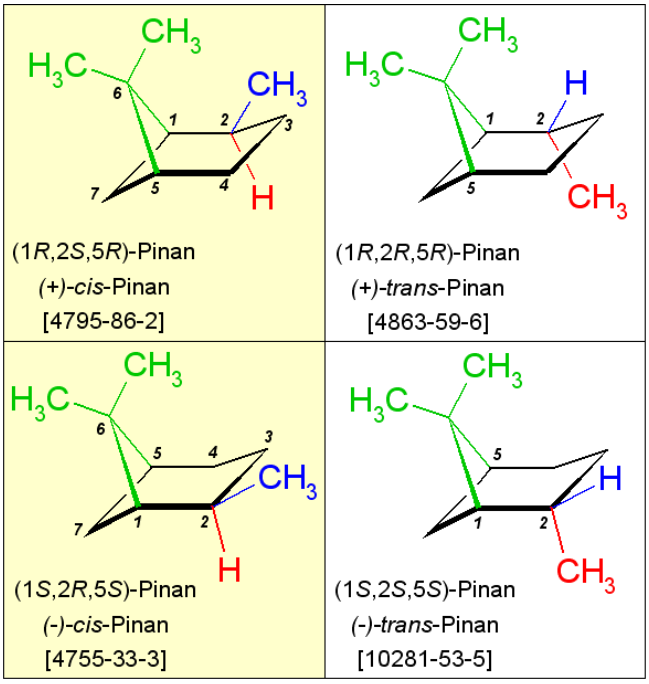

Figure 5. Isomers of pinane, source: wikimedia commons

When analysing the compounds and the identifiers, the authors indicated cases where compounds under the same CID are listed under different CAS and/or Beilstein number with different characteristics, such as boiling point, flashpoint, melting point or name. Thorough analysis of all the 96 cases is not the subject of this paper. Therefore, only several examples are given.

2-methylpenta-1,3-diene

The case will be sketched on a selected compound, see **Figure 3**. Even though it is listed under one CID, there are two different CAS and Beilstein numbers. Nevertheless, the compound at stake is just one *2-methylpenta-1,3-diene*. The difference was found in the name (one of the entries was listed under a name according to the old organic chemistry nomenclature: CAS1118-58-7, the new one CAS926-54-5). Although it may seem the entry CAS926-54-5 is newer, the CAS number is not as great as expected. Melting point has also been added, which is missing under the CAS926-54-5 entry. The boiling points stated in both entries differ from each other by only 0.5 (the CAS926-54-5 entry is more accurate).

It is not possible to say one entry is newer, therefore more detailed than the other. Apparently, the browser draws the information from several databases which are not being updated. A search for such cases may again be motivating for students.

Pinane

This CID number indicates a compound named *pinane*. There are six entries under this particular CID. The common molecular formula is $C_{10}H_{18}$. After a closer look at the entries, the authors found out one of them refers to pinane (CAS6876-13-7, Beilstein5237941), another to cis-pinane (CAS473551, Beilstein1280394) and the four other entries to particular *R* and *S* enantiomers of different optical activity (+/-). Even here, all theoretically possible combinations have not been found in the database. A more thorough study of the molecule (bicyclic molecule based on a six-carbon ring naturally in favoured chair conformer) offers only two positions for methyl group attachment in both of these offer *cis*- and *trans*- stereoisomers. These are listed under different CAS and Beilstein numbers, **Figure 5**.

Examining these compounds' other properties, the authors found that these compounds, with the exception of (+)-*trans*-pinane with a boiling point of 51.5°C, have the boiling point listed as 164, 165, 167, 168, 169°C (The differences are explainable by a typical span in the databases – boiling point is a range of temperatures). Yet it is unclear why the isomers differ in ones of degrees among each other, instead of all having a range of temperature entered. The structure of (+)-*trans*-pinane does not suggest such a big difference in the boiling point. The authors assume this is a mistake in the database.

Another discrepancy was found in the other following categories – the flash and melting point. Only the *cis*-pinane (CAS473552, Beilstein1847301) entry contains both values, while (+)-*cis*-pinane contains only the melting point. The melting point values of the aforementioned are equal.

