

# The Challenges Associated with Developing Data Collections to Support Modelling of Nanomaterial Effects

R. L. Marchese Robinson<sup>1</sup>, A.-N. Richarz<sup>1</sup>, M. T. D. Cronin<sup>1</sup>, T. Puzyn<sup>2</sup>, E. Benfenati<sup>3</sup>, N. Golbamaki Bakhtyari<sup>3</sup>, M. G. Papadopoulos<sup>4</sup>



<sup>1</sup>School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Liverpool, United Kingdom; <sup>2</sup>Faculty of Chemistry, University of Gdansk, Gdansk, Poland; <sup>3</sup>Laboratory of Environmental Chemistry and Toxicology, Mario Negri Institute, Milan, Italy; <sup>4</sup>Institute of Biology, Medicinal Chemistry and Biotechnology, National Hellenic Research Foundation, Athens, Greece

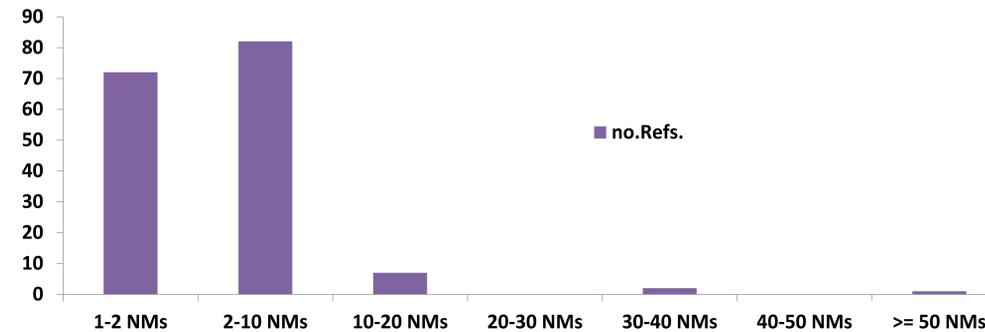
R.L.MarcheseRobinson@ljmu.ac.uk <http://www.nanopuzzles.eu/>

## Introduction

- The **NanoPUZZLES EU project** is one of a number of activities funded by the European Commission to develop a set of computational resources for **predicting the toxicological effects of nanomaterials**.
- The project is investigating category formation/grouping of “similar” nanomaterials, to enable “**read-across**” predictions, and developing **(Q)SAR models/modelling approaches**.
- Both kinds of modelling approach require **high quality data** available in a **suitable electronic format**.
- Data are being collected** and will be submitted to an appropriate, **publicly available**, searchable online database **at the end of the project**.
- In developing this data collection, a number of **challenges are being addressed** in order to achieve the **ultimate goal** of providing a searchable, high quality data collection to modellers.

## Identifying References with Relevant Data

- Primary experimental, English language references investigating **cytotoxicity and genotoxicity of nanomaterials** were searched for using the Web of Knowledge and also identified via earlier work within the NanoPUZZLES and NanoBRIDGES EU projects.
- The most recent articles were prioritised and articles being curated within other EU projects were not considered to avoid duplication of effort.
- Of the 264 references considered to date, 163 **accessible, primary experimental journal articles** (published 1996-2014) and one **scientific report** (NanoCare Project, 2009) were identified which assessed nanomaterials for at least one of the two endpoints noted above.
- Of the 164, most references only reported **studies for a few nanomaterials**, which might **limit their usefulness for supporting modelling**.



## The Need for Standardised Terms

- It is important to ensure data are recorded in a universally consistent fashion e.g. to avoid inconsistencies due to synonyms (left hand side image below).
- Therefore, data extracted from the literature are being identified using terms from **ontologies**, retrieved using BioPortal [1].
- Selecting appropriate terms remains a **challenge** due to **inconsistent definitions of terms** (right hand side image below) and **incompleteness** of existing ontologies.
- The ontologies and ontology terms being used within NanoPUZZLES are recorded using the data collection templates based upon ISA-TAB-Nano (see below).

