Chapter 1

Knowledge Management in the Context of Toxicity Testing

Audrey Sanctorum, Jan Maushagen, Sara Sepehri, Guillaume Vrijens, Joery De Kock, Tamara Vanhaecke, Olga De Troyer and Christophe Debruyne

Abstract

The chapter presents the knowledge management system, developed in the context of an interdisciplinary project called TOXIN, for the toxicity testing domain to facilitate the hazard assessment of new chemical compounds. Tools have been developed to capture existing knowledge captured in Safety Evaluation Opinions documents issued by the Scientific Committee on Consumer Safety in a knowledge graph, to enrich this knowledge with knowledge from other sources, and to access this knowledge efficiently. Ontologies and semantic technology are used to build the toxicological knowledge graph and its tools. The developed knowledge management system is based on the processes for creating, maintaining, and exploiting knowledge graphs defined in the Abstract Reference Architecture. The chapter discusses the approach followed for developing the knowledge management system, and the tools developed to support the different processes of the Abstract Reference Architecture. These tools include end-user tools, as well as more advanced tools for information technology experts.

Keywords: knowledge graph construction, knowledge consumption, knowledge graph enrichment, knowledge graph engineering, data lifting, quality assurance, domain ontology development, end-user development, toxicity testing, jigsaw metaphor

Acronyms

ALP	Alkaline phosphatase
ARA	Abstract Reference Architecture
CAS	Chemical Abstracts Service
CTD	Comparative Toxicogenomics Database
CVS	Comma-Separated Values
EU	European Union
GLP	Good Laboratory Practice (GLP)
GO	Gene Ontology
GO-CAM	Gene Ontology Causal Activity Modeling
HESS	Hazard Evaluation Support System
INCI	International Nomenclature Cosmetic Ingredients
IRI	Internationalized Resource Identifier
IT	Information Technology
MDE	Model-Driven Engineering
NLP	Natural Language Processing
OECD	Organisation for Economic Co-operation and Development
OGG	Ontology of Genes and Genomes
OMG	Object Management Group
OWL	Web Ontology Language
R2RML	Relational Database to RDF Mapping Language
R2RML-F	Extension of R2RML
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RML	RDF Mapping Language
SCCS	Scientific Committee on Consumer Safety
SHACL	Shapes Constraint Language
SM	Subject Matter
SMILES	Simplified Molecular-Input Line-Entry Specification
SPARQL	Standard query language and protocol for RDF triplestores
SQL	Structured Query Language
TOXIN	Non-Animal Methodologies for Toxicity Testing of Chemical Compounds
TXPO	ToXic Process Ontology
UI	User Interface
W3C	World Wide Web Consortium

1. Introduction

Toxicology is a field of science investigating the potentially harmful effects of chemical compounds on living organisms, including humans. An important part of toxicology is the field of toxicity testing, also known as safety assessment, which focuses on investigating the degree of risk associated with chemical substances [3]. Historically, animal testing has formed the basis for safety assessment. Yet, scientific considerations and ethical constraints have driven a worldwide shift towards using animal-free methods for this purpose. This has been enforced by an EU directive N°1223/2009 imposing a full ban on animal testing in the cosmetics field, but also across other sectors; for instance, there is a tendency to address animal-free methods for safety evaluation in pharmaceutical, food, and biocide industries. This has resulted in efforts to develop animal-free methods for evaluating the safety of chemicals, incorporating cell culturing and computational approaches. In general, there are three main toxicity testing methods:

1. *In vivo* toxicity testing involves exposing living animals to a substance to observe its effects on their health and behaviour.

2. *In vitro* toxicity testing involves conducting experiments on cells or tissues taken from a living organism. This method allows for more controlled conditions, as it eliminates the influence of other body systems on the effects of the substance being tested.

3. *In silico* toxicity testing involves using computer models to simulate the effects of a substance on a living organism and how a living organism interacts with a substance. This method relies on mathematical and statistical models and data from in vivo and in vitro studies to predict the toxic effects of a substance.

Although in vitro and in silico testing methods have made advances, in general, different types of methods are combined [26]. For instance, the in vitro approach can be complemented with computational approaches to gather, disclose (i.e., make available in a suitable form) and maintain (i.e., keep up to date) available toxicological information of chemical compounds to avoid redundant testing and/or to predict adverse effects of chemical compounds based on similarity. A wealth of data is available in the toxicological domain, but the available data is in a heterogeneous and varied format, including a diverse range of toxicological and risk assessment reports, regulatory standards, and guidelines. Manual searching through the different documents, reports, and data sources is current practice when information is required. This is a time-consuming process, and, in addition, aggregating knowledge and subsequently searching, analysing, and inferring implicit information from the integrated data is hard. Therefore, one challenge in this context is to provide tools for the efficient access, processing, and analysis of relevant data from the toxicological domain. Developing such tools is one of the objectives of the interdisciplinary project TOXIN[1].

TOXIN is developing a knowledge management system that gathers and organizes information about in vivo tests described in documents issued by the

[1] https://ivtd.research.vub.be/irp-non-animal-methodologies-for-toxicity-testing-of-chemical-compouns-toxin-0

Scientific Committee on Consumer Safety[2] about cosmetic ingredients listed as annexed II, III, IV, V, and VI under Regulation EC No 1223/2009. Each such document, called a Safety Evaluation Opinion, contains information about experiments (also called tests) of a chemical compound on laboratory animals, including information on the outcome of these tests, as well as the authors' opinions about the compound's toxicity. In TOXIN, semantic technology is adopted to build this toxicological knowledge management system. By using semantic technology, the information can be structured flexibly. This is because semantic technology represents and stores information using a graph data model, which is less rigid than the traditional relational model adopted in relational databases. This means that information can be added to entities as needed, avoiding complex data migration. Further, where the schema and data are "tightly coupled" in traditional relational databases, the graph data model in semantic databases are "schemaless"; the "schemas," which are called ontologies later on, are declared as facts themselves, and one can combine several such ontologies. Semantic technology allows us to express the semantics of the information, provides advanced querying mechanisms, and allows us to integrate diverse data from multiple sources into a coherent and meaningful *knowledge graph*. TOXIN's knowledge management system aims to support the non-animal hazard assessment[3] of cosmetic ingredients within the TOXIN project by providing as much knowledge as possible flexibly.

In this chapter, we present the approach followed for developing the knowledge management system, as well as the tools developed for it. The different tools support different processes for creating, maintaining, and exploiting knowledge graphs introduced in [10] by following the Abstract Reference Architecture (ARA). Two tools are provided for defining the knowledge graph of the knowledge management system and filling it with data: an end-user tool, based on the jigsaw metaphor, that allows the manual definition and filling of a knowledge graph by our subject matter experts (i.e., toxicologists), and a tool to automatically import toxicity data previously collected by the toxicologists from Safety Evaluation Opinions in spreadsheets into the knowledge graph. We also discuss integrating multiple other data sources from the field of toxicology into the TOXIN knowledge system to facilitate hazard assessment of new compounds by presenting the relationships integrating the different data sources to the toxicologists. Next, we discuss how we currently tackle aspects of the quality assurance of the knowledge graph. Furthermore, a search and query tool has been developed allowing toxicologists to explore and search in the knowledge graph. We also discuss how integrating the other data sources from the toxicological field can be used to answer questions formulated in the context of toxicity testing.

The chapter is organized as follows: Section 2 presents the background, i.e., introducing the concept of a knowledge graph, as well as the concepts of ontology and vocabulary. Next, the existing ARA framework for engineering knowledge graphs is presented. In Section 3, related work is discussed. Section 4 presents our approach towards knowledge management, and TOXIN's knowledge management system and the various tools developed so far are explained and demonstrated in Section 5. The paper ends with conclusions and future work, which are presented in Section 6.

[2] https://health.ec.europa.eu/scientific-committees/scientific-committee-consumer-safety-sccs_en

[3] Hazard assessment, i.e., evaluating the intrinsic property of a molecule inducing toxicity out of the use context, is the first step of every safety assessment process.

2. Background

In this section, we provide the background of the work. We start by briefly explaining the concepts of a knowledge graph, an ontology, and a vocabulary. Thereafter, we present the ARA framework on which our knowledge management system is based.

2.1 Knowledge Graph, Ontology, and Vocabulary

Knowledge graphs use the concept of a graph to describe knowledge of the real world. In graphs, knowledge is represented with nodes and edges where nodes represent entities for the real world and edges represent the relationships between entities [20]. An edge between two entities is called a triple. A triple is an ordered grouping of a subject, a predicate, and an object. A triple or set of triples can be seen as a directed edge-labelled graph called a data graph. A data graph may be identified by a name. Formally, such a named graph is represented by a pair (n, G) where G is a data graph and n is the unique name (or "name") of the graph. One can also find the so-called *default graph*, which is the only graph without a name.

Graphs can be gathered into a graph dataset. A graph dataset is not composed of triples but rather by quadruples (or *quads*), as the name of the graph (if the graph has a name) is added to the triples: $< g, s, p, o >$ where g is the "name" of the graph and $< s, p, o >$ is the triple in the graph. Concretely, a knowledge graph is any graph dataset or combination of graph datasets.

The Resource Description Framework (RDF) [36] is a W3C recommended abstract graph data model that allows representing data graphs. A set of triples expressed in RDF is called an RDF graph, which can be stored in a named graph.

