Chemosphere 364 (2024) 143078

Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

ECOTOXr: An R package for reproducible and transparent retrieval of data from EPA's ECOTOX database

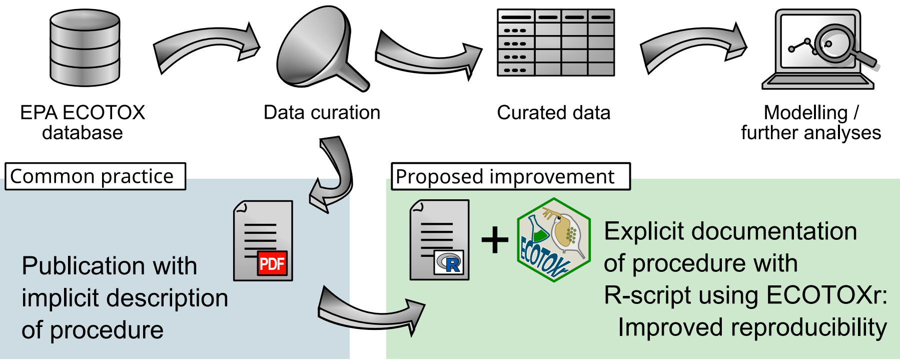
Pepijn de Vries

Dept. Wageningen Marine Research, Wageningen University and Research Center, P.O. Box 57, Den Helder, 1780AB, the Netherlands

HIGHLIGHTS

- Curating ecotoxicological data for analyses is a meticulous and time-consuming task.
- However, its documentation is often descriptive and not always explicit.
- ECOTOXr formalises this process by programming and documenting it as an R script.
- This makes it a tool for reproducible and transparent data curation.

GRAPHICAL ABSTRACT


EPA ECOTOX database
Data curation
Curated data
Modelling / further analyses
Common practice
Publication with implicit description of procedure
PDF
Proposed improvement
R
+
ECOTOX
Explicit documentation of procedure with R-script using ECOTOXr: Improved reproducibility

ARTICLE INFO

Handling editor: Michael Bank

Keywords:
Toxicology database
Ecotoxicology
Data curation
Interoperability
Chemical assessment

ABSTRACT

The US EPA ECOTOX database provides key ecotoxicological data that are crucial in environmental risk assessment. It can be used for computational predictions of toxicity or indications of hazard in a wide range of situations. There is no standardised or formalised method for extracting and subsetting data from the database for these purposes. Consequently, results in such meta-analyses are difficult to reproduce. The present study introduces the software package ECOTOXr, which provides the means to formalise data retrieval from the ECOTOX database in the R scripting language. Three cases are presented to evaluate the performance of the package in relation to earlier data extractions and searches on the website. These cases demonstrate that the package can reproduce data sets relatively well. Furthermore, they illustrate how future studies can further improve traceability and reproducibility by applying the package and adhering to some simple guidelines. This contributes to the FAIR principles, credibility and acceptance of research that uses data from the ECOTOX database.

E-mail address: pepijn.devries@wur.nl.

https://doi.org/10.1016/j.chemosphere.2024.143078
Received 18 March 2024; Received in revised form 24 July 2024; Accepted 9 August 2024
Available online 23 August 2024

1. Introduction

1.1. The ECOTOX database

When analysing or utilising large amounts of existing ecotoxicity data, relational databases play a crucial role. At the present there are multiple alternatives to consider, such as the alternatives listed in Table 1. Note that this overview is not exhaustive as the present study does not intend to review these databases. Instead, the present study focuses on The US Environmental Protection Agency's (US EPA) ECOTOXicology knowledgebase (referred to as ECOTOX database for brevity in the remainder of this text). This database was selected for its wide scope, large coverage and contains detailed information on experimental conditions (Table 1).

The ECOTOX database is a relational database that includes over a million ecotoxicological test results (Table 1), which are invaluable in environmental risk assessment. The database was originally developed in the 1980s to provide regulatory offices access to toxicological information for environmental hazard assessment (Olker et al., 2022).

Currently, the database incorporates the results of (both aquatic and terrestrial) ecotoxicological experiments published in both 'grey' and peer reviewed literature. These data facilitate derivation of safety thresholds (e.g., Eriksson et al. (2007)) and predictions of toxicity of novel substances (e.g., Sheffield and Judson (2019)); untested species (e.g., Dyer et al. (2006)); or mixtures (e.g., De Vries et al. (2022)). This list is not all-embracing, and many other applications are possible (Olker et al., 2022).

The ECOTOX processes match guidelines from Moher et al. (2009) for identifying and curating Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). These steps are formalised in standard operation procedures (SOPs) and used to continuously collect new relevant toxicity data and update the database with this information (Olker et al., 2022).

1.2. Application of ecotoxicological data in risk assessment

When collecting ecotoxicological data, selection criteria usually apply. Especially when they are used to derive safety thresholds such as environmental quality standards, or when they are used to draw conclusions on hazard or environmental risk in general. The data quality is assessed and taken into account in the evaluation for that purpose. Data quality is defined by the reliability and relevance of the tests (Rudén et al. (2017); Peters et al. (2023); Moermond et al. (2016); De Vries (2018)). Data quality of a test is usually evaluated with specific criteria. Some examples of such criteria are: are exposure concentrations measured?; are appropriate controls performed?; are test organisms well described?; etc. (Moermond et al., 2016). It is common practice to either filter out low quality or inadequate data.

The data quality in the ECOTOX database is not evaluated other the criteria specified by the EPA in their SOPs (Olker et al., 2022) (See Table 1). As such, the software presented in the present study does not facilitate this directly.

Table 1
A non-exhaustive overview of databases containing ecotoxicity test results and their properties.

Property	ECOTOXicology Knowledgebase	NORMAN Ecotoxicology Database	OpenFoodTox	ECETOC EAT	EnviroTox
Provider	US EPA[a]	NORMAN[b]	EFSA[c]	ECETOC[d]	HESI[e]
Number of records	1,079,524[f]	81,430	5,934[g]	6,095	80,912
Database quality	PRISMA and SOPs[h]	TS of CEN[i]	Automatic and manual control checks (Dorne et al., 2021)	No longer maintained[j]	Under the auspices of HESI & Environmental Risk Assessment Committee
Test data quality	Responsibility of end user	CRED[k]	Applied guidelines & GLP[l] compliance	Criteria listed in Solbé et al. (1998)	SIFT criteria[m]
Recorded test conditions	detailed	limited	limited	limited	limited
Species covered	Ecologically relevant species of *Protozoa, Chromista, Plantae, Fungi, Animalia*[n]	*Bacteria, Protozoa, Chromista, Plantae, Animalia*	*Plantae, Animalia*	Aquatic species of *Bacteria, Protozoa, Chromista, Plantae, Animalia*[o]	*Eubacteria, Monera, Protoctista, Chromista, Plantae, Animalia*
Full database downloadable	yes	no	yes	yes	yes
Offers web search	yes	yes	yes	no	yes

[a] https://cfpub.epa.gov/ecotox/explore.cfm?sub=Effects, accessed 2024-01-12.
[b] Network of reference laboratories, research centres and related organisations for monitoring of emerging environmental substances (NORMAN, 2024; Dulio et al., 2018); NORMAN contains data curated from the EPA ECOTOX database (Olker et al., 2022).
[c] European Food Safety Authority (EFSA, 2024; Carnesecchi et al., 2023; Dorne et al., 2021).
[d] European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) Aquatic Toxicity (EAT) (ECETOC, 2024; Solbé et al., 1998).
[e] Health and Environmental Sciences Institute (Connors et al., 2019). The EnviroTox database contains data curated from the EPA ECOTOX database.
[f] Excluding accumulation data.
[g] Excluding human health and animal studies.
[h] See main text for details.
[i] Technical Specification (TS) of the European Committee for Standardization (CEN) (Schwesig et al., 2011).
[j] Last release 2003 (ECETOC, 2024).
[k] Criteria for Reporting and Evaluating ecotoxicity Data (CRED, Moermond et al. (2016)) (Only available for test results used in environmental quality targets).
[l] Organization for Economic Cooperation and Development Principles of Good Laboratory Practice.
[m] Stepwise Information-Filtering Tool methodology by Beasley et al. (2015) with criteria specified by Connors et al. (2019).
[n] Except human, monkey, bacteria, viral and yeast.
[o] *Daphnids* and other invertebrates plus fish and other vertebrates.

