J Comput Aided Mol Des (2013) 27:583–603
DOI 10.1007/s10822-013-9664-4

From data point timelines to a well curated data set, data mining of experimental data and chemical structure data from scientific articles, problems and possible solutions

Villu Ruusmann · Uko Maran

Received: 2 March 2013 / Accepted: 2 July 2013 / Published online: 25 July 2013

Abstract The scientific literature is important source of experimental and chemical structure data. Very often this data has been harvested into smaller or bigger data collections leaving the data quality and curation issues on shoulders of users. The current research presents a systematic and reproducible workflow for collecting series of data points from scientific literature and assembling a database that is suitable for the purposes of high quality modelling and decision support. The quality assurance aspect of the workflow is concerned with the curation of both chemical structures and associated toxicity values at (1) single data point level and (2) collection of data points level. The assembly of a database employs a novel "timeline" approach. The workflow is implemented as a software solution and its applicability is demonstrated on the example of the *Tetrahymena pyriformis* acute aquatic toxicity endpoint. A literature collection of 86 primary publications for *T. pyriformis* was found to contain 2,072 chemical compounds and 2,498 unique toxicity values, which divide into 2,440 numerical and 58 textual values. Every chemical compound was assigned to a preferred toxicity value. Examples for most common chemical and toxicological data curation scenarios are discussed.

Keywords Data curation · Data timelines · Toxicity data · *Tetrahymena pyriformis* · QSAR

Introduction

Data quality in general and toxicological data quality in particular is an important issue as it has a direct influence to the hazard and risk assessment and toxicity prediction [1]. Basic principles and existing schemas of assessing toxicological data quality have been recently analysed by Przybylak et al. [2]. They describe four key criteria for quality assessment and assurance of toxicological data. From them two are carried towards improving data quality and can be termed as main components of chemical data curation: (1) the identification and characterization of tested chemical compounds and (2) the consistency of experimental design protocols and experimentally measured values.

Over the last decade a considerable effort has been done to make the development, validation and exploitation of QSAR models rigorous. One of the most visible cooperation initiatives has been the consolidation of validation principles of QSAR models [3] in order to smooth the path towards the use of QSAR for regulatory purposes. Actually, those principles have been well known, but they have been often overlooked while developing and characterizing QSAR models. The cause of ignorance can easily be related to the (not in any particular order) insufficient

Disclaimer on fair use conditions. The data set which accompanies current publication should be regarded as a derivative work or the earlier publications of Prof. T.W. Schultz (University of Tennessee) and his co-workers. Further users are expected to properly credit and reference the most significant of them.
Disclaimer on undiscovered and new data. Authors did their best effort to find all possible *T. pyriformis* acute aquatic toxicity primary publications. Despite of that it can occur that something is missed or new data becomes available. In either case we are most grateful for references to such potential primary publications.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-013-9664-4) contains supplementary material, which is available to authorized users.

V. Ruusmann · U. Maran (✉)
Institute of Chemistry, University of Tartu, Ravila 14A, 50411 Tartu, Estonia
e-mail: uko.maran@ut.ee

education, unrealistic expectations from software, misapplication of QSAR methods, etc.

In general for QSAR development there seems to be a good understanding and agreement about which steps in which order constitute a solid modelling workflow [4]. In this context the quality of QSAR models has been a hot topic for research and discussion. However, the same cannot be said about the quality of data sets used for QSAR. There are only a few works up to date that emphasize the need for better data curation in QSAR workflows [4] and demonstrate how faulty data (incorrect chemical structures) negatively influence the quality of prediction of QSAR models [5, 6] and address the rate of inconsistent experimental values in scientific literature [7]. So, it makes data curation an important aspect of QSAR and any other approach that uses experimental chemical data in elucidating new knowledge [8].

With the advent and proliferation of public domain chemistry and biology databases [9, 10] it is common practice that the data set is obtained by a database search. Such search results promote the illusion that all the results in the query are of equal quality, which is hardly the case. Databases aggregate information from other databases and primary-esque data sources in good faith, and do not take full responsibility for their actual correctness and truthfulness. Exceptions are variety of in-house and commercial databases that curate data, but as rule do not provide detailed data curation workflows. The longer the data propagation chain is, the higher the probability of both random (e.g. transcription) and systematic errors (e.g. incompatibilities between database schemas). So, by accepting the database query results and not following the accompanying references to original data bodies, data modellers wilfully put themselves at the mercy of database builders. The quality of chemistry databases has been reviewed recently [11, 12]. The general conclusions were that chemistry data is much harder to represent and process than other scientific data, and that there are several technological deficiencies that have to be considered (e.g. the lack of a dictionary of most common chemical compound names).

The large majority of experimental results are first published in the scientific literature (e.g. peer-reviewed journals). This suggests that data quality should be addressed at the level of single publications. In short, any (primary-) publication should be able to prove the quality of its data in place. The proof does not need to be overwhelming. It is adequate if simply references are given to other publications that define the endpoint, experimental setup, etc. in full detail. Such references can be used for collecting (primary-) publications during data mining and extraction. The simplest approach is to select publications that share exactly the same endpoint reference. A more advanced approach is to select publications that share one of many "compatible" endpoint references, where the compatibility has been determined beforehand by an expert.

In parallel to Przybylak et al. [2], Fu et al. [13] have recently reviewed data quality issues, (in the context of data governance) in multidisciplinary research fields such as predictive toxicology. They define data quality as fitness for serving particular purpose, and outline three of its most interesting components: (1) accuracy, (2) completeness and (3) integrity. The most important component is accuracy, which refers to data correctness and consistency. Accuracy can be verified and, more importantly, improved via appropriate curation activities.

Data curation has been gaining traction in recent years, but has not found its true foothold yet. Literature search about data curation in the context of QSAR finds only a few results (at the time of writing this manuscript the combination of keywords "data AND curation AND QSAR" matches 7 scientific articles in Thomson Reuters (ISI) Web of Knowledge database). The majority of pioneering work has been done by Tropsha et al. [4, 6] who define five major steps of data curation: (1) removal of inorganics and mixtures, (2) structural conversion and cleaning, (3) normalization of specific chemotypes, (4) removal of duplicates and (5) manual checking. Their data curation activities are based on already existing data collections and therefore effort is mostly focused on two aspects, normalizing and standardizing chemical structure representations and performing data set filtering by structural constraints. While preparing data set for QSAR analysis, all those aspects are very important in order to remove noise from the data set and to make data set fit for modelling method. However it could be worthwhile and useful to differentiate between data filtering and data curation. The main difference is in the handling of problematic and/or suspicious data. Data filtering removes such data points. Data curation corrects such data points using cues from all possible sources so that they could be retained. Uncurable data points are labelled appropriately and also retained. It is possible that some other researchers may be able to correct them. Both of them, data curation and data filtering can be used in sequence or most likely iteratively.

The current work uses published scientific articles as the primary source of information and goes through a complete data curation and data assembly workflow. Data curation issues are dealt systematically, both, from the chemical structure, and the associated target/property point of view. The case is built around the *Tetrahymena pyriformis* acute aquatic toxicity endpoint, also known as the TETRATOX test [14], which has been measured by a single group in a single laboratory and is free of the inter-laboratory variance

component. However, the case is characterized by long history of publication and (re)citation of experimental measurements, which necessitated the introduction of a novel "timeline" approach for data curation, to order all the known experimental values in a consistent and reproducible way. The proposed workflow is implemented as a software solution. Due to the generic nature of the problem it is applicable to other endpoints without any or very little modification. The performed research on literature data curation was largely initiated by the data collection activity while assembling dataset for substituted benzenes with experimental values. Eventually it was extended to all available *T. pyriformis* acute aquatic toxicity data accessible in scientific literature.

Materials and methods

The chemical data curation workflow used in this work is summarized on Fig. 1. Horizontal blocks indicate three types of activities: literature collection, curation of chemical structures, followed by the curation of toxicity values. Vertical blocks provide: flowchart how different data tables merge into curated one, activities performed on each step, agents used for each step and finally software used for the curation. Next chapters provide detailed explanation of each block.

Literature collection of *T. pyriformis* publications

Criteria for the collection of publications

The initial pool of references was populated (Fig. 1: step 1) by querying the Google Scholar scholarly literature search engine [15] using relevant keywords ("*T. pyriformis* toxicity", "Tetrahymena toxicity", "TETRATOX" etc.) and by harvesting the list of selected works of Schultz [16]. All potentially interesting publications electronically retrievable from the network of University of Tartu were downloaded as PDF documents. The initial pool of publications contains only journal articles. All those publications were analyzed and harvested for more references to potentially interesting publications. This cycle was repeated until all reference chains had reached their ends. These additional publications that could not be retrieved electronically were ordered on paper using the interlibrary loan services and digitized locally. Those additional publications included also book chapters and conference proceedings. To the best of our knowledge the complete collection of *T. pyriformis* article publications that have appeared up to the date in scientific literature has been reached.

The earliest publication dates back to 1980 [17] and is the first time when *T. pyriformis* enters the stage as a research object. Up till then *T. pyriformis* was used more like a research tool facilitating the study of other research objects. This publication is remarkable, because it features the first experimental setup for measuring the sublethal response (i.e. the inhibition of population growth), whereas all the earlier setups were measuring the morality of inoculum. The information which is being searched from the publications is experimentally measured *T. pyriformis* acute aquatic toxicity data. The data is usually represented in a tabular form. It is interesting to note that in earlier publications it is pretty common to see the parallel representation of toxicity firstly as the "population growth impairment model" and secondly as the concentration reading of this model for the 50 % inhibition mark. In later publications the first representation is not present. This is probably because the representation of "population growth impairment model" requires considerable amounts of space, but finds no use during QSAR modelling.

All publications were then classified as either primary or secondary publications. A publication was classified secondary publication and removed from the collection if it did not fit to the following conditions. First, does not present any toxicity data. Second, does present toxicity data but, makes it explicit that data had all been extracted from external publications and that the authors were not engaged in performing any experimental measurements.

Data series mining from PDF documents and conversion to QsarDB archives

Next step in literature collection is data extraction and conversion from PDF files (Fig. 1: step 2) that is organized in semi automated way, where data tables are extracted automatically and meta-data manually. This chapter is rather technical, but necessary to show the complexity of PDF documents while working with them. The collection of publications contains PDF documents in different variants: (1) standard format, (2) publisher scanned image format with optional publisher optical character recognition (OCR) data overlay, (3) local scanned image format (300 dpi resolution) with local OCR data overlay (implemented in Adobe Acrobat).

