Toxicology in Vitro 28 (2014) 571–587

Contents lists available at ScienceDirect

Toxicology in Vitro

journal homepage: www.elsevier.com/locate/toxinvit

Automatic sorting of toxicological information into the IUCLID (International Uniform Chemical Information Database) endpoint-categories making use of the semantic search engine Go3R

Ursula G. Sauer [a,1], Thomas Wächter [b,c,1], Lars Hareng [d,1], Britta Wareing [d], Angelika Langsch [d], Matthias Zschunke [c], Michael R. Alvers [c], Robert Landsiedel [d,*]

[a] Scientific Consultancy – Animal Welfare, Neubiberg, Germany
[b] Biotechnology Center, Technische Universität Dresden, Dresden, Germany
[c] Transinsight GmbH, Dresden, Germany
[d] BASF SE, Product Safety – Experimental Toxicology and Ecology, Ludwigshafen, Germany

ARTICLE INFO

Article history:
Received 23 August 2013
Accepted 23 December 2013
Available online 2 January 2014

Keywords:
Semantic search engine
REACH
Registration dossiers
Regulatory toxicology
Data sharing
Alternative methods
3Rs principle

ABSTRACT

The knowledge-based search engine Go3R, www.Go3R.org, has been developed to assist scientists from industry and regulatory authorities in collecting comprehensive toxicological information with a special focus on identifying available alternatives to animal testing. The semantic search paradigm of Go3R makes use of expert knowledge on 3Rs methods and regulatory toxicology, laid down in the ontology, a network of concepts, terms, and synonyms, to recognize the contents of documents. Search results are automatically sorted into a dynamic table of contents presented alongside the list of documents retrieved. This table of contents allows the user to quickly filter the set of documents by topics of interest. Documents containing hazard information are automatically assigned to a user interface following the endpoint-specific IUCLID5 categorization scheme required, e.g. for REACH registration dossiers. For this purpose, complex endpoint-specific search queries were compiled and integrated into the search engine (based upon a gold standard of 310 references that had been assigned manually to the different endpoint categories). Go3R sorts 87% of the references concordantly into the respective IUCLID5 categories. Currently, Go3R searches in the 22 million documents available in the PubMed and TOXNET databases. However, it can be customized to search in other databases including in-house databanks.

1. Introduction

The revised European Union (EU) *Directive 2010/63/EU on the protection of animals used for scientific purposes* (Anon, 2010) has explicitly implemented the 3Rs principle to replace, reduce, and refine animal testing (Russell and Burch, 1959). Accordingly, animal testing may not be performed if "*a scientifically satisfactory method or testing strategy, not entailing the use of live animals*" can be used instead. Furthermore, it has to be ensured that the number of animals used in a procedure is reduced to the minimum, in the same way as any possible pain, suffering, distress, or lasting harm (Article 4 of Anon, 2010). Collecting comprehensive and up-to-date information on available 3Rs methods and on existing information and data related to the scientific topic in question is a prerequisite to fulfilling these legal provisions and to ensuring the indispensability of animal tests. In light of the continuously increasing abundance of information available in the World Wide Web, the gathering of relevant information is an increasingly complex and time-consuming challenge.

Since 2007, the semantic search engine Go3R (available free of charge at www.Go3R.org) has been developed to support scientists in searching for 3Rs relevant information in the Internet (Sauer et al., 2009). Semantic search tools 'understand' what the user is searching for and automatically select and sort relevant pieces of information (Schroeder, 2003). The semantic search paradigm of Go3R indexes documents with bibliometric metadata classes (i.e. authors, journals, cities, countries, year of publication) and with the classes of an ontology that has been specifically created for Go3R (see information box I for ontology-related definitions). In processing search queries, Go3R automatically compares the vocabulary used in the titles, abstracts, and keywords or Medical Subject Headings (MeSH; the indexing thesaurus of the PubMed

* Corresponding author. Address: Experimental Toxicology and Ecology, BASF SE, 67056 Ludwigshafen, Germany. Tel.: +49 (0)621 60 56 203.
E-mail address: robert.landsiedel@basf.com (R. Landsiedel).
[1] Equal first authorship.

http://dx.doi.org/10.1016/j.tiv.2013.12.011
0887-2333/

database; http://www.nlm.nih.gov/[2]) of the documents to the classes of the Go3R ontology and sorts the documents retrieved into the respective recognized classes. The outcome of this sorting is presented to the user alongside the search retrieval in the form of a dynamic 'table of contents' (see Fig. 1, presenting the exemplary search query 'eye irritation'). This table of contents can be used as a navigation tool: By filtering for its respective 'chapters' and 'sub-chapters' (i.e. by clicking onto them on the user interface), broad search results are quickly restricted to relevant information (see Fig. 2, showing subordinate levels of the ontology root concept '3Rs alternative methods' and restriction of the initial search query 'eye irritation' to documents in which the 'bovine corneal opacity and permeability assay' is mentioned).

Information Box I Ontology-related definitions
(Adapted from: Uschold and Gruninger (1996), Boyce and Pahl (2007); and the glossary of the US National Library of Medicine's Unified Medical Language System (UMLS®); available at: http://www.nlm.nih.gov/research/umls/new_users/glossary.html)

Ontology

A framework for representing concepts (things or ideas about things), the relationships that exist between the concepts, and the properties they might have. In the ontology, concepts are linked in a strictly hierarchical (tree-like) structure ensuring that subordinate ('parent–child') relationships not only hold between classes and their direct parents (e.g. 'mouse' and 'rat' as direct subordinate concepts to 'rodent'), but also to all further superordinate concepts (ancestors) in the given ontology branch (e.g. 'mammal', 'vertebrate', 'animal').

