EU US Roadmap Nanoinformatics 2030

Editors:
Andrea Haase, German Federal Institute for Risk Assessment, Germany
Fred Klaessig, Pennsylvania Bio Nano Systems, LLC, USA

Disclaimer

This roadmap has been jointly developed in trustful cooperation among scientists of the European Union, the United States of America and a few other countries. Scientists with different scientific backgrounds working in the field of nanotechnology have cooperated with the main objective to provide as broad an overview as possible about the young and rapidly evolving field of “nanoinformatics”. By no means was the intention to provide all possible details. Instead, interested readers will find plenty of additional references mentioned in each of the chapters that will provide more detailed information.

The opinions expressed in this document are those of the authors and do not necessarily represent the opinions of their respective organizations or the US Government. Mention of product names does not constitute endorsement.

The statements and opinions contained in the individual chapters are solely those of the individual authors and are not legally binding with respect to different regulatory frameworks. In particular it should be noted that some of the terms might be defined and used differently in the US versus the EU, also within different scientific disciplines and within different regulatory frameworks. Therefore, within the definitions sections we attempted to provide an overview, to explain the most important terms, and to highlight some that may have different meanings.

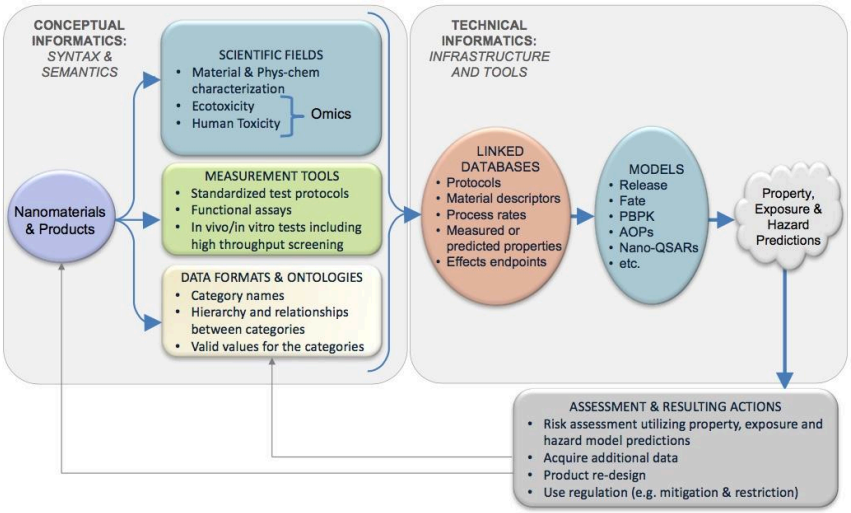
Table of Contents

1. Executive Summary

The Nanoinformatics 2030 Roadmap is a compilation of state-of-the-art commentaries from multiple interconnecting scientific fields combined with issues involving nanomaterial risk assessment and governance. As illustrated in Figure 1, the scientific fields represented include: materials science/physico-chemical characterization; ecotoxicity & human toxicity (including –omics); computational modeling; and informatics. Each has its own history, precepts, test methods, analytical tools, metadata forms, ontologies and criteria for interpreting experimental results. Additionally, each has its own research community. The Nanoinformatics Roadmap adds a separate consideration, namely, capturing the formal environment, health and safety (EHS) data requirements, e.g. good laboratory practice, related to regulatory assessments and governance. Coordination of future research effort and a shared vision, rather than programmatic direction, is the Roadmap's role.

These fields are in different stages of development and have contrasting levels of complexity in terms of information requirements, testing methods, terminology and protocols. Even the more established fields are re-examining testing protocols and accepted data formats to include factors affecting nanomaterials' transformations and the consequent dynamic nature of exposure and dose. Nevertheless, a shared informatics infrastructure can be identified. The technical data storage, data retrieval and theory development capacities required to support modeling functionalities for regulatory guidance can be pursued through a modular growth of the datasets, ontology and structure to meet goals such as a lessened reliance on the vagaries of whole animal testing. With this approach, the nanoEHS community can lay the foundation for an incremental growth building on the structure and ontology developed in earlier projects. Methods can be developed and applied to systematically engineer ontology development and the communication processes that can shepherd the interrelated fields to increasing maturity in terms of protocols, language, testing requirements and integrated data formats.

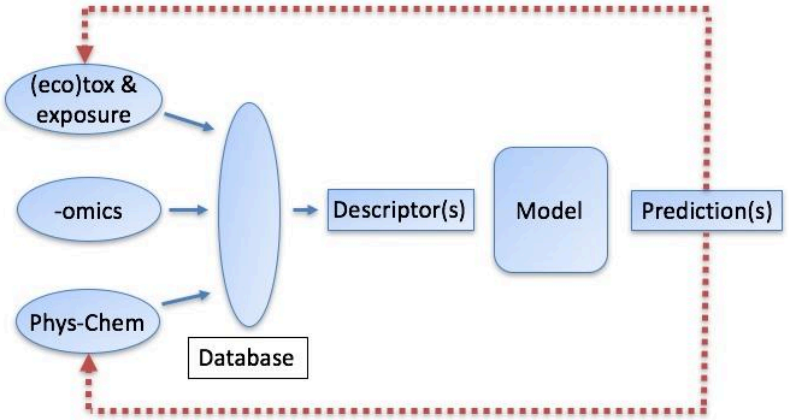
Figure 1: The Roadmap: from disparate fields to an integrated nanoinformatics infrastructure



While each scientific field has its own direction, (eco)toxicity with its important role in ensuring the responsible development of nanomaterials focuses the Roadmap on aligning progress among these fields with the criteria used by regulators for registering chemicals, pesticides and drugs. We recognize that not every adverse outcome pathway (AOP) pursued in the toxicological sciences will be one initiated by a nanomaterial, nor will every physico-chemical property that can be predicted through computer modelling influence toxicity. However, when they do align, there is an imperative that the results be useful to the regulator.

The Nanoinformatics 2030 Roadmap envisages a flow of data from several empirical fields into structured databases for eventual use by computational modelers in predicting property, exposure and hazard values that will support regulatory actions for a target nanomaterial. A very simplified data flow is illustrated in the figure below.

Figure 2: Simplified Data Flow proposed in the NanoInformatics 2030 Roadmap



It is our expectation that current interest in Integrated Approaches to Testing and Assessment (IATA) & Alternative Test Strategies that minimize whole animal testing and the simple desire to have a mechanistic understanding of nanomaterial (eco)toxicity will lead to greater reliance on computational modeling to predict properties for new materials or their toxicities. A growing knowledge commons supporting robust modeling capabilities, which predict properties, exposures and hazards of nanomaterials, would permit fuller implementation of the EU's strategy surrounding safe-by-design. Properties driving either desirable nanomaterial properties or undesirable EHS profiles could be explored as predicltions during early stage research, with later confirmation as a nanomaterial approaches commercialization.

Considering the range of likely readers, probably experts in one field interested in understanding developments in nanoinformatics, the authors have written the Sections to be read by a general audience. The reader can either start with their own field or with the milestones or with the Sections outlining the several nanoinformatics communities. In general, the Roadmap has three categories: an administrative grouping (Executive Summary; Definitions and Context; Objectives); a technically oriented informatics grouping (informatics, materials modeling, statistical computation, omics bioinformatics) and a community of research oriented grouping (stakeholders, database

projects, initiatives and milestones & pilot projects). Each Section is self-standing and, where appropriate, cross-cutting issues are identified.

The Roadmap's Sections do not follow either the complexity in Figure 1 or the simplified data flow in Figure 2. As a guide for the reader, we offer the following commentary connecting the several Sections relying primarily on Figure 2.

Empirical Fields:

- Toxicity (and ecotoxicity) is the subject of a separate Regulatory Research Roadmap. There is a short overview of biological testing from an informatics perspective in the Milestones (Section 12.2).
- The burgeoning field of omics is discussed in Section 8 with emphasis on transcriptomics, the most advanced facet from an informatics standpoint.
- Physico-chemical characterization is interspersed as property representation (Section 5.2) and descriptors (Sections 6.2 and 7.2). As with toxicity, there is a short overview from an informatics perspective in the Milestones (Section 12.3).

Database:

- Informatics involves structured datasets, where the structure is found in the vocabulary used, i.e., a controlled vocabulary, and in the relationships among terms, which is the ontology (Section 5.9). Essentially, the database curator annotates experimental data to maximize its utility beyond that of the original field. In effect, the curator deconstructs the original experiment into components that reflect a physico-chemical understanding of nanomaterials to supplement the biological understanding found in bioinformatics ontologies.
- From a strict data flow standpoint: data collection (Section 5.6) leads to material representation (Section 5.1) and property representation (Section 5.2) that are curated (Section 5.5) using metadata (Section 5.8) so that data can be retrieved (Section 5.7) and exchanged (Section 5.10).
- It is unlikely that there will be one authoritative database, which has led to the development of data transfer formats such as ISA-TAB-nano (or upgrades to ISA-JSON) for exchanging data with other databases or modeling programs (Section 5.10.1). The reasons for multiple databases are many including issues of unpublished data, different foci, proprietary data or even the mundane issue of resources for database maintenance (Section 5.3 and 5.13). In the Roadmap, there is a preference for using extensions compatible with the publicly available ISA standard used in bioinformatics.

Computational Modeling:

- Where informatics deconstructs the nanomaterial and properties, computational modeling re-constructs using those parameters (descriptors, Sections 6.2 and 7.2) viewed as most applicable to the property being predicted. The descriptors may be properties measured for related materials (grouping), or may be concepts found in theories or may be a hypothesis underlying a database query.

- Collecting curated data (Section 5.10) of sufficient extent (size of dataset; replicates; dose-response) has led to several data-filling approaches (Section 6.4) that in turn rely on nanomaterial grouping (Section 6.3).
- Inherent to computational modeling is relating the material description and base physico-chemical properties to the biological outcomes, especially if some descriptors are not readily measurable. This challenge leads to several approaches to deciding on descriptors: in material representation (Section 5.2), in selecting among primarily measured properties (Section 6.2) and use in statistical models to predict properties, QSPR, or biological activity, QSAR, (Section 6.4); in calculating descriptors otherwise difficult to measure from theory and models (Section 7.2) before coupling to biological events (Section 7.5).
- There is of course a need to validate model predictions, which can be done by splitting datasets into training and validation sub-sets for internal consistency or by measuring properties of material libraries known to modify a target property. A modeling overview is given in the Milestones (Section 12.3).

Validation:

- Validation is a critical step if computational model predictions are to find use with regulators, especially for data-filling, the step in pursuing a regulatory action where it is proposed that a prediction is sufficiently valid that making a measurement is not considered necessary.
- The validation requirements are presently unclear, but we can expect that they will be more rigorous for predicting biological outcomes than for nanomaterial properties that have little relevance to toxicity. In toxicity, the mechanism can be termed a mode of action or an adverse outcome pathway (AOP), which is the subject of the Regulatory Research Roadmap and given an overview from an informatics perspective in the Milestones (Section 12.2)
- In all cases, regulators will require that there be a proven relationship among the computational model's algorithm, its domain of applicability (grouping, Section 6.3) and the mechanism underlying the effect induced by the specific property. However, the nature of regulatory requirements will emerge and be communicated through feedback from data-filling exercises (Section 6.4).

