ENABLING SEMANTIC DATA ACCESS FOR TOXICOLOGICAL RISK ASSESSMENT

Erik B. Myklebust[1,2]

Ernesto Jiménez-Ruiz[2,3,4]

Jiaoyan Chen[5]

Raoul Wolf[1]

Knut Erik Tollefsen[1]

[1]Norwegian Institute For Water Research, Oslo, Norway

[2]Department of Informatics, University of Oslo, Oslo, Norway

[3]The Alan Turing Institute, London, United Kingdom

[4]City, University of London, London, United Kingdom

[5]Department of Computer Science, University of Oxford, Oxford, United Kingdom

Abstract

Experimental effort and animal welfare are concerns when exploring the effects a compound has on a organism. Appropriate methods for extrapolating chemical effects can further mitigate these challenges. In this paper we present the efforts to (i) (pre)process and gather data from public and private sources, varying from tabular files to SPARQL endpoints, and (ii) integrate the data and represent them as a knowledge graph with richer semantics. This knowledge graph is further applied to facilitate the retrieval of the relevant data for a ecological risk assessment task (i.e., extrapolation of effect data) where two prediction techniques are applied.

Keywords: Toxicology, Ecology, Risk Assessment, Knowledge Graph, Semantic Web, Effect Prediction

1 Introduction

Expanding the scope of ecological risk assessment models is a long-term goal in ecotoxicological research. However, the limiting factor in risk assessment is often availability of toxicological effect data for a given compound or a given organism (species). The potential use of ten to hundreds of test organisms becomes ethically questionable from an animal welfare perspective. Moreover, collection of these data are labour- and cost-intensive and often require extensive laboratory experiments.

One major challenge in risk assessment processes is the interoperability of data. In this paper we present the effort to enable the (semantic) interoperability of relevant data sources and the application to extrapolate effect data[1].

[1] This paper focuses and extends on the data wrangling challenges (*e.g.*, data access, data preparation and data integration) while our paper in Myklebust et al., 2019 has a special focus on the use of knowledge graph embeddings and machine learning for toxicological effect prediction.

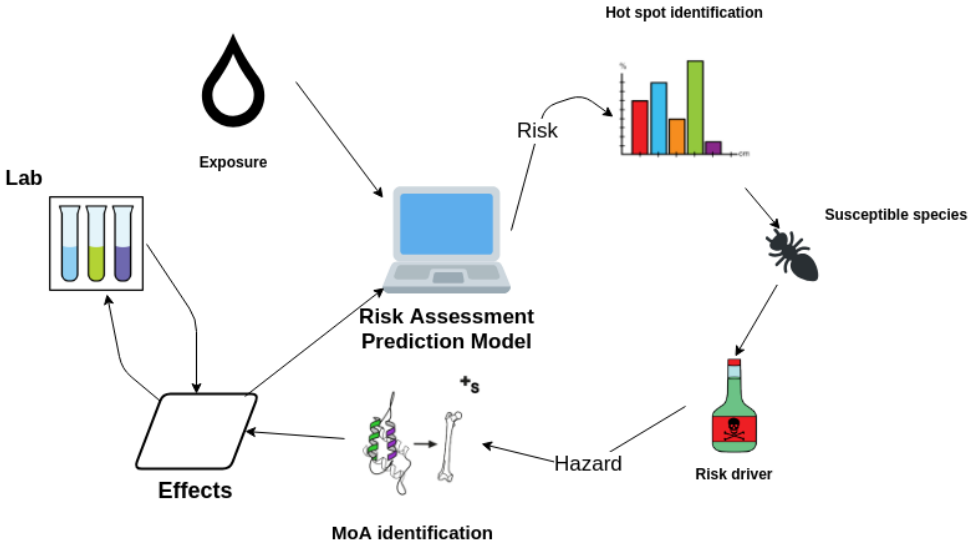

Figure 1. Ecological risk assessment pipeline.

The data sources vary from tabular, to RDF files and SPARQL queries over public linked data. From these sources we create the *Toxicological Effect and Risk Assessment* (TERA) knowledge graph. Certain sources are very large and frequently updated, therefore the knowledge graph is only partially materialized, with the remaining data added upon request to the APIs that are created to interact with TERA.
TERA is used to enable easier data access and toxicological effect extrapolation. Hence, we present data access challenges faced in risk assessment using today's tabular systems, that can be solved by moving to a graph based database. We also present two approaches to extrapolate chemical toxicity on organisms, based on: *(i)* a naive taxonomic distance, and *(ii)* a knowledge graph embedding approach. These approaches show the power of increasing background knowledge modeled by a knowledge graph in ecological risk assessment.
This work is of great importance to future effect and risk assessment approaches within ecotoxicology, and in particular to the work within the Computational Toxicology Program (NCTP)[2] at the Norwegian Institute of Water Research (NIVA), whose strategic goal is designing and developing prediction models to assess the hazard and risks of chemicals and their mixtures where traditional laboratory data cannot easily be acquired.[3] Our contribution will enable the semantic data access across data sets, and facilitate resource-effective and transparent approaches to optimise this work.
This paper is organized as follows. Section 2 provides some background to facilitate the understanding of the subsequent sections. In Section 3 we present the data sources used to construct the TERA knowledge graph. The creation of the knowledge graph is described in Section 4.1, while Section 4.2 discusses the use of APIs and SPARQL queries for data access. Section 5 describes the application of the knowledge graph to effect prediction, while Section 6 elaborates on the contribution of this work and future research directions.

