Final report

Contract No.	Specific Contract 1 (ECHA/2011125) "Multiple framework contract with re-opening of competition for the provision of scientific support services – ONTOLOGY" (ECHA/2011/25)
Document Type:	Final Report
Organisation:	Istituto Superiore di Sanità (ISS), Viale Regina Elena 299,Rome 00161, Italy Fraunhofer Institute for Toxicology & Experimental Medicine (ITEM), Nikolai-Fuchs-Str. 1, 30625 Hannover, Germany
Contributors	Romualdo Benigni Olga Tcheremenskaia Monika Batke

Table of Contents

List of Figures and Tables

1. Inception Phase

1.1 Background

The field of predictive toxicology has an urgent for the development of open, public, computable standardised toxicology vocabularies and ontologies to support the applications required by *in silico, in vitro* and *in vivo* toxicology methods, and by related reporting activities such as the REACH (Registration Evaluation and Authorisation of Chemicals) legislation [1].

All predictive approaches in toxicology actually share the need of highly structured information as a starting point. The definition of ontology and of controlled vocabulary is a crucial requirement in order to: a) standardize and organize the chemical and toxicological databases on which the predictive toxicology methods build on; b) improve the interoperability between toxicology resources; and c) create a knowledge infrastructure supporting R&D and risk assessment.

1.2 Review of related ontologies

At present, several existing ontologies and standards initiatives are available and can contribute –to different extents– to the creation of a toxicology ontology supporting the needs of predictive toxicology and risk assessment. A first essential step of this project was the survey of relevant related projects. The results of this survey are given below.

Computational tools for predictive toxicology include a range of well known machine learning and bioinformatics algorithms, as well as specific cheminformatics procedures, for e.g., descriptor calculation and chemical structure processing. The Blue Obelisk descriptor ontology [2] is the first attempt to provide a formal description of cheminformatics algorithms. The Chemical Information Ontology is another ontology, which has been recently published [3], and it is considered the successor of the Blue Obelisk descriptor ontology.

Several ontologies, covering machine learning and data mining domains (e.g., DAMON [4]) have been developed in the context of Grid services and are available in DAML+OIL language, instead of the more standard OWL language. A number of data mining ontologies have been published recently, i.e.,: the data mining ontology (OntoDM) [5] [6], KDDOnto [7], KDO ontology [8], DMWF Ontology [9], and the e-LICO Data Mining Ontology (DMO) developed in the context of another EU FP7 project [10]. OntoDM is based on the unification of the field of data mining and the growing demand for formalized representation of outcomes of research. It includes definitions of basic data mining entities, such as datatype and dataset, data mining task, data mining algorithms and components thereof (e.g., distance function), etc. It also allows for the definition of more complex entities, e.g., constraints in constraint-based data mining, sets of such constraints (inductive queries) and data mining scenarios. The e-LICO team recently launched Data Mining Ontology Foundry [11] , which is currently populated with e-LICO suite of ontologies for data mining (DMO), model selection and meta-mining (Data Mining Optimization –DMOP) [12]. DMO also includes similar basic data mining entities, and provides means to automatically compose workflows by identifying algorithms with compatible input and output. Finally, collecting details of machine learning experiments in “experiment databases” for subsequent analysis [13, 14] provides distributed storage for all details of predictive toxicology workflows.

Despite the surge of simultaneous activities in developing data mining ontologies, their adoption by the major data mining platforms and tools is still a future goal.

On the other hand, unifying toxicology data description presents additional challenges, and no systematic ontology exists yet. However, a number of ontologies and initiatives can contribute to its development.

As one of the central repositories of large-scale biomedical ontologies, the Open Biomedical Ontology (OBO) Foundry [15] is an important source of ontologies for reuse. Several OBO ontologies could potentially be used as part of the development of a Toxicology Ontology.

The Gene Ontology (GO) [16] project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The GO project has developed three structured controlled vocabularies that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.

Chemical Entities of Biological Interest (ChEBI) [17] is a freely available dictionary of molecular entities focused on “biologically interesting” chemical entities and their activities in a biological context. The molecular entities in question are either natural or synthetic products used to intervene in the processes of living organisms.

The Ontology of Biomedical Investigations (OBI) [18] provides terminology relevant to experimental biological and clinical investigations. This includes a set of 'universal' terms, that are applicable across various biological and technological domains, and domain-specific terms. This ontology supports the consistent annotation of biomedical investigations, regardless of the particular field of study. The OBI addresses the need for a cross-disciplinary approach and represents all phases of experimental processes, as well as the entities involved in preparing for, executing, and interpreting those processes e.g., study designs, protocols, instrumentation, biological material, collected data and analyses performed on that data.

Other existing ontologies of relevance to the toxicology domain include anatomy ontologies such as the Foundational Model of Anatomy (FMA) [19] and the Mouse adult gross anatomy ontology [20]. The FMA Ontology is a knowledge source for biomedical informatics; it is concerned with the representation of classes or types and relationships necessary for the symbolic representation of the phenotypic structure of the human anatomy. Its ontological framework can be applied and extended to the other species. The Mouse adult gross anatomy ontology represents the Anatomical Dictionary for the Adult Mouse. This ontology organizes anatomical structures for the postnatal mouse spatially and functionally, using 'is a' and 'part of' relationships. A browser can be used to view anatomical terms and their relationships in a hierarchical display.

Another toxicology-relevant ontology is the NCI Thesaurus [21], which contains terminology relevant for clinical care, translational and basic research, and public information and administrative activities, with respect to cancer and related diseases and targeted therapies. The NCI Thesaurus provides definitions, synonyms, and other information on nearly 10,000 cancers and related diseases, 8,000 single agents and combination therapies, and a wide range of other topics related to cancer and biomedical research.

Several non-OBO Foundry ontologies and terminology and vocabulary resources were identified as relevant related projects as well. These include large-scale terminology standards such as MeSH [22] and SNOMED [23]. MeSH (Medical Subject Headings) is the controlled vocabulary thesaurus used for indexing articles for PubMed.

SNOMED CT (Systematized Nomenclature of Medicine -- Clinical Terms) is a systematically organised computer-processable collection of medical terminology covering most areas of clinical information such as diseases, findings, procedures, microorganisms, and substances. It allows a consistent way to index, store, retrieve, and aggregate clinical data across specialties and sites of care. It also helps organising the content of medical records, reducing the variability in the way data is captured, encoded and used for clinical care of patients and research.

The eTOX project [24] aims to develop a drug safety database from pharmaceutical industry legacy, toxicology reports, and public toxicology data. Ontology development within the eTOX Project includes the activity to create ontologies for preclinical safety.

1.3 Review of toxicological standard schemas

The survey of ontology-related resources has pointed to two important toxicological standards, that are publicly available and that are close to the needs of generating a Toxicology Ontology: the OECD Harmonized Templates (OECD HTs) [25], and the ToxML (Toxicology XML standard) schema (Leadscope Inc.) [26].

The OECD HTs correspond to the IUCLID5 XML [27] schemas, which are meant to be used by industry when submitting documentation on their chemicals to EU regulatory authorities. For each endpoint, the OECD HTs describe a series of fields defining the information submission requirements of a carcinogenicity experiment. Since they are generic enough to be able to include data on endpoints with different characteristics, in principle the OECD HTs provide a substantial basis for building ontology. However, they are not very formalized; they leave much space for free text entering, and have a strong administration emphasis rather than a scientific focus.

ToxML is an XML data exchange standard based on toxicity controlled vocabulary. The most recent ToxML release has a comprehensive, well-structured scheme for many toxicity studies (including carcinogenicity, *in vitro* mutagenicity, *in vivo* micronucleus mutagenicity, repeated dose toxicity) which fit well the databases included in the Toolbox. It should be emphasized that ToxML is not an ontology, since it contains only elementary relations, has no restriction rules and lacks of systematic definitions. Recently, a semi-automatic partial conversion of the ToxML schema to OWL-DL has been performed, with the purpose of benefitting from the reasoning mechanism of OWL, within the OpenTox project. The resulting ontology is described in detail in the following section, and can be applied to reference and annotate the contents of databases coming from various sources and toxicity studies.

1.4 Review of OpenTox ontology

1.4.1 OpenTox ontology: general considerations

OpenTox (OT) [28] was funded within the EU Seventh Framework program. The implementation of OT currently consists of distributed web services, running at several locations. Two initial OT web-applications have been made available: ToxPredict [29] that predicts the activity of a chemical structure -submitted by the user- in respect to a given toxicity endpoint; and ToxCreate [30] that generates predictive toxicology models from a user-submitted dataset. Within OT, Bioclipse -an Open Source workbench for the life sciences- has been integrated to launch calculations on remote OT services and to provide a rich user interface on the desktop [31].

An initial analysis within OT pointed to the importance of the standardisation of the framework components. This was addressed by adopting an ontology-based approach towards describing both toxicity data and computational procedures. Ontology definition is important for OT, as it supports integrated information processing in a more efficient and reliable manner, thus reducing the cost, maintenance and risk of application development and deployment.

The following related ontologies have been developed for OT: a) Toxicological ontology – listing the toxicological endpoints; b) Organs system and Effects ontology – addressing organs, targets/examinations and effects observed in in vivo studies; c) ToxML ontology – representing semi-automatic conversion of the ToxML schema[26]; d) OT ontology – representation of OT framework components: chemical compounds, datasets, types of algorithms, models and validation web services; e) ToxLink – ToxCast[32] assays ontology and f) OpenToxipedia community knowledge resource on toxicology terminology.

OT components are made available through standardized REST web services, where every compound, data set, and predictive method has a unique resolvable address (URI), used to retrieve its Resource Description Framework (RDF) representation, or to initiate the associated calculations and generate new RDF-based resources.

The services support the integration of toxicity and chemical data from various sources, the generation and validation of computer models for toxic effects, seamless integration of new algorithms and scientifically

sound validation routines and provide a flexible framework, which allows building arbitrary number of applications, tailored to solving different problems by end users (e.g. toxicologists).

The OT toxicological ontology projects may be accessed via the OT ontology development page www.opentox.org/dev/ontology. The OT ontology is available as OWL at http://opentox.org/api/1 1/opentox.owl

1.4.2 OpenTox ontology: general methodology

The construction of formal ontology of OT has followed relatively established principles in knowledge representation, particularly the OBO Foundry principles. The open approach to ontology development supports current and future collaborations with different projects. The DL species of the Web Ontology Language (OWL DL) have been used, as supported by the Protégé OWL editor. An overview of the OT ontology is given on the public area of the OT website [33] together with instructions on how to enter the OT Collaborative Protégé Server and to contribute to existing OT projects on OWL development. Some of the ontologies were manually created from scratch, while others partially reused existing ones and extended them with task-related concepts and relations.

The ToxML ontology is semi-automatically generated from the existing ToxML schema by parsing it to OWL and applying specific rules, which convey the semantics and remove redundant information in the new format.

To convert the XML schema into an OWL structure, several ad hoc rules were implemented:

- For distinguishing classes from properties among the XML fields;
- For introducing object properties – by default in the schema all properties correspond to datatype properties in OWL because they connect an entity to a string value;
- For removal of some container classes, which are not needed in an ontology (Tests, Compounds, etc.). These are necessary in XML because they frame a set of subfields, but in OWL each Test or Compound is a separate Object and many of these objects can exist independently and they are all related to their originating type class;
- To rename classes which appear with the same name in different contexts.

The resulting ontology has a flat structure representing numerous relations, that express the nested structure of the XML schema (see Figure 1.1).

- The IS-A relation is introduced only to a limited number of classes.

Example: ChronicStudies rdfs:subClassOf Study;

- The relations between the classes are obtained from the nested XML structures and encoded as follows: for each tuple of nested fields in the schema F1 and F2 (nested in F1), two new classes <Class F1> and <Class F2> are created in OWL along with an object property <hasClassF2> which expresses the relation between both classes. The feature <hasClassF2> has domain <Class F1> and range <Class F2> (the range and domain could be a union of classes if the nested class appears more than once in the source schema). For example Figure 1.2 compares the corresponding representation of the nested fields Study, Background and Reference Compound in ToxML and OWL.
- Fields which have string values in XML are converted to properties in OWL.

Example: The field Name which has a string value type becomes a datatype property hasName

- Ambiguous labels are unified.

Example: The field Results is used in several different contexts, that is why we rename it to TreatmentResults, in the treatment group.

- Wherever possible, object type properties are introduced instead of datatype ones – thus string values are replaced by named concepts.

Example: In ToxML schema the field Sex is defined as a simple type of type string, which would be converted to property in direct conversion, but in OWL we introduce a new class Sex, thus creating an object property hasSex instead of a datatype one.

- Mapping to Organ ontology and study type classifications can be applied.

Within OT, OpenToxipedia has been developed using Semantic Media Wiki (SMW). OpenToxipedia has been created manually by experts in the fields of *in silico* and experimental toxicology on the basis of known regulatory documents, glossaries, dictionaries and some primary publications. All registered members are welcome to add new entries, suggest definitions and edit the existing resource at www.opentoxipedia.org . OpenToxipedia is curated by toxicology experts within the OT community.

SMW has been chosen for OpenToxipedia representation because it enables automatic processing of the wiki knowledgebase and gives a possibility for data transfer between RDF and SMW through SPARQL. SMW will facilitate the automatic data exchange between OpenToxipedia, the ontologies and OT web services using RDF data. SMW is a collaborative system that supports versioning, RDF export, tools to lock pages by a curator (fixing a validated vocabulary) and the possibility to add annotation without changing the ontology or RDF information.

Figure 1.1 A snippet of the ToxML-based ontology classes

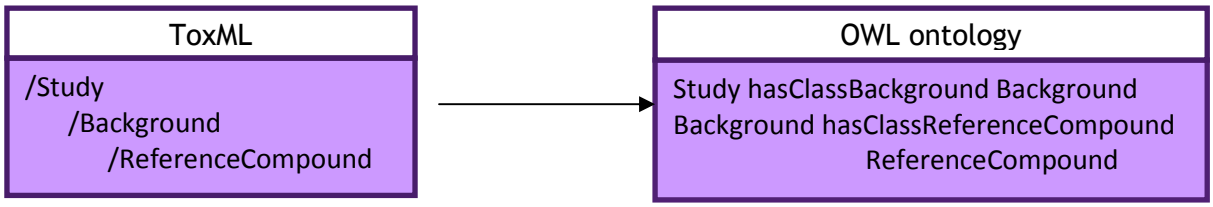


Figure 2.2 Introducing object type properties for each tuple of nested classes

1.4.3 OpenTox ontology: sub-projects

Six ontologies have been made available through the OT Collaborative Protégé Server:

- Toxicological Endpoint ontology;
- Organs system and Effects ontology;
- ToxML ontology;
- OT ontology, representing components of OT web services, framework and algorithm types

- ToxLink (ToxCast assays ontology)
- OpenToxipedia: SMW toxicology knowledge resource.

The OT Toxicological Endpoint ontology currently contains five toxicity study types: carcinogenicity, *in vitro* bacterial mutagenicity, *in vivo* micronucleus mutagenicity, repeated dose toxicity (e.g., chronic, sub-chronic or sub-acute study types) and aquatic toxicity (Figure 1.3))

The purpose of this ontology is to enable the attributes of toxicological dataset entries to be associated with ontology concepts. The OT framework exposes REST web services, corresponding to common OT components. A generic OWL representation is defined for every component (e.g. every OTdataset is a subclass of ot:Dataset, every algorithm is subclass of ot:Algorithm and every model is a subclass of ot:Model). This allows unified representation across diverse data and algorithms, and a uniform interface to data processing services, which take generic ot:Dataset resources on input and generate generic ot:Dataset resources on output.

The main OWL classes are "ToxicityStudyType", "TestSystem" (includes subclasses such as strains, species, sex, route of exposure), "TestResult" (includes subclasses such as toxicity measure, test call, mode of action, target sites).

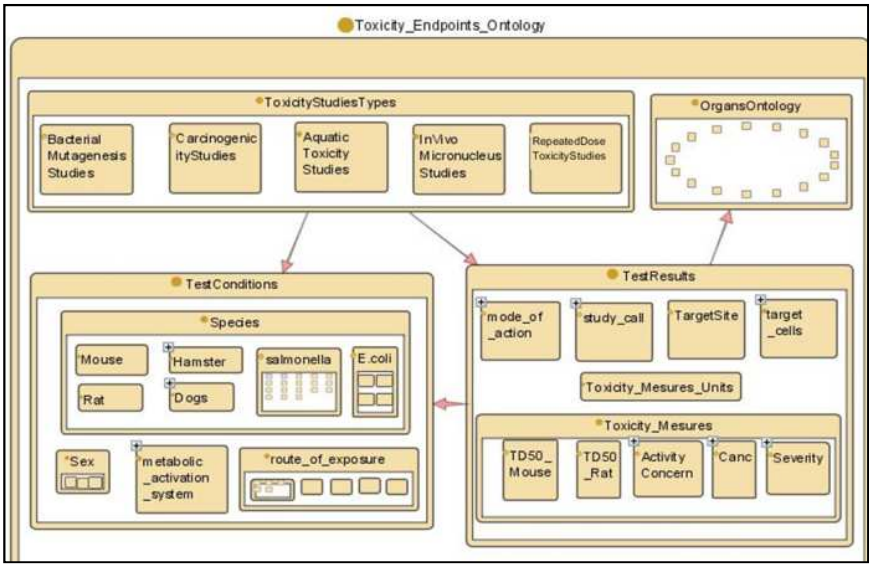


Figure 3.3 OpenTox toxicological ontology structure.

The Organs system ontology has been developed by the Fraunhofer Institute for Toxicology & Experimental Medicine. It is one of the most challenging ontology classes addressing targets, organs and effects observed in *in vivo* studies such as repeated dose toxicity and carcinogenicity experiments. The ontology includes the detailed description of organs starting from organs systems, down to histological components. It was decided to use a hierarchical structure starting with the individual organs systems (e.g. digestive system), instead of orientating the ontology on the examinations performed in guideline studies such as histopathology, necropsy, and clinical observations. The basic structure of the organs ontology is as follows:

```
– Class Organs system – Subclass Organs system
 |– Class Target organs – Subclass Target organs 1 to N
       |– Class Histopathology – Subclasses if needed
```

Currently the Organs system ontology includes 12 organ systems: digestive system, respiratory system, circulatory system, endocrine system, male genital system, female genital system, hematopoietic system, integumentary system, nervous system and special sense organs, urinary system, musculoskeletal system, immune system and lymphatic organs. Synonyms are included to account for differences in terminologies. It focuses on the organs observed in rodents, which are frequently used for toxicity testing. Currently, the Toxicological Effects Ontology comprises neoplastic and non-neoplastic effects observed in repeated dose and cancer studies. This ontology consists of three main parts: classes of effects, linked to pathological effects, which are further linked to detailed diagnostic features as agreed in the International Harmonization of Nomenclature and Diagnostic Criteria for Lesions in Rats and Mice (INHAND) initiative [36]. Its functionality has been initially developed for the respiratory tract. The structure of the combined Organs system and Effects ontology is depicted in Figure 1.4.

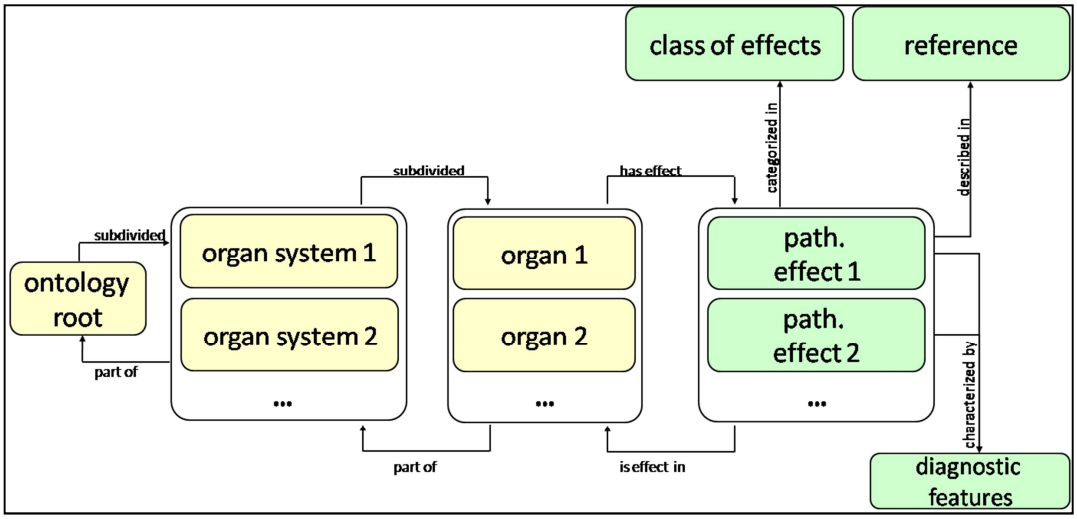


Figure 4.4 Overview of the structure of the combined Organs system and Toxicological Effects ontologies

The ToxML ontology created within OT is a semi-automatic conversion of ToxML schema to OWL-DL. The most recent ToxML release has a comprehensive, well-structured scheme for many toxicity studies (carcinogenicity, *in vitro* mutagenicity, *in vivo* micronucleus mutagenicity, repeated dose toxicity) which fit well the OpenTox purposes. This was verified by manually mapping various existing database entries to the ToxML schema. The purpose of the generation of a ToxML ontology was not only to develop a cross database matching schema, but also to inference on existing facts in the databases.