1.3. Motivation and problem definition

There is an increasing call for open science and data, with numerous initiatives to improve the embeddedness of FAIR principles of scientific data (Findability, Accessibility, Interoperability, and Reusability), which includes environmental sciences (Wilkinson et al., 2016; Brock et al., 2021). These initiatives tend to focus on individual test studies and on how their individual information is stored and shared. The databases listed in Table 1 all strive to adhere to the FAIR principles. However, the same principles should apply to studies that curate and (re)use existing test data from these databases.

The process of collecting or creating, verifying, harmonising, organising and maintaining data is what is commonly referred to as data curation (Connors et al., 2019). Although the software focuses mostly on data collection, it can facilitate the entire curation process, by formalising and documenting it in an R script suggested in section 4.4.

Together with the FAIR principles, reproducibility is an important quality of scientific research for the acceptance of the results (Gentleman and Lang, 2007; Mebane et al. (2019); Wang (2018)). Despite the deemed importance of reproducibility, it has never been quantified how (ir)reproducibility affects conclusions of risk assessment studies. As with the FAIR principles, such studies are inclined to focus on reproducibility of individual experimental studies, not on the process of collecting and evaluation of data.

Many of the aforementioned studies that use the ECOTOX database report either the selected, raised or processed data. Extraction and subsetting from the database prior to the reported data is a meticulous and time-consuming process. However, it is generally given little attention in publications, and is especially important because the database itself is also continuously evolving. Consequently, attempts to reproduce such research are difficult if not impossible. In addition, there is no standardised method for extracting data from the database.

The ECOTOX database is publicly available at https://cfpub.epa.gov/ecotox/, where the user is offered different ways to access the information. The first option is to use a web-based form to search, explore and download specific subsets of the database. Although helpful in many cases, there are limitations: the user can only search in predefined fields; and output is not generated for all available fields in the database (such as the creation and modification date of records). Furthermore, a quota limits the number of records that can be downloaded via this route. Moreover, there is currently no formal Application Programming Interface (API) available which allows convenient access to the online database in a scripted form.

Although an online search of the database is straightforward, it is by definition irreproducible, as users can't control which content is (or is not) available from the online resource. Furthermore, documenting this search process as descriptive text will allow for (mis)interpretation by others. In light of this, it is desirable to document and execute the process of data collection and handling in a scripting language. This is particularly true when the researches need go beyond a quick screening of data.

To promote reproducibility, the present study opted for the second option for accessing data from the database, namely, which is where the user can download the full raw tables from the database archived in a zip file. This provides users with the flexibility that the web based interface does not. As the EPA does not provide any tools or means to build and access a local copy of the database from the raw tables, the present study introduces new software designed to achieve this.

1.4. Implementation and design details

The R language for statistical computing (R Core Team, 2021) is selected as the environment to develop the ECOTOXr package software. This environment is selected as it is free, open source, platform independent and evolved with a focus on data science (Ihaka and Gentleman, 1996; Chambers, 2020; Crawley, 2012). Moreover, it has an active community (Boettiger et al., 2015) and a large number of packages that support a wide spectrum of statistical analyses and modelling techniques.

Before addressing the implementation, this study dissects how desired transparency, reproducibility and user-friendliness is asserted. For this purpose the principles laid out by Poisot (2015) and Wilson et al. (2014) are applied, which were also followed for a similar package (webchem, Szöcs et al. (2020)).

Transparency is partly achieved by publishing on R's central software repository: The Comprehensive R Archive Network (CRAN: https://cran.r-project.org/). This network guarantees official releases are catalogued; the package complies with its strict (quality) policies and is therefore fully documented; the package and its environment is open source; and the software is freely and easily accessible to R users. In addition, the quality of the software is assured by customised, built-in self-tests (Wickham, 2011). These tests are designed to assess whether the software behaves as intended and expected; also after updates. These tests are automatically run with each new release.

Moreover, the source is maintained at GitHub which makes keeping track of source code changes, and issues with the software public and visible. The source code is released under a GNU General Purpose License (https://www.gnu.org/licenses/gpl-3.0.en.html), allowing others to redistribute and modify the code under the same terms. All of this facilitates collaborations. The peer review of the present study contributes to the credibility and transparency of the software.

Versioning of the software for the sake of reproducibility is also important. This is facilitated by both CRAN and GitHub. In addition, several self-contained tests are included within the package, assuring consistent performance when the software is updated (Wickham, 2011). Not only is the data extraction software important in reproducibility, but also the data source. EPA releases a new copy of the database roughly once every four months. The ECOTOXr package stores built and download information as separate plain text files. The end user can easily request this information using the functions get_ecotox_info() and cite_ecotox(). By including this information in reports, the end user ensures that others can recreate the same database and queries, by using the same data and software release.

User-friendliness requires the software to be easily installed and used. As an R package the software should be useable by anyone with only basic knowledge of R syntax. However, it should also allow for users to exploit more advanced knowledge of databases and queries when available.

Part of this is achieved by choosing function names that resemble what they do. Generally, they consist of a verb and a noun (separated by an underscore), for instance search_ecotox is a function that allows searches in the ECOTOX database. There are several exceptions made to the naming convention, most notably: functions that overwrite existing methods such as 'is', 'show', etc.; the function names dbConnectEcotox and dbDisconnectEcotox are chosen to reflect their relationship with their counterparts from the RSQLite package (see also section 1.5). By publishing the package on CRAN, the installation procedure should be appropriate for a novice user. By providing a simple built-in search function and allowing custom (SQL) queries, both laymen and expert user needs can be served. However, to efficiently use the package and take full advantage of all features offered by R and its packages, it is advisable for users to get well acquainted with R (Wickham et al., 2023).

Given the self-imposed constraint of reproducibility, it makes some of the design choices in development easier. With no API to the web-based interface to the database, creating a custom interface is clearly not a stable option. If the web-based interface (which are out of our control) changes, it can easily break software that uses it. In addition, as pointed out before, the online interface does not give access to all the available tables and fields in the database. The best choice is therefore to work with the raw data, which are also available. This unfortunately poses dilemmas of its own. Nonetheless, the ECOTOXr package does offer an experimental feature to perform online searches directly (websearch_ecotox()).

1.5. Dependencies

The EPA provides the database as separate tables with instructions as to how they are related to each other in a database. The database itself still needs to be built from these files. SQLite (Hipp, 2021) is used for that purpose as the structured query language (SQL) database engine. It is selected because it is free, stable, open source, lightweight and is supported in R by the package RSQLite (Müller et al., 2021). However, there is one feature missing from SQLite that is available in many other database engines. Namely, the management of permissions. This means that any user can modify the locally stored database in any way the user sees fit. Although this could be considered to advantage, it could also break reproducibility when the user is not careful. This is pointed out in the manual as well (De Vries, 2024) and should not form a problem when only the package built-in functions are used. In conclusion, the advantages outweigh the disadvantages, the RSQLite package is used in the design of ECOTOXr.

Other packages that are imported by ECOTOXr are ‘crayon’, ‘lifecycle’ (mostly used for user interface aesthetics) and ‘rappdirs’ (used for finding a suitable location for storing the SQLite database). In addition ‘httr2’, ‘jsonlite’, ‘readr’, ‘rvest’, and ‘stringr’ are used for importing and interpreting the raw data files provided by EPA, while building the database.

Tidyverse is a collection of packages that share a common design philosophy, grammar and data structures (Wickham et al., 2019). As ECOTOXr is building on from these packages, it depends on several of them (‘dbplyr’, ‘dplyr’, ‘purrr’, ‘rlang’, ‘tibble’, ‘tidyr’, and ‘tidyselect’). In the present context these packages are used to build a local copy of the database and construct comprehensive search queries.

1.6. Installation and system requirements

Although developmental versions of the package are available on GitHub, it is advisable to work with the latest official release on CRAN. It can easily be installed and loaded by running code snippet 1 in R.

```r
install.packages("ECOTOXr", repos = "CRAN", dependencies = TRUE)
library("ECOTOXr")
```

code snippet 1.

