Toxicology Ontology Perspectives

Barry Hardy [1], Gordana Apic [2], Philip Carthew [3], Dominic Clark [4], David Cook [5], Ian Dix [5,6], Sylvia Escher [7], Janna Hastings [4], David J. Heard [8], Nina Jeliazkova [9], Philip Judson [10], Sherri Matis-Mitchell [5], Dragana Mitic [2], Glenn Myatt [11], Imran Shah [12], Ola Spjuth [13], Olga Tcheremenskaia [14], Luca Toldo [15], David Watson [10], Andrew White [3], and Chihae Yang [16]

[1]Douglas Connect and OpenTox, Zeiningen, Switzerland; [2]Cambridge Cell Networks, Cambridge, UK; [3]Unilever, Sharnbrook, Beds, UK; [4]EMBL-EBI, European Bioinformatics Institute, Cambridgeshire, UK; [5]AstraZeneca, Macclesfield, Cheshire, UK; [6]Pistoia Alliance; [7]Fraunhofer Institute for Toxicology and Experimental Medicine, Hannover, Germany; [8]Novartis, Basle, Switzerland; [9]Ideaconsult, Sofia, Bulgaria; [10]Lhasa Limited, Leeds, UK; [11]Leadscope, Columbus, OH, USA; [12]US EPA, Research Triangle Park, NC, USA; [13]University of Uppsala, Uppsala, Sweden; [14]Istituto Superiore di Sanità, Rome, Italy; [15]Merck KGaA, Darmstadt, Germany; [16]Altamira, Columbus, OH, USA

Summary

The field of predictive toxicology requires the development of open, public, computable, standardized toxicology vocabularies and ontologies to support the applications required by in silico, in vitro, and in vivo toxicology methods and related analysis and reporting activities. In this article we review ontology developments based on a set of perspectives showing how ontologies are being used in predictive toxicology initiatives and applications. Perspectives on resources and initiatives reviewed include OpenTox, eTOX, Pistoia Alliance, ToxWiz, Virtual Liver, EU-ADR, BEL, ToxML, and Bioclipse. We also review existing ontology developments in neighboring fields that can contribute to establishing an ontological framework for predictive toxicology. A significant set of resources is already available to provide a foundation for an ontological framework for 21st century mechanistic-based toxicology research. Ontologies such as ToxWiz provide a basis for application to toxicology investigations, whereas other ontologies under development in the biological, chemical, and biomedical communities could be incorporated in an extended future framework. OpenTox has provided a semantic web framework for the implementation of such ontologies into software applications and linked data resources. Bioclipse developers have shown the benefit of interoperability obtained through ontology by being able to link their workbench application with remote OpenTox web services. Although these developments are promising, an increased international coordination of efforts is greatly needed to develop a more unified, standardized, and open toxicology ontology framework.

Keywords: toxicology, ontology, framework, data resources, software applications

1 Introduction

In this article we will review ontology developments based on a set of perspectives that shows how ontologies are being used in current predictive toxicology initiatives and applications. We also review existing ontology developments in neighboring fields that can contribute to establishing an ontological framework for predictive toxicology. Such an ontology framework offers the potential for unifying existing information and concepts in the development of 21st century toxicological science in support of the Reduction, Refinement, and Replacement (3Rs) principles.

1.1 Definition of ontology

From its original philosophical roots, the word and idea of ontology has been adopted by the sciences as a formal representation of a set of concepts within a knowledge domain and the relationships between those concepts. It is used to reason about the properties of that domain, and it may be used to define the domain. Thus, an ontology[1] is a "formal, explicit specification of a shared conceptualization." An ontology provides a shared controlled vocabulary that can be used to model a domain, i.e., the type of objects and/or concepts that exist, and their properties and relations. The controlled vocabulary is a collection of

Received October 31, 2011; accepted in revised form February 2, 2012

[1] http://en.wikipedia.org/wiki/Ontology_(information_science)

preferred terms that are used to assist in more precise retrieval of content. Controlled vocabulary terms can be used for categorizing content, building labeling systems, and creating style guides and database schemata. A taxonomy is one type of controlled vocabulary. Ontologies are used in artificial intelligence, the Semantic Web[2], software engineering, biomedical informatics, library science, and information architecture as a form of knowledge representation about the world or some part of it.[3] Ontology-based informatics approaches provide very powerful tools to organize and retrieve information.[4]

In biology, the explosion in the amount of data generated through high-throughput techniques has led to the need to organize these data in a logical way; to this purpose, the concept of ontology has been widely adopted. The construction of a biological ontology consists of: a) selection of entities or objects to be included; b) definition of a controlled vocabulary; c) recognition of relationships, interactions, and hierarchies between entities, if any. The definition of ontology and controlled vocabulary is very useful in standardizing and organizing chemical, biological, and toxicological data required for predictive toxicology use cases. It may also be used to define biological effects and entities, systems and their components and interactions, algorithms and models, pathways, and other useful conceptual entities for supporting complex reasoning about concepts, questions, and data. Such an ontology not only supports superior, more reliable, and cost effective computer engineering but also enhances the value of data in a leveraged knowledge management ecosystem. Toxicology ontology could also encourage the Weight of Evidence-based integration of *in vivo* and *in vitro* data from several sources, as well as (Q)SAR, i.e., (Quantitative) Structure-Activity Relationship, data for regulatory processes such as risk assessment under REACH or for classification and labeling (CLP). In addition to alternative testing method representation, a toxicology ontology should include the representation of traditional animal experimental data, for its effective knowledge integration as evidence. Support of a toxicology ontology evolution framework should be adequate for translating the diverse terminology developed and used over time to link study findings across chemicals, labs, and study types.

One of the most successful examples of a systematic description of an area of biology is the Gene Ontology (GO) project.[5] GO is widely used in biological databases, annotation projects, and computational analyses for annotating newly sequenced genomes, text mining, network modeling, and clinical applications, among others. GO has two components: the ontologies themselves, which are the defined terms and the structured relationships between them (GO ontology) and the associations between gene products and the terms (GO annotations). GO provides both ontologies and annotations for three distinct areas of cell biology: molecular function, biological process, and cellular component or location. A GO annotation associates a gene with terms in the ontologies and is generated either by a curator or automatically through predictive methods. Genes are associated with as many terms as appropriate, as well as with the most specific terms available to reflect what is currently known about a gene. When a gene is annotated to a term, associations between the gene and the term's parents are implicitly inferred. Because GO annotations to a term inherit all the properties of the ancestors of those terms, every path from any term back to its root(s) must be biologically accurate or the ontology must be revised. For example, if a gene is known to be specifically involved in "vesicle fusion," it will be annotated directly to that term, and it is implicitly annotated (indirectly) to all of its parents' terms, including "membrane fusion," "membrane organization and biogenesis," "vesicle-mediated transport," "transport," and so on, back to the root node. Thus, a gene annotated to vesicle fusion can be retrieved not only with this term, but also with all of its parent terms, increasing flexibility and power when searching for and making inferences about genes.

1.2 Ontology in predictive toxicology

The application of ontology in predictive toxicology is a relatively recent endeavor in which investigators are developing or applying ontologies towards the solution of toxicology problems. Use of ontology allows data to be organized and combined with metadata and vocabularies for study terms and experimental protocols. Communication between toxicology resources using shared ontology supports toxicology data integration in a more reliable way, thus allowing the construction of more complex queries regarding the data, such as linking data with such biological processes as gene regulation or pathway perturbations. OpenTox designed its framework to extensibly support the creation of a semantic web for toxicology in which ontology plays a key role in resource representations and communications (Hardy et al., 2010). The OpenTox framework supports the development and application of interoperable predictive toxicology software that supports the 3Rs principles, so ontology is a means, not an end, of the OpenTox initiative.

2. Perspectives

2.1 Context for ontology development

The set of perspectives described here offers a review of current developments in the use of ontology in predictive toxicology applications and resources. Many of these developments are a result of recent and emerging initiatives, and they provide insight into what is possible today, as well as future directions towards a more mature and integrated ontological framework, e.g., as envisioned by the Toxicology Ontology Roadmap (Hardy et al., 2012). Perspectives on different resources and independ-

2 http://semanticweb.org/wiki/Main_Page

3 http://en.wikipedia.org/wiki/Ontology_(information_science)

4 http://www.ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-mit-press-(with-citation).htm

5 http://www.geneontology.org/

ent initiatives are reviewed, including OpenTox, eTOX, Pistoia Alliance, ToxWiz, Virtual Liver, EU-ADR, BEL, ToxML, and Bioclipse. By bringing a number of significant independent perspectives together, we anticipated developing collaboration opportunities and a common vision for a framework and roadmap for toxicology ontology development.

Lessons learned from toxicity data modeling and insights for ontology development

Just a decade ago, publicly free structure-searchable databases were nonexistent, as were toxicity databases. In contrast to human genome research, where all the information was freely available, chemistry fell behind until the PubChem Project[6] was initiated in 2004. Databases were scarce and the field was faced with real "lack of data" issues. There were three earlier initiatives in the area of chemically-induced toxicity databases: Lhasa Limited's efforts sponsored by the International Life Science Institute (2000-2004), Leadscope's ToxML[7] LIST (LeadScope In Silico Toxicology) focus group (2001-2006), partially funded by a US NIST ATP grant, and US EPA's DSSTox[8]. The two initiatives from Lhasa Limited and Leadscope concentrated on structure searchable representation of toxicity studies, whereas the DSSTox put much emphasis on chemical structure representations. Later, DSSTox became the chemistry center for other EPA NCCT toxicity and risk assessment databases, including ToxRefDB[9] and ACToR[10]. Collaborations with the US FDA also played an important role in toxicity database activities: one consequence was the agreement between Lhasa Limited and Leadscope to promote ToxML as the data exchange standard since 2006. In those five years, the landscape changed at a remarkable speed.

Toxicity databases have a strategic importance in safety assessment, both in the private sector and in regulatory agencies. Well-designed databases very efficiently provide the information needed for data mining and preparing training sets for computational approaches. End users can save a tremendous amount of time by utilizing data when the data fields are designed logically in line with their workflow and when the curation criteria are documented. Publicly, the main driving forces for common databases have been the international regulatory initiatives of the Canadian Domestic Substance List (DSL)[11], REACH, and the 7th Amendment of the Cosmetic Directive, as well as US Tox21[12]/ToxCast initiatives to challenge the paradigm shift for regulatory efforts. As we are building many local and consortium databases, we are now faced with a slightly different version of the "lack of data" problem, which still persists, though for quite different reasons. Due to the initiatives mentioned above, many databases have been constructed in silos to address local and immediate needs, thereby creating many databases with diverse content and sources. We now have databases that can distribute data ranging from a one-at-a-time query to a mass SQL dump. However, for a user far removed from database administration, these resources remain extremely difficult to utilize and, consequently, the lack of access to data remains the same. The need to distribute data in an easily processable common format can be achieved by a data exchange standard.