An ontology can be (re)used or developed to specify which type of knowledge is stored in the knowledge graph [19]. An ontology describes concepts in a domain, the properties of the concepts, the relations between the concepts, and the domain rules that apply to them. In other words, an ontology aims to model a domain with as much precision as possible. In this way, the ontology can be used as a model for the knowledge to be stored, and the knowledge stored can be considered as an instantiation of the ontology [4, 19]. In this case, the knowledge graph is called an *ontology-based knowledge graph*. Using one or more ontologies to build a knowledge graph has the advantage of providing formal definitions of the knowledge that can be stored, its meaning, and possible restrictions on what can be stored. This not only provides an unambiguous description of the knowledge, but it also allows humans, as well as computers, to process the information and infer new knowledge. These ontologies are expressed in ontology languages, of which the Web Ontology Language (OWL) [1] is standardised by the W3C and supports complex reasoning tasks such as concept satisfiability checking (i.e., can a concept have instances?).

Not all knowledge graphs require expressive ontology languages, which come at additional computational costs. Another way to represent the semantics of the relationships within the data graph is using one or more *vocabularies*, which is a name given to lightweight ontologies. These ontologies are either built with a less expressive ontology language, or restrict the use of a complex ontology language to those constructs deemed "light-weight". A vocabulary can be used to define class hierarchies, relation hierarchies, and relationships between classes. A vocabulary can be represented with RDF Schema (RDFS) [15]. RDFS is an extension of RDF. RDFS is meant to infer implicit information from explicit information. For instance, if a graph contains $< Garfield, is\ of\ type, Cat >$

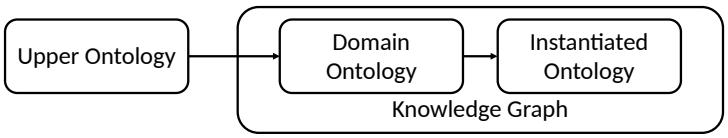

Figure 1.
Ontology-based knowledge graph structure based on the Model-Driven Engineering (MDE) approach (adapted from [4])

and $< \ Cat, is\ a\ subclass\ of, Animal \ >$, then an RDFS reasoner can infer that $< \ Garfield, is\ of\ type, Animal \ >$. While already powerful and often used, it does not support complex reasoning tasks such as checking whether classes (e.g., *Cat*) can have instances or whether there are contradictions in the model or the data.

2.2 Ontology-Based Knowledge Graph

As explained in the previous subsection, an ontology-based knowledge graph is a knowledge graph where one or more ontologies are used to define which type of knowledge can be stored in the knowledge graph [19].

Note that sometimes, the instances of the concepts and relationships defined in the ontology (i.e., the "real" data) are also considered as part of the ontology, removing the strict separation between model and data. However, we follow the approach proposed by Chasseray et al. in [4], where the distinction between model and "data" is kept: a knowledge graph is composed of a *domain ontology* and an *instantiated ontology*[4]. The domain ontology is used to specify the organizational structure of the knowledge graph, and as the name indicates, the instantiated ontology is an instantiation of the domain ontology containing the actual instances (i.e., data).

Furthermore, Chasseray et al. [4] combine ontologies with the OMG's Model-Driven Engineering (MDE) approach. MDE defines four modelling levels: data level, model level, meta-model level, and meta-metamodel level [5]. Following this MDE approach, three levels are considered for ontology modelling by Chasseray et al.: an instantiated ontology (i.e., the data level) is defined by a domain ontology (i.e., the model level), which is an instantiation of an *upper ontology* (i.e., the meta-model level), which defines the concepts and relationships needed to define domain ontologies. Note that the meta-metamodel level of MDE is not used in this approach. We show this three-leveled structure for an ontology-based knowledge graph in Figure 1.

2.3 Abstract Reference Architecture

In [10], the Abstract Reference Architecture (ARA) is introduced to define the main processes and tasks required during the life cycle of knowledge graphs. ARA consists of three layers: *Knowledge Acquisition and Integration Layer*, *Knowledge Storage Layer*, and *Knowledge Consumption Layer*, which correspond to the three major tasks related to using a knowledge graph in organisations: construction, storage, and consumption. See Figure 2 for an illustration. Each of these tasks consists of sub-tasks:

1. ARA distinguishes four sub-tasks in the Knowledge Acquisition and Integration Layer: *Schema Development*, *Data Lifting*, *Data Annotation*, and

[4] or a set of these when the knowledge graph is defined by more than one ontology

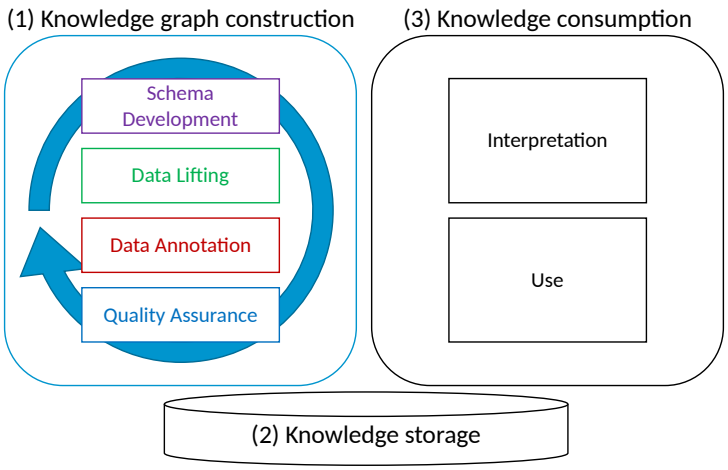

Figure 2.
The architecture of a knowledge graph project depicting the various processes and tasks in knowledge graph engineering, based on [10]

.

Quality Assurance. According to ARA, the first task is Schema Development (i.e., the development of the domain ontology). Data Lifting transforms raw data (e.g., stored in spreadsheets or classical databases) into semantic data. Data Annotation deals with linking and enriching the data with other relevant sources (e.g., other ontologies, knowledge graphs, vocabularies, or even classical databases), resulting in interlinked and contextualised semantic data. Quality assurance is about ensuring that the knowledge graph is of good quality, as one must rely on accurate data[5]. In Figure 2, one can see that the outcome of each of these activities informs the other. For instance, when the schema evolves, the transformation and integration of data into the knowledge graph may need to be updated.

2. The Knowledge Storage Layer deals with the storage of the knowledge graph. Two main architectural options are mentioned in [10]: (1) reusing existing data storage and providing mappings between the ontology and the data schemes of the existing storage, and (2) using a graph-based data store.

3. The Knowledge Consumption Layer provides tools for interested parties to access the knowledge. Examples include search and querying tools.

3. Related Work

3.1 Existing Tools for Supporting Hazard Assessment

Because the aim of the TOXIN knowledge system is to support the hazard assessment of new chemical compounds, we first discuss existing tools in this context.

[5] The difference between the two is that the former refers to mistakes (e.g., typos) and the latter to domain axioms not being respected (e.g., an instance cannot be an element of two disjoint classes).

There are several approaches to hazard assessment, and various tools have been developed to support this process. One such tool is the OECD QSAR Toolbox [11], which is a software application that aims to classify chemicals based on their structural characteristics and potential toxic mechanisms of interaction. In this way, this toolbox can provide some support for hazard assessment and risk assessment.

COSMOS NG [38] is another tool and provides a database of toxicity opinions about several chemicals that can be used for hazard assessment. COSMOS NG also provides in silico tools for analysing toxicity data and performing analyses such as category formation. Compared to TOXIN, the datasets used are different, but even for the common ingredients, only oral exposure studies are included in COSMOS NG and used for in silico predictions, such as DNA binding or physicochemical properties. TOXIN is currently based on existing in vivo-based hazard assessment with extra input from in silico testing (from the OECD QSAR toolbox). Therefore, it can be used to interpret interspecies differences (human in vitro versus laboratory animals in vivo) and to analyse how liver effects in vivo are related to the function or the structure of the ingredients.

VEGA HUB[6] includes several in silico tools for hazard and risk assessment. The VEGA software, which can be downloaded, can be used for hazard assessment using different models. It predicts several toxicity endpoints, such as liver toxicity, skin sensitisation, carcinogenicity, and endocrine disruptors. Another software available on VEGA HUB is Vermeer Cosmolife, which has the particularity to incorporate the context of use (called exposure scenario in toxicology) and, therefore, can be used for risk assessment.

There are other web applications available providing information for a more accurate and animal-free hazard assessment with a focus on one single endpoint, e.g., Vienna Livertox workspace[7], or on multiple endpoints, e.g., SApredictor[8] and ICE[9].

In a broader context, there is a growing interest in developing standardised vocabularies and ontologies that can be used to represent data about different toxicology tests efficiently [17]. The OpenTox initiative [30] aims to provide a framework for integrating and analysing diverse data sources using an ontology to improve the predictability of toxicology models and support decision-making in chemical safety assessment.

Tox21 [32] is a research program that seeks to identify new mechanisms of chemical activity in cells and use this information to prioritise untested chemicals for further evaluation and to develop more accurate predictive models of human response to toxic substances. One purpose is to provide a screening tool that would quickly identify potential hazards amongst a long list of potentially toxic compounds.

The work conducted in TOXIN has similarities with the previously mentioned initiatives. However, it is different in the sense that the main aim of our knowledge management system is to have a tool that can support toxicologists in the in vitro toxicity testing of cosmetic compounds by providing reliable and accurate existing information relevant for hazard assessment. In this way, it is more limited in its scope than the work conducted in Tox21. The OECD toolbox is used in TOXIN in the querying tool to provide a Hazard Evaluation Support

[6] https://www.vegahub.eu/portfolio-types/in-silico-models/

[7] https://livertox.univie.ac.at/

[8] http://www.sapredictor.cn/index.php

[9] https://ice.ntp.niehs.nih.gov/

System (HESS) for repeated dose toxicity studies using toxicological categories within the applicability domain of the tool. The in silico HESS predictions add value to the hazard assessment by pointing to or against a specific type of toxicity based on the structure of the chemicals.