In order to use ToxML for data annotation, it has been necessary to overcome issues raised by the nature of the XML description. In the schema many fields host free text, instead of named concepts which hampers further automatic processing.

Example: In carcinogenicity studies data the survival rate is described as "*MORTALITY, INCREASED*" that are plain text strings not referring to any conceptual knowledge.

For solving this issue, in the ontology the same values are assigned as properties of a concept which is located within a domain knowledge model, thus allowing referencing between known entities. Another problem was the mixed type values (string / integer / interval) describing the same indicator in different studies.

Example: In data from different chronic studies the same indicator survival rate is presented as either of the following strings:

'0/5 (weeks 27-30)' ; '10/10' ; '16/17' ; '5/5' ; '6/6'

The first string in the example '0/5 (weeks 27-30)' represents the number of survived animals among the tested ones during some concrete period of the study between the 27th and 30th week, whereas in all other cases the strings show only the number of survived animals in relation to all tested animals. These are inconsistent data records which show the same indicator value in different value types. Organizing the type values as restricted values in the ontology allows for comparison and further automatic processing which is not feasible in plain string.

Overall, the OT ontology [34] provides a common information model for the most common components, found in any OT application, providing predictive toxicology functionality, namely chemical compounds, datasets of chemical compounds, data processing algorithms, machine learning algorithms, predictive models and validation routines. It is available as OWL at http://opentox.org/api/1 1/opentox.owl and described in detail in [1]. The OpenTox framework exposes REST web services, corresponding to each of these common components. A generic OWL representation allows unified representation across diverse data and algorithms, and a uniform interface to data processing services, which take generic ot:Dataset resources on input and generate generic ot:Dataset resources on output. Specific types of algorithms are described in the algorithm types ontology and more details of descriptor calculation algorithms are specified via the Blue Obelisk ontology [2] of cheminformatics algorithms (e.g. algorithm references, descriptor categories) and extensions, specifically developed to cover algorithms developed by OpenTox developers. Assigning specific information about the datasets, properties and types of algorithms and models is done via linking to the relevant ontologies, for example by sub-classing (rdf:type), owl:sameAs links, or Blue Obelisk ontology bo:instanceOf predicate.

The simultaneous use of OT datasets and compound properties as resources of generic ot:Dataset type and ot:Feature type in the OT ontology, and linking to specific toxicology ontologies, provides a flexible mechanism for annotation. It allows users of OT web services to upload datasets of chemical compounds and arbitrarily named properties of the compounds. The datasets are converted into a uniform ot:Dataset representation and chemical compound properties annotated with the proper terms from toxicology ontologies. The annotation and assigning of owl:sameAs links is currently only done manually, via OT REST web service interface, which modifies the relevant resource representation by adding/modifying triples. In principle, more sophisticated techniques could be applied, and the corresponding RDF representation updated via the same REST interface. This approach is currently used to enter and represent data in OT services and applications. Description of one of the OT API implementations, and examples of RDF representation of various resources is provided in [35].

The sixth ontology project started within OT is the ToxLink ontology representing the ToxCast assays from the US EPA. This development is a collaborative effort of OpenTox with ToxCast [32] to provide an ontological description of *in vitro* HighThroughPut toxicological assays.

OpenToxipedia is an important part of the OT project that supports the definition of terms used in the developed ontologies, OT web services and OT applications such as ToxPredict and ToxCreate. At present, OpenToxipedia contains 862 toxicological terms with descriptions and literature references classified into 26 categories. The terms can be browsed either by category or in alphabetical order. Due to its open collaboration platform, specialists in different toxicology fields can take part in the further creation and curation of terms within OpenToxipedia (see Figure 1.5).

OpenToxipedia provides a compendium for freely-available predictive toxicology resources supporting the application and development of the standards for representation of toxicology data, vocabulary and ontology development needed by OT use cases and web services. The following rules for term management in OpenToxipedia have been developed: (i) Add terms – any registered user (curators receive a message and decide what additions are approved and will become publicly available); (ii) Edit description of terms – curators; (iii)Add remarks – any registered user (curators receive an alert message).

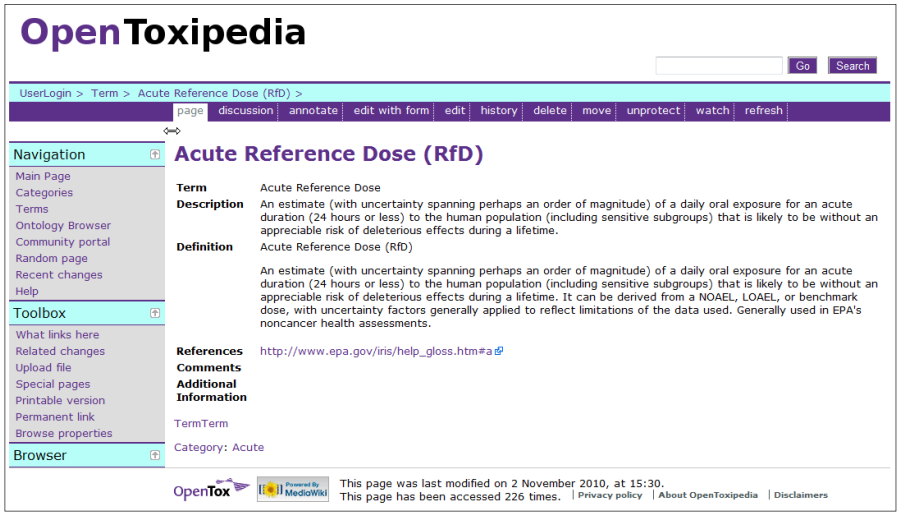


Figure 5.5 The interface for OpenToxipedia

1.5 Conclusions

Currently, no systematic ontology for toxicological effects and predictive toxicology needs is available. The only ontology existing in this field is the OT Toxicological ontology which aims at standardizing and organizing chemical and toxicological databases, and improving the interoperability between toxicology resources processing this data. The OT ontology is based on the ToxML schema and has been developed for the semantic support of the OT web applications. However, it is still a proof-of-concept ontology, and it is difficult to re-build/adapt for the OECD (Q)SAR Toolbox needs because of differences between the ToxML standard schema and the OECD harmonised templates.

Regarding anatomy, even if several ontologies covering this domain area exist, a serious gap is the absence of ontologies for localized histopathology, and more generally, ontologies of micro anatomy. An ontology describing strain names is also missing, which may be due in part to the more general problem of needing a concrete definition for 'strain'. For this reason, an Organs and Effects ontology has been developed within OT. This is closely linked to the INHAND initiative.. INHAND aims at developing -for the first time- an internationally accepted standardized vocabulary for neoplastic and non-neoplastic lesions as well as the definition of their diagnostic features. The description of the respiratory system [36] is already implemented in the OT ontology; additionally, the terms and diagnostic features of the hepatobiliary system were recently published [37].

In conclusion, the neighbouring ontologies mentioned above were available for partial reuse for the development of the ontological support for the Toolbox for all three workpackages of this project. However, significant part of work had to be done from scratch in order to cover all toxicity databases included in the OECD (Q)SAR Toolbox.

1.6 References

1. **REACH** [http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm]
2. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner J, Willighagen EL: **The Blue Obelisk-interoperability in chemical informatics.** *J Chem Inf Model* 2006, **46**:991-998.
3. Hastings J, Chepelev L, Willighagen E, Adams N, Steinbeck C, Dumontier M: **The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web.** *PLoS One* 2011, **6**:e25513.
4. Cannataro M, Congiusta A, Pugliese A, Talia D, Trunfio P: **Distributed data mining on grids: services, tools, and applications.** *IEEE Trans Syst Man Cybern B Cybern* 2004, **34**:2451-2465.
5. **Ontology of Data Mining (OntoDM)** [http://bioportal.bioontology.org/projects/119]
6. Panov P, L.N. S, Dzeroski S: **Towards an Ontology of Data Mining Investigations.** In *Book Towards an Ontology of Data Mining Investigations* (Editor ed.^eds.). pp. 257-271. City: Springer-Verlag; 2009:257-271.
7. Storti E: **Kddonto: An Ontology for Discovery and Composition of Kdd Algorithms.** *In Proc of the ECML/PKDD09 Workshop on Third Generation Data Mining: Towards Service-oriented Knowledge Discovery* 2009:13-24.
8. Zakova M, Kremen, P., Zelezny, F., Lavrac, N.: : **Planning to learn with a knowledge discovery ontology. Second planning to learn workshop at the joint ICML/COLT/UAI Conference pp.** 2008:29-34.
9. Kietz J, Serban, F., Bernstein, A., Fischer, S: **Towards cooperative planning of data mining workflows.** *Proceedings of the Third Generation Data Mining Workshop at the 2009 European Conference on Machine Learning (ECML 2009)* 2009:1-12.
10. **e-LICO: An e-Laboratory for Interdisciplinary Collaborative Research**

in Data Mining and Data-Intensive Science [http://www.e-lico.eu/]

11. **Data Mining Ontology Foundry** [http://www.dmo-foundry.org]
12. Hilario M, Kalousis, A., Nguyen, P., Woznica, A: **A data mining ontology for algorithm selection and meta-mining.** *Proceedings of the ECML/PKDD09 Workshop on 3rd generation Data Mining (SoKD-09)* 2009:76-87.
13. Vanschoren J, Blockeel, H: **A community-based platform for machine learning experimentation.** *Lecture Notes In Computer Science, ECML-2009* 2009, **5782**:750-754.
14. Vanschoren JLS: **Exposé: An ontology for data mining experiments.** *3rd Intl Wshp on Third Generation Data Mining Towards Serviceoriented Knowledge Discovery (SoKD-2010)* 2010:31-46.
15. **The OBO Foundry** [http://www.obofoundry.org/]
16. **Gene Ontology (GO)** [http://www.geneontology.org/]
17. **CHEBI ontology** [http://www.ebi.ac.uk/chebi/]
18. **OBI ontology** [http://obi-ontology.org/page/Main_Page]
19. **Foundational Model of Anatomy (FMA) ontology** [http://sig.biostr.washington.edu/projects/fm/]
20. **Adult Mouse Anatomy Ontology** [http://obofoundry.org/cgi-bin/detail.cgi?id=adult_mouse_anatomy]
21. **NCI Thesaurus** [http://ncit.nci.nih.gov/ncitbrowser]
22. **Medical Subject Headings (MeSH)** [http://www.ncbi.nlm.nih.gov/mesh]

23. **SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms)** [http://www.ihtsdo.org/snomed-ct/]

24. **ETox Project** [http://www.etoxproject.eu]

25. **OECD harmonized templates** [http://www.oecd.org/site/0,3407,en_21571361_43392827_1_1_1_1_1,00.html]

26. **Leadscope ToxML Schema** [www.leadscope.com/toxml.php]

27. **IUCLID 5** [http://iuclid.eu/index.php?fuseaction=home.preregistration]

28. Hardy B, Douglas N, Helma C, Rautenberg M, Jeliazkova N, Jeliazkov V, Nikolova I, Benigni R, Tcheremenskaia O, Kramer S, et al: **Collaborative development of predictive toxicology applications.** *J Cheminform* 2010, **2**:7.

29. **ToxPredict** [http://www.toxpredict.org]

30. **ToxCreate** [http://www.toxcreate.org/]

31. Willighagen E, Jeliazkova N, Hardy B, Grafström R, Spjuth O: **Computational toxicology using the OpenTox application programming interface and Bioclipse.** *J BMC Research Notes* 2011, **accepted**.

32. **ToxCast** [www.epa.gov/ncct/toxcast/]

33. **OpenTox ontology development page** [www.opentox.org/dev/ontology]

34. Tcheremenskaia O, Benigni R, Nikolova I, Jeliazkova N, Escher S, Batke M, Baier T, Poroikov V, Lagunin A, Rautenberg M, Hardy B: **OpenTox predictive toxicology framework: toxicological ontology and semantic media wiki-based OpenToxipedia.** *Journal of Biomedical Semantics* 2012, **3**:S7

35. Jeliazkova N, Jeliazkov V: **AMBIT RESTful web services: an implementation of the OpenTox application programming interface.** *J Cheminform* 2011, **3**:18.

36. Renne R, Brix A, Harkema J, Herbert R, Kittel B, Lewis D, March T, Nagano K, Pino M, Rittinghausen S, et al: **Proliferative and nonproliferative lesions of the rat and mouse respiratory tract.** *Toxicol Pathol* 2009, **37**:5S-73S.

37. Thoolen B, Maronpot RR, Harada T, Nyska A, Rousseaux C, Nolte T, Malarkey DE, Kaufmann W, Kuttler K, Deschl U, et al: **Proliferative and nonproliferative lesions of the rat and mouse hepatobiliary system.** *Toxicol Pathol* 2010, **38**:5S-81S.

1.7 Detailed work packages: Deliverables 1.2-1.4.

Workpackages: three workpackages (one for each toxicological endpoint) has been prepared in detail during the Inception phase.

The final objective of the individual WPs is to develop an ontology for Repeated Dose Toxicity, Reproductive/Developmental Toxicity and Carcinogenicity, following the principles of the Open Biomedical Ontologies Foundry (OBOF) and using the Web Ontology Language (OWL). A first requirement is obviously the compatibility of the Ontology with the existing terminology used in the QSAR Toolbox: the collaboration with the Toolbox contractors ensures that the three ontologies are correctly and efficiently mapped to the Toolbox terminology.

Each WP includes 6 tasks (deliverables):

1. Definition of classes and hierarchical relationships in the ontology structure

This task implies the definition of classes and subclasses for the ontology and its hierarchical relationships, taking into account existing work as agreed in the inception phase. The process has to be flexible in order to be able to cover toxicity databases currently present in the QSAR Toolbox, as well as potential further extensions.

2. Compilation of terms related to the endpoint

Taking into account the results from the review of existing relevant work related to ontology, the contractor compiles a list of terms related to the endpoint, including all terms from the databases present in the latest version of the OECD QSAR Toolbox as well as in related OECD test guideline documents and other identified sources.

3. Definition of synonymous and homonymous

For each term under point 2, a maximum number of synonymous and homonymous terms are to be identified taking into account terms used in test guidelines, databases and other identified relevant sources.

4. Establishment of relationships, interactions and hierarchies between classes, object and numerical properties for each term and rules when existing (internal rules and restriction rules).

The rules and restrictions, for classes, objects and numeric properties, within and across classes are to be established where applicable using the OWL language syntax. .

5. Association of each attribute in a toxicological dataset with an entry in the ontology.

The databases related to the endpoints present in the latest version the OECD QSAR Toolbox are harmonized in terms of terminology, in order to maximize accessibility of experimental data. The integration with existing vocabularies is considered.

The ontology concept has to be as flexible as to be able to be extended when new data become available. In fact, implementation of ontology for covering new use cases and datasets may be necessary in the future, as well as the testing of the updated ontology.

6. Final report on the developed ontology.

1.8 Detailed acceptance plan: Deliverables 1.5

All deliverables of each work packages are to be approved by Ontology Coordination Group.

The contractor (ISS) has first reviewed all subcontractor (ITEM) deliverables. Then all deliverables were presented at milestones to the Coordination Group to identify any significant deficiencies precluding the acceptance of the deliverable.

During the face to face meeting and the regular teleconferences, the Contractor has given an update to the Coordination group..

1.9 Detailed quality assurance plan: Deliverables 1.6

Since ontology is closely linked to the OECD QSAR Toolbox, the contractor has worked in close contact with the Toolbox contractors to ensure the suitability of the ontologies for the OECD (Q)SAR Toolbox.

It was agreed that an external evaluation was needed. OECD has coordinated this evaluation, done by written comment solicited from member countries and selected experts. The contractors agreed to provide a list of names and e-mail address of experts to be contacted for their input.

1.10 Detailed time line for entire project: Deliverables 1.7

Deliverables for Carcinogenicity	Due end of
D 2.1 Definition of classes and general hierarchical relationships in the ontology structure	March
D 2.2 Compilation of terms related to the endpoint	March
D 2.3 Definition of synonymous and homonymous	April
D 2.4 Establishment of relationships, interactions and hierarchies between classes, object and numeric properties for each term and rules	April
D 2.5 Association of each attribute in a toxicological dataset with an entry in the ontology	May
D 2.6 Intermediate ontology (i.e. OWL file) and report (i.e. word file) for carcinogenicity	May

Deliverables Repeat Dose Toxicity	Due end of
D 3.1 Definition of classes and general hierarchical relationships in the ontology structure	May
D 3.2 Compilation of terms related to the endpoint	May
D 3.3 Definition of synonymous and homonymous	June
D 3.4 Establishment of relationships, interactions and hierarchies between classes, object and numeric properties for each term and rules	June
D 3.5 Association of each attribute in a toxicological dataset with an entry in the ontology	July
D 3.6 Intermediate ontology (i.e. OWL file) and report (i.e. word file) for repeat dose toxicity	July

Deliverables Repro/Dev Toxicity	Due end of
D 4.1 Definition of classes and general hierarchical relationships in the ontology structure	July

D 4.2 Compilation of terms related to the endpoint	July
D 4.3 Definition of synonyms and homonyms	August
D 4.4 Establishment of relationships, interactions and hierarchies between classes, object and numeric properties for each term and rules	August
D 4.5 Association of each attribute in a toxicological dataset with an entry in the ontology	September
D 4.6 Intermediate ontology (i.e. OWL file) and report (i.e. word file) for repro/dev toxicity	September

2. Carcinogenicity Ontology

2.1 Definition of classes and general hierarchical relationships in the ontology structure

The ontology modelling and general structure had to satisfy two important principles: since ontology is linked to the OECD QSAR Toolbox[1], the contractor had to work by closely adhering to the Toolbox terminology, and the ontology concept had to be flexible for possible further extensions.

The present ontology had to respect regulatory needs, and was tailored on existing constraints, i.e.,: a) the OECD HTs[2]; and b) the carcinogenicity databases in the Toolbox (e.g., Long-Term Carcinogenicity Bioassay on Rodents Database (ISSCAN) [3] and Carcinogenic Potency Database (CPDB)[4]). The goal of the project was to support and improve data integration and mapping in Toolbox, which is very domain-specific task, with most terms/concepts already having an official terminology (used e.g., in the OECD HTs).

To guarantee the compatibility of ontology with the OECD QSAR Toolbox, the ontology was separated into two main super-classes: the first one is dedicated to HT # 72 Carcinogenicity [5] template, while the second covers the experimental carcinogenicity datasets terminology (Figure 2.1). Using this approach, it is possible to add smoothly any new carcinogenicity dataset to the ontology.

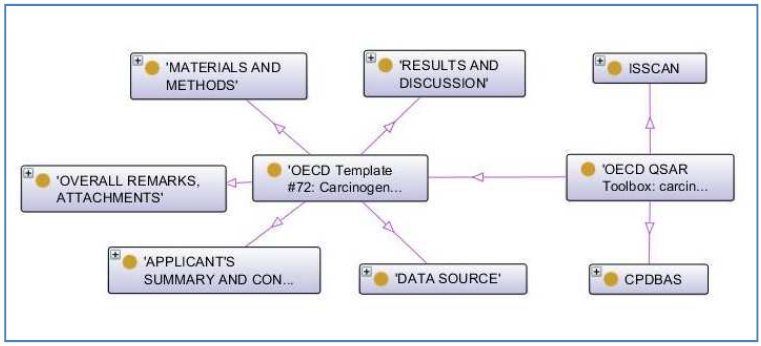


Figure 6.1 General structure of the carcinogenicity ontology.

The conversion of the HT 72 Carcinogenicity into the OWL format was performed semi-automatically, with the Protégé Excel Import Plug-in[6].

The Carcinogenicity ontology based on OECD HT # 72: Carcinogenicity has five main sub-classes: Overall remarks, attachments, Applicant's summary and conclusion, Data source, Materials and Methods, Results and Discussion. The OECD field number and other information contained in the HTs documentation (Figure 2.2) was maintained in order to facilitate the ontology maintenance in case of HTs updates.

