International Journal of Information Management 33 (2013) 567–582

Contents lists available at SciVerse ScienceDirect

International Journal of Information Management

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

Towards model governance in predictive toxicology

Anna Palczewska[a], Xin Fu[b,*], Paul Trundle[a], Longzhi Yang[a], Daniel Neagu[a], Mick Ridley[a], Kim Travis[c]

[a] School of Computing, Informatics and Media, University of Bradford, Richmond Road, BD7 1DP Bradford, UK
[b] Department of Management Science, School of Management, Xiamen University, Xiamen 361005, China
[c] Syngenta Jealott's Hill International Research Centre, Bracknell RG42 6EY, UK

ARTICLE INFO

Article history:
Available online 26 March 2013

Keywords:
Model governance
Data governance
Predictive toxicology
Information representation
Knowledge management
Quality assessment

ABSTRACT

Efficient management of toxicity information as an enterprise asset is increasingly important for the chemical, pharmaceutical, cosmetics and food industries. Many organisations focus on better information organisation and reuse, in an attempt to reduce the costs of testing and manufacturing in the product development phase. Toxicity information is extracted not only from toxicity data but also from predictive models. Accurate and appropriately shared models can bring a number of benefits if we are able to make effective use of existing expertise. Although usage of existing models may provide high-impact insights into the relationships between chemical attributes and specific toxicological effects, they can also be a source of risk for incorrect decisions. Thus, there is a need to provide a framework for efficient model management. To address this gap, this paper introduces a concept of model governance, that is based upon data governance principles. We extend the data governance processes by adding procedures that allow the evaluation of model use and governance for enterprise purposes. The core aspect of model governance is model representation. We propose six rules that form the basis of a model representation schema, called Minimum Information About a QSAR Model Representation (MIAQMR). As a proof-of-concept of our model governance framework we develop a web application called Model and Data Farm (MADFARM), in which models are described by the MIAQMR-ML markup language.

1. Introduction

Efficient access to integrated platforms for toxicological modelling is increasingly important for the chemical, pharmaceutical, cosmetics and food industries. It supports the decision making process for product discovery and development, e.g. drugs, pesticides, cosmetics and food protection. The whole process of product development may last for approximately ten years and is divided into four phases: discovery, profile, evaluation and support. In the first phase, from millions of chemical compounds, thousands are selected according to their biological, chemical or physical properties. This chemical compounds group is profiled against various targets (e.g. biochemical and physiological targets related to metabolism, growth, development, nervous communication) and tens of them pass to the evaluation phase. After the evaluation phase usually only very limited number of chemicals are selected as a product that can be introduced into the market. Thus, many organisations focus on better information organisation and reuse in order to reduce the cost of testing and manufacturing in the product development phase.

Over several years, many different types of computational methods, such as structure–activity relationship (SAR); quantitative structure activity relationship (QSAR); kinetic methods and expert systems have been developed to identify or predict toxic effects on human beings, animals and the environment. A large number and variety of models could be, and are still, created thanks to the continuously increasing amount of available experimental data that covers various domains of chemical space. Currently, good quality models are considered to be a cost efficient alternative to *in vivo* and *in vitro* testing. In order to ensure the safety of humans, animals, and the environment, they may never become a complete substitute for *in vivo* experiments. However, these models can be used to reduce the cost and negative impacts of animal testing. Thus, for domains such as pharmacy, cosmetics or food production experimental toxicity data and toxicity models have become valuable information assets. The collection of data and predictive models and their management is required to support the decision to exclude chemicals that may fail in the profile and evaluation phases.

Having such a wealth of previously developed models at our disposal can bring a number of benefits if we are able to make

* Corresponding author. Tel.: +86 5922181207.
E-mail addresses: a.m.wojak@bradford.ac.uk (A. Palczewska), xfu@xmu.edu.cn (X. Fu), p.r.trundle@bradford.ac.uk (P. Trundle), L.Yang8@bradford.ac.uk (L. Yang), D.Neagu@bradford.ac.uk (D. Neagu), M.J.Ridley@bradford.ac.uk (M. Ridley), kim.travis@syngenta.com (K. Travis).

0268-4012/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijinfomgt.2013.02.005

effective reuse of them. Trained models usually represent a significant investment of time, and may contain high-impact insights into the relationships between particular chemical attributes and specific toxicological effects. In the past, published models in the literature were often unused and unseen within communities because they were not publicly available or not annotated to be suitable for reuse. They are often difficult to restore to a useful form as the published details are either incomplete or the supporting information is missing. Lack of a standard description format for model representation, and the lack of stringent reviewing and authors' carelessness have been identified as the main causes for incomplete model descriptions (Lowe et al., 2010; Novere et al., 2005). Reproducing work to reach the same conclusions is obviously an inefficient use of time in the best case, and in the worst case a different and possibly incorrect conclusion may be reached. In such situations the knowledge that was previously discovered and encapsulated within a predictive model may be lost. To avoid this, the knowledge should be captured together with the human experience of knowledge itself and its use, and the proper management of such knowledge is required (Serna, 2012).

