Plan Overview

A Data Management Plan created using DMPonline

Title: Manually curated transcriptomics data collection for toxicogenomic assessment of engineered nanomaterials

Creator: Ammar Ammar

Principal Investigator: Egon Willighagen

Contributor: Ammar Ammar

Affiliation: Other

Funder: European Commission

Template: Horizon 2020 Template

ORCID iD: 0000-0001-7542-0286

Project abstract:

Toxicogenomics (TGx) approaches are increasingly applied to gain insight into the possible toxicity mechanisms of engineered nanomaterials (ENMs). Omics data can be valuable to elucidate the mechanism of action of chemicals and to develop predictive models in toxicology. While vast amounts of transcriptomics data from ENM exposures have already been accumulated, a unified, easily accessible and reusable collection of transcriptomics data for ENMs is currently lacking. In an attempt to improve the FAIRness of already existing transcriptomics data for ENMs, we curated a collection of homogenized transcriptomics data from human, mouse and rat ENM exposures in vitro and in vivo including the physicochemical characteristics of the ENMs used in each study. [Source: https://doi.org/10.1038/s41597-021-00808-y]

ID: 98453

Start date: 01-06-2021

End date: 01-06-2024

Last modified: 13-04-2022

Grant number / URL: 814572

Manually curated transcriptomics data collection for toxicogenomic assessment of engineered nanomaterials - 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 aimed at improving the FAIRness of already existing transcriptomics data for ENMs. The dataset is a curated collection of homogenized transcriptomics data from human, mouse and rat ENM exposures in vitro and in vivo including the physicochemical characteristics of the ENMs used in each study
- 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/s41597-021-00808-y
- The dataset is ~50 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/5744003
- 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/5744003.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/03-fair-dataset/rdf
  - Zenodo: https://zenodo.org/record/5744003
  - Dataset description and JSON-LD annotation: https://nanocommons.github.io/datasets/overview/5744003.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

Manually curated transcriptomics data collection for toxicogenomic assessment of engineered nanomaterials - Detailed DMP

1. Data summary

State the purpose of the data collection/generation

The dataset resemble an RDF version of a published data set aimed at improving the FAIRness of already existing transcriptomics data for ENMs. The dataset is a curated collection of homogenized transcriptomics data from human, mouse and rat ENM exposures in vitro and in vivo including the physicochemical characteristics of the ENMs used in each study

Explain the relation to the objectives of the project

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.

Specify the types and formats of data generated/collected

The dataset is provided as an RDF model produced and stored in (.nq) format (N-Quads).

Specify if existing data is being re-used (if any)

The existing data will be reused for several purposes like toxicity prediction, nano-QSAR and NanoInChI calculations.

Specify the origin of the data

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/s41597-021-00808-y

State the expected size of the data (if known)

The dataset is ~50 MB in total, fairly small and easy to store and transfer.

Outline the data utility: to whom will it be useful

The dataset will be useful to researchers in the nanosafety domain

2.1 Making data findable, including provisions for metadata [FAIR data]

Outline the discoverability of data (metadata provision)

NA

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?

The data is hosted on GitHub and archived through Zenodo and has a DOI: https://zenodo.org/record/5744003

Outline naming conventions used

For naming conventions, the data is in semantic web format and mapped to known ontologies (e.g. eNanoMapper, BAO and OBO)

Outline the approach towards search keyword

NA

Outline the approach for clear versioning

For versioning, Both GitHub semantic versioning and Zenodo versioning are adopted.

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

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/5744003.html

2.2 Making data openly accessible [FAIR data]

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

The whole RDF data is openly accessible through a SPARQL endpoint:
http://81.169.200.64:8879/sparql

Specify how the data will be made available

The whole RDF data is openly accessbile through a SPARQL endpoint:
http://81.169.200.64:8879/sparql
It is also available for download through Zenodo: https://zenodo.org/record/5744003

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)?

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.

Specify where the data and associated metadata, documentation and code are deposited

Data Access:

- GitHub: https://github.com/ammar257ammar/RDFied-datasets/tree/main/03-fair-dataset/rdf
- Zenodo: https://zenodo.org/record/5744003
- Dataset description and JSON-LD annotation: https://nanocommons.github.io/datasets/overview/5744003.html
- SPARQL query interface: http://81.169.200.64:8090/

Specify how access will be provided in case there are any restrictions

NA

2.3 Making data interoperable [FAIR data]

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

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.

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?

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) [FAIR data]

Specify how the data will be licenced to permit the widest reuse possible

The data is licensed by Creative Commons Attribution 4.0 International

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

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.

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

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.

Describe data quality assurance processes

NA

Specify the length of time for which the data will remain re-usable

Indefinite period

3. Allocation of resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

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.

Clearly identify responsibilities for data management in your project

NA

Describe costs and potential value of long term preservation

NA

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

Manually curated transcriptomics data collection for toxicogenomic assessment of engineered nanomaterials - Final review DMP

1. Data summary

State the purpose of the data collection/generation

Question not answered.

Explain the relation to the objectives of the project

Question not answered.

Specify the types and formats of data generated/collected

Question not answered.

Specify if existing data is being re-used (if any)

Question not answered.

Specify the origin of the data

Question not answered.

State the expected size of the data (if known)

Question not answered.

Outline the data utility: to whom will it be useful

Question not answered.

2.1 Making data findable, including provisions for metadata [FAIR data]

Outline the discoverability of data (metadata provision)

Question not answered.

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?

Question not answered.

Outline naming conventions used

Question not answered.

Outline the approach towards search keyword

Question not answered.

Outline the approach for clear versioning

Question not answered.

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

Question not answered.

2.2 Making data openly accessible [FAIR data]

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

Question not answered.

Specify how the data will be made available

Question not answered.

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)?

Question not answered.

Specify where the data and associated metadata, documentation and code are deposited

Question not answered.

Specify how access will be provided in case there are any restrictions

Question not answered.

2.3 Making data interoperable [FAIR data]

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

Question not answered.

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?

Question not answered.

2.4 Increase data re-use (through clarifying licenses) [FAIR data]

Specify how the data will be licenced to permit the widest reuse possible

Question not answered.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

Question not answered.

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

Question not answered.

Describe data quality assurance processes

Question not answered.

Specify the length of time for which the data will remain re-usable

Question not answered.

3. Allocation of resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these

costs

Question not answered.

Clearly identify responsibilities for data management in your project

Question not answered.

Describe costs and potential value of long term preservation

Question not answered.

4. Data security

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

Question not answered.

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

Question not answered.

6. Other

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

Question not answered.