Master thesis : Knowledge Graph Construction to Facilitate Chemical Compound Hazard Assessment in the TOXIN Project

Auteur : Vrijens, Guillaume
Promoteur(s) : Debruyne, Christophe
Faculté : Faculté des Sciences appliquées
Diplôme : Master : ingénieur civil en science des données, à finalité spécialisée
Année académique : 2022-2023
URI/URL : http://hdl.handle.net/2268.2/16763

UNIVERSITY OF LIÈGE - SCHOOL OF ENGINEERING AND COMPUTER SCIENCE

Knowledge Graph Construction to Facilitate Chemical Compound Hazard Assessment in the TOXIN Project

Master's thesis completed in order to obtain the degree of
Master of Science in Data Science and Engineering
by

Guillaume Vrijens

Supervisor:
DEBRUYNE Christophe

Jury:
DEBRUYNE Christophe
FONTAINE Pascal
LOUPPE Gilles

Academic year 2022-2023

Abstract

This master thesis presents a method for integrating multiple data sources from the field of toxicology into a knowledge graph and linking it with the TOXIN knowledge graph to facilitate the hazard assessment of new compounds. The proposed method uses a hybrid approach, combining an ontology and Linked Data to capture the granularity of the toxicological domain and provide a consistent representation while maintaining the flexibility of Linked Data. The ontology used in the method is the ToXic Process Ontology (TXPO), which offers a structured and reliable representation of the relationships between toxicological processes. The method also incorporates the use of named graphs and provenance information to store different opinions on data and track the integration of different sources. The feasibility and utility of the proposed method for building the knowledge graph are demonstrated through the development of a prototype, the TOXIN enriched knowledge graph (TEKG). Finally, this project illustrates the potential value and usefulness of a knowledge graph such as TEKG for improving access to relevant information, offering a satisfactory representation of the toxicological domain and supporting domain-specific tagging mechanisms.

Acknowledgements

I would like to express my gratitude to my supervisor, Christophe Debruyne, for his guidance, support, and expertise throughout this project. His commitment to the development of this thesis has been instrumental in its successful completion.

I would also like to thank the team of the TOXIN project, particularly Sara Sepehri, for their valuable input, guidance, and review of my work. Their expertise in toxicology and computer science has provided valuable insights and perspectives throughout the course of this project.

Finally, I would like to express my appreciation to the Web & Information Systems Engineering (WISE) Lab at the Vrije Universiteit Brussel (VUB) for providing me with the opportunity to work with them on this project. It has been a truly enriching experience and I am grateful for their support and collaboration.

Table of contents

1 Introduction **1**

1.1 Context . . . 1

1.1.1 Toxicity Testing . . . 1

1.1.2 Toxicological Ecosystem . . . 2

1.1.3 3R Principle . . . 3

1.1.4 TOXIN Project . . . 3

1.2 Problem Statement and Challenges . . . 4

1.3 Methodology and outline . . . 5

2 Background **7**

2.1 Data Modeling . . . 7

2.1.1 Data Graphs . . . 7

2.1.2 Ontologies and Vocabularies . . . 8

2.1.3 Provenance Ontology . . . 10

2.2 Knowledge Graph . . . 10

2.3 Knowledge Graph System . . . 11

2.3.1 Architecture . . . 11

2.3.2 Knowledge Construction and Maintenance . . . 13

2.4 Summary . . . 15

3 Related Work **16**

3.1 Existing Tools for Hazard Assessment . . . 16

3.2 Related Work on Knowledge Graph Enrichment and Data Integration . . . 18

3.3 Summary . . . 19

4 Method and Instantiation **20**

4.1 TOXIN Knowledge Graph . . . 20

4.2 Methodology . . . 22

4.2.1 Specification . . . 22
4.2.2 Modelling and Data Lifting . . . 23
4.2.3 Data Publication . . . 23
4.2.4 Data Curation . . . 24
4.3 The ToXic Process Ontology (TXPO) as Upper Structure . . . 24
4.3.1 Presentation of TXPO . . . 24
4.3.2 Modifications of TXPO . . . 25
4.4 Enrichment of TXPO . . . 26
4.4.1 Genes Integration . . . 27
4.4.2 GO-CAM Integration . . . 28
4.4.3 Biological Pathways Integration . . . 29
4.5 Schema in Context . . . 30
4.6 Structure of the Named Graphs and Provenance . . . 32
4.7 Summary . . . 34

5 Evaluation 36
5.1 Demonstration . . . 36
5.2 Discussion . . . 39
5.2.1 Methodology Relevance . . . 39
5.2.2 Design of the Upper Structure . . . 40
5.2.3 Relevance of the Integrated Data . . . 42
5.2.4 Usage of Named Graphs and Provenance . . . 43
5.2.5 Reproducibility and Transparency of the Method . . . 44
5.3 Summary . . . 45

6 Potential Applications 47
6.1 Search Tool . . . 47
6.2 Bridge between In Vivo and In Vitro Testing . . . 48
6.3 Text Annotation Tool . . . 49
6.4 Summary . . . 50

7 Conclusion 51
7.1 Summary . . . 51
7.2 Achievements . . . 52
7.3 Future Work . . . 53

References 55

Appendix A Internship with the TOXIN project 61

Chapter 1

Introduction

The field of toxicology plays a crucial role in protecting human health and the environment by identifying and assessing the potential risks posed by chemicals. This is often done through the use of toxicology tests, which can provide valuable information about the potential effects of a chemical on living organisms. However, the process of hazard assessment can be complex, as the underlying biological processes are diverse and sometimes not well understood. It involves gathering and evaluating data from multiple sources, including chemical structures, toxicology tests, and biological knowledge. In this thesis, we present a method for integrating multiple toxicological data sources into a centralized structure and finding relations among the data they contain, in order to facilitate hazard assessment for toxicologists.

1.1 Context

1.1.1 Toxicity Testing

Toxicity testing is a field of study that aims to understand the potentially harmful effects of chemical compounds on living organisms. It is an important part of the process of regulating substances that humans and other living being may be exposed to. [21] There are three main toxicity testing methods: in vivo, in vitro, and in silico.

In vivo toxicity testing involves exposing living animals to a substance to observe its effects on their health and behavior. In vitro toxicity testing, on the other hand, involves conducting experiments on cells or tissues that have been removed 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.

In silico toxicity testing involves the use of 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, as well as data from in vivo and in vitro studies, to predict the toxic effects of a substance. Although in silico toxicity testing has made significant advances, it is not yet a universally applicable method for accurately and consistently predicting safety of new compounds and must be used in combination with other methods. [45]

Toxicity testing is used for risk assessment of new compounds, i.e., "the process of gathering all available information on the toxic effects of a chemical and evaluating it to determine the possible risks associated with exposure." [19] It has four steps:

1. Hazard identification: task to identify the potential hazard of the new compound to determine whether the compound is toxic.

2. Hazard characterization: when the previous step gives evidence of hazard, this step tries to determine the dose at which harmful effects are caused by the compound. The combination of this task and the previous one is called hazard assessment.

3. Exposure assessment: the process of estimating the amount of a toxic compound that an individual is exposed to, and the frequency, duration, and intensity of that exposure.

4. Risk characterization: the process of evaluating and interpreting the results of the three previous steps to determine the nature and likelihood of adverse health effects in humans or the environment that may result from exposure to the compound.

1.1.2 Toxicological Ecosystem

The toxicological data ecosystem[1] includes a diverse range of stakeholders, including regulatory agencies, industry, research organizations, and consumer advocacy groups. These stakeholders interact with one another in various ways, including through the development of toxicological data and risk assessments, the establishment of regulatory standards and guidelines, and the dissemination of information to the public.

All of these stakeholders have different purposes and operate within a complex and often dynamic environment. Therefore, the toxicological data ecosystem is heterogeneous and varied. It offers a large collection of data from multiple data sources, but these sources use different conventions, are built from similar but sometimes slightly different knowledge, and evolve

[1] Or toxicological ecosystem for short.

differently over time. For these reasons, there is a range of challenges and uncertainties in order to navigate through the toxicological ecosystem effectively.

1.1.3 3R Principle

The 3R principle is a guiding principle and stands for "Replacement, Reduction, and Refinement." The principle aims to minimize the use of animals in research and testing, while still ensuring that high-quality scientific research can be conducted in order to protect human health and the environment with the same confidence as with animal testing. It encourages the replacement of animal testing with alternative methods whenever possible, the reduction of the number of animals used in research, and the refinement of research methods to minimize any suffering or stress to the animals. The 3R principle is an important ethical consideration in the field of toxicology, as it aims to balance the need for scientific progress with the ethical treatment of animals.

In the context of risk assessment, a recent directive [12] put forward by the European Parliament largely restricts to use of animals for the development of new compounds, especially for cosmetics, where the use of animals is now forbidden. Therefore, new methods have to be developed for risk assessment based on in vitro or in silico testing. To reach this goal, the huge toxicological ecosystem should be able to provide information to have a better understanding of the biological processes behind the effects observed in the different tests, and make the bridge between in vivo tests, on one hand, and in vitro and in silico tests on the other. The challenge is to be able to access and analyze relevant data efficiently in the toxicological ecosystem.

1.1.4 TOXIN Project

Since the EU directive on animal testing, no new compounds have been validated to be used for cosmetics. Therefore, developing alternative methods for evaluating the safety of chemicals that do not involve animal testing is critical for developing better cosmetics in the EU. In this context, the TOXIN project aims to develop non-animal based, human-relevant methods for evaluating the effects of exposure to toxic substances. TOXIN is composed of IT experts, law experts, and toxicology experts, which will be called “domain experts” in this thesis.

For this purpose, the TOXIN project intends to develop methods that use in vitro or in silico testing, based on relevant information from the toxicological ecosystem. As the majority of the human understanding about the toxic effects of different chemicals is based on animal studies

[19], TOXIN is developing a knowledge graph that gathers information about in vivo tests, described in documents dossiers issued by the Scientific Committee on Consumer Safety[2] about cosmetic ingredients included in the Annexes II, III, IV, V and VI of the European Cosmetics Regulation.

This thesis contributes to the TOXIN knowledge graph and is the continuation of an internship with the TOXIN project. The detail of this internship is presented in Appendix A.

1.2 Problem Statement and Challenges

Although in vivo testing is now strongly restricted, the accumulation of toxicological data through animal testing has provided valuable information about the safety of chemicals and their potential effects on humans. The key for using more ethical testing methods as in vitro or in silico is to have a better understanding of the toxicological domain, through previous in vivo data and the large toxicological data ecosystem. To reach this goal, tools need to be developed in order to efficiently access and compare relevant data. The adoption of ontologies to build toxicological knowledge graphs allows one to integrate diverse data into a coherent and meaningful framework in order to generate new scientific understanding [7], and helps for the development of predictive toxicological applications [25].

In particular, hazard assessment involves identifying sets of evidences that demonstrate the presence of toxicity hazards after a toxicity test. These sets of evidences are used to prove the existence of a toxicity hazard and are often well-defined for in vivo tests. However, they are not as well-defined for in vitro tests. This thesis aims to help toxicologists with hazard assessment by bringing together multiple sources of toxicological information into a knowledge graph to facilitate the development of methods for identifying strong toxicity evidences through in vitro tests. By doing so, we hope to improve the ability to assess toxicity hazards in new compounds. The resulting research question is: **Starting from the in vivo data in TOXIN, how could information from the toxicological ecosystem be integrated into a knowledge graph to facilitate hazard assessments of new compounds?** In other words, the research has four parts:

1. Identify potential data sources to integrate.

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

2. Assess whether these resources can be integrated and how to integrate them into a knowledge graph.

3. Propose and implement a method for organizing and integrating these sources into a knowledge graph.

4. Develop a prototype on top of the knowledge graph to demonstrate and discuss its utility.

The research was conducted with the guidance of a domain expert in order to have insights into the needs in their field and the added value of each dataset that we could integrate. Several challenges must be taken into consideration:

- **Granularity:** The granularity of the toxicological domain can vary widely, depending on the specific focus of a particular study or research area. Some data may be at a very detailed level, explaining the study of the molecular and cellular mechanisms of toxicity, while others may refer to more broad-scale effects, such as the impact of toxic substances on entire ecosystems. Overall, the granularity of the toxicological domain reflects the complexity of the subject matter and the wide range of factors that can influence the toxicity of a given compound.

- **Interpretation:** There is no absolute knowledge of the toxicological domain. Different studies or different experts can have different opinions about a subject, and there could be contradictions between the different data sources. Therefore, some knowledge about the toxicological domain can never be certain, and the different interpretations should be taken into account.

- **Big data:** One of the main problems with big data in the toxicological ecosystem is the sheer volume and variety of data, which can make it difficult to store, manage, and analyze effectively. Additionally, there may be issues with data quality and heterogeneity, as well as the need to ensure privacy and security when handling sensitive health and environmental information.

- **Skill gap:** Working with domain experts who have different skills and approaches to problem-solving can be challenging. The key is to find a way to effectively collaborate and share knowledge in order to find the best solution to the problem at hand.

1.3 Methodology and outline

The methodology used to achieve this work is inspired by the guidelines of the Design Science methodology as presented by [28], i.e., design as an artifact, problem relevance, design by

evaluation, research contribution, research rigor, design as a search process, and communication of research.

The guideline entitled “communication of research” is followed all throughout this thesis to explain the work achieved for this thesis. The structure of the report is as follows:

- **Chapter 1 - Introduction:** Presentation of the context, the research question and the challenges of the problem. This chapter assesses the problem relevance of the work.
- **Chapter 2 - Background:** Brief introduction to data modeling with graphs, overview of the knowledge graph data model, and presentation of techniques to build and maintain knowledge graphs.
- **Chapter 3 - Related work:** Introduction of previous work on developing more ethical methods for hazard assessment of chemical compounds, and discussion about related research on knowledge graph construction and enrichment, as well as applications in various fields.
- **Chapter 4 Method and implementation:** This chapter covers the method developed for this thesis and describes the structure and implementation of a prototype that instantiates this method. This follows the design as an artifact guideline.
- **Chapter 5 - Evaluation:** Demonstration and discussion of the utility and functionality of the method along with the prototype. This chapter follows the guidelines’ design by evaluation, research rigor and design by a search process.
- **Chapter 6 - Potential applications:** Presentations of the potential usages of the method and the prototype developed in the field of toxicology. The chapter presents the research contributions.
- **Chapter 7 - Conclusion:** Summary of what was done in this thesis, highlighting the principal achievements and quick overview of some potential future work from this thesis.

Chapter 2

Background

The problem addressed in this thesis is part of the field of data management (DM), i.e., "the practice of collecting, keeping, and using data securely, efficiently, and cost-effectively." [40] More precisely, the processes involved are data modeling and data integration for creating a knowledge graph (KG).

This chapter starts by introducing the basics of data modeling with graphs. Then, it gives an overview of knowledge graph systems and their architectures. Finally, it presents how a KG is constructed and maintained, and the role played by data integration.

2.1 Data Modeling

"Data modeling is the process of creating a visual representation of either a whole information system or parts of it to communicate connections between data points and structures." [31] In graphs, these connections are represented with nodes and edges where nodes are entities and edges are relationships.

2.1.1 Data Graphs

A triple is the combination of a subject, a predicate, and an object $< s, p, o >$ and it can be used to represent a relationship between entities. A triple or set of triples can be seen as a directed edge-labeled graph also called **data graph** G where subjects and objects are nodes, and predicates are edges. Fig. 2.1 shows an example of a triple and the relationship it represents. On the left, there is a sentence in natural language. This sentence is then represented by a relationship in the form of a triple in the middle. And finally, on the right, the triple is in the form of a graph.

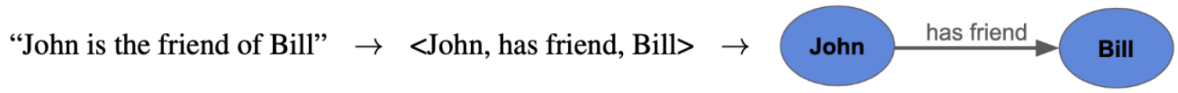


Fig. 2.1 Example of a triple and the relationship it represents in the form of a triple.