From the technical perspective, the PDF data format [18] can be regarded as vector-graphics data format suited for graphical reproduction of information across time and computational environments, which is important from the publishing point of view. This greatly hinders the non-graphical processing of information by programmatic agents, which is important from the further research point of view. In short, PDF documents contain a lot of scientifically valuable information, but it is not easy to take advantage of it in a systemic non-technical way. The topic of parsing PDF documents as well as the topic of

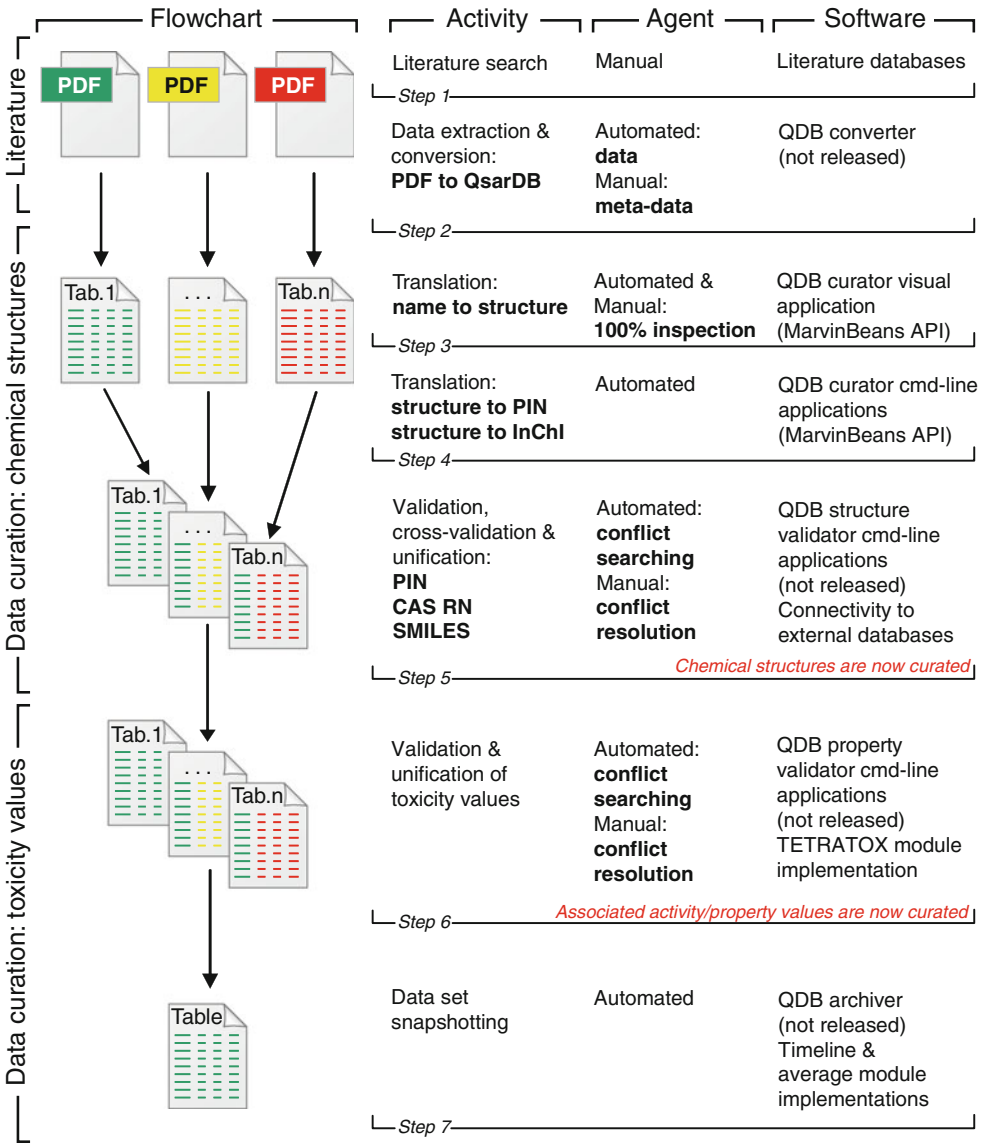

Fig. 1 Chemical data curation workflow

algorithmic recognition and recovery of tabular data structures belong to the field of information science and will not be discussed here in more detail. The description is given only for the general framework for the extraction of tables.

The major drawback of the PDF data format from the perspective of the current research is that it completely lacks the notion of tables as a means of structuring content. To overcome this the table extraction tool was programmed in the Java programming language using the Apache PDFBox library [19]. It parses a PDF document and recognizes and recovers all tabular data structures in a specified page range. After that it maps and transfers the data to another table or table-like data schema. In the current work this other table-like data schema is the QsarDB archive format [20], which has been specifically designed for the purpose of electronically organizing and archiving QSAR data and metadata. Alternative destinations could be generic spreadsheet file formats (CSV, OpenDocument, Microsoft Excel) or relational databases.

The automated data extraction results from the standard PDF document format are guaranteed to be correct. Data extraction results from the scanned image formats may contain random errors that are related to the accuracy of the OCR algorithm. It may be advantageous to reprocess publisher scanned image format documents and replace the publisher OCR data overlay with local OCR overlay in order to reduce the variance, especially if the local OCR algorithm has demonstrably superior accuracy. The overall rate of random errors is negligible (les than on 1/1,000 on symbol basis), because chemical data is typically presented in normal typeface Latin letters and digits. The criticality of random errors is less significant with text fields (e.g. the name of a chemical compound) and more significant with

numeric fields (e.g. the toxicity value of a chemical compound).

While the majority of chemical data in PDFs is contained in tables, the relevant and valuable metadata scattered elsewhere in free text. This includes the description of the endpoint, the specifications of units, the bibliographic reference of the toxicity protocol. Sometimes it is also possible to recover the mathematical equations of QSAR models, which may come in handy during data cross-validation and conflict resolution. All this data was extracted manually and added to QsarDB archives in designated places in the required data format.

Curation of chemical structures and toxicity values

Ranking of attributes for the representation of chemical compounds

Chemical data tables in PDF files typically use one to three table columns for the representation of chemical compound that can be used in chemical structure curation. Those representations can be ranked by relevance for the elucidation of chemical structures as described below.

The chemical name: highest relevance Encodes the chemical structure for human. Consequently, this representation possesses the most "degrees of freedom", because different naming schemes and name variants have been in use. This clearly suggests that, there is a need to translate all existing name representations to a stable and standard representation. The best candidate is the preferred IUPAC name (PIN) [21]. An important factor in favour of PIN is decent support by the latest versions of the MarvinBeans Java cheminformatics library [22], which enables programmatic agents to perform parsing and formatting.

CAS registry number (CAS RN): medium relevance The potential utility of the CAS RN [23] is that it can be resolved to some other chemical compound representation using external on-line services. The only primary service that can source names and chemical structures is the CAS registry database [23]. The full version of CAS registry database is often closed for small-scale academic research. A selection of approximately 7,900 chemical compounds of widespread interest can be reached via the CAS Common Chemistry web portal [24]. However, the sampling with random *T. pyriformis* data sets indicates that the provided coverage by the web portal is too low for the current needs. This leaves secondary services such as the NIH/CADD Chemical Identifier Resolver service [25] or the ChemSpider service [26] as an alternative option. The problem with secondary services is that their results are not symmetrical and normalized (i.e. there is a one-to-one mapping from CAS RN to chemical name, but a one-to-many mapping from chemical name to CAS RN), and that they lack data traceability information. Previous knowledge suggests extra caution with CAS RN, because the reported values may be fitted by researchers rather than being reflective of the labels of reagent bottles. Therefore, in situations where the chemical structure inferred from the chemical name and the CAS RN differ, the one based on the chemical name is typically given preference.

SMILES: lowest relevance Encodes the chemical structure for programmatic consumption [27]. The low ranking is related to its infrequency in PDFs and dubious information content. The problems with SMILES (Simplified Molecular-Input Line-Entry System) line notation are manifold, which may be summarized as follows: (1) competing standards such as the original SMILES [27], Daylight SMILES [28], OpenSMILES [29], (2) competing software implementations, (3) rigid and lossy data model, (4) not guaranteed to produce unique result. In the current work the SMILES line notation will be disregarded as an option for 2D chemical structure representation. It is superseded by the International Chemical Identifier (InChI) [30], which does not suffer from any of the above problems in a major way, at least for small organic molecules.

Identification of chemical compounds

The identification of chemical compounds (Fig. 1: step 3) is performed using a special-purpose visual tool (Fig. 2) which operates on QsarDB archives. This tool and many other tools have been released as open source software in the project's Google Code project page [31]. The user interface of visual tool provides controls for stepping through the contents of the QsarDB archive one chemical compound at a time. The name of the active chemical compound is parsed using the *SystematicNameConverter* component of the MarvinBeans library. If the parsing operation succeeds, the resulting *Molecule* object is displayed using the *MViewPane* component of the MarvinBeans library. It is the responsibility of the human agent ("the chemical curator") to decide based on the visual feedback whether the parsing operation can be declared a success or not (Fig. 2). If the parsing operation fails either by an explicit error condition in the parsing algorithm or by the negative visual feedback, then it is again the responsibility of the human agent to change the chemical name so that the issue is corrected. All changes are recorded to the QsarDB archive.

Translation into stable PIN and InChI representations

Successful identification of chemical compounds gives confidence that the MarvinBeans framework of libraries

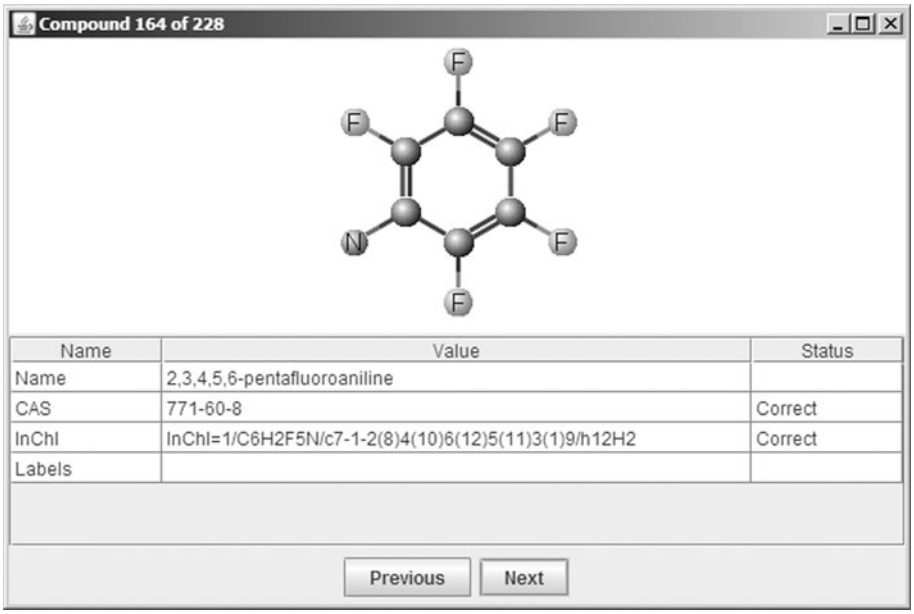

Fig. 2 Screenshot of the QsarDB chemical structure curation GUI application

and tools can be trusted with the further algorithmic processing of *Molecule* object instances without explicit human supervision (Fig. 1: step 4).

The *Molecule* object instance is first converted to a PIN using the *IUPACNamingPlugin* component of the MarvinBeans library. The validity of every generated PIN is verified by checking that it yields an equivalent *Molecule* object instance when parsed using the *SystemicNameConverter* component of the MarvinBeans library (Fig. 3; the final representation is PIN, the indirect pathway proceeds over InChI). This verification revealed less than ten problematic chemical structures in total. For example, the two most prominent cases are penta-substituted anilines `pentafluoroaniline`[1] and `pentachloroaniline` that must be prefixed with position identifiers to produce `2,3,4,5,6-pentafluoroaniline` and `2,3,4,5,6-pentachloroaniline`, respectively, because otherwise the first two halogen atoms would replace the hydrogen atoms of the nitro atom of the aniline group. This particular problem exists in the MarvinBeans library version 5.5.0, but it may or may not exist in other versions.

The *Molecule* object instance is then converted to InChI using the *MolExporter* component of the MarvinBeans library (Fig. 3: direct pathway). Starting from InChI software version 1.03 [30] (included in MarvinBeans library version 5.5.0 and newer) it is possible to generate non-standard InChI representations (prefix "InChI = 1/") in addition to the standard InChI representation (prefix

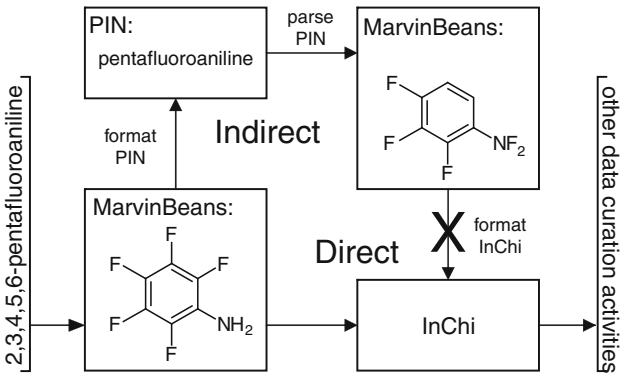

Fig. 3 Conversion of chemical structures from the original representation to standardized representations PIN and InChI. The Molecule instance (depicted *lower left*) is converted to InChI using two pathways. The direct pathway (depicted *below*) uses direct conversion. The indirect pathway (depicted *above*) performs an additional formatting and parsing step over PIN. Both pathways are expected to yield the same InChI. Any differences would indicate a software bug in the MarvinBeans library

"InChI = 1S/"). In the current work the *MolExporter* component is ordered to generate a non-standard InChI representation by specifying the extra creation options "FixedH" and "SUU" (see further details in Supplementary Material).