NanoInformatics Community

While there has been funding for data management on an individual project basis, the use of this information in a regulatory context has been a challenge. In general, nanoinformatics has relied on communities of research, such as those outlined in Section 10. The Roadmap itself is an example of one such community of research. Though initiated in Europe, the Roadmap expands on an earlier U.S. document. The milestones are based on the results of several international workshops whose lead authors were approached during the review process (Section 4). Throughout the process, issues and draft Sections were discussed at European (WG4) and U.S. (nanoWG) teleconferences whose participants have met regularly for several years on nanoinformatics. Colleagues from Canada and Australia participated, as well as those active in ASTM International's E56 and ISO's TC-229. In addition, the EU-US nanoEHS 2016 and 2017 meetings were used for face-to-face discussions.

There are also broader issues that cannot be covered fully in this document. The differing perspectives among stakeholders (Section 9) require a separate examination.

2. Definitions in an Operational Context

Nanotechnology covers a broad array of scientific disciplines, each with a specialized language and at times utilizing different definitions of terms like nanoscale, nanomaterial, etc. Informatics, on the other hand, involves the application of external organizing principles onto the data generated within a scientific discipline. In such situations of countervailing interests, it becomes difficult to offer a coherent glossary of terms and definitions. For the purposes of this Roadmap, and recognizing that readers might appreciate some explanation for those themes beyond their expertise, we instead offer a descriptive overview illustrating their use, i.e. operational definitions.

Nanotechnology is most generally described as the application of scientific knowledge to manipulate and control matter predominantly at the nanoscale, which explains the broad array of stakeholders involved (see Stakeholders in Section 9) when one considers the issues raised when commercializing the resulting products.

Informatics is the application of information and computer science methods for collecting, analyzing, and applying data in a scientific field, e.g. bioinformatics. Thus, nanoinformatics is a systematic methodology to collect, organize, validate, store, share, model, analyze, and apply data involving nanotechnology processes, materials, properties and commercial product implications; to confirm that appropriate decisions were made and that desired outcomes were achieved from the application of the data; and finally to convey experience to the broader community, contribute to generalized knowledge, and update standards and training. The inclusion of the latter point of product commercialization expands the stakeholders to include regulators and the general public interested in nanomaterial environmental, health and safety (nanoEHS), as well as in responsible research and innovation.

The Roadmap combines several aspects of nanoinformatics in a manner that provides operational definitions for a number of concepts (underlined):

1) Data from credible sources are being compiled into structured, electronic datasets, where the data may be publicly available (published) or not (unpublished laboratory data), may be from formal regulatory submissions on specific materials (confidential business information) and may be numerical or pictorial. We anticipate that there will be multiple databases (structured, electronic datasets) administered independently, but with some level of interoperability established.

2) A 'structured, electronic dataset' means that the database can be used to retrieve the original data. The term 'structured' refers to the use of controlled vocabularies, metadata, and ontologies during data entry in order to ensure reasonable recall and precision in collocating findings from related studies. We anticipate that there is a role for data curation in annotating metadata and commenting on data completeness (see

Section 5), and some standardization within the nanoinformatics field will be necessary if data are to be exchanged between databases.

3) Computational techniques for analysis, modelling and theory development also impose issues of standardization in terms of data quantity, robustness, completeness and validity. These issues may differ across stakeholder interests, where the metadata for theory development may be less restrictive, when remaining within a single scientific discipline. Metadata requirements for regulatory purposes may cross disciplines and emphasize following proper test protocols, even where these are not yet validated for use with nanomaterials. We view cross-disciplinary awareness and coordination of these issues as a central impetus to the Roadmap as they will continue to undergo development and refinement throughout the 2030 time frame (Milestones, Section 12).

4) The size of currently available datasets is a particular challenge for computational modelling, raising as it does, issues of database access, data completeness among independent studies, and even model validation. Relative to 'big data' topics, the number of independent studies, the range of nanomaterials studied and the robustness of test protocols are more limited (see Sections 6, 7 & 8)). We anticipate that these fields will advance independently with regulatory validation & acceptance first occurring during data-filling and grouping exercises, the preparation of registration dossiers, and the testing programs under the appropriate regulatory frameworks (e.g. REACH, BPR, Regulation EC 1107/2009, US-EPA etc.) (see Sections 6, 7 and 9).

5) Computational techniques for modeling and theory development eventually lead to predictive capabilities based on descriptive elements (descriptors, Sections 6 & 7) that are the data already present in the 'structured' dataset or are the application of innovative concepts (theory, metadata, mathematical expressions) that are validated by the data already present in the 'structured' dataset. We have provided one physical model of a nanomaterial (Milestones, Section 12.3) to serve as a shared basis for data models incorporated into database ontologies or found as boundary conditions in simulations or computational models.

Provided below are a number of terms with 'operational' definitions that should aid the reader when navigating this Roadmap. It should be emphasized that there are many sources for terms (e.g. ISO, ASTM, published literature) and particular care should be taken when using these terms in a legal or regulatory context. One example of a legal difference between the European Union and the United States is provided for 'chemical substance' in Section 5.

Term	Operational Definition	Roadmap Section
Controlled vocabulary	Standardized list of unique terms and their definitions used to index, annotate, enter and retrieve information.	5
Data Curation	the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time	5
Data Filling	In a regulatory setting, computational methods for estimating a parameter's value for a test material using a base set of known (and	6,7,12

	related) materials and values; implementation requires clear definition of the applicable domain.	
Database	Structured electronic dataset	2,5
Descriptor	Parameters with measured, theoretically or computationally derived values representing the intrinsic (independent of external conditions) or extrinsic (dependent on external conditions) properties of a defined, targeted system and that are also sufficient, mechanistically plausible, relevant and non-redundant for use in a computational model.	6 & 7
Informatics	The application of information and computer science methods for collecting, analyzing, and applying data in a scientific field	All Sections
Metadata	.data describing the content (including indexing terms for retrieval), context and structure of electronic document-based information and their management over time (ISO/TR 18492:2005, term 3.8)	5.8
Nanotechnology	The application of scientific knowledge to manipulate and control matter predominantly at the nanoscale	All Sections
Ontology	Controlled vocabulary extended to include the relationships among terms for the purpose of analysis, computational modeling and theory development	5
Physical Model	Representation of the physical entity that is the basis for a data model, controlled vocabulary and ontology	12
Recall and Precision	The ability to collocate related database entries (recall) that are specific to a query (precision).	2
Structure	source of spatially resolved properties reflecting the relationships among and the manner of arrangement of a complex entity's components	2,5,6 & 7

3. Objectives

Nanotechnology is one of the key technologies of the 21st century. The global nanotechnology market already had a value of $39.2 billion in 2016 and is expected to reach $90.5 billion by 2021 (McWilliams 2016). In addition, public funding sources invested more than $67.5 billion globally during the last decade for research and development (Scientifica Global Funding Report 2015). Nanotechnology is already used for many different applications and the global market is increasing steadily each year. Due to significant funding from both public and private sources, knowledge has increased significantly during the last decades. Several large collaborative projects investigating the environmental and health safety aspects of nanomaterials (nanoEHS) have been finished already, with several more ongoing or starting in early 2018. In addition, there are general toxicology advancements including high throughput and high content methods, which may provide plenty of data within a short term period. Therefore, as also observed in many other scientific disciplines, the amount of recorded data has increased drastically in the last years. Nanotechnology requires integration of knowledge from quite different disciplines such as material science, biology, chemistry, toxicology, medicine, and computational & decision sciences. In parallel, computational approaches are gaining increasing importance and popularity. Therefore, the advancement of nanoinformatics will be crucial for the development and application of sustainable nanotechnology. Within this roadmap the term nanoinformatics will be used for the acquisition, storage and analysis of any data relevant to nanotechnology including the development and the use of specific computational models and decision support systems. The main focus of this roadmap is on environmental and health safety aspects of nanomaterials (nanoEHS).

This roadmap aims to address the following objectives:

Objective 1: Foster community interactions and provide support for different stakeholders

NanoEHS integrates knowledge from many different disciplines (e.g. material scientists, biologists, chemists, toxicologists, risk assessors, computational experts etc.). Different stakeholders (i.e. industry, academia, regulatory agencies, the standardisation community and the civil society) are involved. Each is generating different types of data and each has its own objectives and needs with respect to storage and use of the data.

This roadmap should foster the “self-assembly” of this very heterogeneous community such that different stakeholders get to know each other and become aware of the specific needs and objectives of other stakeholders. In addition, this document provides an overview of the nanoinformatics processes and tools available support different stakeholders in achieving their specific objectives. Therefore, the roadmap will clearly describe the benefits of nanoinformatics at different phases of work within the context of nanoEHS for different stakeholder needs.

Objective 2: Promote capture, preservation and dissemination of all publically-available NM measurement data

A considerable investment has already been made from public as well as industrial sources into nanotechnology development in general but also into nanoEHS specifically. Future resources are limited. Thus, there is a need to make the maximum possible use of existing data, to avoid duplication of work and re-measurements but also to plan new research accordingly to plug gaps in the existing datasets. This both requires and promotes consistency in reporting results. It also ensures that results are secured and data can be assessed later by others. Therefore, knowledge can be increased simply by more detailed data analyses or by meta-analyses, which will be facilitated by an increasing number of *in silico* methods or other methods not yet developed.

This roadmap supports the creation and linkage of repositories to ensure that all publically- funded nanomaterial measurement and modeling results are deposited in accessible repositories, so that they can feed with data the evolving infrastructure of risk assessment and management decision support tools (e.g. SUNDS, caLIBRAte System of Systems). Specifically, it aims to raise public awareness of the benefits of this and embed data-sharing principles and mindsets into all levels of the research community. It describes a step-by-step process to achieve this overarching goal and it explains what kind of infrastructure is needed for this purpose.

Objective 3: Facilitate the (re-)use of existing data

This objective will advance nanotechnology and expedite its commercialization. To pursue optimal data usage, a system should consider **FAIR** data principles and guidelines, based on **F**indability, **A**ccessibility, **I**nteroperability and **R**eusability of data and the algorithms, tools and workflows that operate on it [3].

Encouraging the scientific community and stakeholders to make use of existing data will facilitate:

- a (better) understanding of experimental results through integration of currently disparate datasets;
- the development of different kinds and complexities of models and their validation using existing datasets;
- the prediction of properties and performance/ functionality of nanomaterials;
- the correlation of specific effects with nanomaterial physico-chemical characteristics; and,
- grouping and read-across among nanoforms and bulk analogues and the implementation of Intelligent Testing Strategies for more cost-efficient risk assessment and Safe(r)-by-Design practices.
- the direct use of existing data to fulfill data gaps for risk assessment and regulatory obligations
- encourage consortium and nanomaterial information exchange between interested industry partners reducing cost and animal testing
- capture the breadth and extent of nanomaterial use
- development of appropriate EHS controls.

This enhanced knowledge will support:

- the design of new nanomaterials;
- the establishment of Safe(r)-by-Design Principles;
- decision making regarding the risks of nano-enabled products and processes;
- regulation.

Objective 4: Identify specific milestones and pilot projects in relation to objectives 1-3

This roadmap will identify and describe the key challenges for nanoinformatics covering data storage, data use, dissemination and exploitation for safety assessments and risk management decision making.