Sometimes, a data graph can be identified by a "name." Formally, such **named graph** is represented by a pair (n, G) where G is a data graph and n is the "name" of the graph. Named graphs can be gathered into a **graph dataset** $D = (G_D, N)$ where G_D is the default graph and N is a set of named graphs with distinct names (the set can be empty). [29] Therefore, a graph dataset is not composed of triples but rather by quadruples (or quads) $< g, s, p, o >$ where g is the "name" of the graph and $< s, p, o >$ is the triple, as previously.

The Resource Description Framework (RDF) [59] allows one to represent triples or quads with RDF statements. These statements can be expressed with the Turtle [44] syntax. A set of triples expressed in RDF is called a **RDF graph** where each entity is represented by an IRI[1], a literal or a blank node. A "name" in the form of an IRI can be assigned to an RDF graph to identify it, and a set of RDF named graphs is called a **RDF dataset** (equivalent to the graph dataset). RDF graphs and RDF datasets can be queried through SPARQL queries. [50]

When an IRI is looked up, a description of the entity it represents can be returned in the form of RDF, allowing RDF graphs to link to and reference related entities described in external RDF graphs. This interconnected network of **Linked Data** allows for the creation of a global graph that can be accessed and shared by a wide range of applications and users. [29]

2.1.2 Ontologies and Vocabularies

Data graphs are a powerful tool for representing relationships between data elements, but sometimes the goal is to represent more complex, semantically rich relationships. In these cases, it can be helpful to use more structured tools such as vocabularies and ontologies to provide mechanisms for representing the semantics of the relationships within the data graph. By using these tools, it is possible to create data graphs that are more expressive and meaningful, and that can be more easily understood and used by others. [6]

A vocabulary can be represented with RDF Schema (RDFS) [24]. RDFS is an extension of RDF. RDFS provides a set of new mechanisms to represent classes, hierarchical relations,

[1] An IRI or Internationalized Resource Identifier uniquely identifies an entity. Its format is similar to URLs.

and other features. We can use RDFS to represent class hierarchies, relation hierarchies, and the relationships between classes. RDFS is meant to infer implicit information from explicit information. For instance, if a graph contains <Garfield, is of type, Cat> 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 can have instances, whether there are contradictions in the model or the data, etc. To support these tasks, we need ontologies and ontology languages.

The concept of ontology is thus larger. It refers to "an explicit specification of a conceptualization where a conceptualization is a structured interpretation of a part of the world that people use to think and communicate about the world." [3] In other words, an ontology aims to model a domain with as much precision as possible. To achieve that goal, more precise mechanisms are needed. The Web Ontology Language (OWL) [49] provides tools to represent rich and complex relationships between entities. An ontology expressed in OWL has several useful properties [34]:

1. Extensible and customizable: it is easily possible to add or modify some parts of an ontology.
2. Shareable: well-designed ontologies can be re-used and combined to obtain a larger, more complete representation of a domain.
3. Inter-operable: because of their standard format, they can be shared between systems.
4. Machine-readable: it can be understood and interpreted by a computer program.

OWL is based on a decidable subset of first-order logic that is capable of expressing a wide range of relationships and axioms. It can be used for many complex reasoning tasks, such as relationship inference, consistency checking, classification and explanation generation. However, there are certain types of axioms that OWL is not able to model. Rule languages based on Horn clauses, such as Datalog [46], allow for the definition of a broader range of rules, but do not provide support for reasoning tasks beyond inference and are not designed to be compatible with OWL. The Semantic Web Rule Language (SWRL) [30] is another option that allows for the definition of several rules that can be added to an ontology, and it is compatible with OWL. It is important to be careful when using these rules, as they can make the reasoning process intractable if not used carefully.

A controlled vocabulary provides a consistent way to gather information based on a well-defined organization. For example, [39] is a controlled vocabulary in the form of guidelines to

deposit on the Digital Repository of Ireland. These guidelines help to integrate new information into the general repository by giving predefined constraints that the information's structure must satisfy, e.g., author fields are "free strings", but they impose a controlled format so that the DRI platform can process them. The use of ontologies can help in the construction of controlled vocabularies by defining constraints through axioms. To resume, vocabularies are lightweight ontologies that only define basic axioms, and a controlled vocabulary provides an organization to represent information and can be expressed with an ontology.

2.1.3 Provenance Ontology

Provenance information describes the origin and history of entities in a database. Such information is important to explain how the data was obtained and it can be used to add value to the data and to help verify its integrity, especially with online data that can be manipulated by a variety of different people. [10]

Provenance ontologies, such as PROV-O [47], are important tools to describe provenance information of entities in an interoperable and meaningful way. They provide a standardized framework for capturing the relationships between entities involved in the provenance, which can be used for various applications. Depending on the context, different depths of provenance information are needed, i.e., the more complex the provenance information expected to be represented, the larger the number of entities and relations needed to represent them. To represent simple provenance information, PROV-O provides a starting point category composed of three classes and nine predicates. A schema of this category is illustrated in Fig. 2.2.

PROV-O also proposes more complex structures to describe in more detail the provenance of a knowledge graph, which will not be discussed in this thesis.

2.2 Knowledge Graph

The definition of a knowledge graph (KG) varies slightly from one source to another. [29] provides a fairly extensive survey of the different definitions of a knowledge graph. In this thesis, we will adopt their definition, which defines a KG as "a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities." [29] This definition is quite abstract but it captures well the flexibility of the concept. Concretely, a KG is any graph dataset or combination of graph datasets. A KG can be built following one or several ontologies. If it is

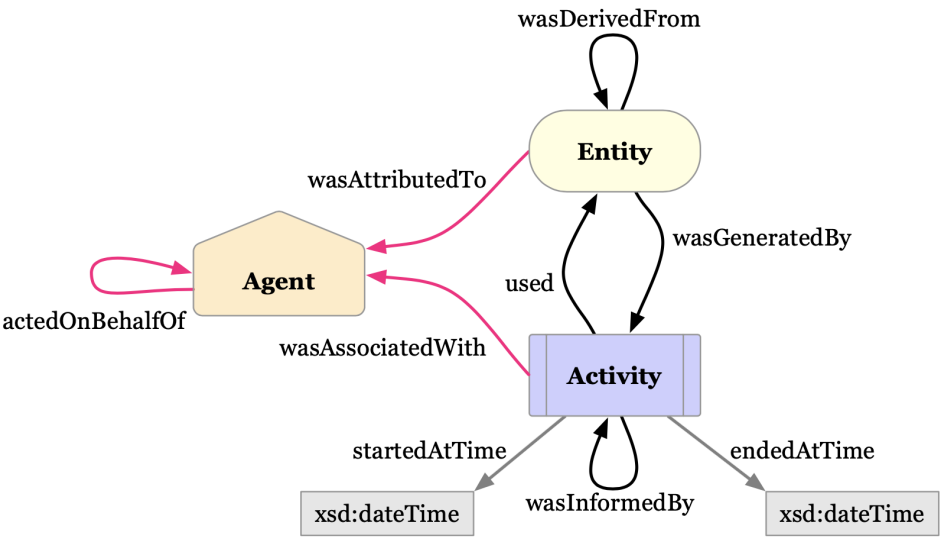


Fig. 2.2 The three Starting Point classes and the predicates that relate them. [10]

the case, it will be called an ontology-based KG.

The usage of knowledge graphs to represent information presents two interesting advantages. Firstly, they efficiently store and integrate large and heterogeneous data sources. Secondly, they bring additional benefits as input or output to machine learning or artificial intelligence algorithms. For example, neural language processing or computer vision tasks are often more efficient with KGs. [9]

2.3 Knowledge Graph System

While the KG refers to the data and knowledge, a knowledge graph system (KGS) encompasses all the tasks, design decisions, integration, etc., that are needed in the KG's development, maintenance, and storage.

2.3.1 Architecture

A popular architecture for designing a KGS is the Abstract Reference Architecture (ARA) presented in [41]. The architecture is represented in Fig. 2.3.

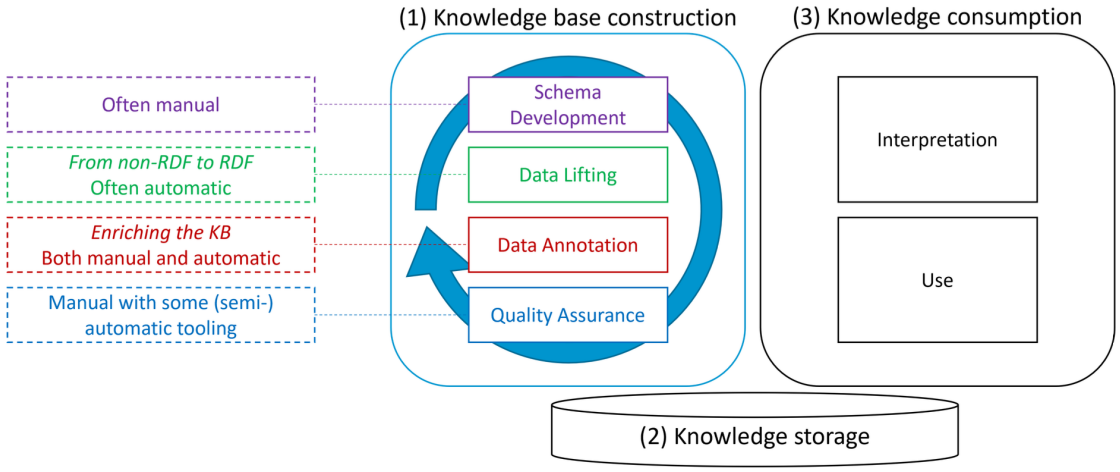


Fig. 2.3 Architecture of a knowledge graph project. This image came from [48] and adapted the Abstract Reference Architecture (ARA) [41].

This is the architecture used in the TOXIN project to develop their knowledge graph, and is described in [48]. It is divided into three layers: knowledge graph construction, knowledge storage, and knowledge consumption.

Knowledge Graph Construction Layer

This layer represents the knowledge life cycle, i.e., the cycle through which knowledge is created, stored, maintained, and shared within an organization. This cycle is composed of four steps: schema development, data lifting, data annotation, and quality assurance. Schema Development is the development of the structure and the vocabulary of the KGS, which can be done through an ontology. Data Lifting involves transforming raw data into semantic data and interlinking the KGS with other heterogeneous sources, while Data Annotation enriches the KGS. It contextualizes data points by linking them to their corresponding concept or class defined in the schema. Finally, Quality Assurance ensures the accuracy of the knowledge graph. Automation is possible for some tasks, but others require manual intervention. Each task informs the other to ensure the knowledge graph is up-to-date and consistent.

Knowledge Storage Layer

According to [41], the goal of this layer is to provide storage and access to data. They propose two approaches to this question. The first one is to keep the data stored “as it is” and design

some pipelines for accessing it. The second one is to store the data as graphs with a data model like RDF.

Knowledge Consumption Layer

This layer's concern is how to access and use the knowledge graph for an end-user. It has two sub-parts: interpretation and use. The interpretation sub-part is responsible for providing a summary of the contents of the KG for end-users, in order to help them understand what is contained in the KG and what types of questions it can answer. The use sub-part focuses on how to search and query the KG in order to efficiently access the information it contains. This includes a search mechanism for finding relevant information within the KG, as well as a query generation and answering process for formulating and executing queries to retrieve specific pieces of information.

2.3.2 Knowledge Construction and Maintenance

The knowledge construction and maintenance tasks correspond to the first layer of the architecture presented in Sec. 2.3.1. As specified in [41], it is responsible for extracting, integrating, and extending the KG under a specific schema or vocabulary. In the previous section, we presented the general architecture of ARA for data management in a KGS. Here the focus is on a methodology that can be used for the integration of external sources into the KG. To reach this goal, they describe a general pipeline that follows an iterative and incremental life cycle. The steps involved in this cycle are represented in Fig. 2.4. Nowadays, the general opinion is that model-based approaches such as this pipeline are the most adapted for the construction of a knowledge system. [53] The different steps of the pipeline, as described by the authors, are presented hereafter.

Specification

The goal of this step is to answer the questions "Why are we building this KG?" and "What should be in the KG?" For the first question, the basis of an agreement should be reached between the domain expert and the developer in order to define what the knowledge graph system should be able to do. This can be done through Competency Questions [23], i.e., natural language questions that represent the tasks that the KG (or KGS) should be able to complete. The second question is about the exploration of the available resources. It is interested in the identification and analysis of possible data sources in order to see if they bring added value to the KG and if it is possible to integrate them.

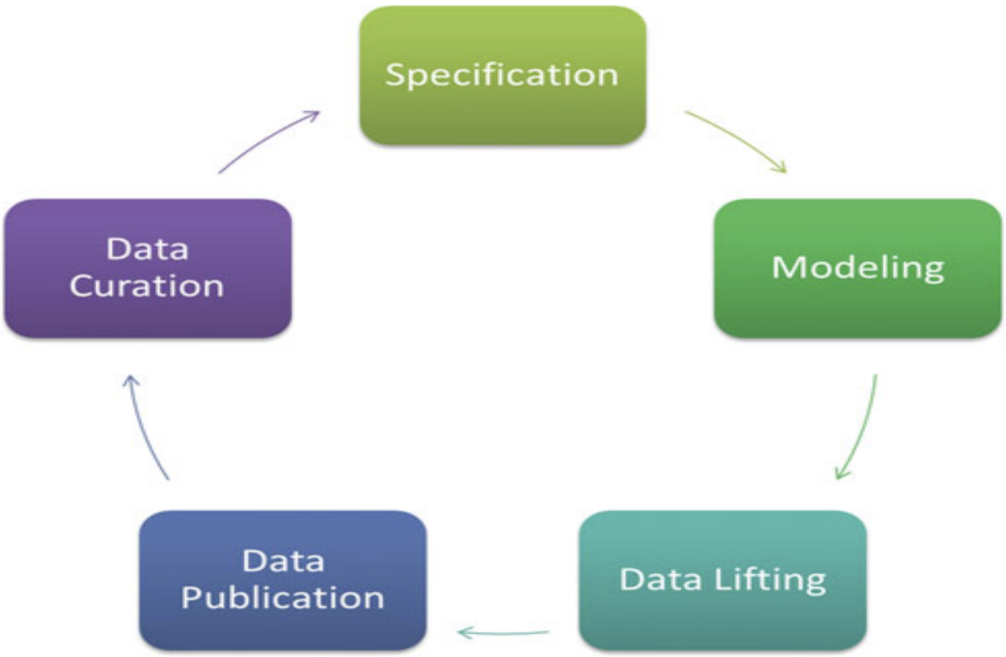


Fig. 2.4 Knowledge construction and maintenance life cycle from [41].

Modeling

Modeling is the task of representing the domain as well as possible. This can be done with the development of an ontology. For this purpose, reuse is highly encouraged. The modeling should encompass as much existing structure as possible in order to promote knowledge sharing. This is an important part of ontological commitment, i.e., an agreement between different agents to use shared structures and vocabularies consistently and coherently [22]. This guarantees consistency but not completeness. Therefore some parts of the ontology or sometimes the entire ontology should be created from scratch, which is the field of ontology engineering.

Data Lifting

The activity of data lifting is to take the sources found in the specification and integrate them into the KG. This task of data integration has two facets:

1. Transformation: it is the task to transform a data source content into graph data (RDF triples). There are two requirements: it should be possible to answer the same queries with the graph data and the original data and the new triples should be compatible with the schema that is already defined. Several tools exist to support this task like R2RML [54] or RML [17].

2. Linking: it is the task to discover links between entities inside the KG or with external KGs. This can be done manually but there are also tools to do this like Silk framework [58] or the alignment API [13].

Data Publication

This is the task of making the KG available.

Data Curation

It is concerned with cleaning, maintaining, and preserving data for reuse. It has two steps: the identification of the mistakes (obsolete links, ...) and the resolution of these errors in order to keep the KG correct and up-to-date.

2.4 Summary

This chapter first discusses the basics of data modeling with graphs, introducing key concepts such as triples, named graphs, and RDF datasets. The second part of the chapter discusses knowledge graphs and knowledge graph systems, including the Abstract Reference Architecture (ARA) for designing a knowledge graph system, and a methodology for the construction and maintenance of KGs. The background Chapter provides an overview of the key concepts and considerations involved in the creation and use of knowledge graphs. These concepts are necessary for the understanding of the work presented in this thesis. Now that they are introduced, we will present the existing research and tools related to the research question.