Scientists in computational toxicology have consistent and strong opinions about their data needs from a "perfect" database: "*I'd like a database that is internet-based, allows me to slice and dice according to my own needs, lets me update and modify, and most of all is free*." This dream database, in fact, describes the process of a scientist transforming information into knowledge. To enable this ultimate purpose of a database, we will need to go further than just an exchange standard. Ontology conceptualizes and formalizes these processes. In addition, ontology-driven databases also can guide the knowledge discovery and formation.

Chemical ontology

Most computational toxicologists often deal with both sides of a knowledgebase, i.e., development and use. Through ontology, the aggregation of effects can be represented and the hidden relationships of effects and mechanistic knowledge can potentially be discovered. Ontology has been strongly promoted in biology, for example, through gene ontology, whereas the concept of ontology is still underdeveloped in chemistry, although ChEBI[13] has now undertaken this important task. Toxicity is the manifestation of biological responses to stressors, which are represented by chemical structures in chemically-induced toxicity. The modes of action between chemical structures and biological targets (e.g., proteins) often can be well described by substructural patterns, conformations, and shapes. Systematically constructing the logical objects for the chemical units, reflecting both structural configuration and physicochemical properties, will lead us closer to our goal – linking chemical structural features to biological activities. It is important that we move beyond the confines of connection tables for representing structural features. In this sense, the new subgraph representation method, CSRML (Chemical Subgraph Representation Mark-up Language)[14] can represent the biologically active substructures by physicochemical parameters, including

6 PubChem, http://pubchem.ncbi.nlm.nih.gov

7 Leadscope ToxML Schema, http://www.leadscope.com/toxml.php

8 DSSTox, http://www.epa.gov/ncct/dsstox/index.html

9 ToxRefDB, http://www.epa.gov/ncct/toxrefdb

10 ACToR, http://www.epa.gov/actor/

11 CDSL, http://www.ec.gc.ca/lcpe-cepa/

12 US Tox21, http://www.epa.gov/ncct/Tox21/

13 ChEBI, http://www.ebi.ac.uk/chebi/

14 CSRML, http://acscinf.org/docs/meetings/240nm/presentations/240nm77.pdf CSRML

charge or electronegativity differences, in addition to conventional annotations of atom and bond specifications.

2.2 The Pistoia Alliance and vocabulary services

The primary purpose of the Pistoia Alliance[15] is to streamline non-competitive elements of the life science workflow by the specification of common standards, business terms, relationships, and processes. Pistoia Alliance goals directed towards this purpose are:
– to allow this framework to encompass and support most pre-competitive work between the organizations;
– to support life science workflow prior to regulatory submission;
– to work with other standards organizations.

Pistoia's foundation in 2007 grew from a shared frustration by large pharmaceutical organizations with the lack of standards and interoperability in the life sciences informatics sector. Meanwhile, the development of web services and Web 2.0 provided an opportunity for decoupling proprietary data from technology. Publicly available structural and biological databases increasingly allow for a non-IP related analysis, and they can serve as a scientific test suite.

Vocabulary Standards Initiative

A key aspect of scientific work involves naming and identifying a variety of entities: genes, sequences, diseases, tissues, anatomy, cell lines, bioprocesses, species, phenotypes, compounds, drugs, mode of action, people, places, adverse events, targets, reagents. Moreover, the science is continually evolving, and so the vocabularies are in constant flux. Connecting data across content providers, academic and internal organizational systems is frequently difficult. Often there is significant branching and ab initio vocabulary development, even within each organization. As a result, no common language is formed.

Vocabulary Standards Challenges

While organizations can create ontologies internally, or in public collaboration, what is lacking is central governance and a service infrastructure to:
– manage vocabulary interfaces;
– manage change processes;
– promote standards;
– manage vocabulary mapping services;
– promote use;
– respond to feedback.

Leaving this to the external environment to fix is not an option; to improve productivity and manage costs industry needs to be much more proactive by:
– moving to a model of vocabularies as pre-competitive assets for consumers and suppliers alike;
– moving to a shared hands-on service for pre-competitive vocabularies, removing the need to do this in each individual organization;
– consolidating vocabularies, providing the benefit of shared costs in generation and maintenance across sectors;
– separating vocabularies from data and content management and provision.

Target Ontology Initiative

The Target Ontology Initiative addresses the following issues:
– representation of a molecular drug target in structured databases is currently ad hoc;
– single-protein targets are linked via Entrez Gene, but this is not an agreed standard;
– multi-protein targets, complexes, and many biological entities are poorly described, often simply by raw text.

The project focuses on industry and suppliers to describe a specification for reporting drug targets within a structured content framework. The output will be a specific set of "rules" regarding the representation of complex molecular targets. The aim is not to define a list of all known targets and to name them, but rather to publish a specification as a recommendation to suppliers and industry to adopt the specification along with industry-generated mappings for pre-existing targets.

2.3 Ontology development in support of predictive toxicology use cases and services

OpenTox[16] has worked on the provision of integrated, high-quality toxicity data for *in silico* predictive toxicology model development, validation, and database queries, implementing it in a data infrastructure that is accessible by the OpenTox framework, including algorithm, model, and validation services that are semantically defined. OpenTox creates dictionaries and ontologies that describe the relations between chemical and toxicological data and experiments, as well as developing novel techniques for the retrieval and quality assurance of toxicological information.

The OpenTox approach to ontology allows for efficient mapping of complementary data coming from different datasets into a unifying structure having a shared terminology and representation. The definition of ontology and controlled vocabulary in OpenTox is required so as to standardize and organize high-level concepts, chemical information, and toxicological data. Distributed OpenTox services exchanging communications need to have unambiguous interpretations of the meaning of any terminology and data they exchange.

The definition of ontology and controlled vocabulary is extremely important in the construction of the OpenTox data infrastructure. It contributes to the necessary standardization and rational organization of data, thus facilitating both vertical (e.g., within one toxicological endpoint) and horizontal (e.g., through different endpoints) retrievals. The definition consists of two main steps: first, the selection of the toxicological endpoints to be included; second, the definition of the type and extent of information for each endpoint and their internal relationships and hierarchies.

[15] http://www.pistoiaalliance.org/

[16] OpenTox, http://www.opentox.org/

A collaborative Protégé ontology development environment, established for the development and review of toxicological ontologies, is located on the OpenTox server.[17]

As toxicological data may come from different, heterogeneous sources, a deployed ontology unifying the terminology and the resources is critical for the rational and reliable organization of the data and its automatic processing. Up to now, the following related ontologies have been developed for OpenTox: Toxicological ontology – listing the toxicological endpoints; Organ system ontology – addressing targets/examinations and organs observed in *in vivo* studies; ToxML ontology – representing semi-automatic conversion of the ToxML schema to an ontology; ToxLink – ToxCast assays ontology; OpenTox ontology – representation of OpenTox framework components: chemical compounds, datasets, algorithms, models, and validation web services; Algorithms ontology – types of algorithms. Besides being defined in an ontology, OpenTox components are made available through standardized web services, where every compound, dataset or 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.

Whenever possible, we are trying to integrate relevant information of neighboring ontologies, such as the Foundational Model of Anatomy (FMA), Ontology for Biomedical Investigation (OBI), NCI Thesaurus, SNOMED Clinical Terms, together with the ToxML (Toxicology XML standard) schema[18].

The OECD Harmonised Templates

The OECD Harmonised Templates correspond to the IUCLID5 XML schemata, which are meant to be used by industry when submitting the documentation on their chemicals to EU authorities. For each endpoint, the OECD Harmonised Templates define a series of fields, e.g., 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 harmonized templates provide a substantial basis for building an ontology. However, they are not well formalized, leaving much space to free text entering, and they have a strong administration emphasis rather than a scientific focus.

OpenTox data infrastructure

The current OpenTox data infrastructure provides:

- a universal database structure design, allowing for storage of multi-faceted life sciences data;
- an ontology allowing for efficient mapping of similar and/or complementary data coming from different datasets into a unifying structure having a shared terminology and meaning;
- integration of multiple datasets with proven high-quality physico-chemical and/or experimental toxicity data;
- built-in heuristics for automatic discovery of 2D chemical structure inconsistencies;
- extensive support for structure-, substructure- and similarity-based searching of chemical structures;
- an OpenTox standards-compliant dataset interface that allows query submission and results retrieval from any OpenTox standards-compliant web service;
- transparent access to and use of life sciences data hosted at various physical locations and incorporating a variety of distributed software resources through the OpenTox framework.

The OpenTox prototype database includes ECHA's list of pre-registered substances,[19] along with high-quality data from consortium members (e.g., ISS ISSCAN,[20] IDEA AMBIT[21]) and third parties (e.g., JRC PRS list,[22] EPA DSSTox,[23] ECETOC skin irritation (ECETOC, 1995), LLNA skin sensitization (Gerberick et al., 2005), Bioconcentration factor (BCF) Gold Standard Database[24]). Additional data for chemical structures has been collected from various public sources (e.g., Chemical Identifier Resolver,[25] ChemIDplus,[26] PubChem[27]) and further checked manually by experts. The database provides means to identify the origin of the data, i.e., the specific inventory a compound originated from. The data is currently publicly available and accessible via initial implementation of REpresentational State Transfer (REST) web services,[28] as defined in the OpenTox Framework design and its implementations.

Toxicological endpoints

OpenTox includes an OWL (Web Ontology Language)[29] ontology of toxicological endpoints, which corresponds to the endpoint classification of REACH guidance documents[30] and allows for unique mapping between endpoints from various in-

17 Collaborative OpenTox Protégé ontology environment, http://www.opentox.org/dev/ontology/collaborative_protege

18 Leadscope ToxML Schema, http://www.leadscope.com/toxml.php

19 http://apps.echa.europa.eu/preregistered/pre-registered-sub.aspx

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

21 http://ambit.sourceforge.net

22 http://ihcp.jrc.ec.europa.eu/our_labs/computational_toxicology/information-sources/ec_inventory/PRS_processed_file_sdf.zip

23 http://www.epa.gov/ncct/dsstox/

24 http://ambit.sourceforge.net/euras

25 http://cactus.nci.nih.gov/chemical/structure

26 http://chem.sis.nlm.nih.gov/chemidplus

27 http://pubchem.ncbi.nlm.nih.gov

28 http://apps.ideaconsult.net:8080/ambit2

29 http://www.w3.org/TR/owl-features

30 http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_r6_en.pdf?%20vers=20_08_08

ventories. The OpenTox toxicological ontology at the moment contains five toxicity study types: carcinogenicity, *in vitro* bacterial mutagenesis, *in vivo* micronucleus, repeated dose toxicity (e.g., chronic, sub-chronic, or sub-acute toxicity), and aquatic toxicity. Using this ontology, each attribute in a toxicological dataset can be associated with an entry to the ontology. The main OWL classes are "ToxicityStudyType," "TestSystem" (includes subclasses such as strains, species, sex, route of exposure, etc.), "TestResult" (includes subclasses such as toxicity measure, test call, mode of action, etc.). The aquatic toxicity ontology was based on the requirements of the EU Directive 92/69/EEC, i.e., acute toxicity for fish (method C.1.), acute toxicity for *Daphnia* (C.2.), and the algal growth inhibition test (C.3.).