2 Background

Ecological risk assessment and effect prediction. Ecotoxicology is a multidisciplinary field that studies the potentially adverse toxicological effects of chemicals on individuals, populations, communities and ecosystems. In this context, risk is the result of the intrinsic hazards of a substance combined with an estimate of the environmental exposure (*i.e.*, the product of exposure and effect (hazard)).

[2] `http://www.niva.no/nctp`
[3] See `http://www.niva.no/radb`

Figure 1 shows the risk assessment pipeline used at NIVA. *Exposure* data is gathered from analysis of environmental concentrations of one or more chemicals, while *effects* (*hazards*) are characterized for a number of species in the laboratory as a proxy for more ecologically relevant organisms. These two data sources are used to calculate risk, using so-called assessment factors to extrapolate a risk quotient (RQ; ratio between exposure and effects). The RQ for one chemical and/or the mixture of many chemicals is used to identify chemical(s) with the highest RQs (risk drivers), susceptible species (or taxa), identify relevant modes of action[4] (MoA) and characterize detailed toxicity mechanisms for one or more species (or taxa) as described in more detail in Tollefsen (2017). Results from these predictions can generate a number of new hypotheses that can be investigated in the laboratory or studied in the environment.

The effect data is obtained during laboratory experiments, where the sub-population of a single species is exposed to a gradient of concentrations of a chemical. Most commonly, mortality rate, growth, development or reproductive output are measured over time.

To give a good indication of the toxicity to a species, these experiments are conducted with a concentration range spanning from no effect (0%) to complete effect (100%) when this is pragmatically possible. Hence some compounds will be more toxic than others and variance in susceptibility between species may provide a distribution of the effective concentration for one specific compound.

Ecological risk assessment require large amounts of effect data to efficiently predict risk for the ecosystems and ecosystem components (*e.g.*, species and taxa). The data must cover a minimum number of the chemicals found when analysing water samples, along with covering species and taxa present in the ecosystem. This leads to an immense search space that is close to impossible to encompass in its entirety and risk assessment is thus often limited by lack of sufficient high quality effect data. It becomes essential to extrapolate from known to unknown combinations of chemical-species pairs, which in some degree can be overcome by predicting the effects themselves through the use of quantitative structure-activity relationship models (QSARs). These models have shown promising results for use in risk assessment, *e.g.*, Pradeep et al. (2016), albeit have limited application domain (coverage), both in terms of compounds and species. Use of read-across and selection of proxy compounds that are chemically similar, display similar toxicity or have similar MoA and toxicity mechanisms are therefore becoming an intuitively attractive solution with increasing popularity (*e.g.*, Netzeva et al. (2008) and Wu et al. (2010)). Development of computational approaches that identify data that can be used for identifying proxy compounds to be used for read-across and data gap filling, is key to facilitate rapid, cost-effective, reliable and transparent predictions of new effects. This can populate risk assessments with high quality data. We contribute in this regard by creating a semantic layer, *i.e.*, a knowledge graph, to enable extraction of this high quality data.

Knowledge graphs. We follow Arnaout et al. (2018) in the notion of a RDF-based knowledge graph which is represented as a set of RDF triples $\langle s, p, o \rangle$, where s represents a subject (a class or an instance), p represents a predicate (a property) and o represents an object (a class, an instance or a data value *e.g.*, text, date and number). RDF entities (*i.e.*, classes, properties and instances) are represented by URIs (Uniform Resource Identifier). A knowledge graph consits of a terminology and a assertions box (TBox and ABox). The TBox is composed by RDF Schema constructs like class subsumption ((*e.g.*, `ncbi:taxon/6668` `rdfs:subClassOf` `ncbi:taxon/6657`)) and domain and range for properties (*e.g.*, `ecotox:concentration` `rdfs:domain` `ecotox:Chemical`).[5] The ABox contains relationships among entities and semantic type definitions (*e.g.*, `ecotox:taxon/28868` `rdf:type` `ecotox:Taxon`).

SPARQL Queries. RDF-based knowledge graphs can be accessed by SPARQL query language.[6] The common SPARQL constructs used in this work are:

[4] The functional or anatomical change in an organism due to exposure to a compound is called MoA.

[5] The OWL 2 ontology language provides more expressive constructors. Note that the graph projection of an OWL 2 ontology can be seen as a knowledge graph (*e.g.*, Agibetov et al. (2018)).

[6] `https://www.w3.org/TR/rdf-sparql-query/`

Proportion	Endpoint	Endpoint description
0.21	NR	Not reported
0.17	NOEL	No-observable-effect-level
0.16	LC50	Lethal concentration for 50% of test population
0.14	LOEL	Lowest-observable-effect-level
0.05	NOEC	No-observable-effect-concentration
0.05	EC50	Effective concentration for 50% of test population
0.04	LOEC	Lowest observable effect concentration
0.03	BCF	Bioconcentration factor
0.02	NR-LETH	Lethal to 100% of test population
0.02	LD50	Lethal dose for 50% of test population
0.11	Other	

Table 1. Frequency of experimental results in ECOTOX.

1. *Property paths* express multiple edges in a graph. *e.g.*, alternate path (*e.g.*, `rdfs:label | foaf:name`), path sequence (*e.g.*, `rdf:type / rdfs:subClassOf`), inverse relations (*e.g.*, `^rdf:type`), and any combination of these.
2. *Filter* is used to filter the results from a query. We use this in Listing 4.1 along with *isBlank* and negation to filter for classes that are not blank nodes.
3. A *blank node* is a node where the identifier is not explicitly given. This allows us to use temporary nodes in queries. *e.g.*, In Listing 4.5 we use `[rdfs:label ''Langtjern'']` to represent a node (a lake in this case) with label *Langtjern*.