Chapter 3

Related Work

The aim of this thesis is to develop a method for integrating multiple toxicological data sources into a centralized structure and finding relations among the data they contain, in order to facilitate hazard assessment for toxicologists. In the life-science domain, the potential of knowledge graphs has been asserted to build a better understanding of the domain and facilitate the development of machine learning and AI tools [5].

In this chapter, we will review related work in two areas: existing tools in the toxicological ecosystem for hazard assessment, and existing work about the construction of knowledge graphs by integrating and linking multiple data sources.

3.1 Existing Tools for Hazard Assessment

The purpose of hazard assessment is to identify potential hazards and assess the risk they pose to humans or the environment. The process involves gathering and evaluating data from various sources, including chemical structures, toxicology tests, and exposure data. As such, the TOXIN project gathered and organized toxicological data in order to facilitate hazard assessment.

From this data, there are several approaches to hazard assessment, and various tools have been developed to support this process. One such tool used by the TOXIN project is the OECD QSAR Toolbox [16], which is a software application that aims to classify chemicals into categories based on their structural characteristics and potential toxic mechanisms of interaction. The Toolbox is designed to support hazard assessment and risk assessment by providing a convenient and standardized way to classify chemicals based on their potential hazards. COSMOS NG [61] is another interesting tool, providing a database of toxicity opinions about several

chemicals that can be used for hazard assessment, and in silico tools for analyzing toxicity data and performing analyses such as category formation.

Using tools such as the OECD QSAR Toolbox and COSMOS NG can be an important part of the hazard assessment process. However, it is important to recognize that these tools are not always sufficient to accurately assess the toxicity of a chemical. There are a number of factors that can affect the accuracy of hazard assessments based on these tools, including the quality and reliability of the data used to train the models, the relevance of the data to the chemical being evaluated, and the limitations of the tools themselves. If there are limited similarities between the chemical being evaluated and those for which data is available, or if the available data is of low quality or was not collected according to established guidelines, the accuracy of the hazard assessment may be compromised.

In a broader context, there is a growing interest in using in vitro and in silico testing for hazard assessment and supporting the 3R principle. One of the challenges in this area is the need to develop standardized vocabularies and ontologies that can be used to efficiently represent data about different toxicology tests [25], including in silico, in vitro, and in vivo tests. The OpenTox initiative [55] aims to address this challenge by providing a framework for integrating and analyzing diverse data sources using an ontology, with the goal of improving the predictivity of toxicology models and supporting decision-making in chemical safety assessment.

OpenTox developed the ToxPredict tool[1], which is a web-based platform that uses machine learning algorithms to predict the toxic effects of chemicals based on their chemical structure and properties. Although they state that the predictions are computed using various data sources and different machine learning algorithms, it seems that the predictions are solely based on existing toxicity tests and classification methods such as QSAR toolbox, without exploiting the underlying biological processes affected.

Tox21 [57] is a research program that seeks to identify new mechanisms of chemical activity in cells and use this information to prioritize untested chemicals for further evaluation and to develop more accurate predictive models of human response to toxic substances. The ultimate goal of Tox21 is to create comprehensive models of biological interactions in the organism that can be used to generate high-throughput predictions for risk identification purposes. Therefore, their purpose is to provide a screening tool that would be able to quickly identify potential

[1] https://www.opentox.net/library/toxicity-prediction

hazards amongst a long list of potentially toxic compounds.

The work of this thesis has similarities with Tox21 as it seeks to provide a better understanding of the toxicological domain. However, while Tox21 is working to create a tool that can handle a large volume of work efficiently to efficiently identify existing compounds that could be toxic, this thesis is focused on creating a tool that can support toxicologists in their work and help them to provide reliable and accurate results for hazard assessment, while developing new compounds.

The main objective of this thesis is to find ways to link multiple data sources together by identifying relationships between the entities they contain. There have been numerous studies focused on linking different types of entities in the toxicological domain, such as [62] developing a comprehensive map of disease-symptom relations, [52] developing a method for representing the organization of human cellular processes in a network and mapping diseases onto this network, and [42] providing an overview of existing work on integrating genes, pathways, and phenotypes to better understand the effects of gene mutations. While these studies have made valuable contributions, they are specific in nature and may not necessarily be applicable to the broader task of integrating all available data sources. The focus of this thesis is to integrate all relevant sources together, even if some possible relationships are missed or may be subject to debate, allowing domain experts to consider them in their analysis.

3.2 Related Work on Knowledge Graph Enrichment and Data Integration

The management and integration of heterogeneous data sources into a single, centralized model is a common problem that has arisen with the emergence of big data. This problem affects various domains, including toxicology, and solutions developed in these other domains may be useful or adaptable for addressing the problem in toxicology.

There are various techniques for constructing a knowledge graph and integrating data into it. Chapter 6 of [29] provides an overview of these techniques, including the use of human collaboration and text sources, as well as tools such as R2RML [54] for converting CSV files into RDF graphs. The field is very vast and the choice of appropriate techniques and tools to use depends on the specific problem at hand. In this section, we present examples

of solutions for building a knowledge graph in various domains that address different challenges.

Dacura [43] is a data curation platform that helps organizations manage, curate, and share structured data, with a focus on creating and using high-quality data assets. It offers tools and services for data modeling, data integration, and data governance. Originally developed to facilitate the gathering and sharing of archaeological information, Dacura relies on human input and techniques to find consensus among multiple annotators.

Bio2RDF [4] is another example of a platform for integrating and standardizing data from multiple sources. It is specifically geared towards the creation of Linked Data versions of biological databases, enabling data from these sources to be more easily linked, queried, and integrated with other data. Bio2RDF uses a combination of manual curation and automated techniques to extract and convert data from various sources, and it provides a range of tools and services for data integration, including data mapping, data cleansing, and data transformations.

While these tools are effective in their respective domains, a more specialized and toxicology-focused approach may be necessary for effectively integrating and relating data sources in the field of toxicology.

3.3 Summary

In this chapter, we reviewed related work in toxicology and data integration, including existing tools for hazard assessment such as the OECD QSAR Toolbox and COSMOS NG. While these tools can be useful in assessing chemical toxicity, we also highlight their limitations and the need for additional data in some cases. We also examined the work of the OpenTox initiative and its ToxPredict tool, as well as the Tox21 program, which addresses the same problem but from different angles. Finally, we consider similar efforts in other fields to integrate and link multiple data sources in support of various applications and quickly describe Dacura and Bio2RDF.

Now that we reviewed the relevant related work and introduced the necessary background information, the following chapter will present our method for integrating multiple data sources and facilitating hazard assessment.

Chapter 4

Method and Instantiation

This chapter presents the contribution of this thesis in the form of a method to build a knowledge graph by integrating heterogeneous data sources from the toxicological ecosystem. The instantiation of this method is called the TOXIN Enriched Knowledge Graph (TEKG). The purpose of this knowledge graph, alongside its knowledge graph system, is to integrate several resources into a general architecture to help toxicologists to access more easily data relevant to their research, and help them to put concepts in relation in order to facilitate the development of new hazard assessment methods based on in vitro or in silico testing.

This chapter will be presented in a bottom-up fashion. First, the TOXIN knowledge graph is described in more detail. Then, we will present the ToXic Process Ontology and its integration, which will be the foundation of the prototype's structure. After that, we describe the integration of various data sources into the KG. Finally, the general schema of the KG is exposed as well as the organization of the different sub-graphs (or named graphs) within the general KG.

4.1 TOXIN Knowledge Graph

As explained in Section 1.1, the TOXIN project has integrated information on in vivo tests, described in documents dossiers[1]. Each dossier contains information about experiments (also called tests) of a compound on laboratory animals. The information includes the quantity of the compound tested, the way it was inoculated, the species on which the compound was tested, and so on. They also include information on the outcome of these tests. The dossiers also include opinions (written by the authors of the dossier) about the compound's toxicity after the test (guidelines used, effects observed,...). In other words, not only do dossiers gather

[1] An example of such a dossier can be accessed at https://ec.europa.eu/health/scientific_committees/consumer_safety/docs/sccs_o_195.pdf

information about experiments, published in papers, for instance, that information is then used to formulate an opinion. The data contained in these dossiers are stored in the TOXIN KG. This KG, currently under development, leads to more efficient access to data for toxicologists. A simplified illustration of the TOXIN KG's structure is represented in Fig. 4.1.

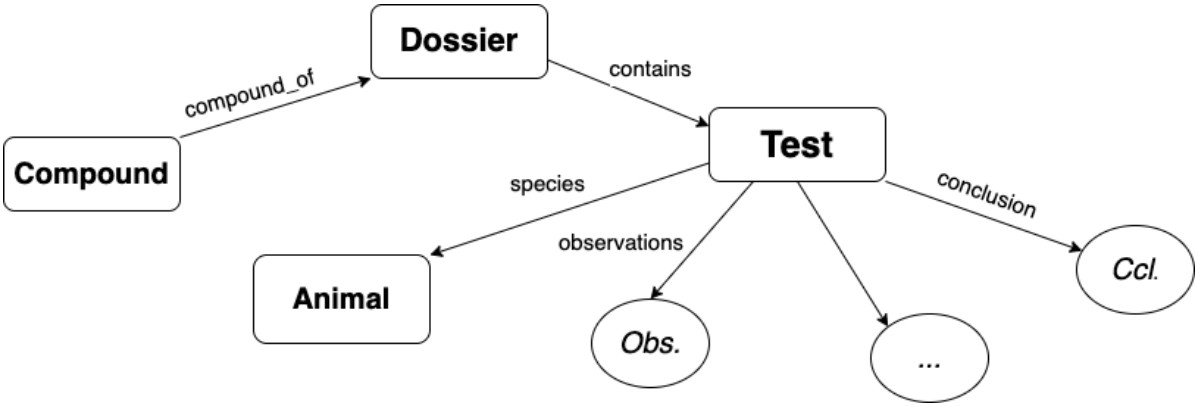


Fig. 4.1 Simplified schema of the TOXIN KG structure that shows the relations between a compound, a dossier, and the information related to this dossier.

In this figure, rectangles represent resources identified by IRIs and circles represent literals. The schema illustrates that a dossier is associated with two resources: a compound (the compound that was tested) and one or more tests. Each test is associated with a species (the animal species on which the tests were made). Moreover, tests are linked to literals that describe the experiment, such as the observations and the conclusion after the test. Thanks to this KG, domain experts can efficiently access relevant data, such as, for example, all the dossiers in which a specific compound was tested, etc.

This representation is quite "human-oriented" as the purpose of the KG is to help domain experts to access the information. The majority of the information about dossiers is in the form of literals that contain textual data.

For the integration and linkage of the TOXIN KG with other data sources, this representation is not very suitable. Consequently, another representation said "machine-oriented" would be used further on. This new representation divides the raw text data into "atomic text entities". For example, instead of having a paragraph describing all the observations made in a dossier, there would be a set of clear observations in the form of a short sentence (e.g., "Increase of cholesterol in the blood"). The advantages of this representation will be explained later in this section.

Concerning the transformation of the data from one representation to the other, it can be done manually with people separating the text into sub-parts or automatically with some sort of tagging mechanism. This problem is however out of the scope of this thesis.

4.2 Methodology

As explained in Section 1.3, the goal of design science research is to create a new innovative artifact. "IT artifacts are broadly defined as constructs (vocabulary and symbols), models (abstractions and representations), method (algorithms and practices), and instantiation (implemented and prototype systems)." [28] In this thesis, the artifact developed is a KG, which integrates different data sources into a unique structure. The artifact is developed in the form of a method to build such a system, along with the implementation of a prototype of it (TEKG) in order to demonstrate its feasibility and its utility. The construction of the KG was based on the methodology presented in Section 2.3.2. This section will explain how the different steps were achieved in this project.

4.2.1 Specification

This task is done in three parts. The first step is to identify, in collaboration with the domain expert, the desired capabilities of the KGS through Competency Questions [23]. These questions should represent tasks that the KGS will be able to complete and should be realistic and feasible given the available resources. The second step involves identifying and analyzing potential data sources for integration into the KGS. These data sources should be chosen based on their relevance and potential to contribute valuable information to the KGS. Finally, the selected data sources and their intended integration into the KGS should be discussed and agreed upon with the domain expert in order to ensure that the final KGS meets his needs and provides value.

As a starting point, we define three competency questions that the KGS should be able to answer:

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 competency questions, defined in collaboration with the domain expert, represent the three desired features for the KGS and they will come back later for the evaluation of our

prototype. Concerning the data sources, six different resources have been integrated into the KG. These sources will be presented later, alongside their integration into the general structure.

4.2.2 Modelling and Data Lifting

These two tasks are different but strongly linked, as integrating a new data source may cause the model to be extended or revisited. The goal of data modeling is to define the core structure of the enriched KG. We are using the term "model" as it both includes a definition of the ontology (i.e., the concept and relations), as well as knowledge organization (i.e., how the different integrated pieces will be organized into (named) graphs. This model has to be able to give an overview of the different sources that will be integrated. This overview has to take into account some properties of the domain that will be represented.

In the toxicological domain, we highlighted the difficulty of representing the granularity within the biological processes. The model has to be able to accept and represent data at the cellular level, the molecular level, and the organic level, and even, if possible, to represent relations between all these levels. For this purpose, the ToXic Process Ontology was used. It will be presented in detail in the next section.

After the definition of this core structure, the integration of each data source generally involves the two tasks. Actually, for each integration process, the first step is to adapt the model for it to represent well the data that will be integrated. And the second step is to “lift” the data from the source and incorporate it into the KGS (by transforming it or through links).

4.2.3 Data Publication

There are different means to make a KGS available. Currently, this project is only a prototype that can be stored on a computer as a triplestore with an Apache Jena Fuseki server [20], and can be accessed through a SPARQL endpoint. This endpoint can be used to query the KGS with SPARQL queries and to visualize the relationships within the KG with Ontodia [38]. Even if this approach is not user-friendly, it has the advantage of being able to demonstrate to end-users the kind of links and entities that can be searched within the KG, and it is all we need for the development of a prototype.

The current method for accessing the KGS may be challenging for users without a background in IT. As the final goal of this project is to create a tool for toxicologists who may not

be familiar with these types of technologies, it is important to design an intuitive interface to make the tool more accessible and user-friendly. This is, however, not covered in this thesis.

4.2.4 Data Curation

Data curation is a key task in the maintenance of a knowledge graph. As its concern is maintenance and not creation, it is not a pressing issue that needs immediate attention in this project. However, our design of the KGS does include two features that will aid in data curation efforts in the future: the inclusion of provenance information and the use of named graphs to incorporate various integrations separately. These features will help to organize and manage the KG over time. We will come back to these two features after the presentation of the complete KG.

4.3 The ToXic Process Ontology (TXPO) as Upper Structure

To create the prototype, the first modeling task is to establish a general representation of the domain using some kind of upper structure. An ontology is a suitable tool for this purpose, as it provides a systematic representation of the concepts and relationships within a domain. While it would be possible to develop a new ontology from scratch, this would require extensive expertise and time. Fortunately, there are numerous existing ontologies in the toxicological domain that can be reused to save time and effort.

4.3.1 Presentation of TXPO

The ToXic Process Ontology (TXPO) [60] 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. Their focus is on the liver, as liver toxicity is the most frequent cause of the withdrawal of a new drug after testing. Nonetheless, the ontology is still in development, and the authors plan to extend it to other organs and other toxic processes.

TXPO is part of the Open Biomedical Ontologies (OBO) Foundry initiative [51], whose purpose is to develop a set of principles to guide ontology development. The resulting family of ontologies is easy to combine and share, and it proposes well-formed representations of the biological domain. The ontology reuse is a key guideline to be part of OBO, and as such,

TXPO contains several parts of other ontologies.

The structure of TXPO is composed of three layers:

1. The top layer is the Basic Formal Ontology (BFO) [1]. Its use is one of the guidelines of OBO and provides general entities and relationships. The use of BFO simplifies the shareability of ontologies within the OBO community.

2. The intermediate layer is composed of biomedical entities and is the combination of parts of existing ontologies. It contains "anatomic structures from Uber-anatomy ontology (UBERON), cells from the Cell Ontology, organisms from the NCBI Taxonomy, compounds from Chemical Entities of Biological Interest (ChEBI), biological processes and cellular components from the Gene Ontology (GO), qualities from the Phenotype And Trait Ontology (PATO), some molecule families from INOH, genes from the Ontology of Genes and Genomes (OGG), and diseases from the Disease Ontology." [60] While this description pertains to biology rather than computer science, this citation illustrates the granularity inside TXPO. We do not need to go into the details of all of these concepts for the purpose of this thesis.