Over several years, various predictive toxicology systems (e.g. AMBIT,[1] InkSpot,[2] OpenTox,[3] OCHEM[4] or JRC QSAR DB[5]) have been proposed in an attempt to address the aforementioned problems. Predictive toxicology systems are high quality data warehouses that support the model development process and collaboration between various institutions. To make models more reusable sources of information, various model representations and ontologies have been implemented for each toxicology system. This allows users to build models or workflows and reuse them only within a particular system (Cartmell et al., 2005; Cassano et al., 2010; Hardy et al., 2010; Sushko et al., 2011). To make use of existing models, users are required to register with the system and also submit data that they use for predicting a given endpoint. This discourages modellers to use such predictive toxicology systems to some extent. Often, the data in use is confidential and modellers do not fully trust the existing systems. Additionally, model exchange across different platforms is challenging, due to the various model representation formats.

Once the predictive models have been built, it is important to consider global, effective and efficient ways of representing and (re)using them. A good model representation would contribute to further model management tasks, such as model validation, model identification, model comparison and model ranking. In this paper we aim to address these problems and provide a concept of model governance in predictive toxicology. Model governance extends the principles of data governance (DGI, 2010). The key aspects of model governance include: policies and standards for information (data and models) representation, information quality and information security. All refer to the efficient management of the information as enterprise assets (Sweden, 2008). The current Organisation for Economic Cooperation and Development (OECD) guidance (OECD, 2011) for model validation and data quality in ITS (Integrated Testing Strategies) frameworks under the Registration, Evaluation, Authorisation & restriction of Chemicals (REACH) (REACH, 2012) regulatory framework can be combined with organisation strategies in order to provide definitions, standards and policies that allow efficient data and model governance. The principles are summarised as:

- Defined Endpoint (Principle 1): The intent of this principle is to ensure clarity in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions.
- Unambiguous Algorithm (Principle 2): The intent of this principle is to ensure transparency in the model algorithm that generates predictions of an endpoint from information on chemical structure and/or physicochemical properties. It is recognised that, in the case of commercially-developed models, this information is not always made publicly available. However, without this information, the performance of a model cannot be independently established, which is likely to represent a barrier for regulatory acceptance.
- Defined Domain of Applicability (Principle 3): The need to define an applicability domain expresses the fact that (Q)SARs are reductionist models which are inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions.
- Appropriate Measures of Goodness-of-Fit, Robustness and Predictivity (Principle 4): The wording of the principle is intended to simplify the overall set of principles, but not to lose the distinction between the internal performance of a model (as represented by goodness-of-fit and robustness) and the predictivity of a model (as determined by external validation).
- Mechanistic Interpretation (Principle 5): It is recognised that it is not always possible, from a scientific viewpoint, to provide a mechanistic interpretation of a given (Q)SAR, or there even be multiple mechanistic interpretations of a given model. The absence of a mechanistic interpretation for a model does not mean that a model is not potentially useful in the regulatory context. The intention of this principle is not to reject models that have no apparent mechanistic basis, but to ensure that some consideration is given to the possibility of a mechanistic association between the descriptors used in a model and the endpoint being predicted, and also to ensure that this association is documented.

The above principles can be used as guidance for the creation of a model representation. They cover the minimum information that is required for model evaluation according to its development and usage. The model information content can be encapsulated in a metadata object and meta-analysis can provide well-grounded statements on the degree of confidence one may have in the model's predictions (Wu et al., 2011). In this paper we introduce six rules that define the Minimum Information about a QSAR Model Representation (MIAQMR), and we also propose the MIAQMR-ML markup language that encapsulates the above principles in a semantic way. For proof-of-concept purposes, this model representation is implemented within the Model and Data Farm Governance Framework (MADFARM) prototype system that currently is validated and tested internally in the Product Safety Department, Syngenta Ltd., UK.

The paper is organised as follows. Section 2 includes a review of existing frameworks for model development in predictive toxicology. In Section 3, a discussion of the novel concept of model governance is provided. In Section 4 data and model quality are discussed. The model object representation called MIAQMR (Minimum Information about a QSAR Model Representation) is proposed in Section 5. Section 6 presents the implementation of the MADFARM system. Section 7 concludes the paper and proposes future research directions.

[1] http://ambit.sourceforge.net.
[2] http://www.inkspotscience.com.
[3] http://www.opentox.org.
[4] http://ochem.eu.
[5] http://qsardb.jrc.ec.europa.eu/qmrf/.