Moreover, the extended syntax of SPARQL enables the use of complex property paths (*e.g.*, a path of minimum 1 to maximum *n* `rdfs:subClassOf` relations is represented as `rdfs:subClassOf{1,n}`), sub-queries (*e.g.*, Listing 5.1), aggregations (*e.g.*, AVG in Listing 5.1) and more.[7]

Ontology alignment. Finding the corresponding mappings between a source and a target ontology or knowledge graph is called ontology alignment (Euzenat et al., 2013). The equivalence mappings are represented as triples among the entities of the source and target (*e.g.*, `ncbi:taxon/13402 owl:sameAs ecotox:taxon/Carya`).

3 Data sources

The TERA knowledge graph is constructed from a number of sources, including tabular data, RDF triples and SPARQL endpoints.

Effect data. The largest publicly available repository of effect data is the ECOTOXicology knowledgebase (ECOTOX) developed by the US Environmental Protection Agency (U.S. EPA, 2019). This data is gathered from published toxicological papers and limited internal experiments. The dataset consists of 940*k* experiments using 12*k* compounds and 13*k* species, implying a compound-species pair converge of maximum $\sim 0.6\%$. The resulting endpoint from an experiment is categorised in one of a plethora of predefined endpoints. For example, the *LC*50 endpoint implies lethal concentration for 50% of the test population. Table 1 shows the most frequent endpoints in ECOTOX. For endpoints such as *EC*50, an effect must be defined in conjunction with the endpoint. Mortality, chronic, and reproductive toxicity are common effect outcomes to characterise the *effective concentration* of a compound upon a given target species.

[7] `https://www.w3.org/wiki/SPARQL/Extensions`

test_id	reference_number	test_cas	species_number
1068553	5390	877430 (2,6-Dimethylquinoline)	5156 (Danio rerio)
2037887	848	79061 (2-Propenamide)	14 (Rasbora heteromorpha)

Table 2. ECOTOX database tests examples.

result_id	test_id	endpoint	conc1_mean	conc1_unit
98004	1068553	$LC50$	400	mg/kg diet
2063723	2037887	$LC10$	220	mg/L

Table 3. ECOTOX database results examples.

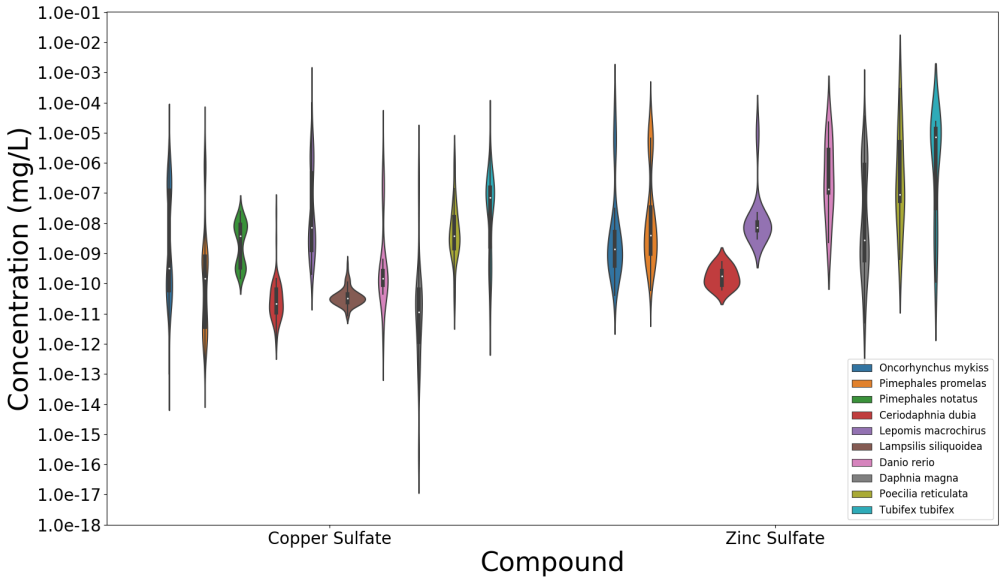

Figure 2. Example of effective concentrations of two sulfates[8] on ten species (where data is available).

Tables 2 and 3 contains an excerpt of the ECOTOX database. ECOTOX includes information about the compounds and species used in the tests. This information, however, is limited and additional (external) resources are required to complement ECOTOX.

Currently, the ECOTOX database is used in ecotoxicological risk assessment as reference data when predicting risk for an ecosystem. Since most compounds have multiple experiments per species, the mean and standard deviation of the effect to a species can be calculated. However, if there is only one experiment for a compound-species pair we cannot calculate a standard deviation, such that the hazard characterisation becomes featureless. Therefore, predicting new effects is important to represent the natural variability of the effect data, as shown in Figure 2. For certain distributions in Figure 2, they consists of two log-normal distributions with different means. This comes down to many factors, such as duration (*e.g.*, lower concentration implies longer experiment duration for the same effect), effected compounds due to test conditions (*e.g.*, temperature, pH, PKa, ionic content), and organism traits (*e.g.*, strain, age, life stage, and size).

Compounds. The ECOTOX database use an identifier called CAS Registry Number assigned by the

[8] Sulfate is a functional group. Here, it binds to metals to create salts.