It will also identify and describe specific pilot projects covering short (i.e. within the next 3-5 years), medium (i.e. within the next 5-10 years) and long-term (> 10 years period) needs as key stepping stones / demonstrators to reach the three described objectives.

4. Introduction

This roadmap is a timely continuation of several previous efforts, namely of three workshops, a few workshop reports, and the US Nanoinformatics 2020 Roadmap. As this roadmap builds and extends those, they should be briefly mentioned here.

The **Nanoinformatics 2020 Roadmap** (4) was based on a **2010 workshop** involving ~ 73 participants, mainly from USA with some representatives of the EU's Action Grid effort (5). The following topics were discussed during this workshop and accordingly described in the roadmap. Many of them remain pertinent:

1. Data collection and curation needs:
   - Minimal information standards for nano-data sets (completeness & quality);
   - Inter-laboratory studies (ILS) for test protocol and data completeness validation;
   - Standardized characterization is needed community-wide; and
   - How much information is needed to trigger a "recognized hazard"?
2. Tools and methods for data innovation, analysis and simulation needs:
   - A complete map of data collection and curation workflows to guide the development of nanoinformatics;
   - A mechanism for federated searches to utilize existing nanotech databases;
   - Getting the science right; and
   - Getting the right data.
3. Tools, training, and education perspectives:
   - Data Accessibility and information sharing;
   - Context is critical for effective information sharing; and
   - Competing socio-cultural incentives impact data sharing.

The Nanoinformatics 2020 Roadmap listed available resources at that time and also proposed several pilot projects.

In **2011**, **COST** (European Cooperation in Science and Technology) sponsored a **workshop in Maastricht** with ~90 attendees on the use of QSAR methods to model biological effects of nanomaterials (www.cost.eu/events/qntr). The resulting paper by Winkler et al. (6) proposed 14 milestones grouped in 2-year, 5-year and 10-year time horizons. For the most part, the milestones reflected:

- a need to generate sufficient data for model development;
- an acceptance of 'surrogate' assays useful for modeling if not for regulation;
- an expectation that understanding protein corona formation would provide the necessary mechanistic information; and
- a view of informatics as a needed infrastructure for data accessibility.

The roadmap also benefited from Prof. Winkler's more recent commentary (7). While progress was noted, especially the availability of benchmark test materials, there remain insufficient data resulting in a need for surrogate or fast screens, for improved nano-specific descriptors and for an exploration of chemical grouping. The update gave greater stress to data curation, informatics, data consolidation and standardized testing.

In **2014**, the **U.S. National Science Foundation** (US NSF) funded a workshop held prior to the Sustainable Nanotechnology Organization meeting in Boston on the general theme of defining the fundamental science needed to support nanoEHS. The resulting paper by Grassian et al. (8) identified mechanistic data gaps that when resolved would enable a predictive biological response capability.

In **2015**, the **first European Modelling Conference, CompNanoTox,** took place in Benahavis, Spain. This conference was jointly organized by all European modelling and database projects funded at that time (i.e. NanoPUZZLES, ModENPTox, PreNanoTox, MembraneNanoPart, MODERN, eNanoMapper) together with the EU COST action TD1204 MODENA. The resulting paper by Banares et al (9) described the most important current challenges with respect to nanomaterials modelling. This paper described for instance shortcomings with respect to material characterization, a lack of suitable, validated toxicity assays and a lack of mechanistic understanding of nanomaterial toxicity.

This roadmap builds on these documents. In chapters 5, 6, 7 and 8 the state of the art and the current challenges with respect to data collection and data curation (Section 5), nanochemoinformatics modelling (Section 6), materials modelling (Section 7) and nanobioinformatics (Section 8) are described. This is followed by a description of the "nanoinformatics community and stakeholders", the currently ongoing nanoinformatics activities, available databases, interesting projects and integrating activities etc. (Sections 9 to 11). This leads into Section 12 describing suggested milestones and several useful pilot projects grouped according to their time-horizon as short-term, mid-term or long-term projects, which are listed and described from several perspectives, i.e. the perspective of material characterization, the perspective of toxicologists, of modellers and regulators.

5. Data collection and curation

Nina Jeliazkova[1], Christine Ogilvie Hendren[2], Danail Hristozov[3], Lucian Farcal[4], Nikolay Kochev[1,5], Philip Doganis[6], Peter Ritchie[7], Barry Hardy[4], Frederick Klaessig[8], Egon Willighagen[9]

[1] Ideaconsult Ltd, Sofia, Bulgaria
[2] Center for the Environmental Implications of NanoTechnology (CEINT), Duke University, Durham, NC, USA
[3] Greendecision Srl, Italy
[4] Douglas Connect GmbH, Basel, Switzerland
[5] Department of Analytical Chemistry and Computer Chemistry, University of Plovdiv, Plovdiv, Bulgaria
[6] National Technical University of Athens, Greece
[7] Institute of Occupational Medicine, Edinburgh, UK
[8] Pennsylvania Bio Nano Systems, LLC, USA
[9] Department of Bioinformatics, NUTRIM, Maastricht University, NL

A major challenge for the nanoEHS community is the establishment of common languages, standards and harmonized infrastructures with applicability to the needs of the different stakeholders. The complexity of nanomaterials, and their physicochemical properties and interactions with biological and environmental systems, leads to increased uncertainty in the applicability of experimental data for regulatory purposes. Thus, recent community efforts have focused on building databases that support computational modeling and decision frameworks for nanomaterial environmental health and safety (nanoEHS) assessment and risk management. Those based on open standards, open source, common languages, and that have an interoperable design are desirable.

Another major challenge for the nanoEHS community is linked to data quality and data curation. The nanomaterial data curation topic has been the focus of multiple collaborative efforts and publications [cite doi:10.3762/bjnano.6.179 and follow-ups,]. Specific recommendations regarding terminology, (meta)data requirements, computational tools, and recommendations regarding the role of organizations and scientific communities have been published [10.1039/c5nr08944a]. The terminology recommendation includes defining community agreed data completeness and quality criteria . One of the key findings is that the data completeness and quality will depend on specific user or stakeholder needs. Hence it is critical to identify the relevant scientific, regulatory, societal and industrial use cases. Building and adopting common vocabularies or ontologies address the provenance metadata requirements to represent materials and studies, manufacturer supplied identifiers, composition, impurities, as well as experimental protocols, experimental errors, etc. As investigators will vary in their knowledge of informatics, it is desirable to have standardized templates for data entry based on minimum information checklists and ISA-TAB [ref] and ISA-TAB-Nano specifications [ref]. However, user-friendly templates for data logging captures only one data source, a specific laboratory, when there are also other data sources such as journal articles, proprietary studies, or independently maintained databases. While challenges for nanomaterial data curation workflows are extensively described in [10.3762/bjnano.6.189], the broader experience of extracting and compiling literature

data, leads to another recognized task of integration of, and exchange between, existing structured databases.

Nanomaterial entries (information) are found not only in dedicated nanomaterial databases, but also generic chemical, toxicology and toxicogenomics databases as well as regulatory, databases like those supporting REACH dossiers, molecular modelling and image processing tools [10.5281/zenodo.375637].

To summarize, unstructured nano-related data are relatively abundant, and rapidly generated, but also quite dispersed across many different sources. Combining data from various sources is hampered by the lack of programmatic access and the absence (or infrequent use) of a common representation of nanomaterials and related experimental data. It has to be noted that while common vocabularies are being developed, the nanoinformatics community has not yet arrived at a commonly agreed "conceptual schema", or agreed on how to represent the common concepts of the domain and their relationships.

5.1 Challenges: Material representation

The representation, processing, and communication of information about objects are at the core of any information system and informatics in general. The representation of chemical and biological objects is fundamental for the interdisciplinary field of bioinformatics. Chemoinformatics is a well-established field which supplies tools for representing, processing and solving problems with chemical molecules in general. The term nanoinformatics was introduced to delineate the activities specific to managing and processing information about nanomaterials. An adequate computer representation of the objects is required in order to handle biological, chemical, or nanomaterial information, and to enable the building of information systems. There are also literally thousands of different descriptor that can be measured or calculated, but only a subset are likely relevant to a specific EHS aspect or application. Descriptors encompass physical and chemical identity (size, shape, chemical composition, particle architecture) associated with material representation, intrinsic properties and extrinsic properties (Sections 6.2, 7.2.1, 7.2.2).

For cheminformatics (Section 6), the central object is the chemical structure, following the origin of the "chemoinformatics" in the context of drug design. There are several levels of chemical structure representations, which reflect different chemistry models or theories. For example, graph theoretical approaches (e.g. constitutional, topological, 3D, conformational representation) are not easily combined with quantum chemical approaches (Section 7) [REF]. The structure formalization is the starting point for all other activities and is reductionistic by its nature because only particular aspects of the chemical reality are formalized. The most popular method of representing chemical structures is the chemical graph, which is the basis of representing structures by connection tables, linear notations as SMILES and InChI, de-facto standard chemical formats such as SDF. Even those chemical databases using the same chemical graph concepts may differ in database technology and physical database schema. Unfortunately, the graph theoretic representation of well defined chemical structures is

ill-suited as a single representation of nanomaterials: it is not able to distinguish all aspects of the NM structure, also partly because that structure may not always be known. As a result, it is difficult to distinguish between properties of a nanoscale and bulk material with the same chemical structure. The quantum chemistry formalisms are also able to capture aspects of the nanomaterials and are used to study material functionality and structure (*cite EMMC, nano-hub, commercial tools, the modelling section below*), but may also suffer from a lack of knowledge about the structure. Relating nanomaterial identity, characterisation and biological properties often requires less detailed representation than the quantum chemistry level, and there are several parallel attempts in this direction.

There is a *need for an agreed conceptual representation of a (nano)material* compatible with the emerging regulatory consensus that nanomaterials are to be handled as an extension of chemical substances.[1]. The REACH definition of a substance encompasses all forms of substances and materials on the market, including nanomaterials. A substance may have complex composition. The definitions of the terms "substance" and "material" are discussed in [2], comparing ISO, EU REACH and general scientific definitions of the terms. *Note: The reader is reminded that terms may have different definitions in other jurisdictions. In the United States, molecular identity defines a chemical substance for TSCA, while for REACH in Europe, impurities and residual catalysts are included.*

The Nano Particle Ontology (NPO) defines a nanomaterial (NPO_199) as equivalent to a chemical substance (NPO_1973 or CHEBI_59999) that has as a constituent a nano-object, nanoparticle, engineered nanomaterial, nanostructured material, or nanoparticle formulation. The OECD Harmonized Templates represent nanomaterials as substances, consisting of components, additive and impurities, and the recent IUCLID6 implementation extends the representation to handle nanoforms. Describing the ENM composition requires description of many components (also termed constituents) and the complex relations between components. For example a nanomaterial may consist of core and one or more layers (shells, coatings) around the core.