(Q)SAR model quality relies crucially on the clarity of endpoints and experimental protocol used, as well as the ability to communicate this information in an unambiguous way, both in the model development and in model application. The current common practice usually includes a textual description of the materials and methods used for acquiring experimental data, as well as literature references, while the model description is a separate entity. Providing an automatic and unique way of describing and linking the endpoint information in a formal way, ready for software processing with minimal human intervention, is one of the big challenges that OpenTox's distributed web services framework tries to address. Using the ontology, each attribute in a toxicological dataset can be associated with an entry to the ontology, therefore allowing a unique mapping between endpoints in various and heterogeneous datasets. The mapping of chemical compound properties stored in the OpenTox prototype database with the endpoints ontology and the information regarding which properties are predicted by models is available via the OpenTox model service. It is used to automatically recognize which endpoints have predictive models available, and it ensures consistency of the endpoint terminology used across the set of distributed OpenTox services.

The ontology has been used for mapping the relevant OpenTox prototype database dataset fields with the ECHA endpoints. This allows for dynamic linking of models, datasets, and endpoints through OpenTox-compliant operations. The Ontology service[31] stores the endpoints ontology, along with other OpenTox-relevant ontologies (opentox.owl,[32] algorithm types,[33] and Blue Obelisk Descriptor Ontology[34]), as well as dynamic information on available models, algorithms, and features provided by OpenTox services. The OpenTox-based ToxPredict application[35] queries[36] the Ontology service by standard SPARQL[37] interface for all available models and retrieves associated information about endpoint modeled, algorithms used to create models, as well as independent and target variables used in models.

Organs, target site, and effects ontologies

The OpenTox Organ Ontology developed by the Fraunhofer Institute for Toxicology & Experimental Medicine (FhG ITEM) is very closely linked to the INHAND initiative (International Harmonization of Nomenclature and Diagnostic Criteria for Lesions in Rats and Mice). INHAND aims to develop an internationally accepted standardized vocabulary for neoplastic and non-neoplastic lesions, as well as the definition of diagnostic features for organs and organ systems observed in rodent *in vivo* studies. The description of the respiratory system and the liver has recently been published (Renne et al., 2009; Thoolen et al., 2010).

The organ ontology is one of the most challenging ontology classes, addressing targets, examinations, and organs observed in *in vivo* studies. It is not endpoint-specific, as organs are species-specific but not endpoint-specific. Endpoint specificity will be defined subsequently when the organ ontology is linked to the toxicological endpoint ontology to address different toxicity types such as repeated dose toxicity, carcinogenicity, or reproductive toxicity. The ontology includes the detailed description of organs, starting from organ systems down to histological components. A hierarchical structure was used, starting with the organ system (e.g., respiratory system) instead of orientating the ontology on the examinations performed in guideline studies, such as histopathology, necropsy, and clinical observations.

At the moment, the OpenTox Organ Ontology contains 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. Currently, the organ ontology is orientated to rodent studies and organs present in humans, so that organs up to histological components of three species are documented. Other species, such as dogs, are also widely used in *in vivo* toxicity studies and, later on, need to be represented as well. Species specificity will be obtained linking the organ ontology to the species ontology documented in the toxicological endpoint ontology, e.g., the gall bladder, usually a part of the digestive system, is not present in rats. Synonyms are included to account for differences in terminologies. As ontology development is a dynamic process, more sources of vocabulary need to be included from industry, as well as public, and non-public databases, to derive a comprehensive ontology.

The concept of an Effect Ontology also has been developed. Effects are endpoint-dependent. The Effect Ontology currently comprises neoplastic and non-neoplastic effects observed in repeated dose and cancer studies. It consists of classes of ef-

31 http://apps.ideaconsult.net:8080/ontology

32 http://opentox.org/data/documents/development/RDF%20files/OpenToxOntology/view

33 http://opentox.org/data/documents/development/RDF%20files/AlgorithmTypes/view

34 http://opentox.org/data/documents/development/RDF%20files/BlueObeliskOntology/view

35 http://toxpredict.org

36 http://opentox.org/data/documents/development/RDF%20files/JavaOnly/query-reasoning-with-jena-and-sparql

37 http://www.w3.org/TR/rdf-sparql-query/

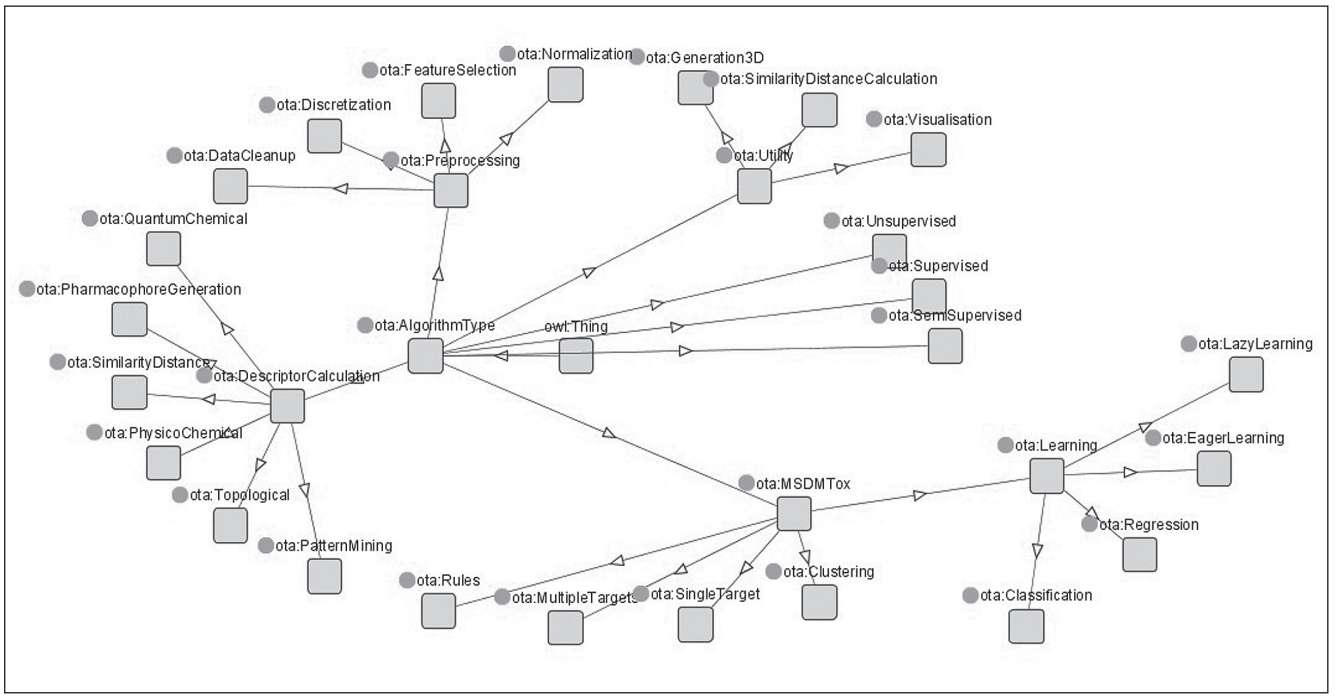

Fig. 1: OpenTox algorithm resources modeled in the OpenTox Algorithm Ontology

fects, as well as effects linked to detailed diagnostic features, as agreed in the INHAND process. Synonyms are included. For the respiratory tract, the effects and diagnostic features are already linked to the organs and organ systems in which they occur. Links and restrictions of effects and diagnostic features to the organ ontology are a complex procedure, e.g., the combination of the diagnostic features 1, 2 and 5 may characterize a special effect in one organ, whereas the combination of the diagnostic features 2, 4 and 8 characterizes the same effect in another organ or organ compartment. Moreover, the discrimination of effects versus histopathological components can be controversial.

OpenTox ontology for linked resources

The OpenTox ontology[38] provides a shared information model for the most common components found in any application providing predictive toxicology functionality, namely chemical compounds, datasets of chemical compounds, data processing and machine learning algorithms, predictive models, and validation routines. The OpenTox framework exposes REST web services, corresponding to each of these common components. A generic OWL representation is defined for every component (e.g., every OpenTox dataset is a subclass of ot:Dataset, every algorithm is a 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 only 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 even more details of descriptor calculation algorithms are specified via the Blue Obelisk[39] ontology of cheminformatics algorithm (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 subclassing (rdf:type), owl:sameAs links, or Blue Obelisk ontology bo:instanceOf predicate. The simultaneous use of OpenTox datasets and compound properties as resources of generic ot:Dataset type and ot:Feature type in the OpenTox ontology, and linking to specific toxicology ontologies, provides a flexible mechanism, allowing users of OpenTox services to enter the data with arbitrarily named properties and further annotate them with the proper terms from toxicology ontologies. This approach is currently used to enter and represent data in OpenTox services and applications.

A graphical overview of the current OpenTox Algorithm ontology is shown in Figure 1. A formal OWL representation of the algorithm ontology is available on the OpenTox website.[40] The plan is to extend this ontology in the future to a full description of every algorithm, including references, parameters, and default

[38] OpenTox Ontology, http://www.opentox.org/api/1.1/opentox.owl

[39] Blue Obelisk Ontology, http://www.ncbi.nlm.nih.gov/pubmed/16711717

[40] OpenTox Algorithm Types in OWL format, http://opentox.org/data/documents/development/RDF%20files/AlgorithmTypes/view