---

[1] http://www.oecd.org/chemicalsafety/assessmentofchemicals/theoecdqsartoolbox.htm

[2] http://www.oecd.org/ehs/templates/

[3] http://www.iss.it/ampp/dati/cont.php?id=233&lang=1&tipo=7

[4] http://potency.berkeley.edu/index.html

5 http://www.oecd.org/ehs/templates/carcinogenicity.htm

[6] http://protegewiki.stanford.edu/wiki/Excel_Import

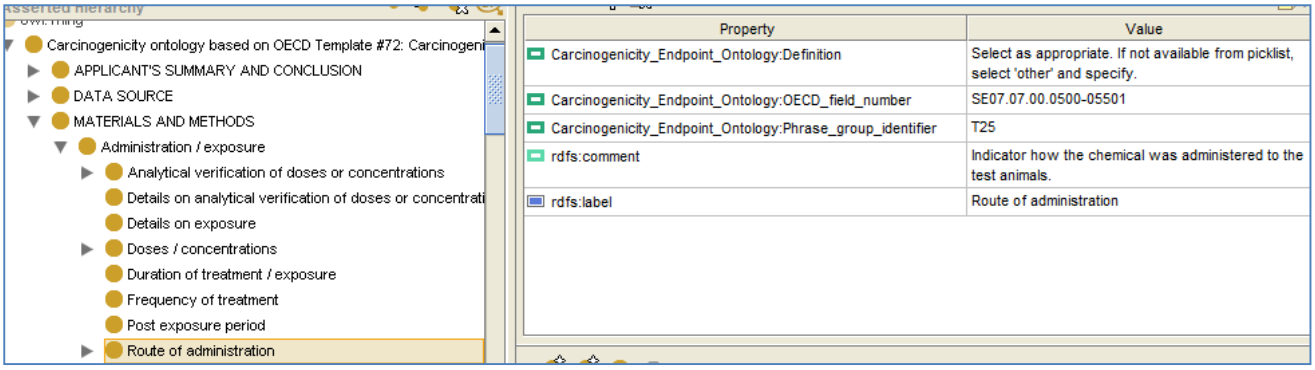


Figure 7.2 Protégé Screenshot: the HT's fragment in OWL format.

2.2 Compilation of terms related to the endpoint

Following the semi automatically conversion of OECD Template #72: Carcinogenicity into the OWL format, the additional terms have been collected from different sources including:

-OECD picklists[7]: the standard phrase description as well as the picklist' number have been also imported (Figura 2.3);

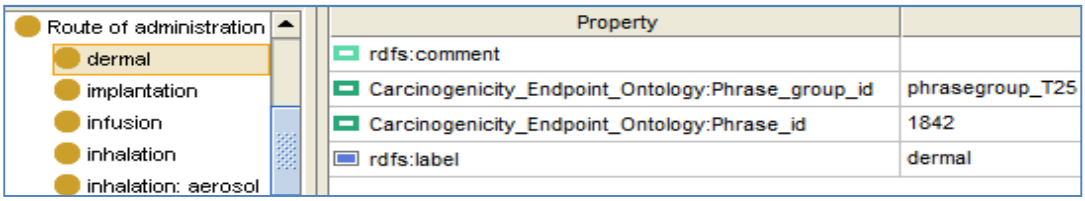


Figure 8.3 Protégé Screenshot: example of the OECD HT picklist imported into the OWL format.

-OECD Predefined Tables, which contain rich text fields for terms not defined by the picklists; for example, the Clinical observation predefined table[8] contains the extension for the OECD HT term "haematology";

-toxicological ontology elaborated within the OpenTox project[9], including the ITEM Organs system and Effects ontology based on the INHAD initiative[10] . Within this project, the above has been further extended for the Repeated Dose toxicity ontology;

When no internal resources were available, terms from existing ontologies freely available at the Bio-portal ontology depository (http://bioportal.bioontology.org/) were partially re-used.;

-SNOMED CT[11];

-NCI Thesaurus[12] ;

-Clinical Terms Version 3 (CTV3) [13];

---

[7] http://www.oecd.org/ehs/templates/picklists.htm

[8] http://www.oecd.org/ehs/templates/45080116.htm

[9] www.opentox.org

[10] http://www.goreni.org/gr_back_inhand.php

[11] http://purl.bioontology.org/ontology/SNOMEDCT

[12] http://purl.bioontology.org/ontology/NCIt

[13] http://purl.bioontology.org/ontology/RCD

-Mouse_pathology [14]

-Mouse adult gross anatomy [15]

The Protégé Bioportal Import plug-in[16] has been used for the direct import of external ontologies class into Protégé.

The full vocabularies and relative documentation from the carcinogenicity experimental datasets Carcinogenic Potency Database (CPDB)[17],[18] and Long-term Carcinogenicity Bioassay on Rodents Database (ISSCAN)[19] have been collected into the class "OECD QSAR Toolbox: carcinogenicity datasets" (Figure 2.3).

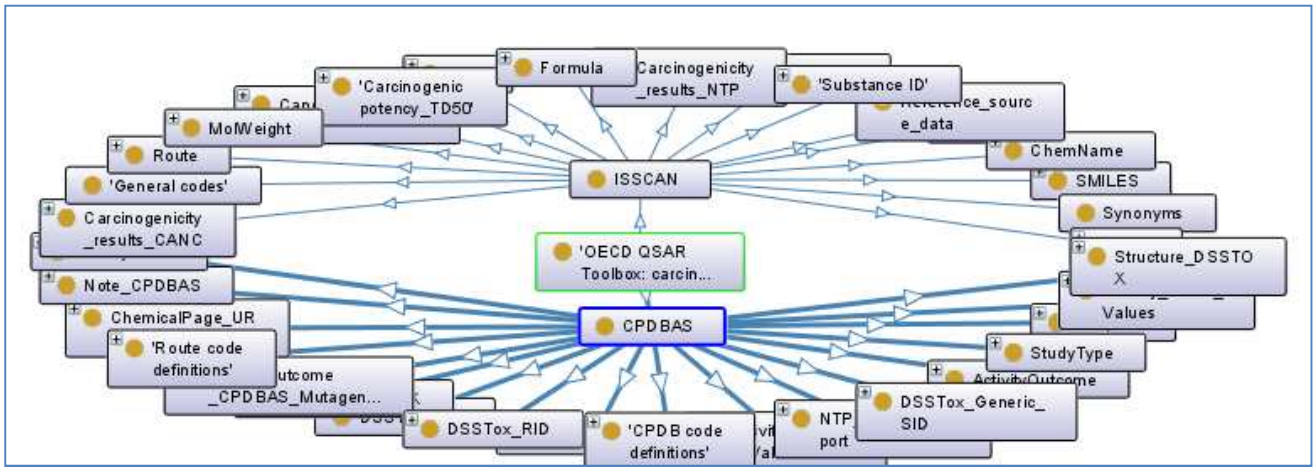


Figure 9.4 Protégé Screenshot: collection of terms relative to the experimental datasets.

Overall about 6 000 terms are implemented in the carcinogenicity ontology.

2.3 Definition of synonyms and homonyms

Synonyms and homonyms collection is needed for understanding how the same terms are defined and annotated in different sources, e.g. in the different datasets, the OECD HTs, the external biomedical ontologies. The OHTs terms, if available, are used as standard names.

The synonyms definition for single terms for different datasets was performed by introducing the OWL annotation property: "has_synonym" and "p4:synonym" for organs ontology. An example of such synonyms definition are shown on Figure 2.5.

Synonyms added both manually and automatically when imported from the external resource using the Protégé Bioportal Import plug-in (Figure 2.6).

In some cases the definition of synonyms was not even needed, as the same term may be used in different sources.

---

[14] http://purl.bioontology.org/ontology/MPATH

[15] http://purl.bioontology.org/ontology/MA

[16] http://protegewiki.stanford.edu/wiki/BioPortal_Import_Plugin

[17] http://potency.berkeley.edu/index.html

[18] http://www.epa.gov/ncct/dsstox/StandardChemFieldDefTable.html

[19] http://www.iss.it/ampp/dati/cont.php?id=233&lang=1&tipo=7

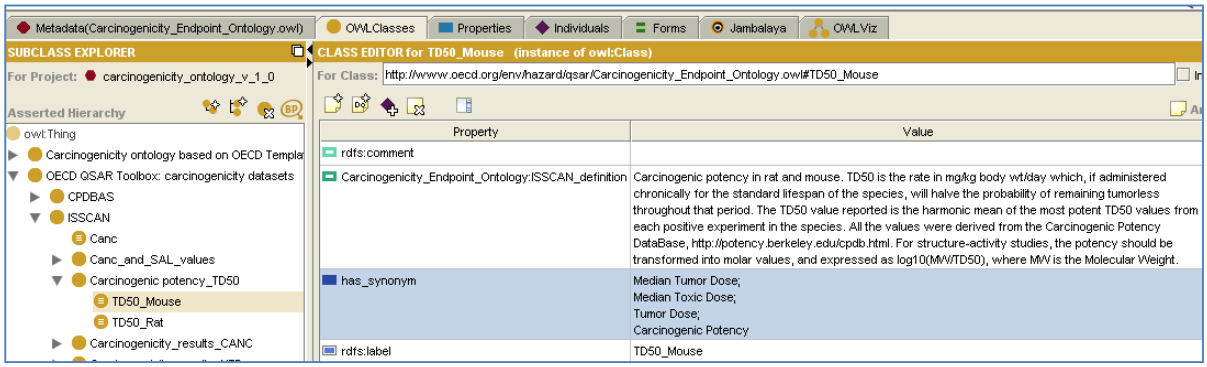


Figure 10.5 Protégé Screenshot: example of synonyms collection for the Carcinogenic potency TD50.

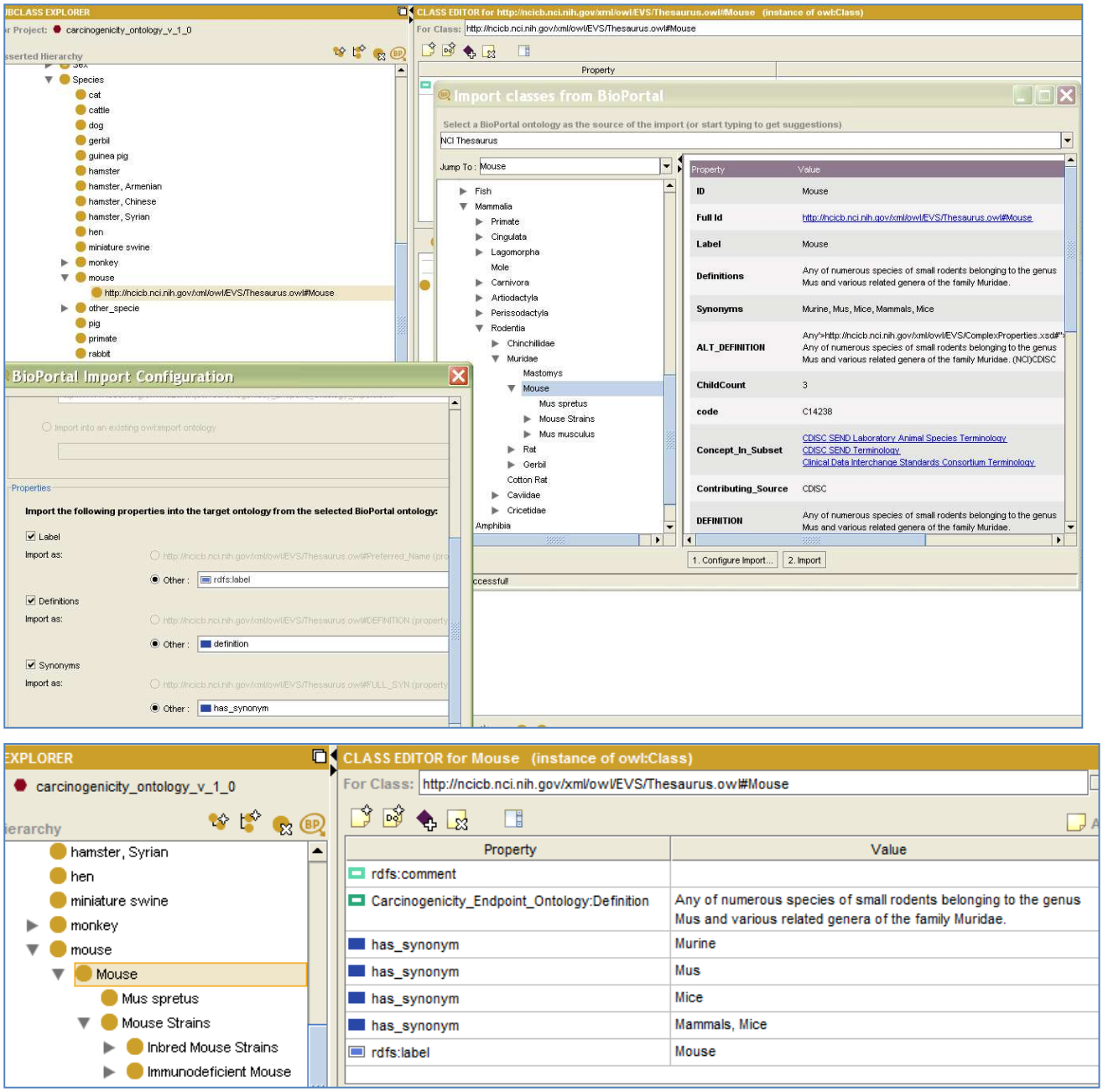


Figure 11.6 Protégé Screenshot: example of synonyms collection with the Bioportal Import plug-in.

2.4 Establishment of relationships, interactions and hierarchies between classes, object and numeric properties for each term and rules

The classes and their hierarchy in the present ontologies correspond to the OECD Harmonised templates and the OECD (Q)SAR toolbox terminology. The classes relative to the experimental datasets vocabulary have hierarchy corresponding to the original database organization.

Relationships between ontological classes and restriction rules can be introduced into the OWL using the object properties. The Figure 2.7 shows the full list of object properties created during the carcinogenicity ontology development.

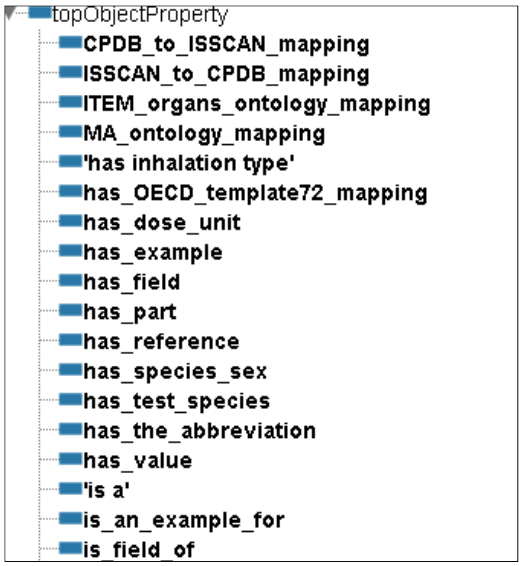


Figure 12.7 Protégé Screenshot: object properties created during the carcinogenicity ontology development.

An example of relationships and rules in ontology is show in Figure 2.8. The term from the ISSCAN database Mouse_Female_Canc has been restricted using the object properties “has_test_species” and “has_species_sex” only to the female mouse

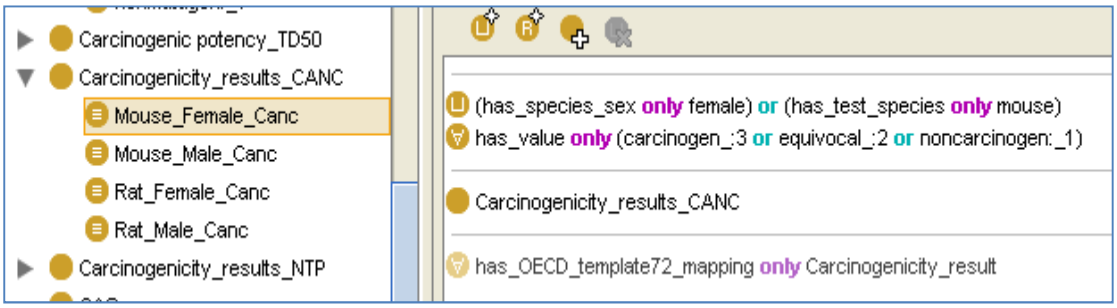


Figure 13.8 Protégé Screenshot: example of restriction rules in OWL.

To indicate the equivalent terms coming from different datasets and annotated originally to the different terminology, the inverse property: CPDB_to_ISSCAN_mapping ↔ ISSCAN_to_CPDB_mapping was created. This property helps to map the same term one to each other. The Figure 2.9 shows the mapping between the equivalent terms from ISSCAN and CPDB datasets: “carcinogen_3” is mapped to the” active”, while

"noncarcinogen_1" is mapped to the "inactive". Such mapping enables the automatic alignment between two databases.

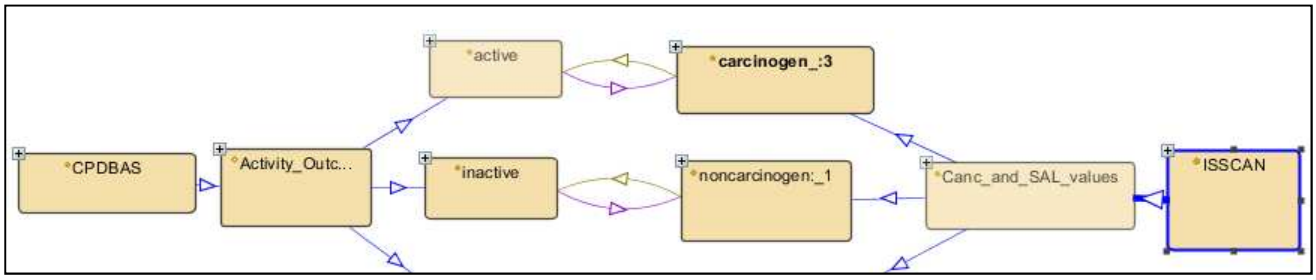


Figure 14.9 Protégé Screenshot: example of mapping between datasets.

The Figure 2.9 shows an example of using the numeric property "has_numeric_value" for "ActivityScore" term collected from the CPDB dataset.

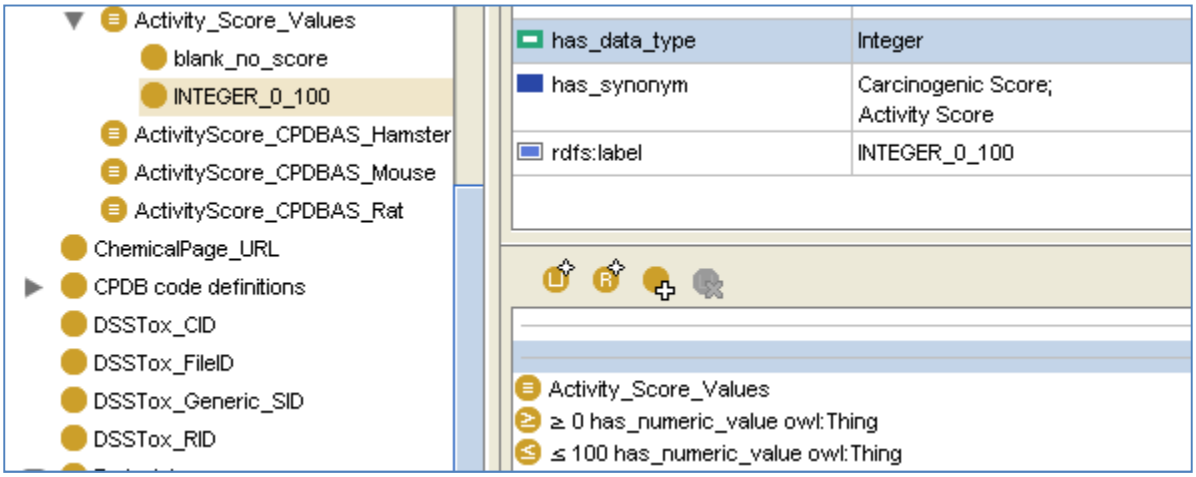


Figure 15.10 Protégé Screenshot: example of numeric property use in OWL.

2.5 Association of each attribute in a toxicological dataset with an entry in the ontology

The exact mapping between the same terms coming from ISSCAN, CPDBAS and the OECD Template #72: Carcinogenicity has been performed associating each attribute in a toxicological datasets with an entry in the ontology.

The OWL property "has_OECD_template72_mapping" has been introduced for mapping the datasets entries to the ontology created on the basis of extended OECD Template #72: Carcinogenicity.