Chemical Abstracts Service to identify compounds. The CAS numbers are proprietary, however, Wikidata (Vrandecic et al., 2014) (indirectly) encodes mappings between CAS numbers and open identifiers like *InChIKey*, a 27-character hash of the International Chemical Identifier (InChI) which encodes chemical information uniquely[9] (Heller et al., 2015). Moreover, chemical features can be gathered from the chemical information dataset PubChem (Kim et al., 2018) using the open identifiers.
The classification of compounds in PubChem only concerns permutations of compounds. Therefore, we use the (Ch)EBI SPARQL endpoint to access the ChEMBL dataset, which enables us to create a more extensive classification hierarchy.

Taxonomy. ECOTOX contains a taxonomy, however, this only considers the species represented in the ECOTOX effect data. Hence, to enable extrapolation of effects across a larger taxonomic domain, we introduce the NCBI taxonomy (Sayers et al., 2008). This taxonomy data source consists of a number of database dump files, which contains a hierarchy for all sequenced species, which equates to around 10% of the currently known life on Earth. For each of the taxa (species and classes), the taxonomy defines a handful of labels, most commonly used are the *scientific* and *common* names. However, labels such as *authority* can be used to see the citation where the species was first mentioned, while *synonym* is a alternate *scientific* name, that may be used in literature. Other data include the *gencodes* of species and the *host*, where applicable, *e.g.*, for bacteria.

Species traits. As an analog to chemical features, we use species traits to expand the usability of the knowledge graph. The traits we have included in the knowledge graph are the habitat, endemic regions, and presence. This data is gathered from the Encyclopedia of Life (EOL) (Parr et al., 2014), which is available as tabular files. Moreover, EOL uses external definitions of certain concepts, and mappings to these sources are available as glossary files. In addition to traits, researcher may be interested in species that have different conservation statuses, *e.g.*, if the population is stable or declining, etc. This data can also be extracted from EOL.

4 Data wrangling

We perform data wrangling to prepare and incorporate the relevant data to enable the correct research decisions. To facilitate the integration of new data sources (regardless of format), we use semantic technologies within the data wrangling tasks. The use of semantic technologies in the form of a knowledge graph gives us flexibility without committing to a concrete structure (*i.e.*, schema).

4.1 Data preparation and integration

We have created four APIs for wrangling and incorporating effect, taxonomy, and chemical data into the TERA knowledge graph. These APIs also provide (predefined) methods to access the knowledge in TERA. Figure 3 shows how the data sources integrate into the APIs and how the APIs map among each other. Excluding the SPARQL endpoints,[10] the data can be downloaded from the sources websites.[11]
The APIs for data wrangling and data access are available from the following GitLab repository: `https://gitlab.com/Erik-BM/rappt`.

Species API. This API uses data from various tabular sources to describe the species taxonomy and features.

1. The integration of the the NCBI Taxonomy into the knowledge graph is split into several sub-tasks.

[9] While InChI is unique, InChiKey is not, although collisions are few (Willighagen, 2011)

[10] Wikidata: `https://query.wikidata.org/sparql`
ChEBI: `https://www.ebi.ac.uk/rdf/services/sparql`

[11] ECOTOX: `https://cfpub.epa.gov/ecotox/`
PubChem: `https://pubchemdocs.ncbi.nlm.nih.gov/downloads`
NCBI Taxonomy: `https://www.ncbi.nlm.nih.gov/guide/taxonomy/`

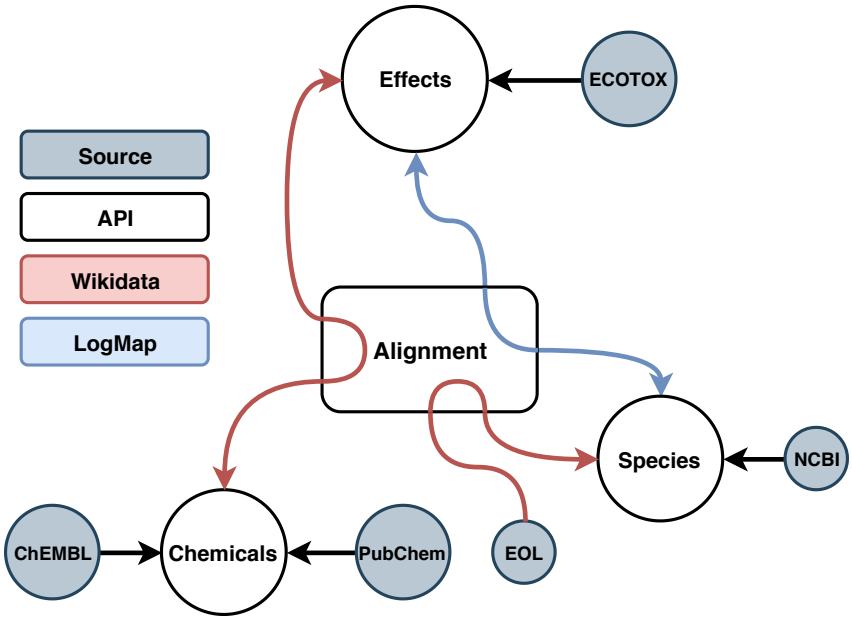

Figure 3. Data sources and colour-coded elements of the TERA knowledge graph.

a) Loading the hierarchical structure included in *nodes.dmp*. The columns of interest are the taxon identifiers of the child and parent taxon, along with the rank of the child taxon and the division where the taxon belongs. We use this to create triples like `(v)` and `(vi)` in Table 4.

b) To aid alignment between NCBI and ECOTOX identifiers, we add the synonyms found in *names.dmp*. Here, the taxon identifier, its name and name type are used to create triples similar to `(vii)` in Table 4. Note that a taxon in NCBI can have a plethora of names while a taxon in ECOTOX usually have two, *i.e.*, common name and Latin name.

c) Finally, we add the labels of the divisions found in *divisions.dmp*. In addition, we add disjointness axioms among all divisions, *e.g.*, Triple `(ii)` in Table 4.