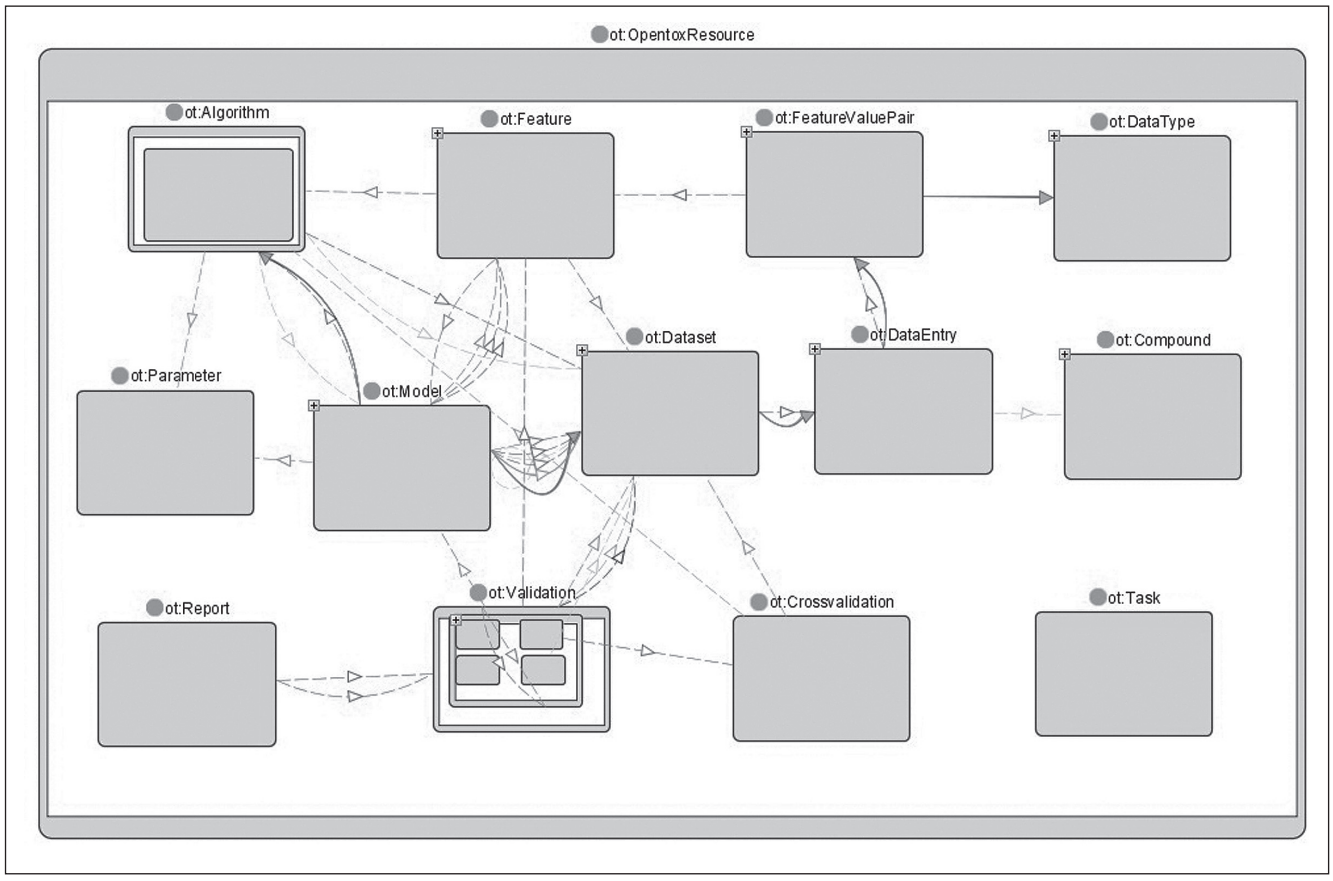

Fig. 2: OpenTox resources modeled in the OpenTox Ontology

values. This will be achieved by adopting the Blue Obelisk ontology and is currently work-in-progress. The RDF representation of an Algorithm contains metadata described by the Dublin Core Specifications[41] for modeling metadata (DC Namespace) and the OpenTox namespace. The establishment of an ontological base for the services facilitates the extension of the services and the introduction of new algorithms and new algorithm classes.

OpenTox Application Programming Interfaces and Ontologies

To assure reliable interoperability between the various OpenTox web services, a well-defined Application Programming Interface (API) is required. The OpenTox APIs specify how each OpenTox web service can be used and what the returned resources look like. It further specifies the HTML status codes returned in the case of successful operations as well as error codes. The initial specifications for the OpenTox APIs have been defined and are available on the OpenTox website.[42] The initial objects already specified are Endpoint, Structure, Structure Identifiers, Feature Definition, Feature, Feature Service, Reference, Algorithm, Algorithm Type, Model, Dataset, Validation Result, Applicability Domain, Feature Selection, and Reporting.

All current OpenTox web services adhere to the REST web service architecture for sharing data and functionality among loosely-coupled, distributed heterogeneous systems. The choice of employing web services allows the complete framework to operate in different locations, independent of operating systems and underlying implementation details.

Figure 2 shows the OpenTox resources modeled in the OpenTox ontology. These resources are provided by the various OpenTox web services. The links between the components reflects interaction between the respective web services. The model web service provides access to (prediction) models. Models are created via the algorithm web service, which supports different types of algorithms (e.g., supervised learning, feature selection, descriptor calculation, and data cleanup). Building a model normally will require various parameters, one or several datasets, and a set of features.

Datasets are stored in the dataset web service. A dataset contains data entries, which are chemical compounds, as well as their feature values. Features are defined as objects represent-

[41] Dublin Core Metadata Initiative, http://dublincore.org/

[42] Current Specifications of OpenTox Interfaces, http://www.opentox.org/dev/apis

ing a property of a compound, including assays, descriptors and calculated features, endpoints, and predictions. Different representations for chemical compounds can be accessed from the compound web service. The feature web service provides the available features (e.g., structural features, chemical descriptors, assay responses, endpoints).

The validation web service evaluates and compares the performance of prediction models. Simple training-test-set-validation is supported as well as cross-validation. The validation result contains quality statistical figures and reports (available in html or pdf formats) that visualize the validation results. The task web service supports long-running, asynchronous processes. The ontology web service provides meta information from relevant ontologies (which can be accessed using SPARQL queries[43]), as well as lists of available services. Approaches to Authentication and Authorization are specified in version 1.2 of the OpenTox API.

All OpenTox resources have representations providing information about the type of resource and what the service accepts as input, such as tuning parameters. Most algorithms and model resources in OpenTox are available in multiple representations. The Resource Description Framework (RDF) representation, and in particular its XML formatted variant, was chosen as the master data exchange format, for the following reasons:

– RDF is a W3C recommendation: RDF-related representations such as rdf/xml and rdf/turtle are W3C recommendations, so they constitute a standard model for data exchange;
– RDF is part of Semantic Web Policy: RDF as a representation for a self-contained description of web resources contributes to the evolution of the Semantic Web; a web where all machines can "understand" each other;
– RDF is designed to be machine-readable.

REACH endpoints

The OpenTox data infrastructure prioritizes support of toxicological endpoints for which data are required under the REACH regulation. In current toxicological testing, these endpoints are addressed by both *in vitro* and animal experiments carried out according to OECD guidelines.

The toxicological endpoints considered by REACH are the following (Lilienblum et al., 2008): Skin irritation, skin corrosion; eye irritation; dermal sensitization; mutagenicity; acute oral toxicity; acute inhalative toxicity; acute dermal toxicity; subacute toxicity (oral or inhalation); subchronic toxicity (oral or inhalation); reproductive toxicity screening; developmental toxicity; two-generation reproductive toxicity study; toxicokinetics; and carcinogenicity.

The OECD guidelines for testing of chemicals[44] are published on the Internet. Whereas there is no official list of OECD endpoints (test guidelines are developed according to the needs of member countries), and no official OECD approach to toxicity testing, interesting background information on criteria for toxicity testing has been developed as SIDS (Screening Information Data Set).[45]

OECD Guidelines

Use of ontologies provides support in OpenTox for the OECD Guidelines for (Q)SAR validation.[46] Use of the toxicological endpoint ontology supports satisfaction of the first principle of a "defined endpoint." Using this ontology, each attribute in a toxicological dataset can be associated with an entry to the ontology, therefore allowing a unique mapping between endpoints in various and heterogeneous datasets. To address the second principle of an unambiguous algorithm, OpenTox provides unified access to documented models and algorithms, as well as to the source code of their implementation. Algorithm template descriptions and the algorithm type ontology allow a clear definition of what type of algorithm(s) is used to construct a model.

OpenToxipedia

We have created OpenToxipedia as a collaborative resource for the entry and editing of toxicology terms, supported by a semantic media wiki.[47] The semantic media wiki environment for OpenToxipedia was developed as a community-based, predictive toxicology knowledge resource with a vocabulary of toxicology terminology used in ontology development. The controlled vocabulary created in Year 1 was updated and transferred to a semantic media wiki format and made available from OpenToxipedia at www.opentoxipedia.org. OpenToxipedia allows creating, adding, editing, and storing terms used in toxicology terminology. The rules for working with the terms also were created. At present, OpenToxipedia contains 862 toxicological terms with description and literature references classified into 26 categories. The terms can be browsed either by category or in alphabetical order. Terms used in OpenTox applications can be linked to using SPARQL queries to the media wiki.

ToxPredict

The ToxPredict application[48] demonstrates the use of OpenTox ontology in connecting multiple web resources to execute the use case of accessing available data and predictions for a chemical structure specified by the user. Links to training datasets, algorithms, and descriptor calculation REST services are provided. Algorithm, Model, and Feature services are registered into

[43] SPARQL Query Language for RDF: W3C Recommendation 15 January 2008, http://www.w3.org/TR/rdf-sparql-query/

[44] Guidelines for the Testing of Chemicals, http://www.oecd.org/document/40/0,3343,en_2649_34377_37051368_1_1_1_1,00.html

[45] Chapter 2 of the Manual for Investigation of High Production Volume (HPV) Chemicals, http://www.oecd.org/document/7/0,3343,en_2649_34379_1947463_1_1_1_1,00.html

[46] OECD Validation Principles, http://www.oecd.org/dataoecd/33/37/37849783.pdf

[47] OpenTox Community Resource for Toxicology Vocabulary and Ontology: OpenToxipedia, http://www.opentox.org/opentoxipedia

[48] http://www.toxpredict.org/

the Ontology service, which provides RDF triple storage with SPARQL, allowing various queries. The models provide information about the independent variables used, the target variables (experimental toxicity data), and predicted values. All these variables are accessible via the OpenTox Feature web service, where each feature can be associated with a specific entry from the existing endpoint ontology. The association is usually done during the upload of the training data into the database. The endpoint, associated with the model variables, is automatically retrieved and displayed in the user interface. This provides an automatic and consistent way of complying with the first OECD validation principle of using a "Defined endpoint."

The ToxPredict application queries the ontology service for all available models, along with the associated information about algorithms used in the model, descriptors, and endpoints. The list of models may include models provided by different partners and running on several remote sites. The Ontology service serves as a hub for gathering a list of available models and algorithms from remote sites. There could be multiple instances of the ToxPredict application, configured to use different ontology services, and therefore allowing for a different subset of models to be exposed to end users.

The interplay of multiple services, running on remote sites, provides a flexible means for the integration of models and descriptors, developed by different organizations and running in different environments. Identification of algorithms and models via web URIs ensure the compliance with the OECD validation principle 2 of "an unambiguous algorithm," as well as repeatability of the results of the model building. Extensive meta-information about the algorithm, and about the models themselves, is accessible via web URIs and the OpenTox API.

Drug design

Work was initiated in spring 2010 to design interoperation, including ontology-based data exchange between drug discovery predictions and screening experiments. OpenTox, a collaborative semantic-based electronic notebook system (CERF), and Synergy collaboration services, communicating through a Petals enterprise service bus were used together to support integrated deployment and testing on the FP7 Synergy pilots on anti-malarial drug design[49] and predictive toxicology, were initiated in 2010.