Nanomaterial representations (descriptions or identities) may differ across databases. For example, the NECID database defines the material by its core only for the purpose of handling exposure scenarios, while the CEINT database introduces an additional concept of "instance" meaning the point in time when the NM transits to the next life cycle stage and warrants measurement of its chemical or biological properties as well as those of the system. The "instance" is considered critical by the CEINT group in order to allow investigation of the dynamic nature of nanomaterials including the transformations and kinetic processes that have been proven to significantly affect their fate and effects. NanoMILE took a similar approach, linking "aged" nanomaterial properties to the initial pristine properties, and compared the toxicity of the both. NanoFASE is building on the NanoMILE and CEINT approaches, such that the characteristics of nanomaterials after "reaction" in different environmental compartments (soil, water, sediment, wastewater treatment or uptake and excretion by organisms are all considered as different instances,

[1] https://euon.echa.europa.eu/nanomaterials-are-chemical-substances

[2] Roebben, G.; Rasmussen, K.; Kestens, V.; Linsinger, T. P. J.; Rauscher, H.; Emons, H.; Stamm, H. *J. Nanopart. Res.* 2013, *15,* 1455. doi:10.1007/s11051-013-1455-2

unless experimentally (and in due course predicted) to be identical to the outcome from the previous compartment.

The basis of many chemical databases is the direct link between the chemical structure (as chemical composition) and properties, which is well aligned to supporting modelling. However, the concept of assigning measured properties to chemical structures is yet another approximation, not directly applicable to material data representation. Instead, measured properties have to be assigned to legally-defined 'chemical substances' (ENM as a subclass of substances), in line with the IUPAC definition. This approach is also applicable where information on chemical substances as produced by industry is required. Flexibility with respect to cases where the measured property is a property not of the entire material, but only one of its components (e.g. surface layer composition) is also relevant.

5.2 Challenges: Property representation

Besides the materials themselves, a nanoinformatics data curation framework must capture the physical and chemical attributes of NMs, including the notions of mixtures, particle size distribution, differences in amount of surface modification, manufacturing conditions, and batch effects. It must also capture the potential for evolution of many of these properties, such as changes in surface speciation, loss of coating, acquisition of an environmental or biological corona, and so forth, once the nanomaterial is embedded into a product, is released into the environment or comes into contact with biological organisms. Finally, the biological attributes (e.g. toxicity pathways, effects of ENM coronas, modes-of-action), interactions (cell lines, assays), and a wide variety of measurement approaches. A number of analytic techniques have been adopted and developed to characterize nanomaterials physicochemical properties. The selected pilot project on dissolution illustrates the complexity of just one type of measurement. With expanding insight into the factors determining toxicity, this list of properties is growing. *In vitro* characterization includes many endpoints for hazard identification. High throughput cellular assays and omics data & kinetics are becoming increasingly important in nanomaterial assessment. A common requirement for all types of users is to link the nanomaterial entries to those studies in which toxicology or biological interference of the nanomaterial has been studied, in addition to an accurate physicochemical characterisation. Thus, the properties and their representation should remain consistent with the descriptors used by ECHA (2017) and EPA (2017) for "nanoforms" and "nanoscale forms", respectively, but with more detail.

Supporting such heterogeneous datasets is a significant challenge; however, it is not unique to nanoinformatics. The potential solution is to organize the experimental data around the fundamental concepts of "test" and "measurement" [3]. There is evidence of database developers adopting this approach, although the very terms of "test", "assay", "experiment", "endpoint" are often used inconsistently across different players. The

[3] Roebben, G.; Rasmussen, K.; Kestens, V.; Linsinger, T. P. J.; Rauscher, H.; Emons, H.; Stamm, H. J. Nanopart. Res. 2013, 15, 1455. doi:10.1007/s11051-013-1455-2

OECD guideline defines the “test” or “test method” as the experimental system used to obtain the information about a substance. The term “assay” is considered a synonym. The term “testing” is defined as applying the test method. The endpoints recommended for testing of nanomaterials by the OECD Working Party on Manufactured Nanomaterials (OECD WPMN) use the terms and categories from the OECD Harmonized Templates. The NPO distinguishes between the endpoint of measurement (e.g., particle size, NPO_1694) and the assay used to measure the endpoint (e.g., size assay, NPO_1912), where the details of the assay can be further specified (e.g., uses technique electron microscopy, NPO_1428). This structure is generally the same as the one supported by the OHT (e.g., in the OHT granulometry type of experiment several size-related endpoints can be defined, as well as the equipment used, the protocol and specific conditions). The CODATA UDS requires specification of how each particular property is measured. ISA-Tab-Nano also allows for defining the qualities measured and detailed protocol conditions and instruments. The level of detail in the OHT, CODATA UDS, ISA-Tab-Nano and available ontologies differ, which is due to their different focus.

Examples

- zeta potential - entries for zeta p. property (NPO_1302), measured property (ENM_0000092), calculated property (ENM_8000111).
- materials - is material with the old NM-100 (ENM_9000201) and new JRC code JRCNM01000a (ENM_9000074) the same entity or not (not in the eNanoMapper ontology, per JRC advise)
- same term used in two (or more) ontologies in different context (example: biological process)
- how to describe COMET assay (OBI_0302736) and COMET FPG assay - same protocol, or different protocol with FPG= yes/no ? Or with a protocol parameter “enzyme=FPG” or enzyme=”None”
- is TEM a protocol, experiment, or measurement instrument ?
- Ontology annotation of specifically treated cells (e.g. THP-1 cells with macrophage properties). If the cell is annotated with THP-1 and the induced cellular change is only described in the protocol, the subsequent data analysis should take into account the protocol details as well.
- how to define “dispersion agent”
- how is “toxicological endpoint” defined and how is it linked or not linked with specific assays
- Are new classes/definitions required for chemical composition (or about discrepancies between ontology concepts)

[TBD] exposure databases specifics

Product databases have been reviewed[4] in a study identifying three databases with nano-specific products[5] and two general products databases. This review presents a methodology for identifying consumer products that contain nanomaterials, proposes a data model, and has developed and populated a database, containing 200 products. Assigning products to categories, as well as identifying where and what amount of nanomaterials are used in particular product is a genuine challenge: for instance, the sample preparation may change the particle size distribution, and therefore most product databases include products based on labels "nano" used by the manufacturer, rather than any analytical evidence.

[TBD] product databases - may need an update the sources cited below are not very recent.

5.3 Challenges: Data management plans

Research Data Management Plans (RDM and DMPs) are common act, but vary greatly in content. There is an increasing level of guidance, e.g. the ELIXIR-NL overview: https://www.dtls.nl/research-data-management/data-management-knowledge-tools/.
Having a project-level DMP matters as too frequently issues of data sharing come late in the project, slowing down project completion and limiting knowledge sharing. Data management is a cornerstone of collaboration: how, when, with what frequency, in what format are data archived and exchanged, and how, when, with what frequency data curation is done. The growing interest in DMPs has resulted in many suggested tools (see the aforementioned list) and literature, such as several articles in the "Ten Simple Rules" series about cultivating collaboration [REF,REF], creating DMPs [REF], and care of data [REF]. The above initiatives should serve to strengthen the efficiency with which data is archived and retrieved for research purposes and ensure that everyone that uses well annotated and coordinated archived data can collaborate efficiently.

Besides interactive access and archiving, data curation has received considerable attention [REF,REF]. A group of scientists from the US and the EU wrote a series of articles on the topic [REF], for example, dealing with how data completeness and quality could be estimated [REF], and the interoperability of the data (manuscript in preparation).
Given the importance of DMP for collaboration within a project consortium and after the project, it is surprising that these plans are not consistently peer-reviewed. Second, wider acceptance would be achieved if the DMP were an activity and not a deliverable: not just is the DMP an active document, it also needs auditing during the project and not

[4] S. Wijnhoven and M. Bakker, "Development of an inventory for consumer products containing nanomaterials," Centre for Substances and Integrated Risk Assessment, Tech. Rep., equally 2010. [Online]. Available: http: //ec.europa.eu/environment/chemicals/nanotech/pdf/study inventory.pdf

[5] Woodrow Wilson database, "Consumer Products An inventory of nanotechnology-based consumer products currently on the market."2011. [Online]. Available: http://www.nanotechproject.org/inventories/ consumer

left to the project end. Peer review could focus on ensuring these features, in addition to the proposed methods for data management.

5.4 Data Curation

Data curation, as defined in Section 2 (ref#), encompasses all of the activities that are necessary throughout the process of extracting, organizing, and entering data and knowledge into discretized formats within digital resources, and is central to the process of enabling data integration regardless of the size, scope or purpose of a given project/tool. Various aspects of curation, including its centrality to nanoinformatics, workflow, and data completeness and quality, have been addressed in a series of papers called the Nanomaterial Data Curation Initiative (NDCI), developed through the US National Cancer Informatics Program's Nanotechnology Working Group (NCIP NanoWG)- (Hendren et al. 2015 http://www.beilstein-journals.org/bjnano/articles/6/179, Powers et al. 2015 (DOI 10.3762/bjnano.6.189), Marchese-Robinson et al. 2016 (DOI 10.1039/C5NR08944A)).

5.4.1 Data Quality and Completeness

Based on a survey of 24 nanoinformatics resource representatives and the subsequent development of broad and flexible definitions for both data quality and completeness, Marchese-Robinson *et al.* report that these concepts are best understood in terms of their fit for a given purpose (DOI 10.1039/C5NR08944A).

Data quality may be considered of the potential correctness and trustworthiness of datasets, though there are a wide variety of metrics by which these attributes may be measured, including reproducibility, precision, uncertainty. Of critical importance is that due to the pivotal role curation plays in integrating data, "data quality" can be affected by compliance anywhere across the knowledge life cycle from initial experimental design and execution through transcription from a publication or database into the target resource.

The completeness of data and associated metadata may be considered to include the extent of nanomaterial characterization along with surrounding media and experimental conditions to support specific post-analyses, or relative to conforming to a minimal information checklist. Data driven modelling methods function best with large, diverse data sets with good property coverage, chemical diversity. There is a strong need for a systematic approach to generating data for nano-bio interactions as advocated by Bai et al. recently (cite Bai, X.; Liu, F.; Liu, Y.; Li, C.; Wang, S.; Zhou, H.; Wang, W.; Zhu, H.; Winkler, D.A; Yan, B. Toward A Systematic Exploration of Nano-Bio Interactions, Toxicol. Appl. Pharmacol, 2017; 323, 66–73.)

Because these concepts continue to evolve and will inherently vary by the purpose and scope of a given resource, data completeness and quality aspects of pilot projects are best conveyed by explanations of the processes, both technological and workflow related, that are in place to address these issues and to ensure consistency.

5.4.2 Data Curation Process

The process of curating data is currently highly resource intensive in terms of management, workflow, sourcing and ontology. As standards for ontology and minimal information requirements may be developed over time, curation processes and tools may accordingly converge. However, in the meantime this process should be defined for each resource to understand the implications on data sourcing, extraction, quality, completeness, and utility for purpose. (doi:10.3762/bjnano.6.189)

5.5.1.1. OECD Harmonized Templates

The OECD Harmonized Templates (OHTs) are structured (XML) data formats for reporting summary data on safety-related studies on chemical substances. The OHTs and the supporting IT tool (IUCLID6, iuclid.eu) are used for preparing substance dossiers for REACH and for other regulatory frameworks operating in Europe; The substance identification section is compliant to "ECHA guidance for i

5.5 Getting data in - data sources and data entry

It is important to understand the variety of data sources (e.g. literature, intermediate laboratory formats, or raw data), the criteria for inclusion in the resource, and how they are parsed. In addition to the human decision-making aspects, the technological components of curation should be characterized; it is key to understand both manual and automated data exchange formats and web- or desktop-enabled data entry tools.