3. The lower layer contains entities related to the toxicological domain. The key entity in this layer is the Process entity, which contains natural biological processes and sequences of processes along with the toxic processes[2] that can affect them.

4.3.2 Modifications of TXPO

TXPO is imported into TEKG using the propriety `owl:imports` defined in [49]. This property allows one to import an ontology, facilitating reuse. TEKG's upper structure is the TXPO ontology, and it is from this structure that the other data sources are integrated. TXPO has been developed to be used as a structure in order to build ontology-based knowledge graphs. In other words, TXPO defines a set of classes and axioms to represent a part of the toxicological domain. After that, the ontology-based knowledge graph is constituted of objects that are instances of the classes defined in the ontology. The goal of the classes and axioms is to guide

[2]For later, we make the distinction between different "types of toxic processes": adverse effects, toxic effects, toxic courses, and diseases. Adverse effects are the immediate, negative consequences of exposure to a toxic substance. These effects may be physical, such as skin irritation or vomiting, or they may be behavioral, such as changes in sleep patterns or changes in appetite. Toxic effects refer to the long-term, negative consequences of exposure to a toxic substance. These effects may be physical, such as organ damage or cancer, or they may be behavioral, such as changes in cognitive function or changes in emotional state. Toxic courses refer to the overall progression of toxic effects over time. Diseases are the end result of toxic courses, and are often severe or even fatal conditions.

the construction of the KG and to allow different reasoning tasks, such as inference, within the KG.

However, our usage of TXPO is different. We want to travel the ontology as if it "was the KG", i.e., to see the classes as objects and the axioms as relationships. With this point of view, a relationship represents a potential link between two objects, while with the previous point of view, an axiom represents a restriction on the relationship between two individuals from two classes.

To reach this goal, TXPO was slightly modified by transforming all the "direct" axioms into relationships. For example, the axiom "Non-alcoholic fatty liver disease (NAFLD) is a subclass of something that has part some hyper-function of fatty acid biosynthesis (HFAB)" means that any instance of NAFLD can have part some HFAB. We transform this axiom into the direct relation "NAFLD can have part HFAB". For more complex axioms, with conjunction or disjunctions, for example, we separate the complex axioms into simple ones when possible, and use the same technique as before. Concerning the classes, they are not modified but are viewed as objects.

For representation, we use Ontodia, and one limitation is that it cannot represent axioms. Therefore, these modifications not only simplify the process of querying the ontology, but also its representation. It is important to note that these modifications do not change the semantic meaning of the ontology. All the triples relative to the importation and the modification of TXPO are contained in a dedicated named graph.

4.4 Enrichment of TXPO

Now that the general model for the KG is defined, the process of integrating various data sources into the KG using the methodology we have previously outlined can begin. The toxicological ecosystem is vast and contains many data sources that would be valuable to incorporate into the KG. To start, we have selected three types of sources that are of particular interest to toxicologists. By integrating these sources into the KG, it should be possible to answer the three competency questions defined during the specification step. Moreover, in order to be able to easily recognize the different integrations, the triples relative to each of them are stored in dedicated named graphs.

In the integration of external data sources, 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 during this project, we have designed our own IRIs to describe them. This choice allows us to clearly distinguish between original resources and those added as part of this project, while also providing transparency and traceability in the integration process. Other approaches that provide a uniform set of IRIs could be used but it is not in the scope of this project.

4.4.1 Genes Integration

TXPO contains a set of human genes that are imported from the Ontology of Genes and Genomes (OGG) [26]. The focus of this ontology is to offer classes and relationships to represent genes and genomes in different organisms. TXPO only imports the genes related to human organisms, but integrating the genes from other species would be straightforward as it would be a similar exercise.

TXPO also contains entities and relationships from the Gene Ontology (GO) [2]. The goal of this ontology is to represent the functions of genes, i.e., to provide a set of entities and relationships to describe the biological roles of genes and proteins. The main use of this ontology is to regroup and compare gene products and proteins from different species or individuals, depending of their role in the organism. GO is divided into three categories: biological process, molecular function, and cellular component. This separation allows for taking into account the granularity of the domain.

With the purpose to make in relation 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. Most of these natural processes are related to GO terms from the “biological process” category. Finally, it is possible to link these GO terms to the gene products involved in it through GO annotations.

A GO annotation represents 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. The field of annotation of genes is in constant evolution. As such, an annotation is not guaranteed to be reliable. Therefore, each annotation is associated with a proof, which can be more or less strong depending on the confidence in the annotation.

For the integration of such annotations, we created a class: “annotation”, and three predicates: “has annotation”, “has gene product” and “has proof.” With these new elements, the structure of the integration of an annotation is done as presented in Fig. 4.2, where the blue rectangles represent classes and the yellow circle represents an attribute.

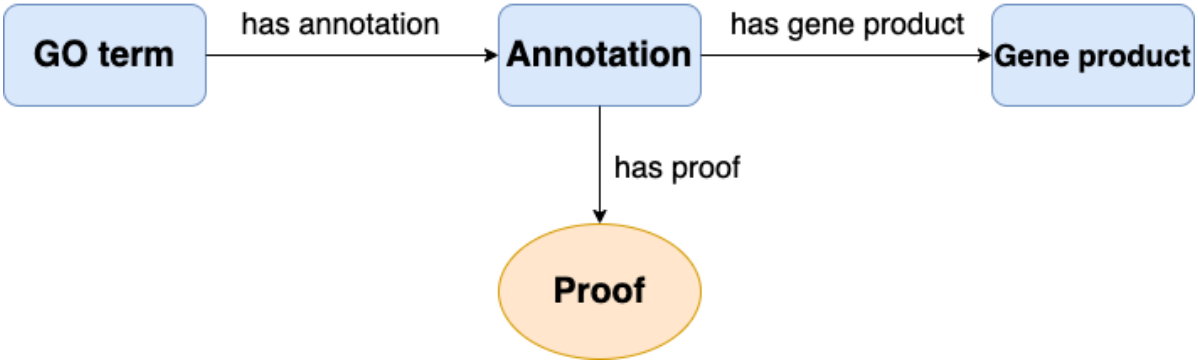


Fig. 4.2 Structure of the integration of a gene product annotation.

In this project, the integration of annotations from two sources has been done. First, OGG provides annotations for each of its genes. These annotations are in the form of a string with pairs (GO term, proof). The first integration is then to travel the string and add annotation links in the form of the structure presented previously between the gene and the GO terms present in TXPO.

The second integration is concerned with the annotations present in the gene ontology resource[3]. 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 the source UniProt [11]. To reach this goal, the annotations, alongside their proofs, associated with each GO term are obtained through the gene ontology REST API. Then, the corresponding triples are added to the KG following the same structure as previously.

The triples are stored in three different named graphs: one that contains the definition of the new entity and predicates, one for the integration of OGG and one last for the integration of UniProt.

4.4.2 GO-CAM Integration

The Gene Ontology Causal Activity Modeling (GO-CAM) [56] is a framework that introduces models to connect GO annotations (plus some information from other ontologies) following

[3] http://geneontology.org

a defined schema. These models are interesting to represent the causal relationships between different biological processes. In TEKG these models are used to represent relations between the different processes present in the KG. For example, a GO-CAM model could show that the hyper-function of a biological process positively regulates another process. These kinds of relationships allow toxicologists to track a toxic effect from its starting point to all the other elements that are indirectly affected.

To integrate GO-CAM into TEKG, we query first all the GO-CAM models related to the biological processes included in TXPO through their REST API. Then, we create a new class to contain all these models and link them to their corresponding biological processes with a newly created predicate “in GO-CAM model”.

4.4.3 Biological Pathways Integration

Biological pathways are sequences of actions that happen in a cell or organism. These pathways can involve many different types of processes, including chemical reactions, transport of substances across cells, gene expression, and other activities. They are therefore much more precise and complete than GO-CAMs. GO-CAMs can be used to have a larger overview of the interactions between biological entities and pathways, giving a deeper understanding of these interactions. Two well-known pathways repositories are Reactome [37] and Kegg [33].

To integrate pathways from both databases, we used associations from the Comparative Toxicogenomics Database (CTD) [14]. CTD brings together biological data by manually curating and linking information from published literature. It offers several files containing the relationships between different biological entities, and for this work, we used the disease-pathway association file.

The associations come in the format of a CSV file with five columns. The two first ones are the disease and its identifier in the Medical Subject Headings (MeSH)[4] thesaurus, the next two are the pathway name and its identifier in Kegg or Reactome, and the last one is the gene that is common to the disease and the pathway, which justify the link. To integrate the information in the CSV file, we first transform it into an RDF graph using RML, presented in Chapter 2. Then we integrate the newly formed RDF graph into TEKG by introducing the disease as instances of the class “disease” and pathways as instances of the class “pathway." These two classes are already defined in TXPO. Finally, links between pathways and biological processes are

[4]MeSH is a controlled vocabulary for biomedical information. https://www.nlm.nih.gov/mesh/meshhome.html

established using the Silk framework, also introduced in Chapter 2. As previously, the new instances and links associated with them are stored in dedicated (named) graphs.

4.5 Schema in Context

Now that the various components of the enrichment have been presented, the schema can be situated within a broader context, so we examine the potential connections that can be made with the TOXIN KG. An illustration of the KG's schema is represented in Fig. 4.3.

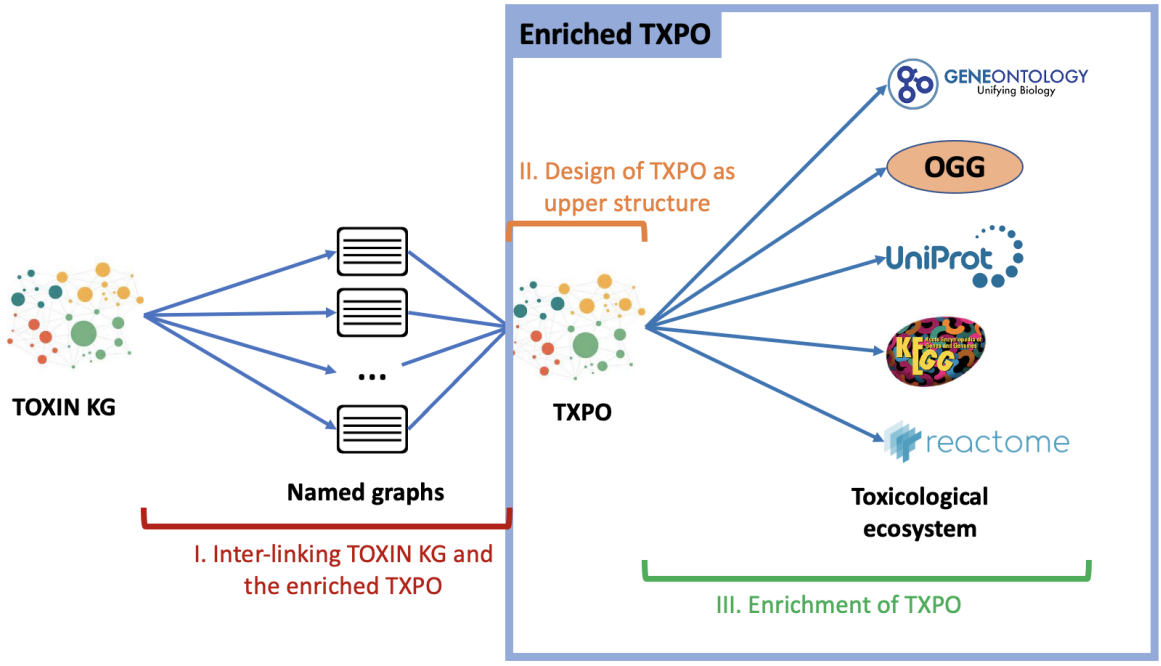


Fig. 4.3 Schema of TEKG with the relations between TOXIN KG and the enriched TXPO.

The schema starts from the TOXIN KG in the "machine-oriented" representation and is divided into three large sub-parts. The sub-parts II and III describe the creation of the enriched TXPO and are the ones that were presented previously. The sub-part one is concerned with linking the TOXIN KG and the enriched TXPO. Two kinds of links can be made between these two KGs. 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, it is possible to infer "indirect" links between the dossier and the toxic effects or diseases that were affecting 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 the links that it contains. In this way, it would be possible to query only some of these links depending on who made them and on what ground.

The links can be found manually. The direct links can be constructed by anybody as the task is simply to link the same observations in two data graphs but the "indirect" links need the intervention of domain experts, which have the knowledge to infer the toxic effects from the observations. Currently, observations are in the form of large paragraphs describing some components of the experimentation and the effects observed, and are thus difficult to use for an example. Therefore, to illustrate these links, a simplified example given by the domain expert is depicted in Fig. 4.4.

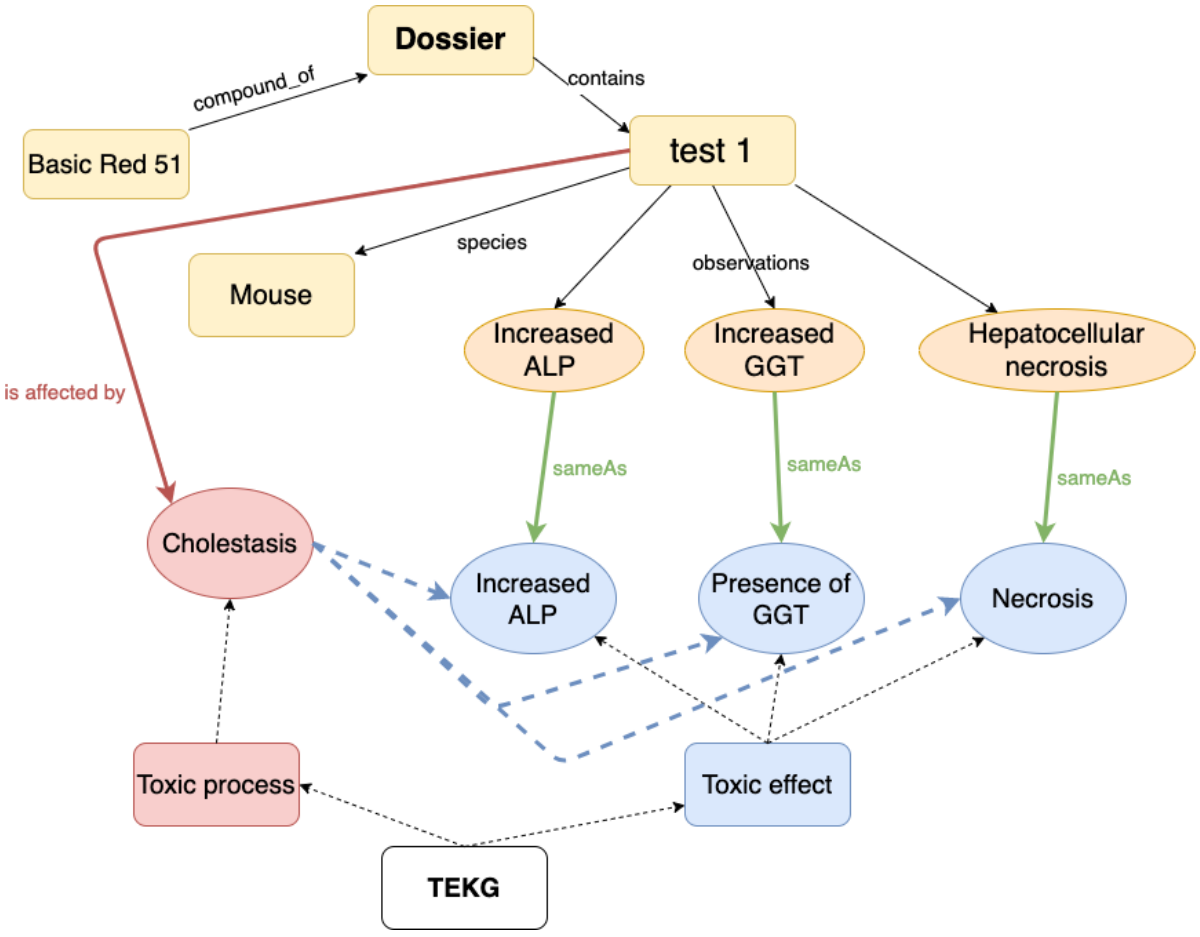


Fig. 4.4 Example of links between TOXIN KG and other resources within the TEKG.