3.2 Related Work on Linking Data Related to Toxicology

Numerous studies focused on linking data related to toxicology: [39] focuses on a comprehensive map of disease-symptom relations, [29] presents a method for representing the organisation of human cellular processes in a network and mapping diseases onto this network, and [25] provides an overview of existing work on integrating genes, pathways, and phenotypes to understand the effects of gene mutations better. While these studies are valuable contributions to the toxicology domain, they are specific. Moreover, their approaches may not necessarily apply to the broader task of integrating different available data sources.

4. Approach

As already explained in the introduction, our approach to knowledge management and the tools developed for it are based on the Abstract Reference Architecture (ARA).

Figure 3 provides an overview of the tasks defined for our knowledge management system. As described by ARA, the process starts with the Knowledge Graph Construction process that consists of four tasks: Ontology Development, Data Lifting, Data Annotation & Enrichment, and Quality Assurance.

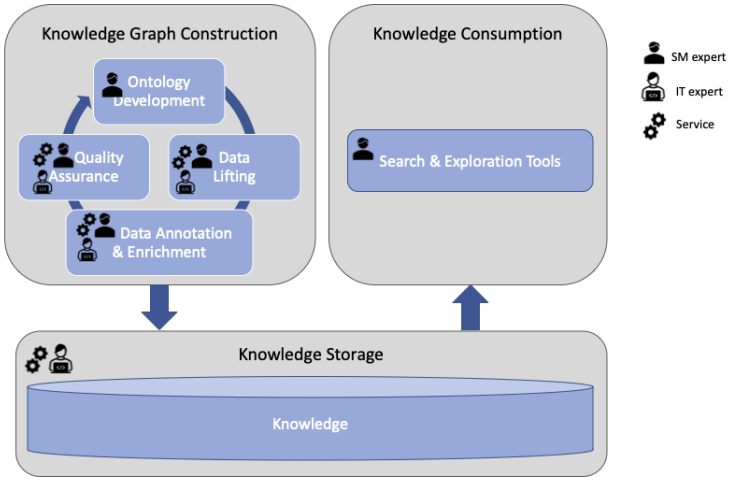

Figure 3.
TOXIN's knowledge Management approach adapted from [27]

Ontology Development is concerned with developing the knowledge graph's schema, which is done by means of an ontology, created with an ontology language. As already indicated, we follow [19] and [4] in the sense that a knowledge graph is the combination of a schema (i.e., an ontology) and an instantiation of that ontology (i.e., the data).

The term data lifting is used in ARA to denote the activities related to populating the knowledge graph. We have foreseen different ways to populate the knowledge graph in our knowledge management system. Subject matter experts can enter data manually. However, importing data from non-RDF sources, i.e., from spreadsheets, is also possible.

In ARA, Data Annotation comprises the activities of linking concepts and data with other relevant sources (e.g., other ontologies, knowledge graphs, vocabularies, or even classical databases), resulting in interlinked semantic data. We have called this task "Data Annotation & Enrichment" to better emphasise that data annotation also includes linking the data to existing sources. Some data annotation activities can be done manually by subject matter experts, but linking for the purpose of enrichment can be more complicated and may need the help of IT experts, although tools can also be developed for (semi-) automatic data annotation & enrichment.

Quality assurance is about ensuring that no mistakes are introduced into the knowledge graph and its ontology. This is a complex issue. In general, it concerns two aspects [34]: (1) the question of whether the knowledge graph has been built correctly, i.e., according to the requirements, and (2) the question of whether the right knowledge graph has been built, i.e., does the ontology correctly reflects the domain and does it contain correct data? Some quality control can be performed by subject matter experts, more technical aspects can be controlled by IT experts but manual quality control is time-intensive. Tools could be helpful in this respect.

The Knowledge Storage process deals with the storage of the knowledge graph. In this process, we decide how to store and service the data to the various applications built on top of this knowledge graph. Since we build a knowledge graph with semantic technologies (i.e., RDF as the graph data model and Semantic Web ontology languages), we use a triplestore. A triplestore is a name commonly given to RDF graph databases.

The Knowledge Consumption provides query, search & exploration functionality.

For the tools, we have followed an end-user approach as much as possible [27]. This is done because the direct use of semantic technology is often complicated for subject matter experts who are not technologically skilled. IT experts can be called in, but for specialised domains, such as the toxicology domain, it may take a long time before IT experts have familiarised themselves with the domain. In addition, an IT expert needs to stay available for the complete lifetime of the knowledge management system as knowledge systems tend to evolve over time, e.g., new properties, relationships and concepts may be needed, and new data must be added. Furthermore, during the Knowledge Consumption process, the assistance of IT experts may be needed, e.g., for the formulation of (new) queries or for the development of (new) reasoning support. To avoid being largely or completely dependent on IT experts, we developed end-user tools where possible. This means that subject matter experts who are not skilled in Computer Science (in our case, toxicologists) should be able to use these tools with some minimal training. Ideally, all tasks in ARA should be accessible to subject matter experts; however, this seems not feasible for some tasks. In particular, automatic Data Lifting, automatic Data Annotation & Enrichment, Quality Assurance, and Knowledge Storage may require the assistance of IT experts. For the ontology development and the manual data input, we used the jigsaw metaphor [13]. The purpose of using this metaphor was to hide the technicalities of the semantic technology.

In Figure 3, three different icons are used to indicate who can perform the tasks. If a subject matter (SM) expert can perform the task, the SM-expert icon is used, for instance, an SM expert can query the knowledge storage. The IT-expert icon indicates that an IT-expert should perform the task or at least be involved. The service icon is used to indicate that a (part of a) task can be done automatically. For instance, Data Lifting, Data Annotation & Enrichment and Quality Assurance have all three icons meaning that these tasks can be partially done by subject matter experts, partially by IT-experts, and partially automated.

5. TOXIN's Knowledge Management System

We present the tools developed for the knowledge management system of TOXIN in this section. Recall that the ultimate goal of this knowledge management system is to support toxicologists in the hazard assessment of new compounds through in vitro tests by bringing together multiple sources of toxicological information into a knowledge graph and allowing them to query and search the graph.

The first aim was to gather information about in vivo tests, described in documents dossiers, called Safety Evaluation Opinions, issued by the Scientific Committee on Consumer Safety (SCCS) about cosmetic ingredients in a knowledge graph. Each dossier contains information about experiments (also called tests) of a compound on laboratory animals. The information includes the quantity of the compound tested, how it was inoculated, the species on which the compound was tested, and so on. The dossiers also include information on the outcome of these tests, as well as the authors' opinions about the compound's toxicity. The data contained in these dossiers are stored in an ontology-based knowledge graph to provide more efficient access to this data for toxicologists.

The second aim is to enrich this knowledge graph with information from other relevant sources in the toxicology domain.

Because the development of the knowledge management system is based on ARA, we describe the tool support for the different tasks in ARA described in Section 4.

5.1 Ontology Development Support

While ontology development is one of the tasks within our approach, we want to emphasise that the purpose of this step is not to allow toxicologists to create a full-fledged domain ontology but to set up an ontology-based knowledge graph, where the role of the ontology is to define the organisational structure of the knowledge graph. The ontology that needs to be developed corresponds more with a vocabulary rather than a highly axiomatised ontology. Therefore, the expressiveness of the developed tool is kept limited with the aim to curb the learning curve for the toxicologists.

The tool developed for defining a domain ontology is a web-based application built on top of Apache Jena. The jigsaw metaphor [13], used in the tool to hide the technicalities of the semantic technology used, is implemented via the Google Blockly JavaScript library. The tool is not limited to the toxicology domain but can also be used for other domains to define the ontology for an ontology-based knowledge graph. It is described in detail in [27, 28]. Here, we provide a short description to show the reader how the tool is used to develop the ontology of TOXIN's knowledge management system.

A predefined jigsaw block is provided to allow toxicologists to define the domain concepts in the ontology by themselves. In this way, toxicologists can easily define the domain concepts for which they will maintain information in the knowledge graph. An example of such a domain concept is "Repeated Dose Toxicity", which is used extensively in Safety Evaluation Opinions. Figure 4 shows the web page for defining this domain concept. On the left-hand side of the page, one can see the predefined jigsaw block (i.e., top-level blue block "Domain Concept") that allows specifying the domain concept. Properties can be added by dragging and dropping the empty *Property* block from the *Custom Properties* tab in the menu at the left and filling the slots. Properties have a name, a value type, and an optional default value. In the example, the first property of the domain concept is named "OECD test nr" with value type "number" and no default value. Note that we also have the *Default Properties* tab in the left sidebar menu, which provides predefined blocks for recurring properties, such as the property "year". Re-using these recurring property blocks will improve the consistency of the ontology (as the recurring properties will always be defined in the same way) and speed up the specification process.