Figure 2.11 shows an example of mapping of the datasets entries (ISSCAN:"Route" and CPDBAS:"Route code definitions") to the ontology entry ("Route of administration") and respective sub-classes.

Since no picklist for the Organ term in the OECD HT is available, it has not been necessary to import an ontology containing the target organ list. We have imported two external ontologies that are ITEM organs ontology from the OpenTox Project and the Mouse adult gross anatomy ontology http://purl.bioontology.org/ontology/MA. The ITEM organ ontology is the main organs and effects description rescurce, however for the moment also the Mouse adult gross anatomy ontology link was maintained. We have introduced the OWL properties "ITEM_organs_ontology_mapping" and "MA_ontology_mapping" for mapping

the Target organs class. Figure 2.12 shows the mapping of CPDBAS entry “ovary” to the ITEM organs ontology “p4:ovarium” and to the MA ontology “ovary”.

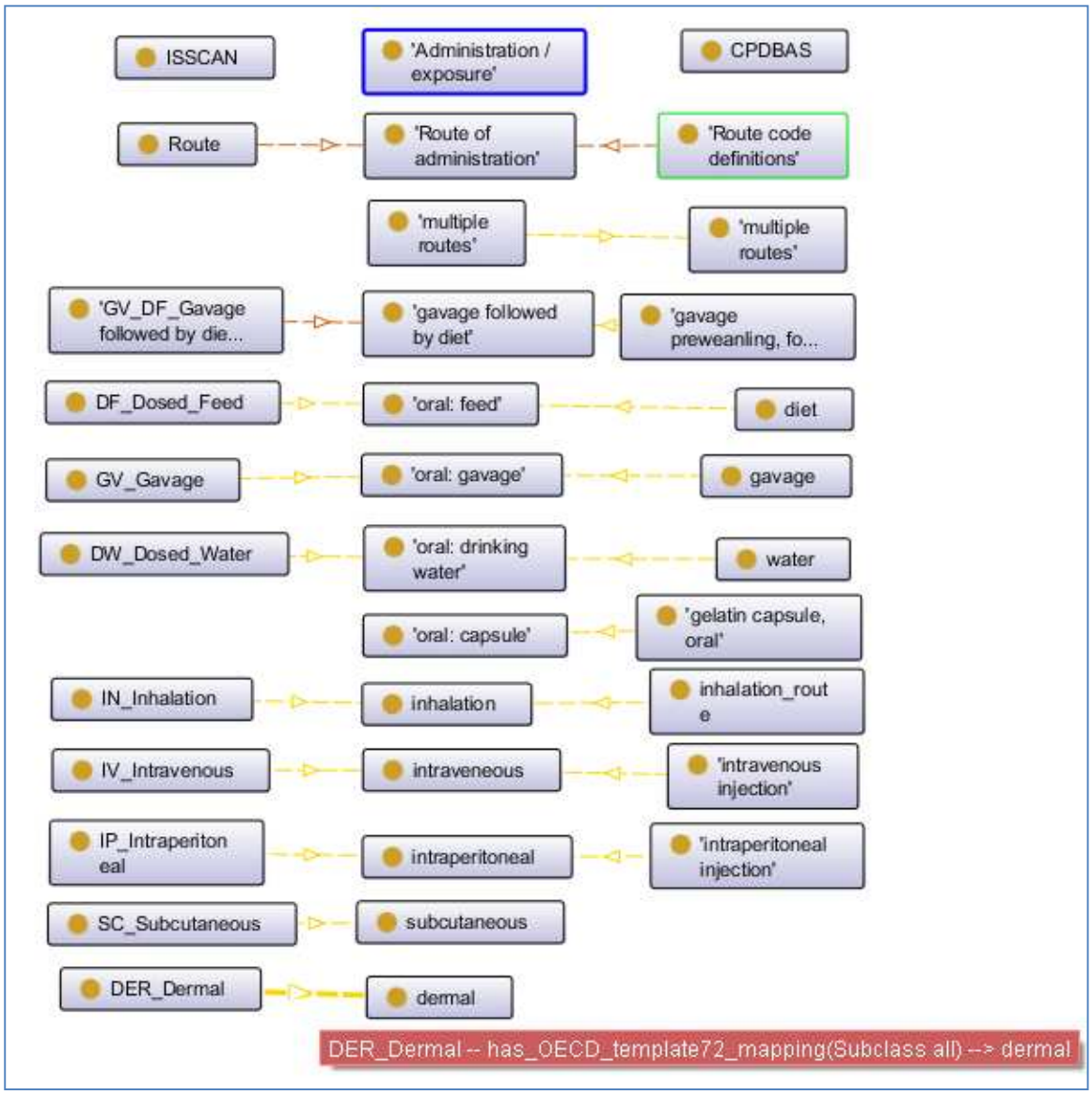


Figure 16.11 Example of mapping of the ISSCAN and CPDB datasets entries to the ontology.

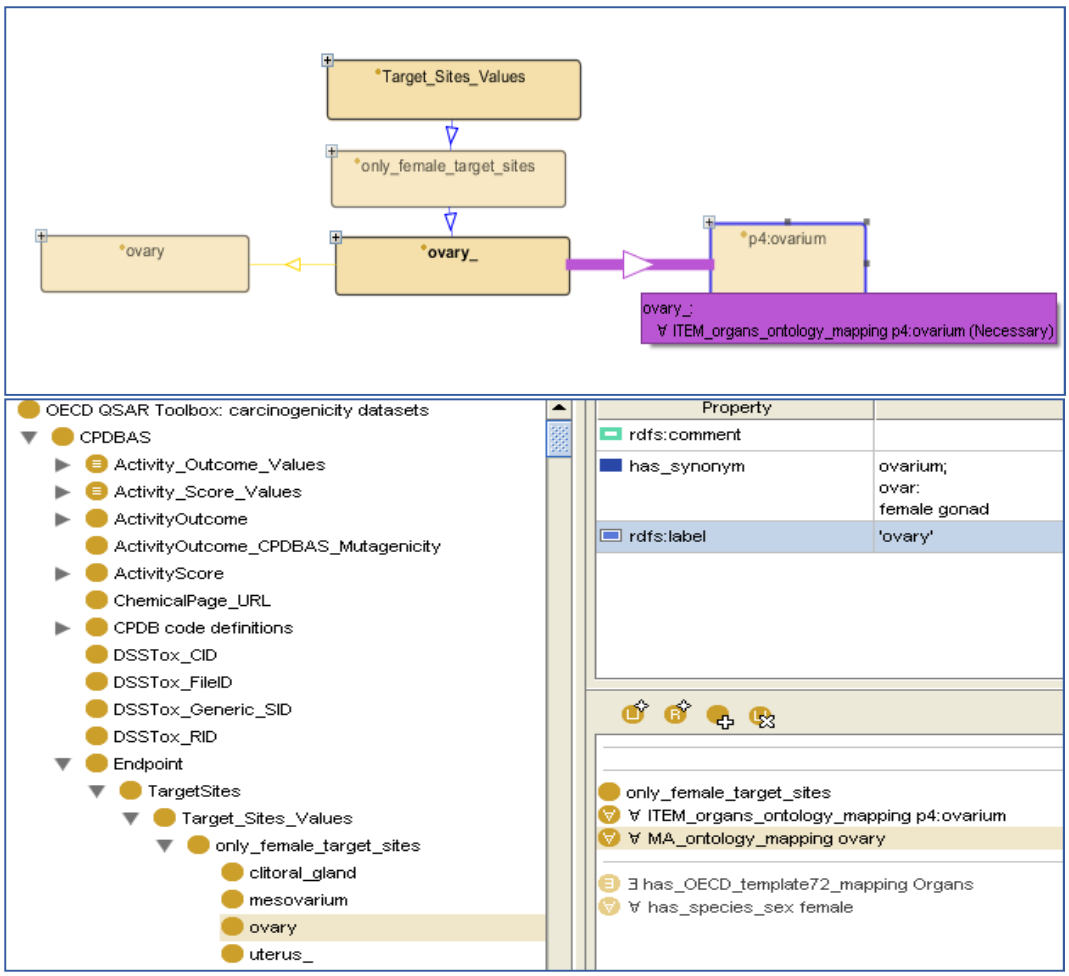


Figure 17.12 Example of mapping of the ISSCAN and CPDB datasets entries to the ontology.

In a number of cases, besides the mapping to the ontology it has been necessary to indicate the equivalent terms between datasets. For this purpose, the inverse property: CPDB_to_ISSCAN_mapping ↔ ISSCAN_to_CPDB_mapping was introduced as explained in more details in the previous Section 2.4, Figure 2.9.

2.6 Intermediate ontology

The OWL file with the carcinogenicity ontology version 1.1 can be opened for view/edits using the open source OWL editor Protégé version 3.4.7 or 3.4.8 (for more details, see the Appendix A)

The main sub-classes of the first ontology class "Carcinogenicity ontology based on OECD Template #72: Carcinogenicity" are Administrative data, Applicant's summary and conclusion, Data source, Materials and Methods, Results and Discussion, that aim at describing in detail the carcinogenicity experimental study. The second ontology class is "OECD QSAR Toolbox: carcinogenicity datasets", that contains detailed description of the toxicity datasets included into the OECD QSAR Toolbox software (CPDBAS and ISSCAN databases). For more details, see section 2.1.

The ontology has more than 6000 terms related to the carcinogenicity endpoint, which have been collected from the OECD HTs, OECD picklists and predefined tables, OECD QSAR Toolbox carcinogenicity datasets and external related ontological resources (section 2.2).

Synonyms and homonyms collection was performed. The synonyms definition for single terms for different datasets has been done introducing the OWL annotation property: "has_synonym" (section 2.3).

Establishment of relationships, interactions and hierarchies between classes, objects and numeric properties was performed using the OWL functionalities (section 2.4).

The exact mapping between the same terms coming from ISSCAN, CPDBAS datasets and the OECD HT #72: Carcinogenicity has been performed associating each attribute in a toxicological datasets with an entry in the ontology (section 2.5).

Even if the ontology addresses three specific datasets of the Toolbox, the general concept is remaining flexible. It will be possible to import other carcinogenicity databases and map them to the ontology, or update the ontology in the case of further extension of the Toolbox Software. The owl-file describes the hierarchies and relations using RDF/XML-based serializations, thus apart from using Protégé the file can also be opened in respective editors and used as an xml-template for other operations and functions.

2.7 Additional comments (OECD (Q)SAR Toolbox relevance, Harmonised Template relevance, modification made according to the reviewers' comments)

All carcinogenicity toxicity databases of the QSAR Toolbox with their terminology were implemented in the current version of this ontology. The entities of each database are mapped to the respective classes of the OHTs. The OHTs are implemented with terms, field numbers, explanations, help texts, and pick lists. The original terminology of the databases and OHTs is completely maintained. The concept of the ontology is flexible and open for future additional database to be mapped or for changes that may be necessary within the current implementation.

The original terminology of OHTs and databases vocabulary had to be maintained in order to fulfill the requirements of the contract and to facilitate the implementation in the QSAR Toolbox. The terminology does not meet completely the principle of naming conventions which is one of the OBO Foundry best practice principles[20]. This was one of the major aspects discussed by the reviewers. The Referees' points are acknowledged, but -being this ontology to be used within the QSAR Toolbox- this conflict cannot be solved at present. The ontology developed within this project is a very domain-specific task, with limited degrees of freedom. As a result, at this point all naming convention rules cannot be fully respected because most terms have already their official names (used for example in the OECD HTs).

The above consideration is valid also for the modeling problems discussed by the reviewers. We kept the original HT terminology and the structure of the HT's picklists. Remodeling the hierarchy and names (as shown in the worked example given in Figure 2.13) would destroy the HTs structure.

Our point-to-point responses to the reviewers are given in the Appendix B. The updated version of the OWL file contains some local modifications such as improving of naming when is possible, better definitions of some classes, correction of errors.

[20] http://obofoundry.org/wiki/index.php/Main_Page

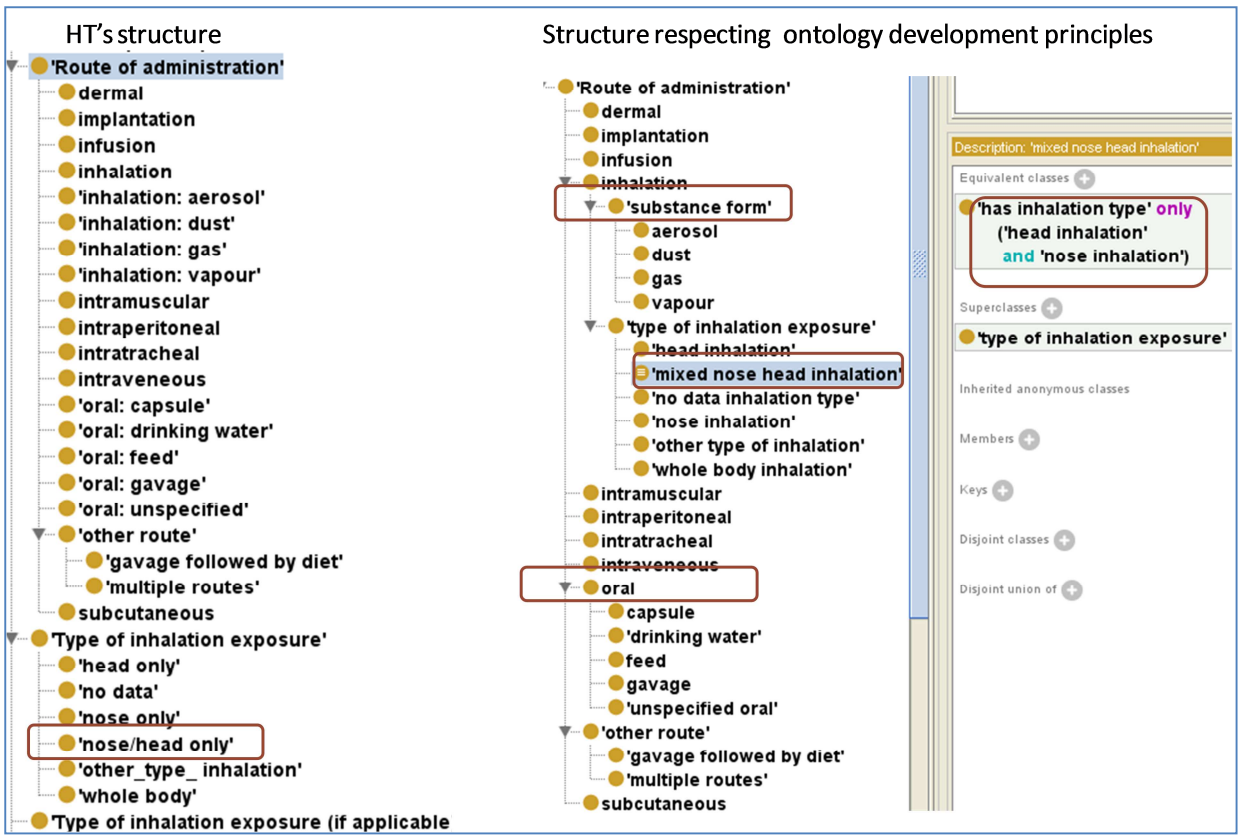


Figure 18.13 Example of possible remodelling and renaming of ontology classes.

The complete conversion of the OECD HTs into the OWL semantic-based format is suitable for future use for automatic experimental data mapping. To facilitate the use of OWL based HTs we preserved all HT's identifiers and connected the HT's terms directly to the corresponding picklits. When the picklists were not available, the extensions using available external resources were performed.

A long term goal may be the introduction of the ontology as the basis of data exchange and harmonization within the OECD QSAR Toolbox. The short term goal of the ontology is to help the integration and standardization of experimental data (e.g. improving the experimental trees, other local improvements). A proposals for the OECD (Q)SAR Toolbox trees modification is given in Appendix C.

3. Repeated Dose Toxicity Ontology

The ontology for Repeated Dose Toxicity comprises and relates terms used in the currently three different databases of the QSAR Toolbox and global standards as there are the OECD Harmonized templates (OHTs[21]) and the INHAND terminology (International Harmonization of Nomenclature and Diagnostic Criteria for Lesions in Rats and Mice[22]).

Currently the QSAR Toolbox contains three databases (DB) for repeated dose toxicity:

- RepDose DB – developed by the partner Fraunhofer ITEM, [23]
- HESSDB – developed by New Energy and Industrial Technology Development Organization (NEDO), Japan, [24]
- Tox Ref DB – developed by US EPA,[25]

The OHTs define a structure on reporting toxicological studies but it has to be emphasized that they are not ontologies as they miss the definition of relationships and the naming conventions e.g. by OBO foundry (Open Biological and Biomedical Ontologies) are not followed (e.g. usage of plural). For the purpose of the repeated dose ontology, three templates in their versions of February 2012 were implemented: OHTs #67 – 69: Repeated dose toxicity: oral, inhalation, dermal[26,27,28].The terms of the indicated OECD pick lists were added as required[29]. These OHTs for repeated dose are mainly focused on describing the study design and report but they are very limited in defining terms for organs and corresponding effects. No pick lists for organs or effects are available but some terms for organ systems, organs and clinical observations are included in the predefined tables[30, 31, 32].

We tried as much as possible to maintain the original terminology of the OHTs but as the organ and effect terminology is very sparse, we used the terminology developed by INHAND. The organ ontology based on INHAND has already been developed within the OPENTOX project [33](EU Seventh Framework, Health-F5-2008-200787, 2008-2011)). The terminology of organs and their corresponding effects can be assessed by registered pathologists (ESTP, STP, BSTP, JSTP). Within our ontology, we included commonly used terms for further description of organs up to the cellular level. The terminology developed for effects by INHAND is published related to organ systems. Within the OPENTOX project we were able to implement the effects on the respiratory tract[1]only, within the current project the hepatobiliary system was available as publication [2] and

---

21 www.oecd.org/ehs/templates

22 www.goreni.org/gr_back_inhand.php

23 www.fraunhofer-repdose.de

24 www.nedo.go.jp/kankobutsu/pamphlets/kouhou/2008gaiyo_e/73_82.pdf

25 http://www.epa.gov/ncct/toxrefdb/

26 www.oecd.org/dataoecd/57/1/47299664.html

27 www.oecd.org/dataoecd/57/5/47299823.html

28 www.oecd.org/dataoecd/57/50/47300168.html

29 http://www.oecd.org/document/9/0,3746,en_21571361_43392827_45625801_1_1_1_1,00.html

30 www.oecd.org/ehs/templates/repeatdosetoxicityoral.htm

31 www.oecd.org/ehs/templates/repeateddosetoxicityinhalation.htm

32 www.oecd.org/ehs/templates/repeateddosetoxicitydermal.htm

33 www.opentox.org

the terms of the urinary system, central and peripheral nervous system, mammary, zymbal's, preputial, and clitoral glands as well as of the male reproductive system were available as pre-published lists. The full publications were available shortly after the due date of deliverable *3.2 Compilation of terms related to the endpoint* in June and August ([3], [4], [5], [6]). With respect to the INHAND ontology developed in OPENTOX and the continuation within this project, we had a minor change in the concept of implementation: within the OPENTOX version we added the diagnostic features as description to each effect, but having discussed this with the developers of the INHAND nomenclature (Dr. Susanne Rittinghausen, head of pathology Fraunhofer ITEM),we decided to give the link to GoRENI.org instead. The diagnostic features as published now are subjected to change in the future as new knowledge and diagnostic technics may arise. These changes will always be implemented in the GoRENI.org web service and thus having the link will enable the ontology user to assess the actual information instead of using an old diagnostic feature.

The terminologies of the databases for repeated dose toxicity implemented in the QSAR Toolbox were gathered from version 2.3. Additionally, the vocabulary of the ToxRefDB as provided by M. Martin (personal communication, Version of August 2011 received by email January 2012) and the thesaurus of the HESS-NEDO database as published [7] were considered. The published thesaurus of HESS-NEDO database comprises the same terms as the database in the QSAR Toolbox. The ToxRef DB version in the QSAR Toolbox refers to the data sets provided by EPA related to their respective publications: Chronic & Cancer Endpoints [8], Developmental Toxicity Endpoints [9], and Reproductive Toxicity Endpoints [10]. In these data sets, effect classes were implemented for the purpose of the respective publications and underlying statistical evaluations. These effect classes were not originally part of the ToxRef DB and are not implemented in the vocabulary. The ToxRef DB in its entirety is freely available for download from the EPA website [34] and contains the vocabulary as provided by M. Martin. Within the current ontology, the ToxRef DB terminology as implemented in the QSAR toolbox and in addition the vocabulary for organs and effects are provided. The latter is not mapped to OHTs or other databases in order not to cause confusion about the terminology currently contained in the QSAR Toolbox. Some terms are, however, similar to other sources and thus these terms then belong to several classes including classes of ToxRef DB vocabulary and HESS-NEDO DB for instance.