At the top, the illustration shows a dossier, the compound being analyzed (Basic Red 51), and a test performed with this compound. The test has three observations, represented in orange. Some entities from the TEKG are also pictured at the bottom. The dashed arrows represent indirect relationships, which are included for clarity to avoid showing all the intermediate links. The direct links between the TOXIN observations and TEKG entities are represented in green, while the indirect links are shown in red. As we can see, the three observations have corresponding entities in the TEKG, even if they are not exactly the same. From these three observations, an expert might infer that the animal is experiencing cholestasis (shown by the red link). Finally, the blue arrows show the link between cholestasis and the three observations within the TEKG.

Another approach is to do these links automatically. For the direct links, it is straightforward. The task is to infer the relations `owl:sameAs` between TOXIN KG and TEKG, and tools exist to find these relations such as Silk and Alignment API presented in Chapter 2. Concerning the "indirect" links, some rule-based mechanisms can be put in place. This is not in the scope of this work but it is interesting for future work.

4.6 Structure of the Named Graphs and Provenance

This section explains the general structure of the different named graphs and the links between them. It also explains how the provenance is handled within and between them. A simplified structure is illustrated in Fig. 4.5.

In this figure, each rectangle (except the two scripts on the left) represents a (named) graph. The boxes that are outside the TEKG cylinder are data from other sources that have been integrated into it, and can therefore be accessed from TEKG. The image is divided into three parts: the TOXIN KG in blue that already existed before this project, the inter-linking part in grey, and the enrichment part in yellow.

The inter-linking part represents how the links between TOXIN KG and other TEKG entities are stored in different named graphs depending on the expertise at the origin of these links. The dashed grey lines illustrate these connections. The purple box on the top is there to store all the provenance information about the different named graphs containing the links. It uses the starting point category of PROV-O introduced Chapter 2 and keeps the information about who made the links and all the revisions of these links if any. The inter-linking script handles this part of TEKG and is responsible for making the links. It is expected to be used

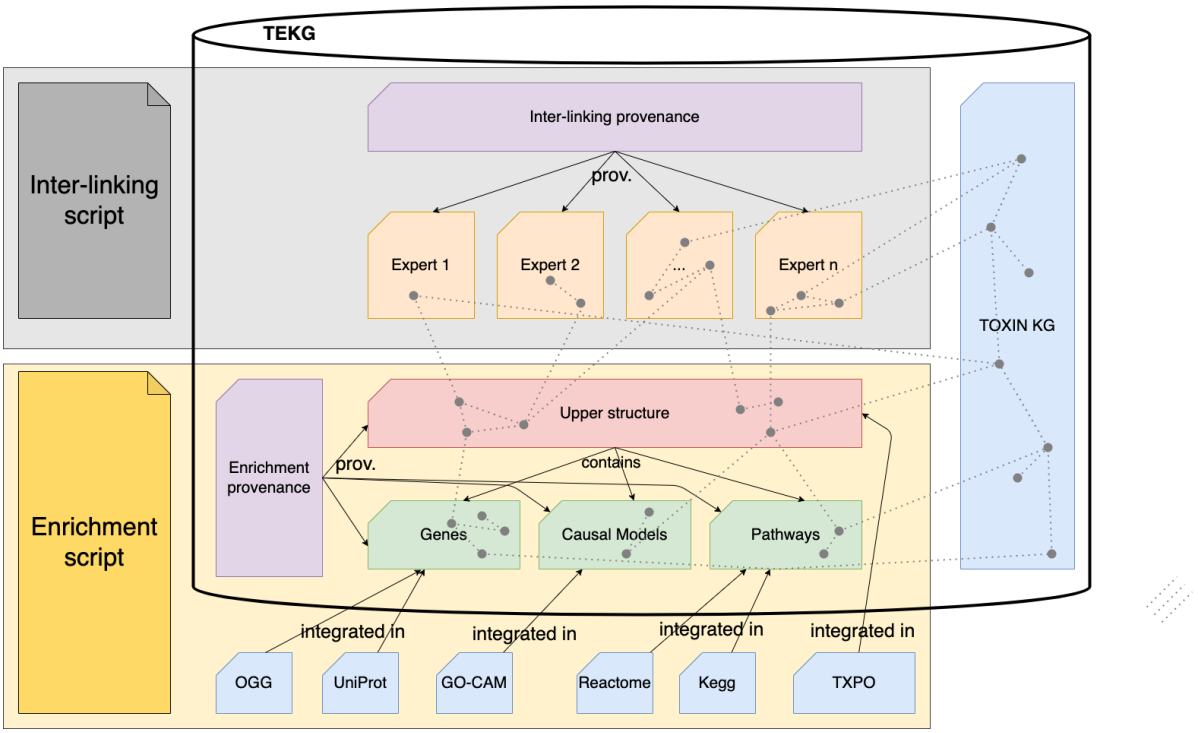


Fig. 4.5 Organization of the different named graphs within the database.

frequently as each action of adding/removing/modifying a link needs to use this script.

The enrichment part shows a simplified representation of the different named graphs in the core of TEKG (the enriched TXPO). The red box (the upper structure) contains all the links responsible for the integration and modification of TXPO. The green boxes contain the triples responsible for the integration of the three categories of information (genes, GO-CAMs and pathways). In the reality, each of these boxes is composed of a few named graphs (one for the definition of the new predicates and entities of this category, and one for each source), but for the sake of clarity, they are assembled into the three big categories on this image.

The purple box, here again, stores all the provenance information about the different integrations. As previously, the starting point category from PROV-O is used. This allows the tracking of the different activities that modify the KG. The provenance keeps track of who modified a named graph, and from what data were these modifications made. As an example, if only the provenance of TEKG is taken into account, the resulting provenance graph is represented in Fig. 4.6.

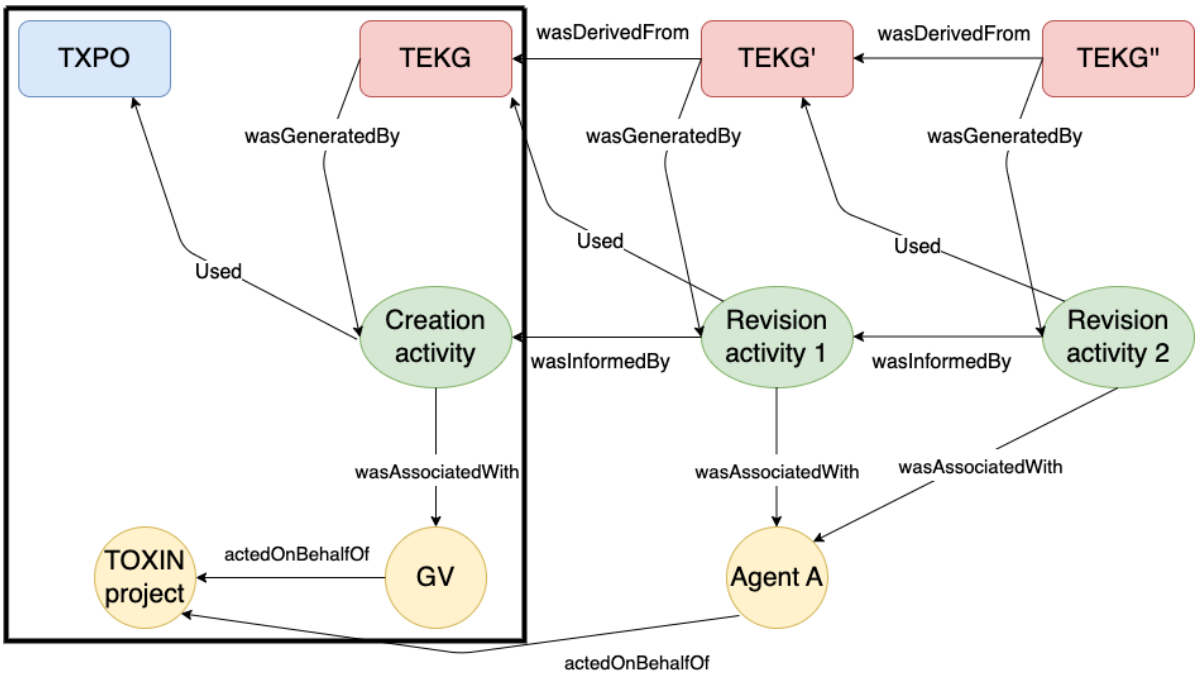


Fig. 4.6 Example of provenance information that could be stored in the provenance graph.

This image shows the creation of a named graph, TEKG, and its evolution in time. The information in the black rectangle is present at the creation of the KG and the rest of the schema shows how the information could evolve in time. The example shows that TEKG was created during the "creation" activity, which was handled by the agent "GV", and used the TXPO to be built. After that, each revision of the KG uses the previous version of it and is done by another agent. This can help with evaluating the reliability and credibility of the information and understanding the context in which it was generated.

4.7 Summary

The purpose of this chapter is to describe our contribution to the TOXIN project, which is to develop a method for integrating multiple data sources into a KG following a general structure, and link information from the resulting KG with the information contained in the TOXIN KG, which yield our TOXIN enriched knowledge graph. This method corresponds to schema development, data annotation and data lifting in the construction layer of the ARA architecture described in Section 2.3.1. These three tasks allow us to build a knowledge graph that enriches the data already present in the TOXIN project and helps researchers. For other applications, additional tasks may need to be considered in order to further develop the architecture, such as data quality assurance and data storage. This is out of the scope of this thesis but it will be

discussed for future work.

As it is the starting point of the thesis, we first described the TOXIN KG and outlined that its format is not well-adapted for automatic linkage or integration with other data sources. Then, we described a 5 steps methodology followed in this project for developing the method, which is: specifying the design of the KG, modeling the general structure, lifting data from other sources into it, publishing it, and finally curating the data for maintenance.

After that, we described the method, alongside the TOXIN Enriched Knowledge Graph (TEKG), the prototype built to demonstrate it. The method has two parts. The first one presents the creation of an upper structure with TXPO and the enrichment of the graph from this upper structure by the integration of several data sources. The second part describes the inter-linking process between TOXIN KG and the data integrated. Finally, the final structure of the KG and the handling of provenance information are also discussed.

The next chapter's goal is twofold: firstly, to evaluate the method presented through a demonstration with the prototype developed, and secondly to discuss the different steps of the design and compare them to other potential design choices.

Chapter 5

Evaluation

This chapter provides an evaluation of the method proposed in the previous chapter. In Design Science, evaluation is crucial because it allows researchers to assess the effectiveness, usefulness, and impact of the solution that has been developed. **??** For this thesis, the functionality of the method is evaluated through structural testing to ensure that it meets the requirements and goals of the design.

For this purpose, we will start with a small demonstration of possible queries that TEKG can answer, and then discuss the different parts of the method and spot their strengths and limitations, and how they meet the challenges presented in Section 1.2. As a reminder, the purpose of the method, along with the TEKG, is to propose a solution to facilitate access to relevant toxicological data and help toxicologists with hazard assessment.

5.1 Demonstration

To demonstrate the method's functionality, the idea is to show that the prototype can provide useful information to toxicologists in TOXIN. More specifically, the information should facilitate the hazard assessments of new compounds. The three competency questions defined for the specification step in Section 4.2 can fulfill this role as they were defined to respond to the primary needs of the domain experts. The answers to the competency questions are found and presented using Ontodia [38] in Fig. 5.1, 5.2 and 5.3. Note that Ontodia is a tool to interact with RDF graphs and is therefore merely a way to navigate knowledge graphs. Other tools to navigate and explore the knowledge graph are beyond the scope of this thesis.

Fig. 5.1 illustrates the answer to the competency question: "Knowing some adverse effects observed in a subject, what diseases or toxic processes may affect this subject?" At the

bottom of the image, different adverse effects that could be observed during a toxicological test are presented, 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.

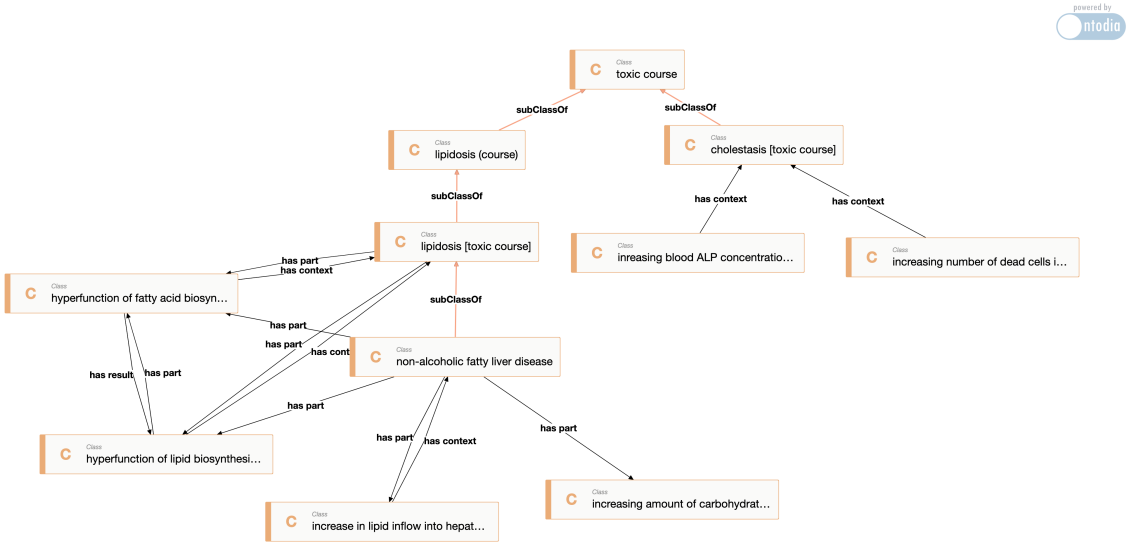


Fig. 5.1 Illustration of how TEKG can be used to answer the competency question "Knowing some adverse effects observed in a subject, what diseases or toxic processes may affect this subject?".

Moreover, the prototype also represents relationships between toxic processes. It is, therefore, possible to see what specific toxic courses may cause an adverse effect and the relationships between a specific toxic course and other ones. An example is shown on the left of the image, where three key pieces of information are represented:

- "NAFLD" is a subclass of a more general toxic course that is "lipidosis."
- The two increases are adverse effects that are specific to NAFLD, and the two hyperfunctions are adverse effects that NAFLD inherits from lipidosis, i.e., that are there because NAFLD is a special case of lipidosis.
- We also see the relations between the two hyperfunctions as their effects are linked.

Overall, the prototype, not only allows for finding toxic processes related to an adverse outcome, but also for the examination of the relationships between adverse effects, toxic effects, toxic courses, and diseases, enabling the identification of the characteristics of specific toxic

courses and their connections to one another.

Fig. 5.2 presents how could be answered the question "Which biological processes or pathways are affected by a certain disease?" It starts from the NAFLD on the left and links it to biological processes, pathways and GO-CAMs[1] on the right. The disease is linked to different pathways from different sources. The upper part shows the link between NAFLD and "fatty acid biosynthetic process", which was present in TXPO, and two GO-CAM models where this biological process takes part. The lower part shows the link between NAFLD and two other pathways from Reactome and Kegg. Finally, the upper and lower parts are related by the predicate `owl:sameAs` that was found with Silk to find related entities in the KG.

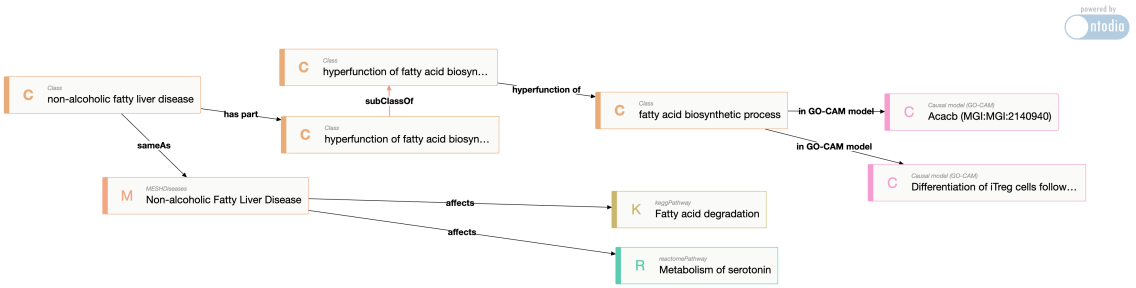


Fig. 5.2 Illustration of how TEKG can be used to answer the competency question "Which biological processes or pathways are affected by a certain disease?".

