Integrating Semantic Technologies in Environmental Risk Assessment: A Vision

Erik B. Myklebust[1,2], Ernesto Jimenez-Ruiz[2,3], Zofia C. Rudjord[1], Raoul Wolf[1] and Knut Erik Tollefsen[1]

[1]Norwegian Institute for Water Research (NIVA), Gaustadalléen 21, 0349 Oslo, Norway
[2]Department of Informatics, University of Oslo, P.O. box 1080 Blindern, 0316 Oslo, Norway
[3]Alan Turing Institute, British Library, 96 Euston Road, London NW1 2DB, United Kingdom
E-mail contact: erik.b.myklebust@niva.no

1. Introduction

Extending the scope of risk assessment models is a long-term goal in ecotoxicological research. However, biological effect data is only available for a few combinations of species–chemical pairs. We aim at designing tools and methods to extrapolate from known to unknown combinations, in order to facilitate risk assessment predictions on a population basis.

In the present study we consider the use of semantic technologies[1] to improve risk assessment analysis. Linking openly available taxonomies and chemical databases with curated data from experts will improve the coverage and access to data. This knowledge can then be transferred into rulesets within an ontology[2]. This ontology can be as straightforward as species taxonomy, or more involved to include interactions between species.

2. Methods

Risk assessment requires the use of different metadata, e.g., from taxonomies, chemicals, or toxicity data. In addition, data collected by researchers in the field or the lab is used as direct input. A partial goal of our work is to effectively use semantic technologies to link these datasets together.

First, we collect metadata from different sources and create a common language for communication among sources. This collection of (meta-)data will form the basis of a knowledge graph[3]. We add an ontological layer onto the knowledge graph, which includes rules and axioms that either validate (e.g., to rule out unreasonable data extrapolation) or expand the data (e.g., infer implicit facts from explicit knowledge). Second, we design a system to extract new knowledge from the knowledge graph. This can be done using both a deductive (based on theory) and an inductive (based on data) approach using probabilistic predictions.

Reasoning is the use of ontological rules to extract new relationships in the knowledge graph. The new data will depend on rules we define for the particular dataset, and will also require expert insight when defining the rules. These rules can be, 1) hierarchical rules, e.g., *"if x is of type* Solibacteres *then x is of type Bacteria",* 2) general rules, e.g., *"if a chemical affects some species then that chemical is toxic",* 3) domain rules, e.g., *"if x affects something, x is a type of Chemical",* and 4) range rules, e.g., *if y is affected by something, y is a type of Species"*. In some cases, domain knowledge is not available, this is where the probabilistic approach is implemented.

Machine learning and especially knowledge graph completion can aid the decision-making process when relating between species and chemicals. Algorithms uses relations between entities to create embeddings (contextual dimensionality reduction) of entities into a lower dimensional vector space. E.g., *{Chemical A, affects, Species X}* and *{Chemical A, affects, Species Y}* are defined in the knowledge graph. We can expect *Species X* and *Species Y* to be close in the vector space, since they are related to the same entity (by the same relation, e.g., belonging to the same genus). If the distance between two entities is sufficiently small, we can use data involving *Species X* as it were attached to *Species Y* and *vice versa*. Moreover, we rank the findings by most probable, and select candidates only if the findings are above an *a priori* determined threshold.

Explaining the extrapolated data is a major part of the work. When defining ontological rules, we explain data addition by those rules. However, when resorting to machine learning models for deducing data, explanation (or reason) for the prediction is essential for trusting the results of risk assessment. We will apply a method

for comparing predicted data with the knowledge graph to get an explanation for the possibility rating. This way, experts can verify that the prediction functionality is consistent with expected values.

The methods described above will increase the size of confidence intervals in the risk assessments. However, the quality of the overall assessment for a population will increase.

3. Case studies

Developing robust case studies is key to successful system development. We need cases with a complete qualitative performance analysis, compared to previous single compound analysis. As there is currently no unified framework for ontological risk assessment, expert judgement is crucial to verify results of semantic ontology approaches. A simple case study taking advantage of a limited chemical applicability domain represented by data from pesticide monitoring of Norwegian rivers and creeks[4], NCBI taxonomy classification and nomenclature[5] and effects data from the US EPA Ecotoxicology Database (ECOTOX)[6] will be used to illustrate the process of integrating semantic technologies into environmental risk assessment.

4. Conclusion

We have outlined a vision for integration of semantic technology in environmental risk assessment (*Figure 1*). This approach will yield new insight by combining and inferring from available data and subsequently improve the quality of risk assessment.

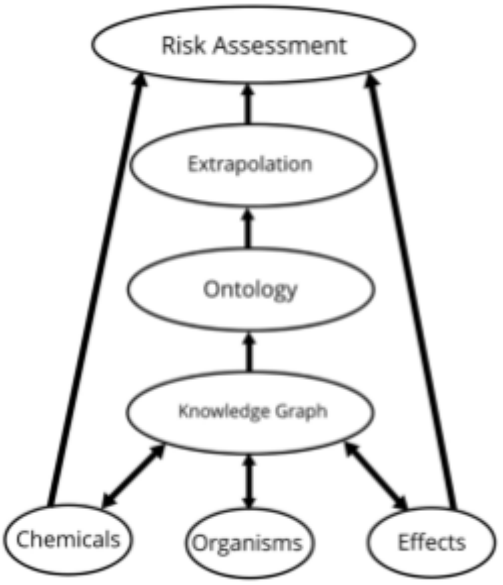

Figure 1. Workflow from sources to risk assessment. Available effects and toxicity data flows straight to risk assessment while incomplete data passes through the extrapolation process.

5. References

[1] Pascal Hitzler, Markus Krötzsch, Sebastian Rudolph: Foundations of Semantic Web Technologies. Chapman and Hall/CRC Press 2010, ISBN 9781420090505

[2] Markus Krötzsch, Description Logic Rules. Studies on the Semantic Web 8, IOS Press 2010, ISBN 978-1-60750-654-6, pp. 1-263

[3] Ehrlinger, Lisa & Wöß, Wolfram. Towards a Definition of Knowledge Graphs. Posters&Demos@SEMANTiCS 2016.

[4] https://www.nibio.no/en/subjects/environment/the-norwegian-agricultural-environmental-monitoring-programme-jova

[5] The National Center for Biotechnology Information, NCBI Taxonomy, https://www.ncbi.nlm.nih.gov/guide/taxonomy/, accessed 28.11.18.

[6] US Environmental Protection Agency, Ecotoxicology Database, https://cfpub.epa.gov/ecotox/, accessed 28.11.18.

Acknowledgments - This work is an collaboration between NIVA's Computational Toxicology Program (NCTP, https://www.niva.no/en/projectweb/nctp) and the University of Oslo's Department of Informatics. Funding: PhD scholarship (Research Council of Norway, RCN), the MixRisk project (RCN 237889), the AIDA project (UK Government's Defence & Security Programme in support of the Alan Turing Institute), SIRIUS Centre for Scalable Data Access (RCN 237889), and the BIGMED project (IKT 259055).