Validation of CAS RN

CAS RN has been ignored so far in the chemical structure data curation workflow. They are brought up to date by two activities (Fig. 1: step 5). The first activity is the validation in isolation using the built-in "check digit verification" [32] mechanism. This relies on a simplistic checksum calculation algorithm, but can identify on average 9 out of

[1] Two chemical naming systems are used through the manuscript: (i) Standardized chemical names, i.e. Preferred IUPAC Name (PIN) and (ii) original chemical names from referenced publications. They are distinguished by the font: PIN is in font Courier+Italic and original names in font Courier.

10 compromised digit sequences. Invalid digit sequences are immediately corrected by querying the NIH/CADD Chemical Identifier Resolver service [25] after the PIN or InChI. If the query returns more than one result (see above, the comment of secondary CAS RN databases) then the one which is the most similar to the original CAS RN is selected. The second activity is the validation of CAS RN against PIN and InChI representations. The most promising approach today seems to be the CAS RN resolution to InChI using the NIH/CADD Chemical Identifier Resolver service [25]. The standardized InChI is compared against the temporary externally resolved InChI in terms of the main layer and the stereochemistry layer.

Unification of chemical structures

In practice when chemical data tables are copied from one publication to another the quality of data can be degraded. The information about the chemical structure of chemical compounds becomes more opaque, because the authors either knowingly or unknowingly tend to omit specific features. For example, when chemical structures are prepared for the calculation of molecular descriptors, it is common that the salt and hydrate forms of a chemical compounds are replaced with its "raw" form. The unification of all observations of a chemical compound attempts to undo (i.e. reverse) the effects of such degradation as much as possible.

The first step of the unification pathway (Fig. 1: step 5) is grouping chemical compounds into working sets by the main layer of the InChI. If the chemical structure contains more than one fragment only the first (largest) is taken into the consideration (for example in case of 2-aminophenol hydrochloride only 2-aminophenol fragment is considered significant). The chemical structures which probably suffer from the loss of fragment(s) (e.g. omitted salt and hydrate forms) are restored. If there is stereoisomerism involved, then the working sets are re-grouped into subsets by the stereochemistry layer of the InChI. Again, the chemical structures which probably suffer from the loss of stereochemistry configuration (e.g. omitted geometric isomerism of double bonds) are restored.

Unification of toxicity values

Workflow now continues with the curation of toxicity values, with the second step of the unification pathway that is about identifying and correcting toxicity value errors (Fig. 1: step 6). It is perfectly normal to expect that a chemical compound may have different target activity/property values associated with it over time. For example, *T. pyriformis* acute aquatic toxicity data has been experimentally measured according to five different toxicity protocols [17, 33–36] in a time period which spans over three decades. The variability of toxicity values could be explained in two ways: (1) re-testing the chemical compound under the same toxicity protocol, (2) re-testing it under a different (i.e. updated) toxicity protocol. One of the advantages of *T. pyriformis* acute aquatic toxicity data is that the variability does not have an inter-laboratory component, because almost all experimental measurements have been conducted in a single laboratory [14].

The handling of toxicity values errors is well exemplified by sign errors. The automated screening process employs a programmatic agent that searches for chemical compounds that report toxicity values that are of equal magnitude but of opposite sign. The correct sign can be judged by the human agent by observing the relation of either value with some well-established quantitative (structure activity-) relationship such as the baseline toxicity regression line [37] or an ad hoc single-parameter regression line for closer homologues. Other kinds of toxicity value errors can be identified on an individual basis, but not probably corrected. However, it is always advisable to document all inconsistencies and "gut feelings" for future reference.

Data assembly

Timeline approach

The central problem of data assembly is how to handle the situation where a chemical compound has different target activity/property values associated with it in different publications. The "classical" solution is to select whatever was published most recently (as represented by the date of publication) depending on availability and access. This approach is easy to implement for human agents and can be applied to any literature collection of publications. However, such data sets lack inherent stability. The result of "classical" method however also strongly depends on the experience of human component.

In the current work the situation of multiple values in different publications is tackled using a novel so-called "timeline" approach, which helps to analyse the appearance of experimental values in the literature for one compound and performs the selection of experimental values in a consistent and reproducible way (Fig. 1: step 7). The timeline algorithm is applied on a chemical compound basis. It is composed of the following procedure:

1. Ordering the collection of publications by the date of publication.
2. Identification of unique toxicity values. Pruning the collection of publications so that for each unique toxicity value only its very first appearance is retained.

3. Reporting the last unique toxicity (for each compound).
4. Reporting the average of all unique toxicity values (for each compound).

Last two reported values are suitable for QSAR modelling. The last unique toxicity value (#3) is considered "natural", because it can be directly related to experimental measurement. The average of all unique toxicity values (#4) is considered "synthetic". When the latter is presented in chemical data tables then it should be accompanied by appropriate informational comment.

The timeline algorithm can be applied to any literature collection of publications, but the best results are obtained when the following two conditions are well satisfied: (1) the collection extends to as early dates as possible, (2) the collection is complete between the start date and the end date. In the current work both assumptions are met with great confidence. It is important to note that "timeline" approach establishes the unique ranking of publications for every chemical compound in the dataset. The literature collection of publications starts with the year of 1980 and to our best knowledge it includes all primary publications between the time period 1980–2011. When there are new primary publications coming out that naturally belong to this literature collection of publications then it will be straightforward to extend the affected timelines with new toxicity values. See also note in Supplementary Material regarding step size for timeline and on comparison of toxicity values regarding their uniqueness.

It has been estimated that the variance of *T. pyriformis* toxicity protocols is about 0.3 log units [38]. In the light of this it becomes evident that there is not much practical sense in trying to distinguish toxicity values below certain threshold distance, which could be set around 0.01–0.03 log units. Therefore one can link together all toxicity values which are less than 0.01 log units apart from one another. The resulting chain of toxicity values may span a range of 0.02 or more log units. For example, the members of the following chain of toxicity values [0.21, 0.2148, 0.22] are considered to represent the same unique toxicity value when applying the timeline algorithm. One should note that literature provides experimental values sometimes up to the four decimal places, resulting from mathematical procedures of fitting curves or taking averages, and are provided with excessive accuracy not considering the experimental error. However despite of this we have chosen to keep all decimal places after comma for values as explicitly provided by the primary publication. This is needed to allow full data reproduction while building up data timelines (data chronology) and because of the data curation orientation of current manuscript. Users of this data should perform appropriate rounding, similarly to the final result described in the chapter "Description of the final dataset".

Results and discussion

Literature collection of *T. pyriformis* primary publications

The literature collection contains 86 primary publications [17, 33–36, 38–118] (Table 1). While analyzing literature the aim was to remove all secondary publications. In practice, however, it was discovered that three of them [60, 86, 92] should be retained, because they contain new information, which comes in the form of added arithmetic precision of toxicity values. It is difficult to justify how three secondary publications [60, 86, 92] are able to report toxicity values using three decimal places, when their "parent" (i.e. the referred to) primary publications report toxicity values using two decimal places. The plausible explanation can be that the data has been retrieved from an external undisclosed and/or unpublished source, most likely from a spreadsheet file shared between research groups via private communication means. During the literature search a number of secondary publications were discovered (not included in the literature collection of primarily publications) that habitually appended an extra zero to all toxicity values in order to create the illusion of "higher arithmetic precision" data.

As expected, the central group in publications is lead by Prof. T.W. Schultz who is affiliated with 83 out of 86 publications, including the primary authorship of 42 publications. The only publication that is truly "independent" is the recent work of Böhme et al. [115], whereas the other two are the already familiar secondary publications [86, 93]. Böhme et al. [115] declare to have employed in-house *T. pyriformis* bioassay (based on the TETRATOX toxicity protocol [36]) to measure the toxicity values of four previously missing chemical compounds. However, literature search finds `oct-2-ynal` [99], `3-phenylprop-2-ynal` [99, 110] and `pent-1-en-3-one` [59, 102, 107, 113] in earlier

Table 1 Literature collection of 86 primary publications

Years	Reference(s)	Years	Reference(s)	Years	Reference(s)
1980	[17]	1993	[52, 53]	2003	[92–98]
1982	[39]	1994	[54, 55]	2004	[99–102]
1983	[33]	1995	[56–59]	2005	[103–108]
1985	[40–43]	1996	[35, 60–62]	2006	[109]
1986	[44]	1997	[36, 63–66]	2007	[110–113]
1987	[45]	1998	[67–71]	2008	[114]
1989	[46, 47]	1999	[72–77]	2010	[115–117]
1990	[34, 48, 49]	2000	[78]	2011	[118]
1991	[50]	2001	[79–85]		
1992	[51]	2002	[38, 86–91]		

literature, leaving `2-methylprop-2-enal` as a sole truly "independent" chemical compound.

Available data in PDF documents

The collection of QsarDB archives corresponding to the literature collection of publications is available in the QsarDB repository [119]. The aim was to reproduce the original chemical data tables as closely as possible. The only exception was the CAS RN, which was corrected either by replacing or interchanging one or two digits so that all values would pass the validation via the "check digit verification" [32]. The collection of QsarDB archives serves as source for the alternative data curation and data assembly workflows.

Table 2 provides the summary of information in the data tables of PDF documents. It can be seen that the information coverage is not uniform. By design, chemical names and toxicity values have to be fully covered. Chemical name can be cross-validated against CAS RN most of the time, but not against SMILES, which is too infrequent (Table 2). Toxicity value has no counterpart with sufficient coverage. Some cross-validation can be performed via toxicity value references, which can be either table-level (reference available only for table) or table row-level (reference available for each chemical in the table) (Table 2). This reference information can be used by programmatic agents to search for random transcription errors. However, processing of table-level references may become challenging if the referenced publications report disagreeing toxicity values (e.g. some chemical compounds appear to be taken from one publication, some from another).

Chemical data tables often include relevant molecular descriptors, especially when there is a need to support the discussion about embedded QSAR model(s). The majority of *T. pyriformis* acute aquatic toxicity QSAR models employ the hydrophobicity represented by the 1-octanol/water partitioning coefficient ($\log K_{ow}$) as the leading molecular descriptor (Table 2). For example, the baseline toxicity concept [37] rests entirely on the nearly linear relationship between the hydrophobicity and the acute aquatic toxicity of a chemical compound. Hydrophobicity values could be retrieved from experimental databases [120] or calculated from the chemical structure using specialized software [121, 122]. The general opinion appears to be that the experimental hydrophobicity values suit the modeling needs better than the calculated hydrophobicity values. This is mostly because the dominating algorithms implement atom- or fragment-based summation methods, which tend to over-simplify the underlying phenomenon. In the current study the hydrophobicity values (and some other easily reproducible molecular descriptor values) are used for chemical structure elucidation. A conflicting or an ambiguous chemical structure can be often settled by enumerating all candidates, and comparing their locally retrieved or locally calculated molecular descriptor values against what is reported in the chemical data table.

Sometimes it is desirable to partition the complete collection of publications into more focused sub-collections. The first possible sub-collection is the toxicity protocol reference (Fig. 4a; Table 2; 78 appearances), which is the bibliographic reference of the publication which introduces and characterizes a particular experimental setup. The second sub-collections is the toxicity endpoint (Fig. 4b; Table 2; 84 appearances) which characterizes a particular experiment. There is a considerable interdependence between two sub-collections (Table S1, 'S' in front of the table number designates table in Supplementary Material) in a sense that newer toxicity protocols employ shorter test durations.