2. The EOL data can be downloaded as a uniform format regardless of the relation to be added to the knowledge graph. Therefore, our approach is universal for endemic, habitat, and other data from EOL.

a) Each dataset for a EOL relation contains a *glossary.tsv* and a *data.tsv* file.

b) The glossary is used to convert the strings (columns Measurement Type and Value) given in the data to URIs. We map the identifiers given in the data to NCBI URIs (see data alignment API section) and create triples using the NCBI URI as subject, with Measurement Type and Value (from EOL) as predicate and objects, as shown in Triples `(viii)` and `(ix)` in Table 4.

c) In addition, EOL gives hierarchies for the Measurement Values in a two column format with parent and child node. Therefore, we can simply add subsumption axioms using these child-parent pairs, as shown in Triple `(xvii)` in Table 4.

Chemical API. The combination of RDF and SPARQL endpoints form the basis for the chemical API:

1. The downloaded *turtle* files (standard format to store RDF graphs) from PubChem can be directly used as they already include RDF triples. Triple `(x)` in Table 4 is an example from these files.

2. To complete the class hierarchy where PubChem provides the bottom level, we query the ChEBI SPARQL endpoint using the query shown in Listing 4.1. Here, we use the values found in `<current>`, to find superclasses that have an edge of type `rdfs:subClassOf` or `rdf:type` to `<current>`. The query is iterated replacing `<current>` with the superclasses resulting from the query (`<current>` can also be replaced with a list). Triple `(xi)` in Table 4 shows an example of result of this query.

#	subject	predicate	object
(i)	ecotox:group/Worms	owl:disjointWith	ecotox:group/Fish
(ii)	ncbi:division/2	owl:disjointWith	ncbi:division/4
(iii)	ncbi:division/2	rdfs:label	‘‘Mammals’’
(iv)	ecotox:taxon/34010	rdfs:subClassOf	ecotox:taxon/hirta
(v)	ncbi:taxon/687295	rdfs:subClassOf	ncbi:taxon/513583
(vi)	ncbi:taxon/687295	ncbi:rank	ncbi:Species
(vii)	ncbi:taxon/687295	ncbi:scientific_name	‘‘Coleophora cornella’’
(viii)	ncbi:taxon/35525	eol:habitat	ENVO:00000873
(ix)	ncbi:taxon/35525	eol:presentIn	worms:Oostende
(x)	compound:CID10198308	rdf:type	obo:CHEBI_134899
(xi)	obo:CHEBI_134899	rdfs:subClassOf	obo:CHEBI_37919
(xii)	compound:CID10198308	pubchem:formula	‘‘$C_7H_6O_6S$’’
(xiii)	ecotox:effect/001	ecotox:compound	ecotox:chemical/115866
(xiv)	ecotox:effect/001	ecotox:species	ecotox:taxon/26812
(xv)	ecotox:effect/001	ecotox:endpoint	ecotox:LC50
(xvi)	ecotox:taxon/33155	owl:sameAs	ncbi:taxon/311871
(xvii)	eol:freshwaterPond	rdfs:subClassOf	ENVO:00000033

Table 4. Example triples from the TERA knowledge graph

3. Since the chemical data is much larger than any of the other data sources used, we do not load chemical features on initialization, but upon request. We use the PubChemPy (Swain, 2014) library to query the PubChem REST API. Triples such as `(xii)` in Table 4 is a results of an API request.

```
SELECT ?class {
    VALUES ?s { <current> }
    ?s rdfs:subClassOf | rdf:type ?class .
    FILTER (!isBlank(?class))
}
```

Listing 4.1. Query superclasses from ChEBI.

Effect API The tabular data in ECOTOX requires significantly more cleaning than the other data.

1. ECOTOX contains metadata about the species and compounds used in the experiments. We use this information to aim alignment between the effect and the background data.
   a) Species metadata in *species.txt* include common and Latin names, along with a (species) ECOTOX group. This group is a categorization of the species based on ECOTOX use cases. We filter the species names, *e.g.*, *sp.*, *var.* (*i.e.*, unidentified species and variant) are removed along with various missing value short hands used in the metadata.
   b) The full hierarchical lineage is also available in the *species.txt* file. Each column represent a taxonomic level, *e.g.*, *genus* or *family*. If a column is empty, we construct a intermediate classification, *e.g.*, say *Daphnia magna* has no genus classification in the data, then its classification will be Daphniidae genus (family name + genus, actually called *Daphnia*). We construct these classifications to ensure the number of levels in the taxonomy is consistent. This consistency will help when aligning to the NCBI data. Note that when adding triples such as `(iv)` in Table 4, we also add a classification based on the column to aid easier querying for a specific taxonomic level.
   c) Chemical metadata in *chemicals.txt* is handled similarly, the data includes chemical name and a (compound) ECOTOX group.

2. The effect data consist of two parts, a test definition and results associated with that test. Note that a test can have multiple results.
   a) The important aspects of a test is the compound and the species used, other columns include metadata, but these are optional and often empty. Each result gives an endpoint (see Table 1), an effect (*e.g.*, chronic or mortal), and a concentration and unit at which the endpoint and effect where recorded.
   b) We construct a node of type result (*e.g.*, `ecotox:effect/001`) and link each result component to it, examples can be seen in `(xiii)-(xv)` in Table 4.