3.1 Definition of classes and general hierarchical relationships in the ontology structure

The first deliverable *3.1 Definition of classes and general hierarchical relationships in the ontology structure* resulted in the general structure of the ontology based on the OHTs and the RDT (Repeated Dose Toxicity) DBs of the QSAR Toolbox (Figure 3.1). The INHAND based organs and effects ontology was implemented in the OHTs as shown in **Error! Reference source not found.**. It also includes all effects observed in organs for which no INHAND manuscript was available by the end of this project: “effects in organs without INHAND manuscript”.

The object properties shown in **Error! Reference source not found.** were implemented and used to define relationships.

34 http://actor.epa.gov/actor/faces/Download.jsp

Table 19.1 Object properties implemented in the RDT ontology for definition of relationships

Object property	Inverse object property
has part	is part
has effect	is effect in
has example	is an example for
has mapping	
has reference	
produces	is produced by

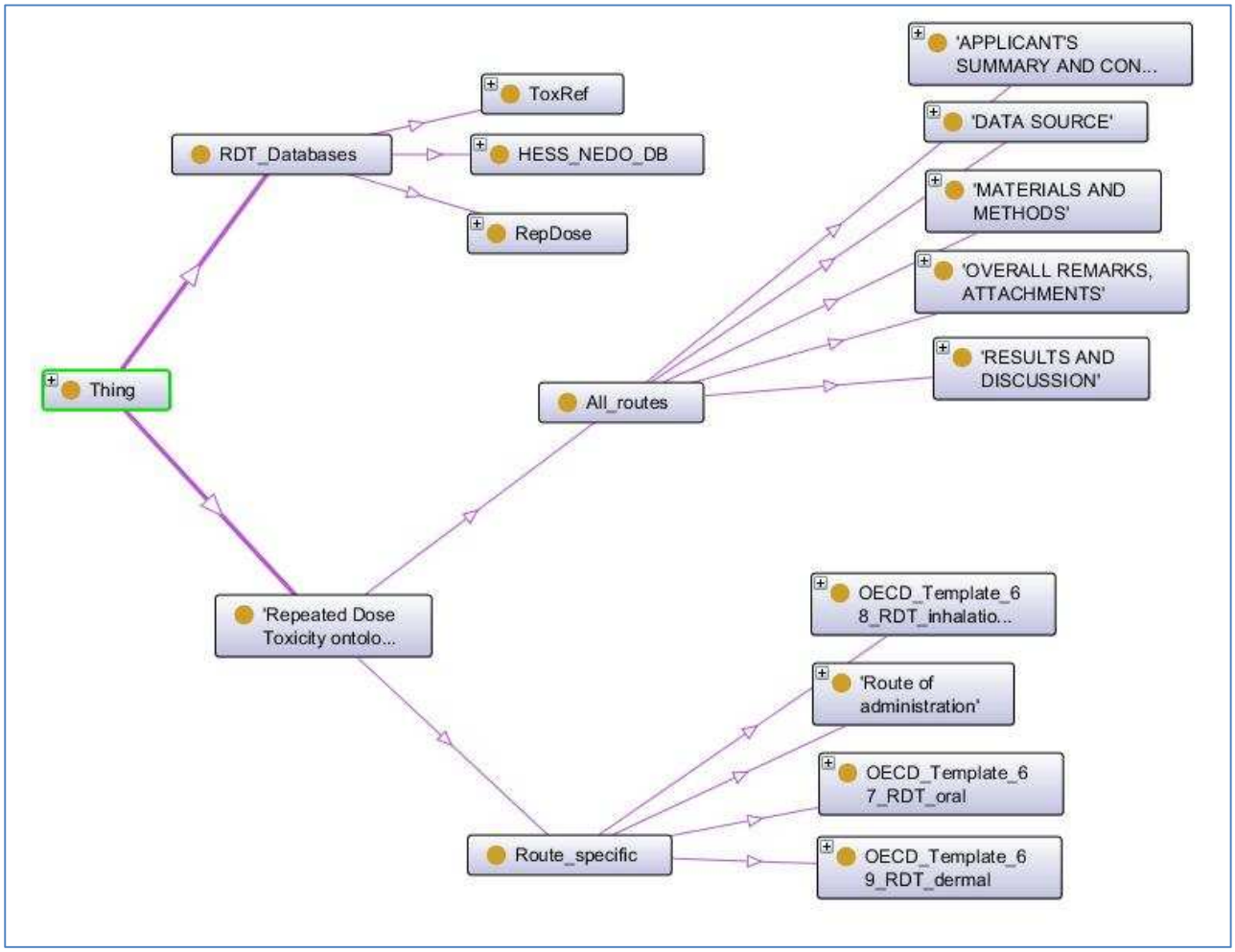


Figure 20.1 *Main classes of the repeated dose toxicity ontology based on the OHTs and the databases ToxRef, HESS-NEDO and RepDose included in the QSAR Toolbox.*

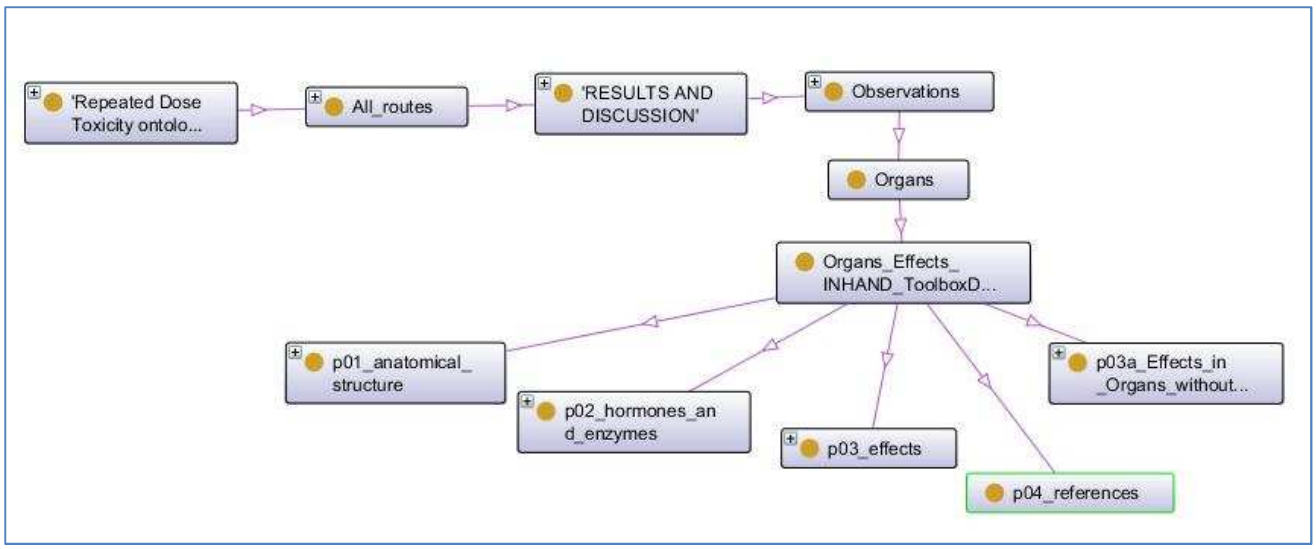


Figure 21.2 *Implementation of organs and effects ontology based on INHAND and the repeated dose toxicity databases of the QSAR Toolbox within the structure of the OHTs.*

3.2 Compilation of terms related to the endpoint

Terms were collected from: OHTs, INHAND manuscripts and pre-published lists, vocabulary of ToxRef DB, RDT DBs of the QSAR Toolbox.

The OHT #67, #68, and #69 have 87 terms in common and 91, 94 and 93 terms were collected per OHT respectively. For each term, the field number (specific identifier per term and template),the explanations and help text were implemented. Wherever indicated, the respective pick list terms were added to the ontology. Overall 15 pick lists are used by all three templates and the respective templates indicate the use of 23, 25 and 24 pick lists per OHT, sometimes at multiple sites.

The INHAND manuscripts and pre-published lists comprise terms for 940 single effects. In addition to these terms, synonyms and modifiers are given.

The vocabulary of the ToxRef DB includes 16 345 terms for different sites of observation.

Within the QSAR Toolbox, the databases use terms to describe the study design, report and effect. In RepDose DB these are 434 terms, HESS_NEDO has 700 terms and additional modifiers and the repeated dose toxicity version of the ToxRef contains 47 terms.

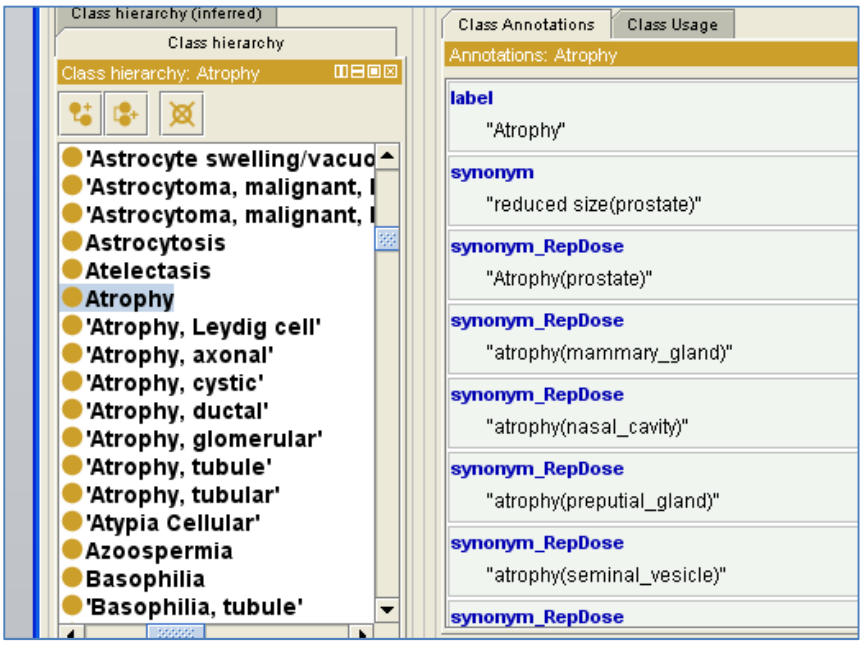
Overall about 20 000 terms are implemented in the repeated dose toxicity ontology.

3.3 Definition of synonyms and homonyms

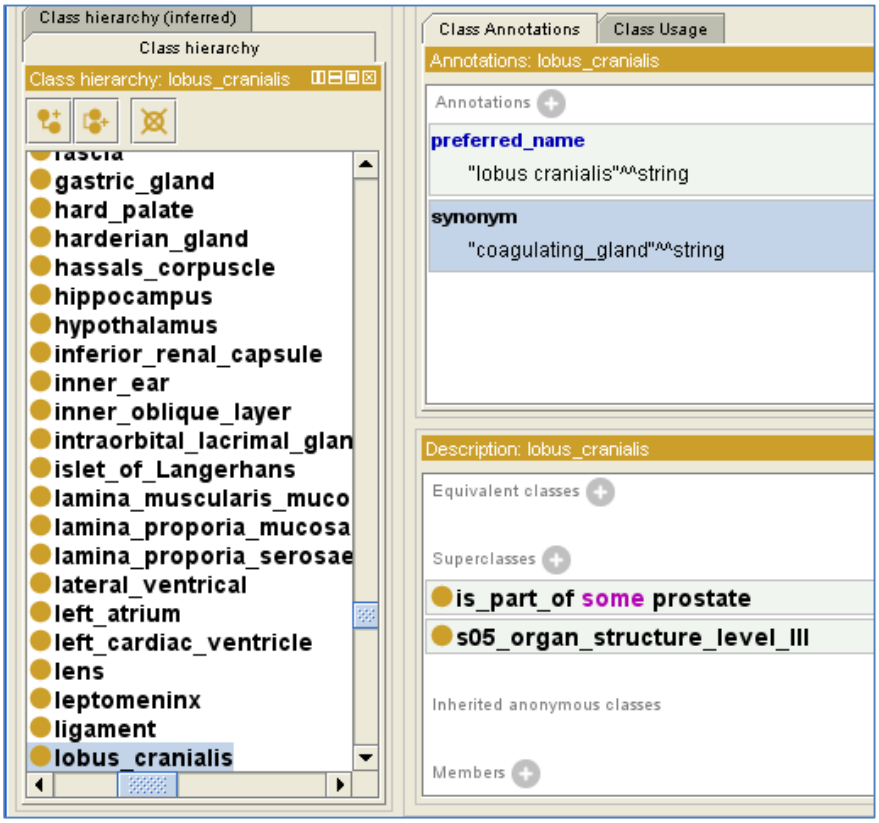
The terminology of the three RDT databases implemented in the QSAR Toolbox has to be standardized as only a harmonized terminology enables simultaneous use of the DBs for toxicity prediction. The OHTs terms or, if not available, the INHAND terminology, are used as standards. If the respective organ system has not yet been published by INHAND, the terms, e.g. for effects on female reproductive organs, were collected from the RDT databases of the QSAR Toolbox and the most frequent term was selected as preferred name and the others setas synonyms. The synonyms definition for single terms for different datasets has been done introducing the web ontology language (OWL) annotation property: “has_synonym” for synonyms from the INHAND manuscripts, “has_synonym_ToxRef” for synonyms from the ToxRef DB, “has_synonym_Hess-Nedo” and “has_synonym_RepDose” for the respective databases. In addition to the synonyms term, the respective organ for which this synonym is used is given in brackets. Two examples of such synonymous definition are shown in

Error! Reference source not found.: (a) Atrophy definition with different annotations for synonyms and (b) the lobuscranialis an example from the organs ontology for organs synonymous definition.

In some cases the definition of synonyms was not even needed as seen in **Error! Reference source not found.**, as the same term may be used in different sources.

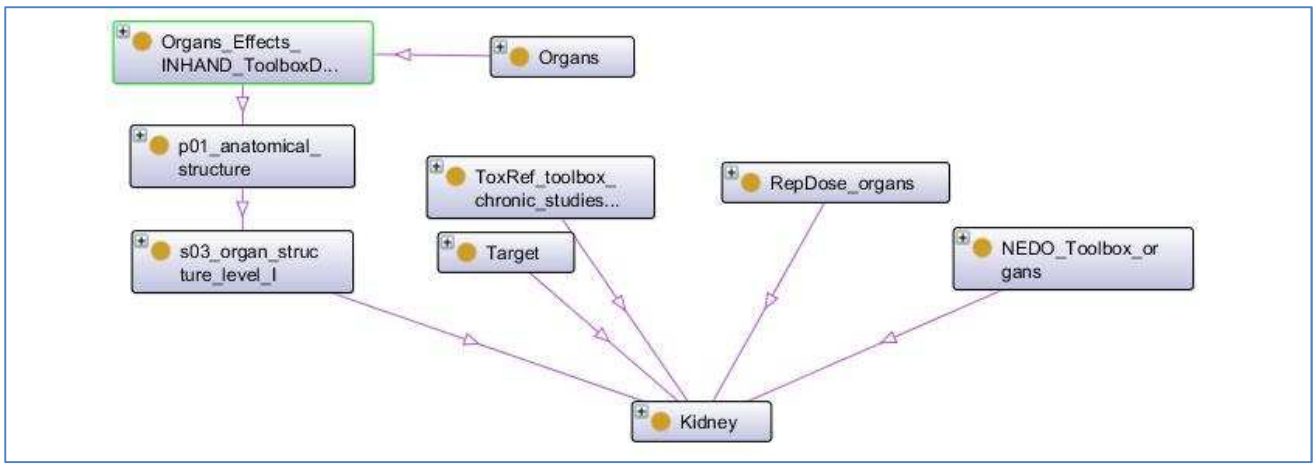


a)

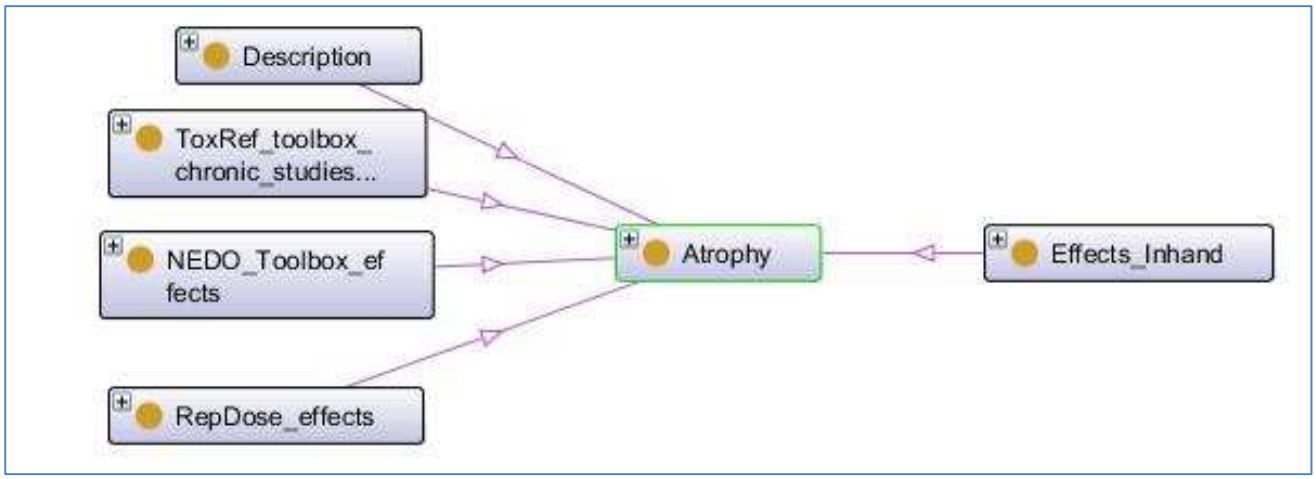


b)

Figure 22.3. Two examples of synonyms definition: (a) Atrophy and (b) Lobus cranialis in the respective effect and organs ontology.



a)



b)

Figure 23.4. The same term used within different sources is indicated by subclass relation and makes the definition of synonyms obsolete.

3.4 Establishment of relationships, interactions and hierarchies between classes, object and numeric properties for each term and rules

The definition of classes as described in **Error! Reference source not found.** includes the implementation of a general hierarchy. Furthermore the implemented object properties (**Error! Reference source not found.**) are used to define respective relationships.

Error! Reference source not found. Error! Reference source not found.shows an example of rule using the object properties “is_effect_in” and “is_example_for” for Adenocarinoma in the INHAND based ontology.