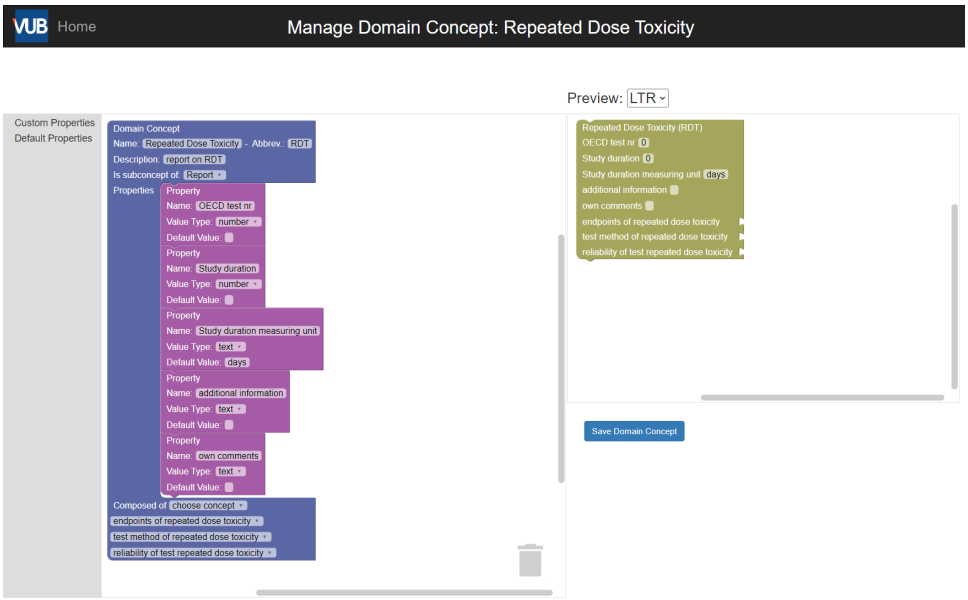

Figure 4.
Screenshot of the domain concept definition page

A domain concept can be composed of other domain concepts. In Figure 4, we see that the toxicologist specified that the *Repeated Dose Toxicity* concept is further composed of the *endpoints of repeated dose toxicity*, the *test method of repeated dose toxicity* and the *reliability of test repeated dose toxicity* concept. This is shown in the blue Domain Concept block under the "Composed of" field. The "Composed of" dropdown allows adding other composing concepts. The dropdown "Is subconcept of" is used to indicate that a concept is a sub-concept of another concept. This can be used to introduce hierarchies between concepts.

The ontology development process generally goes as follows (see Figure 5). The toxicologists use general jigsaw blocks to create the domain ontology, i.e., to define the domain concepts and their relationships needed for the knowledge graph. For each defined domain concept, a domain-specific jigsaw block will be generated. These generated jigsaw blocks can then, in turn, be used later on

by toxicologists to compose and fill the knowledge graph (explained in the next section (Section 5.2.1)). On the right in Figure 4, a preview of the generated jigsaw block for the domain concept defined on the left is given. In this case, the jigsaw block contains three puzzle connectors on the right side, one for each composing concept, and five slots, one for each property.

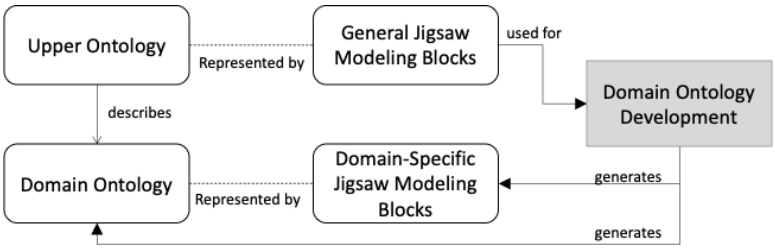

Figure 5.
Ontology Development Process Using the Jigsaw Metaphor (taken from [27])

5.2 Data Lifting Support

The current purpose of the Data Lifting task of the TOXIN knowledge management system is to capture, in the knowledge graph, information about in vivo tests described in Safety Evaluation Opinions issued by the Scientific Committee on Consumer Safety (SCCS) about cosmetic ingredients.

5.2.1 Manual Data Lifting

The first way to do this is by manually entering the information while reading a Safety Evaluation Opinion. For this, a toxicologist can use the domain-specific jigsaw blocks generated for the different domain concepts defined in the ontology. For example, when a toxicologist wants to enter the information from a particular Safety Evaluation Opinion document, they use the domain-specific jigsaw blocks created during the Ontology Development Process for Safety Evaluation Opinions (described in the previous section) to compose a so-called *dossier* (representing the opinion). They do so by connecting the relevant puzzle blocks and filling in the value fields in the blocks (see Figure 6 for a (partial) example dossier). The jigsaw blocks can only be composed in a restricted way (i.e., as defined in the ontology) and validation for data fields is provided. Figure 6, for example, shows how a toxicologist filled in the dossier on "Tetrabromophenol Blue" by adding the report block "Repeated Dose Toxicity" block (created as shown in Figure 4) and connecting it with other domain concepts to complete the report. After that, the toxicologist filled in the different property values corresponding to the information found in the safety evaluation opinion[10].

In order to save time during the manual entering of data and also to ensure that similar data is always entered in the same way, the tool allows the user to save block structures so that they can be reused in multiple dossiers as is. For example, the "Repeated Dose Toxicity" block shown in Figure 6 could be saved as a block structure. Then this block and all its sub-components (i.e., blocks attached to its right) will appear in the tab "saved block structures" depicted in Figure 7.

[10] https://health.ec.europa.eu/system/files/2021-08/sccs_o_232_0.pdf

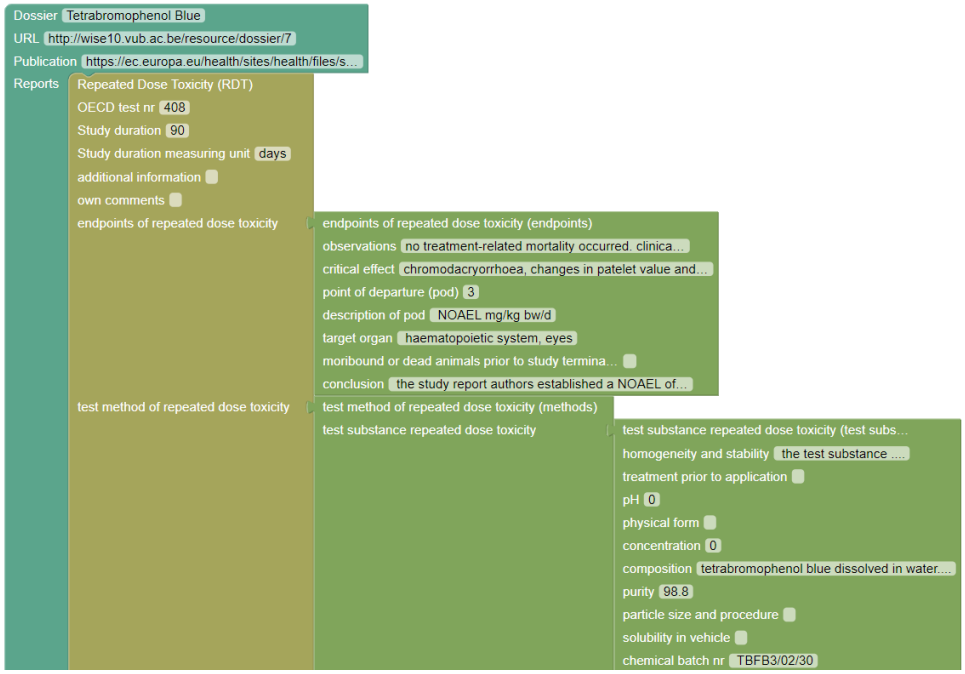

Figure 6.
Jigsaw Block composed for the dossier "Tetrabromophenol Blue"

These saved block structures can be dragged and dropped as a whole, which saves time when composing another dossier.

5.2.2 Automatic Data Lifting

In the past, the toxicologists of the TOXIN project used spreadsheets to structure and store information from Safety Evaluation Opinions. Because a considerable amount of time was spent on creating these spreadsheets, it was decided to develop a tool to import the data into the knowledge graph automatically.

We have used R2RML to transform the spreadsheets into RDF. R2RML is a W3C Recommendation for transforming relational data into RDF. Although spreadsheets are not relational databases, once stored as comma separated values (CSV) files, they can be considered as containing relational data (i.e., rows with attributes). R2RML engines (and dialects) such as RML [12] and R2RML-F [9] provide support for CSV files. We have chosen to adopt R2RML-F as this particular engine loads the CSV files into an in-memory relational database, which allows manipulating the data in the records with SQL prior to generating RDF.

The only requirement is that the CSV files are well-formed (e.g., the first row contains the names of attributes and no duplicates are allowed). This requirement is ensured by the people curating the spreadsheets. The advantage of using R2RML is that when new spreadsheets need to be transformed into RDF, one only needs to specify new R2RML mappings (which can be done by an IT expert). So far, data from 93 scientific opinions dealing with 88 different cosmetic ingredients, published between 2009 and 2019 by SCCS, were imported this way.

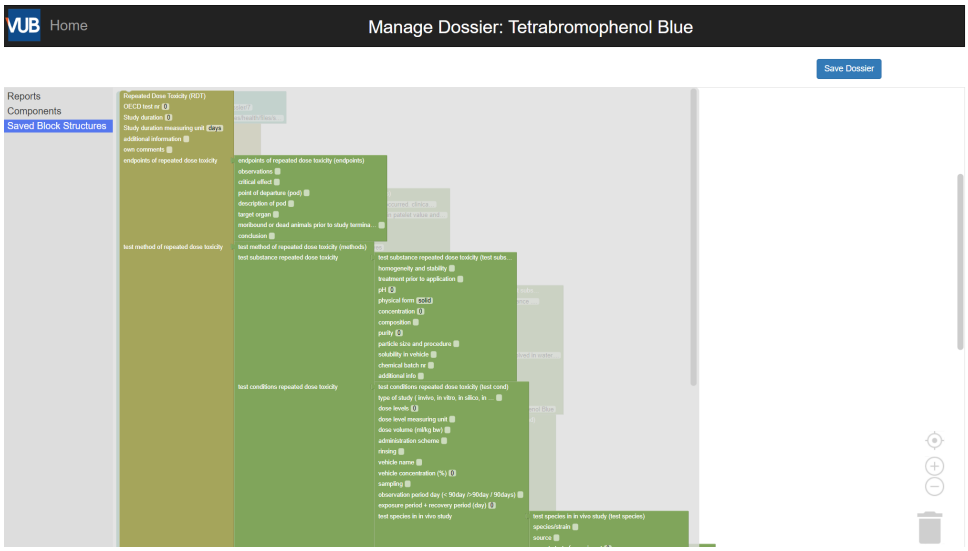

Figure 7.
Screenshot of the dossier creation page's saved block structure tab

5.3 Data Annotation & Enrichment Support

The current purpose of the task is to facilitate access to relevant toxicological data and provide answers to specific questions that the toxicologists formulated (see below). In a later stage, the annotations and enrichment could be used for AI-based reasoning.