**ChEBI Ontology**  
 zinc oxide  
 Preferred name: "zinc oxide"  
 hasRelatedSynonym:"ZnO, OZn..."

**BioAssay Ontology**  
 LC50  
 Definitions: "It is 50% lethal concentration (the concentration that kills 50% of treated cells)."  
 EC50  
 Definitions: "The effective concentration of an agonist, which produces 50% of the maximum possible response for that agonist."

Hu et al. 2009: "...nano-sized metal oxides, ZnO..."  
 Wang et al. 2014: "...zinc oxide (N14)..."

Hu et al. 2009: "LD50 (the dosage ... proved to be fatal to 50% of the bacterium)"  
 Puzyn et al. 2011: "EC50 - the effective concentration of a compound that brings about a 50% reduction in bacteria viability"

## Data Quality and Usefulness Criteria

- Earlier work by Lubinski et al. [2] suggested that at least 10 data points, from a single source, are required for (Q)SAR modelling: as the **literature survey (above) illustrates**, achieving this criterion in practice is a **challenge**.
- Additional quality and usefulness criteria are being developed.
- Quality scores for physicochemical/structural (meta)data**, similar to the Nanomaterial Registry “compliance levels” [3], have been developed with points assigned to items of (meta)data based on the following criteria.
  - “Completeness”. All else being equal, the quality of the (meta)data is higher if more items of *relevant* (meta)data are reported.
  - “Standardisation”. Where possible, terms should be taken from ontologies.
  - “Precision”. Nanomaterial structures can be specified to varying degrees of “precision”, e.g. “dextran coat” vs. “hydrocarbon coat”: more “precise” descriptions are more valuable.
  - “Quality assurance”. Those (meta)data which have been checked for errors can be considered of higher quality.

## Data Collection Templates Based on ISA-TAB-Nano

- Data collection templates were developed **based upon a proposal for a global standard for nanomaterial data exchange: ISA-TAB-Nano** [4,5].
- These templates were created in Excel and Python code is being developed to **allow the data to be uploaded to online databases** and read into modelling software e.g. KNIME ([www.knime.org](http://www.knime.org)).
- Columns/rows were pre-defined **based upon relevant items of (meta)data** capturing key physicochemical measurements, toxicological data and experimental conditions.

**Investigation file template ("i\_ref\_ID\_based.on.DOI.or.URL.xls")**

**Material file template ("m\_Description of the nanomaterial ...xls")**

**size determination by DLS Assay file template ("a\_Investigation Identifier\_PC\_size\_DLS.xls")**

**Physicochemical Study file template ("s\_Investigation Identifier\_PC.xls")**

## References

- [1] P. L. Whetzel et al. Nucleic Acids Res., vol. 39, no. Web Server issue, pp. W541–W545, 2011.
- [2] L. Lubinski et al. SAR QSAR Environ. Res., vol. 24, no. 12, pp. 995–1008, 2013.
- [3] M. L. Ostraat et al. Int. J. Nanomedicine, p. 7, 2013.
- [4] D. G. Thomas et al. BMC Biotechnol., vol. 13, no. 1, p. 2, 2013.
- [5] <https://wiki.nci.nih.gov/display/ICR/ISA-TAB-Nano>. [last accessed 9th April 2014].

## Conclusions

- Obtaining suitable data for modelling nanomaterials’ effects from the public literature requires a number of challenges to be addressed.
- Most publications only assess a small number of materials – which may restrict our ability to develop models.
- Standardised recording of data is being addressed.
- Careful attention needs to be given to which (meta)data need to be recorded and how to record these (meta)data to support database development and modelling.



## Acknowledgements



The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 309837 (NanoPUZZLES project).

The following researchers (LJMU) assisted with literature analysis: Fabian Steinmetz, Katarzyna Przybylak, Mark Hewitt, Mark Nelms, Claire Mellor.