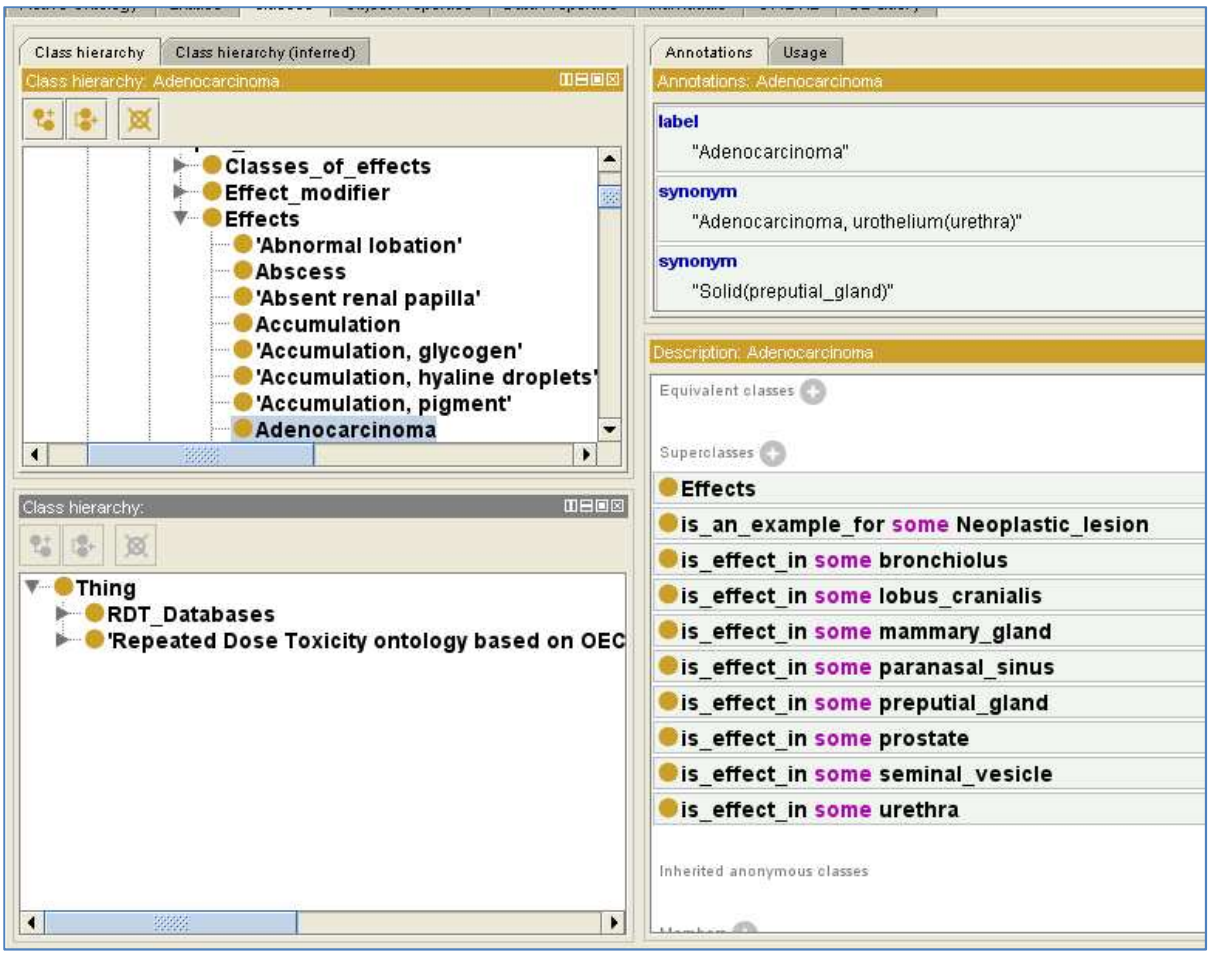


Figure 24.5. An example of rule using the object properties “is_effect_in” and “is_an_example_for” for Adenocarcinoma as effect in the INHAND based ontology for organs and effects.

3.5 Association of each attribute in a toxicological dataset with an entry in the ontology

The exact mapping between the same terms coming from ToxRef, HESS-NEDO, RepDose DB and the OECD Template #67-69: repeated dose has been performed by associating each attribute in a toxicological dataset with an entry in the ontology.

The OWL property “has_mapping” has been used for this purpose. In some cases, the same term is used in several databases or the OECD vocabulary, thus the term belongs to several classes representing its sources. In these cases no mapping is needed. The naming of the OWL property for mapping has been changed according to the reviewers’ comments.

Error! Reference source not found.shows an example of mapping of the datasets entries (RepDose:”rat”, HESS-NEDO: “rat” and ToxRef:“rattusnorvegicus”) to the ontology entry “rat”.

Figure 26. shows an example of mapping of ToxRefDB entries to the OHT based ontology.

Error! Reference source not found. shows an example of mapping the effects of several databases to the OHT.

Since there is no pick list for the organ or for the effect term in the OHT, it was necessary to set up an ontology containing the target organ list. We have based our ontology on the available terms in OHTs, manuscripts from INHAND and the databases collected in the Toolbox. The organs are defined within p01_anatomical_structure (**Error! Reference source not found.**) and have mainly been based on the ontology developed within the OPENTOX Project with some additions from the “Mouse adult gross anatomy ontology”

http://purl.bioontology.org/ontology/MAand the OHTs. **Error! Reference source not found.**shows an example of mapping of HESS-NEDO entry “Adrenal” to the organs ontology “adrenal_gland” which is also an organ of RepDoseDB.

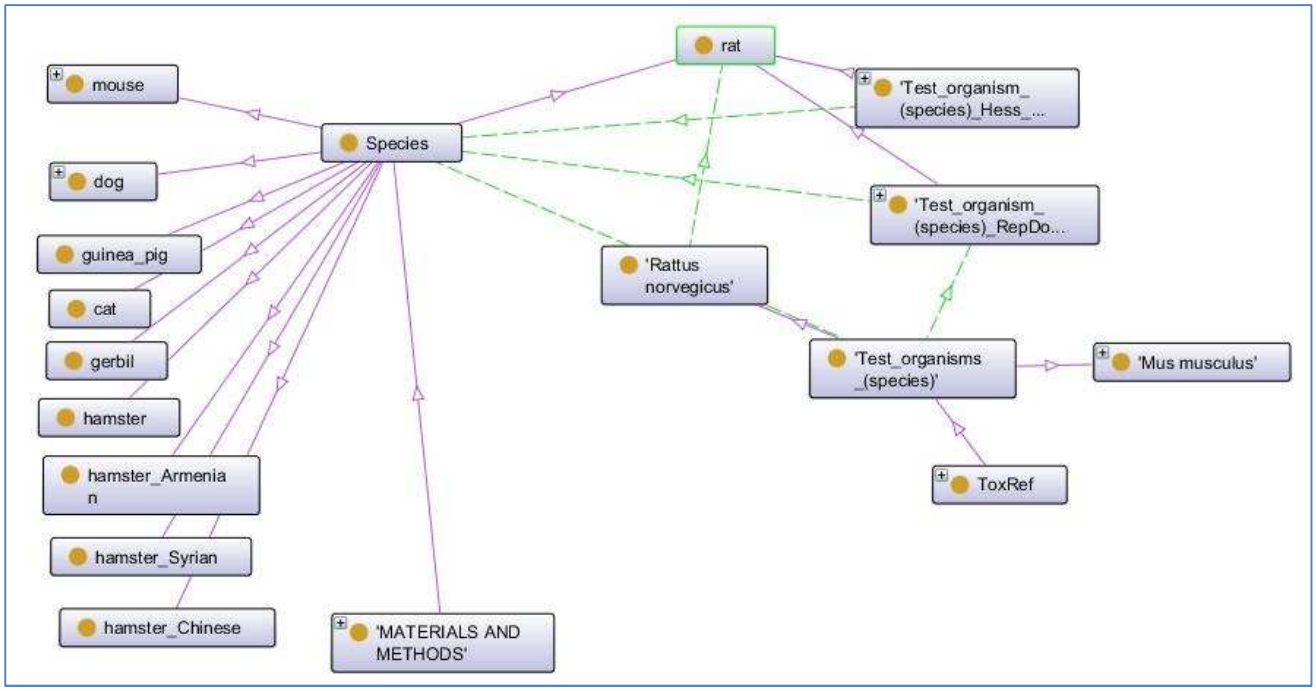


Figure 25.6 Example of mapping of the datasets entries (RepDose:”rat”, HESS_NEDO:“rat” and ToxRef: “rattusnorvegicus”) to the ontology entry (“rat”).

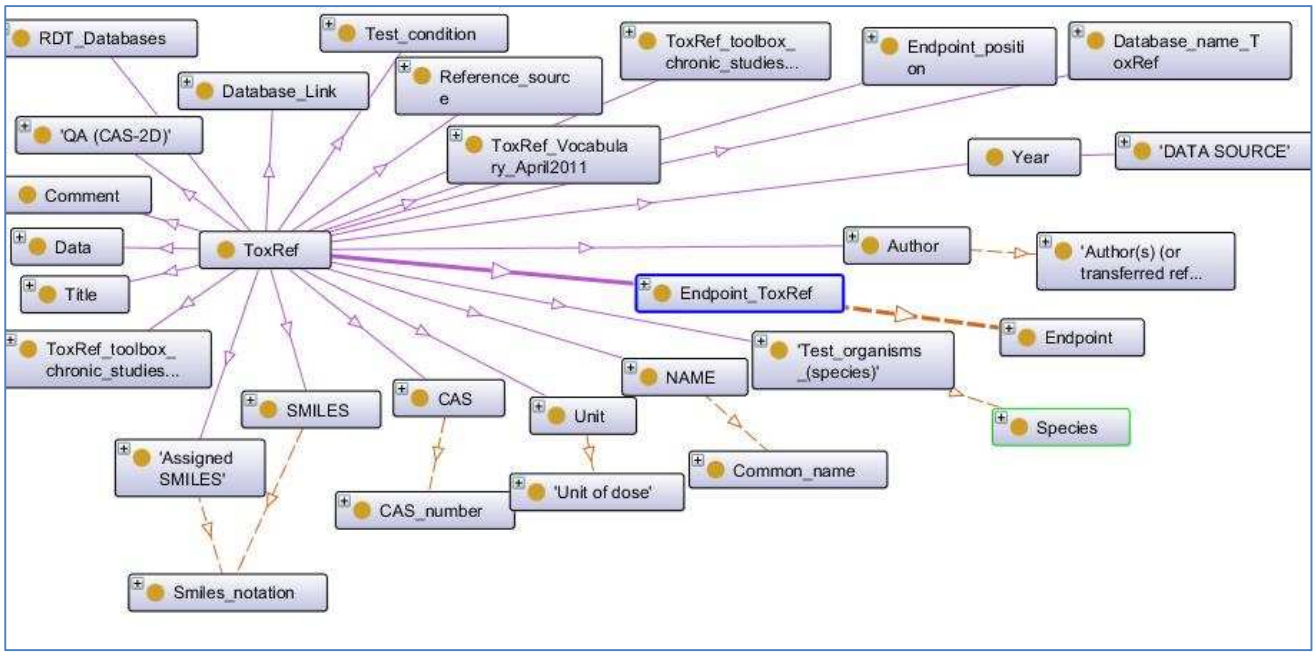


Figure 26.7 Example of mapping of ToxRef entries to the ontology based on the OHT

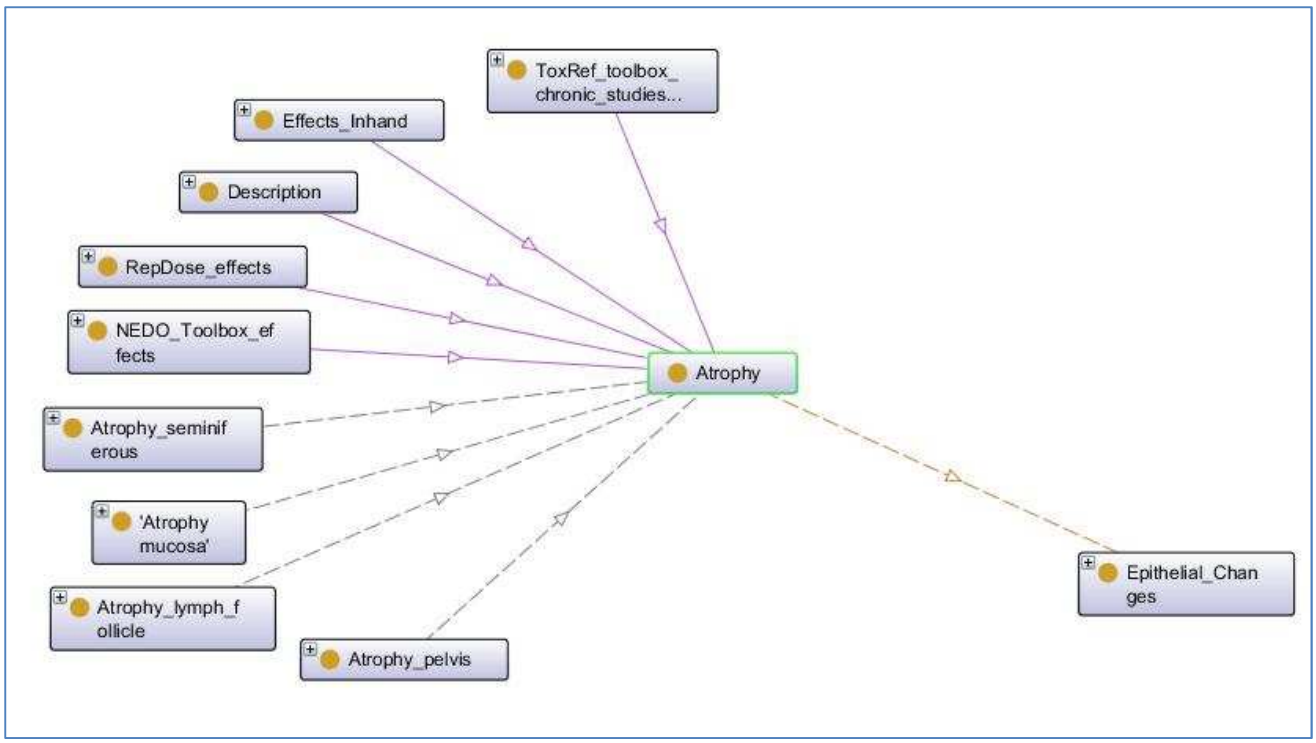


Figure 27.8 Example of mapping of effects of several databases entries to the OHT

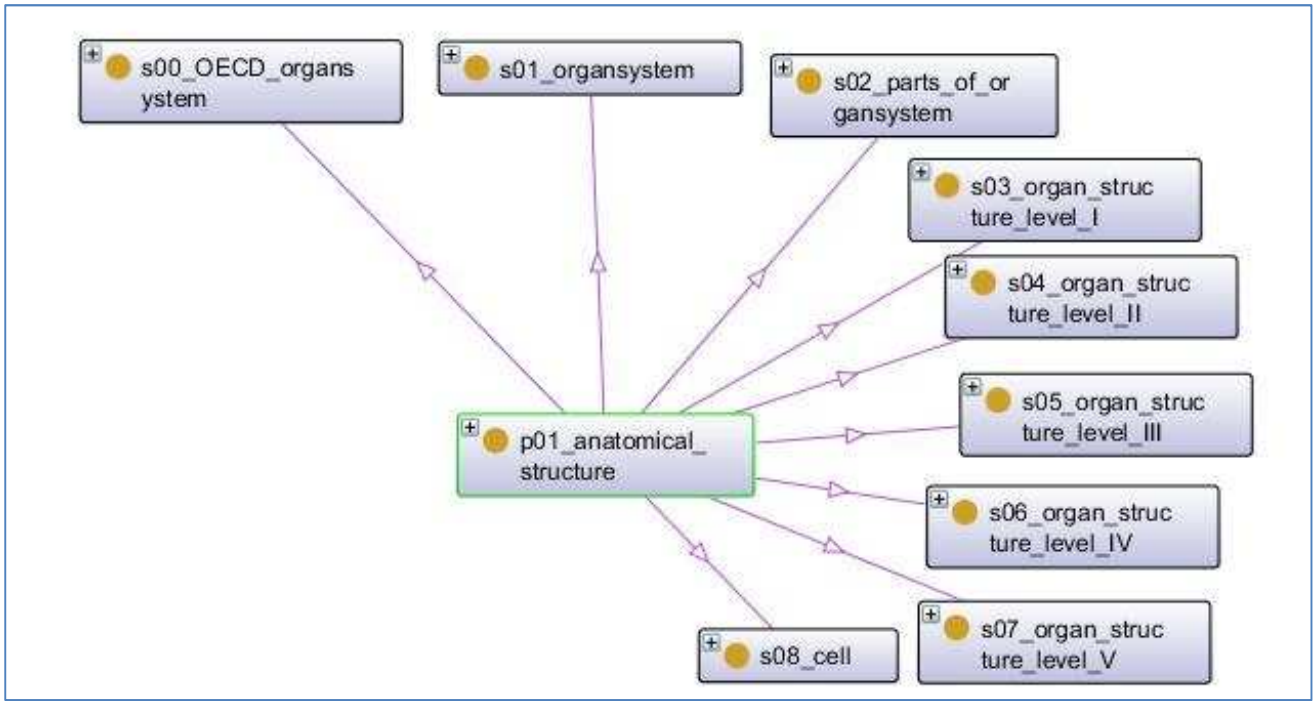


Figure 28.9 Organ ontology structure

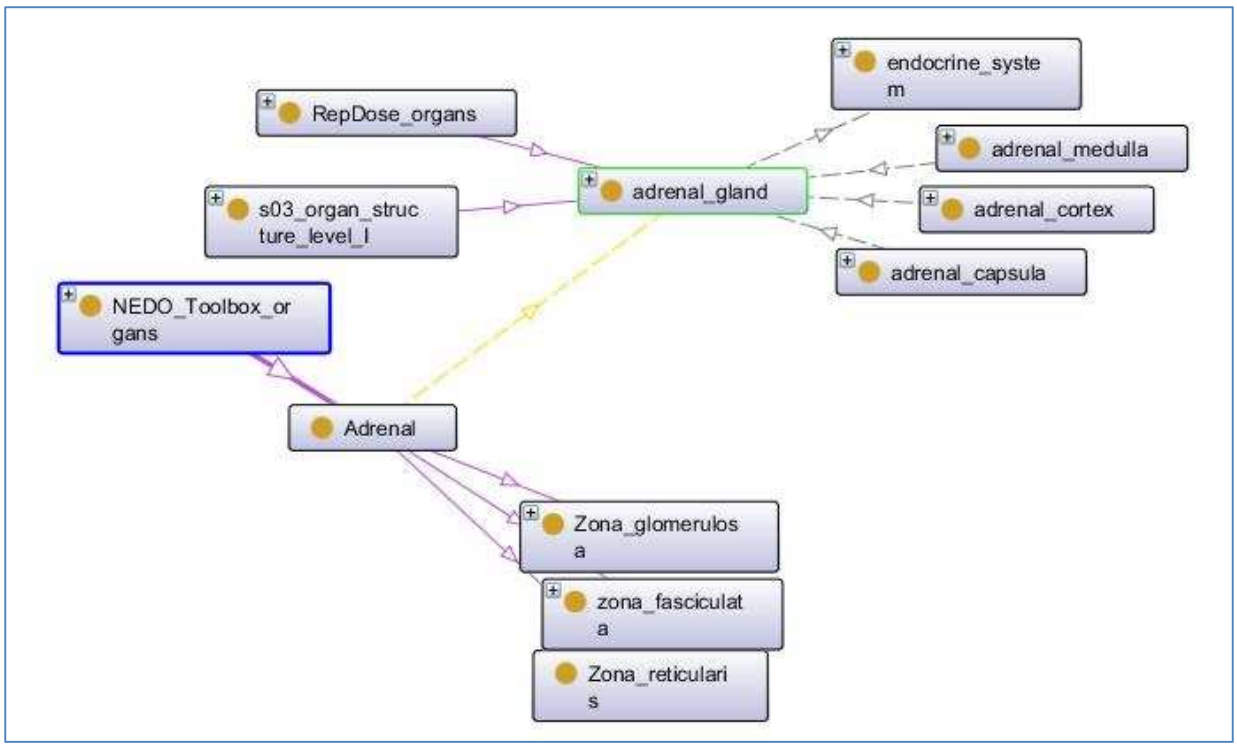


Figure 3.10 Mapping of Hess-Nedo entry "Adrenal" to the OPENTOX and INHAND based organ ontology entry "adrenal_gland", which is also an entry in RepDoseDB.

3.6 Intermediate ontology

The OWL file can be opened for view/edits using the open source OWL editor Protégé version 3.4.7 http://protegewiki.stanford.edu/wiki/Protege_3.4.7_Release_Notes.

We have tried as much as possible to maintain the original terminology of the Repeated dose toxicity: oral, inhalation, dermal (the latest versions[35,36,37])

As it has been planned in the inception phase, the general classes of the OHTs have been expanded into the subclasses adding the relating terminology from different sources:

a) OECD picklists:

-http://www.oecd.org/document/9/0,3746,en_21571361_43392827_45625801_1_1_1_1,00.html

b) existing ontologies freely available at the Bio-portal ontology repository[38]:

-Mouse adult gross anatomy[39]

c) we have re-used partially the toxicological ontology elaborated within the OPENTOX project (www.opentox.org) including the ITEM Organs system and Effects ontology based on the INHAND initiative. The

35 http://www.oecd.org/dataoecd/57/1/47299664.html

36 http://www.oecd.org/dataoecd/57/5/47299823.html

37 http://www.oecd.org/dataoecd/57/50/47300168.html

38 http://bioportal.bioontology.org/

39 http://purl.bioontology.org/ontology/MA

remainder of the OPENTOX ontology was not adaptable to the OECD HT because of being based on a different standard which is the Leadscope ToxML schema[40].

The main sub-classes of the first ontology class "Repeated Dose Toxicity ontology based on OECD Templates #67, #68, #69: Repeated dose: oral, inhalation, dermal" are "All routes" and "Route specific". ,"All routes" the homologous terms of the #67-69: Applicant's summary and conclusion, Data source, Materials and Methods, Overall Remarks, Attachments and Results and Discussion that aim at describing in detail the toxicological study. "Route specific" contains terms coming from single templates and Route of administration of template #67 and #68.