Over the time there have been total of five toxicity protocols [17, 33–36] in use. Arranging them into a timeline (Fig. 4a) indicates a fairly ordered succession. The timeline reveals interesting detail that there is a considerable delay (typically 2 years) between the time when a toxicity protocol is published and the time when the measurement results start to flow into publications. Every toxicity protocol has had its time period of dominance ranging from (1) 1980–1982, (2) 1983–1992, (3) 1993–1996, (4) 1997–1998, (5) 1999–2011 (Fig. 4a). For example, the most recent toxicity protocol was published in 1997 [33] and became dominant in 1999. It differs from its predecessor by slightly higher pH and increased volume of inoculating culture (Table S2).

Table 2 The summary of recovered chemical data and metadata

Chemical data or metadata	Count	Chemical data or metadata	Count
Chemical name	86	Toxicity value	86
CAS RN	77	Toxicity value reference (table-level)	11
SMILES	7	Toxicity value reference (table row-level)	12
Toxicity protocol reference	78	Hydrophobicity value	65
Toxicity endpoint	84	QSAR model(s)	54

Fig. 4 Timetables of **a** toxicity protocols and **b** toxicity endpoints (*Capital letters* in the *boxes* indicated colour pattern for *black* and *white* representation: *Magenta*, *Red*, *Orange*, *Yellow*, *Green*, *Blue*, and *Pink*)

Curation scenarios for chemical structures

The complete collection of curation cases, together with the needed evidence, is presented in the Supporting Material (Table S3). In total there were 64 curation scenarios for chemical structures that makes 3.09 % from all 2,072 chemicals. The detailed analysis of curation cases allows grouping them into five general scenarios, summarized in following chapters and discussed in more detail in Supporting information.

Invalid or deficient chemical names

The largest number of structure curation cases is related to invalid or deficient chemical names. Typical examples are the violation of valence rules and the omission of substituent locations. Such conflicts can be solved with high confidence within single data table by performing a CAS RN resolution. Good illustrative example is `nonylphenol`, that in chemical industry designates a mixture of branched ortho- and para-isomers [123] but in all 18 publications appears to be para-isomer when CAS RN and the toxicity value are compared (see discussion and Table S4A in Supporting information). The second example of conflict resolution is based on the evidence which is unavailable at the level of a single publication, but is readily available at the level of the complete literature collection of publications. This is given by `dichlorodinitrobenzene` that appears in only one publication [67] but can be in variety of isomers depending on the substitution pattern (see Table S4B and related discussion in Supplementary Material).

Conflicts between chemical name, CAS RN and SMILES

The majority of observed conflicts occur between chemical name and CAS RN. As pointed out above, SMILES is too infrequent to portend serious trouble (Table 2). However when all three are present in publications in various combinations the conflict can not be resolved with the help of literature as in case of `3-(2-bromoacetyl)thiophene` (see Table S5A, and discussion in Supplementary Material). The case was discussed with the measuring laboratory, where the backtracking of records yielded that this chemical compound should be identified as *`2-bromo-1-(thiophen-3-yl)ethan-1-one`* (i.e. the 3-isomer) [124].

Conflicts when chemical name and CAS RN are valid but represent a different chemical

More substantial conflicts arise when both chemical name and the CAS RN are valid, but represent different chemical compound. Typically such cases are related to disagreements over the nature of the molecular skeleton or substituent locations. Literature search also reveals that there are some primary publications (e.g. [35, 75, 79, 80]) that contain relatively more such conflicts. It is likely that these chemical data tables were typed in manually using spreadsheet software, and the misplacement of substituents comes from transcription errors. Also, the auto-fill and auto-completion capabilities of modern spreadsheet software may easily lead to data intermingling. One may call such compounds "phantom" chemicals, i.e. chemicals that appear in the tables but are not measured in reality. A good example is `2-sec-butylphenol`, that after careful analysis it turns out to be `4-sec-butylphenol` (see Table S6A and respective discussion in Supplementary Material). Similar example is for `2,5-dinitro-1-naphthol` that had experimental value fitting to its weak acid respiratory uncouplers analogue of 2,4-dinito isomer (see Table S6B and related discussion in Supplementary Material). Indeed, the measuring laboratory confirmed our findings that they have only experimentally measured the 2,4-dinitro isomer and have two results associated with it [124]. The largest number of inconsistencies between the chemical name and the CAS RN are found with three `tetrachlorophenol` isomers that is in detail explained in Supplementary Material and Table S6C.

It was pointed out earlier that in some conflict situations, while choosing between two chemical structures, can be resolved by calculating molecular descriptors. For instance the chemical name `cinnamyl isothiocyanate` appears both in the text and in the chemical data table of one publication [108]. However, this chemical name is in conflict with the associated CAS RN of 19495-08-0 and SMILES of "c1ccccc1C = CC(= O)N = C=S", which correspond to `cinnamoyl isothiocyanate`. Literature search does not find any more appearances of either candidate. The evidence about the presence or absence of the carbonyl group in the chemical structure can be obtained in local data table by attempting to reproduce the reported hydrophobicity ($\log K_{ow}$) values. The original CLOGP version 3.55 software is unavailable, but the probation and estimation using KOWWIN version 1.68 software predicts the presence of a highly hydrophobic functional group. Also, both the original and estimated $\log K_{ow}$ values of `benzoyl isothiocyanate` and `cinnamyl isothiocyanate` differ by about 0.7 units [108], which means that these two chemical structures differ from one another by a fragment that consists of two olefinic carbonyl atoms. The identity of *`3-phenylprop-2-enonyl isothiocyanate`* was later confirmed by the measuring laboratory [124]. Another example where calculated molecular descriptors are of considerable utility is the case of elaborating the chemical structure of `dibutyl isophthalate`, described in more detail in Supplementary Material.

Restoration of auxiliary information of chemical structures

When chemicals are transferred from one publication to other it is very easy to loose auxiliary information, like compound being in salt form or "neutralized" vs. "non-neutralized". So therefore one needs to find publication of its earliest appearance. Good example for the first case is `2,4-Diaminophenol` that appears in publications in "raw" and in dihydrochloride salt form (see Table S7A and analysis in Supplementary Material) and finally, based on evidence, is qualified as the last one. Aminoalkanols provide example for "neutralized" or "non-neutralized" forms, where nature of `N,N-diethylethanolamine` is decided based on the toxicity ranges of neutralized and non-neutralized amino alcohols (see Table S7B and related analysis in Supplementary Material).

Restoration of stereochemistry information

The determination of geometric and optical isomers is effective only over all data tables. Geometric isomerism is a common feature, which is easy to process due to actionable shortcuts. For small organic molecules it is often the case that every individual isomer has a well-established trivial name. Trivial names are the most useful when they appear in pairs "that belong together" or are supported by the CAS RN. Geometric isomerism, even if there is only one stereocentre in the chemical structure, manifests itself readily in the target activity/property. As a rule, the cis-isomer is more compact than the trans-isomer, which has a direct effect on their partitioning properties and, in turn, on the baseline toxicity component of their overall toxicity. Optical isomerism is a less common feature and is more tedious to process.

Example of chemical compound is the one with the stereochemistry-less systematic name of `2-butenal` and solved by its trivial name of `crotonaldehyde` in another publication (see Table S8A and analysis in Supplementary Material). In given data tables the use of trivial names is especially widespread with unsaturated short-chain aliphatic carboxylic acids and their esters (Table S9), which can be explained by their relative compactness. The second example is provided by the *`2-Methylpent-2-enal`* where it was deemed necessary to remove existing stereochemistry information (see Table S8B and respective analysis in Supplementary Material).

Curation scenarios for toxicity values

In the case of multiple laboratory measurements the curation of toxicity values is needed to find out whether data point deviates from the general trend provided by the similar compounds. Often the curation of toxicity values is more needed to spot out random errors caused by the manual transfer of numerical values between scientific works and recycling of publications for building different combinations chemical data series. Those random errors are very human in nature and are typically the following:

- Changing the sign, the negative value to positive (and vice versa) by omitting the leading minus sign.
- Interchanging digits (inside the number).
- Changing a digit (inside the number), incrementing or decrementing the value by one unit. Also, changing a digit (in decimal fraction) to another digit that is visually similar to it.
- Interchanging the locations of data cells in chemical data tables.

Curation scenarios for toxicity values are summarized in following chapters and discussed with additional detail is Supplementary Material. In total there were 11 curation scenarios for toxicity values that make 0.53 % from all 2,072 chemicals and more prominently one per curated chemical structure. Based on the given dataset one can conclude that the problems with the representation of chemical structures in scientific literature are much more common than the problems with the representation of experimental values.

Broken value references

Recycling and reusing chemical and experimental data from one publication to another can cause so called broken references that are most likely the caused by the errors in data management. One of the most fruitful operations to find such inconsistencies is the cross-validation of toxicity values according to their toxicity value references, i.e. comparing table and row level references in publication where toxicity value is provided. Table-level toxicity references are however often proved to be ambiguous, because a particular chemical compound may appear in several referenced publications. Additionally, they fail if a reported toxicity value represents a previously known or a new experimental measurement. Table row-level toxicity value references do not suffer from any of the above concerns; however they can point into nothingness in a sense that the chemical data table that is referred to simply does not contain a matching row. In the current literature collection of publications there are more than three broken references found in 2 publications [83, 86]. Examples of broken value references are given in Supplementary Material (Table S10 and the respective discussion).

Changing the sign of a number

Changing the sign of the number by accident or systematically may not be with severe influence if the change is within the estimated variance of *T. pyriformis* toxicity protocols (about 0.3 log units). However it may influence decisions if this range is exceeded. The example that provides typical human error is the one by Jaworska and Schultz [52] that contains a toxicity value of −0.67 for `2-nitrophenol` and 0.19 for `4-methylphenol`. The literature search finds massive evidence (not shown) that both toxicity values should be of opposite sign. These two chemical compounds appear one after another in the published chemical data table ("Table 1", chemicals "48" and "49", respectively) [52]. Therefore, the cause of the problem is a single minus sign that has been inserted one row too high. Next examples are not that trivial and assume some toxicological background from the data curator. First is provided by `prop-2-yn-1-ol`, where the difference of over 2 log units is too much and as `propargylic alcohol` is known to act by a pro-electrophile mechanism of action it must have negative sign (see Table S11A and respective analysis in Supplementary Material). Second example (Table S11B), where more characteristic value for the polar narcosis mechanism of action is elucidated is given by the `2-[(hydroxyimino)methyl]phenol`, showing also that when mistake is made in some central primary publication it travels to the secondary publication and is not spotted easily afterwards.

Interchanging or changing digits inside a number

Interchanging or changing digits inside a number are most likely the human errors but may also cause severe problems. An example where the ordering of digits has changed between two consecutive publications is provided by `3-amino-4-hydroxybenzene-1-sulfonic acid` (see Table S12A and respective discussion in Supplementary Material). An example of a randomly reduced toxicity value is provided by `4,6-dinitrobenzene-1,2,3-triol`, where analysis of literature shows that the toxicity of this most reactive chemical compound should be about 1.0 log unit higher and this brings it out from outlier status in original publication and puts into the perspective with similar compounds (Table S12B and respective discussion in Supplementary Material).

Deviation from trend line

The acute aquatic toxicity of a chemical compound has a toxicodynamic component (i.e. the uptake of a xenobiotic from the environment) and a toxicokinetic component (i.e.

the physicochemical interaction of the xenobiotic at the site of action) [125]. The majority of chemical compounds act by non-reactive (i.e. narcosis) mechanisms of action which are limited by the toxicodynamic component. Their toxicity is proportional to their hydrophobicity and exhibits a nearly linear relationship in a fairly wide range. When employing the 1-octanol/water partitioning coefficient ($\log K_{ow}$) as a hydrophobicity scale, then the linear relationship holds true in a $\log K_{ow}$ range from -1 to 5 units [89].