Before the package can be used to its full potential, a local copy of the database needs to be built from the files provided by the EPA. This only needs to be performed once (although this could be repeated when a new release of the database becomes available) and can be done automatically with code snippet 2.

```r
download_ecotox_data()
```

code snippet 2.

The user could also stipulate a specific target path for downloading and building the database with this function, otherwise a default path is used. The database itself requires roughly 1.2 GB of free disk space. However, while building, the zip archive and the raw tables also need to be stored. Hence, installation would require at least twice as much free disk space.

If for some reason the download via R fails, the remote database can be downloaded manually from https://cfpub.epa.gov/ecotox/. A local database can then be built from that download using a call to function below on the files extracted from the downloaded zip archive (see code snippet 3).

```r
build_ecotox_sqlite()
```

code snippet 3.

When working in R, data is normally loaded into RAM memory. An advantage of storing the data in a database is that R doesn’t have to load all data into the computer’s memory. Instead, a query can be sent to the database driver which only returns the requested data to R and is stored in memory. While building, data from the raw tables are transferred consecutively to the database with a maximum of 50.000 rows per iteration, in order to spare memory (as the complete database might not fit in the memory of all systems).

Installation, building the database and extracting information from it was successfully tested on both a Windows 10 and Linux machine (Windows 10 build 18363 x64 Intel(R) Core(TM) i5-8265U CPU @ 1.6 GHz 1.8 GHz 8 GB RAM; Linux 5.10.0-6-amd64 SMP Debian 5.10.28-1 (2021-04-09) x86_64 Intel(R) Xeon(R) Gold 6154 CPU @ 3 GHz 377 GB RAM). It is expected to run on lower end machines as well.

1.7. Usage

Once the database is built, it is fairly simple to retrieve specific records from the database. Consider R code snippet 4.

```r
search_ecotox(
  list(
    latin_name    = list(terms  = "Daphnia", method = "contains"),
    chemical_name = list(terms  = "benzene", method = "exact")
  )
)
```

code snippet 4.

A list is used to concatenate different search terms. The name of each element in the list corresponds with the field name in the database that is being searched. In this case records for species for which the Latin name contains *Daphnia* and are tested with the chemical benzene are retrieved. The user doesn't need to have extensive prior knowledge on the database structure in order to use the R function search_ecotox. Internally this function constructs an SQL statement, based on the provided arguments, for which the database is queried. Using the underlying Database Interface (DBI) functions (R Special Interest Group on Databases (R-SIG-DB) et al., 2022), the user is also allowed to send custom queries to the database. But in order to do so, knowledge of the database structure is required, which can be obtained from the ECOTOX manual (Olker, 2022) and accompanying data dictionary as included with the ASCII download of the EPA database (https://www.epa.gov/ecotox/). An advantage of the latter approach is that the tailored queries might perform better in terms of processing time.

1.8. Aim

There is not a clean and firmly reproducible method available for extracting data from the database. This is what triggered the development of the ECOTOXr package (De Vries, 2024) described in the present study. This package aims to provide the means to effectively extract any information from the ECOTOX database; and allowing the user to do so in a transparent, reproducible and user-friendly fashion.

Both as a quality check and to demonstrate the added value of using the presented package, three case studies are performed. The first two will attempt to recreate data extractions from earlier studies. The third will compare a search using the presented software with a search of the ECOTOX website. The development and design of the software are already presented in the introduction. Consequently, the Methods and Results section will focus on the three case studies. The discussion will start with lessons learned from the case studies and end with implications for the software and future applications and recommendations.

2. Methods

De Vries and Murk (2013) and Szöcs et al. (2020) were selected for the aforementioned two case studies. Both these studies indicate that they used data from the ECOTOX database, but without applying the software presented here. The ECOTOXr package should be able to reproduce the data used in these studies to some extent, making these suitable case studies to evaluate the software with respect to reproducibility. A list of endocrine disrupting substances (Akerman and Blankinship, 2015) was selected for the third case study to produce a large data set with little overlap with the first two case studies. The third case study was used to compare with online search results. Methods applied to each case study are described in separate sections below. R scripts and data to reproduce the results are provided as Supplemental Information (S1 and S2). In the present study, the ECOTOX release of June 13th, 2024 was used.

2.1. Case study 1

De Vries and Murk (2013) extracted 50% lethal concentrations (LC50) in specific concentration units (μg/L and tenfolds of this unit) from the ECOTOX database, in order to test compliance of the data with Benford's law (Benford, 1938; Newcomb, 1881). This law tests the distribution of first digits in a dataset against an expected (Benford) distribution of digits in a natural dataset as an indication of anomalies in the tested set. The authors of the original study also included No Observable Effect Concentrations (NOEC) in their analyses, but for simplicity the present study focuses on LC50 data only. In order to reproduce this data the search terms shown in code snippet 5 were defined.

```r
units <- c(
  apply(
    expand.grid(
      mass   = c("pg", "ng", "ug", "mg", "g", "cg", "dg", "kg"),
      volume =
        apply(expand.grid(
          prefix = c("", "10", "100", "10 ", "100 "),
          vol    = c("cc", "nL", "uL", "mL", "cL", "dL", "L",
                     "mm3", "cm3", "dm3", "m3")
        ), 1, paste, collapse = "")
    ), 1, paste, collapse = "/"
  ), "ppb", "ppt", "ppm")

search_terms <- list(
  endpoint      = list(terms = "LC50",        method = "contains"),
  conc1_mean_op = list(terms = c("", "None"), method = "exact"),
  conc1_unit    = list(terms = units,         method = "contains")
)
```

code snippet 5.

The search was performed by providing the list of search terms as argument to a call to the search_ecotox function. Note that only alphanumerical characters were used as search terms. As other characters may not reproduce well on varying operating systems (De Vries, 2024).

```r
## calling LC50 data from the webchem package:
webchem::lc50
```

code snippet 6.

Only LC50 values without an operator (conc1_mean_op, i.e. less than, or greater than) were included, in order to conform with the study of De Vries and Murk (2013). The LC50 unit (conc1_unit) search terms were based on the units reported in the data set from that study (De Vries and Murk, 2013).

In addition to the default database output fields (list_ecotox_fields()), unique database record identifiers plus date creation and modified fields were added to the output. The creation date was used to subset the extracted data. Only records that were created on or before March 15th, 2012 were included in the present study. This is because De Vries and Murk (2013) used a copy of the ECOTOX database with that release date. Records that have been modified since that release date were labelled.

The data originally used by De Vries and Murk (2013) in their analyses was downloaded from the paper's supplemental information for comparison. As mentioned before, this data set was subset to LC50 records in the present study. In their paper only data required for their analyses were reported. This makes it hard to link individual records from their study to those of the present study. The only identifier that can be used is a combination of LC50 magnitude, its unit and the year of publication. Unfortunately, this does not result in unique identifiers as some LC50 values may occur multiple times in the same publication year. In any case, these identifiers were used to match records from both studies to each other.

Using all this information, the number of records were counted for each publication year in both data sets. The number of matched (and non-matched) records were also counted for both data sets. The number of records that were modified since publication by De Vries and Murk (2013) were also counted for the data extracted with ECOTOXr.

```r
ecotox_lc50  <- search_ecotox(
  list(
    endpoint   = list(terms = c("EC50", "LC50"), method = "contains"),
    latin_name = list(terms = "Daphnia magna",   method = "exact"),
    effect     = list(terms = c("ITX", "MOR"),   method = "contains")
  ),
  c(list_ecotox_fields(), "results.obs_duration_mean", "results.obs_duration_unit",
    "results.result_id"))
```

code snippet 7.

2.2. Case study 2

The 'webchem' package comes with aggregated data collected from the ECOTOX database, namely acute (48 h) LC50 data for the species *Daphnia magna* exposed to insecticides (Szöcs et al., 2020) (see code snippet 6). This data set contains two column names: cas and value. The latter contains the value of the exposure concentration presumably in µg/L.

The second case attempted to recreate this data set and compare it to the data set within 'webchem'. This was done with the intention to study how accurately a data set could be reproduced given the (lack of) available documentation. This will also help to understand what kind of documentation would be required to reproduce a data set.