5.5.1 File Formats and Templates

The following section describes several existing approaches to support data entry for regulatory purposes (OECD HT), research data in bioinformatics (ISA-TAB, ISA-JSON) and its extensions for NM (ISA-TAB-Nano), as well NANoREG data logging templates [ref].

dentification and naming of substances under REACH and CLP" and requires specification of detailed chemical composition (including impurities and additives), concentrations of each constituent (typical concentration and range concentration), and links to chemical structures and identifiers. Each substance is assigned a unique identifier (UUID), which is specific to the company, submitting the dossiers. The common list of reference substances (also assigned UUID) are used to link company-specific substance entries to the same reference substance and chemical structures. Details on

manufacturing can be submitted in the relevant section. The experimental data is arranged hierarchically, within four endpoint groups (physicochemical, ecotox, environmental fate and toxicology) at the top. Each endpoint group contains several tens of templates for reporting specific endpoints (e.g. melting point under physchem group, aquatic toxicity under toxicology group), and the experimental data are reported separately for each substance in substance dossiers. Specifying the testing protocols with all associated details is mandatory. The protocols used in the regulatory context are established, e.g. OECD guidelines. The OHTs contain vocabularies in the form of pick-lists for some of the specified fields. A substance can be marked as nanomaterial, but there is no support for describing ENM specifics at the composition level. However, the surface composition (coating, core, functionalisation, along with the method of measurement), as well as ENM characterization can be specified as additional physicochemical endpoint study records (thirteen templates), which include granulometry (particle size distribution), agglomeration/aggregation, crystalline phase, crystallite and grain size; specific surface area; zeta potential; aspect ratio/shape, dustiness, porosity, pour density, catalytic and photocatalytic activity and radical formation potential vector quiral. The full list of OHTs is available at www.oecd.org/ehs/templates/templates.htm. Nanomaterials are covered by the substance definition of REACH, and the REACH provisions apply to them. NMs can be registered as nanoform(s) in the dossier of the corresponding non-nanoform of a substance or as distinct substance.

5.5.1.2 ISA-TAB, ISA-TAB-nano and ISA-JSON

ISA[6] is a metadata framework to manage an increasingly diverse set of life science, environmental and biomedical experiments that employ one or a combination of technologies. The framework provides means to describe complex experiments in a form of directed acyclic graph and is built around the concepts of Investigation (the project context), Study (a unit of research) and Assay (analytical measurements), arranged in as three hierarchical layers. The actual experimental readouts are stored in an additional data layer. It developed by S. Sansone's group at the University of Oxford e-Research Centre. ISA-Tab is the legacy format, relying on tab delimited files. The latest specification (Feb 2017) defines an Abstract Model , implemented in two format specifications ISA-Tab and ISA-JSON (JavaScript Object Notation). The new ISA-JSON specification includes a JSON schema and an ecosystem of tools used for creating, validating and visualizing documents and is designed around the concept of "core" ISA schema and "extensions". It is expected that different communities will develop extensions specific to their interests. The eNanoMapper project developed a (nano)material extension for ISA-JSONv1 [*cite D3.4, github*]. A separate helper JSON schema is implemented for definition of all components of the nanomaterial. The composition of a nanomaterial may contain one or several components. Each component has a role (core, coating, etc.) and linkages to other constituents. The linkage describes the relation between two components. For example, two components may be covalently bonded, one being embedded or encapsulated within another constituent etc.

[6] https://media.readthedocs.org/pdf/isa-specs/latest

The default approach for representation of chemical compounds in ISA-Tab[7] is an ontology entry, which typically points to a single chemical structure. This is insufficient for describing substances of complex composition such as nanomaterials; hence, a material file was introduced to address this need in ISA-Tab-Nano[8]. The latest ISA-Tab-Nano 1.2 specification recommends using the material file only for material composition and nominal characteristics, and to describe the experimentally determined characteristics in regular ISA-Tab assay files.

The ISA-Tab-Nano project is an effort of the National Cancer Institute (NCI), National Cancer Informatics Program (NCIP) and Nanotechnology Informatics Working Group (US Nano WG) and an attempt to extend the ISA-Tab format by introducing a separate file for describing the (nano)material components. The ISA-Tab-Nano is documented in a publication [2] and in the US Nano WG wiki2, which included sample spreadsheets, but no tools to parse the files and to enforce the specification. For this reason, the practical use of ISA-Tab-Nano is not straightforward, as demonstrated by the efforts of the FP7 NanoPuzzles project [3] and the introduction of "ISA-Tab-logic" templates by the FP7 NANoREG project.

5.5.1.3. NanoSafety cluster Excel templates

NANoREG data logging templates for the environmental, health and safety assessment of nanomaterials are developed under the JRC's leadership and in the frame of the EU-funded FP7 flagship project NANoREG. A team of experts in different fields (physical-chemistry, in vivo and in vitro toxicology) has produced a set of easy-to-use templates aimed at harmonising the logging of experimentally-produced data in the field of nano- environmental, health and safety (nanoEHS). The templates are freely available to the nanoEHS community (Common Creative Licence – Share alike) [ref] as jump start towards the harmonisation, sharing and linking of data, with the purpose of bringing benefits to the data management at European level and beyond. They have a common first part to identify the sample under investigation; a second part aimed at recording basic information on the dispersion method adopted and to record the essential parameters used to fully describe an assay (the experimental settings); and a third one to log the experimental results. The experimental parameters, their values, together with the Standard Operating Procedure (SOP) linked to a given template, allow to critically evaluate and/or to compare the results of a given assay performed in different laboratories. This approach should also allow reproducing the assay at a later stage. The structure adopted for the templates tries to reflect the ISA-TAB logic, already widely used in 'omics' studies, while addressing the low user-friendliness of ISA-TAB files, which limits its applicability in a "laboratory environment".

---

[7] Sansone, S.-A.; Rocca-Serra, P.; Field, D.; Maguire, E.; Taylor, C.; Hofmann, O.; Fang, H.; Neumann, S.; Tong, W.; Amaral-Zettler, L.; Begley, K.; Booth, T.; Bougueleret, L.; Burns, G.; Chapman, B.; Clark, T.; Coleman, L.-A.; Copeland, J.; Das, S.; de Daruvar, A.; de Matos, P.; Dix, I.; Edmunds, S.; Evelo, C. T.; Forster, M. J.; Gaudet, P.; Gilbert, J.; Goble, C.; Griffin, J. L.; Jacob, D.; Kleinjans, J.; Harland, L.; Haug, K.; Hermjakob, H.; Ho Sui, S. J.; Laederach, A.; Liang, S.; Marshall, S.; McGrath, A.; Merrill, E.; Reilly, D.; Roux, M.; Shamu, C. E.; Shang, C. A.; Steinbeck, C.; Trefethen, A.; Williams-Jones, B.; Wolstencroft, K.; Xenarios, I.; Hide, W. *Nat. Genet.* 2012, *44,* 121–126. doi:10.1038/ng.1054

[8] Thomas, D. G.; Gaheen, S.; Harper, S. L.; Fritts, M.; Klaessig, F.; Hahn-Dantona, E.; Paik, D.; Pan, S.; Stafford, G. A.; Freund, E. T.; Klemm, J. D.; Baker, N. A. *BMC Biotechnol.* 2013, *13,* 2. doi:10.1186/1472-6750-13-2

In the summer of 2017, the Center for the Environmental Implications of NanoTechnology (CEINT) lead a stakeholder input process to expand the ISA-Tab-Nano logic templates and to propose two new functional assay templates capturing data on attachment efficiency () and dissolution rate. The expansions that the templates would be poised to incorporate additional meta-data regarding sample preparation, instances of characterization, and media characteristics necessary to track nanomaterial transformations (http://ceint.duke.edu/research/nikc/isa-tab-nano). The various adoptions and adaptations of ISA-TAB-Nano, which was from the start intended as a flat file sharing format, provide a spreadsheet based solution for informing and organizing comparable datasets which is consistent, but not convenient. The templates represent an important incremental step toward harmonization of data, but one that must be surpassed in straightforwardness and ease of use to attract sufficient utilization for amassing significant data.

A different type of Excel templates, developed by the Institute of Occupational Medicine (IOM) (http://www.iom-world.org/) have been used by a number of NSC projects (NANOMMUNE, NANOTEST, ENPRA, MARINA, NANOSOLUTIONS).

5.5.1.4 Semantic Web formats

The semantic web has been introduced as the next generation world wide web, aimed at integrating data and knowledge from different online information sources [REF]. To implement this idea of a semantic web, the W3 Consortium has developed the Resource Description Framework (RDF, https://www.w3.org/RDF/) and a series of complementary standards to work with RDF, such as serialization formats like JSON-LD, RDF/XML, and Turtle [REF,REF,REF]. Because ontologies can also be expressed in RDF, for example with the Web Ontology Language (OWL) [REF], it is increasingly adopted as implementation for the FAIR data requirements. This RDF approach is being adopted by the eNanoMapper project and data provided by the eNanoMapper database can be downloaded as RDF data [REF], using the eNanoMapper ontology. With the semantic web serialization, eNanoMapper proposed an approach for data completeness testing and for answering scientific questions [REF].

5.5.1.5. Format conversions

ISA provide documentation and tools for conversion between ISA-TAB, ISA-JSON and ISA-RDF formats [ref]. Tools for conversion between several data formats (Excel templates, ISA-TAB, ISA-JSONv1, OECD HT and semantic formats) have been developed by eNanoMapper [REF]. These tools also enable automatic generation of ISA-JSON files from supported input formats (e.g. NANoREG templates) . If needed, the ISA-JSON files can be translated into legacy ISA-TAB via the tools provided by the ISA team. Export to ISA-JSON is enabled for each data collection of the eNanoMapper database.

5.6 Getting data out - support for data analysis

A number of recommendations (computational and strategic) for data curation promoted by [doi:10.3762/bjnano.6.179] relate to the ability of a data management solution support data analysis, data mining and seamless integration with modelling tools. The first level of support is to be able to download a user selected subset of the data to be further processed by a modelling package. The next level is the ability to export data programmatically, allowing integration into third party systems and workflow engines (e.g. KNIME). Another level of integration is providing unified access to data and analysis tools in addition to the data querying facilities. This could be done by either wrapping a selected set of statistical / machine learning packages into the database application, or using remote modelling or prediction services by submitting computational tasks and obtaining results transparently to the user. All these approaches have pros and cons and have been reviewed several times in the context of safety assessment of chemicals [10.1517/17425255.2012.685158, 10.1002/minf.201600082].

5.7 Metadata

Metadata are, very broadly speaking, “data about the data”. The distinction between data and metadata can vary widely across different disciplines; for example, in some cases metadata is conceived of only as the bibliographic information that allows tracing the source of the information set, where in other cases, the term might apply also to quantitative data that describe how (standard methods) or when (temporal specificity) a measurement was taken. Without focusing on a single definition and for the purpose of this roadmap, we consider metadata to be another lens through which to examine whether the data being recorded include sufficient information to later sort, evaluate, compare and analyze effectively. Moreover, it is important to note the need for fit-for-purpose considerations with regard to data and metadata, regardless of how one distinguishes between these. Whether there is sufficient information to support a desired combination, comparison and analysis of a dataset depends entirely on what research questions and relationships are being investigated [DOI: 10.1039/C5NR08944A].