The image in Fig. 5.3 shows how from a biological process, it is possible to access the gene products playing a role in it. The origin of the gene products is either OGG or Uniprot. All these gene products are linked through an annotation following the schema of Fig. 4.2 in which the proof of the link is stored. The image demonstrates that it is possible to answer the question: "The functioning of what gene or protein is impaired by some toxic process?" by first finding the biological processes affected by the toxic process and then the genes involved in them.

During the integration process, we made the decision to retain the IRI of the authoritative resources whenever possible. This means that a majority of the entities in the knowledge graph are represented by an IRI that allows for access to additional information about them in the

[1] As a reminder, biological process, pathway and GO-CAM are similar concepts as their purpose is to describe biological interactions between biological entities. The biological process is the name given to a more or large phenomenon that is observed, a GO-CAM represents the basic interactions between biological entities within this biological process, and a pathway is a precise, complete map of the interactions that happen during the biological process.

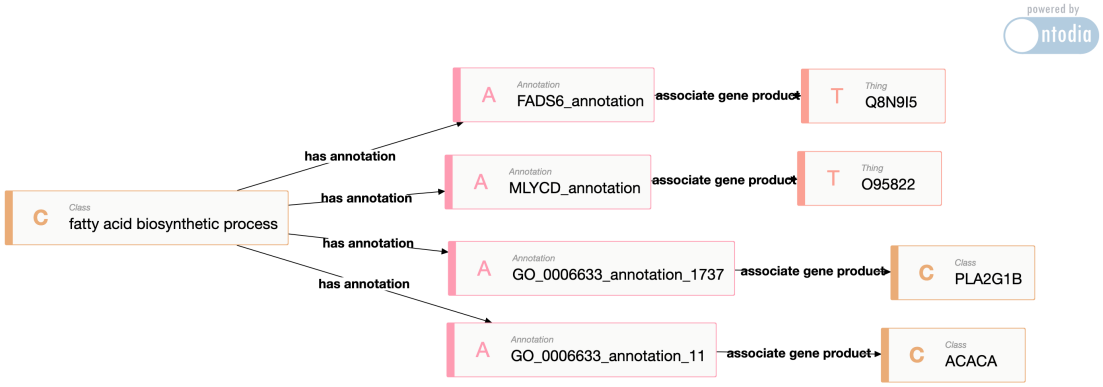


Fig. 5.3 Illustration of how TEKG can be used to answer the competency question "The functioning of what gene or protein is impaired by some toxic process?".

original database. An example of this can be seen in Fig. 5.4.

On the left of the figure, we can see three entities of TEKG represented with Ontodia. When the IRIs of these entities are searched on a browser, the pages on the right of the image appear. This illustrates that not only can we navigate through the entities in the knowledge graph, but we can also access the source data through the IRI representing an entity.

5.2 Discussion

The demonstration shows that TEKG fulfills the tasks it was built for. In this section, we will discuss in detail the design of the method with its advantages and its limitations.

5.2.1 Methodology Relevance

The methodology used in this project has several key advantages and some potential disadvantages that should be considered when evaluating its suitability for integrating and linking multiple data sources within the toxicological domain. One major advantage is the specification step to communicate with domain experts and incorporate their input into the design of the methodology, which helps to ensure its utility and relevance to the toxicological domain. This helps to meet the skill gap challenge by checking that the integration processes that are planned will be useful.

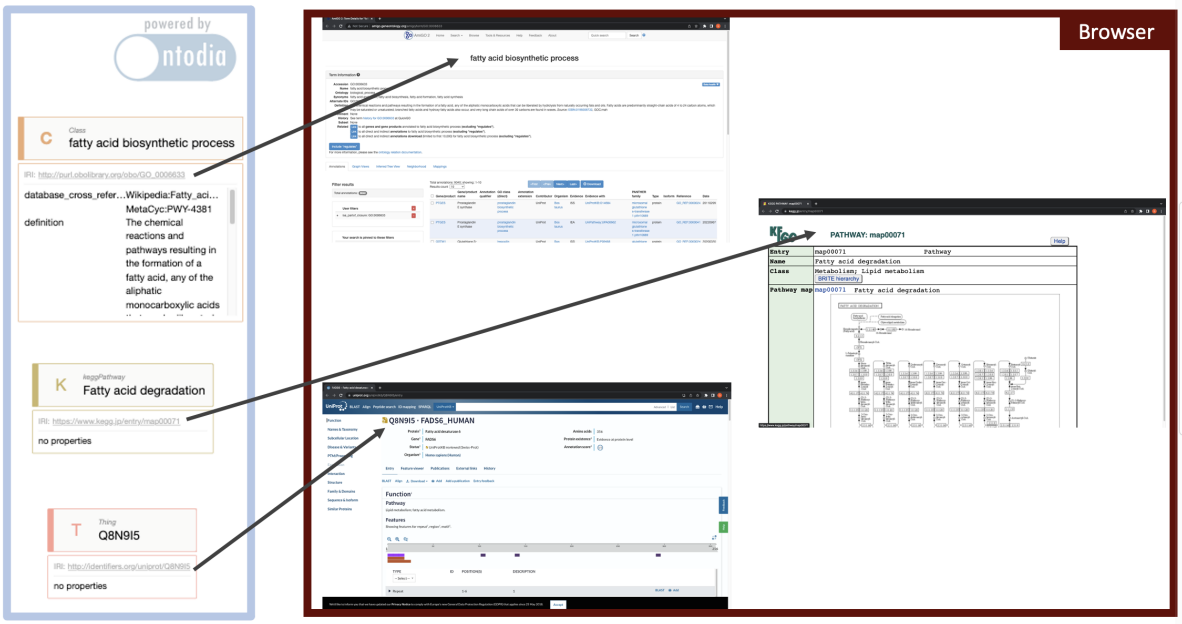


Fig. 5.4 Illustration of the possibility to access the external sources from which data was integrated into TEKG through the IRIs.

Another benefit of the cyclic methodology used in this project is that it allows for the continuous evaluation and improvement of the knowledge graph. By following a process of specification, modeling, and data integration, we can ensure that the current state of the KG is relevant and consistent, and that it maintains a clear structure. Additionally, the attention to data curation from an early stage will help later to ensure the quality and value of the data within the knowledge graph.

On the other hand, there are also some potential disadvantages to this methodology that should be considered. As the proposed method is specific to TOXIN, one potential disadvantage is the time and effort required to develop and implement the methodology to other use cases, which may not be feasible for all projects as it needs the intervention of a domain expert and a computer scientist for each integration. Additionally, the methodology may not be suitable for all types of data or all domains. For example, if the structure of a data source is too far from the one already defined for other sources, it may be too complex or time-consuming to implement.

5.2.2 Design of the Upper Structure

In this project, the upper structure of the knowledge graph was designed with an ontology to provide a general framework for integrating multiple data sources within the toxicological domain. This “ontology-based approach” helps to ensure that the data is organized in a consistent

and logical manner, making it easier to understand and use.

One alternative approach is to gather the different data sources together and combine their structures to create a Linked Data model. The advantage is that such models are designed to facilitate the integration and linking of data from multiple sources, which can be particularly useful in the toxicological domain where data is often dispersed across many different sources. They are also more flexible and adaptable than "ontology-based approaches".

However, with Linked Data models, the integration of large and diverse data sources into a model that captures all the necessary information and relationships in a meaningful way is challenging. Moreover, there is a risk that some similar entities may not be well linked, leading to a lack of understanding and meaning in the overall structure and a model prone to errors and inconsistencies. Therefore, our design of TEKG starts with an "ontology-based approach" to define the upper structure as it is more adapted to capture the complexity and granularity of the toxicological domain with fewer errors. After that, the different data sources are integrated into this general structure using the Linked Data approach to facilitate the integration and to have more flexibility.

While it is possible to build an ontology from scratch, it can be a time-consuming process, particularly in a domain as complex and nuanced as toxicology. Reusing existing ontologies, such as TXPO, is highly recommended [8] because it can help to reduce the time and effort required and ensure that the upper structure is well-founded and meaningful.

The primary purpose of the upper structure in this project is to provide a general structure for diseases, toxic courses, toxic effects and adverse effects. This corresponds to the purpose of searching for a toxic effect and then looking up all the information linked to it. Using TXPO rather than other ontologies or vocabularies such as MeSH (Medical Subject Headings) or MONDO (disease ontology) has several advantages in this project:

- TXPO is a comprehensive ontology specifically designed for the toxicological domain, which means it covers a wide range of concepts and relationships relevant to toxicology. This makes it a good choice for integrating and linking multiple data sources within the toxicological domain, as it provides a detailed and coherent structure for organizing the data.

- TXPO is a well-established and widely used ontology, with a large community of users and developers. This means that it is likely to be well-maintained and updated over time, which is important for ensuring the long-term reliability and usefulness of the knowledge graph. Moreover, TXPO is still under development, so even more interesting entities and relationships should be added to it over time.
- TXPO has been designed to be compatible with other ontologies and vocabularies (it is part of the OBO foundry), which makes it easier to integrate data from multiple sources and to link the knowledge graph with other datasets. This is particularly important in the toxicological domain, where data is often dispersed across different sources and may use different terminology and standards.

Concerning the structure of TXPO, it was originally designed as an ontology and thus has a traditional class and axiom-based structure. In order to use it as a knowledge graph, we treated its classes as objects and its axioms as predicates. This allowed us to represent the relationships between different entities in the toxicological domain more flexibly and to link them to external data sources.

One potential disadvantage of this approach is that it may not fully utilize the capabilities of an ontology and may not adhere to all of the best practices for designing ontologies. Additionally, we had to add our own IRIs for the new data that we integrated into the KG. This may be problematic for the re-usability of our KG. The question of the uniformization of the IRIs could be studied in future work.

5.2.3 Relevance of the Integrated Data

The integrations performed in this project were chosen for their relevance to the toxicological domain and their potential to provide valuable information for hazard identification and assessment. This relevance has been assessed in collaboration with the domain expert, by their capacity to provide information that could help to answer the competency questions. The integration of TXPO plays a central role as it helps to answer the three questions. Reactome, Kegg and CTD provide the additional information about pathways, and OGG and Uniprot offer information about human genes. These integrations should be viewed as a starting point, as the toxicological domain is vast, and a lot of other useful information could be integrated.

Integrating big data sources into the knowledge graph presents a number of challenges, especially in ensuring the accuracy, utility, and completeness of the data. In this project, relevance and utility to the toxicological domain of the integrations are assumed as they were chosen with

the help of a domain expert. However, the accuracy and completeness of the data remain a concern, as the toxicological domain is constantly evolving and it is difficult to thoroughly evaluate the quality of the data without a thorough examination by a domain expert. Ensuring the accuracy, utility, and completeness of the data is part of quality assurance over time and corresponds to a task of the first layer of the ARA architecture presented in Section 2.3.1. As explained earlier, it is outside of the scope of this thesis but it will be discussed for future work on this subject.

Finally, even with useful and accurate data in the KG, it does not guarantee that the knowledge graph brings value, as it is sometimes very difficult to access relevant data in the huge ecosystem. For example, after the different integrations in TEKG, the entity "cellular component" is linked to more than 1000 GO-CAMs and genes. It is expected as this entity is a very general one in the hierarchy but the larger the number of integrations, the larger the number of such problems. Therefore, an efficient interface to access relevant data should be developed in parallel to the integrations to have an efficient tool. This problem will be discussed for further work.

5.2.4 Usage of Named Graphs and Provenance

Using named graphs and incorporating provenance information is important for several reasons in this project. Firstly, during the interlinking process, it allows for the representation of different opinions or perspectives on the data. All the data associated with a particular opinion is stored in a dedicated named graph. By using named graphs to store data associated with specific opinions, it is possible to easily query and retrieve information linked to a specific opinion or to exclude information linked to a particular opinion. This allows for the representation of different viewpoints and interpretations of the data, and enables the effective handling of contradictions or inconsistencies by tracking their origin through the provenance information. In addition, if an automatic process is used to do some links, the resulting links can be stored in a dedicated named graph that can be easily accepted or disregarded later depending on the validity of the links found.

Secondly, as explained earlier, the utility and accuracy of the data is not well established, especially with the number of integrations growing up. By putting the data associated with each integration in dedicated named graphs and keeping their provenance information, it is easily possible to only take into account integrations that are interesting for a specific problem, or to exclude integrations where the information has errors or is not compatible with some other data.

Thirdly, incorporating provenance information can help with the evolution of the information integrated into the knowledge graph. As new data become available or existing data is updated, the provenance information can be used to track the changes and ensure that the information remains accurate and reliable. Overall, it is particularly useful when integrating data from multiple sources, as it allows for the preservation of the original context of the data and its evolution over time.

Finally, there are different possibilities for the management of provenance information. In this project, all the provenance information is gathered into two separate named graphs. One named graph for the interlinking process and one other for the enrichment process. This "general approach" allows having a general "provenance graph", which purpose is to represent the evolution of the data collected in the knowledge graph from its origin to its current state. For the prototype, this solution is appropriate as the data has not evolved much and there is therefore a small amount of provenance information.

However, imagining that the knowledge graph evolves over a long period, with a lot of different added data and contributors (it is especially the case for the interlinking process), some other management of the provenance information may be more efficient. One of these other management techniques is to create a new "provenance named graph" for each new update of a named graph. This "divided approach" is the method proposed by [15] where they define their structure with two layers: a deployed layer containing the up-to-date version of the named graphs, and a provenance layer with sequences of pairs named graph and corresponding provenance named graph, which gather the whole history of the content in the different named graphs.

The general approach has the advantage be the simplest and offers a general view of the provenance with an easy access to the provenance information. On the other hand, the divided approach offers a better separation of the data. As often, only the deployed layer is accessed regularly, while the provenance layer is rarely accessed, it is possible to handle the two parts separately in memory. However, with this approach, access to general information is more complex as the different provenance named graphs have to be accessed. Therefore, each approach has its advantages and its inconvenient and should be chosen depending on the purpose of the KG.

5.2.5 Reproducibility and Transparency of the Method

The reproducibility and transparency of the method, along with TEKG, are important aspects to consider. The method used in this project is designed to be reusable because it is modular

and flexible, which means that it can be applied to other similar domains or contexts as long as the necessary modifications are made. This can help to save time and resources, as the same approach can be used to integrate multiple data sources into a general structure in the toxicological domain.

Additionally, the method is designed to be transparent, which means that all of the steps and decisions made during the integration process are clearly documented through, for example, provenance information, and can be easily understood by others. This transparency helps to ensure that the method can be easily reproduced and evaluated by others, and helps to build confidence in the results obtained using the method.

5.3 Summary

The purpose of this thesis was to develop a method for integrating and linking multiple data sources within the toxicological domain, with the goal of facilitating access to relevant data and helping toxicologists with hazard assessment. The method proposed in this thesis, along with its prototype called TEKG, involves a structured approach to modeling and integrating data, with the involvement of domain experts to ensure its relevance and utility.

The functioning of TEKG was demonstrated, showing that it can provide useful answers to competency questions relevant to toxicologists. The evaluation of the method through structural testing showed that it meets the requirements and goals of the design. However, the method has some potential disadvantages, including the time and effort required to develop and implement it, as well as its potential inapplicability to certain types of data or domains.

The design of the upper structure of TEKG, using the TXPO ontology, allows for the organization of data in a consistent and logical manner, while also enabling the representation of different opinions or perspectives on the data through the use of named graphs and provenance information. The integrations of multiple data sources in TEKG add value to the knowledge graph by providing a more comprehensive view of the toxicological domain, and the use of named graphs and provenance information helps with the management of the evolution of the data and facilitates data curation. However, some data quality tools and an efficient interface should be developed in order to get as much value as possible from these integrations and not be overwhelmed by the amount of data.

Overall, the transparency and reproducibility of the method and prototype make it a promising solution for addressing the challenges of accessing and integrating toxicological data. In the following chapter, we will present the final version of TEKG and discuss its potential applications and uses within the toxicological domain.

Chapter 6

Potential Applications

The previous chapter provided a demonstration and a discussion of the work conducted in this thesis. We demonstrated the links that have been created from toxic processes to gene products potentially affected by them, and we discussed the different design choices made during the development of TEKG. In this chapter, we will explore the potential applications and uses of the TEKG.

