Plan Overview

A Data Management Plan created using DMPonline

Title: Towards a generalized toxicity prediction model for oxide nanomaterials using integrated data

Creator: Ammar Ammar

Principal Investigator: Egon Willighagen

Contributor: Ammar Ammar

Affiliation: Other

Funder: European Commission

Template: Horizon 2020 DMP

ORCID iD: 0000-0001-7542-0286

Project abstract:

A generalized toxicity classification model for 7 different oxide nanomaterials is presented in this study. A data set extracted from multiple literature sources and screened by physicochemical property based quality scores were used for model development. Moreover, a few more preprocessing techniques, such as synthetic minority over-sampling technique, were applied to address the imbalanced class problem in the data set. Then, classification models using four different algorithms, such as generalized linear model, support vector machine, random forest, and neural network, were developed and their performances were compared to find the best performing preprocessing methods as well as algorithms. The neural network model built using the balanced data set was identified as the model with best predictive performance, while applicability domain was defined using k-nearest neighbours algorithm. The analysis of relative attribute importance for the built neural network model identified dose, formation enthalpy, exposure time, and hydrodynamic size as the four most important attributes. As the presented model can predict the toxicity of the nanomaterials in consideration of various experimental conditions, it has the advantage of having a broader and more general applicability domain than the existing quantitative structure-activity relationship model. [Source: https://doi.org/10.1038/s41598-018-24483-z]

ID: 96221

Start date: 01-06-2021

End date: 01-06-2024

Last modified: 23-03-2022

Grant number / URL: 814572

Copyright information:

Towards a generalized toxicity prediction model for oxide nanomaterials using integrated data - Initial DMP

1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation**
- **Explain the relation to the objectives of the project**
- **Specify the types and formats of data generated/collected**
- **Specify if existing data is being re-used (if any)**
- **Specify the origin of the data**
- **State the expected size of the data (if known)**
- **Outline the data utility: to whom will it be useful**

- The dataset resemble an RDF version of a published data set extracted from multiple literature sources and screened by physicochemical property based quality scores containing physicochemical characterization and toxicological endpoint measurements for 7 different oxide nanomaterials.
- As part of NanoSolveIT project WP1, we aim to provide a knowledge base for nanosafety related data by integrating existing datasets using semantic web technologies and semantic modeling. This dataset will be integrated into the under developing knowledgebase.
- The dataset is provided as an RDF model produced and stored in (.nq) format (N-Quads).
- The existing data will be reused for several purposes like toxicity prediction, nano-QSAR and NanoInChI calculations.
- The original data, from which the RDF version is derived, is provided as a supplementary material of a peer-reviewed article with the DOI: https://doi.org/10.1038/s41598-018-24483-z
- The dataset is ~4 MB in total, fairly small and easy to store and transfer.
- The dataset will be useful to researchers in the nanosafety domain

2. FAIR data

2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision)**
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?**
- **Outline naming conventions used**
- **Outline the approach towards search keyword**
- **Outline the approach for clear versioning**
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

- The data is hosted on GitHub and archived through Zenodo and has a DOI:

https://zenodo.org/record/5743204

- For naming conventions, the data is in semantic web format and mapped to known ontologies (e.g. eNanoMapper, BAO and OBO)
- For versioning, Both GitHub semantic versioning and Zenodo versioning are adopted.
- For metadata standards, the dataset itself contains metadata in RDF format using ontologies like VoID. Also, the datasets is described with a web page and annotated using JSON-LD following the caLIBRAte quality criteria (as part of the NSDRA framwork). The dataset description page: https://nanocommons.github.io/datasets/overview/5743204.html

2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**
- **Specify how the data will be made available**
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**
- **Specify where the data and associated metadata, documentation and code are deposited**
- **Specify how access will be provided in case there are any restrictions**

- The whole RDF data is openly accessbile through a SPARQL endpoint: http://81.169.200.64:8879/sparql
- Both human and software can access the data since it is machine readable. The data is directly queryable using a user interface and can be accessed programmatically using any software agent that can parse semantic format data and communicate with SPARQL query language.
- Data Access:

- - GitHub: https://github.com/ammar257ammar/RDFied-datasets/tree/main/01-moesm/rdf
  - Zenodo: https://zenodo.org/record/5743204
  - Dataset description and JSON-LD annotation: https://nanocommons.github.io/datasets/overview/5743204.html
  - SPARQL query interface: http://81.169.200.64:8090/

2.3 Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?**

- Data/Metadata interoperability is natively availably due to adopting the semantic web approach. Using SPARQL query language, data can be queried, transformed, mapped to other formats, models and standards.
- All datasets produced to be integrated into the NanoSolveIT nanosafety knowledgebase will use the same standards/ontologies.

2.4 Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible**
- **Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed**
- **Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why**
- **Describe data quality assurance processes**
- **Specify the length of time for which the data will remain re-usable**

The data is licensed by Creative Commons Attribution 4.0 International
The data is open for anyone without authentication and authorizations. It is licensed with open access license and made available with a globally unique DOI for indefinite period.

3. Allocation of resources

Explain the allocation of resources, addressing the following issues:

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs**
- **Clearly identify responsibilities for data management in your project**
- **Describe costs and potential value of long term preservation**

This work is part of Task 1.3 (WP1) in the NanoSolveIT project which 48 man months for Maastricht University are allocated for the whole work package.

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

The data is openly available and so the original datasets so there is not confidential or sensitive data here.

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics

deliverables. Include references and related technical aspects if not covered by the former

Not applicable

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

https://data.europa.eu/data/datasets/open-research-data-the-uptake-of-the-pilot-in-the-first-calls-of-horizon-2020?locale=en