The compound is more complicated and therefore suitable to be examined in higher secondary school grades. It could be examined by younger students too, however the activity would lose its interdisciplinary aspect between chemistry and IT.

Pellandrene

The compound under this CID appears in the database under five different CAS numbers, only two of them have a Beilstein number assigned. They represent a compound called *a*-phellandrene. The only systematic name which enables the creation of the formula is CAS4221-98-1, Beilstein 2487824. Boiling point values differ from 155°C to 176°C. Two entries (CAS34448-33-4 and CAS99-83-2) of the same name *alpha pellandrene* contain further information; the first about the melting point, the second about the flash point. The compound named *p*-menthadiene (CAS1329-99-3) also contains a boiling point record. It differs by 18°C from the value under CAS34448-33-4. In education, this could easily be approximated for the students with the use of a well-known example of alcohol. The difference between the boiling point of methanol and ethanol is about 14°C, which enables the distillation of one liquid from another. This is the process of liquor production. Discrepancy in the boiling point value as big as in case of pelladrene would lead to severely jeopardising health, as poisonous methanol not distilled from the ethanol mixture would cause the death of a consumer.

The databases offer a great number of possibilities which enable teachers to focus on more complicated (complex) cognitive processes instead of wasting time by making the students look up some properties in books or random websites. (Nevertheless, the authors of this study assume this skill is still also very important. The main focus, however, should not be put on this.) Searching the online databases, however, leads to several ambiguities among the search results.

This could lead to further investigation of the correct values, which is more suitable for chemistry classes. Also, the search engine's functionality could be examined in order to trace a possible source of the discovered discrepancies, which is more suitable for IT classes. An important aspect is the emphasis on critical data analysis when searching for information.

Strong points and limitations of the research

Although the mentioned search tools are being used widely, the duplicities and discrepancies have not been given appropriate attention. Connecting this discovery with the need to practice students' ability to search for information and evaluate them critically is a novel approach.

The results of this paper are affected by the authors' focus on only one group of chemical compounds. More examples could be found if all the compounds in the databases were reviewed. Also, complete analysis and selection of compounds where the discrepancies found exceed normal tolerances (therefore meaning an error) could set certain examples for the students to examine. This would save the time and a teacher would be able to guide their students' learning process better.

Advantages and disadvantages of the approach

The aforementioned approach offers several directions where to guide students' learning process. As far as problem solving, information literacy development or critical thinking are concerned, the activities are suitable for all, preferably secondary school, students. However, the compounds comparison activity, it is advised to be used for the group of students who focus on chemistry or science as too much emphasis on formulas and structure of compounds can easily demotivate students with different interests. This approach serves as a demonstration of information literacy development within a scientific subject. Bearing STEM conception in mind it therefore brings the two important components together in terms of the unifying thought.

CONCLUSION

Using Wolftam Alpha as a search tool also proved feasible and suitable for educational purposes. Based on the search analysis it is possible to distinguish:

- duplicities,
- obsolete or out-of-date information,
- error.

They may be caused by the human factor or by non-current information being kept in the database and new simply added.

These, when not identified and reported, may cause misunderstandings not only at school but also in pure chemistry. Also, it could cause problems in terms of safety (see the above mentioned NFPA labels). If the databases are supposed to serve science and scientists, it is necessary to keep them updated.

However, the educational potential of this discovery is also interesting. Students motivated by unexpected discrepancies are led to further investigate the results. This enables them to focus either on the chemical/physical properties and compound identification or on the IT part, in terms of the possible origins of the discrepancies discovered.

The number of compounds found may differ as new compounds are being discovered and enlisted into databases. **Figure 2** serves as an example. During final corrections to this paper, the authors found the error has been detected and this entry corrected in the database.

The work's further steps could be in detailed inspection of the hydrocarbon-search results, with the aim of searching for errors and unexpected values among the physical properties of the same compound. The authors also tend to elaborate the idea of combinatorics-organic chemistry into more depth.

Endnotes

The number 1498 represents compounds found by the authors during the work on this paper. As research in organic chemistry produces new compounds, the number may be greater by the time the paper is published.

The database mentioned in the text is available on this link: http://reference.wolfram.com/language/note/ChemicalDataSourceInformation.html and http://demonstrations.wolfram.com/.