The above linear relationship provides specific quality criteria. The simplest procedure is to verify that the toxicity of chemical compounds increases as the number of methylene groups CH_2 increases. A more advanced procedure would also verify that the increase happens in legitimate step sizes. The precise step size can be estimated using the baseline toxicity regression equation [114], but it should be about the same for most common chemical classes. For example, it is very hard to justify a single toxicity value that lies well below the baseline toxicity regression line. It is equally hard to justify a single toxicity value that lies well above the baseline toxicity regression line in a high hydrophobicity region where $\log K_{ow}$ value is greater than 3.5 or 4.

The search for homologous series is suitable for programmatic agents. The implemented programmatic agent is conditioned to "terminally" extend unbranched hydrocarbon chains. A single chemical compound may participate in several homologous series if it contains several extensible hydrocarbon chains. For example, `methyl formate` may be extended both along the alcohol residue (e.g. `methyl formate`, `ethyl formate`) and the carboxylic acid residue (e.g. `methyl formate`, `methyl acetate`). The agent raises a warning if the toxicity values of two successive members differ less than the mandatory threshold of 0.10 log units. All warnings are reviewed by the human agent and acted upon as appropriate. Possible example is provided by five-membered homologous series for 1-chloroalkanes and 1-bromoalkanes (see Table S13 and respective discussion).

Data assembly and consolidated dataset

After all data curation activities have been completed and individual data tables are corrected it is possible to start working towards assembling a new homogeneous data set.

Description of the final dataset

All observations are arranged into a data table (Table S17, in Supplementary Material). The table includes 2,060 unique InChI-s and 2,072 unique PIN-s. The difference is due to the neutralized and non-neutralised 12 aminoalkanols that have identical InChI and CAS RN. The neutralization of aminoalakanols is performed by the measuring laboratory in agreement with the toxicity protocol and its details may vary. The neutralization is known to have direct effect on the *T. pyriformis* toxicity [71, 89], which is why both forms are considered separate toxicochemical entities. In the data table, the neutralized form has a qualifier "neutralized" appended to its PIN, whereas the non-neutralized form has no qualifier. The rows in the table are sorted in ascending order by the chemical formula (modified Hill system, ordering of isomers is unspecified).

The number of columns in full table (available up on request) is determined by the number of publications. The hands-on experience with the data has revealed that the literature collection of 86 primary publications contains several (at least 10) redundant publications. In this context, a publication is redundant if it does not bring in any new information, that is, does not report any toxicity values that had not been reported by any of earlier publications. There is no need to worry about purging redundant publications, because all of the following data assembly procedures are independent of the size of the collection of publications (i.e. they are data intensive, not data extensive procedures). For example, the calculation of average and median toxicity values for a chemical compound is performed on the basis of its data timeline, which is size invariant.

The majority of reported toxicity values are numerical values, but there are also two kinds of text values to designate special cases. The first text value is "N/A" (a shorthand for "Not Applicable/Available"), designating a toxicity value that could not be measured due to technical difficulties. This text value appears only in the latest publication [118], which is concerned with the reactivity and toxicity of aromatic compounds transformable to quinone-type Michael acceptors. The technical difficulties were related to colour interference, which prevents toxicity quantification via the standard spectrophotometric method [36]. Technical difficulties are typically related to the instability of a chemical compound in aqueous solution over the test period (reactivity, decomposition via hydrolysis, volatility etc.). The normal practice is to publish nothing about such failed experiments.

The second text value is "NTAS" (a shorthand for "Not Toxic At Saturation", also known as "No Effect At Saturation"), which designates a toxicity value that could not be measured because the chemical compound was either too hydrophilic (i.e. highly soluble in water) or too hydrophobic (i.e. insoluble in water), so that it did not attain a stationary concentration required for the construction of a concentration–response curve [36]. This phenomenon is often linked with the baseline toxicity concept in a complementary way. Namely, the toxicity of a chemical compound is considered non-quantifiable, if its

hydrophobicity value falls outside of the hydrophobicity range where the linear relationship is valid. When employing the 1-octanol/water partitioning coefficient ($\log K_{ow}$) as a hydrophobicity scale then these are all chemical compounds whose $\log K_{ow}$ value is less than −1 or greater than 5 [89]. This notion is used to optimize and prioritize experimental measurements. However, extra care is warranted with less common chemical structures, because the predicted $\log K_{ow}$ values may not be very reliable in bending regions. Toxicity values of "NTAS" do not lend themselves easily for QSAR modelling.

The majority of "N/A" and "NTAS" text values are already presented as such in chemical data tables. However, during the conversion of PDF documents to QsarDB archives, the following inexact numerical values in two publications [109, 112] were replaced with "NTAS" text values (Table S14). In theory, it would be possible to "save" these inexact numerical values by enabling the use of open-interval notation. In practice, they are too infrequent (less than 0.5 % of all observations) to merit adding extra complexity to existing data management procedures.

Four out of seven chemical compounds that are listed in Table S14 are "saved" for quantitative analysis indirectly, because literature search finds them in other publications where they have exact numerical values (Table S15). However, taking the above into account extra caution is recommended with trusting them.

There are 9,449 toxicity values over all primary publications, which divide into 94 text values and 9,355 numerical values. When the duplicates are disregarded (according to the extended numerical comparison procedure, see below) then there are 2,498 unique toxicity values, which divide into 58 text values and 2,440 numerical values. These and above numbers can be used to describe current literature collection of publications. First, the ratio of toxicity values to chemical compounds stands at 4.56 (9,449/2,072 = 4.560). Second, the ratio of unique toxicity values to chemical compounds stands at 1.21 (2,498/2,072 = 1.206). It shows that on average, every chemical compound has been experimentally measured 1.21 times and (re)cited 3.35 times (4.56 − 1.21 = 3.35) in time period between 1980 and 2011. The average number of experimental measurements is surprisingly low, which suggests that for the majority of chemical compounds there is only one result available, which in turn suggests that extensive data analysis of repeated measurements is not required for given data assembly.

Data timelines for the single compounds

Data timelines of single compounds do provide means to analyse occurrence of toxicity values over the time. For example, `phenol` appears in 30 publications (Fig. 5). The application of the timeline algorithm (see the chapter "Materials and Methods") identifies five unique toxicity values, which have been first published in the years of 1985 [41], 1994 [54], 1996 [35, 62], 2001 [83] and 2002 [91] (Fig. 5). The expectation was to see an orderly phasing-in and phasing-out of toxicity values (analogously to the relatively orderly succession of toxicity protocol references and toxicity endpoints seen on Fig. 4a,b, respectively). Instead of this, they are (re)cited in no particular order (Fig. 5). For example, the most "fit for survival" toxicity value for `phenol` appears to be the second unique toxicity value of −0.21, which effectively outcompetes latest three toxicity values. Most likely the spread of a particular toxicity value depends on the accessibility of the publication. For example, the last unique toxicity value of −0.46 appears in one publication [91], which is concerned with

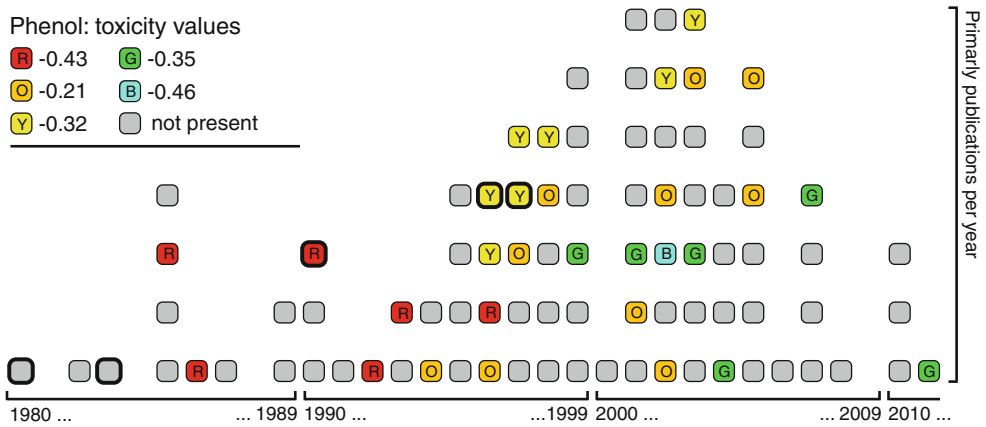

Fig. 5 Timetable of reported toxicity values for `phenol`: *Red* "−0.43". −0.4310 [41, 44, 51, 60], −0.43 [34, 52], *Orange* " −0.21". −0.2114 [54], −0.21 [38, 69, 79, 87, 103, 105], −0.208 [61, 65, 92]; *Yellow* " −0.32". −0.321 [62, 86], −0.32 [35, 36, 63, 67, 96]; *Green* " −0.35". −0.3533 [83], −0.35 [75, 97, 100, 110, 118]; *Cyan* " −0.46". −0.46 [91] (*Capital letters* in the *boxes* indicated colour pattern for *black* and *white* representation)

the regression comparisons of *T. pyriformis* and *P. reticulata* toxicity, without any reuse in other later publications.

Similar data timelines are created for all 2,072 chemical compounds. Table 3 gives the summary of all data timelines for the original numerical comparison procedure and four extended numerical comparison procedures. The length of the data timeline is the number of unique toxicity values (both text and numerical values) that it contains. The length of 0 means no numerical values (i.e. there are text values "N/A" and "NTAS"). The extended procedure contains an additional post-processing step, which joins unique numerical toxicity values whose difference is less than or equal to the specified separation threshold value (Δ_{sep}). The need for post-processing was inspired by the fact that original data timelines seemed "noisy", because too many of them contained unique toxicity values that were only 0.01 or 0.02 log units apart. The best balance between signal and noise was achieved by employing the separation threshold value of 0.02 log units (Table 3). For example, the separation threshold values of 0.01 and 0.02 log units do not affect the original length of `phenol` data timeline, whereas the separation threshold value of 0.03 log units would reduce it from 5 to 3 (Table 3). A trial run was also performed for the separation value threshold of 0.30 log units, which, as noted above, is the estimated variance of *T. pyriformis* toxicity protocols. For 95 % of chemical compounds (1,957/2,072 = 0.945) the length of data timelines collapses to 1 under this scenario (Table 3). This suggests that the majority of chemical compounds could be adequately quantified after any of their experimentally measured toxicity values, provided that they are curated toxicity values. Obviously, this simplification applies better for global data sets, not for small mechanism-oriented data sets. The decision whether to use "latest" unique measured value or "average" of unique values over the entire timeline has to be made by the modeller depending on the nature of the QSAR modelling task at hand. For example, if the target is local QSAR (e.g. investigating the mechanism or mode of action of a narrow class of chemical compounds) then "latest" unique measured values are likely to provide higher quality. However, if the target is global QSAR (e.g. investigating maximally diverse set of chemical compounds), then the "average" of all unique measured values (over the entire timeline) provides better signal-to-noise ratio.

The shortening of data timelines by joining adjacent unique toxicity values creates new value type which is closed-interval numerical value. For example, the original data timeline of `dodecan-1-ol` consists of 5 unique toxicity values of 2.08, 2.07, 2.18, 2.16 and 2.15 (Table 3). The extended timeline ($\Delta_{sep} = 0.02$) consists of only 2 unique toxicity values of [2.07, 2.08] and [2.15, 2.18] (Table 3). The closed-interval notation is suitable for data storage and management, but not for data analysis. The most straightforward approach to convert such values to ordinary numerical values would be to calculate the centre (i.e. the midpoint) of the interval. In order to avoid confusion, it is very important to employ proper annotation in chemical data tables to distinguish between "natural" and "synthetic" associated target activity/property values. Otherwise it may happen that average values are unknowingly averaged many times more, which may greatly reduce their information content.