Unfortunately, the 'webchem' package does not document the procedure followed to curate the data in detail. Therefore, a similar and sensible procedure was devised in the present study.

A definition of what is considered to be an insecticide was required for this purpose. The present study assumed that this entails all substances that were presently listed as insecticide by Chemical Abstracts Service (CAS) registry number in the British Crop Production Council's (BCPC) pesticide compendium as available from http://www.bcpcpesticidecompendium.org/(accessed on 2021-11-22). A list was composed using the compendium and supplemented with molar weight required for concentration conversions. Molar weights were obtained from the Chemical Translation Service (https://cts.fiehnlab.ucdavis.edu/, accessed on 2021-11-22), this is provided as Supplemental Information (S2) to this paper. Note that the CAS is known to form ambiguous chemical identifiers in some cases. In the present case study it was the only point of entry and therefore had to be used.

Next a search was performed using the ECOTOXr package to produce an initial similar data set. This was done by performing a wide search in the local database for 50% mortality effect concentrations for *Daphnia magna*, where the search results needed to be condensed to match with the selection used in 'webchem' (see code snippet 7).

Tests were selected from this data with observation duration equalling 48 h. Another selection criterion was only tests with reported effect concentrations were selected. Furthermore, these concentrations needed to be reported in the test medium and not in the organism. All concentrations were then converted to µg/L. All data processing (selection steps and conversion) are performed in R (see Supplemental Information (S1 and S3)).

It was difficult to assert which selection criteria were applied by the authors of the 'webchem' package from this point forward. Therefore, the comparison was based on the current selection described here. As an

Table 2
Number of LC50 records as extracted by De Vries and Murk (2013), and those collected in the present study with the ECOTOXr package. The number of records is also counted for the cases where the (non-unique) identifier introduced in the present study matches across the dataset from De Vries and Murk (2013) and collected here, which is shown per row. Percentages between brackets express the amount of records relative to the total number in the bottom row (which is by definition 100%). For data collected in the present study it is also determined whether or not the record was modified in the ECOTOX database since the publication by De Vries and Murk (2013).

Number of records	De Vries and Murk (2013)		present study (not modified since 2013)		present study (modified since 2013)		present study (All)	
Not matched against other data set	7,231	(8%)	6,734	(9%)	1,229	(12%)	7,963	(9%)
Matched against other data set	81,551	(92%)	69,806	(91%)	8,936	(88%)	78,742	(91%)
All	88,782	(100%)	76,540	(100%)	10,165	(100%)	86,705	(100%)

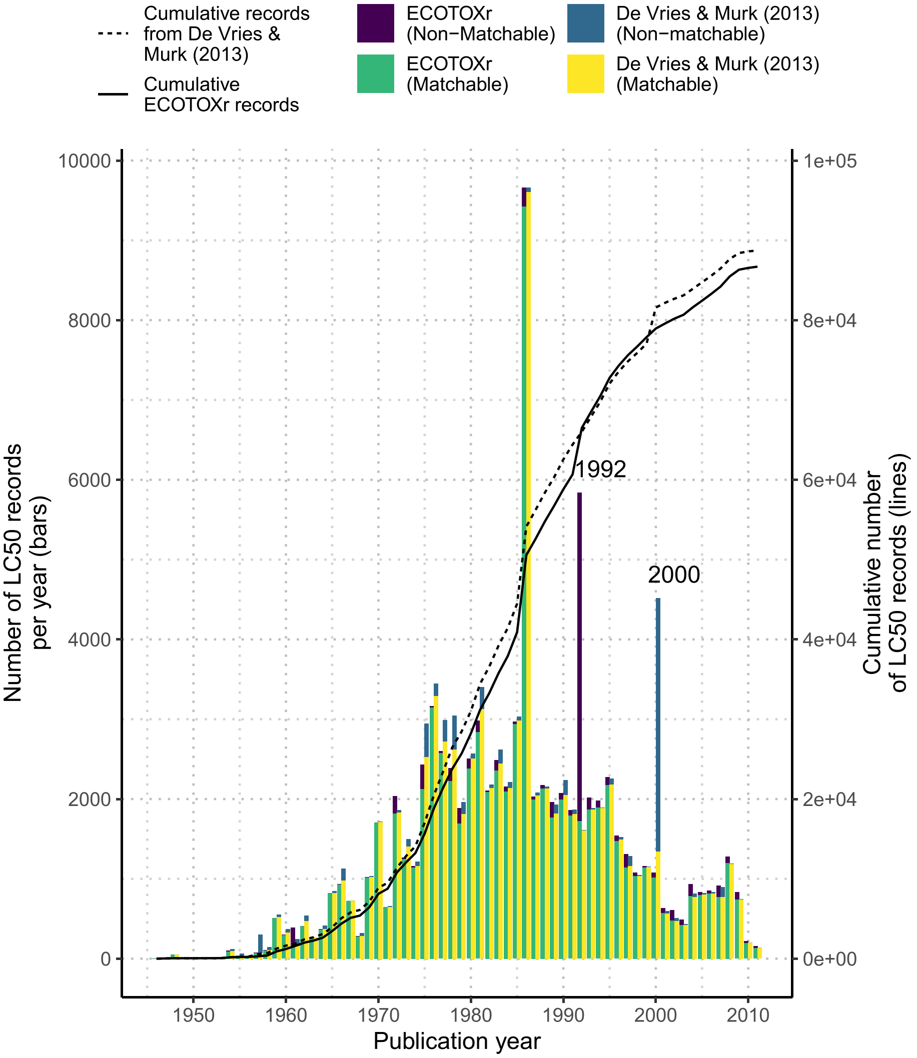


Fig. 1. Histogram of LC50 records per publication year (bars). In the histogram records from both studies (present and from De Vries and Murk (2013)) are shown side by side for each year. Whether records could be matched to each other is indicated by differently coloured stacked bars. Cumulative total number of records for both studies are shown as lines.

Table 3
Number of insecticides listed in 'webchem' and those extracted using ECOTOXr.

	Insecticides not retrieved with ECOTOXr (local and websearch)	Insecticides retrieved with ECOTOXr (local and websearch)
Insecticides included in 'webchem'	3	121
Insecticides not included in 'webchem'	0	70

additional internal check, the same search is also repeated with online (websearch_ecotox) using the same list of substances and conditions.

The 'webchem' package states that the effect concentrations were aggregated per substance, but not how. In the present study the geometric mean was taken of the records retrieved with ECOTOXr, the number of records were counted and the minimum and maximum values were determined. These values were plotted against those obtained from 'webchem'.

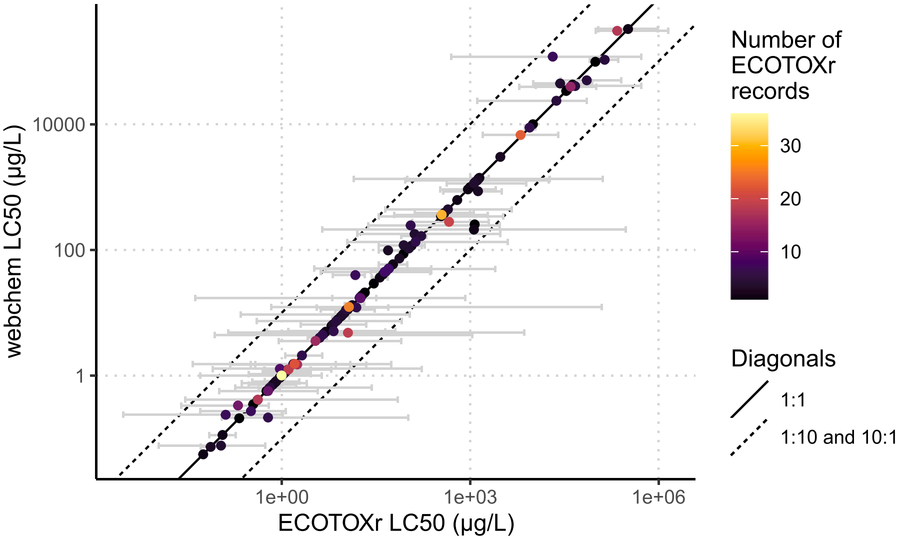


Fig. 2. A comparison of acute *Daphnia magna* (*D. magna*) insecticide toxicity (48 h LC50) between data reported by 'webchem' and extracted with ECOTOXr, both from the ECOTOX database. Each marker represents the geometric mean *D. magna* LC50 for a specific CAS registration entry. The grey scale of the markers indicate the number of resulting ECOTOX records ('webchem' only reports a single aggregated value for each CAS entry).