2.4 Ontology development within the eTOX project

As part of the IMI eTOX project, Novartis has been leading the activity to create ontologies for preclinical safety. Novartis has decided to take a different approach to ontology development for preclinical safety. Other projects in this area often link a finding with an organ, i.e., "Liver necrosis" is treated as a single term. This means that the term "necrosis" must be entered many times into an ontology, as it can occur in different tissues. It was decided to separate the finding from the anatomy so that the term necrosis stands alone and the term liver stands alone. This was done for the following reasons: 1) the ontologies will be easier to maintain and 2) the ontologies will allow more flexible mapping of findings for later computer modeling. For example, using a machine learning approach the feature "finding" (e.g., necrosis) will be treated separately from the feature "anatomical region" or "organ" (e.g., liver), which will allow the machine learning algorithm to automatically compare compounds causing necrosis across all tissues vs. all findings in liver.

Vocabulary curation tool for ontology mapping

A vocabulary curation tool was created at Novartis prior to the beginning of the eTOX project.[50] This application was modified from an open source project: Open Biological and Biomedical Ontologies (OBO) editor, OBO-Edit.[51] OBO-Edit is developed by the Berkeley Bioinformatics and Ontologies Project, and is funded by the Gene Ontology Consortium. A modified OBO-Edit application was built on the pre-release version of OBO-Edit, and the functionality created will be incorporated into the production version. This is an important requirement to allow all people working on toxicology ontologies to add terms to the ontology in a simple way.

Anatomy ontology

The anatomy ontology is the list of terms and relationships that could describe the anatomical location or organ type from any animal used in preclinical safety experiments. Starting with the Adult Mouse Anatomy Ontology created at the Jackson Labs, investigators mapped all the terms in use in Novartis preclinical databases to this ontology as synonyms and expanded it with new terms where appropriate. This ontology is ready for release to the eTOX project members to allow them to map their terms onto the ontology.

Microscopic pathology findings ontology

The microscopic pathology findings ontology is one of the most difficult ontologies to build as it has to be built from scratch. Although an attempt was made to find an existing ontology that fits the needs of scientists at Novartis, one could not be found that fit the vocabularies used internally. Therefore, ontology developers worked with pathologists at Novartis to come up with the backbone for the ontology and mapped all the existing terms to that. The backbone is in the final stages of curation and subsequent efforts will map the existing terms (11,000 terms in the Novartis database) to that. This part of the project has been more time consuming than originally expected and significant pathology sources still have to be investigated. It is anticipated, however, that once it is completed, it will serve as a model for an industry standard ontology that all organizations can use in the future.

Cell and tissue type ontology

Novartis also is working on a cell and tissue type ontology to complement the anatomy ontology, which is intended to be

[49] Scientists Against Malaria, http://www.scientistsagainstmalaria.net/

[50] eTOX, http://www.etoxproject.eu/

[51] OBO-Edit, http://oboedit.org/

shared with the eTOX members. The rationale behind separating this ontology from the main anatomy ontology is similar to that for separating findings from organs, namely that the same cell types occur in many different tissues so maintaining a separate ontology will be easier. For example, epithelial cells exist everywhere in the body. So creating a child in the anatomy ontology with epithelial cells will link to many different terms throughout the body – skin epithelium, lung epithelium, vascular epithelium, etc. Managing those links will be almost impossible. Simple keyword searches/text mapping searches in the cell type ontology will provide the links here, and since this ontology is rather small it should not result in a loss of performance. For data modeling, this should provide sufficient information when there is a cell type specific effect that has been annotated by the pathologist.

2.5 EU-ADR

EU-ADR[52] is a Research and Development project funded by the Information and Communication Technologies (ICT) area of the European Commission under the FP7 Program, which aims to explore and understand adverse drug reactions by integrative mining of clinical records and biomedical knowledge. The overall objective of the EU-ADR project is the design, development, and validation of a computerized system that exploits data from electronic healthcare records and biomedical databases for the early detection of adverse drug reactions. EU-ADR specific objectives are: to detect events, to relate these events to drugs, to develop hypotheses that explain adverse events, to detect adverse events earlier, and to avoid false positives. EU-ADR makes extensive use of existing standards and uses ICD9-CM, ICD10, READ CODES, and ICPC2 terminologies together with the UMLS system. The use of the above mentioned thesauri is limited to making initial suggestions to local database sites. These local sites, which are familiar with their own coding system and how that system is used, will then take these suggestions as a starting point. EU-ADR is a collaboration with intent to share and aggregate the results of local analysis: a process of "harmonization" that is essential to ensure all databases share a common understanding about the events of interest. The experiences and methodologies matured in the EU-ADR are important for the Toxicology Ontology Roadmap (Hardy et al., 2012), due to the interdependency between human safety and toxicology, and also as a demonstration of the practical application of ontologies in the safety/toxicology domain.

2.6 Harmonized ontologies for toxicology pathways and mechanisms

CCNet's ToxWiz ontology[53] was developed to organize and manage information about toxic effects and to help provide as much insight as possible about the mechanisms underlying these toxic effects and to elucidate underlying biological pathways. This ontology is a set of controlled vocabularies used to describe and interconnect three major categories: effects (toxic and biological effects – including a full set of histopathology terms covering all organs); target proteins/genes, and chemical compounds to each other (Fig. 3).

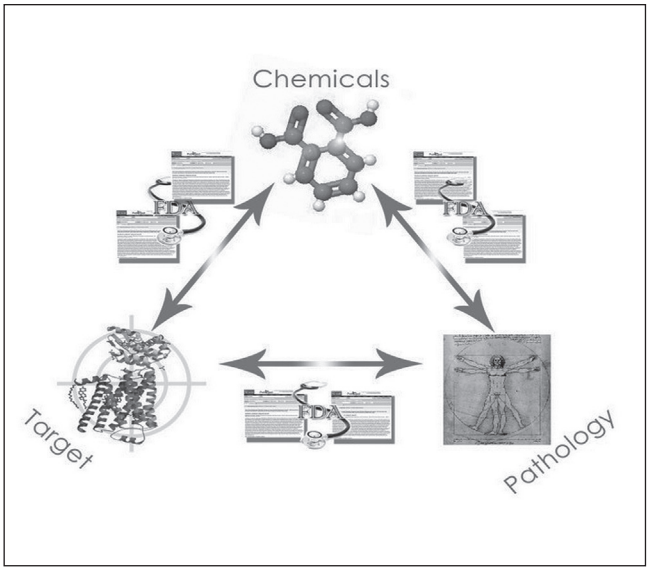

Fig. 3: Uses of the CCNet ToxWiz Ontology
The toxicology ontology is implemented in an online resource to support data integration and literature curation on toxic effects with a mechanistic focus. Furthermore, it can also serve as a cross-reading tool. The ontology enables getting quick insights (with the background intelligent search engine with an ontology backbone) for a chemical, listing all the reported toxicity endpoints and, in parallel, asking what all the reported metabolizing enzymes/nuclear hormone receptors or disregulated genes are for this chemical, in order to obtain a quick idea about possible mechanisms. Furthermore, one can ask, for a specific chemical and a specific histopathology, what all the reported associated genes and proteins are and how they fit into known metabolic and signaling biological pathways.

The purpose of the ontology is to facilitate prediction of toxic effects (prospective analysis), to help explain causes of toxic effects (retrospective analysis), to elucidate modes of action and create hypotheses for modes of action and mechanisms, to support the extraction process of knowledge relevant to toxicology from toxicology reports and literature, and to enable integration and transfer of findings to clinical observations (Fig. 4).

Using this ontology, chemicals can be linked to a histopathological observation, with different levels of granularity of the observation (hierarchy) (Fig. 5). In order to obtain a hypothesis about underlying biological pathways and mechanisms, histopathologies can be linked back to associated chemicals, at the

[52] EU-ADR, www.euadr-project.org

[53] ToxWiz, http://www.toxwiz.com/

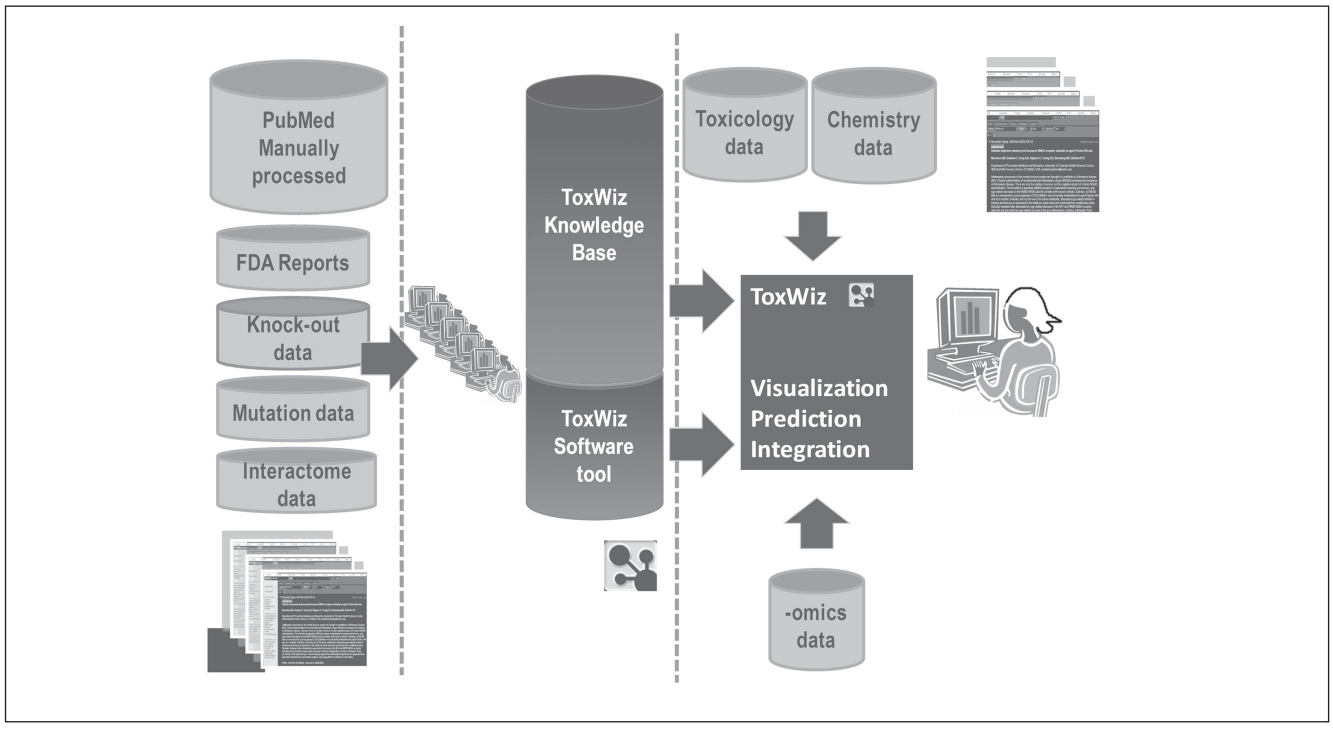

Fig. 4: Purposes of the CCNet ToxWiz Ontology

The CCNet toxicology ontology is generic and can be integrated into a variety of databases as a backbone for data integration around histopathological endpoints. It does not require mechanistic information, but it is designed to readily provide a framework for integrating it, should that information become available for some areas or organs. In this picture, we describe how CCNet toxicology ontology serves to enable i) (left) interpretation, normalization, and integration of the information derived from scientific literature and regulatory documents into the internal ToxWiz Knowledge Base, ii) (in the middle) serves as a backbone of search engines and predictive algorithms, iii) (on the right) for enabling integration of customer toxicology information on this framework at the client site, as well as analysis of toxicology screening data (including many types of *in vivo*, *in vitro* and *-omics* data) and for making predictions or knowledge-based mechanistic hypotheses.

same time linking histopathologies to associated entities such as nuclear metabolizing enzyme, nuclear receptor, or a disregulated gene (proteins/genes in generic terms enables studying biological pathways and mechanisms).