As described previously, TEKG is a structured representation of toxicological data that integrated and interlinked multiple sources to facilitate access to relevant information and support hazard assessment. This chapter will analyze the ways in which TEKG can be utilized in the toxicological domain. More specifically, we will discuss the potential of TEKG as a search tool, a bridge for putting in relation in vivo and in vitro data through common biological knowledge, and as a domain ontology for toxicological text annotation.

6.1 Search Tool

The primary use of TEKG is as a component for a search tool for toxicologists and other stakeholders interested in accessing relevant information about toxicological processes, diseases, and effects. By organizing and linking data from multiple sources in a structured and logical manner, TEKG provides a comprehensive and easily navigable resource for finding relevant information about a wide range of topics within the toxicological domain. Currently, the demonstration provided in this thesis used the Ontodia library [38] to interact with the KG and show its functionalities. The TOXIN project is conducting research for the development of a domain-specific search tool for improving the consumption of the data in their knowledge graph system [48]. TEKG could be a valuable resource in the development of this tool.

TEKG “centralizes” information by carefully integrating and interlinking pieces of different sources. TEKG, therefore, serves as a centralized tool that allows users to find more easily relationships between various data sources and access these sources from a single endpoint instead of manually searching through multiple databases. This was demonstrated in the previous chapter by making a link from the disease NAFLD to the genes affected by it and by showing how it is possible to access the data about these genes in the authoritative resources with their IRIs.

Additionally, by incorporating provenance information and using named graphs to represent different opinions or viewpoints, TEKG allows users to filter results based on the source and reliability of the data, and makes the search very flexible. In the vast toxicological ecosystem, finding specific information and related data can be like looking for a needle in a haystack. TEKG will be a vital tool for helping toxicologists in this task.

6.2 Bridge between In Vivo and In Vitro Testing

Another potential application of TEKG is as a bridge for putting in relation in vivo and in vitro testing through common biological knowledge. During the inter-linking step, the in vivo data, such as toxic effects observed in toxicology tests on animals, is linked directly or indirectly to toxic processes, (natural) biological processes affected and genes involved. This allows for a deeper understanding of the mechanisms behind the observed adverse effects and the conclusions about toxicity in the in vivo tests.

The data gathered in TEKG can be categorized into three large domains: toxic processes, natural processes, and modules. The modules represent the different elements that participate in a process, such as gene products, molecules, and chemical compounds. To illustrate these three domains, an illustration of a small subpart of TEKG is presented in Fig. 6.1 using Ontodia.

The toxicological processes are on the left, the natural processes are in the middle, and the modules are on the right. Generally, an in vivo test will be linked to one or more toxic processes observed on the animal. For example, a link between a test T and the toxic process "lipidosis" could be made. This link implies that during T, the fatty acid biosynthesis process may have been affected and the production of the two gene products Q9UGI9 and G8N9I5 may have been affected too. These links are more complex in reality, but this example offers a simplified way to illustrate the kind of relations that TEKG can represent (and thus find with graph-query languages) for in vivo data. These relationships can also be used the other way

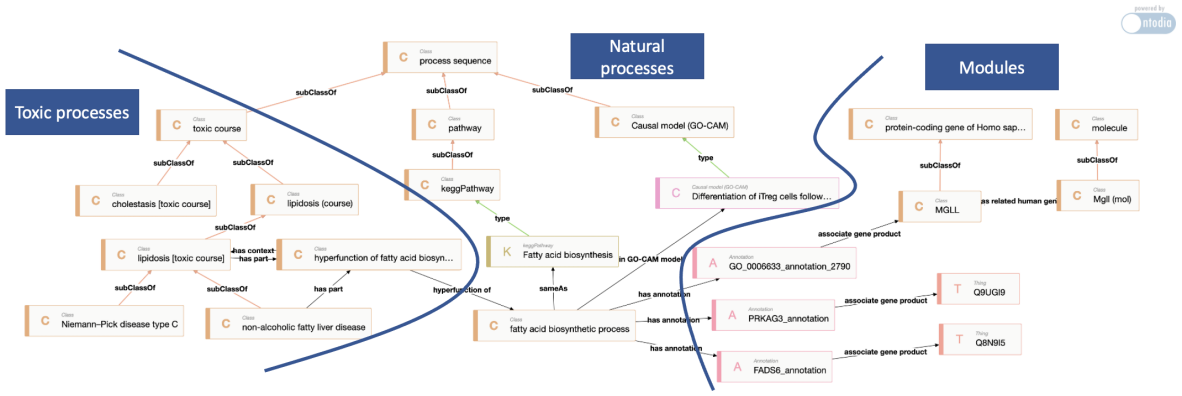


Fig. 6.1 Subpart of TEKG represented with Ontodia and separation of the entities into three categories.

around by for example querying all the toxic processes that affect a targeted gene, and even the toxicological tests where this gene was potentially affected.

In vitro tests examine the effects of a compound on specific tissues in a specific organ. Therefore, the toxicity observed in vitro is often at another level than the one observed during in vivo tests. For example, a toxic effect observed in vitro could be the overproduction of some enzymes or gene products such as Q9UGI9 or G8N9I5. Using TEKG, we can see how the effects observed in an in vitro test may be related to toxic processes observed in in vivo tests, providing a more comprehensive and contextualized understanding of the data. By linking these two types of data, TEKG helps to bridge the gap between in vitro and in vivo testing and gain a deeper understanding of the underlying processes behind toxic effects. This deeper understanding should help domain expert to build new methods for hazard assessment using in vitro testing.

6.3 Text Annotation Tool

One potential application of TEKG is in the field of text annotation for toxicological documents. Text annotation is the process of labeling or marking specific words or phrases in a text. It is often used in natural language processing (NLP) tasks, such as information extraction, machine translation, and text classification, to provide context and structure to the text being analyzed.

Several techniques exist to do text annotation. The one that is of interest for this project is ontology-based data annotation for which different techniques exist. For example, [36] proposes the tool OnTeA, which creates a semantic version of text documents based on a domain

ontology, and [18] proposes a method for automatically annotating text segments in a document with concepts from a domain ontology. Overall, ontology-based data annotation allows for easier searching and information retrieval within the text and enables the identification of relationships and connections between different concepts within the text.

The tools above can help identify and tag relevant terms within a text using an ontology explicitly designed for the toxicological domain. TEKG can serve as such an ontology, providing a common vocabulary and structure for organizing data and ensuring a standardized and consistent approach to text annotation. Using TEKG for text annotation could significantly improve the effectiveness and efficiency of information retrieval and analysis within the toxicological domain.

6.4 Summary

To summarize, we described TEKG's potential in three contexts. One potential use is as a component in search tools, which provides a centralized and easily navigable resource for finding relevant information in the toxicological domain. It also allows users to filter results based on the source and reliability of the data.

Another potential application is as a bridge for linking in vivo and in vitro data through common biological knowledge. Overall, TEKG helps researchers understand the mechanisms behind the adverse effects observed during in vivo tests and develop new hazard assessment methods through in vitro testing.

Finally, TEKG can be used as a domain ontology for toxicological text annotation, enabling the accurate and efficient tagging of relevant terms within a text. This can improve the effectiveness and efficiency of information retrieval and analysis within the toxicological domain.

Overall, this section explored the possibilities and potential of using TEKG in the toxicology domain. The results of this project demonstrate the potential value and usefulness of a knowledge graph such as TEKG for improving access to relevant information and supporting hazard assessment. This project achieved its goal of creating a prototype (both the method and resulting KG) and exploring its potential applications and benefits. Even though the potential uses were beyond the scope of this thesis, TEKG may serve as a foundation for future development and research in the TOXIN project.

Chapter 7

Conclusion

In this final chapter, we will summarize the work accomplished in this thesis and highlight our most significant achievements. We will also provide an overview of potential future work on the subject.

7.1 Summary

Current methods for chemical safety assessment often involve the use of in vivo testing and categorization tools that are based on previous toxicology tests. However, in order to comply with the 3R principle and European regulations, there is a need for new approaches to chemical safety assessment based on in vitro or in silico testing. To achieve this, it is important to have a greater understanding of the underlying toxicological processes in order to better relate and interpret in vitro, in silico, and in vivo test results and develop more effective hazard assessment methods.

The purpose of the method and prototype described in this thesis is to create a knowledge graph that integrates multiple data sources within the toxicology domain and links this information with the TOXIN KG, which gathers in vivo data from previous tests. The KG is intended to provide a centralized, structured resource that can be used to store and access integrated toxicological information, and to facilitate the development of new methods for hazard assessment. The prototype, known as the TOXIN Enriched Knowledge Graph (TEKG), is a proof-of-concept implementation of this method that demonstrates its feasibility and potential utility.

The method addresses several major challenges. Firstly, to address the problem of granularity in toxicology, the method uses the TXPO ontology, which provides a structured and reliable representation of the relationships between toxicological processes at various levels

of granularity. Secondly, the method is based on a cyclic methodology that provides support for the constant verification and evaluation of the validity and usefulness of the development with the help of a domain expert. This helps to bridge the skills gap between domain experts and IT experts who develop the KG. Finally, the challenges of handling a large amount of data being integrated and potential incoherences or discrepancies between resources are addressed through the use of named graphs and provenance information.

TEKG is reproducible, transparent, and can be used as a component for a domain-specific search tool, a way to better understand and link in vivo, in vitro, and in silico data, and as a base for a tagging mechanism to find relevant toxicological concepts in a text. Overall, the results of this project demonstrate the potential value and usefulness of TEKG for improving access to relevant information and supporting hazard assessment.

7.2 Achievements

In this thesis, we have **developed a method** for integrating multiple toxicological data sources and linking them with the TOXIN knowledge graph to facilitate the hazard assessment of new compounds. Our key contributions include **demonstrating the benefits** of using a hybrid approach that starts with an ontology, and afterward utilizes Linked Data to integrate different data sources. This approach allows for the **capture of granularity** in the toxicological domain and **provides a consistent representation of the domain** through the ontology while retaining the flexibility of Linked Data. We have shown that the TXPO ontology is well-suited for this task and should be considered by others pursuing similar projects.

Additionally, we have **demonstrated the usefulness of using named graphs and provenance information** for storing data separately and keeping track of the origin of each set of data. This enables us to easily access specific parts of the complete data as needed, making it more adaptable. Additionally, by linking errors to specific sources or only using trusted subsets of data, it becomes easier to track and correct any inconsistencies. Finally, we have **demonstrated the feasibility and utility of our approach** to help toxicologists with hazard assessment, and this work could be used as the foundation for further development and research in the field of toxicology. Possible future development will be discussed in the next section.

Working on this project has been a great opportunity to collaborate with a real-life research team and to be faced with the challenges of understanding domain needs and proposing effective solutions. It has allowed us to improve our communication skills and work on all aspects

of building a tool, from design to implementation. It has been a rewarding and challenging experience.

7.3 Future Work

While TEKG serves as a proof-of-concept prototype, there is still significant work to be done in order to fully realize the potential of the approach. In particular, the next steps in the development of TEKG should focus on building a robust and scalable knowledge graph system that can handle an increasing volume of data and generate a larger amount of relevant links.

To build a strong knowledge graph system, it is necessary to address all three layers of the ARA architecture presented in Chapter 2, knowingly knowledge base construction, knowledge storage and knowledge consumption. While the work presented in this thesis primarily focused on the knowledge base construction, and in particular on schema development, data lifting and data annotation, further development is needed to ensure the quality of the data being integrated, as well as efficient storage and consumption mechanisms. This will allow the system to effectively handle large volumes of data and provide more useful insights for toxicological research, which will be essential for improving the accuracy and reliability of hazard assessment.

To ensure the quality and reliability of the knowledge graph system, various quality assessment tools can be developed. These could include mechanisms for verifying the accuracy and consistency of data, as well as techniques for detecting and correcting errors or inconsistencies. Chapter 7 of [29] presents an overview of tools that could be used for this purpose. This could involve the use of automated checks and validation processes, as well as manual review and verification by domain experts. In addition, it may be necessary to develop approaches for dealing with missing or incomplete data, and for handling multiple conflicting sources of information. The separation of the KG into several named graphs and the creation of provenance information should greatly facilitate this task.

With the increasing amount of data integrated into the KG and the corresponding increasing amount of provenance information, the issue of data storage should be considered. One option to address this problem is to design some filtering tool that makes it possible to only store relevant "metadata" about the different sources, while the core of the biological data is kept in the original sources. For provenance and updates history, some research for a better organization and separation the data could be considered. For example, [15] proposes an organization where the data is separated into a deployed layer with the up-to-data information and a

provenance layer with the previous versions of the data along with their provenance information.

As the knowledge graph grows in size and complexity, it will be important to develop efficient interfaces for querying and accessing the data (knowledge consumption). In particular, tools such as [27] that can help users find relevant relationships between entities in an RDF graph could be useful for developing such interfaces. In order to fully realize the potential of the knowledge graph system in toxicology, it will be necessary to continue developing and refining these types of tools into a user-friendly interface.

Once the KGS is well built and robust, more information could be integrated into it. For the moment, the data in the KG is focused on humans, i.e., we included information on human genes, human processes, etc. One potential area for future development of the KGS is the integration of data related to different species and the differences between their respective biological processes, in order to increase precision. Another possibility is the integration of adverse outcome pathways (AOPs), which are structured organizations of existing knowledge illustrating causal pathways from the initial molecular perturbation triggered by various stressors, through key events at different levels of biology, to the ultimate health or ecotoxicological adverse outcomes. [32] developed a tool to build such AOPs from medical publications and state that they provide valuable support for various needs in risk assessment.

Another possible improvement for the KG would be to enhance the mechanisms for linking entities from different sources together, as well as linking TOXIN KG test data with toxic effects. While these links can be established manually by a domain expert, it is a time-consuming process. Alternatively, rule-based mechanisms or machine learning algorithms could be employed to make these links semi-automatically and even to identify sets of evidence that demonstrate toxicological effects based on the toxicological data in the KG. To achieve this task, different tools already exist and could be adapted for this project. [35] gives an overview of possible tools in this area.

References

[1] R. Arp, B. Smith, and A. D. Spear. *Building Ontologies with Basic Formal Ontology*. The MIT Press, 2015. ISBN 9780262527811. URL http://www.jstor.org/stable/j.ctt17kk7vw.

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

[3] Willem B. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. PhD thesis, University of Twente, Netherlands, 1997.

[4] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5):706–716, 2008. ISSN 1532-0464. doi: 10.1016/j.jbi.2008.03.004.

[5] C. Bendtsen and S. Petrovski. How data and AI are helping unlock the secrets of disease. Accessed Jan. 3, 2023 [Online]. URL https://www.astrazeneca.com/what-science-can-do/topics/data-science-ai/how-data-and-ai-are-helping-unlock-the-secrets-of-disease.html.

[6] J.-L. Binot. Semantic data. [Lecture notes], 2021. URL https://people.montefiore.uliege.be/binot/.

[7] R. R. Boyles, A. E. Thessen, A. M. Waldrop, and M. A. Haendel. Ontology-based data integration for advancing toxicological knowledge. *Current Opinion in Toxicology*, 16: 67–74, 2019. ISSN 2468-2020. doi: 10.1016/j.cotox.2019.05.005.

[8] V. A. Carriero, M. Daquino, A. Gangemi, A. G. Nuzzolese, S. Peroni, V. Presutti, and F. Tomasi. The Landscape of Ontology Reuse Approaches. In *Applications and Practices in Ontology Design, Extraction, and Reasoning*. IOS Press, 2020. doi: 10.3233/SSW200033.

[9] V. K. Chaudhri, C. Baru, N. Chittar, X. L. Dong, M. Genesereth, J. Hendler, A. Kalyanpur, D. B. Lenat, J. Sequeda, D. Vrandečić, and K. Wang. Knowledge graphs: Introduction, history, and perspectives. *AI Magazine*, 43(1):17–29, 2022. doi: 10.1002/aaai.12033.

[10] J. Cheney, L. Chiticariu, and W.-C. Tan. Provenance in Databases: Why, How, and Where. *Foundations and Trends in Databases*, 1(4):379–474, 2009. doi: 10.1561/1900000006.

[11] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa1100.

[12] Council of European Union. Directive 2010/63/eu of the european parliament and of the council of 22 september 2010 on the protection of animals used for scientific purposes, 2010. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32010L0063.