2.3. Case study 3

As a third case, data was collected on 52 chemicals selected for the tier 1 screening in the EPA's Endocrine Disruptor Screening Program (EDSP) (Borgert et al., 2011; Akerman and Blankinship (2015)). The search was performed using the CAS numbers listed by Akerman and Blankinship (2015). Data was collected using the R package presented in the present study. In addition, the same set of chemicals is explored via the ECOTOX website ({https://cfpub.epa.gov/ecotox/explore.cfm?sub=Chemicals, accessed July 3rd, 2024). Using the form on the website, a custom chemical group was created by entering the list of CAS numbers. The number of resulting records was noted. Downloading the complete set of results was not possible due to the limit of 10,000 records of the website.

Taxonomical information (class and subphylum) of the test species was collected. Additional test information was collected, the first being whether specific controls were applied, matching the following codes in the database: vehicle (V); concurrent (C); satisfactory (S) (legacy term); multiple types (M); multiple controls entered (ME/); and baseline (B). The second being whether exposure concentrations were measured, matching the following codes in the database: measured (M or M/); some measured values (X); chemical analysis reported (Z); and multiple methods (MULT). These aspects were selected as they are often used in the evaluation of test data quality (see section 1.2). Furthermore, it was noted whether records were created, modified or published after the year 2014. This would give an indication of the amount of new or modified data since the original presentation of the tier 1 screening (Akerman and Blankinship, 2015). The latter information could not be obtained via the website.

3. Results

Results presented here can be reproduced with the R script provided as Supplemental Information (S1 and S3). Identifiers of the records extracted from the ECOTOX database with this script are also provided as Supplemental Information (S4). The specific results for all three case studies are presented below.

3.1. Case study 1

The number of records retrieved by De Vries (De Vries and Murk, 2013) and those retrieved in the present study are listed in Table 2 and shown per publication year in Fig. 1. The retrieved data sets differ by 2,077 records in size. This sounds like a lot, but on a total of 88,782 records, it is only 2.3% (Table 2). Still, 9% of the records from De Vries and Murk (2013) and 8% from the present study, could not be matched to each other (using the custom non-unique identifier introduced above; Table 2). The observed mismatch could very well be explained given that over 12% of all records retrieved in the present study have been modified since publication by De Vries and Murk (2013). However, it is probably not the only explanation.

Looking at the cumulative records count in Fig. 1 (solid and dashed lines), there seems to be a fluctuating offset, suggesting a mismatch between the recorded publication dates obtained in both studies. This is

Table 4
Summary of records collected for the 52 tier 1 chemicals in the EDSP. Number of records (effect concentrations), tests and publications are shown per class of *Vertebrata* together with the range of publication years. Between brackets is the percentage of records that was added or modified after 2014 in the database. For the publications this percentage shows the percentage of documents published after 2014. Percentages of tests and records that include controls and measured concentrations are also reported.

Species class	*N* Records (>2014, Measured, Control)	*N*Tests (>2014, Measured, Control)	*N* Publications (>2014)	Publication years
Actinopterygii	28,015 (36%, 21%, 86%)	14,212 (13%, 20%, 77%)	2,185 (10%)	1917–2022
Amphibia	5,268 (42%, 30%, 97%)	2,724 (27%, 29%, 94%)	416 (14%)	1960–2021
Aves	4,340 (25%, 7%, 79%)	2,392 (21%, 2%, 82%)	341 (7%)	1957–2021
Cephalaspidomorphi	11 (0%, 0%, 100%)	10 (0%, 0%, 100%)	2 (0%)	1957–2010
Chondrichthyes	2 (0%, 0%, 0%)	2 (100%, 0%, 0%)	1 (0%)	1976–1976
Mammalia	14,562 (21%, 3%, 98%)	9,449 (18%, 3%, 98%)	1,446 (1%)	1955–2020
Reptilia	444 (66%, 2%, 98%)	121 (49%, 4%, 96%)	40 (42%)	1975–2022
Unspecified	34 (9%, 18%, 65%)	34 (9%, 18%, 65%)	18 (0%)	1966–2012

perhaps most apparent for the bars of the publication years 1992 and 2000 (Fig. 1).

3.2. Case study 2

Table 3 lists the number of substances for which data is retrieved in the present study versus those included in the 'webchem' package. There are three substances listed under the 'webchem' package that were not retrieved with the ECOTOXr package (Table 3). This is probably because 'webchem' used a different, yet undisclosed definition, of insecticides. Furthermore, data selection criteria were also unavailable. On the other hand, the ECOTOXr package retrieved data for 70 substances that were not listed in 'webchem', again for the same reason.

There are 121 substances present in both data sets. The effect concentrations for those substances are shown in Fig. 2. In general the effect concentrations obtained from 'webchem' and those derived from data collected in the present study correspond well. These values differ much less than an order of magnitude (Fig. 2).

The data sets correspond well with each other and the difference is less than an order of magnitude in general. The internal check with the online search yields the exact same substances as the search in the local database. Moreover, the online search yields three additional records. These are all records describing effect concentrations in μg/org, which are not included in the local search. The local search extracted one extra record, that was not retrieved with the online search. Possibly, it was not returned as the unit (μl/l) may not have been converted to a standardised unit on the server of the EPA, as indicated on the ECOTOX website (https://cfpub.epa.gov/ecotox/help.cfm), section 'ECOTOX Unit Conversion Logic'). Unfortunately, this cannot be corroborated. If these four records are omitted, the resulting effect concentrations have exactly a 1:1 relationship between the local and online search results. More detailed results for the online versus local search comparison is presented in the Supplemental Information (S4).

3.3. Case study 3

Searching for the 52 tier 1 EDSP chemicals yielded 177,634 records. This number is identical for both the local and the online search. Also the number of records per substance (i.e., CAS number) for both searches are identical. Table 4 shows a summary of the number records and tests collected for the chemicals tested on *Vertebrata*, Supplemental Information (S4) lists all retrieved records and tests. Table 4 shows that most tests are performed with *Actinopterygii* (ray-finned fish) and *Mammalia*. The percentage of records and tests that were added to or modified in the database since 2014 is variable across classes. In general, the percentage of modified records is larger than the percentage of modified tests. Both are also greater than the percentage of references published after 2014. The percentage of tests and records that include controls is relatively high (>65% in most cases) but variable across classes. Tests and records with measured concentrations generally occur less frequently ($\leq$30%).

4. Discussion

The present study demonstrates a new software package for transparent and reproducible data extractions from the ECOTOX database. The package was tested in three case studies where data extractions from earlier studies were recreated with the software presented here. Resulting data sets within each case study were compared to each other.

4.1. Specific lessons learned from the case studies

4.1.1. Case study 1

When the data that is extracted in the present study is compared to that collected by De Vries and Murk (2013), a large discrepancy is found for data published in the years 1992 and 2000 (case study 1, Fig. 1). There is a large mismatch for both years, similar in size, but in counterpart studies. There are several possible explanations: the publication year of one or more references is or were mislabelled in the ECOTOX database, in the study by De Vries and Murk (2013) or in both. Alternatively, other errors could be present in either the database or the study by De Vries and Murk (2013); or a combination of these aspects.

One way of explaining this, is by determining whether a particular publication underpinning the data is responsible for a large part of the inconsistency between the two data sets. The literature reference from the ECOTOX database with the largest number of mismatches between both sets is published by the US EPA (EPA, 1992). This publication is associated with 4,043 records that do not align. This accounts for the large difference observed in the years 1992 and 2000 (Fig. 1).

Hypothetically, when De Vries and Murk (2013) authored their publication, this literature reference was registered in the database under the publication year 2000, and was later corrected to 1992. Such corrections may have taken place for other publications as well, accounting for other discrepancies.