CCNet's ontology is derived from more than 150 man-years of expert curation and toxicology information extraction from safety reports and literature. It describes histopathology with 912 terms and classifies a further 500,000 synonyms, terms with common or similar meanings used by pathologists. During development, the ontology underwent six iterations over eight years to ensure harmonization between projects, leading to increased interoperability support.

CCNet ontology of toxic endpoints: starting endpoint and development

In creating and developing the CCNet ontology, requirements were to capture all available knowledge from the literature and toxicology reports so as to support the human way of interpreting and recording such findings, to adequately capture toxicology test results in pre-clinical testing, as well as systemic long term toxicity, together with efforts to classify and define a spectrum of histopathology findings.

The ToxWiz ontology strives to be interoperable with other ontology efforts. As a starting point, the following activities were pursued:

1. Hierarchy of organs and tissues for model organisms derived by consultation of various sources (e.g., NCBI, ENSEMBL, UniProt)
   System – Organ – Components/Tissue/Cell-type
   (e.g., Hepatic system – Liver – Bile duct/Kupffer Cell)
   N.B. includes synonyms (Biliary = Bile duct, etc.)
2. Set of histopathological observational keywords and principles, by consultation with toxicologists, e.g., Frank Bonner (e.g., Hypertrophy, Hyperplasia, Neoplasia, Carcinoma, Necrosis, Fibrosis, Inflammation, etc.)
3. Several principles of progression in time/pathology
   (e.g., Carcinoma is a subset of Neoplasia, Hyperplasia precedes Neoplasia)

Ontology is only useful if it is used and serves a purpose. Using ontologies inevitably leads to changes and improvements. Over the first two years several necessary changes in the CCNet ontology were incorporated in response to user feedback. This included finer definitions of certain pathologies, e.g., Carcinoma (Adenocarcinoma, Squamous cell carcinoma, Metastatic

carcinoma, etc.), and some other broad terms to cover pathological observations, e.g., primary and secondary as qualifiers. Generally speaking, terms are added to capture better what experimentalists are saying. This does not mean that the ontology standard is a moving target. The changes and improvements only increase the granularity and are captured by categories at different levels.

2.7 Ontology-based probing of biological effects in a Virtual Liver

The Virtual Liver (v-Liver) project[54] is aimed at providing decision support tools for assessing chemical safety. The risk of chemical-induced injury has traditionally been estimated using animal testing studies, but this is infeasible because there are thousands of chemicals in commerce. Furthermore, it is difficult to translate adverse effects from rodents to humans because the underlying pathways to toxicity are poorly defined. The v-Liver project is addressing this problem by focusing on two specific issues: an ontology for representing toxicology findings to describe chemical effects, and a systems model to quantitatively estimate their dose-dependence in humans.

The v-Liver project has designed a domain-specific ontology to formally express disparate evidence about chemical effects using description logic and to reason about this information in the context of toxicity. Domain knowledge is combined with published experimental findings about chemicals from the literature, high-throughput screening experiments, and legacy rodent toxicology studies, organized in an OWL knowledge base (KB). The v-Liver KB captures computable assertions about chemical effects and enables plausible physiological explanations of these effects. The important distinction with related efforts is that the semantics are domain-specific, i.e., they attempt to describe how chemicals perturb functions across levels of biological organization so that molecular changes can be translated to clinically relevant outcomes (see Fig. 6). In doing so, the v-Liver ontology aims to encode the relevant semantics required to formalize "toxicity pathways," mode of action, and mechanism of action. Currently, the v-Liver KB focuses on evaluating the human relevance of *in vivo* toxicology endpoints and non-genotoxic pathways to rodent liver cancer (Shah and Wambaugh, 2010).

The immediate utility of the v-Liver ontology and KB is to evaluate chemical safety using bioactivity data (e.g., from quantitative high-throughput screening (qHTS) experiments, or *-omics* profiling) by investigating the relevant "pathways to toxicity." This provides end users contextual information about a chemical to support decisions regarding additional targeted assays in environmental chemical testing or prioritization in safety evaluation for lead compounds. For a subset of chemicals where there is sufficient evidence for health risks, the ontology is also being used in an agent-based systems model to quantitatively simulate the dose and time-dependence of hepatic effects.

Category	Example Toxic Endpoint Cluster	Includes	Description
System	Hepatotoxicity	Gall bladder, liver, hepatic system	General observations (e.g. clinical, etc.)
Organ/tissue	Liver toxicities	Liver	Pathology report for any toxicity in liver
Organ/tissue	Liver hypertrophy	Liver	Pathology report of specific toxicity
Cells	Hepatocyte neoplasia	Cells or cell lines	Result from cell-based assays

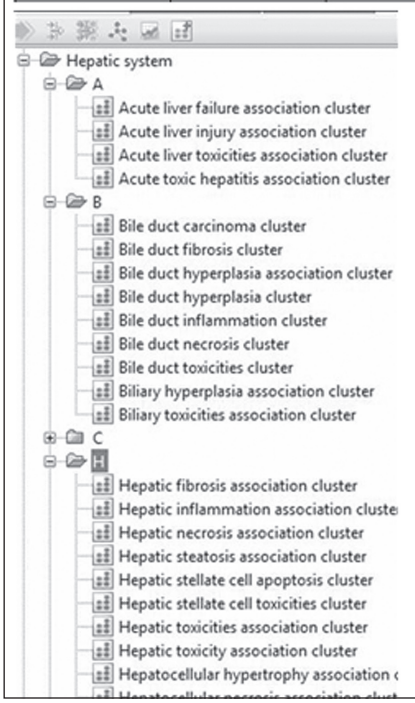

Fig. 5: Example of the hierarchical structure of the CCNet ToxWiz Ontology

On top is an example of four different levels of hierarchy and different granularity for describing a biomedical observation or toxicity effect. The first three categories from top down are *in vivo* observations: i) the entire organ system, such as hepatotoxicity without specifying if it is liver or gall bladder (these observations mostly come from the clinical observations by medical doctors), ii) organ-specific observations mostly from the clinic, iii) more specific, mostly histopathological findings from *in vivo* preclinical tests. The fourth level is an extra category designed to capture observations from *in vitro* cell assays that we can try to relate to the above *in vivo* observations. Below is a one level view of a computed form of an ontology and a subset of liver toxicities designed for several different purposes (to capture the histopathology of liver, report biomarker candidates, etc.).

[54] V-Liver, http://www.epa.gov/ncct/virtual_liver/ V-Liver http://www.epa.gov/ncct/virtual_liver/

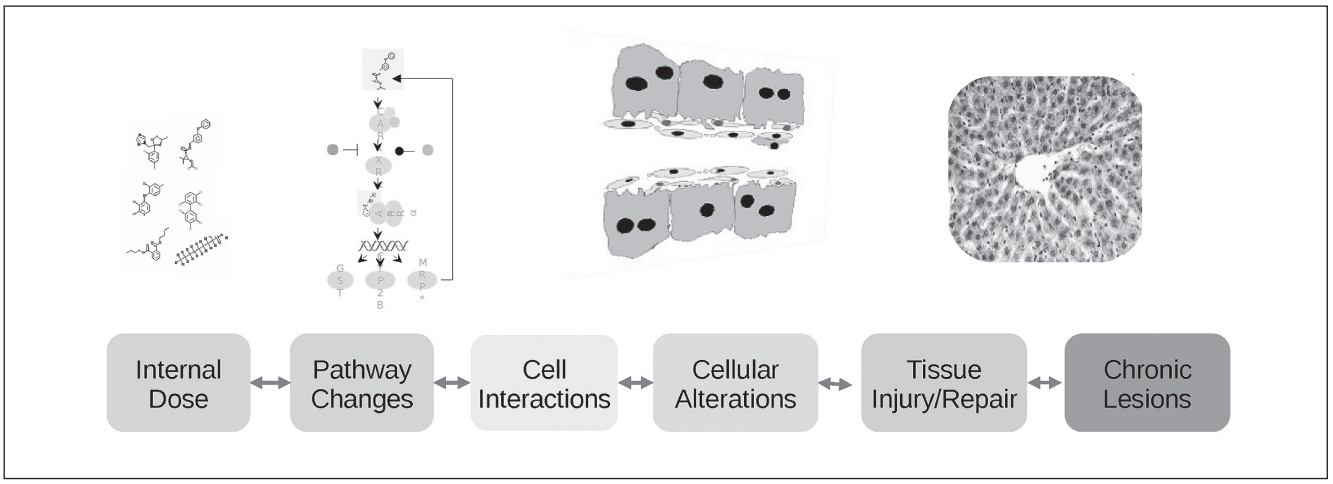

Fig. 6: Probing of biological effects using the V-liver ontology
Chemical-induced perturbations begin with the distribution of a chemical to the target tissues where they produce dose-dependent changes in pathways that can ultimately lead to histological lesions.

2.8 Ontology support of interoperable application development

The interoperable architecture of OpenTox (Hardy et al., 2010) facilitates application integration with independent third party applications. Bioclipse developers recently demonstrated this by developing plugins for the Bioclipse workbench application (Spjuth et al., 2007) that are capable of querying OpenTox SPARQL endpoints using the OpenTox ontology, allowing users to discover and interact with data, algorithms, and predictive models, which can be available locally or from distributed OpenTox services.

A plugin for Bioclipse was constructed to query OpenTox SPARQL endpoints to discover the available descriptor implementations for setting up (Q)SAR datasets. The user gets a visual presentation of the available descriptors and can choose to combine OpenTox descriptors with local implementations to set up an interoperable and reproducible (Q)SAR dataset (Spjuth et al., 2010), which was achieved by adding support for the Blue Obelisk Descriptor Ontology[55] to OpenTox descriptor algorithms. Full cheminformatics functionality, including structure drawing and compound browsing, is provided via the Chemistry Development Kit (CDK) (Steinbeck et al., 2003).