Data alignment API. We use various techniques to align the datasets described above.

ECOTOX-NCBI (Species). There does not exist a complete and public alignment between ECOTOX species and the NCBI taxonomy. Therefore, we have used the LogMap (Jiménez-Ruiz et al., 2011, 2012) ontology alignment system to align the two vocabularies. There exists a partial mapping curated by experts through the ECOTOX search interface,[12] we have gathered a total of 929 mappings for validation purposes. LogMap's lexical indexation gave us 5,472 possible NCBI entities to map to ECOTOX. Around 40% of the ECOTOX (instance) vocabulary was mapped to NCBI covering all 929 expert curated mappings. Hence, an estimated recall of 100%. The TERA knowledge graph include the LogMap mappings as additional equivalence triples, *e.g.*, Triple *(xvi)* in Table 4.

EOL-NCBI (Species). To be able to use the EOL data we need to align the EOL identifiers with NCBI, this can be done through Wikidata as shown in query in Listing 4.2. This query use the Wikidata properties *instance of* (`wdt:P31`), *Encyclopedia of Life ID* (`wdt:P830`), and *NCBI Taxonomy ID* (`wdt:P685`), along with the class *taxon* (`wd:Q16521`).

```
SELECT ?species ?ncbi ?eol WHERE {
        ?species wdt:P31 wd:Q16521 ;
                 wdt:P830 ?eol ;
                 wdt:P685 ?ncbi .
}
```

Listing 4.2. EOL and NCBI identifiers.

ECOTOX-PubChem (Compounds). To enable the interaction between the Chemical API and the effect data we create a mapping between CAS and InChIKey using the SPARQL query shown in Listing 4.3 on the Wikidata endpoint. This query use the Wikidata properties and classes `wdt:P31`, `wdt:P235`, `wdt:P231`, and `wd:Q11173`, which has labels *instance of*, *InChIKey*, *CAS Registry Number*, and *chemical compound*.

```
SELECT DISTINCT ?compound ?inchikey ?cas WHERE {
    ?compound wdt:P31 wd:Q11173 ;
              wdt:P235 ?inchikey ;
              wdt:P231 ?cas .
}
```

Listing 4.3. Compound CAS and InChIKey identifiers.

Requesting chemical features from PubChem requires us to convert InChIKeys to CIDs, fortunately this mapping is available through the PubChem REST API, an example request using PubChemPy is shown in Listing 4.4.

[12] `https://cfpub.epa.gov/ecotox/search.cfm`

```
from pubchempy import get_compounds
inchikey = "MMOXZBCLCQITDF-UHFFFAOYSA-N" # DEET
r = get_compounds(inchikey,"inchikey")
r = [c.to_dict(properties=['cid']) for c in r]
cid = [c['cid'] for c in r] # 4284
```

Listing 4.4. Converting from InChIKey to CID for the pesticide DEET using PubChemPy.

4.2 Data access

The knowledge in TERA can be accessed via SPARQL queries or via predefined APIs. The (final) output will depend on the required task, and can be given either as a graph or in tabular format.

SPARQL queries. For researchers competent in SPARQL the most powerful method for accessing data in TERA is via SPARQL queries. TERA provides an improved and intuitive method for accessing case study data over the current tabular data base structure. We will here give an example of the usability of TERA in extracting data for a risk assessment case study.
The first step in a risk assessment is to define a case study, in Listing 4.5 we define our study area as the lake *Langtjern*. Thereafter, we can extract the compounds and concentrations, at which, the species in the lake experiences lethal effects. The concentrations can then be compared with water samples (*exposure*) from *Langtjern* to see if the endangered species are under threat of going extinct.[13]

```
SELECT ?s ?c ?conc WHERE {
    ?s      eol:habitat eol:Freshwater ;
            eol:presentIn [rdfs:label ''Langtjern''] ;
            eol:conservationStatus eol:endangered .
    []      rdf:type ecotox:Result ;
            ecotox:endpoint ecotox:LC50 ;
            ecotox:effectType ecotox:ACUTE ;
            ecotox:compound ?c ;
            ecotox:concentration ?conc ;
            ecotox:species ?s .
}
```

Listing 4.5. Example query for selecting all species, compounds, and concentrations, where the species is endangered and lives in the freshwater lake Langtjern.

APIs. In addition to SPARQL queries for extracting data from the knowledge graph, the TERA APIs povide predefined methods which enable access to the data without being proficient in SPARQL,[14] but rather prefer a scripting language (*e.g.*, Python).

1. In addition to classification, sibling, and name queries, the Species API has methods for fuzzy querying of identifiers based on close matched names. This is a necessary feature, since the name definition may vary from user to user.
2. Since the Chemical API use the most varied sources, we need to convert between them, therefore, the API can convert between CAS, InChIKey, PubChem ID (called CID) and internal identifiers to interact with the NIVA internal databases. If these identifiers are not sufficient the user can query Wikidata directly.

[13] The comparison can be done with another (case study) API. However, this uses only private data and therefore is not included here.
[14] Methods are, for the most part, abstractions of SPARQL queries.

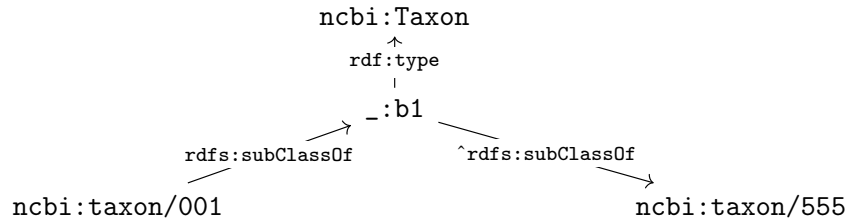

Figure 4. Example sibling graph for query in Listing 5.1.