The number of appearances of a chemical compound in the literature collection of publications (Table S16) gives broad estimate how much they have been used in gaining scientific knowledge. It is significant that over 25 % of all chemical compounds (556/2,072 = 0.268) only appear in one primary publication and consequently it can be hypothesized that they found no use also in secondary publications. In another words, 25 % of all chemical compounds have had their toxicity values published and

Table 3 The summary of data timelines under different numerical comparison procedures

Length	Original	Extended ($\Delta_{sep} = 0.01$)	Extended ($\Delta_{sep} = 0.02$)	Extended ($\Delta_{sep} = 0.03$)	Extended ($\Delta_{sep} = 0.30$)
0	50	50	50	50	50
1	1,582	1,640	1,678	1,699	1,957
2	334	294	275	261	64
3	80	70	58	55	1
4	21	16	9	6 (E)	0
5	3 (A)	2 (C)	2 (D)	1 (F)	0
6	2 (B)	0	0	0	0
Total	2,651	2,562	2,498	2,465	2,138

A—`quinoline`, `phenol`, `dodecan-1-ol`, B—`pentachlorophenol`, `octan-1-ol`, C—`phenol`, `octan-1-ol`, D—`phenol`, `octan-1-ol`, E—`2-phenylethan-1-ol`, `nitrobenzene`, `4-chloro-2-nitrophenol`, `1-chloro-4-nitrobenzene`, `1-methyl-3-nitrobenzene`, `propan-2-one`, F—`octan-1-ol`

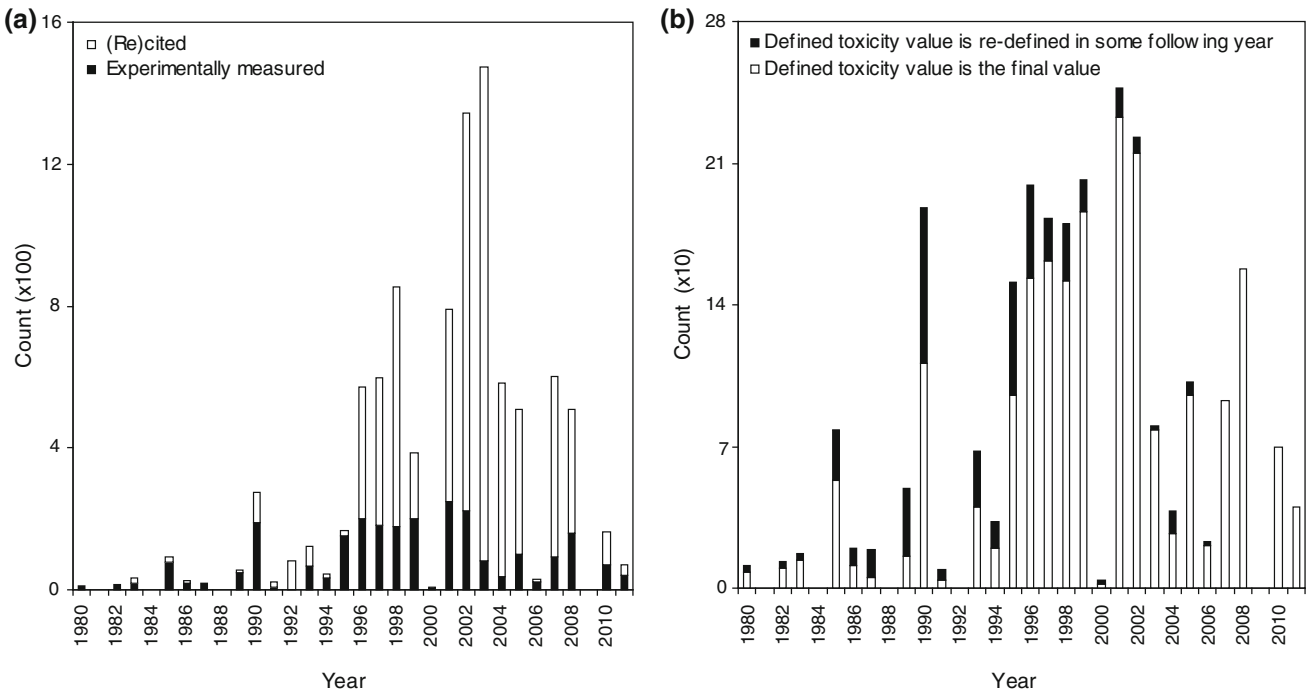

Fig. 6 The division of chemical compound appearances between **a** type of publication and **b** type of experimental measurement

discussed at one point in time, and have most likely contributed very little towards the broader understanding of acute aquatic toxicity afterwards.

In further analysis of 9,449 toxicity values the ratio between experimental measurements and (re)citation of toxicity values exhibits three characteristic patterns (Fig. 6a) of appearance of new toxicity values. In time period between 1980 and 1995 the ratio is strongly trending (656/308 = 2.130) towards the introduction of new toxicity values. In chemical terms, the experimental measurements were aimed at providing the adequate coverage of the most popular chemical classes. In time period between 1996 and 2000 the ratio is balanced (768/1,647 = 0.466). Data sets from the previous time period have been enhanced with less common chemical structures. Also, there have been attempts to bridge different chemical classes by poly-functional chemical structures. In time period between 2001 and 2011 the ratio is strongly reducing (1,074/4,996 = 0.215) towards the other end, the referencing of existing toxicity values. The focus has shifted from the experimental aspects to the theoretical aspects (i.e. the methodology of QSAR modelling). The average size of data sets has increased rapidly and often exceeds 500 chemical compounds in order to support novel statistical and data mining methods. Clearly, most of the increase is attributable to data reuse. The spike around 2002 and 2003 is largely driven by the EU 5FP project IMAGETOX [126], which included *T. pyriformis* as a model organism for acute aquatic toxicity. The vast majority of the experimental measurements were carried out under the umbrella of the TETRATOX project [14] that was conceived with the goal to experimentally measure about 100 chemical compounds per year [35]. Assuming that this goal has been feasible for the past 30 years there should be over 3,000 chemical compounds experimentally measured by now. This number is roughly 30 % greater than the number of chemical compounds in the current work, which gives hope that there can be additions to the data set in the coming years. The rate of publication (Fig. 6b) can be described with the same three time intervals as above. The ratio between the published and experimentally measured toxicity values is 0.41 ((387 + 269)/1,600 = 0.410) for the first time period between 1980 and 1995. This ratio nearly quadruples to 1.54 ((655 + 113)/500 = 1.536) for the second time period between 1996 and 2000, and finally eases off to "normal" 1.07 ((1,030 + 44)/1,100 = 1.074) for the third time period between 2001 and 2011.

The current experience shows that it is a valid practice to combine toxicity values that have been measured after different toxicity protocols (e.g. Figure 4a) and have different toxicity endpoints (e.g. Figure 4b). This is supported by the practice of the measuring laboratory since the beginning of releasing experimental values, so that all the toxicity values that are available in the present literature collection of publications are qualitatively inseparable from one another based on any objective or subjective criteria. Data timelines may provide a more revealing

answer to this question. Namely, as pointed out above, there are total 426 (2,498 − 2,072 = 426) toxicity values that may be regarded as obsolete (Fig. 6b; black stacks), because they have been superseded by newer experimentally measured toxicity values.

It is difficult to precisely quantify the human effort behind the current work (or make projections for new data sets). Broad estimates can be drawn by the following procedure. The first step, data set preparation, depends on the number of publications. Here, the retrieval, digitization and basic curation (e.g. visual inspection of chemical structures) of about two hundred articles consumed between two to four man-months. The second step, searching the collection of publications for possible conflicts is fully automatic and essentially runs in constant time. The third step, the analysis and resolution of conflict situations, depends on the nature of the endpoint and the experience and motivation of the scientist. Here, it consumed about two man-months, which is considered rather optimistic.

Conclusions

The scientific publishing process does not have a mechanism for making sure that the presented data is correct. Therefore data curation is important issue for data analysis tasks such as QSAR, which derives new knowledge from existing data and/or information. The current publication describes a systematic and reproducible workflow for the assembly of curated data sets based on scientific literature (PDF files). The focus of data curation was set on two most common problem areas which are correctness of chemical structure representations and associated toxicity values. The workflow is implemented as a collection of graphical and command-line utilities and is demonstrated on *T. pyriformis* acute aquatic toxicity endpoint. The current work clearly shows that scientific literature includes inconsistencies in both problem areas. The biggest contributor to the deterioration of data quality over time appears to be human factor (e.g. random errors, etc.).

Data curation can be described as a two-stage process. The first stage takes care of identifying the correct structure of chemical compounds and translating them to standardized human- and machine-readable representations. The second stage takes care of identifying the correct associated target activity/property value of chemical compounds. Both stages could be described as iterative, because they involve performing the same set of actions over and over again until all unhealthy structures and experimental values have been found and curated. Different persons are likely to produce different data sets (starting from the same data), because it takes considerable chemical and biological expertise to determine optimal end conditions. Similarly to over-training during data modelling, there is a risk of over-curation either by incompetence or over competence.

The curation of chemical structures aims to recover as much information as possible about every chemical compound. The most common chemical structure curation scenarios included (1) fixing invalid or deficient chemical names, (2) resolution of chemical name and CAS RN conflicts, (3) restoration of lost auxiliary and stereochemistry information. Among other things, it was possible to identify and eliminate several "phantom" chemical compounds. The goodness of the performed analysis ultimately determines at which level of theory it will be possible to calculate molecular descriptor values and perform other QSAR modelling work.

The curation of toxicity values aims to identify and correct numerical values that do not match known experimental measurement results. Incorrect values appear to come into existence mainly by transcription errors such as (1) changing the sign and (2) changing the order of digits. The curation of toxicity values helped to bring several chemical compounds that were previously thought to be outliers, back to normal state, which can be considered one of major achievements.

The assembly of a ready to use data set was performed using a novel so-called "timeline" approach. The underlying algorithm is capable of generating data sets with varying signal-to-noise ratios. In the current work the optimal results were obtained by applying relatively rigorous criteria. Additional strength of data timelines is in providing valuable insights into data development history, which may come in handy during the design of QSAR modelling exercises. Data timelines have the added benefit that they are easily extensible when new data becomes available. For example, the current data set can be updated incrementally with new chemical compounds when new primary publications about *T. pyriformis* acute aquatic toxicity are published. The size of the current curated data set is 2,072 chemical compounds (Tabel S17), which is roughly 30 % more than the size of previously published largest and most diverse *T. pyriformis* data sets.

Acknowledgments Estonian Science Foundation (Grant 7709) and Estonian Ministry for Education and Research (Grant SF0140031Bs09) for financial support. Authors are grateful to Prof. T.W. Schultz (University of Tennessee) for his assistance in resolving selected data points. Authors are thankful to Dr. Sulev Sild (University of Tartu, Estonia) for the discussion at final stages of manuscript preparation.