An additional plugin was constructed for Bioclipse to allow querying OpenTox services for available predictive models, again using the OpenTox ontology over SPARQL endpoints. This integrates OpenTox models with local models in Bioclipse and allows for decision support where public data/models are combined with local data/models (see Fig. 7). Hence, Bioclipse is turned into a rich interface for toxicology predictions with a responsive graphical user interface (GUI) that updates prediction results instantly upon changes to the chemical structures (Spjuth et al., 2011).This opens up the possibility for scientists to explore hypotheses on how predictions would be affected by certain structural modifications, as well as the capability of showing additional information about the predictions in a user-friendly workbench (such as highlighting substructures deemed important by the model).

Bioclipse is equipped with a scripting language for automating and sharing analyses. The interoperable architecture of OpenTox allowed for integration in the scripting language of data discovery and downloading, algorithm discovery, and calculations, together with submitting chemical structures for prediction to one or many of the OpenTox models. The integration of OpenTox in the Bioclipse Scripting Language (BSL) is one way of standardizing a toxicology protocol, and such scripts, preferably, can be shared using services like MyExperiment (Goble et al., 2010).

2.9 ToxML

Gaining access to, and exchanging, sufficiently-detailed toxicological study data is a particular problem in (Q)SAR model development work, but it is increasingly important in all areas as the exchange of detailed data increases. Work was initiated a few years ago by LeadScope Inc. and Lhasa Limited to develop a standard, ToxML, for the exchange of biological and chemical data, including chemical structures. This led to the publication by LeadScope Inc. of an illustrative standard covering numerous *in vivo* end-points. More recently, a project has been set up to broaden the coverage of ToxML and, it is hoped, to promote its wide acceptance.

Purpose of the ToxML standard

ToxML is a schema, or format, for electronic exchange of data between software applications. This distinguishes it from internal database schemas and from designs for the representation of data in user interfaces. Different database management systems have their own database schemas and user interface functionality, and ToxML is not intended to change or standardize them. Each database management system will need to have its own

[55] Blue Obelisk Descriptor Ontology, http://qsar.sourceforge.net/dicts/qsar-descriptors/index.xhtml Blue Obelisk Descriptor Ontology http://qsar.sourceforge.net/dicts/qsar-descriptors/index.xhtml

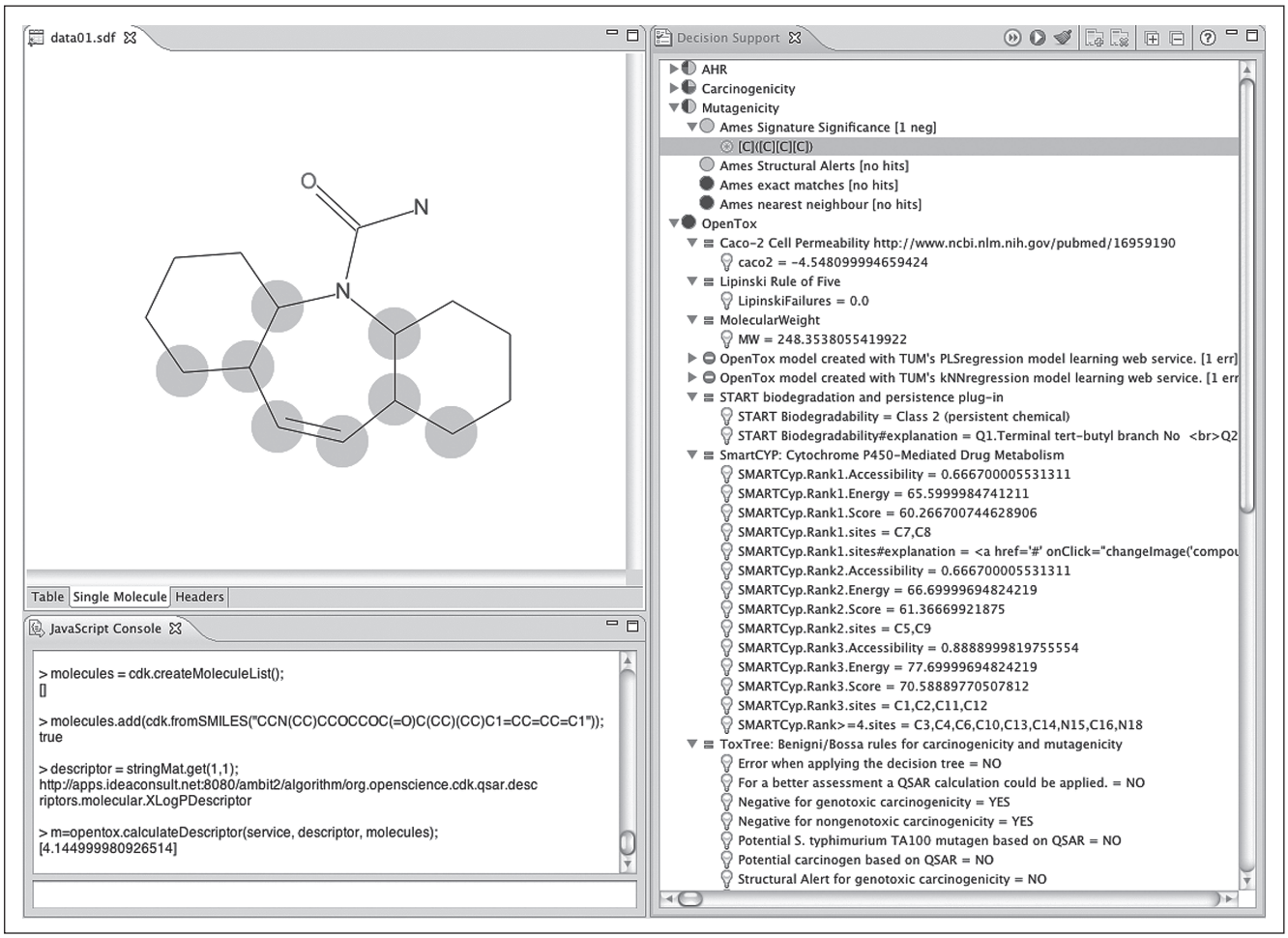

Fig. 7: Interoperable application developed by Bioclipse
Screenshot from Bioclipse showing the drug Carbamazepine, an anticonvulsant and mood stabilizing drug used primarily in the treatment of epilepsy and bipolar disorder. In the right panel are displayed the results from various predictive models, both OpenTox models and local Bioclipse models. Selecting models can highlight functionally important residues in the original chemical structure (if applicable).

import/export component to connect with ToxML files, but this should be invisible to the end user.

Technical implementation of ToxML

The language used in ToxML files is currently XML, and this will continue to be the case. It has the advantage of being widely – and increasingly – used by a broad range of computer software. This makes it familiar to software developers, and it also means that many of the software development tools they use include ready-made components for reading and writing XML files.

An alternative would be the SD file format, which is widely used by chemistry software and, consequently, by toxicology applications that make use of chemical structures. A disadvantage of the SD file format is that it is designed primarily for exchange of chemical structures and, as published, it does not easily support the communication of associated hierarchical data. A hierarchical SD file format has been developed by Lhasa Limited for use in connection with their Vitic database management system, and this was considered as an alternative to ToxML. Two disadvantages remain, however – records relating to chemical structures are in a fixed format in an SD file, which does not suit modern computing practice well, and the SD file format is peculiar to cheminformatics, whereas XML is more widely used.

The chemical information in an SD file and a Molfile takes the same form. A Molfile contains a single chemical entity, whereas an SD file can contain many. Chemical structures in the current ToxML are represented by embedded Molfiles, but this representation will be replaced by one based on XML – most probably a representation developed by Molecular Networks GmbH for regulatory bodies in the USA and in the public domain.

Within OpenTox, work has recently been completed that provides a conversion capability between ToxML and a semantic representation (RDF), thus enabling the automatic provision and integration of data in the ToxML format within the semantic web of distributed resources supported by OpenTox. Application of authorization and authentication policies on web services, hence, can provide controlled access to data contained within a database such as Leadscope's within a distributed semantic web context.

Collaborative development of ToxML

Leadscope has donated the current ToxML standard fully into the public domain, and Lhasa Limited has agreed to fund and manage the launch of a project to develop and maintain it. Creating a universal schema would be a huge job, would take a long time, and might or might not deliver well what was needed in each specialized area of toxicology. So a collaborative project is being set up. This will help in sharing the work and, more importantly, it will allow people to develop the standard incrementally for those areas where it is needed when it is needed, and only those areas. Researchers will be encouraged to deliver proposals for additions and changes to ToxML via a wiki website. It will be a so-called curated wiki site: contributors will not modify the published standard directly; rather, their proposals will be assessed and collated by one or more curators who will control the contents of the published standard.

At the time of this writing, a draft constitution for the long-term project is in place, a first meeting of an Advisory Board (at which the constitution will be formally ratified) was organized, technical meetings are in progress on early modifications to the existing standard, and a contract for the development of the software framework for collaborative development of the standard has been completed.

2.10 Biological Expression Language

The Biological Expression Language (BEL)[56] is a language for representing scientific findings in the life sciences in a computable form. BEL is designed to represent scientific findings by capturing causal and correlative relationships in context, where context can include information about the biological and experimental system in which the relationships were observed, the supporting publications cited, and the process of curation.

BEL is intended as a knowledge capture and interchange medium, supporting the operation of systems that integrate knowledge derived from independent efforts. The language is designed to be use-neutral, facilitating the storage and use of structured knowledge for inference by applications through a knowledge assembly process that can create computable biological networks. While BEL does not prescribe any particular assembly process or any particular knowledge format for the output of an assembly process, a suite of software components called the BEL Framework provides everything necessary to create, compile, assemble, and deliver computable knowledge models to BEL-aware applications. This makes the BEL language and the BEL framework a unique system for knowledge capture and representation that scales with the increasing complexity of scientific facts.

3 Review of existing ontologies of relevance to predictive toxicology

Biomedical ontologies have become essential tools in managing data throughout many of the life science disciplines, and as a result, an ever-increasing number of ontologies are being developed to address different needs. Many of the domain areas being tackled by the different ontology development communities overlap to greater or lesser extents, leading to a sometimes confusing proliferation of conflicting resources. Re-use of existing ontologies, where possible, and the careful delineation of scope are key to managing this proliferation and to reducing the overhead required for ontology development by each of the participating projects. This idea is central to the OBO Foundry (Smith et al., 2007), a coordinated community that collects and provides peer review for ontologies, agreeing to abide by principles favoring openness, scope delineation, and mutual reuse. Reuse of OBO Foundry ontologies, as well as several non-OBO Foundry ontologies, will be very helpful in establishing a Toxicology Ontology, while a few key gap areas where no existing ontologies yet exist will require new development. Many of the existing commercially available terminologies would also be very useful to this effort, but their current licensing models present some challenges for an open project; however, these challenges could potentially be overcome through the development of new business models involving all stakeholders.