First, simple manual data annotation is possible while defining the ontology and entering data. For a dossier, a link to the Safety Evaluation Opinion file for which the dossier is created should be given. For a domain concept, an IRI referring to a relevant source can be provided (see Figure 8). When references to other ontologies or RDF datasets are provided, users effectively create Linked Data[11].

Furthermore, a method for integrating multiple toxicological data sources and linking them with the TOXIN knowledge graph has been developed and applied [35]. The method starts with identifying the desired capabilities of the enriched knowledge management system by means of so-called Competency Questions [14]. Currently, the following competency questions have been formulated:

1. Knowing some adverse effects observed in a subject, what diseases or toxic processes may affect this subject?

2. Which biological processes or pathways are affected by a certain disease?

3. The functioning of what gene or protein is impaired by some toxic process?

These questions helped to select the potential sources for the enrichment.
In collaboration with the toxicologists, the following sources were selected: TXPO [37], OGG [18], Uniprot [6], Reactome [23], Kegg [21], and CTD [8]. We

[11] Linked Data is an initiative in which one published RDF data according to specific best practices that result in interconnected data stored on different servers; a Web of data.

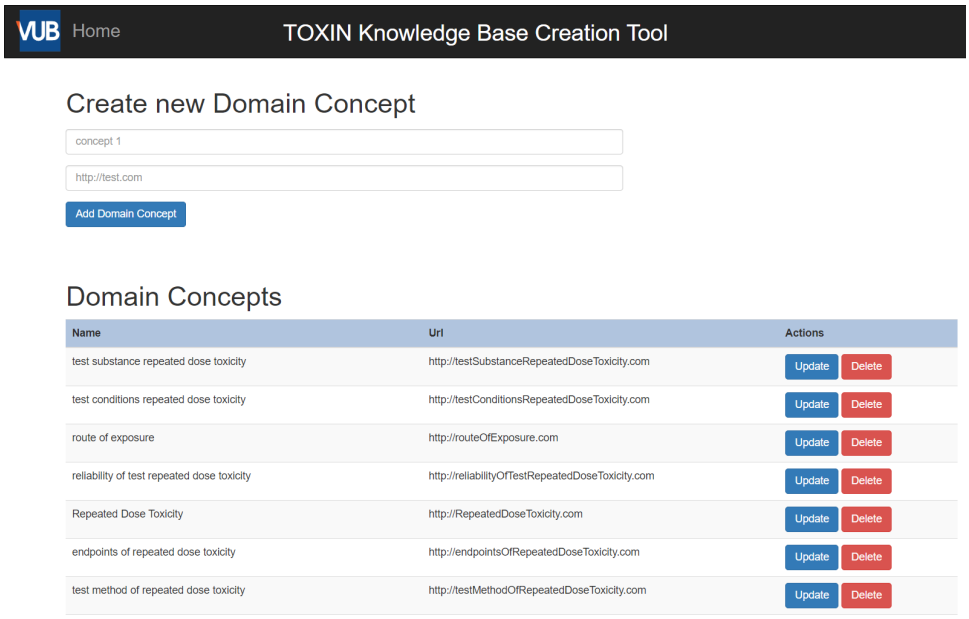

Figure 8.
Web page for domain concept creation and modification showing IRI's to related sources

furthermore identified the Gene Ontology (GO) [2] and Gene Ontology Causal Activity Modeling (GO-CAM) [31]. GO is a comprehensive and structured data source that classifies and describes genes and gene products based on their biological functions, cellular locations, and molecular activities. GO is a resource that many of the other initiatives reference in their data.

The ToXic Process Ontology (TXPO) is an ontology designed to represent causal relationships between toxic processes. Its purpose is to clarify the toxicological mechanisms from latent to toxic manifestations in order to help in drug development. Similar to TOXIN, their current focus is on the liver, as liver toxicity is the most frequent cause of the withdrawal of a new drug after testing.

TXPO contains a set of human genes that are imported from the Ontology of Genes and Genomes (OGG). This ontology focuses on offering classes and relationships to represent genes and genomes in different organisms. TXPO only imports the genes related to human organisms. TXPO also contains entities and relationships from the Gene Ontology (GO) [2]. The goal of this ontology is to represent the functions of genes. To create relationships between a toxic process and the genes or proteins affected by this process, GO is the perfect intermediate. Some links already exist in TXPO between a toxic process and the natural processes that it affects. For this, GO annotations are used that represent the link between a GO term, i.e., a biological role and a gene product (gene or protein) that assumes this role in the organism. Each annotation is associated with a proof, which has a weight representing the confidence in the annotation. We considered the integration of annotations from two sources: OGG and from the gene ontology resource[12]. This resource regroups annotations made for a large variety of species and from several different sources. We chose to integrate only human gene products from UniProt.

[12] http://geneontology.org

GO-CAM is developed by the GO community and is a modelling approach that builds upon GO. GO-CAM introduces models to connect GO-annotations to represent causal relationships between gene products and biological processes, providing a more detailed and explicit representation of molecular events. The GO-CAM approach allows for the representation of specific regulatory interactions and signalling pathways, enabling researchers to analyse and interpret biological data in a more precise and context-dependent manner. For example, a GO-CAM model could show that the hyper-function of a biological process positively regulates another process. These relationships allow toxicologists to track a toxic effect from its starting point to all the other elements that are indirectly affected. GO-CAM is thus an interesting resource to represent the different relationships between biological processes.

Reactome and Kegg are two well-known pathways repositories. To integrate pathways from both repositories, we used associations from the Comparative Toxicogenomics Database (CTD). CTD combines biological data by manually curating and linking information from published literature. It offers several files containing the relationships between different biological entities, and for TOXIN, we used the disease pathway association file.

To perform the integration of (parts of) the different sources, we defined an upper structure, which is based on the TXPO ontology that has been developed to be used as a structure to build ontology-based knowledge graphs. While TXPO offers a set of classes and axioms, i.e., an ontology, we wanted to use the identifiers of these classes as individuals. Luckily, using semantic technology allows us to reuse these identifiers in a different ontology – our upper structure.

For the integration, we have chosen to maintain the original IRIs to uniquely identify the resources they describe and maintain the authority of the original sources. However, for the additional links and entities created for the integration, we have designed our own named graphs to describe them. This choice allows us to clearly distinguish between original resources and those added as part of the integration. This architecture is illustrated in Figure 9.

the decision to retain the authoritative resources' IRI whenever possible allows access to additional information about entities in the knowledge graph from the original data source. An example of this can be seen in Figure 10. On the left of the figure, we can see three entities of the enriched knowledge graph represented with Ontodia (see Section 5.6 for more details about Ontodia). When the IRIs of these entities are visited, e.g., by clicking on them and loading the resource in a Web browser, the resources on the right of the image appear. This illustrates that we can access the source data through the IRI representing an entity. In other words, not only are the resources conceptually integrated, we even have "integrated" systems thanks to the distributed graph data model offered by RDF.

Concerning the linking of the TOXIN knowledge graph and the enriched TXPO, two kinds of links can be made. Firstly, direct links can be made between an effect observed in a dossier and the same effect present in the enriched TXPO. Secondly, from the observations and the conclusions in a dossier, a domain expert (who has the knowledge to infer the toxic effects from the observations) can infer "indirect" links between the dossier and the toxic effects or diseases that affect the test animal. There is no absolute knowledge about how to link some observations and a toxic effect, and errors can occur. Therefore, the links are stored in different named graphs. Each graph corresponds to an individual, a group, or a particular knowledge responsible for defining the links that it contains. In this way, it is possible to query only some of these links depending on who made them and on what ground. This type of linking is a manual

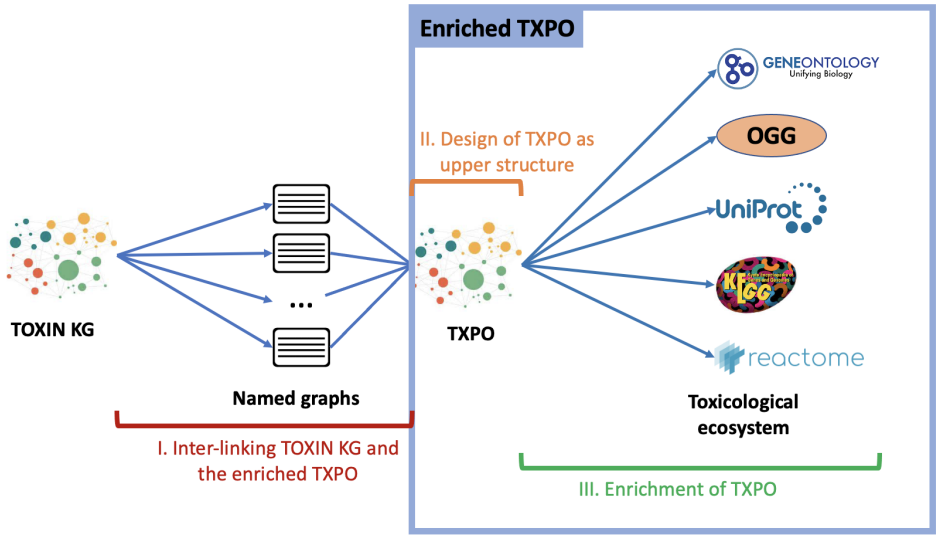

Figure 9.
Schema of TEKG with the relations between TOXIN KG and the enriched TXPO.

process. Another approach is to create these links automatically. For the direct links, it is straightforward. The task is to infer the relations `owl:sameAs` between the TOXIN knowledge graph and the enriched TXPO, and tools exist to find these relations, such as Silk [33] and Alignment API [7]. Concerning the "indirect" links, some rule-based mechanisms could be put in place. This is subject to future work.