REFERENCES

Abramovich, S. (2014). Revisiting mathematical problem solving and posing in the digital era: Toward pedagogically sound uses of modern technology. *International Journal of Mathematical Education in Science and Technology, 45*(7), 1034-1052.

Bell, R. L., Gess-Newsome, J., & Luft, J. (2008). *Technology in the secondary science classroom*: NSTA Press.

Beneš, P., Rusek, M., & Kudrna, T. (2015). Tradice a současný stav pomůckového zabezpečení edukačního chemického experimentu v České republice. *Chemické Listy, 109*(2), 159-162.

Eilks, I., Wittec, T., & Pietzner, V. (2009). A Critical Discussion of The Efficacy of Using Visual Learning Aids From The Internet To Promote Understanding, Illustrated With Examples Explaining The Daniell Voltaic Cell. *Eurasia Journal of Mathematics, Science and Technology Education, 5*(2), 145-152.

Ersoy, M., & Akbulut, Y. (2014). Cognitive and affective implications of persuasive technology use on mathematics instruction. *Computers & Education, 75*, 253-262. doi: 10.1016/j.compedu.2014.03.009.

Filippoupoliti, A., & Koliopoulos, D. (2014). Informal and non-formal education: An outline of History of Science in museums. *Science & Education, 23*(4), 781-791. doi: 10.1007/s11191-014-9681-2

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for Life in Digital Age*. Amsterdam: Springer International Publishing.

Greenhow, C., & Lewin, C. (2016). Social media and education: reconceptualizing the boundaries of formal and informal learning. *Learning, Media and Technology, 41*(1), 6-30. doi:10.1080/17439884.2015.1064954

Jančaříková, K. (2017). Teaching Aids and Work with Models in e-Learning Environments. *The Electronic Journal of e-Learning, 15*(3), 244-258.

Kim, M., & Aktan, T. (2014). How to Enlarge the Scope of the Curriculum Integration of Mathematics and Science (CIMAS): A Delphi Study. *Eurasia Journal of Mathematics, Science and Technology Education, 10*(5), 455-469. doi:10.12973/eurasia.2014.1115a.

Král, L., & Řezníčková, D. (2013). The proliferation and implementation of GIS as an educational tool at gymnasiums/grammar schools in Czechia. *Geografie, 118*(3), 265-283.

Lebedeva, O., & Zaitseva, L. (2014). *Question Answering Systems in Education and their Classifications.* Paper presented at the Joint International Conference on Engineering Education & International Conference on Information Technology.

Libman, D., & Huang, L. (2013). Chemistry on the Go: Review of Chemistry Apps on Smartphones. *Journal of Chemical Education, 90*(3), 320-325. doi:10.1021/ed300329e

OECD. (2016). Pisa 2015 Results (Volume 1): Excellence and Equity in Education Paris: PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264266490-en

Pence, H. E., & Williams, A. J. (2016). Big data and chemical education. *Journal of Chemical Education, 93*(3), 504-508. doi:10.1021/acs.jchemed.5b00524

Rusek, M. & Dlabola, Z. (2011). "Projectivity" of Projects and Ways of its Achievement. *Project-Based Education in Chemistry and Related Fields, IX*, 12-23.

Rusek, M., Stárková, D., Chytrý, V., & Bílek, M. (2017). Adoption of ICT Innovations by Secondary School Teachers and Pre-service Teachers within Education. *Journal of Baltic Science Education, 16*(4), 510-523.

Sha, L., Looi, C. K., Chen, W., & Zhang, B. H. (2012). Understanding mobile learning from the perspective of self-regulated learning. *Journal of Computer Assisted Learning, 28*(4), 366-378. doi:10.1111/j.1365-2729.2011.00461.x

Weisstein, E. (2014). Computable Data, Mathematics, and Digital Libraries in Mathematica and Wolfram Alpha. In S. M. Watt, J. H. Davenport, A. P. Sexton, P. Sojka & J. Urban (Eds.), *Intelligent Computer Mathematics: International Conference, CICM 2014, Coimbra, Portugal, July 7-11, 2014. Proceedings* (pp. 26-29). Cham: Springer International Publishing.