Another explanation is formed by duplicate records, which have been removed from the database at some point and may also contribute to inconsistencies. Looking at the only identifier of the records currently available (a combination of publication year, effect concentration and concentration unit), the potential for duplicate records is definitely there. Records matching between the two data sets contain duplicated identifiers. Records collected with ECOTOXr has 41% duplicated identifiers and the data set from De Vries and Murk (2013) has 43% duplicated identifiers. Note that these percentages only show the duplicates of non-unique identifiers used in the present study. The percentage of actually duplicated records is expected to be much lower. However, De Vries and Murk (2013) already revealed that the database contains duplicated records. This could partially explain the differences observed here.

The crux is that one can only truly tell what causes discrepancies when records can be unambiguously linked together based on unique identifiers. This could be achieved by storing the record identifier from the ECOTOX database, preferably in combination with the registered modification date (if available), which is advocated in the present study.

4.1.2. Case study 2

The data extracted in the present study corresponds well with data from the 'webchem' package (case study 2, Fig. 2). However, there are differences as not all values align with the diagonal (1:1). This is not surprising given the variability in the data, indicated by the error bars in Fig. 2. All minimum and maximum ranges intersect with the 1:1 line. Attempts to explain differences observed here by comparing them with a variation in test conditions and experimental set up were not successful. The 'webchem' package could benefit from the ECOTOXr package to formalise, document and implement the data processing to obtain the aggregated LC50 values. This would make their data curation process more transparent and reproducible.

4.1.3. Case study 3

The third case study demonstrates how the R package is capable of retrieving the same records as the online search. The software allows the user to extract all data at once, whereas the website limits downloads to 10,000 records, requiring the user to subset the data. It also shows that the software is able to extract fields that cannot be retrieved from the website (namely dates on which records are created and modified). In addition, meta-information on the tests extracted from the database (measured concentrations and control types), can be extracted by the software which assisting research when evaluating data quality. Some suggestions on this matter is presented in section 4.4.

4.2. General considerations

The comparison of ECOTOXr with earlier studies (Szöcs et al., 2020; De Vries and Murk, 2013) demonstrates that there is generally a good correspondence between the data extracted from the ECOTOX database with ECOTOXr and by the respective authors, indicating that most of the data is reproduced. The comparison between the local and online search in case study 3 resulted in identical numbers of records. However, the comparison of the first two case studies show that there are differences that could not be explained with available information.

Three hiatuses are identified as a potential explanation: search and filtering procedures are not fully explicit; there is no way to unambiguously link records in the studies to specific records in the ECOTOX database; and the ECOTOX database has been modified over time (i.e., in the time between the original publication and the ECOTOX release used in the present study). ECOTOXr can easily tackle the first two issues. This can be achieved by documenting search and filtering process as reproducible R code and storing original identifiers from the database the retrieved data. These identifiers serve as tracers to the corresponding records in any later version of the database. This is approach is reliable because identifiers from the database are unique, robust and are unambiguously linked to a specific record (Olker et al., 2022).

What ECOTOXr cannot accommodate is tracking modifications of the database. This is because database management is outside the sphere of influence of this package. However, when record identifiers are documented well, users can compare earlier extractions with later iterations of the database in order to establish whether a record has been modified.

4.3. Alternatives and related software

There are alternative databases, some of which are listed in Table 1. There is no comprehensive review of (the extent of) their usage in literature. Such a review is beyond the scope of the present study, Olker et al. (2022) has listed several practical usage examples of the ECOTOX database. It is currently impossible to measure the relevance of each of the databases in the scientific arena. This is something that practice will have to demonstrate, or needs to be addressed in future research.

Another alternative that is not listed in Table 1 is International Uniform Chemical Information Database (IUCLID, see e.g. Evangelisti et al. (2023)). It is not included here because it has an inherently different role. It provides a uniform format for exchanging chemical information. This is particularly helpful for setting up and maintaining dossiers in management of chemicals. This may not always meet with research needs.

In addition to the databases there are also different software packages to access them. The present study focuses on R packages. There is an alternative R package, called 'webchem' (Szöcs et al., 2020), which uses ecotoxicological data from the EPA ECOTOX database, as shown in the case study above. In fact, this package allows retrieval of all sorts of chemical properties from a wide range of online resources. However, this package is limited only to summarised acute toxicity of insecticides to *Dapnia magna*. In contrast, ECOTOXr offers access to all chemicals and species in the database, but is limited to that database only. Obviously, the two packages can also be used to merge information. Naturally, the level of proficiency in the R language will affect the user's efficiency of application of the presented (and alternative) R packages.

The 'standartox' package was developed by Scharmüller et al. (2020) in parallel to the ECOTOXr package presented here. 'standartox' uses a very different strategy for disseminating the information from that used by ECOTOXr. This makes both methods complementary and will serve different (research) needs. The 'standardtox' package provides a selection of processed US EPA ECOTOX data via the online standartox database. The R package interacts online with it via an API. This database harmonises the reporting units, whereas the ECOTOXr package currently only reports units as is. This allows for full control over the data and for easier tracing to source material.

Recently, the 'REcoTox' package was posted to github (https://github.com/tsufz/REcoTox, applied by Kramer et al. (2024)). Currently, it seems to be a maturing package that is as yet not extensively documented and is not available via the CRAN repository. This makes it difficult to compare with the ECOTOXr package and its features presented here. It does, however, provide features to process the concentration from the ECOTOX database to standardised units, which the ECOTOXr package currently does not offer.

4.4. Future developments

Presently, the ECOTOXr package returns database records in the same format as stored in the database. In many cases this means that information is stored as text whereas it represents something else, such as numeric data. The package would benefit from implementing standardised routines for cleaning and formatting the data. At present, the package only includes methods for formatting and converting the CAS registry numbers. As these identifiers are known to be ambiguous, better support for other chemical identifiers are desirable. The database already contains Distributed Structure-Searchable Toxicity identifiers (DTXSID, Grulke et al. (2019)), which can be linked to generally accepted International Chemical Identifiers (InChI) keys via CompTox (Williams et al., 2017). The ECOTOXr package already indirectly supports this. The example provided as Supplemental Information (S5), illustrates how to search for trivial chemical names, rather then its systematic name or CAS number.

Currently, retrieved concentrations are returned as reported by (and in the units of) its source material. It would be an improvement to provide standardised routines to convert the reported units to a single specific unit. In some cases that would require external information such as molar weights or density of the applied test medium. The ECOTOX website provides concentrations converted to standardised units for many of the aquatic records. This process and the challenges associated with converting all units to a standardised unit are described in the unit conversion logic section of the ECOTOX website (https://cfpub.epa.gov/ecotox/help.cfm). Should the effort be undertaken to convert to single specific unit with ECOTOXr Package, it could be beneficial to follow similar unit conversion logic. Rather than developing this from scratch, the ECOTOXr could also benefit from features currently developed in other packages under similar licensing (e.g., REcoTox).

In computational toxicology (Grulke et al., 2019) it will be helpful to link the ECOTOX data to other sources of information. In fact, interoperability by combining different sources and exchanging of information is an important aspect of the FAIR principle (Wilkinson et al., 2016). Such information entails, for instance, species meta-information or chemical characteristics. The present set-up of the package already allows the linking of other sources but will require some effort. In that light it is good to point out that the ECOTOX database contains reference identifiers to the Distributed Structure-Searchable Toxicity (DSSTox) database (Grulke et al., 2019). This database contains a wide variety of measured and predicted physico-chemical properties of substances.

Vice versa, it can also be helpful to link computed chemical properties to toxicological information from the ECOTOX database. For this purpose, the Comptox dashboard can be a helpful instrument (Williams et al., 2017). In fact, the ECOTOXr already provides an experimental feature for searching this dashboard (websearch_comptox()).

Not only chemicals but species can be queried using alternative identifiers as well. The National Center for Biotechnology Information Taxonomical Identifier (NCBI TaxID) is available for species in the database. This can be used to collect meta-information, such as genetic and protein information, for those species (https://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi). The software package will benefit from further automation of combining data from different providers.