5.4 Quality Assurance Support

Currently, the support for quality assurance is limited in our tool to manual quality checking and the validation of the structure of the data in the knowledge graph by means of SHACL shapes. With SHACL [22], a W3C Recommendation, one can validate RDF graphs, i.e., one can validate the structure of triples in a Closed World Setting. SHACL provides a set of "core" constructs for declaring rules (value- and data type checking, cardinality, value ranges, comparisons,... which can be combined with a set of logical operators). Validating the knowledge graph with SHACL is especially valuable in the case of the automatic import of the spreadsheets' data. A SHACL shape is a subset of an RDF graph, which can be declared, and to which one can add constraints.

Because we cannot expect that toxicologists can formulate constraints in SHACL, we looked for an approach by which the SHACL-shapes and constraints can be generated. How this is achieved is explained in [27].

5.5 Knowledge Storage

While quite a few triplestores are available (both free, commercial, and free for research purposes), we have adopted Apache Fuseki[13] for managing the storage of triples, named graphs, and SPARQL endpoints. The ontology is stored in one named graph. The data which has been lifted are stored in another. Data that has

[13] https://jena.apache.org/documentation/fuseki2/

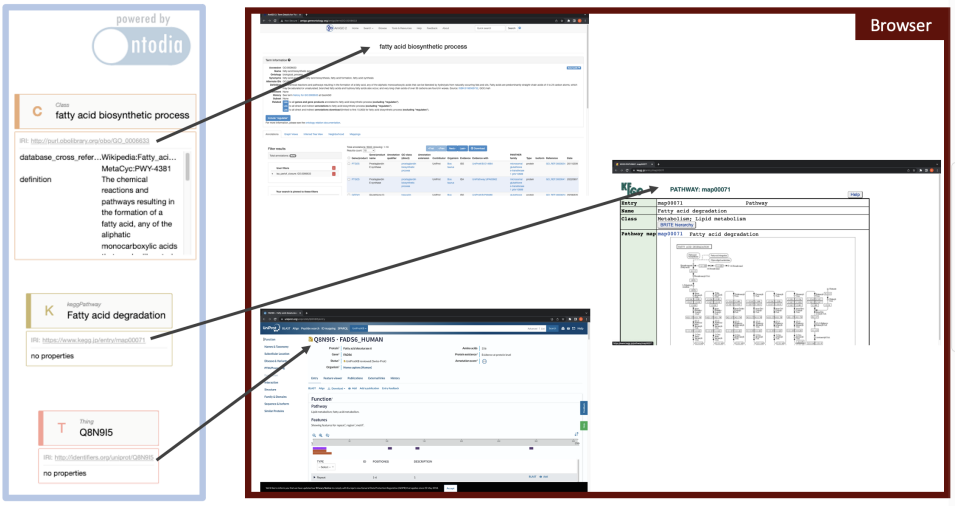

Figure 10.
Illustration of the possibility of accessing the external sources from which data was integrated through the IRIs.

been integrated from other sources with the enrichment scripts are also stored in dedicated named graphs. This allows us to separate the ontology and the data, and to ensure that data from various sources (and possibly with different interpretations) are not mixed up.

5.6 Knowledge Consumption Support

Knowledge consumption concerns the access and use of the knowledge graph by end users. The primary use of TOXIN's knowledge graph is to provide liver-specific toxicological information. For this purpose, a web tool has been developed that allows one to search TOXIN's knowledge graph from two perspectives; from the perspective of chemical compounds and from the perspective of health effects.

In the first perspective, one can search the knowledge graph for dossiers (i.e., Safety Evaluation Opinions) about specific compounds by means of the CAS number, INCI name, or SMILES string of the compound. For example, HC Yellow n° 2 can be found using the INCI name "HC Yellow n° 2", the CAS no 4926-55-0, or the SMILES string "C1=CC=C(C(=C1)NCCO)[N+](=O)[O-]". As a result, a summary of information on the searched compound is given; an interface to the OECD QSAR toolbox is provided to select information from this toolbox for the compound, e.g., the Hazard Evaluation Support System (HESS) in silico prediction; and a list of the dossiers (i.e., opinions) that deal with the compound. The dossiers are grouped by the endpoints (acute toxicity, repeated dose toxicity, skin sensitisation, mutagenicity, and carcinogenicity). Clicking on such a dossier will show the information stored for the dossier, such as the OECD guideline number, dose levels, and GLP-compliance. The UI for this perspective is shown in Figure 11.

In the second perspective, the user can search for compounds with dossiers mentioning a specific toxicological outcome by selecting a health endpoint. For the moment, the endpoints acute toxicity, repeated dose toxicity, and toxicokinetics are supported by the tool. As a result, in the "compound view" all relevant dossiers ordered by compound are listed (see Figure 12 below); the "Opinions view" lists the relevant dossiers (i.e., opinions) directly (similar to the

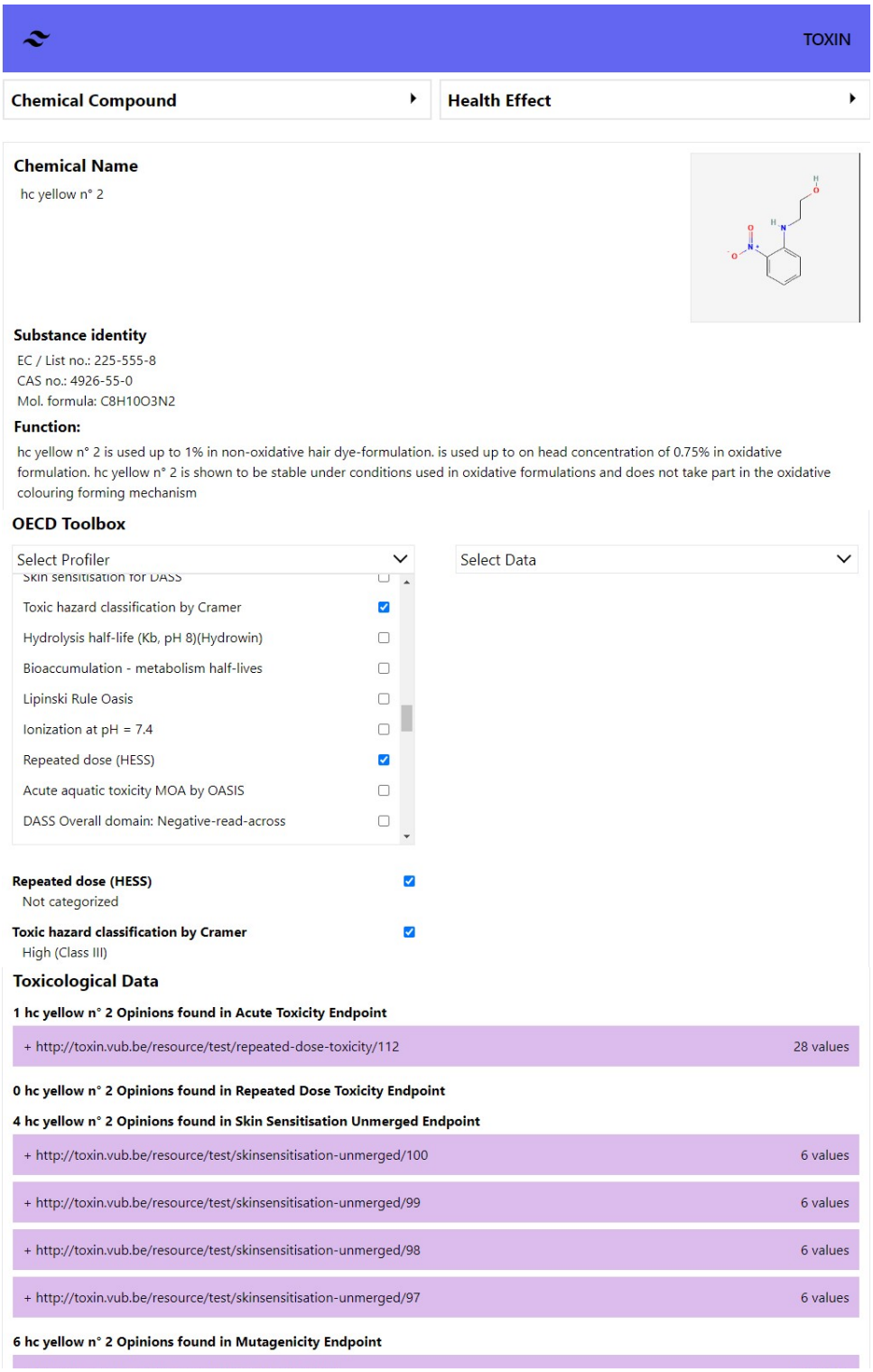

Figure 11.
Partial screenshot of the search result of the chemical compound "HC Yellow n° 2"