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

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

[15] C. Debruyne, G. Munnelly, L. Kilgallon, D. O'Sullivan, and P. Crooks. Creating a Knowledge Graph for Ireland's Lost History: Knowledge Engineering and Curation in the Beyond 2022 Project. *J. Comput. Cult. Herit.*, 15(2), 2022. ISSN 1556-4673. doi: 10.1145/3474829.

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

[17] A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, and R. Van de Walle. RML: a generic language for integrated RDF mappings of heterogeneous data. In *Proceedings of the 7th Workshop on Linked Data on the Web*, volume 1184, 2014. URL http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf.

[18] S. R. El-Beltagy, M. Hazman, and A. Rafea. Ontology Based Annotation of Text Segments. In *Proceedings of the 2007 ACM Symposium on Applied Computing*, page 1362–1367. Association for Computing Machinery, 2007. ISBN 1595934804. doi: 10.1145/1244002.1244296.

[19] Agency for Toxic Substances and Disease Registry. Module 3 - Risk Assessment. *Toxicology Curriculum for Communities Trainer's Manual*, 2015. URL https://www.atsdr.cdc.gov/training/toxmanual/modules/3/index.html.

[20] Apache Software Foundation. Apache Jena Fuseki, 2022. URL https://jena.apache.org/.

[21] M. E. Gallagher. Toxicity testing requirements, methods and proposed alternatives. *Environs*, 26(2):253–273, 2003. URL https://environs.law.ucdavis.edu/volumes/26/2/gallagher.pdf.

[22] T.R. Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5):907–928, 1995. ISSN 1071-5819. doi: 10.1006/ijhc.1995.1081.

[23] M. Grüninger and M. S. Fox. *The Role of Competency Questions in Enterprise Engineering*, pages 22–31. Springer US, Boston, MA, 1995. ISBN 978-0-387-34847-6. doi: 10.1007/978-0-387-34847-6_3.

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

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

[26] Y. He, Y. Liu, and B. Zhao. OGG: a Biological Ontology for Representing Genes and Genomes in Specific Organisms. *ICBO*, pages 13–20, 2014.

[27] P. Heim, S. Hellmann, J. Lehmann, S. Lohmann, and T. Stegemann. RelFinder: Revealing Relationships in RDF Knowledge Bases. pages 182–187, Berlin/Heidelberg, 2009. Springer. doi: 10.1007/978-3-642-10543-2_21.

[28] A.R. Hevner, S.T. March, J. Park, and S. Ram. Design Science in Information Systems Research. *Management Information Systems Quarterly*, 28:75–105, 2004. doi: 10.2307/25148625.

[29] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutiérrez, S. Kirrane, José E. Labra G., R. Navigli, S. Neumaier, A.-C. Ngonga Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J.F. Sequeda, S. Staab, and A. Zimmermann. *Knowledge Graphs*. Number 22 in Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool, 2021. ISBN 9781636392363. doi: 10.2200/S01125ED1V01Y202109DSK022.

[30] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean. SWRL: A semantic web rule language combining OWL and RuleML. W3C Recommendation, W3C, 2004. URL http://www.w3.org/Submission/SWRL/.

[31] IBM Cloud Learn Hub. Data Modeling. Accessed Dec. 6, 2022 [Online]. URL https://www.ibm.com/cloud/learn/data-modeling#toc-types-of-d-NeQpje_a.

[32] F. Jornod, T. Jaylet, L. Blaha, D. Sarigiannis, L. Tamisier, and K. Audouze. AOP-helpFinder webserver: a tool for comprehensive analysis of the literature to support adverse outcome pathways development. *Bioinformatics*, 38(4):1173–1175, 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab750.

[33] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1): D353–D361, 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw1092.

[34] B. Kasenchak and A. E. Lehnert. Introduction to ontology concepts and modeling. Accessed Dec. 6, 2022 [Online], 2021. URL https://boxesandarrows.com/introduction-to-ontology-concepts-and-modeling/.

[35] H. Köpcke and E. Rahm. Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, 69(2):197–210, 2010. ISSN 0169-023X. doi: 10.1016/j.datak.2009.10.003.

[36] M. Laclavik, M. Šeleng, E. Gatial, Z. Balogh, and L. Hluchý. Ontology based Text Annotation - OnTeA. In *Information Modelling and Knowledge Bases XVIII*, volume 154, pages 311–315, 2006. URL https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d1ba15ad9d4689f5a7ebccf8249a5c73ec26203e.

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

[38] D. Mouromtsev, D. Pavlov, Y. Emelyanov, A. Morozov, D. Razdyakonov, and M. Galkin. The simple web-based tool for visualization and sharing of semantic data and ontologies. In *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*, volume 1486 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015. URL http://ceur-ws.org/Vol-1486/paper_77.pdf.

[39] Digital Repository of Ireland. Dublin Core and the Digital Repository of Ireland v.3. 2016. doi: 10.7486/DRI.2z119b06h.

[40] Oracle. What Is Data Management? Accessed Dec. 6, 2022 [Online]. URL https://www.oracle.com/be/database/what-is-data-management/#data-management-best-practices.

[41] J. Pan, G. Vetere, J.M. Gomez-Perez, and H. Wu. *Exploiting Linked Data and Knowledge Graphs in Large Organisations*. Springer Cham, 2017. ISBN 978-3-319-45652-2. doi: 10.1007/978-3-319-45654-6.

[42] I. Papatheodorou, A. Oellrich, and D. Smedley. Linking gene expression to phenotypes via pathway information. *Journal of Biomedical Semantics*, 6(1):17, 2015. ISSN 2041-1480. doi: 10.1186/s13326-015-0013-5.

[43] P. Peregrine, R. Brennan, T. Currie, K. Feeney, P. Francois, P. Turchin, and H. Whitehouse. Dacura: A New Solution to Data Harvesting and Knowledge Extraction for Archaeology. *Historical Methods*, 2017. doi: 10.13140/RG.2.2.35803.26405.

[44] E. Prud'hommeaux and G. Carothers. RDF 1.1 Turtle. W3C Recommendation, W3C, 2014. URL https://www.w3.org/TR/2014/REC-turtle-20140225/.

[45] K. Rajpoot, N. Desai, H. Koppisetti, M. Tekade, M.C. Sharma, S.K. Behera, and R.K. Tekade. Chapter 14 - In silico methods for the prediction of drug toxicity. In *Pharmacokinetics and Toxicokinetic Considerations*, volume 2 of *Advances in Pharmaceutical Product Development and Research*, pages 357–383. Academic Press, 2022. ISBN 978-0-323-98367-9. doi: 10.1016/B978-0-323-98367-9.00012-3.

[46] S. Rudolph, M. Krötzsch, and P. Hitzler. Description Logic Reasoning with Decision Diagrams: Compiling SHIQ to Disjunctive Datalog. *International Semantic Web Conference*, pages 435–450, 2008. doi: 10.1007/978-3-540-88564-1_28.

[47] S. Sahoo, T. Lebo, and D. McGuinness. PROV-O: The PROV Ontology. W3C Recommendation, W3C, 2013. URL https://www.w3.org/TR/2013/REC-prov-o-20130430/.

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

[49] G. Schreiber and M. Dean. OWL Web Ontology Language Reference. W3C Recommendation, W3C, 2004. URL https://www.w3.org/TR/2004/REC-owl-ref-20040210/.

[50] A. Seaborne and S. Harris. SPARQL 1.1 Query Language. W3C Recommendation, W3C, 2013. URL https://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

[51] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. Goldberg, K. Eilbeck, A. Ireland, C. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. Scheuermann, N. Shah, P. Whetzel, and S. Lewis. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25: 1251–1255, 2007. doi: 10.1038/nbt1346.

[52] R. Stoney, D.L. Robertson, G. Nenadic, and J.-M. Schwartz. Mapping biological process relationships and disease perturbations within a pathway network. *npj Systems Biology and Applications*, 4(1):22, 2018. ISSN 2056-7189. doi: 10.1038/s41540-018-0055-2.

[53] R. Studer, R.V. Benjamins, and D. Fensel. Knowledge engineering: Principles and methods. *Data Knowledge Engineering*, 25(1):161–197, 1998. ISSN 0169-023X. doi: 10.1016/S0169-023X(97)00056-6.

[54] S. Sundara, R. Cyganiak, and S. Das. R2RML: RDB to RDF Mapping Language. W3C Recommendation, W3C, 2012. URL https://www.w3.org/TR/2012/REC-r2rml-20120927/.

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

[56] P. Thomas, D. Hill, H. Mi, D. Osumi-Sutherland, K. Auken, S. Carbon, J. Balhoff, L.-P. Albou, B. Good, P. Gaudet, S. Lewis, and C. Mungall. 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. doi: 10.1038/s41588-019-0500-1.

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

[58] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. 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*, volume 538 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009. URL http://ceur-ws.org/Vol-538/ldow2009_paper13.pdf.

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

[60] Y. Yamagata and H. Yamada. 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. doi: 10.1038/s41598-020-71370-7.

[61] C. Yang, M.T.D. Cronin, K.B. Arvidson, B. Bienfait, S.J. Enoch, B. Heldreth, B. Hobocienski, K. Muldoon-Jacobs, Y. Lan, J.C. Madden, T. Magdziarz, J. Marusczyk, A. Mostrag, M. Nelms, D. Neagu, K. Przybylak, J.F. Rathman, J. Park, A-N Richarz, A.M. Richard, J.V. Ribeiro, O. Sacher, C. Schwab, V. Vitcheva, P. Volarath, and A.P. Worth. COSMOS next generation – A public knowledge base leveraging chemical and biological data to support the regulatory assessment of chemicals. *Computational Toxicology*, 19, 2021. ISSN 2468-1113. doi: 10.1016/j.comtox.2021.100175.

[62] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma. Human symptoms–disease network. *Nature Communications*, 5(1), 2014. ISSN 2041-1723. doi: 10.1038/ncomms5212.

Appendix A

Internship with the TOXIN project

This thesis was coupled with an internship with the TOXIN project conducted at the Web & Information Systems Engineering (WISE) Lab[1], which is a research unit of the Vrije Universiteit Brussel (VUB) that focuses on innovative information systems and human-computer interaction. The internship was part of the TOXIN project[2], which aims to establish non-animal based, human-relevant strategies to assess repeated dose toxicity. This project brings together expertise in in vitro experimental toxicology (In Vitro Toxicology and Dermato-Cosmetology department), computational information systems (WISE lab), and ethics and legislation (Law, Science, Technology & Society department).

Currently, the role of the WISE Lab is to develop a computational information system for the TOXIN project. Specifically, they want to create an ontology-based knowledge base for safety data of cosmetic compounds. This knowledge base will be at first composed of existing safety evaluations issued by the Scientific Committee on Consumer Safety[3]. The knowledge graph should be built and later accessible using a user-friendly tool that is based on the jigsaw metaphor to hide the technicalities of semantic technology. This tool will be used to support toxicologists in constructing and accessing the knowledge base.

During my internship, I focused on finding and analyzing external resources in the toxicological domain that could be adapted to bring added value to the TOXIN project. My first task was to gain a thorough understanding of the research domain and the challenges of the project, and then to locate relevant information that was accessible and could be integrated

[1] https://wise.vub.ac.be
[2] https://wise.vub.ac.be/project/toxin
[3] https://health.ec.europa.eu/scientific-committees/scientific-committee-consumer-safety-sccs_en

into the project. My first accomplishment during the internship was the transformation of the OpenToxipedia vocabulary from OpenTox [55] into an RDF graph, which purpose is to clearly define the terms used in the knowledge base developed by TOXIN. This RDF graph has finally been utilized as domain-specific information added to the "toxicological Tickle tool" for learning purposes in the toxicological domain. The Tickle tool[4] has been originally developed to create a variety of innovative tools to engage young people in learning and to determine whether these tools can help to lower the rate of school dropouts. The "toxicological Tickle tool" is simply the adaptation of this tool for the toxicological domain.

The internship was the foundation for my thesis work, providing me with the opportunity to define the problem and develop the method for my research. It involved regular meetings, both with the WISE research team and with the consortium where I was asked to present updates on my research. I also participated to the Knowledge Graphs For Data Integration (KG4DI)[5] kick-off session, where I presented a poster about my work. The poster is represented in Fig. A.1. Overall, this internship provided me with the opportunity to work on a project that has the potential to advance safety assessment of chemicals without the use of animal testing, and to contribute to the development of innovative information systems in the field of toxicology.

[4] https://wise.vub.ac.be/tickle/
[5] https://u0152642.pages.gitlab.kuleuven.be/kg4di-fwo-network/

Master thesis on the TOXIN project

LIÈGE université — VUB VRIJE UNIVERSITEIT BRUSSEL

Integrating datasets into a knowledge graph for hazard assessment in the toxicological domain

Guillaume Vrijens – Montefiore Institute, University of Liège

Context

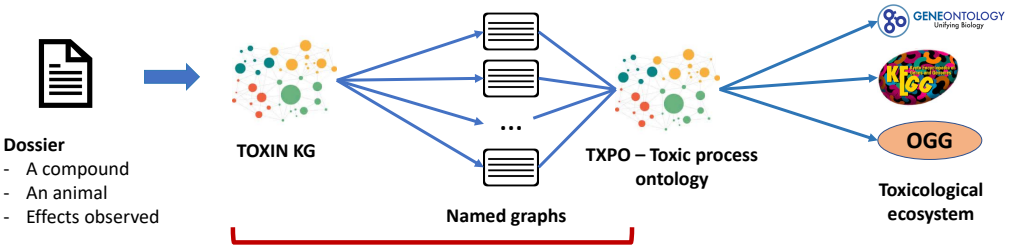
- **Hazard assessment (HA)** is the identification and characterization of hazards linked to a new chemical compound.
- **In vivo HA:** Tests on animals and then analyse of the toxicity on their organisms.
- **In vitro HA:** Tests on molecules or cells in a glass.
- **3R rule:** Replacement, Reduction, Refinement → need a better understanding of the biological processes that are influenced by the compound.

Challenges

- **Granularity:** Different levels of interaction (organ, cell, molecule).
- **Amount of data:** Huge toxicological ecosystem available.
- **Interpretation:** The knowledges evolve in time and different sources can have different conventions.

Contribution

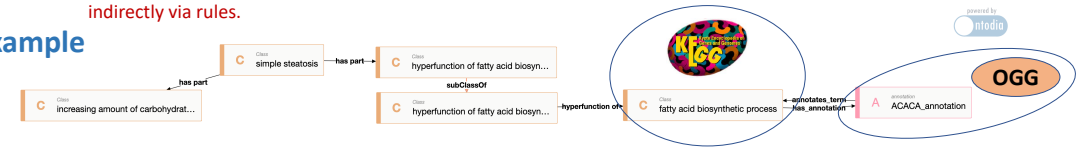
→ TOXIN developed a Knowledge Graph (KG) that gathers information about in vivo experiments contained in reports. This KG, combined with the huge toxicological ecosystem, can help make links between old in vivo data and new in vitro and in silico data by detecting correspondences between adverse outcomes and biological processes involved.



Inter-linking TOXIN and TXPO through named graphs.
- Each graph corresponds to an expert "interpretation" or a current state of toxicological knowledge.
- A link can be made between a dossier and a toxic effect either directly when the same effect occurs in two datasets or indirectly via rules.

Enrichment of TXPO with different resources from the toxicological ecosystem.

Example



Summary

- The aim of the TOXIN KG is to provide a tool for toxicologists that simplifies access to relevant biological data in order to provide hazard assessment.
- **We have proposed a method and tools to integrate information from various datasets starting from TOXIN KG.**
- **Our method took into account the need to different interpretations using named graphs.**

Conclusion

- TXPO gives a good representation of different toxic processes through different biological levels (molecule, cell,…).
- The enrichment adds useful information to the KG. This process is delicate because only relevant data must be added, without contradictions.
- This KG may be used in NLP tasks to find relevant toxicological concepts in a text.

Future work

- This project is still in progress. The inter-linking process is in development and other resources could be integrated.
- To facilitate the utilization of the KG, an efficient interface should be designed.
- Machine learning could be used to find comparisons between compounds and effects to early identify some potential hazards.

Acknowledgements

- Christophe Debruyne
- Sara Sepehri

Fig. A.1 Poster presented at the KG4DI kick-off.