The database does not contain direct information on data quality (see section 1.2), such as CRED scoring (Moermond et al., 2016). The process of scoring data quality and adequacy could also be facilitated by the software presented here. This can be achieved by adding tables to the database schema, containing meta information which refers to the unique identifiers of test results. Adding additional tables neither affects nor compromises the data originally collected from the EPA database. The software presented here can be used to add, export, share and maintain such additional data. In addition, the database contains information (see for instance Table 2 of Olker et al. (2022), and case study 3) that can assist researchers with the evaluation of data quality and partly automate the process.

5. Conclusions and recommendations

The ECOTOXr package is successful in producing specific and traceable subsets of the ECOTOX database. Given the design of the software, the dissemination of the ECOTOX database and the lessons learned from the case studies presented here, it is concluded that transparency and reproducibility is optimised when:

- An official package release from the CRAN repository is used and cited (including the version of the package and the database used: cite_ecotox()).
- It is asserted that the local copy of the database is clean and unaltered (build a fresh copy when in doubt by using download_ecotox_data()).
- Only non-accented alphanumerical characters are used in search terms (as other characters may not reproduce well on different systems).
- The scripting code used for searching and subsetting the database is documented and it includes database record identifiers (and possibly the date modified).
- Record identifiers are stored and included in the process.

By documenting data selection and processing in R scripts following the steps above, researchers ensure that they contribute to the FAIR use, credibility and acceptance of their work.

CRediT authorship contribution statement

Pepijn de Vries: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data is either provided as supplemental info, or available from the EPA ECOTOX database and mentioned papers.

Acknowledgements

The Wageningen Marine Research department of Wageningen University & Research is acknowledged for funding the work presented in this paper. Anonymous reviewers are acknowledged for valuable suggestions and feedback, which helped improve the quality of the manuscript. Geke Lianne den Otter is acknowledged for proofreading the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemosphere.2024.143078.

References

Akerman, G., Blankinship, A., 2015. EDSP Weight of Evidence Conclusions on the Tier 1 Screening Assays for the List 1 Chemicals. Official Memorandum. United States Environmental Protection Agency. https://downloads.regulations.gov/EPA-HQ-OPP-2012-0330-0039/content.pdf.

Beasley, A., Belanger, S.E., Otter, R.R., 2015. Stepwise information-filtering tool (SIFT): a method for using risk assessment metadata in a nontraditional way. Environ. Toxicol. Chem. 34 (6), 1436–1442. https://setac.onlinelibrary.wiley.com/doi/abs/10.1002/etc.2955.

Benford, F., 1938. The law of anomalous numbers. Proc. Am. Phil. Soc. 78 (4), 551–572. http://www.jstor.org/stable/984802.

Boettiger, C., Chamberlain, S., Hart, E., Ram, K., 2015. Building software, building community: lessons from the rOpenSci project. J. Open Res. Software 3 (1), 8–e8.

Borgert, C.J., Mihaich, E.M., Quill, T.F., Marty, M., Levine, S.L., Becker, R.A., 2011. Evaluation of EPA's tier 1 endocrine screening battery and recommendations for improving the interpretation of screening results. Regul. Toxicol. Pharmacol. 59 (3), 397–411. https://www.sciencedirect.com/science/article/pii/S0273230011000055.

Brock, T.C.M., Elliott, K.C., Gladbach, A., Moermond, C., Romeis, J., Seiler, T.-B., Solomon, K., Peter Dohmen, G., 2021. Open science in regulatory environmental risk assessment. Integrated Environ. Assess. Manag. 17 (6), 1229–1242. https://setac.onlinelibrary.wiley.com/doi/abs/10.1002/ieam.4433.

Carnesecchi, E., Mostrag, A., Ciacci, A., Roncaglioni, A., Tarkhov, A., Gibin, D., Sartori, L., Benfenati, E., Yang, C., Dorne, J.L.C., 2023. Openfoodtox: EFSA's chemical hazards database. https://doi.org/10.5281/zenodo.8120114.

Chambers, J.M., 2020. S, R, and data science. The R Journal 12 (1), 462–476. https://doi.org/10.32614/RJ-2020-028.

Connors, K.A., Beasley, A., Barron, M.G., Belanger, S.E., Bonnell, M., Brill, J.L., de Zwart, D., Kienzler, A., Krailler, J., Otter, R., Phillips, J.L., Embry, M.R., 2019. Creation of a curated aquatic toxicology database: envirotox. Environ. Toxicol. Chem. 38 (5), 1062–1073. https://doi.org/10.1002/etc.4382.

Crawley, M.J., 2012. The R Book. John Wiley & Sons, Ltd., Chichester.

De Vries, P., 2018. Targeted Selection of Existing Aquatic in Vivo Bioassay in Ecotoxicological Hazard Quantification. Wageningen University. https://doi.org/10.18174/430498. Ph.D. thesis.

De Vries, P., 2024. ECOTOXr: download and extract data from US EPA's ECOTOX database. R package version 1.0.9. https://CRAN.R-project.org/package=ECOTOXr.

De Vries, P., Jak, R.G., Frost, T.K., 2022. Comparison of substance-based and whole-effluent toxicity of produced water discharges from Norwegian offshore oil and gas installations. Environ. Toxicol. Chem. 41 (9), 2285–2304. https://setac.onlinelibrary.wiley.com/doi/abs/10.1002/etc.5414.

De Vries, P., Murk, A.J., 2013. Compliance of LC50 and NOEC data with Benford's Law: an indication of reliability? Ecotoxicol. Environ. Saf. 98, 171–178.

Dorne, J., Richardson, J., Livaniou, A., Carnesecchi, E., Ceriani, L., Baldin, R., Kovarich, S., Pavan, M., Saouter, E., Biganzoli, F., Pasinato, L., Zare Jeddi, M., Robinson, T., Kass, G., Liem, A., Toropov, A., Toropova, A., Yang, C., Tarkhov, A., Georgiadis, N., Di Nicola, M., Mostrag, A., Verhagen, H., Roncaglioni, A., Benfenati, E., Bassan, A., 2021. EFSA's OpenFoodTox: an open source toxicological database on chemicals in food and feed and its future developments. Environ. Int. 146, 106293. https://www.sciencedirect.com/science/article/pii/S0160412020322480.

Dulio, V., van, Bavel, Bertand Brorström-Lundén, E., Harmsen, J., Hollender, J., Schlabach, M., Slobodnik, J., Thomas, K., Koschorreck, J., 2018. Emerging pollutants in the EU: 10 years of NORMAN in support of environmental policies and regulations. Environ. Sci. Eur. 30 (1), 5. https://doi.org/10.1186/s12302-018-0135-3.

Dyer, S.D., Versteeg, D.J., Belanger, S.E., Chaney, J.G., Mayer, F.L., 2006. Interspecies correlation estimates predict protective environmental concentrations. Environ. Sci. Technol. 40 (9), 3102–3111. https://doi.org/10.1021/es051738p pMID: 16719118.

ECETOC, 2024. TR 091 – ECETOC Aquatic Toxicity (EAT) database – EAT database. https://www.ecetoc.org/publication/tr-091-ecetoc-aquatic-toxicity-eat-database-eat-database/. (Accessed 12 January 2024).

EFSA, 2024. OpenFoodTox: Chemical Hazards Database. https://www.efsa.europa.eu/en/microstrategy/openfoodtox. (Accessed 5 January 2024).

EPA, 1992. Pesticide Ecotoxicity Database (Formerly: Environmental Effects Database (EEDB)). U.S. Environmental Protection Agency. Environmental Fate and Effects Division, U.S.EPA, Washington, D.C.

Eriksson, E., Baun, A., Mikkelsen, P.S., Ledin, A., 2007. Risk assessment of xenobiotics in stormwater discharged to Harrestrup Å, Denmark. Desalination 215 (1), 187–197 mEDAWATER International Conference on Sustainable Water Management, Rational Water Use, Wastewater Treatment and Reuse. https://www.sciencedirect.com/science/article/pii/S0011916407004109.