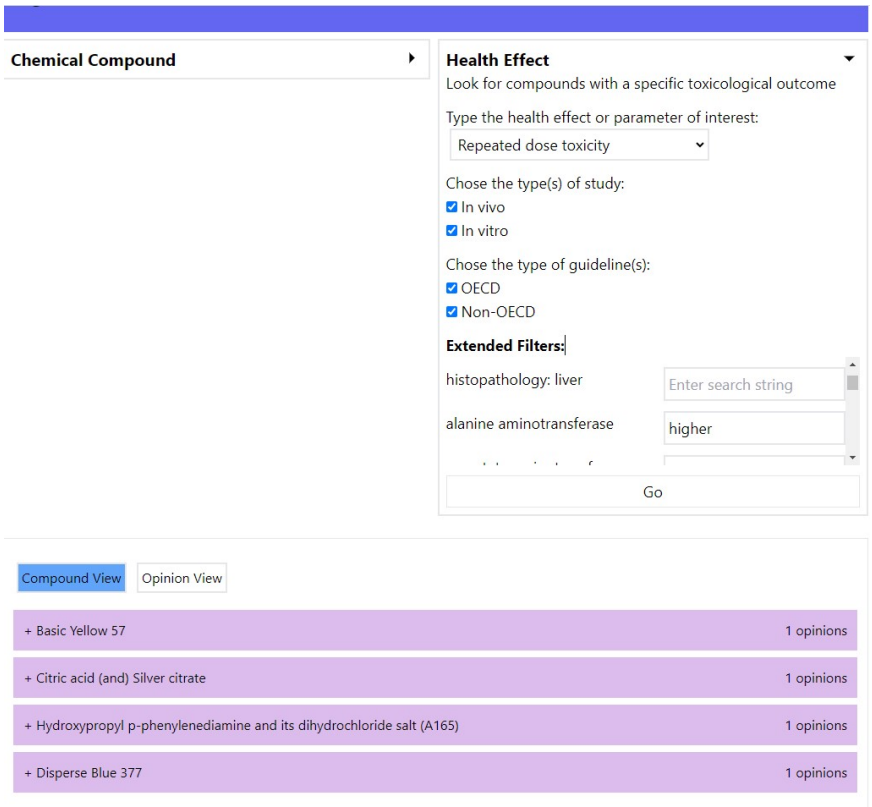

Figure 12.
Screenshot of the search by Health Effect: "Repeated Dose Toxicity"

bottom part of Figure 11). The user filter the output on the type of test (in vivo and/or in vitro) and filter by whether OECD guidelines were used. Moreover, extended filters can search for values in certain parts coming from the opinions. The UI for this perspective is shown in Figure 12 for the "Repeated dose toxicity" health effect.

A dedicated tool for querying and searching the enriched Knowledge graph described in Section 5.3 has yet to be developed, but an existing tool, Ontodia [24], has been used to show that the competency questions formulated by the toxicologists can be answered. Ontodia is a tool allowing one to visually explore a triple-based knowledge graph using diagrams and faceted browsing, providing a way to navigate knowledge graphs (note that there exist other tools that could also be used for this purpose, such as WebVOWL[14]). In Figure 13, we illustrate the answer to the competency question "Knowing some adverse effects observed in a subject, what diseases or toxic processes may affect this subject?" Figure 13 presents different adverse effects that could be observed during a toxicological test, such as "Increasing blood ALP concentration." These adverse effects are linked to toxic courses or diseases with the predicates 'has part' and 'has context'. These predicates allow one to query all the toxic courses for which an adverse effect could be observed. Moreover, relationships between toxic processes are represented. However, it not only allows for finding toxic processes related to an adverse outcome, but it also allows for the examination

[14] http://vowl.visualdataweb.org/webvowl.html

of the relationships between adverse effects, toxic effects, toxic courses, and diseases.

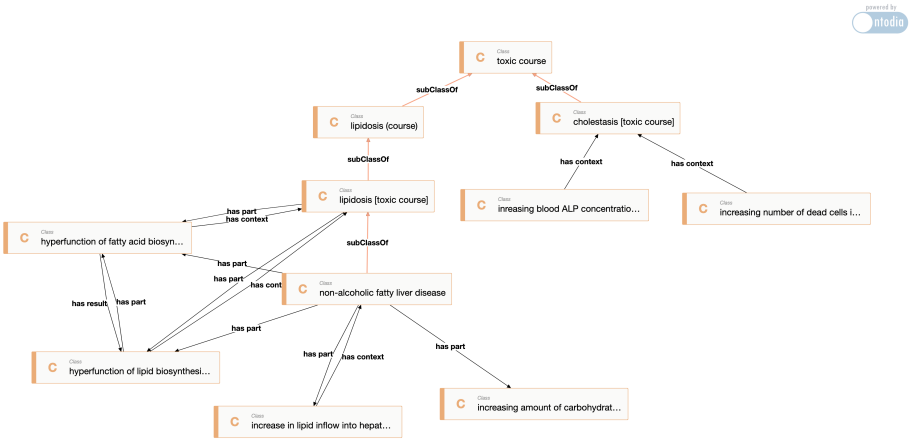

Figure 13.
Illustration of how the enriched knowledge graph can be used to answer the competency question "Knowing some adverse effects observed in a subject, what diseases or toxic processes may affect this subject?".

6. Conclusions

In this chapter, we presented the knowledge management system developed in the TOXIN project to support toxicologists in the animal-free hazard assessment of new cosmetic compounds. The developed knowledge management system is based on the tasks defined for creating, maintaining, and exploiting knowledge graphs in the Abstract Reference Architecture (ARA), which defines the main processes and tasks required during the life cycle of knowledge graphs. The knowledge system is developed as a knowledge graph by means of semantic technology, and tools were developed for the following ARA tasks: the Ontology Development, Data Lifting, Data Annotation & Enrichment, Quality Assurance, and Knowledge Consumption.

In addition, special attention has been paid to the fact that subject matter experts, i.e., toxicologists, should be able to perform most of the tasks by themselves. Where possible, the tools shield the non-IT users as much as possible from the technical aspects of the technology used and they are able to use these tools with some minimal training. For the tasks for which this was impossible, automatic tools were developed where possible.

The different developed tools have been described. Two tools are provided for defining the knowledge graph and populating it with data. First, we have developed an end-user tool, based on the jigsaw metaphor, that allows the manual definition and population of a knowledge graph by toxicologists. In addition, we developed a tool to automatically import toxicity data previously collected by the subject matter experts in spreadsheets into the knowledge graph. A search and query tool has been developed that allows toxicologists to explore and search in the knowledge graph by means of a simple web-based user interface and without the need to use the technical query language SPARQL. Furthermore, we integrated multiple data sources from the field of toxicology

into the TOXIN knowledge graph to further support the hazard assessment of new compounds. For exploring this enriched knowledge graph, an existing tool, i.e., Ontodia, is currently used. Finally, we investigated an approach to deal with some aspects of the quality assurance of the knowledge graph.

For future work, we are considering the use of Natural Language Processing (NLP) techniques for different tasks. We noticed that more than searching in the knowledge graph by only using text matching is needed, especially when one wants to search for knowledge from different sources. For example, if a toxicologist wants to know which chemical compounds raise the value of a particular observation. They might search using the term "increase", however, given that different terminology is used to describe this effect, such as "growth" or "change", they will miss a few search results. In addition, sometimes results of tests are entered all together as one large text block rather than as separate results. Sometimes a value was even considered as a concept or the other way around when entering the knowledge, which means that searching should not be limited to the values of properties but also the names of concepts and properties should be considered.

Last but not least, other future work concerns the addition of a data provenance layer [16] to our knowledge graph, which will allow to trace who has added which information and when.

Acknowledgments

The TOXIN project is financially supported by Vrije Universiteit Brussel under Grant IRP19.

Some funding came from Cosmetics Europe and the European Chemical Industry Council (CEFIC).

The research of Audrey Sanctorum has been funded by an FWO Postdoc Fellowship (1276721N) of the Research Foundation Flanders.

Author details

Audrey Sanctorum[1*], Jan Maushagen[1,2], Sara Sepehri[2], Guillaume Vrijens[3], Joery De Kock[2], Tamara Vanhaecke[2], Olga De Troyer[1] and Christophe Debruyne[3]

1 WISE Lab, Vrije Universiteit Brussel, Brussels, Belgium

2 Research Group of *In Vitro* Toxicology and Dermato-Cosmetology (IVTD), Vrije Universiteit Brussel, Brussels, Belgium

3 Montefiore Institute, University of Liège, Liège, Belgium

*Address all correspondence to: Audrey.Sanctorum@vub.be

References

[1] OWL 2 web ontology language document overview (second edition). W3C recommendation, W3C (Dec 2012), https://www.w3.org/TR/2012/REC-owl2-overview-20121211/

[2] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nature genetics **25**, 25–29 (2000). https://doi.org/10.1038/75556

[3] Ball, N., Bars, R., Botham, P.A., Cuciureanu, A., Cronin, M.T., Doe, J.E., Dudzina, T., Gant, T.W., Leist, M., van Ravenzwaay, B.: A framework for chemical safety assessment incorporating new approach methodologies within reach. Archives of Toxicology **96**(3), 743–766 (2022)

[4] Chasseray, Y., Barthe-Delanoë, A.M., Négny, S., Le Lann, J.M.: A Generic Metamodel for Data Extraction and Generic Ontology Population. Journal of Information Science (2021). https://doi.org/10.1177/0165551521989641

[5] Colomb, R.M., Raymond, K., Hart, L., Emery, P., Welty, C., Xie, G.T., Kendall, E.F.: The Object Management Group Ontology Definition Metamodel. In: Ontologies for Software Engineering and Software Technology, pp. 217–247. Springer (2006). https://doi.org/10.1007/3-540-34518-3_8

[6] Consortium, T.U.: UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research **49**(D1), D480–D489 (2021). https://doi.org/10.1093/nar/gkaa1100

[7] David, J., Euzenat, J., Scharffe, F., Trojahn, C.: The Alignment API 4.0. Semantic Web **2**, 3–10 (2011). https://doi.org/10.3233/SW-2011-0028

[8] Davis, A.P., Murphy, C.G., Saraceni-Richards, C.A., Rosenstein, M.C., Wiegers, T.C., Mattingly, C.J.: Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. Nucleic Acids Research **37**, D786–D792 (2022)