As an example of how meta-data vary between studies or contexts, one can consider human toxicology and ecotoxicology studies. For human tox, the metadata consists mainly of pristine particle characterisation data, test methodology, and dosing protocols, which are then related to the “primary” observational data on detailed sub-lethal endpoints. In contrast, while the observed endpoints of ecotoxicology studies can often be much more simple, e.g. survival, the relevant meta-data required to describe the exposure will generally be significantly more extensive. For ecotoxicity the exposure system (the environmental compartment components) may interact with the nanomaterial resulting in transformations in the material form actually encountered by the receptor [DOI: 10.1021/es300839e; doi.org/10.1016/j.envint.2015.01.013]. In fact, realistically, actual exposures to materials in the environment for plants, animals and humans alike will contend with similar transformations, both before reaching and after entering the organism, such that the relevant form will be dependent on surrounding media, the exposure pathway, and other external governing factors. In practice, these transformed particles are difficult to measure in situ using routine techniques; yet, the

true form of a material that a receptor encounters and is the exposure relevant to understanding a resulting toxic response.

Because nanomaterial transformations are such a pivotal determinant of the outcome(s), it is not enough to know what you put into your ecotox system, and what media it was you put it in. There are multiple system dependencies that determine the transformations, so the meta-data requirements are extensive for capturing enough parameters to be able to model the fate and ultimately the exposure driving the observed effects. The importance of this can be seen in such examples as low dose chronic nanomaterial exposures in complex systems, where given only information on what material was added to the system, the toxic responses could not be predicted. In this case, an absence of detailed meta-data describing all biotic and abiotic system constituents and temporal variations in environmental conditions such that interactions can be interrogated would absolutely preclude interpretation of the results.

5.8 Ontologies,

Also consider recommendations in Winkler, D.A. Issues in, and examples of computational design of 'safe-by-design' nanomaterials, in Computational Nanotoxicology: Challenges, pitfalls and perspectives, Gajewski, A, Puzyn, T. (Eds) Pan Stanford Publishing 2017.

Ontologies are tools to formalize the language we use to exchange knowledge. Its need in nanosafety community has been clearly stated and resulted in a project call within the EU FP7 programme in 2012. This section describes a few aspects of current ontologies useful for nanosafety research. Example applications of ontologies in nanoinformatics includes the use of the Gene Ontology [REF] and the annotation of data in databases [REF] (see also the eNanoMapper database).

To make it easier to reuse the common language, various tools are available to use ontologies. Table X shows general ontology tools, but it is important to realize that many specific tools use ontologies too. For example, a database may use the ontology to provide faceted searching.

Table X: An overview of generic ontology tools.

Ontology Tool	**Description**
BioPortal http://bioportal.bioontology.org/	Searchable registry of ontologies.
OBO Foundry http://obofoundry.org/	Community project to develop and maintain ontologies in biology.
Ontology Lookup Service	Searchable registry of ontologies.
Protégé	Desktop software to view, search, and edit OBO and OWL ontologies.

Webulous https://www.ebi.ac.uk/efo/webulous/	Platform of a server and a Google Spreadsheet plugin that allows using ontologies in spreadsheet.
Ontology Slimmer	Java library that support remixing of existing ontologies. Used to create the eNanoMappper ontology.

5.8.1 NanoParticle Ontology (NPO)

The NPO was created out of the need to standardize data description in cancer nanotechnology research and enable searching and integration of diverse experimental reports. It covers various aspects of ENM description and characterisation, including chemical components in ENM, ENM type, physicochemical properties, experimental methods and applications in cancer diagnosis, therapy and treatment [10.1016/j.jbi.2010.03.001].

5.8.2 eNanoMapper ontology

The eNanoMapper ontology is a typical application ontology aimed at addressing needs of the community [REF]. This is in contrast to the demanding work of defining internally consistent ontology (see for example [REF]). Instead, by reusing (and occassionally extending) existing ontologies we are able to reflect that various sub-domains of the nanosafety community. The current ontology [REF] builds on several other ontologies, including the Basic Formal Ontology, the NanoParticle Ontology, the BioAssay Ontology, the Chemical Information ontology, the ontology of Chemical Entities of Biological Interest. The ontology releases are build by an automated environment that selects parts of these ontologies and integrates them into an ontology with exactly one ontology term for each concept. Guidance documents demonstrate how other controlled vocabularies map to this ontology, including a list of OECD nanomaterials [REF] and the JRC representative nanomaterials [REF].

The ontologies existing at the time of the eNanoMapper project that were related to modeling offered only fragmented coverage, with term definitions that were quite often oriented at the specific work or needs of the ontology they were a part of. In order to better describe nanoinformatics modelling actions and results, 162 terms were added to the eNanoMapper ontology, describing experimental and calculated (Image Analysis and algorithm-derived) descriptors, the processes that lead to their generation, modeling, statistics and algorithms [eNanoMapper D2.4 Descriptor Calculation Algorithms and Methods http://www.enanomapper.net/deliverables/d2/D2.4_Ontology_final_release_with_Annexes.pdf].

5.8.3 CHEMINF ontology

The Chemical Information (CHEMINF) ontology was set up to improve the interoperability of chemical information and data [REF]. It reuses concepts from other ontologies, like the BFO, SIO, and CHEBI and extends this with the notion that there is information about chemical compounds. This includes a chemical graph, names, identifiers, etc. Importantly, it also formalizes how to capture the difference between measured and calculated properties. The eNanoMapper ontology uses this ontology for nanomaterial identifiers and for computed properties.

5.8.4 BioAssay ontology (BAO)

The BioAssay Ontology (BAO) aims to address the need for describing and annotating biological assays in a standardized way. Experimental data is organized in "measure groups". A measure group can be annotated with an endpoint, screened entity (e.g. chemical or nanomaterial), assay method and participants (e.g. biological macromolecule). A bioassay may contain multiple measure groups. The measure groups could be combined to create "derived" measure groups (e.g. IC_{50} is a derived measure from dose response data) [10.1177/1087057111400191]. BAO has been used for annotation of a large number of HTS assays in PubChem [10.1186/1471-2105-12-257] and is used in Open Access ChEMBL database with chemical-protein affinity data. BAO is not a nanomaterial-specific ontology, but provides a useful data model for describing bioassays for arbitrary screened entities. The description of the screened entities is expected to come from elsewhere.

5.9 Data exchange

5.9.1 Data sharing

There is significant momentum towards greater access to journal articles, databases and government reports that will allow interested parties and the public in general to have a fuller range of nanoEHS data available for examination. While impediments will certainly lessen, it is unlikely that there will be full access to all data without some requirements being placed on data sharing. From that standpoint, those administering a database should establish an appropriate policy similar to steps they will take for ensuring data security (avoiding intrusions or unauthorized changes to data entries). The data user should, in turn, realize that the data accessed may be incomplete and use professional judgement accordingly.

Offering some examples of limitations that might be placed on data access is appropriate. Where academic colleagues will wait for the peer review process to be completed before releasing data, the industrial colleagues will wait for a patent to be allowed. For both, there may be issues of attribution, which would encompass authorship on papers that utilize an investigator's dataset or payment in the case of a company-sponsored study for a REACH dossier. Competitive pressures and anti-trust laws will influence company decisions, while project proposals, thesis requirements and intent to patent and commercialize may be prominent for some academics. For many of these examples, the remaining data access impediments can be resolved through setting time limits on data

embargoes, but for others, especially those data critical to a regulatory decision, industry will argue for confidential business information or trade secret status.

In terms of data sharing, the experiences with model organisms are illustrative of the above considerations. As described by Leonelli and Ankeny (ref. Studies in History and Philosophy of Biological and Biomedical Sciences 43 (2012) 29–36), the *C. elegans* and *Arabidopsis thaliana* communities of research have been more successful than their *Drosophila* and *Mus musculus* counterparts in standardizing on specific strains of those species, central stock source and sharing of information. Smaller community size and a more pressing need to leverage limited research funding are advantages to *C. elegans* and *Arabidopsis thaliana* progress, while the disruptions of selecting one strain for preferred study to suppliers and investigators attached to strains not selected is a disadvantage to the *Drosophila* and *Mus musculus* communities. As a multi-disciplinary effort, great care has been taken that the Nanoinformatics 2030 Roadmap itself be a tool fostering community interactions through both its description of current challenges and its suggested milestones.

Another important step towards the advancement of knowledge through sharing of nano datasets will be accomplished through the wide availability of online modelling capabilities. The current picture, where users first find nanomaterial data online, must download the datasets in order to process them offline for modelling and then possibly reupload any results (if they ever do so), makes little sense and severely slows down the advancement of knowledge. Online modelling (or Cloud modelling) infrastructure that makes available both nano-specific modelling and mathematical modelling tools is necessary to bring sophisticated tools and methodologies to a wider audience with a more moderate learning curve, ease of use and reduced or no costs. Although of course is dependent on appropriate and responsible data curation activities to ensure that high quality and complete datasets are provided, and that each study is screened appropriately. Otherwise creating validated and accurate models in a cloud based manner becomes impossible. Augmented by advanced Nanoinformatics tools, datasets will be enriched, allowing better decision making at a shorter cycle time. A global scope platform that provides access to mathematical modelling and nano-specific functionalities is Jaqpot Quattro (http://jaqpot.org), developed within the eNanoMapper project. Apart from a variety of algorithms for regression and clustering, users can perform Read Across, Optimal Experimental Design and Interlaboratory Comparison (Chomenidis et.al., 2017), supporting through both knowledge extraction from existing datasets and intelligent generation of consistent new data. There can be diverse motivations and requirements for each group of users (i.e. academia, industry etc.) that wishes to perform modelling work. At the same time, there can also be diverse platforms with clearly defined features that suit each group's purpose. The first such stakeholder-driven platform for nanomaterials risk modelling and risk management decision making is the SUNDS system that was developed by the EU FP7 SUN project. This online platform and the web-based System of Systems of the EU H2020 caLIBRAte project are growing in parallel to eventually form an integrated, interoperable data and modelling decision support infrastructure. This internet-based infrastructure will be capable of making efficient use of the available data for predictive modelling of possible risks from both legacy and novel nanomaterials, as well as for the assessment and management of these risks according to regulatory requirements.

5.9.2 Open Science

The European Commission have adopted the notion that concepts like Open Science and FAIR data benefit the European industries (SMEs and LEs) [REF]. The FP7 and H2020 have adopted policies around Open Access and Open Data publishing, with great respect of sustainability of existing industries. Open Science is about being able to reuse existing knowledge and finding its origin in the American Open Source community [REF]. They noted in the late nineties that the basic rights of being able to use and reuse disseminated knowledge, modify knowledge (curate it, extend it), and redistribute the outcome of that reuse. This section describes some initiatives important to the nanoinformatics community.

5.9.2.1 European Open Science Cloud (EOSC) and research data management

[citation] The European Commission is promoting open science data, supported by freely accessible infrastructure. OpenAire integrates institutional repositories and also provides the Zenodo repository to upload research output (datasets and publications) files up to 50GB. Zenodo is hosted at CERN and funded by the EU and CERN and provides integration with DropBox & GitHub. Users can define collections and communities, and configure the uploaded files for restricted access and embargo periods.