The second ontology class is "RDT Databases", which includes detailed description of the toxicity datasets that are included into the OECD QSAR Toolbox software. These are HESS-NEDO, RepDoseDB and ToxRefDB. For the ToxRefDB, in addition to the Toolbox entries, the vocabulary (Version August 2011) is contained in the ontology but not mapped: ToxRef_Vocabulary_August2011.

Synonyms and homonyms collection has been performed. The synonyms definition for single terms for different datasets has been done introducing the OWL annotation property: "synonym/has_synonymHess-Nedo/RepDose/ToxRef".

Establishment of relationships, interactions and hierarchies between classes, object and numeric properties has been done using the OWL functionalities as given in **Error! Reference source not found.**.

The exact mapping between the same terms coming from HESS-NEDO, RepDose, ToxRef and the OECD Template #67-69 has been performed associating each attribute in a toxicological dataset with an entry in the ontology.

Even if the ontology addresses three specific datasets of the Toolbox, the general concept is remaining flexible. It will be easy to import other repeated dose databases and map them to the ontology, or update the ontology in the case of further extension of the Toolbox Software. The owl-file describes the hierarchies and relations using RDF/XML-based serializations, thus apart from using Protégé the file can also be opened in respective editors and used as an xml-template for any other operations.

3.7 Additional comments (OECD (Q)SAR Toolbox relevance, Harmonised Template relevance, modification made according to the reviewers' comments)

All repeated dose toxicity databases of the QSAR Toolbox with their terminology have been implemented in the current version of this ontology. The first level of subclasses within each database in the ontology represents the columns of the data tables for the respective database in the QSAR Toolbox. The terminology found in the single fields of these columns is on the second level of subclasses. The entities of each database are mapped to the respective classes of the OHTs. The OHTs are implemented with terms, field numbers, explanations, help texts, and pick lists. The original terminology of the databases and OHTs is completely maintained. The concept of the ontology is flexible and open for any future additional database to be mapped or changes within the current implementation to be made.

The original terminology of OHTs, databases and INHAND vocabulary had to be maintained in order to fulfill the requirements of the contract and to facilitate the implementation in the QSAR Toolbox. As the terminology does not meet the principles of naming conventions of current ontology developments, this was one of the major aspects discussed by the reviewers. Their points are fully understood but with the purpose of this ontology to be used in the QSAR Toolbox, this conflict cannot be solved at present.

40http://www.leadscope.com/toxml.php

As the concept of the ontology and its format (XML) is open for future additions, changes and developments new databases with their terminology, next INHAND publications and changes in the OHTs can be considered in follow-up projects.

3.8 References

1. Rennen R, Brix A, Harkema J, Herbert R, Kittel B, Lewis D, March T, Nagano K, Pino M, Rittinghausen S, Rosenbruch M, Tellier P, Wohrmann T: **Proliferative and nonproliferative lesions of the rat and mouse respiratory tract**. *ToxicolPathol*2009, **37**: 5S-73S, doi:10.1177/0192623309353423;
2. Thoolen B, Maronpot RR, Harada T, Nyska A, Rousseaux C, Nolte T, Malarkey DE, Kaufmann W, Küttler K, Deschl U, Nakae D, Gregson R, Vinlove MP, Brix AE, Singh B, Belpoggi F, Ward JM: **Proliferative and nonproliferative lesions of the rat and mouse hepatobiliary system**. *ToxicolPathol*2010, **38**: 5S-81S, doi:10.1177/0192623310386499
3. Frazier KS, Seely JC, Hard GC, Betton G, Burnett R, Nakatsuji S, Nishikawa A, Durchfeld-Meyer B, Bube A:**Proliferative and nonproliferative lesions of the rat and mouse urinary system**. *ToxicolPathol*2012, **40 (4):** 14S-86S, doi:10.1177/0192623312438736
4. Kaufmann W, Bolon B, Bradley A, Butt M, Czasch S, Garman RH, George C, Gröters S, Krinke G, Little P, McKay J, Narama I, Rao D, Shibutani M, Sills R:**Proliferative and nonproliferative lesions of the rat and mouse central and peripheral nervous systems**. *ToxicolPathol*2012, **40 (4)**: 87S-157S, doi:10.1177/0192623312439125
5. Rudmann D, Cardiff R, Chouinard L, Goodman D, Küttler K, Marxfeld H, Molinolo A, Treumann S, Yoshizawa K; for the inhand mammary, zymbal's, preputial, and clitoral gland organ working group: **Proliferative and nonproliferative lesions of the rat and mouse mammary, zymbal's, preputial, and clitoral glands**. *ToxicolPathol*2012, **40 (6)**: 7S-39S, doi:10.1177/0192623312454242
6. Creasy D, Bube A, de Rijk E, Kandori H, Kuwahara M, Masson R, Nolte T, Reams R, Regan K, RehmS, Rogerson P, Whitney K: **Proliferative and nonproliferative lesions of the rat and mouse male reproductive system**. *ToxicolPathol*2012, **40 (6)**: 40S-86S, doi:10.1177/0192623312454337
7. Nishikawa S, Yamashita T, Imai T, Yoshida M, Sakuratani Y, Yamada J, Maekawa A, Hayashi M: **Thesaurus for histopathological findings in publically available reports of repeated-dose oral toxicity studies in rats for 156 chemicals**. *J Toxicol Sci*2010,**35(3)**:295-8.
8. Martin M T, Judson R S, Reif D M,Kavlock R J, and Dix D J: **Profiling Chemicals Based on Chronic Toxicity Results from the U.S. EPA ToxRef Database**, *Environmental Health Perspectives*2009, doi:10.1289/ehp.0800074
9. Knudsen T B, Martin M T, Kavlock R J, Judson R S, Dix D J, Singh A V: **Profiling the Activity of Environmental Chemicals in Prenatal Developmental Toxicity Studies using the U.S. EPA's ToxRefDB**, *Reproductive Toxicology*2009, doi:10.1016/j.reprotox.2009.03.016
10. Martin M T, Mendez E, Corum D G, Judson R S, Kavlock R J, Rotroff D M, Dix D J: **Profiling the Reproductive Toxicity of Chemicals from Multigeneration Studies in the Toxicity Reference Database (ToxRefDB)**, *Toxicological Sciences*2009, doi:10.1093/toxsci/kfp080

4. Reproductive/Developmental Toxicity Ontology

The ontology for developmental and reproductive toxicity comprises and relates terms used in the currently two different databases of the QSAR Toolbox and global standards as there are the OECD Harmonized templates (OHTs[41]) and the Devtox.org terminology (the Devtox Project[42][1]). For the general organ system the respective INHAND[43] based ontology of the repeated dose toxicity ontology developed within this project was integrated. The mouse gross anatomy and development ontology[44] was added to assess possible overlap, but has a different general structure.

Currently the QSAR Toolbox contains two databases (DB) for developmental and reproductive toxicity:

- ILSI-Devtox DB – developed by ILSI[45]
- Tox Ref DB – developed by US EPA[46]

The OHTs define a structure on reporting toxicological studies but it has to be emphasized that they are not ontologies as they miss the definition of relationships and the naming conventions e.g. by OBO foundry (Open Biology and Biomedical Ontology) are not followed (e.g. usage of plural). For the purpose of the developmental and reproductive toxicity ontology, three templates in their versions of February 2012 were implemented: OHTs #73 – 74: Toxicity to reproduction/ Developmental toxicity / teratogenicity/ Toxicity to reproduction: other studies[47].The terms of the indicated OECD pick listswere added as required[48]. These OHTs for developmental and reproductive toxicity are mainly focused on describing the study design and report but they are very limited in defining terms for organs and corresponding effects. No pick lists for organs, developmental and reproductive parameters or effects are available but some terms for organ systems, organs, clinical observations and reproductive performance are included in the predefined tables[5].

We tried as much as possible to maintain the original terminology of the OHTs but as the organ and effect terminology is very sparse we used the terminology developed by Devtox.org and INHAND. The organ ontology was developed within the repeated dose ontology and is described further in section 3 of this report.The Devtox.org vocabulary contains detailed effect terminologies for external, visceral and skeletal effects as well as maternal-fetal findings. The version from August 16th 2012 was implemented as a standard in this ontology[2] [2].

The terminologies of the databases for developmental and reproductive toxicity implemented in the QSAR Toolbox were gathered from version 2.3 (ToxRef DB) and version 3.0 (ILSI DB) as well as from the original ILSI DB[4]. Additionally, the vocabulary of the ToxRefDB as provided by M. Martin (personal communication, Version of August 2011 received by email January 2012) was considered. The terminology used by the developers of the ILSI DB is a previous version of the Devtox.org vocabulary published in 1997 [3]. The ToxRef DB version in the QSAR Toolbox refers to the data sets provided by EPA related to their respective publications: Chronic & Cancer Endpoints [4], Developmental Toxicity Endpoints [5],and Reproductive Toxicity Endpoints [6]. In these data sets effect classes were implemented for the purpose of the respective publications and underlying statistical evaluations. These effect classes are not originally part of the ToxRef DB and are not implemented in

41www.oecd.org/ehs/templates/

42http://www.devtox.org/

43www.goreni.org/gr_back_inhand.php

44http://bioportal.bioontology.org/ontologies/1010

45http://www.ilsi.org/ResearchFoundation/Pages/DevelopmentalToxicityDatabase.aspx

46http://www.epa.gov/ncct/toxrefdb/

47http://www.oecd.org/ehs/templates/toxicitytoreproduction.htmhttp://www.oecd.org/ehs/templates/developmentaltoxicityteratogenicity.htm

http://www.oecd.org/ehs/templates/toxicitytoreproductionotherstudies.htm

48http://www.oecd.org/document/9/0,3746,en_21571361_43392827_45625801_1_1_1_1,00.html

the vocabulary. The ToxRef DB in its entirety is freely available for download from the EPA website[49] and contains the vocabulary as provided by M. Martin. Within the current ontology, the ToxRef DB terminology, as implemented in the QSAR toolbox, and in addition the vocabulary for organs and effects areprovided. The latter is not mapped to OHTs or other databases in order not to cause confusions about the terminology currently contained in the QSAR Toolbox. Some terms are, however, similar to other sources and thus these terms then belong to several classes including classes of ToxRef DB vocabulary and HESS-NEDO DB for instance.

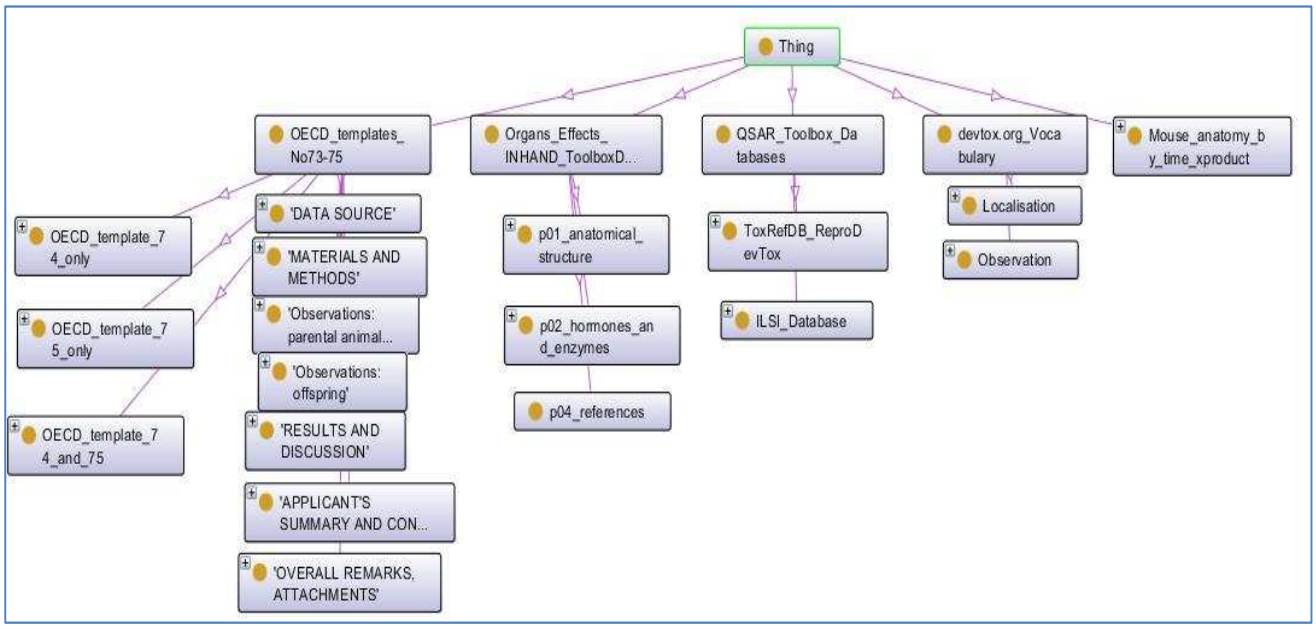
4.1 Definition of classes and general hierarchical relationships in the ontology structure

The first deliverable *4.1 Definition of classes and general hierarchical relationships in the ontology structure* resulted in the general structure of the ontology based on the OHTs and the reproductive and developmental toxicity DBs of the QSAR Toolbox (Figure 4.1).

The object properties shown in **Error! Reference source not found.** were implemented and used to define relationships.

Table 29.1 *Object properties implemented in the developmental and reproductive ontology for definition of relationships*

Object property	Inverse object property
has part	is part of
has effect	is effect in
has example	is an example for
has mapping	
has reference	
produces	is produced by



49 http://actor.epa.gov/actor/faces/Download.jsp

Figure 30.1 Main classes of the developmental and reproductive toxicity ontology based on the OHTs, Devtox.org vocabulary, Organs and Effects INHAND ontology, mouse anatomy ontology and the databases ToxRef and ILSI Devtox DB included in the QSAR Toolbox.

4.2 Compilation of terms related to the endpoint

Terms were collected from: OHTs, Devtox.org, ILSI DB and from the ToxRef DB as implemented in the QSAR Toolbox. The implementation of the ILSI DB was performed in parallel in this ontology as well as in the QSAR Toolbox. The final implementation in the QSAR Toolbox was provided to the ontology project group after the first deliverable of the developmental and reproductive ontology. Within this first delivery, the ILSI DB is structured according to the database content. In the final ontology the fields implemented in the QSAR Toolbox are indicated and mapped.

The OHT #73, #74, and #75 have 67 terms in common and 106, 83 and 72 terms were collected per OHT respectively. For each term the field number (specific identifier per term and template),the explanations and help text were implemented. Wherever indicated the respective pick lists terms were added to the ontology. Overall 23pick lists are used by all three templates and the respective templates indicate the use of 24, 26, and 26pick lists per OHT sometimes at multiple sides.

The Devtox.org vocabulary comprises terms for 1710 single effects at different localisations.
The vocabulary of the ToxRef DB includes 16 345 terms for different sites of observation.
Within the QSAR Toolbox the databases use terms to describe the study design, report and effect. In ILSI DB these are about 1500 terms and the version of the ToxRef contains 150 terms.
The INHAND organs ontology contains about 600 terms from organ systems to cellular level.
The mouse gross anatomy and development ontology has been added with about 13 700 entries.
Overall about 34 500 terms are implemented in the repeated dose toxicity ontology.

4.3 Definition of synonyms and homonyms

The terminology of the two RDT databases implemented in the QSAR Toolbox has to be standardized as only a harmonized terminology enables simultaneous use of the DBs for toxicity prediction. In principle the OHTs terms or if not available the Devtox.org or INHAND terminology are used as standards. The synonyms definition for effects of the ToxRef DB as implemented in the QSAR Toolbox was not possible as the terms used are not further specified effect groups such as “epididymis pathology and weight” which were only mapped to the respective organs and report classes. The terminology of effects in the ILSI DB was used for synonym definition of the Devtox.org vocabulary. These synonyms are implemented as OWL annotation property: “has_synonym_ILSI”. In addition to the synonyms term, the respective organ for which this synonym is used is given in brackets. An example of such synonyms definition is shown in **Error! Reference source not found.**: Absent with different annotations for synonyms.

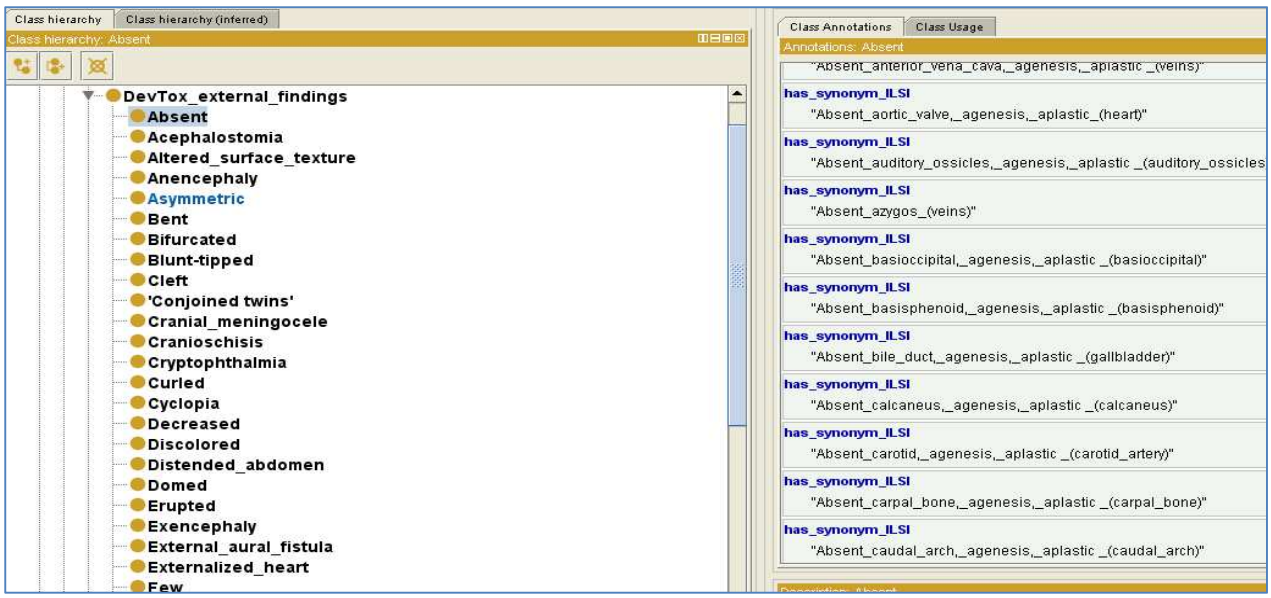


Figure 31.2 An example of synonyms definition: Absentfrom the Devtox.org vocabulary with different organ-specific synonyms in the ILSI DB.

4.4 Establishment of relationships, interactions and hierarchies between classes, object and numeric properties for each term and rules

The definition of classes as described in section 4.1 includes the implementation of a general hierarchy. Furthermore the implemented object properties (**Error! Reference source not found.**) are used to define respective relationships.

Error! Reference source not found. shows an example of rule using the object properties “is_effect_in” and “has_effect” for Cyst, a term used in the Devtox.org and the original ToxRef DB vocabulary.

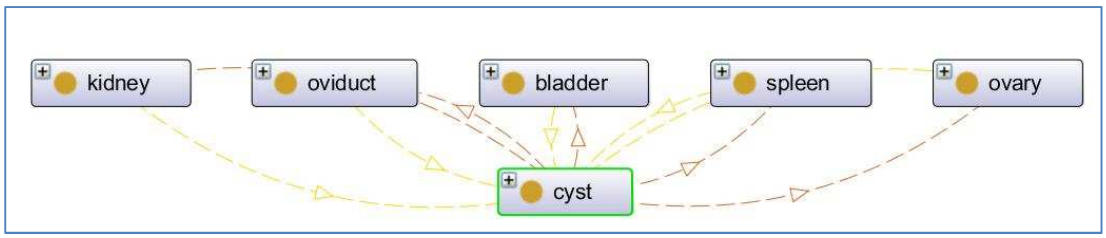


Figure 32.3 An example of rule using the object properties “is_effect_in” and “has_effect” for cyst (yellow line ‘has_effect’, orange line ‘is_effect_in’)

4.5 Association of each attribute in a toxicological dataset with an entry in the ontology

The exact mapping between the same terms coming from ToxRef and ILSI Devtox DBs and the OHTs #73-75 has been performed by associating each attribute in a toxicological dataset with an entry in the ontology.

The OWL property “has_mapping” has been used for this purpose. In some cases, the same term is used in several databases or the OECD vocabulary, thus the term belongs to several classes representing its sources. In these cases no mapping is needed.

Figure 4.4 shows an example of mapping of the datasets entries for species from the ToxRef DB and the use of the same terms in the ILSI Devtox DB and OHTs.

Figure 4.5 shows an example of mapping of ToxRefDB entries to the OHT based ontology.