Concept (or class)

The fundamental unit of meaning in the knowledge source. A concept represents a single meaning and contains all textual labels from any source that express that meaning in any way, whether formal or casual, verbose or abbreviated (e.g. 'BCOP' and 'OECD TG (test guideline) 437' as textual labels for the 'bovine corneal opacity and permeability assay'). All of the textual labels within a concept are synonymous.

Root concept

The highest-level concept of a given thematic branch of the ontology.

Term

The textual label assigned to a concept.

Go3R is based upon prior work by the *Technical University Dresden* and *Transinsight GmbH Dresden* to develop GoPubMed (Doms and Schroeder, 2005; www.gopubmed.org), a semantic search tool to search general biomedical information provided in the PubMed database. Go3R has been designed to retrieve information on alternative methods in all areas of biomedical research. However, due to the main topic of the Go3R research project, the search engine currently has a special focus on retrieving information on 3Rs methods in the realm of regulatory toxicity testing. Accordingly, Go3R has been aligned to search the databases *PubMed* and *Toxicology Data Network* (TOXNET; http://toxnet.nlm.nih.gov/). Addressing the increasing economic importance of nanotechnological developments and the resulting amount of research investigating the safety of nanomaterials (Oomen et al., 2013), an additional 'nano'-ontology (Fig. 3) was created and linked to Go3R (Sauer et al., 2011). Thereby, the 3Rs relevant concepts of the Go3R ontology can be combined with concepts of the nanotechnology and nanotoxicology domains (such as 'endpoints methods for nanomaterial characterization and testing'; Fig. 3).

[2] *Note:* All websites were accessed in June 2013.

Depending on the vocabulary used in title and abstract, 3Rs-related documents of relevance for regulatory toxicity purposes might be sorted into a number of different Go3R ontology branches, i.e. '3Rs methods in the life sciences' (see Fig. 2 for subordinate levels of concepts to this root concept), 'cultured cells, tissues, etc.', '*in vitro* experimental design', 'cell culture technology, etc.', '3Rs method types' (listing different test systems), or '3Rs toxicity testing strategies' (see Fig. 1 for overview of root concepts). When *in vitro* studies are compared to *in vivo* tests or reduction and refinement methods are referred to, the ontology branches 'animal species', 'animal test method', and '*in vivo* experimental design' are also of relevance. Of note, the vocabulary used in different documents to describe one specific type of study is not uniform, but is incumbent upon the respective author's choice. Therefore a document containing information on a given toxicological endpoint (e.g. 'eye irritation') might be sorted into a combination of any of these (or even further) ontology branches. To allow searching for endpoint-specific information, documents from possibly many different branches need to be aggregated and structured by endpoint-specific topics to prevent users from manually checking multiple branches.

Against this background it was the aim of the present study to extend the Go3R search engine to automatically sort all relevant documents by toxicological endpoint *regardless of the vocabulary used by the respective authors*. Even if it did not explicitly mention the endpoint under consideration or the respective test method, a document should still be recognized as relevant for the given endpoint, e.g. due to a specific pattern of vocabulary mentioning specific cell lines, cellular endpoints, or endpoint detection methods relating to the respective endpoint, and it should be sorted accordingly.

The categorization scheme of the 5th version of the *International Uniform Chemical Information Database* (IUCLID5) was selected as a template for this sorting task. IUCLID5 (http://iuclid.eu) plays a central role in collecting, storing, submitting, and exchanging data in fulfilling the data submission requirements of, e.g. the EU *regulation on the Registration, Evaluation, Authorisation, and Restriction of Chemicals* (REACH; Anon, 2006), the *OECD Cooperative Chemicals Assessment Programme* (http://www.oecd.org/chemicalsafety/risk-assessment/oecdcooperativechemicalsassessmentprogramme.htm), or the EU *regulation concerning the making available on the market and use of biocidal products* (Anon, 2012). The IUCLID categorization scheme is closely linked to the OECD harmonized templates (http://www.oecd.org/ehs/templates/) for structuring data entry systems to report the results of tests determining human health and environmental effects of substances.

Chapter 7 of the IUCLID5 scheme provides the structure to submit toxicological information. This section encompasses 12 distinct endpoint-specific categories with sub-sorting into a total of 30 sub-categories (Table 1 and Fig. 4). Detailed guidance has been published on how to select appropriate information and sort it into the respective endpoint-specific categories (see: ECHA and OECD, 2007; ECHA, 2012a,b).

In regard to chemical substances, the REACH regulation prescribes that all substances manufactured or imported in quantities above 1 tonne per year have to be registered with the European Chemicals Agency (ECHA). Information requirements for the registration dossiers, making use of the IUCLID categorization scheme, increase with increasing tonnage. More than 100,000 registration dossiers for a total of 30,000 chemical substances are foreseen by the end of May 2018 (estimations of the *German Chemical Industry Association*, VCI – Verband der Chemischen Industrie; based upon the numbers of pre-registered substances). By this date, the so-called phase-in substances, i.e. substances that were already manufactured or imported before the implementation of the REACH regulation, are to be registered.