While Zenodo serves mainly archival purposes, the pan European collaborative data infrastructure (EUDAT) provides generic data services, such as storage and computing services to European researchers and research communities, and offers a joint metadata service integrating metadata from different communities into easily searchable and open catalogues. There is a number of services implementing cloud facilities: B2ACCESS (Authentication and Authorisation, identity provider, implemented by Unity IDM); B2DROP offering cloud services using own cloud, B2SHARE providing file sharing; B2STAGE – file transfer services, based on iRods data management system and GridFTP; B2SAFE providing replication and data management policies; B2FIND implementing metadata search, and finally BHOST allowing custom applications to be integrated within the EUDAT infrastructure.

5.9.2.2 Infrastructure for open science

Repositories versus databases… Confidential information protection , etc

Federated designs with Open APIs and Open ontologies: making the end-user aware of tools, and making the tools user friendly

Planning for federated knowledgebase/repository infrastructure (rather than one-size most-correct omniscient repository) is challenging but rewarding: from the data-supply side, supports interoperability among specialists, permits the flexibility of evolution in technology, increases data persistence-robustness, distributes the FTE for curating the repositories for completeness and by so doing improves overall quality of data available for further research and reuse. [US-NCI Alliance data sharing?] From the data consumer

side, API and consistent ontology facilitate casting the broadest search net, and catching harmonized returns for clear answers.
Interoperability with FAIR data ...

Open licenses as a legal promise for collaboration….

5.10 Other challenges

....developing field, quickly growing etc.

output of the projects is increasing, and we're not even able to (practically) keep up with the list of research articles [ref] and capturing in databases the material they report about.

5.11 Sustainability

Objective 2 of this roadmap addresses the overarching goal that all publically funded research data should be deposited in a sustainable database or knowledge resource. The sustainability of databases and knowledge resources created by different research and development activities is a complex multifactorial goal. What does this mean in practice? If, as part of a publicly-funded nanoEHS project, a laboratory has conducted valuable experiments which have yielded valuable results, that laboratory and others should be able to access those results in the future, e.g., five years after the project ends, and make sense and use of them in a reliable way. What do we have to do to achieve this goal with regards to nanoinformatics? The following elements are key for the success:

1) Agreement on best practices at the start of project with regards to experimental design and data management planning, including consideration of the end use of the data.
2) Data generated throughout the project should be well documented with regards to protocols, templates and metadata, and data processing workflows. Provision of data access, including review and testing, to the nanoEHS knowledge infrastructure, by the curator should be accomplished in a timely manner during the project (even if authorization controls are needed).
3) Education and training on data science for project team members should be completed early in projects. Interdisciplinary interactions between younger scientists within networks should be supported. (This will be a core task addressed by NanoCommons, the H2020-funded research infrastructure for nanoinformatics, which has a workpackage dedicated to training as part of its community building activities, and will also operate a Helpdesk offering support to the community in all aspects of nanoinformatics, starting in early 2018).
4) The FAIR principles should be followed with regards to access to scientific data resources (refer to objective 2)

5) Data resource completion (e.g., according to FAIR) and including a resource review should be delivered alongside the reporting and publication of the scientific results of projects.
6) A cluster and community wide data governance framework should be established to facilitate data sharing and interactions around data. For example, a simplified process and legal framework for data sharing between projects and programs would be beneficial.

However, clearly a more comprehensive vision would be to establish longer term knowledge infrastructure programs, which are actually required to ensure sustainability of scientific resources beyond the end of specific, individually funded projects. Such infrastructure programs can address issues of engineering, robustness, performance, quality control, review, maintenance, and support of nanoinformatics projects, which are often not addressed sufficiently during research projects, and are often not currently addressed at all after the completion of projects. OpenRiskNet is such an example where data services of relevance to safety assessment will be driven by the needs of the nanoEHS community. The infrastructure project has the NSC as a customer. International cooperation between EU and US programs should support the development of interoperable services, common data templates and shared data curation and are an opportunity for infrastructure programs to align, harmonize and avoid unnecessary costs from duplication.

Longer term community infrastructure programs such as NanoCommons (starting 2018) provide a common ground for the international community to work together on sustainability of community resources and aid the development and incorporation of a common language (ontology), best practices and knowledge sharing supporting excellence and governance. Programs such as NanoCommons should also be an opportunity to strengthen international cooperation between EU and US scientists working on related informatics problems, and to interact and collaborate with establishments and agencies (such as ECHA and US EPA) on the long-term provision of access to information resources to all stakeholders.

A mechanism for fostering a good progression from development of new methods, tools, ontology and best practices to efforts within standards groups (such as ISO, ASTM, OECD) to develop standards and test methods used within industry and obtaining regulatory acceptance should be developed. Although it can be said that some tests in their current form are considered acceptable, or are acceptable with minor adaptation (RIP-oN and ECHA guidance R7a-c appendices). [KP1] Such guidance could be included in documents specifically for difficult to test substances, much in the same manner the OECD have the “Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures” and others. Simply adding to existing frameworks eases cost and time, and makes the implementation more effici.

All initiatives should involve a strong consultation with industry and societal stakeholders so as to ensure that resources are created that satisfy needs and have utility.

6. Data Analysis: Nanochemoinformatics and statistical modelling

Tomasz Puzyn[1], Geert Verheyen[2], Sabine Van Miert[2], Baoshan Xing[3], Sarfraz Iqbal[1], Qing Zha[4], Vladimir Lobaskin[5], Gianpietro Basei[6]

[1] University of Gdansk, Gdansk, Poland
[2] University of Antwerpen, Antwerpen, The Netherlands
[3] University of Massachusetts, MA, USA
[4] Chinese Academy of Sciences, Shenyang, China
[5] University College Dublin, Dublin, Ireland
[6] Greendecision Srl, Italy

6.1 Introduction

The term 'nanochemoinformatics' refers to the application and appropriate adaptation of chemoinformatic methods for solving nanotechnology-related questions. Nowadays, nanochemoinformatic methods are mainly developed in the regulatory context of risk assessment, including hazard assessment and exposure assessment. This is because such methods as Quantitative Structure-Activity Relationships (QSAR) modeling for conventional (i.e. non-"nano") chemicals have already found increasing acceptance, primarily in devising an integrated testing strategy, but under some frameworks as a basis or alternative for toxicity testing with animals. However, the application of these methods is not limited to nanomaterial safety but also covers a broad range of questions regarding their functionality.

The name "chemoinformatics" came from "chemical information" understood as the information about chemical structure of chemicals. The information on different aspects of chemical structure can be encoded by a set of quantitative characteristics (e.g. the number of functional groups of a given type, the angle between two selected rings), which are generally referred to as **descriptors**.

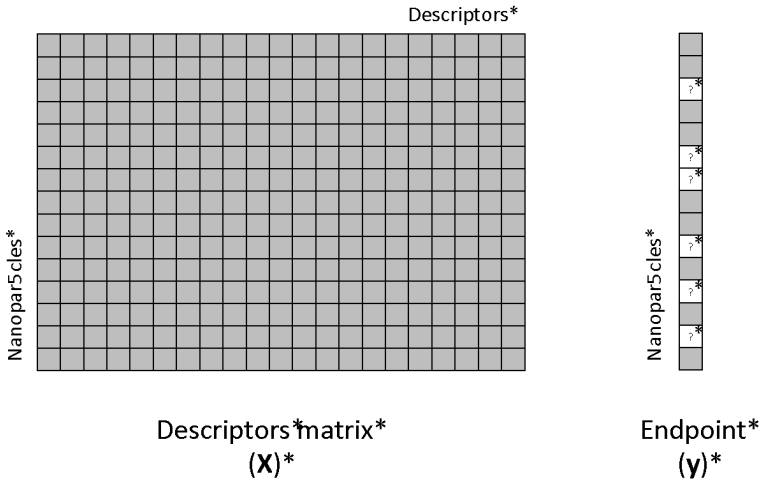


Figure 1. Nanochemoinformatics data. Typically, nanochemoinformatics data sets consist of various descriptors assembled in a descriptor matrix, which then are later on put into relation with information regarding specific data on toxicity

Nanochemoinformatics data are usually collected in matrices (tables), where rows represent individual NMs and columns correspond to descriptors (Figure 1). Such a matrix (usually referred as **X**-matrix) can be then used for analyzing similarities between structures of NMs (profiling), which mathematically refers to searching for similarities between the row vectors in the matrix. NMs can be clustered (grouped) together by analyzing the similarity of their descriptors by means of various hierarchical and non-hierarchical unsupervised algorithms (e.g. Hierarchical Cluster Analysis, Principal Component Analysis, Density-Based Spatial Clustering). In any case, care must be taken on the assumptions (e.g. normality, linearity) each algorithm employs for the analysis and the conclusions reached to be valid statistically. This is why linearity (e.g. Durbin-Watson test) and normality (e.g. Shapiro-Wilks test, Q-Q plots) checks should be performed prior to analysis for selecting the most appropriate algorithm.

However, the major role of nanochemoinformatics in hazard and exposure assessment is for filling gaps in the existing data. Such techniques help reduce bias originating from smaller datasets and increased difficulty in data handling and analysis, as long as the assumptions they employ are not violated [Applications of multiple imputation in medical studies: from AIDS to NHANES]. In such cases, an additional vector representing the endpoint data to be filled (**y**-vector) is used. The underlying idea is to use the descriptor matrix **X** and the existing elements of the endpoint vector **y** to estimate the absent elements of the endpoint vector **y** (indicated with "?" in Figure 1). Restated, a base set of descriptors (X) are used to estimate data-elements of an incomplete descriptor (Y). There are currently three data filling approaches, namely:

(i) (Quantitative) Structure-Activity Relationships methods (in case of nanomaterials often abbreviated as Nano-QSAR, Quantitative Nanostructure-Activity Relationships, QNAR or Quantitative Nanostructure-Toxicity Relationships, QNTR);
(ii) trend analysis and
(iii) read-across.

In the next sections we discuss current state-of-the-art and further developments necessary for making the existing nanochemoinformatic methods more useful from the regulatory and application points of view.

6.2 Descriptors

In nanochemoinformatics, the descriptors encode the information about the composition, structure, and properties of the NM. The descriptors of NMs refer to (Mills et al 2014):

- physical and chemical identity of NMs (i.e. size, shape, particle architecture, chemical composition of that architecture, e.g. core and coatings),
- intrinsic properties of NMs (e.g. crystal structure/crystallinity,, purity, surface area and rugosity, porosity, surface functionalities),
- extrinsic (system-dependent) properties of NMs (e.g. electrophoretic mobility/zeta potential, corona, degree of aggregation/agglomeration, dissolution, surface reconstruction, sorption, surface reactivity and persistence).

In some cases, data on NM activity such as toxicity endpoints (e.g. mutagenicity, EC_{50}/IC_{50}) might be used as descriptors as well, as the term has broad use in the modeling field. However, since this is not a purely chemical type of information, such data found the application in Quantitative Activity-Activity Relationships (QAAR) modeling. Descriptors can be **experimentally measured properties**, usually physicochemical properties, and **theoretical descriptors**, which are derived from the electronic, atomistic and molecular structure of the NM and its immediate environment. In Section 6, the emphasis is on descriptors as experimentally measured properties and in Section 7, the emphasis is on theoretical descriptors. For the purpose of predictive modelling, any quantitative characteristic that can be consistently measured or calculated in a controlled and reproducible way can serve as an NM descriptor.