Figure 4.6 shows an example of mapping the effects of several databases to the OHT.

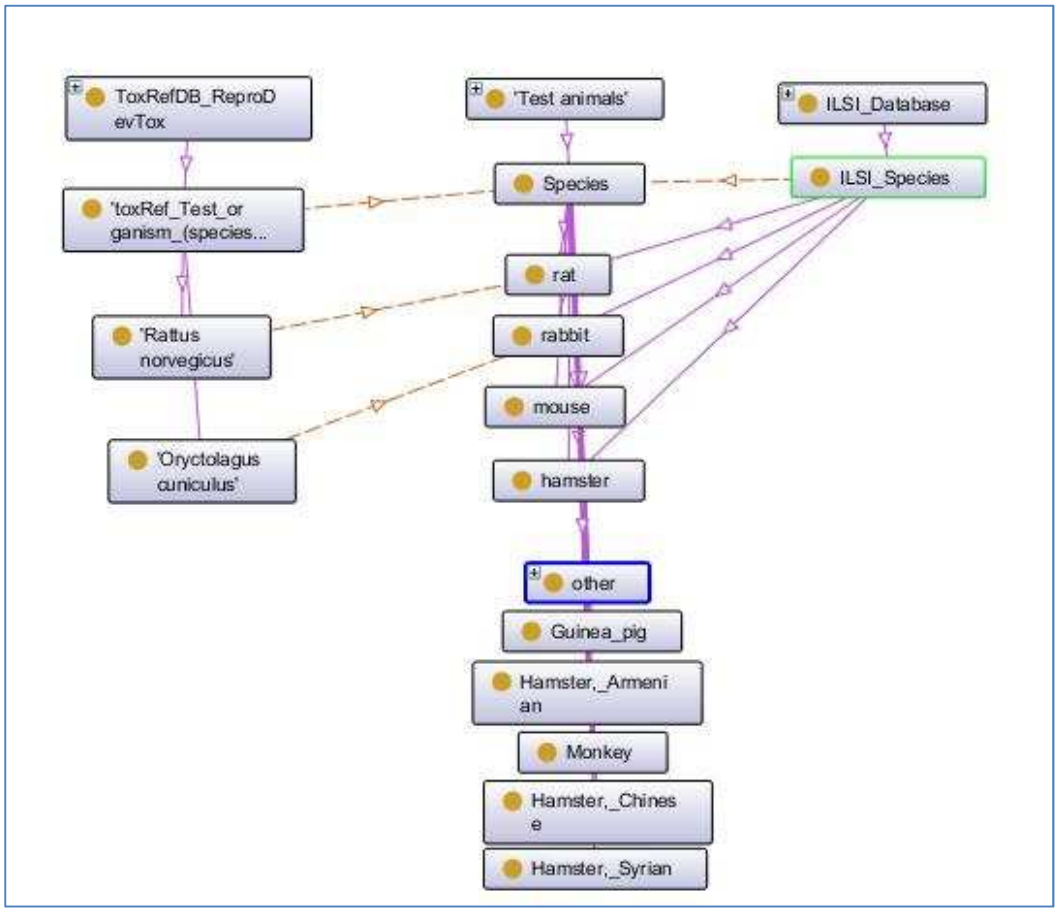


Figure 33.4 Example of mapping of the datasets entries for species: the terms for species of ILSI Devtox DB and the OHTs are the same (violet lines), whereas the terms of the ToxRef DB are different and therefore are mapped (orange lines).

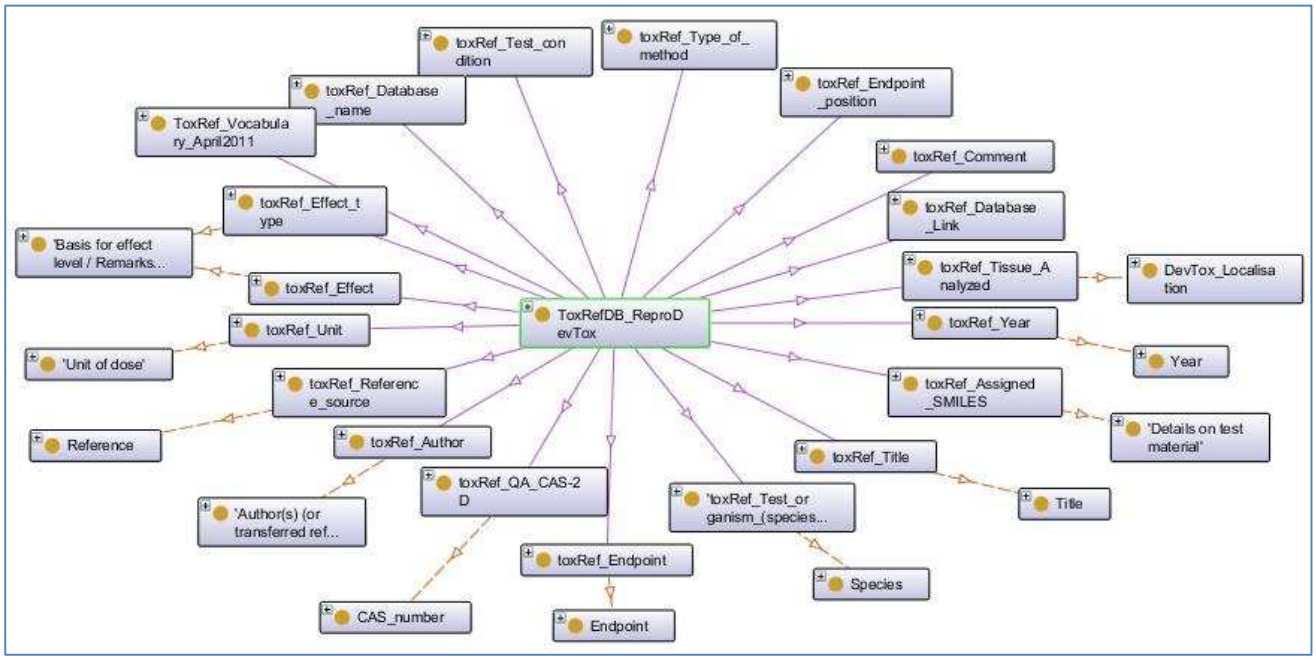


Figure 34.5 Example of mapping of ToxRef entries to the ontology based on the OECD HT

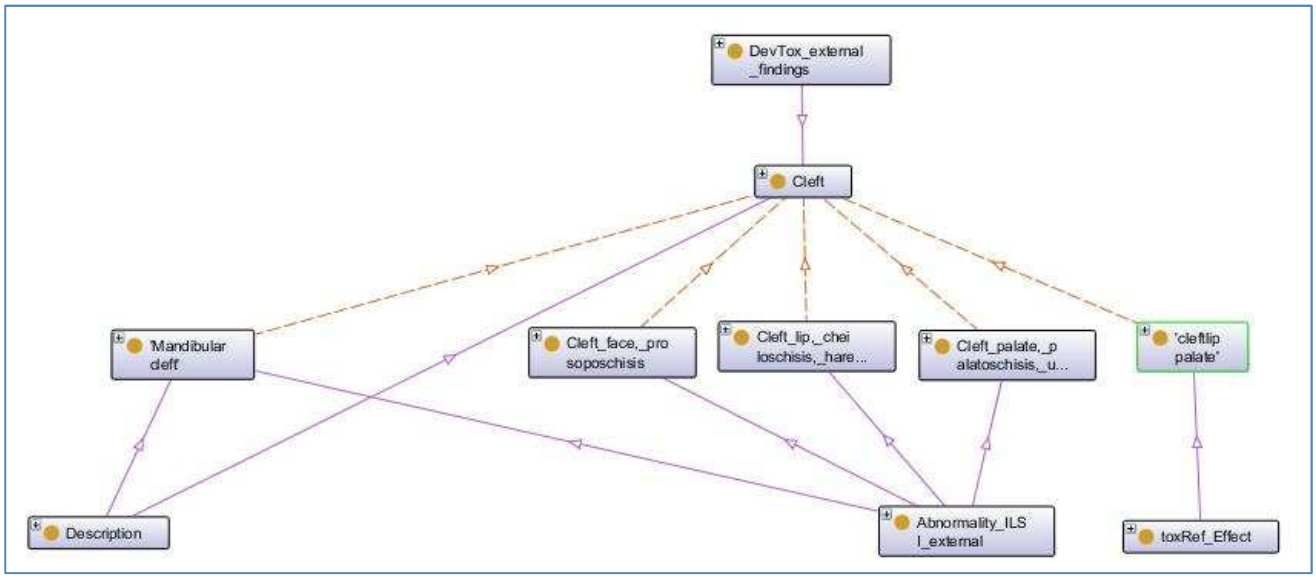


Figure 35.6 Example of mapping (orange lines) of effects of several databases to the DevTox.org Vocabulary: the term "Cleft" and associated terms in the ILSI Devtox DB ("Abnormality ILSI external") and the ToxRef DB ("toxRef_Effect" as implemented in the Toolbox and "Description" as in the full vocabulary of the ToxRefDB).

4.6 Intermediate ontology

The OWL file can be opened for view/edits using the open source OWL editor Protégé version 3.4.4, 3.4.7 and 4.1 which can be found at: http://protegewiki.stanford.edu/wiki/Category:VersionOfApplication

The respective applications can be downloaded after registration from: http://protege.stanford.edu/download/download.html

We have tried as much as possible to maintain the original terminology OHTs #73 - 74: Toxicity to reproduction/ Developmental toxicity / teratogenicity/ Toxicity to reproduction: other studies[50].

As it has been planned in the inception phase, the general classes of the OHTs have been expanded into the subclasses adding the relating terminology from different sources:

a) OECD picklists:

-http://www.oecd.org/document/9/0,3746,en_21571361_43392827_45625801_1_1_1_1,00.html

b) The DevTox.org Vocabulary as a globally harmonized standard for localisations and findings in developmental toxicity studies[51].

c) Existing ontologies freely available at the Bio-portal ontology repository[52]:

---

50 http://www.oecd.org/ehs/templates/toxicitytoreproduction.htmhttp://www.oecd.org/ehs/templates/developmentaltoxicityteratogenicity.htm

http://www.oecd.org/ehs/templates/toxicitytoreproductionotherstudies.htm

51 http://www.devtox.org

-Mouse gross anatomy and development[53]

c) As planned in the inception phase, we have re-used partially the toxicological ontology elaborated within the OPENTOX project[54] including the ITEM Organs system based on the INHAND initiative. The remainderof the OPENTOX ontology was not adaptable to the OECD HT because it was based on the different standard which is the LeadscopeToxML schema[55].

The main sub-classes of the "Reproductive and Developmental Toxicity ontology" are "DevTox_Vocabulary", "Mouse anatomy by time xproduct", "OECD templates No73-75", "Organs_INHAND" and "QSAR Toolbox Databases". "OECD templates No 73-75" contains the homologous terms of the #73-75 in "Applicant's summary and conclusion", "Data source", "Materials and Methods", "Overall Remarks, Attachments" and "Results and Discussion" that aim at describing in detail the toxicological study. Template specific terms are contained in the respective classes: "OECD template 74 and 75", "OECD template 74 only", and "OECD template 75 only". The ontology class "QSAR Toolbox Databases" includes detailed description of the reproductive and developmental toxicity datasets that are included into the OECD QSAR Toolbox software. These are "ILSI Devtox DB" and "ToxRef DB". For the ToxRefDB, in addition to the Toolbox entries, the full vocabulary (Version August 2011) is contained in the ontology but not mapped: "ToxRef_Vocabulary_August2011".Synonyms and homonyms collection has been performed. The synonyms definition for single terms for different datasets has been done introducing the OWL annotation property "has_synonym_ILSI".

Establishment of relationships, interactions and hierarchies between classes, object and numeric properties have been done using the OWL functionalities as given in **Error! Reference source not found.**.

The exact mapping between the same terms coming from ILSI Devtox DB, ToxRef DB and the OECD Template #73-75has been performed associating each attribute in a toxicological dataset with an entry in the ontology.

Even if the ontology addresses three specific datasets of the Toolbox, the general concept is remaining flexible. It will be easy to import other databases and map them to the ontology, or update the ontology in the case of further extension of the Toolbox Software. The owl-file describes the hierarchies and relations using RDF/XML-based serializations, thus, apart from using Protégé, the file can also be opened in respective editors and used as an xml-template for any other operations.

4.7 Additional comments (OECD (Q)SAR Toolbox relevance, Harmonised Template relevance, modification made according to the reviewers' comments)

All developmental and reproductive toxicity databases of the QSAR Toolbox with their terminology have been implemented in the current version of this ontology. The first level of subclasses within each database in the ontology represents the columns of the data tables for the respective database in the QSAR Toolbox. The terminology found in the single fields of these columns is on the second level of subclasses. The entities of each database are mapped to the respective classes of the OHTs. The OHTs are implemented with terms, field numbers, explanations, help texts, and pick lists. The original terminology of the databases and OHTs is completely maintained. The concept of the ontology is flexible and open for any future additional database to be mapped or changes within the current implementation to be made.

The original terminology of OHTs, databases and Devtox.org vocabulary had to be maintained in order to fulfill the requirements of the contract and to facilitate the implementation in the QSAR Toolbox. As the terminology does not meet the principles of naming conventions of current ontology developments, this was one of the

---

52http://bioportal.bioontology.org

53http://purl.bioontology.org/ontology/EMAP

54www.opentox.org

55http://www.leadscope.com/toxml.php

major aspects discussed by the reviewers. Their points are fully understood but with the purpose of this ontology to be used in the QSAR Toolbox, this conflict cannot be solved at present.

As the concept of the ontology and its format (XML) is open for future additions, changes and developments, new databases with their terminology, standard vocabularies for reproductive toxicity testing (if available) and changes in the OHTs can be considered in follow-up projects.

4.8 References

1. Solecki R, Heinrich V, Rauch M,Chahoud I, Grote K,Wölffel B,Buschmann J,Morawietz G, Kellner R,Lingk W:**The DevTox Site: Harmonized Terminology and Database.** In: Charlene A. McQueen, Comprehensive Toxicology 2010, **volume 12**, pp. 339-346, Oxford: Academic Press.
2. Solecki R, Barbellion S, Bergmann B, Bürgin H, Buschmann J, Clark R, Comotto L, Fuchs A, Faqi AS, Gerspach R, Grote K, Hakansson H, Heinrich V, Heinrich-Hirsch B, Hofmann T, Hübel U, Inazaki TH, Khalil S, Knudsen TB, Kudicke S, Lingk W, Makris S, Müller S, Paumgartten F, Pfeil R, Rama EM, Schneider S, Shiota K, Tamborini E, Tegelenbosch M, Ulbrich B, van Duijnhoven EA, Wise D, Chahoud I: **Harmonization of description and classification of fetal observations: Achievements and problems still unresolved: Report of the 7th Workshop on the Terminology in Developmental Toxicology Berlin, 4-6 May 2011**.ReprodToxicol. 2012 Jul 8. [Epub ahead of print]
3. Wise LD, Beck SL, Beltrame D, Beyer BK, Chahoud I, Clark RL, Clark R, Druga AM, Feuston MH, Guittin P, Henwood SM, Kimmel CA, Lindstrom P, Palmer AK, Petrere JA, Solomon HM, Yasuda M, York RG:**Terminology of developmental abnormalities in common laboratory mammals (version 1)**. Teratology1997; **55(4)**:249-92.
4. Martin MT, Judson RS, Reif DM, Kavlock RJ, and Dix DJ: **Profiling Chemicals Based on Chronic Toxicity Results from the U.S. EPA ToxRef Database**, Environmental Health Perspectives 2009, doi:10.1289/ehp.0800074
5. Knudsen TB, Martin MT, Kavlock RJ, Judson RS, Dix DJ, Singh AV: **Profiling the Activity of Environmental Chemicals in Prenatal Developmental Toxicity Studies using the U.S. EPA's ToxRefDB**, Reproductive Toxicology2009, doi:10.1016/j.reprotox.2009.03.016
6. Martin MT, Mendez E, Corum DG, Judson RS, Kavlock RJ, Rotroff DM, Dix DJ:**Profiling the Reproductive Toxicity of Chemicals from Multigeneration Studies in the Toxicity Reference Database (ToxRefDB)**, Toxicological Sciences2009, doi:10.1093/toxsci/kfp080

5. Appendixes

5.1 Appendix A: Protégé and ontology user's guide

All three ontologies have been developed in OWL format using the Protégé ontology editor[56]. Protégé is a free, open source software, supported by a strong community of developers and academic, government and corporate users, employing Protégé for knowledge solutions in areas as diverse as biomedicine, intelligence gathering, and corporate modelling.

To browse and edit the present ontologies, a recommended version to download is Protégé version 3.4.8[57] Detailed information on how to install Protégé-OWL is available at the Protégé wiki webpage[58] . In order to load an OWL file (e.g., one of the ontologies created within this project), choose: File, Open, specify the location of the OWL file in the "Open Project" dialog, and click OK.

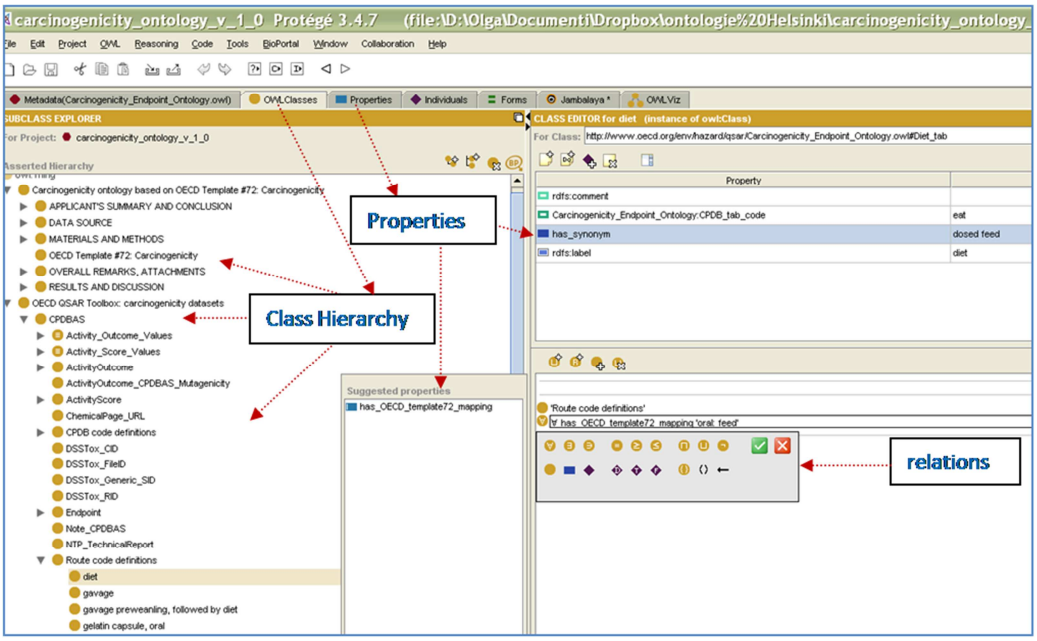


Figure 36.1 Carcinogenicity ontology in Protégé: subclasses tree, properties, restriction rules application.

As it displayed in Fig. 5.1, the classes have been organised into a superclass-subclass hierarchy, also known as a taxonomy. The classes and their hierarchy in our ontologies correspond to the OECD Harmonised templates and the OECD (Q)SAR toolbox terminology. The classes relative to the experimental datasets vocabulary have hierarchy corresponding the original database organization. For example, "Route code definitions" class from the CPDB dataset has different subclasses: "diet", "gavage", etc…

Properties are used in OWL to describe and define classes. For example, property "has_synonym" defines the synonym of class: in Figure 5.1 "dosed feed" is the synonym of "diet" route.

Properties in OWL are used also to create restrictions or relations between classes. For example the property "has_OECD_template72_mapping" is used to relate the term of the experimental dataset "diet" to the "oral:feed" term of the OECD HT (see Figure 5.1).
The ontology component can be easily browsed using Protégé. For more detailed information we recommend the practical guidance [59] describing how to build ontologies using Protégé.

---

[56] http://protege.stanford.edu

[57] http://protege.stanford.edu/download/registered.html

[58] http://protegewiki.stanford.edu/wiki/Protege-OWL_3_FAQ#How_do_I_install_Protege-OWL.3F

[59] http://home.skku.edu/~samoh/class/sw/ProtegeOWLTutorial.pdf

5.2 Appendix B: Point to point responses to the reviewers' comments

Please see the supplemental file for the Appendix B.

5.3 Appendix C: Proposals for the OECD (Q)SAR Toolbox trees modification

Please see the supplemental file for the Appendix C.