References

1. Nendza M, Aldenberg T, Benfenati E, Benigni R, Cronin MTD, Escher S, Fernandez A, Gabbert S, Giralt F, Hewitt M, Hrovat M, Jeram S, Kroese D, Madden JC, Mangelsdorf I, Rallo R, Roncaglioni A, Rorije E, Segner H, Simon-Hettich B, Vermeire

T (2010) Data quality assessment for in silico methods: a survey of approaches and needs. In: Cronin MTD, Madden JC (eds) Silico toxicology: principles and applications. The Royal Society of Chemistry, Cambridge, pp 59–117
2. Przybylak KR, Madden JC, Cronin MTD, Hewitt M (2012) Assessing toxicological data quality: basic principles, existing schemes and current limitations. SAR QSAR Environ Res 23:435–459
3. OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models, 37th joint meeting of the chemicals committee and working party on chemicals, pesticides and biotechnology (2004). http://www.oecd.org/dataoecd/33/37/37849783.pdf Accessed 10 Dec 2012
4. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inf 29:476–488
5. Young D, Martin T, Venkatapathy R, Harten P (2008) Are the chemical structures in your QSAR correct? QSAR Comb Sci 27:1337–1345
6. Fourches D, Muratov E, Tropsha A (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 50:1189–1204
7. Zhao CY, Boriani E, Chana A, Roncaglioni A, Benfenati E (2008) A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). Chemosphere 73:1701–1707
8. Clark RD, Waldman M (2012) Lions and tigers and bears, oh my! Three barriers to progress in computer-aided molecular design. J Comput Aided Mol Des 26:29–34
9. Bolton EE, Wang Y, Thiessen PA, Bryant SH (2008) PubChem: integrated platform of small molecules and biological activities. In Cornell W (ed) Annual reports in computational chemistry, volume 4, American Chemical Society, Washington, DC, Chapter 12. http://pubchem.ncbi.nlm.nih.gov/. Accessed 10 Dec 2012
10. ChemaIDpluss database. http://chem.sis.nlm.nih.gov/chemidplus/. Accessed 10 Dec 2012
11. Williams AJ, Ekins S (2011) A quality alert and call for improved curation of public chemistry databases. Drug Discov Today 16:747–750
12. Williams AJ, Ekins S, Tkachenko V (2012) Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. Drug Discov Today 17:685–701
13. Fu X, Wojak A, Neagu D, Ridley M, Travis K (2011) Data governance in predictive toxicology: a review. J Cheminf 3:24
14. TETRATOX web-site. http://www.vet.utk.edu/TETRATOX/ Accessed 10 Dec 2012
15. Google Scholar. http://scholar.google.com/. Accessed 10 Aug 2012
16. Selected Works of Terry W Schultz. http://works.bepress.com/terry_schultz/doctype.html#article. Accessed 10 Aug 2012
17. Schultz TW, Cajina-Quezada M, Dumont JN (1980) Structure-toxicity relationships of selected nitrogenous heterocyclic compounds. Arch Environ Contam Toxicol 9:591–598
18. ISO 32000-1:2008, Document management—portable document format—Part 1: PDF 1.7. http://www.iso.org/iso/catalogue_detail.htm?csnumber=51502. Accessed 10 Dec 2012
19. Apache PDFBox—Java PDF Library. http://pdfbox.apache.org/. Accessed 10 Dec 2012
20. QSAR DataBank. http://www.qsardb.org/. Accessed 10 Dec 2012
21. IUPAC project, Preferred names in the nomenclature of organic compounds. http://www.iupac.org/nc/home/projects/project-db/project-details.html?tx_wfqbe_pi1[project_nr]=2001-043-1-800. Accessed 10 Dec 2012
22. MarvinBeans Java chemoinformatics library, version 5.5.0. http://www.chemaxon.com/products/marvin/. Accessed 10 Dec 2012
23. Chemical Abstracts Service Registry. http://www.cas.org/content/chemical-substances. Accessed 10 Dec 2012
24. CAS Common Chemistry web service. http://www.commonchemistry.org/. Accessed 10 Dec 2012
25. NIH/CADD Chemical Identifier Resolver service. http://cactus.nci.nih.gov/chemical/structure/documentation. Accessed 10 Dec 2012
26. ChemSpider web service. http://www.chemspider.com/About-Services.aspx. Accessed 10 Dec 2012
27. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comp Sci 28:31–36
28. Daylight SMILES, Daylight Chemical Information Systems, Inc., Laguna Niguel (CA) USA. http://www.daylight.com/smiles/. Accessed 10 Dec 2012
29. OpenSMILES. http://www.opensmiles.org/. Accessed 10 Dec 2012
30. InChI Trust Website. http://www.inchi-trust.org/. Accessed 10 Dec 2012
31. QsarDB tools. http://qsardb.googlecode.com/. Accessed 10 Dec 2012
32. Check Digit Verification of CAS Registry Numbers. http://www.cas.org/content/chemical-substances/checkdig. Accessed 10 Dec 2012
33. Schultz TW (1983) Aquatic toxicology of nitrogen heterocyclic molecules: quantitative structure-activity relationships. In Nriagu JO (ed) Aquatic toxicology. Wiley, New York, pp 401–424
34. Schultz TW, Lin DT, Wilke TS, Arnold LM (1990) Quantitative structure-activity relationships for the Tetrahymena pyriformis population growth endpoint: a mechanism of action approach. In: Devillers J, Karcher W (eds) Practical applications of quantitative structure-activity relationships (QSAR) in environmental chemistry and toxicology. Joint Research Centre, Italy, pp 241–262
35. Schultz TW (1996) Tetrahymena in aquatic toxicology: QSARs and ecological hazard assessment. In: Berger S, Pauli W (eds) Proceedings of the international workshop on a protozoan test protocol with tetrahymena in aquatic toxicity testing. German Federal Environmental Agency, Germany, pp 31–65
36. Schultz TW (1997) TETRATOX: tetrahymena pyriformis population growth impairment endpoint—a surrogate for fish lethality. Toxicol Mech Meth 7:289–309
37. Könemann H (1981) Quantitative structure-activity relationships in fish toxicity studies Part 1: relationship for 50 industrial pollutants. Toxicology 19:209–221
38. Cronin MTD, Aptula AO, Duffy JC, Netzeva TI, Rowe PH, Valkova IV, Schultz TW (2002) Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to Tetrahymena pyriformis. Chemosphere 49:1201–1221
39. Schultz TW, Cajina-Quezada M (1982) Structure-toxicity relationships of selected nitrogenous heterocyclic compounds II. Dinitrogen molecules. Arch Environ Contam Toxicol 11:353–361
40. Schultz TW, Applehans FM (1985) Correlations for the acute toxicity of multiple nitrogen substituted aromatic molecules. Ecotox Environ Safe 10:75–85
41. Schultz TW, Riggin GW (1985) Predictive correlations for the toxicity of alkyl- and halogen-substituted phenols. Toxicol Lett 25:47–54
42. Schultz TW, Moulton BA (1985) Structure-activity relationships for nitrogen-containing aromatic molecules. Environ Toxicol Chem 4:353–359
43. Schultz TW, Moulton BA (1985) Structure-activity relationships of selected pyridines: I. Substituent constant analysis. Ecotox Environ Safe 10:97–111
44. Schultz TW, Holcombe GW, Phipps GL (1986) Relationships of quantitative structure-activity to comparative toxicity of selected

phenols in the Pimephales promelas and Tetrahymena pyriformis test systems. Ecotox Environ Saf 12:146–153
45. Schultz TW, Applehans FM, Riggin GW (1987) Structure-activity relationships of selected pyridines: III. Log Kow analysis. Ecotox Environ Saf 13:76–83
46. Schultz TW, Dawson DA, Lin DT (1989) Comparative toxicity of selected nitrogen-containing aromatic compounds in the Tetrahymena pyriformis and Pimephales promelas test systems. Chemosphere 18:2283–2291
47. Schultz TW, Arnold LM, Wilke TS, Moulton MP (1989) Relationships of quantitative structure-activity for normal aliphatic alcohols. Ecotox Environ Saf 19:243–253
48. Cajina-Quezada M, Schultz TW (1990) Structure-toxicity relationships for selected weak acid respiratory uncouplers. Aquat Toxicol 17:239–252
49. Schultz TW, Wyatt NL, Lin DT (1990) Structure-toxicity relationships for nonpolar narcotics: a comparison of data from the tetrahymena, photobacterium and pimephales systems. Bull Environ Contam Toxicol 44:67–72
50. Schultz TW, Wilke TS, Bryant SE, Hosein LM (1991) QSARs for selected aliphatic and aromatic amines. Sci Total Environ 109:581–587
51. Schultz TW, Lin DT, Wesley SK (1992) QSARs for monosubstituted phenols and the polar narcosis mechanism of toxicity. Quality Assur Good Pract Regul Law 1:132–143
52. Jaworska JS, Schultz TW (1993) Quantitative relationships of structure-activity and volume fraction for selected nonpolar and polar narcotic chemicals. SAR QSAR Environ Res 1:3–19
53. Schultz TW, Tichy M (1993) Structure-toxicity relationships for unsaturated alcohols to Tetrahymena pyriformis: C5 and C6 analogs and primary propargylic alcohols. Bull Environ Contam Toxicol 51:681–688
54. Bryant SE, Schultz TW (1994) Toxicological assessment of biotransformation products of pentachlorophenol: tetrahymena population growth impairment. Arch Environ Con Tox 26:299–303
55. Schultz TW, Kissel TS, Tichy M (1994) Structure-toxicity relationships for unsaturated alcohols to Tetrahymena pyriformis: 3-alkyn-1-ols and 2-alken-1-ols. Bull Environ Contam Toxicol 53:179–185
56. Cronin MTD, Bryant SE, Dearden JC, Schultz TW (1995) Quantitative structure-activity study of the toxicity of benzonitriles to the ciliate Tetrahymena pyriformis. SAR QSAR Environ Res 3:1–13
57. Dearden JC, Cronin MTD, Schultz TW, Lin DT (1995) QSAR study of the toxicity of nitrobenzenes to Tetrahymena pyriformis. QSAR Comb Sci 14:427–432
58. Jaworska JS, Hunter RS, Schultz TW (1995) Quantitative structure-toxicity relationships and volume fraction analyses for selected esters. Arch Environ Contam Toxicol 29:86–93
59. Schultz TW, Sinks GD, Hunter RS (1995) Structure-toxicity relationships for alkanones and alkenones. SAR QSAR Environ Res 3:27–36
60. Piršelová K, Baláž Š, Schultz TW (1996) Model-based QSAR for ionizable compounds: toxicity of phenols against Tetrahymena pyriformis. Arch Environ Con Tox 30:170–177
61. Cronin MTD, Schultz TW (1996) Structure-toxicity relationships for phenols to Tetrahymena pyriformis. Chemosphere 32:1453–1468
62. Schultz TW, Bearden AP, Jaworska JS (1996) A novel QSAR approach for estimating toxicity of phenols. SAR QSAR Environ Res 5:99–112
63. Bearden AP, Schultz TW (1997) Structure-activity relationships for Pimephales and Tetrahymena: a mechanism of action approach. Environ Toxicol Chem 16:1311–1317
64. Jaworska JS, Hunter RS, Gobble JR, Schultz TW (1997) Structure-activity relationships for diesters: aquatic toxicity to Tetrahymena. In: Schüürmann G, Chen F (eds) Quantitative structure-activity relationships in environmental sciences. SETAC Press, New York, pp 277–283
65. Schultz TW, Sinks GD, Cronin MTD (1997) Identification of mechanisms of toxic action of phenols to Tetrahymena pyriformis from molecular descriptors. In: Schüürmann G, Chen F (eds) Quantitative structure-activity relationships in environmental sciences. SETAC Press, New York, pp 329–342
66. Schultz TW, Sinks GD, Cronin MTD (1997) Quinone-induced toxicity to Tetrahymena: structure-activity relationships. Aquat Toxicol 39:267–278
67. Bearden AP, Schultz TW (1998) Comparison of Tetrahymena and Pimephales toxicity based on mechanism of action. SAR QSAR Environ Res 9:127–153
68. Cronin MTD, Gregory BW, Schultz TW (1998) Quantitative structure-activity analyses of nitrobenzene toxicity to Tetrahymena pyriformis. Chem Res Toxicol 11:902–908
69. Schultz TW, Sinks GD, Bearden AP. QSAR in aquatic toxicology: a mechanism of action approach comparing toxic potency to Pimephales promelas, Tetrahymena pyriformis, and Vibrio fischeri. In: Devillers J (ed) Comparative QSAR. Taylor & Francis, UK, pp 51–109
70. Schultz TW, Bearden AP (1998) Structure-toxicity relationships for selected naphthoquinones to Tetrahymena pyriformis. Bull Environ Contam Toxicol 61:405–410
71. Sinks GD, Carver TA, Schultz TW (1998) Structure-toxicity relationships for aminoalkanols: a comparison with alkanols and alkanamines. SAR QSAR Environ Res 9:217–228
72. Akers KS, Sinks GD, Schultz TW (1999) Structure-toxicity relationships for selected halogenated aliphatic chemicals. Environ Toxicol Pharmacol 7:33–39
73. Muccini M, Layton AC, Sayler GS, Schultz TW (1999) Aquatic toxicities of halogenated benzoic acids to Tetrahymena pyriformis. Bull Environ Contam Toxicol 62:616–622
74. Schultz TW, Cronin MTD (1999) Response-surface analyses for toxicity to Tetrahymena pyriformis: reactive carbonyl-containing aliphatic chemicals. J Chem Inf Comp Sci 39:304–309
75. Schultz TW (1999) Structure-toxicity relationships for benzenes evaluated with Tetrahymena pyriformis. Chem Res Toxicol 12:1262–1267
76. Schultz TW, DeWeese AD (1999) Structure-toxicity relationships for selected lactones to Tetrahymena pyriformis. Bull Environ Contam Toxicol 62:463–468
77. Seward JR, Schultz TW (1999) QSAR analyses of the toxicity of aliphatic carboxylic acids and salts to Tetrahymena Pyriformis. SAR QSAR Environ Res 10:557–568
78. Seward JR, Sinks GD, Schultz TW (2000) Population growth kinetics of Tetrahymena pyriformis exposed to selected pyridines. Europ J Protistol 36:139–149
79. Cronin MTD, Schultz TW (2001) Development of quantitative structure-activity relationships for the toxicity of aromatic compounds to Tetrahymena pyriformis: comparative assessment of the methodologies. Chem Res Toxicol 14:1284–1295
80. Cronin MTD, Manga N, Seward JR, Sinks GD, Schultz TW (2001) Parametrization of electrophilicity for the prediction of the toxicity of aromatic compounds. Chem Res Toxicol 14:1498–1505
81. DeWeese AD, Schultz TW (2001) Structure-activity relationships for aquatic toxicity to Tetrahymena: halogen-substituted aliphatic esters. Environ Toxicol 16:54–60
82. Schultz TW, Sinks GD, Miller LA (2001) Population growth impairment of sulfur-containing compounds to Tetrahymena pyriformis. Environ Toxicol 16:543–549
83. Seward JR, Sinks GD, Schultz TW (2001) Reproducibility of toxicity across mode of toxic action in the Tetrahymena population growth impairment assay. Aquat Toxicol 53:33–47