3.1 Existing ontologies that can be reused for predictive toxicology

OBO Foundry ontologies

As one of the central repositories of large-scale biomedical ontologies, the OBO Foundry is a promising source of ontologies for reuse. As of February 2011, the OBO Library (the full repository of ontologies made available via the OBO Foundry website, regardless of review status) consisted of 99 ontologies. The following OBO ontologies are potentially of use for the development of a Toxicology Ontology:

- The popular Gene Ontology, which includes the biological processes and molecular functions that gene products are involved in, as well as the sub-cellular locations at which they act;
- the ChEBI Ontology, which includes "biologically interesting" chemical entities and their activities in a biological context;
- the Ontology of Biomedical Investigations, which provides terminology relevant to experimental investigations, that is, instruments and protocols, data types and analyses on the data, and the related Units of Measurement ontology;
- anatomy ontologies, in particular the Mouse anatomy ontology and the Foundational Model of Anatomy;
- the cell type ontology and the Brenda tissue ontology;
- the NCI Thesaurus, 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 Human Disease ontology, which gives a comprehensive ontology of diseases, including genetic, environmental, and infectious diseases for wide applications in clinical research and medicine.

[56] Biological Expression Language, http://www.selventa.com/technology/bel-framework

Non-OBO Foundry ontologies

Several non-OBO Foundry ontologies and terminology and vocabulary resources are identified as essential for reuse in a predictive toxicology context. These include the large-scale terminology standards such as ICD-9/10, MeSH and the UMLS, SNOMED, and the NCI CTAE. MedDRA currently has valuable vocabulary with wide use in pharmaceutical toxicology pipelines. Other relevant commercial resources include Goreni, Pharmapendium and GVK-BIO. For a unifying toxicology ontology project, engagement with service providers such as MedDRA to negotiate cost-effective options for reuse will be an important goal in service of the community's needs.

Other existing ontologies of relevance in the toxicology domain include ToxHunter and ToxWiz (commercial offerings), and ToxML and OpenTox (open, community efforts). For vocabulary relating to aspects of chemistry and bioactivity, PubChem and DrugBank are highlighted as important open databases. Finally, while not an ontology per se, the Minimum Information for Annotation of a Bioactive Entity (MIABE) standardization effort is highlighted as important, since toxicity is a kind of bioactivity, and it can be expected that MIABE standards should also apply in the case of toxicity (although several additional requirements would also apply).

3.2 In support of mechanism and pathway-based approaches

The utility of an ontology or vocabulary resource is strongly tied to the intended application. Several vastly different applications make use of ontologies and vocabulary resources, for example text mining, hierarchical classification and clustering, and approaches to automatic knowledge discovery. Mechanism and pathway-based approaches to predictive ontology pose particular challenges in that they involve inherently complex, structured information, and what counts as a mechanism or pathway can differ from community to community. There is ambiguity on the level of detail required: for example, in some cases inferred protein-chemical interaction networks are sufficient, while in other cases detailed models of how a particular pathway brings about a particular outcome is required. A key question, therefore, which needs to be addressed by the community as a whole, is what counts as a pathway in the context of predictive toxicology?

Software such as Cell Designer could be used for pathway modeling and linking to existing ontologies. Furthermore, pathways annotated with ontology terms could be brought onto a Semantic Web for Toxicology via the SBML2SMW utility[57] (Mathäß et al., 2010).

3.3 Key gaps in existing ontologies

While there is a bewildering array of existing ontologies that need to be incorporated in support of predictive toxicology, some gaps remain, nevertheless, where it is desirable that ontologies should be created in the public domain. While many other species-specific anatomy ontologies are well developed, canine anatomy currently seems to be overlooked by other anatomy ontology communities. The existence of species-specific anatomy ontologies is itself a matter subject to debate, as some advocate using a cross-species reference ontology for anatomy instead. However, even if such an approach is favored, a gap remains in anatomy resources for the representation of dog anatomy.

Related to anatomy is the need for good ontologies describing phenotypic abnormalities. While several ontologies covering this domain area already exist, they are hindered somewhat by a lack of shared community agreement on how to define phenotypes and their coverage falls significantly short of what is required to adequately perform modeling in predictive toxicology.

Another serious gap is the absence of ontologies for localized histopathology, and more generally, ontologies of microanatomy. An ontology describing strain names also is missing, which may be due in part to the more general need for a concrete definition of "strain."

Finally, the lack of ontologies describing developmental stages is highlighted as a challenge for the community. For example, the developmental stages for disease progressions, physiology, and anatomy and phenotypes, are not well described in ontologies for disease, physiology, anatomy and phenotypes respectively. Many challenges lie in the representation of development over time, while also accurately representing the state at any one time. This is a challenge for the ontology community as a whole, but it has particular bearing on the toxicology community. Consider that many substances are in particularly toxic to the developing fetus in certain stages, and yet, ontology annotation of this important fact remains elusive.

4 Conclusions

In this review we have attempted to describe numerous ontology development activities in progress in the field of predictive toxicology and in neighboring areas of science. A significant set of resources is already available to provide a foundation for an ontological framework for 21st century mechanistic-based toxicology research. Ontologies such as ToxWiz already provide a basis for application to toxicology investigations, whereas other ontologies under development in the biological, chemical, and biomedical communities could be incorporated into an extended future framework. OpenTox has provided a semantic web framework for the implementation of such ontologies into software applications and linked data resources. Bioclipse developers have shown the benefit of interoperability obtained through ontology by being able to link their workbench application with remote OpenTox web services. Although these developments are promising, there is an important need for an increased international coordination of efforts to develop a more unified, standardized, and open toxicology ontology framework (Hardy et al., 2012).

The toxicology community should not embark on projects that involve unnecessary investment in the creation of ontology that

[57] http://code.google.com/p/sbml2smw

can be found elsewhere. Certainly, the reuse of existing ontologies is not without overhead of a different kind: in some cases, the fact that different communities primarily developing different ontologies adopt different perspectives causes unexpected difficulties, and, in other cases, technological or institutional legacy hinders adaptation to novel requirements. Adequate and careful requirement elicitation and subsequent ecosystemic analysis of the potential "fit" between requirements and available offerings is therefore essential to ensure the success of ontology reuse in an integrating or interoperable ontology for toxicology. Although some gaps have already been identified, a more careful analysis is needed to pinpoint exactly those gaps that are both essential for a toxicology ontology and are unlikely to be addressed within the required timeframe by other communities; resources should be directed to these gap areas to "plug the holes." Another challenge in the adoption of existing ontologies is the level of pre-coordination of terms, i.e., complex terms such as "acute kidney disease" and "chronic kidney disease" present in ontologies as a whole, rather than the different semantic components such as "acute," "chronic," and "kidney disease" being available separately for post-coordination, which is indeed the strategy favored by Novartis in its eTOX ontology development efforts. This is an issue to be addressed by the ontology community as a whole, since approaches favoring post-coordination reduce the maintenance overhead for ontology maintainers and ease the adoption of the ontology by new communities. Finally, one of the core issues identified that will determine success (or ensure failure) is the level of buy-in, required from several different institutional "layers" – industry, academia, standardization organizations, regulatory agencies – to ensure optimal adoption. Strategies such as the involvement of publishing organizations in requiring public domain ontology annotations in primary literature also can help to facilitate the large-scale change that will be required.

References

ECETOC Technical Report No. 66 (1995). Skin Irritation and Corrosion: Reference Chemicals Data Bank.
http://www.ecetoc.org/technical-reports.

Gerberick, G. F., Ryan, C. A., Kern, P. S., et al. (2005). Compilation of historical local lymph node assay data for the evaluation of skin sensitization alternatives. *Dermatitis 16*, 157-202.

Hardy, B., Douglas, N., Helma, C., et al. (2010). Collaborative development of predictive toxicology applications. *J Cheminform. 2*, 7.

Hardy, B., Apic, G., Carthew, P., et al. (2012). A toxicology ontology roadmap. *ALTEX 29*, 129-137.

Goble, C. A., Bhagat, J., Aleksejevs, S., et al. (2010). Myexperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res. 38*, Web Server issue (Jul 2010), W677-682.

Lilienblum, W., Dekant, W., Foth, H., et al. (2008). Alternative methods to safety studies in experimental animals: role in the risk assessment of chemicals under the new European Chemicals Legislation (REACH). *Regulat. Toxicol. 82*, 211-236.

Mathäß, T., Haase, P., Kitano, H., and Toldo, L. (2010). SBML2SMW: Bridging System Biology with semantic web technologies for biomedical knowledge acquisition and hypothesis elicitation. In A. Facchiano and P. Romano (eds.), Proceedings of NETTAB-BBCC 2010 Biological Wikis, November 29 - December 1, 2010, Napoli, Italy. http://www.molgen.mpg.de/~lappe/stehr/NETTAB-BBCC-2010-Proceedings.pdf.

Renne, R., Brix, A., Harkema, J., et al. (2009). Proliferative and nonproliferative lesions of the rat and mouse respiratory tract. *Toxicol. Pathol. 37*, 5-73.

Shah, I. and Wambaugh, J. (2010). Virtual tissues in toxicology. *J. Toxicol. Environ. Health B Crit. Rev. 13*, 314-328.

Smith, B., Ashburner, M., Rosse, C., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol. 25*, 1251-1255. http://www.nature.com/nbt/journal/v25/n11/full/nbt1346.html

Spjuth, O., Helmus, T., Willighagen, E. L., et al. (2007). Bioclipse: an open source workbench for chemo- and bioinformatics. *BMC Bioinformatics 8*, 59.

Spjuth, O., Eklund, M., Helgee, E. A., et al. (2011). Integrated decision support for assessing chemical liabilities. *J. Chem. Inf. Model. 51*, 1840-1847.

Spjuth, O., Willighagen, E. L., Guha, R., et al. (2010). Towards interoperable and reproducible QSAR analyses: Exchange of datasets. *J. Cheminform. 2*, 5.

Steinbeck, C., Han, Y., Kuhn, S., et al. (2003). The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci. 43*, 493-500.

Thoolen, B., Maronpot, R. R., Harada, T., et al. (2010). Proliferative and nonproliferative lesions of the rat and mouse hepatobiliary system. *Toxicol. Pathol. 38*, 5-81.

Acknowledgements

OpenTox – An Open Source Predictive Toxicology Framework, www.opentox.org, was funded under the EU Seventh Framework Program: HEALTH-2007-1.3-3 Promotion, development, validation, acceptance and implementation of QSARs (Quantitative Structure-Activity Relationships) for toxicology, Project Reference Number Health-F5-2008-200787 (2008-2011). The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement n° 115002, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution.

Correspondence to

Barry Hardy, PhD
Douglas Connect
Baermeggenweg 14
4314 Zeiningen
Switzerland
Tel: +41 61 851 0170
e-mail: Barry.Hardy@douglasconnect.com