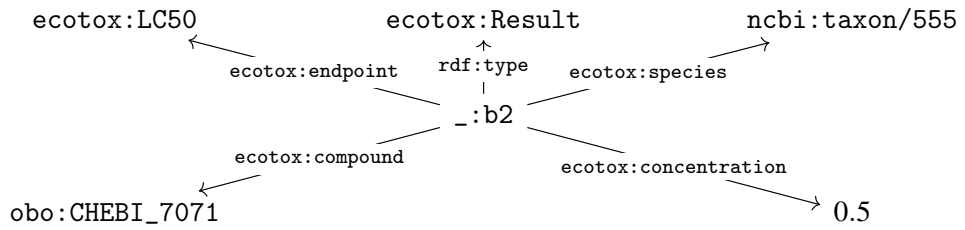

Figure 5. Result found using sibling of `ncbi:taxon/001` *from Figure 4.*

3. As mentioned, the chemical features are not included in the knowledge graph, purely for practical reasons. Therefore, fetching features from PubChem is a method in the API. We also include methods for other properties available in PubChem, such as chemical fingerprints, which is a string of bits representing the presence or absence of selected chemical properties.
4. The Effect API has several methods for mapping between species identifiers (complementing LogMap mappings). These methods use the species names to query the Wikidata SPARQL endpoint and fetch the mappings between identifiers.

5 Effect prediction with TERA

The TERA knowledge graph has been applied to extrapolate effects. We are currently using a deterministic and a probabilistic technique.

Deterministic effect prediction. A taxonomic distance approach where we assume similar compounds and species have a small graph distance. An example of this is shown in Listing 5.1. Here, we construct new triples based on as a mean of the closest taxa in the taxonomy to `ncbi:taxon/001` (compound is kept constant). By changing m, the number of taxa used can be fine tuned. When $m = 1$ the effects are extrapolated from the siblings, while when $m > 1$ the effects are extrapolated from the (*m-1*)th cousins. Figures 4 and 5 shows an example of the select query in Listing 5.1. In Figure 4 we can see that `ncbi:taxon/555` is the sibling of `ncbi:taxon/001`. Then we find all results where `ncbi:taxon/555` is used as test species. Here, there is only one result. Therefore, the construct query will create triples equal to triples in Figure 5 by replacing `ncbi:taxon/555` with `ncbi:taxon/001`. Similarly, one can use the method to estimate effects on taxonomic groups, *e.g.*, the class of crustaceans. This method is intuitive and explainable, and is therefore favoured in ecotoxicological research. However, as shown in Myklebust et al. (2019) this method has it limitations, namely that taxonomic distance is a limited proxy for similarity between species with regard to the effect of a compound or compound combination.

Probabilistic effect prediction. To aid the effect prediction a knowledge graph embedding approach is proposed in Myklebust et al. (2019). The knowledge graph is split into two separate parts: one considers the taxonomy and species features, while the other considers the chemical classifications and features. This method then embeds each part of the knowledge graph into a vector space. The embeddings can be

optimized using different methods. The most intuitive method is to represent the relations as the difference (in the vector space) between subject and object of a triple (called TransE (Bordes et al., 2013)). Thereafter, a machine learning model uses the embeddings to learn from the known effects to estimate new effects of compounds on species. This model take the effect data into account, in addition to the knowledge graph, when the compounds and species are embedded into the vector space, which results in improvement in the representation, and hence, higher performance.

```
CONSTRUCT {
      [] rdf:type ecotox:Result ;
         ecotox:species ncbi:taxon/001 ;
         ecotox:compound obo:CHEBI_7071 ;
         ecotox:endpoint ?e ;
         ecotox:concentration ?concmean .
} WHERE {
    SELECT AVG(?conc) as ?concmean WHERE {
    # Finding cousins of ncbi:taxon/001
    ?s rdfs:subClassOf{1,m} [
                            rdf:type ncbi:Taxon ;
                            ^rdfs:subClassOf{1,m} ncbi:taxon/001
                            ] .
    # Extracting results for cousins found above.
    [] rdf:type ecotox:Result ;
          ecotox:species ?s ;
          ecotox:compound obo:CHEBI_7071 ;
          ecotox:endpoint ?e ;
          ecotox:concentration ?conc .
    }
}
```

Listing 5.1. Effect extrapolation from siblings or cousins.

Effect prediction is an ongoing research line. There are trade-offs for both methods, namely performance against explainability. Therefore, we are exploring more complex models and aiming to use the TERA knowledge graph to perform semantic explaining of them (*e.g.*, Lécué et al. (2018)).

6 Discussion and Conclusion

We have created a knowledge graph called TERA and accompanying tools. This knowledge graph aims at covering the knowledge and data relevant to the ecotoxicological domain. We have also shown the applications of the knowledge graph, including data retrieval and effect prediction. These applications show the benefits of having a integrated view of the different knowledge and data sources.

Knowledge graph. The TERA knowledge graph is by itself an important contribution to NIVA and the hazard and risk assessment community. Different knowledge and data sources are integrated into TERA, which aims at consolidating the relevant information to the ecological risk assessment domain. The adaption of a RDF-based knowledge graph enables the use of an extensive range of Semantic Web infrastructure (*e.g.*, reasoning engines, ontology alignment systems, SPARQL query engines). The accompanying tools enables us to draw conclusions on the effect data from background knowledge, and extrapolate on it.