84. Seward JR, Cronin MTD, Schultz TW (2001) Structure-toxicity analyses of Tetrahymena pyriformis exposed to pyridines—an examination into extension of surface-response domains. SAR QSAR Environ Res 11:489–512
85. Sinks GD, Schultz TW (2001) Correlation of Tetrahymena and Pimephales toxicity: evaluation of 100 additional compounds. Environ Toxicol Chem 20:917–921
86. Baláž Š, Lukacova V (2002) Subcellular pharmacokinetics and its potential for library focusing. J Mol Graph Model 20:479–490
87. Aptula AO, Netzeva TI, Valkova IV, Cronin MTD, Schultz TW, Kühne R, Schüürmann G (2002) Multivariate discrimination between modes of toxic action of phenols. Quant Struct-Act Relat 21:12–22
88. Kaiser KL, Niculescu SP, Schultz TW (2002) Probabilistic neural network modeling of the toxicity of chemicals to Tetrahymena pyriformis with molecular fragment descriptors. SAR QSAR Environ Res 13:57–67
89. Schultz TW, Cronin MTD, Netzeva TI, Aptula AO (2002) Structure-toxicity relationships for aliphatic chemicals evaluated with Tetrahymena pyriformis. Chem Res Toxicol 15:1602–1609
90. Seward JR, Cronin MTD, Schultz TW (2002) The effect of precision of molecular orbital descriptors on toxicity modeling of selected pyridines. SAR QSAR Environ Res 13:325–340
91. Seward JR, Hamblen EL, Schultz TW (2002) Regression comparisons of Tetrahymena pyriformis and Poecilia reticulata toxicity. Chemosphere 47:93–101
92. Cottrell MB, Schultz TW (2003) Structure-toxicity relationships for methyl esters of cyanoacetic acids to Tetrahymena pyriformis. Bull Environ Contam Toxicol 70:549–556
93. Schüürmann G, Aptula AO, Kühne R, Ebert RU (2003) Stepwise discrimination between four modes of toxic action of phenols in the Tetrahymena pyriformis Assay. Chem Res Toxicol 16:974–987
94. Netzeva TI, Schultz TW, Aptula AO, Cronin MTD (2003) Partial least squares modelling of the acute toxicity of aliphatic compounds to Tetrahymena pyriformis. SAR QSAR Environ Res 14:265–283
95. Netzeva TI, Aptula AO, Chaudary SH, Duffy JC, Schultz TW, Schüürmann G, Cronin MTD (2003) Structure-activity relationships for the toxicity of substituted poly-hydroxylated benzenes to Tetrahymena pyriformis: influence of free radical formation. QSAR Comb Sci 22:575–582
96. Ren S, Frymier PD, Schultz TW (2003) An exploratory study of the use of multivariate techniques to determine mechanisms of toxic action. Ecotox Environ Saf 55:86–97
97. Schultz TW, Netzeva TI, Cronin MTD (2003) Selection of data sets for QSARS: analyses of tetrahymena toxicity from aromatic compounds. SAR QSAR Environ Res 14:59–81
98. Schultz TW, Tucker VA (2003) Structure-toxicity relationships for the effects of N- and N, N'-alkyl thioureas to Tetrahymena pyriformis. Bull Environ Contam Toxicol 70:1251–1258
99. Dimitrov S, Koleva Y, Schultz TW, Walker JD, Mekenyan O (2004) Interspecies quantitative structure-activity relationship model for aldehydes: aquatic toxicity. Environ Toxicol Chem 23:463–470
100. Schultz TW, Netzeva TI (2004) Development and evaluation of QSARs for ecotoxic endpoints: the benzene response-surface model for Tetrahymena toxicity. In: Livingstone DJ, Cronin MTD (eds) Predicting chemical toxicity and fate. CRC Press, Boca Raton, FL, pp 265–284
101. Schultz TW, Seward-Nagel J, Foster KA, Tucker VA (2004) Population growth impairment of aliphatic alcohols to Tetrahymena. Environ Toxicol 19:1–10
102. Schultz TW, Yarbrough JW (2004) Trends in structure-toxicity relationships for carbonyl-containing α, β-unsaturated compounds. SAR QSAR Environ Res 15:139–146
103. Aptula AO, Jeliazkova NG, Schultz TW, Cronin MTD (2005) The better predictive model: high q2 for the training set or low root mean square error of prediction for the test set? QSAR Comb Sci 24:385–396
104. Aptula AO, Roberts DW, Cronin MTD, Schultz TW (2005) Chemistry-toxicity relationships for the effects of di- and trihydroxybenzenes to Tetrahymena pyriformis. Chem Res Toxicol 18:844–854
105. Gagliardi SR, Schultz TW (2005) Regression comparisons of aquatic toxicity of benzene derivatives: tetrahymena pyriformis and Rana japonica. Bull Environ Contam Toxicol 74: 256–262
106. Netzeva TI, Schultz TW (2005) QSARs for the aquatic toxicity of aromatic aldehydes from Tetrahymena data. Chemosphere 61:1632–1643
107. Schultz TW, Netzeva TI, Roberts DW, Cronin MTD (2005) Structure-toxicity relationships for the effects to Tetrahymena pyriformis of aliphatic, carbonyl-containing, α, β-unsaturated chemicals. Chem Res Toxicol 18:330–341
108. Schultz TW, Yarbrough JW, Woldemeskel M (2005) Toxicity to Tetrahymena and abiotic thiol reactivity of aromatic isothiocyanates. Cell Biol Toxicol 21:181–189
109. Schultz TW, Yarbrough JW, Koss SK (2006) Identification of reactive toxicants: structure-activity relationships for amides. Cell Biol Toxicol 22:339–349
110. Schultz TW, Hewitt M, Netzeva TI, Cronin MTD (2007) Assessing applicability domains of toxicological QSARs: definition, confidence in predicted values, and the role of mechanisms of action. QSAR Comb Sci 26:238–254
111. Schultz TW, Yarbrough JW, Pilkington TB (2007) Aquatic toxicity and abiotic thiol reactivity of aliphatic isothiocyanates: effects of alkyl-size and –shape. Environ Toxicol Pharmacol 23:10–17
112. Schultz TW, Ralston KE, Roberts DW, Veith GD, Aptula AO (2007) Structure-activity relationships for abiotic thiol reactivity and aquatic toxicity of halo-substituted carbonyl compounds. SAR QSAR Environ Res 18:21–29
113. Yarbrough JW, Schultz TW (2007) Abiotic sulfhydryl reactivity: a predictor of aquatic toxicity for carbonyl-containing α, β-unsaturated compounds. Chem Res Toxicol 20:558–562
114. Ellison CM, Cronin MTD, Madden JC, Schultz TW (2008) Definition of the structural domain of the baseline non-polar narcosis model for Tetrahymena pyriformis. SAR QSAR Environ Res 19:751–783
115. Böhme A, Thaens D, Schramm F, Paschke A, Schüürmann G (2010) Thiol reactivity and its impact on the ciliate toxicity of α, β-unsaturated aldehydes, ketones, and esters. Chem Res Toxicol 23:1905–1912
116. Roberts DW, Schultz TW, Wolf EM, Aptula AO (2010) Experimental reactivity parameters for toxicity modeling: application to the acute aquatic toxicity of SN2 electrophiles to Tetrahymena pyriformis. Chem Res Toxicol 23:228–234
117. Schultz TW, Sparfkin CL, Aptula AO (2010) Reactivity-based toxicity modelling of five-membered heterocyclic compounds: application to Tetrahymena pyriformis. SAR QSAR Environ Res 7:681–691
118. Bajot F, Cronin MTD, Roberts DW, Schultz TW (2011) Reactivity and aquatic toxicity of aromatic compounds transformable to quinone-type Michael acceptors. SAR QSAR Environ Res 22:51–65
119. QsarDB collection of TETRATOX primary publications. http://hdl.handle.net/10967/7. Accessed 15 Dec 2012
120. LOGKOW™, A databank of evaluated octanol-water partition coefficients, Sangster Research Laboratories, Montréal, QC, Canada. http://logkow.cisti.nrc.ca/logkow/. Accessed 15 Dec 2012

121. ClogP, BioByte Corp. Claremont (CA), USA. http://www.biobyte.com/bb/prod/clogp40.html. Accessed 15 Dec 2012
122. Estimation Program Interface (EPI) Suite, U.S. Environmental Protection Agency, Washington (DC), USA. http://www.epa.gov/oppt/exposure/pubs/episuite.htm. Accessed 15 Dec 2012
123. Nonylphenol, Wikipedia, The free encyclopedia. http://en.wikipedia.org/wiki/Nonylphenol. Accessed 15 Dec 2012
124. Personal communication with Prof. Schultz TW, College of Veterinary Medicine, The University of Tennessee, 2407 River Drive, Knoxville, TN 37996 July 2012
125. Schultz TW, Cronin MTD, Walker JD, Aptula AO (2003) Quantitative structure-activity relationships (QSARs) in toxicology: a historical perspective. J Mol Struc-THEOCHEM 622:1–22
126. Intelligent Modelling Algorithms for the General Evaluation of TOXicities (IMAGETOX), EU 5-th FP, # HPRN-CT-1999-00015, duration 2001–2004, participating institutions: Mario Negri Institute for Pharmacological Research (Milan, Italy), Liverpool John Moores University (UK), Umweltforschungszentrum Leipzig-Halle GmbH (Germany), Polytechnic of Milan (Italy), National Institute of Chemistry (Ljubljana, Slovenia), Utrecht University (Netherlands), University of Tartu (Estonia)