[9] Debruyne, C., O'Sullivan, D.: R2RML-F: towards sharing and executing domain logic in R2RML mappings. In: Auer, S., Berners-Lee, T., Bizer, C., Heath, T. (eds.) Proceedings of the Workshop on Linked Data on the Web, LDOW 2016, co-located with 25th International World Wide Web Conference (WWW 2016). CEUR Workshop Proceedings, vol. 1593. CEUR-WS.org (2016)

[10] Denaux, R., Ren, Y., Villazón-Terrazas, B., Alexopoulos, P., Faraotti, A., Wu, H.: Knowledge architecture for organisations. In: Pan, J.Z., Vetere, G., Gómez-Pérez, J.M., Wu, H. (eds.) Exploiting Linked Data and Knowledge Graphs in Large Organisations, pp. 57–84. Springer (2017). https://doi.org/10.1007/978-3-319-45654-6_3

[11] Dimitrov, S., Diderich, R., Sobanski, T., Pavlov, T., Chankov, G., Chapkanov, A., Karakolev, Y., Temelkov, S., Vasilev, R., Gerova, K., Kuseva, C., Todorova, N., Mehmed, A., Rasenberg, M., Mekenyan, O.: QSAR Toolbox - workflow and major functionalities. SAR and QSAR in Environmental Research **27**, 203–219 (2016). https://doi.org/10.1080/1062936X.2015.1136680

[12] Dimou, A., Sande, M.V., Colpaert, P., Verborgh, R., Mannens, E., de Walle, R.V.: RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In: Bizer, C., Heath, T., Auer, S., Berners-Lee, T. (eds.) Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014. CEUR Workshop Proceedings, vol. 1184. CEUR-WS.org (2014)

[13] Gozzi, R.: The Jigsaw Puzzle as a Metaphor for Knowledge. ETC: A Review of General Semantics **53**(4), 447–451 (1996)

[14] Grüninger, M., Fox, M.S.: The Role of Competency Questions in Enterprise Engineering, pp. 22–31. Springer US, Boston, MA (1995). https://doi.org/10.1007/978-0-387-34847-6_3

[15] Guha, R., Brickley, D.: RDF schema 1.1. W3C Recommendation, W3C (2014), https://www.w3.org/TR/2014/REC-rdf-schema-20140225/

[16] Gupta, A.: Data Provenance. In: Liu, L., Özsu, M.T. (eds.) Encyclopedia of Database Systems, p. 608. Springer US (2009). https://doi.org/10.1007/978-0-387-39940-9_1305

[17] Hardy, B., Apic, G., Carthew, P., Clark, D., Cook, D., Dix, I., Escher, S., Hastings, J., Heard, D., Jeliazkova, N., Judson, P., Matis-Mitchell, S., Mitic Potkrajac, D., Myatt, G., Shah, I., Spjuth, O., Tcheremenskaia, O., Toldo, L., Watson, D., Yang, C.: Toxicology ontology perspectives. ALTEX. Alternatives zu Tierexperimenten **29**(2), 139–156 (2012). https://doi.org/10.14573/altex.2012.2.139

[18] He, Y., Liu, Y., Zhao, B.: OGG: a Biological Ontology for Representing Genes and Genomes in Specific Organisms. ICBO pp. 13–20 (2014)

[19] Hepp, M.: Ontologies: State of the art, business potential, and grand challenges. In: Hepp, M., Leenheer, P.D., de Moor, A., Sure, Y. (eds.) Ontology Management, Semantic Web, Semantic Web Services, and Business Applications, Semantic Web and Beyond: Computing for Human Experience, vol. 7, pp. 3–22. Springer (2008). https://doi.org/10.1007/978-0-387-69900-4_1

[20] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutiérrez, C., Kirrane, S., Labra G., J.E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.C., Polleres, A., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., Zimmermann, A.: Knowledge Graphs. No. 22 in Synthesis Lectures on Data, Semantics, and Knowledge, Morgan & Claypool (2021). https://doi.org/10.2200/S01125ED1V01Y202109DSK022

[21] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K.: KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research **45**(D1), D353–D361 (2016). https://doi.org/10.1093/nar/gkw1092

[22] Knublauch, H., Kontokostas, D.: Shapes Constraint Language (SHACL) (2017), https://www.w3.org/TR/shacl/

[23] Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., D'Eustachio, P.: Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Research **37**, D619–D622 (2008). https://doi.org/10.1093/nar/gkn863

[24] Mouromtsev, D., Pavlov, D.S., Emelyanov, Y., Morozov, A.V., Razdyakonov, D.S., Galkin, M.: The simple web-based tool for visualization and sharing of semantic data and ontologies. In: Villata, S., Pan, J.Z., Dragoni, M. (eds.) Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located with the 14th International Semantic Web Conference (ISWC-2015), Bethlehem, PA, USA, October 11, 2015. CEUR Workshop Proceedings, vol. 1486. CEUR-WS.org (2015)

[25] Papatheodorou, I., Oellrich, A., Smedley, D.: Linking gene expression to phenotypes via pathway information. Journal of Biomedical Semantics **6**(1), 17 (2015). https://doi.org/10.1186/s13326-015-0013-5

[26] Rajpoot, K., Desai, N., Koppisetti, H., Tekade, M., Sharma, M., Behera, S., Tekade, R.: Chapter 14 - In silico methods for the prediction of drug toxicity. In: Pharmacokinetics and Toxicokinetic Considerations, Advances in Pharmaceutical Product Development and Research, vol. 2, pp. 357–383. Academic Press (2022). https://doi.org/10.1016/B978-0-323-98367-9.00012-3

[27] Sanctorum, A., Riggio, J., Maushagen, J., Sepehri, S., Arnesdotter, E., Delagrange, M., De Kock, J., Vanhaecke, T., Debruyne, C., De Troyer, O.: End-user engineering of ontology-based knowledge bases. Behaviour & Information Technology **41**(9), 1811–1829 (2022). https://doi.org/10.1080/0144929X.2022.2092032

[28] Sanctorum, A., Riggio, J., Sepehri, S., Arnesdotter, E., Vanhaecke, T., De Troyer, O.: A Jigsaw-Based End-User Tool for the Development of Ontology-Based Knowledge Bases. In: Fogli, D., Tetteroo, D., Barricelli, B.R., Borsci, S., Markopoulos, P., Papadopoulos, G.A. (eds.) Proceedings of IS-EUD 2021,8th International Symposium on End-User Development. Lecture Notes in Computer Science, vol. 12724, pp. 169–184. Springer (July 2021). https://doi.org/10.1007/978-3-030-79840-6_11

[29] Stoney, R., Robertson, D., Nenadic, G., Schwartz, J.M.: Mapping biological process relationships and disease perturbations within a pathway network. npj Systems Biology and Applications **4**(1), 22 (2018). https://doi.org/10.1038/s41540-018-0055-2

[30] Tcheremenskaia, O., Benigni, R., Nikolova, I., Jeliazkova, N., Escher, S., Batke, M., Baier, T., Poroikov, V., Lagunin, A., Rautenberg, M., Hardy, B.: OpenTox predictive toxicology framework: Toxicological ontology and semantic media wiki-based OpenToxipedia. Journal of biomedical semantics **3 Suppl 1**, S7 (2012). https://doi.org/10.1186/2041-1480-3-S1-S7

[31] Thomas, P., Hill, D., Mi, H., Osumi-Sutherland, D., Auken, K., Carbon, S., Balhoff, J., Albou, L.P., Good, B., Gaudet, P., Lewis, S., Mungall, C.: Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nature Genetics **51**, 1429–1433 (2019). https://doi.org/10.1038/s41588-019-0500-1

[32] Thomas, R., Paules, R., Simeonov, A., Fitzpatrick, S., Crofton, K., Casey, W., Mendrick, D.: The US Federal Tox21 Program: A strategic and operational plan for continued leadership. ALTEX - Alternatives to animal experimentation **35**(2), 163–168 (2018). https://doi.org/10.14573/altex.1803011

[33] Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk - A link discovery framework for the web of data. In: Proceedings of the WWW2009 Workshop on Linked Data on the Web, LDOW 2009, Madrid, Spain, April 20, 2009. CEUR Workshop Proceedings, vol. 538. CEUR-WS.org (2009), http://ceur-ws.org/Vol-538/ldow2009_paper13.pdf

[34] Vrandecic, D.: Ontology evaluation. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 293–313. International Handbooks on Information Systems, Springer (2009). https://doi.org/10.1007/978-3-540-92673-3_13

[35] Vrijens, G.: Knowledge Graph Construction to Facilitate Chemical Compound Hazard Assessment in the TOXIN Project. Master's thesis, School of Engineering and Computer Science - University of Liège (2023)

[36] Wood, D., Cyganiak, R., Lanthaler, M.: RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation, W3C (2014), https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

[37] Yamagata, Y., Yamada, H.: Ontological approach to the knowledge systematization of a toxic process and toxic course representation framework for early drug risk management. Scientific Reports **10**, 14581 (2020). https://doi.org/10.1038/s41598-020-71370-7

[38] Yang, C., Cronin, M., Arvidson, K., Bienfait, B., Enoch, S., Heldreth, B., Hobocienski, B., Muldoon-Jacobs, K., Lan, Y., Madden, J., Magdziarz, T., Marusczyk, J., Mostrag, A., Nelms, M., Neagu, D., Przybylak, K., Rathman, J., Park, J., Richarz, A.N., Richard, A., Ribeiro, J., Sacher, O., Schwab, C., Vitcheva, V., Volarath, P., Worth, A.: COSMOS next generation – A public knowledge base leveraging chemical and biological data to support the regulatory assessment of chemicals. Computational Toxicology **19** (2021). https://doi.org/10.1016/j.comtox.2021.100175

[39] Zhou, X., Menche, J., Barabási, A.L., Sharma, A.: Human symptoms–disease network. Nature Communications **5**(1) (2014). https://doi.org/10.1038/ncomms5212