Evangelisti, M., Parenti, M.D., Varchi, G., Franco, J., vom Brocke, J., Karamertzanis, P.G., Del Rio, A., Bichlmaier, I., 2023. A non-clinical and clinical IUCLID database for 530 pharmaceuticals (part I): methodological aspects of its development. Regul. Toxicol. Pharmacol. 142, 105416. https://www.sciencedirect.com/science/article/pii/S0273230023000843.

Gentleman, R., Lang, D.T., 2007. Statistical analyses and reproducible research. J. Comput. Graph Stat. 16 (1), 1–23. https://doi.org/10.1198/106186007X178663.

Grulke, C.M., Williams, A.J., Thillanadarajah, I., Richard, A.M., 2019. EPA's DSSTox database: history of development of a curated chemistry resource supporting computational toxicology research. Computational Toxicology 12, 100096. https://www.sciencedirect.com/science/article/pii/S2468111319300234.

Hipp, R.D., 2021. SQLite. https://www.sqlite.org/index.html.

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. J. Comput. Graph Stat. 5 (3), 299–314. https://www.tandfonline.com/doi/abs/10.1080/10618600.1996.10474713.

Kramer, L., Schulze, T., Klüver, N., Altenburger, R., Hackermüller, J., Krauss, M., Busch, W., 2024. Curated mode-of-action data and effect concentrations for chemicals relevant for the aquatic environment. Sci. Data 11 (1), 60. https://doi.org/10.1038/s41597-023-02904-7.

Mebane, C.A., Sumpter, J.P., Fairbrother, A., Augspurger, T.P., Canfield, T.J., Goodfellow, W.L., Guiney, P.D., LeHuray, A., Maltby, L., Mayfield, D.B., McLaughlin, M.J., Ortego, L.S., Schlekat, T., Scroggins, R.P., Verslycke, T.A., 2019. Scientific integrity issues in environmental toxicology and chemistry: improving research reproducibility, credibility, and transparency. Integrated Environ. Assess. Manag. 15 (3), 320–344. https://setac.onlinelibrary.wiley.com/doi/abs/10.1002/ieam.4119.

Moermond, C.T., Kase, R., Korkaric, M., Ågerstrand, M., 2016. CRED: criteria for reporting and evaluating ecotoxicity data. Environ. Toxicol. Chem. 35 (5), 1297–1309. https://setac.onlinelibrary.wiley.com/doi/abs/10.1002/etc.3259.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 339. https://www.bmj.com/content/339/bmj.b2535.

Müller, K., Wickham, H., James, D.A., Falcon, S., 2021. RSQLite: 'SQLite' Interface for R. R package version 2.2.8. https://CRAN.R-project.org/package=RSQLite.

Newcomb, S., 1881. Note on the frequency of use of the different digits in natural numbers. Am. J. Math. 4 (1), 39–40. http://www.jstor.org/stable/2369148.

NORMAN, 2024. NORMAN ecotoxicology database. https://www.norman-network.com/nds/ecotox/, 2024-01-05.

Olker, J., 2022. ECOTOXicology Knowledgebase System User Guide - Version 5.5. U.S. Environmental Protection Agency (EPA), Duluth, Minnesota, USA. https://nepis.epa.gov/Exe/ZyPDF.cgi/P10164D9.PDF?Dockey=P10164D9.PDF.

Olker, J.H., Elonen, C.M., Pilli, A., Anderson, A., Kinziger, B., Erickson, S., Skopinski, M., Pomplun, A., LaLone, C.A., Russom, C.L., Hoff, D., 2022. The ECOTOXicology Knowledgebase: a curated database of ecologically relevant toxicity tests to support environmental research and risk assessment. Environ. Toxicol. Chem. 41 (6), 1520–1539. https://setac.onlinelibrary.wiley.com/doi/abs/10.1002/etc.5324.

Peters, A., Beking, M., Oste, L., Hamer, M., Vuaille, J., Harford, A.J., Backhaus, T., Lofts, S., Svendsen, C., Peck, C., 2023. Assessing the relevance of environmental exposure data sets. Integrated Environ. Assess. Manag. 20 (4), 1004–1018. https://setac.onlinelibrary.wiley.com/doi/abs/10.1002/ieam.4881.

Poisot, T., 2015. Best publishing practices to improve user confidence in scientific software. Ideas in Ecology and Evolution 8, 50–54. https://doi.org/10.4033/iee.2015.8.8.f.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

R Special Interest Group on Databases (R-SIG-DB), Wickham, H., Müller, K., 2022. DBI: R database interface. R package version 1 (1.3). https://CRAN.R-project.org/package=DBI.

Rudén, C., Adams, J., Ågerstrand, M., Brock, T.C., Poulsen, V., Schlekat, C.E., Wheeler, J. R., Henry, T.R., 2017. Assessing the relevance of ecotoxicological studies for regulatory decision making. Integrated Environ. Assess. Manag. 13 (4), 652–663. https://setac.onlinelibrary.wiley.com/doi/abs/10.1002/ieam.1846.

Scharmüller, A., Schreiner, V.C., Schäfer, R.B., 2020. Standartox: standardizing toxicity data. Data 5 (2). https://www.mdpi.com/2306-5729/5/2/46.

Schwesig, D., Borchers, U., Chancerelle, L., Dulio, V., Eriksson, U., Farré, M., Goksoyr, A., Lamoree, M., Leonards, P., Wegener, J.-W., Lepom, P., Leverett, D., O'Neill, A., Robinson, R., Silharova, K., Tolgyessy, P., Slobodnik, J., Tutundjian, R., Westwood, D., 2011. A harmonized european framework for method validation to support research on emerging pollutants. TrAC, Trends Anal. Chem. 30 (8), 1233–1242 climate-Change Impacts on Water Chemistry. https://www.sciencedirect.com/science/article/pii/S0165993611001518.

Sheffield, T.Y., Judson, R.S., 2019. Ensemble QSAR modeling to predict multispecies fish toxicity lethal concentrations and points of departure. Environ. Sci. Technol. 53 (21), 12793–12802. https://doi.org/10.1021/acs.est.9b03957 pMID: 31560848.

Solbé, J., Mark, U., Buyle, B., Guhl, W., Hutchinson, T., Kloepper-Sams, P., Länge, R., Munk, R., Scholz, N., Bontinck, W., Niessen, H., 1998. Analysis of the Ecetoc aquatic toxicity (EAT) database I – general introduction. Chemosphere 36 (1), 99–113. https://www.sciencedirect.com/science/article/pii/S0045653597100236.

Szöcs, E., Stirling, T., Scott, E.R., Scharmüller, A., Schäfer, R.B., 2020. webchem: an R package to retrieve chemical information from the web. J. Stat. Software (13), 1–17. Articles 93. https://www.jstatsoft.org/v093/i13.

Wang, N., 2018. Increasing the reliability and reproducibility of aquatic ecotoxicology: learn lessons from aquaculture research. Ecotoxicol. Environ. Saf. 161, 785–794. https://www.sciencedirect.com/science/article/pii/S0147651318305311.

Wickham, H., 2011. testthat: get started with testing. The R Journal 3 (1), 5–10. https://doi.org/10.32614/RJ-2011-002.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. J. Open Source Softw. 4 (43), 1686. https://doi.org/10.21105/joss.01686.

Wickham, H., Çetinkaya-Rundel, M., Grolemund, G., 2023. R for Data Science, second ed. O'Reilly Media, Sebastopol, CA https://r4ds.hadley.nz/.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR guiding principles for scientific data management and stewardship. Sci. Data 3 (1), 160018. https://doi.org/10.1038/sdata.2016.18.

Williams, A., Grulke, C., Edwards, J., Mansouri, K., Patlewicz, G., Shah, I., Wambaugh, J., , R.S, J., Richard, A., 2017. The comptox chemistry dashboard: a community data resource for environmental chemistry. J. Cheminf. 9 https://doi.org/10.1186/s13321-017-0247-6.

Wilson, G., Aruliah, D.A., Brown, C.T., Chue Hong, N.P., Davis, M., Guy, R.T., Haddock, S.H.D., Huff, K.D., Mitchell, I.M., Plumbley, M.D., Waugh, B., White, E.P., Wilson, P., 2014. Best practices for scientific computing. PLoS Biol. 12 (1), 1–7. https://doi.org/10.1371/journal.pbio.1001745.