Value for NIVA. The data integration efforts and the construction of the TERA knowledge graph goes in line with the vision of NIVA's Section for Environmental Data Science. The availability and accessibility of the best knowledge and data will enable optimal decision making. The applications falls into one of

the main research lines of NIVA's Computational Toxicology Program (NCTP) to enrich risk assessment models with improved effect prediction and easier access across data sources.

Future work. The main goal of the near future is to integrate the TERA knowledge graph and the tools presented here into the NIVA risk assessment pipeline. This will help to assess the day-to-day usage of the knowledge graph. Moreover, it will provides feedback on missing features. Later, we will focus on improving the effect prediction models. These models will benefit from expanding TERA with sources previously not included, especially expert curated data and rules.

Acknowledgements

This work is supported by grant 272414 from the Research Council of Norway (RCN), the MixRisk project (RCN 268294), the AIDA project, The Alan Turing Institute under the EPSRC grant EP/N510129/1, the SIRIUS Centre for Scalable Data Access (RCN 237889), the Royal Society, EPSRC projects DBOnto, MaSI^3 and ED^3, and is organized under the Computational Toxicology Program at NIVA. We would also like to thank Martin Giese and Zofia C. Rudjord for their contribution in different stages of this project.

References

Agibetov, A., E. Jiménez-Ruiz, M. Ondresik, A. Solimando, I. Banerjee, G. Guerrini, C. E. Catalano, J. M. Oliveira, G. Patanè, R. L. Reis, and M. Spagnuolo (2018). "Supporting shared hypothesis testing in the biomedical domain." *J. Biomedical Semantics* 9 (1), 9:1–9:22.

Arnaout, H. and S. Elbassuoni (2018). "Effective Searching of RDF Knowledge Graphs." *Web Semantics: Science, Services and Agents on the World Wide Web* 48 (0). ISSN: 1570-8268.

Bordes, A., N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko (2013). "Translating Embeddings for Modeling Multi-relational Data." In: *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., pp. 2787–2795.

Euzenat, J. and P. Shvaiko (2013). *Ontology Matching, Second Edition*. Springer. ISBN: 978-3-642-38720-3.

Heller, S. R., A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi (2015). "InChI, the IUPAC International Chemical Identifier." *Journal of Cheminformatics* 7 (1), 23. ISSN: 1758-2946.

Jiménez-Ruiz, E. and B. Cuenca Grau (2011). "LogMap: Logic-Based and Scalable Ontology Matching." In: *10th International Semantic Web Conference*, pp. 273–288.

Jiménez-Ruiz, E., B. Cuenca Grau, Y. Zhou, and I. Horrocks (2012). "Large-scale Interactive Ontology Matching: Algorithms and Implementation." In: *the 20th European Conference on Artificial Intelligence (ECAI)*. Montpellier, France: IOS Press, pp. 444–449.

Kim, S., J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton (2018). "PubChem 2019 update: improved access to chemical data." *Nucleic Acids Research* 47 (D1), D1102–D1109. ISSN: 0305-1048.

Lécué, F. and J. Wu (2018). "Semantic Explanations of Predictions." *CoRR* abs/1805.10587. arXiv: 1805.10587. URL: http://arxiv.org/abs/1805.10587.

Myklebust, E. B., E. Jiménez-Ruiz, J. Chen, R. Wolf, and K. E. Tollefsen (2019). "Knowledge Graph Embedding for Ecotoxicological Effect Prediction." In: *Int'l Sem. Web Conf. (ISWC)*.

Netzeva, T. I., M. Pavan, and A. P. Worth (2008). "Review of (Quantitative) Structure–Activity Relationships for Acute Aquatic Toxicity." *QSAR & Combinatorial Science* 27 (1), 77–90.

Parr, C. S., N. Wilson, P. Leary, K. Schulz, K. Lans, L. Walley, J. Hammock, A. Goddard, J. Rice, and M. Studer (2014). *The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth.*

Pradeep, P., R. J. Povinelli, S. White, and S. J. Merrill (2016). "An ensemble model of QSAR tools for regulatory risk assessment." *Journal of cheminformatics* 8, 48–48.

Sayers, E. W., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye (2008). "Database resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 37 (suppl_1), D5–D15. ISSN: 0305-1048.

Swain, M. et al. (2014). *PubChemPy: Python wrapper for the PubChem PUG REST API.* [Online; accessed 15.08.2019]. URL: `https://pubchempy.readthedocs.io/`.

Tollefsen, K. E. (2017). *NIVA Risk Assessment Database (RAdb)*. URL: `www.niva.no/radb`.

U.S. EPA (2019). *ECOTOXicology knowledgebase (ECOTOX)*. URL: `https://cfpub.epa.gov/ecotox/`.

Vrandecic, D. and M. Krötzsch (2014). "Wikidata: a free collaborative knowledgebase." *Commun. ACM* 57 (10), 78–85.

Willighagen, E. (2011). *InChIKey collision: the DIY copy/pastables*. URL: `https://chem-bla-ics.blogspot.com/2011/09/inchikey-collision-diy-copypastables.html`.

Wu, S., K. Blackburn, J. Amburgey, J. Jaworska, and T. Federle (2010). "A framework for using structural, reactivity, metabolic and physicochemical similarity to evaluate the suitability of analogs for SAR-based toxicological assessments." *Regulatory Toxicology and Pharmacology* 56 (1), 67 –81. ISSN: 0273-2300.