The development of chemoinformatics (eco)toxicity models for chemicals relies heavily on the availability of appropriate chemical structure descriptors that tie relevant aspects of the molecular structure and physicochemical properties to the compound under investigation. Well-defined and robust descriptors are important for correct modelling and classification purposes. The base set of descriptors (the X-matrix) should satisfy the following criteria:

- allow a structural interpretation
- have a good correlation with at least one property
- not be trivial correlations of other base set descriptors
- exhibit gradual changes value with gradual changes in molecular structure
- be not restricted to a too small class of substances

Descriptor quality and relevance are even more important for NMs than for their bulk counterparts, requiring a larger number and different types of descriptors to account for properties due to factors such as their smaller size and larger surface-to-area ratio. Minimum data sets of NMs' descriptors required for predictive modelling encompass information on their chemical composition and intrinsic properties, which are specific for the NM and independent of the system. The system influencing extrinsic properties can be the matrix of a specific product (i.e. a specific formulation) or a specific biological

environment. Many datasets that are currently available for NMs are incomplete and unsystematic (Wang et al., 2014).

For chemicals, a hierarchy of descriptors can be derived already from the molecular structure. Molecular descriptors typically relate to steric and electronic properties of the compound and can be measured experimentally or computationally. Depending on the information content, descriptors are usually classified according to their dimensionality in 0D, 1D, 2D, 3D or 4D descriptors (Willighagen et al., 2006). 0D or constitutional descriptors don't take the molecular structure into account (e.g. molecular weight, atom number counts,…); 1D descriptors capture bulk properties like Log Kow; 2D descriptors are derived from molecular connectivity and 3D descriptors take the 3-dimensional geometry of the molecule into account. The 4D descriptors are used to describe the interaction field of the molecule or to describe different conformations of the molecule.Creo que un descriptor importante es N cantidad de átomos.

In the case of NMs, the composition and structure often do not reflect the most relevant properties for the activity, which may be entirely controlled by the engineered or spontaneously modified interface. These interfacial properties can be context-dependent and affected by the surrounding matrix. Therefore, the primary descriptors (composition and intrinsic) may not be the best suited to predict toxicological behaviour. Moreover, the NM properties can be interdependent and changing one property can result in the change of several other ones (Lynch et al., 2014). To tease out these relationships, well defined and good experimental data should be available to allow the development of models (and descriptors) that describe the relationship and that can subsequently be used to classify related NMs. One approach suggested by Lynch et al. (2014) is to identify 3 overarching descriptors (based on principal components analysis of observed variables) that describe intrinsic properties, extrinsic properties and composition aspects of the nanoparticle and that can be related endpoints to be modelled.

Among NMs, some of the most extensive research has been done on metal oxides. Ying et al. (2015) discern bare and coated metal oxide NMs in a toxicity study. For coated metal oxide nanoparticles, the structural descriptors used are those of the descriptors of the organic surface modification as this is mainly considered to be the key-factor to influence the toxicity and it can be referred to as a an organic chemicals QSAR study. For bare metal oxide NMs, the experimental descriptors covered morphological structural properties such as size distribution, shape, porosity, etc. and physicochemical properties such as zeta potential, pKa, surface charge, etc. Several technologies are available and are developed to measure and extract these properties (e.g. Bigdeli et al., 2014). Depending on the type of nanoparticle, different parameters may be more relevant. Additional descriptors can be derived from these measurements, such as surface/volume diameter, aspect ratio or sphericity (Gajewicz et al., 2015).

As part of the ITS-NANO project a gap analysis on the available knowledge required to (i) assess the risks of NMs and (ii) to develop an intelligent testing strategy, was conducted. As part of the outcome of the evaluation, risk analysis of NMs can use a similar approach as the traditional risk assessment paradigm used for chemicals, but it emphasizes the need for thorough physicochemical characterisation of nanomaterials compared to

practices that are currently in use, in order to take account of NM-specific or NM-relevant factors such as size, shape and surface characteristics. The ITS-NANO vision envisages that in the distant future (> 15 years) risk assessment will be increasingly reliant on modelling/ *in silico* approaches, with focused physicochemical, exposure and hazard testing only if additional information is required (Stone et al., 2014). These approaches will have to take into account the whole life cycle of nanomaterials from manufacture, use and their disposal, as well as the influence of the system on the nanomaterials' physicochemical properties at each stage during its life cycle and the consequences for potential (eco)toxicity and biological effects.

In contrast to descriptors for classic chemicals:

a) a matrix of nanodescriptors for chemoinformatic analysis rarely consists of the purely calculated (computational) descriptors only (experimentally-derived descriptors are additionally needed);
b) the experimentally-derived descriptors should take into account not only intrinsic, but also system-dependent (extrinsic) properties of the studied nanoparticles;
c) computational descriptors cannot be simply calculated from a single molecular model because of hardware limitations (separate simplified models representing various aspects of the structure, e.g. surface, aspect ratio, are needed).

Therefore, the most important challenges for further studies include:

1. the development of new descriptor sets (preferably computational) that enable comprehensively describe various aspects of the nano-structure;
2. the extension of currently used descriptor sets into system-dependent properties;
3. the development of simplified computational methods and/or molecular models (e.g. coarse-grain molecular mechanics) that enable calculating descriptors in the most efficient way.

6.3 Unsupervised chemoinformatics techniques for similarity analysis, profiling and grouping

Unsupervised techniques involve the use of statistical techniques for similarity analysis, profiling and grouping of chemicals in chemoinformatics. Specifically, these methods aim at discovering underlying patterns and relations in the dataset when data is not labeled (i.e.: there is no prior knowledge on data classification or categorization) (Bishop 2006). These techniques rely on computing different numerical chemical related parameters such as chemical descriptors. Such approaches can potentially be used in nanochemoinformatics for the identified categories of nanomaterials. A short description of few of them is given below.

6.3.1 Principal Components Analysis (PCA)

PCA is a statistical unsupervised learning technique that transforms a set of observations of possibly correlated variable into a set of values of linearly

uncorrelated variables called Principal Components (PCs) (ref). This technique help exploring the strong patterns in a chemical related data. The application of PCA for grouping of nanomaterials toxicity has already been suggested by Lynch I et. al. (2014) in an opinion article. As an example, Lynch et. al. (2014) initially suggested three principal components to be utilized to describe each nanomaterial, namely, intrinsic properties (inherent), extrinsic properties (interaction with media, molecular coronas etc.), and composition (proposition of a separate parameter e.g. inherent molecular toxicity). Each of these PCs has multiple contributors (observed variables as descriptors) and the relative contribution of these will vary for different nanomaterials. The schematic illustration of the use of PCA as applied to determination of the primary descriptors of NMS toxicity is shown in the following figure (Figure 2), taken from Lynch I et. al. (2014).

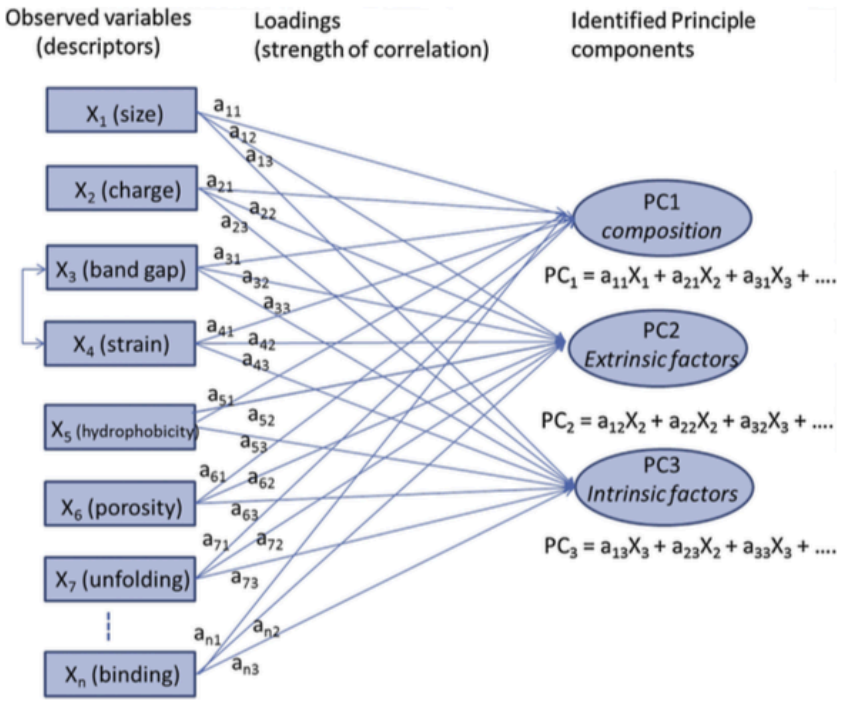


Figure 2: The schematic illustration of the use of PCA as applied to determination of the primary descriptors of NMS toxicity taken from Lynch I et. al. (2014).

6.3.2 Clustering

Clustering is another unsupervised learning technique that is very useful to explore the structures in a collection of data (Ref). In other words, this process consists of organizing objects – chemicals – into different groups having some similarities. In algorithms of clustering, the chemicals are collected which are 'similar' between themselves and are 'not similar' to the chemicals belonging to other chemical clusters. Alternative clustering algorithms include:

i) Exclusive clustering;
ii) Overlapping clustering;
iii) Hierarchical clustering;
iv) Probabilistic clustering (ref).

Exclusive clustering: in this class of clustering algorithms the data are grouped in an exclusive way, so that if a certain data point belongs to a definite cluster then it cannot be included in another cluster.

Overlapping clustering: these algorithms use fuzzy sets to cluster data, so that each object may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value.

Hierarchical clustering: this algorithm is based on the union between the two nearest clusters. The starting condition is realized by setting every data point as a cluster. After several iterations final clusters are realized.

Probabilistic clustering: it relies on a completely probabilistic approach.

The two most common clustering techniques are K-means (an exclusive clustering technique) and hierarchical clustering.

Clustering techniques are useful in initial steps of exploratory data analysis, to provide insights to modelers about similarities in both outcomes and descriptors. Moreover, these algorithms are a powerful set of tools to assist the categorization of chemicals into groups, and to further subgroup. Indeed, clustering methods have already been adopted in nanochemoinformatics as an initial step in the development of QSAR models to examine if chemicals that shown similarity in descriptors presented similar biological activity (Fourches2010, Epa2012, Fourches 2016), and to provide grouping of nanoparticles in different toxicity classes and then use those clusters to predict toxicity of untested materials (Gajewicz2015).

6.3.3 Self-organizing Maps

A Kohonen Self Organizing Map (SOM) is a special type of Artificial Neural Network (ANN) that it is used, like PCA, to reduce dimensionality of data, providing a representation of the input space through a lattice (usually one or two dimensional). The SOM method, likewise K-means, assigns data points (chemicals) to prototype vectors of the same size of the total number of descriptors, corresponding to a cell of the lattice. These vectors (called weight vectors or codes) are iteratively updated in such a way that they "self-organize" in a smoothed way: weight vectors of neighboring nodes in the lattice will thus be similar.

Specifically, the general algorithm to train a SOM works as follows:

1. Randomly initialize weight vectors corresponding to each node of the lattice.
2. Select at random an observation (a chemical) from the dataset.
3. Find the node in the lattice whose prototype vector in the lattice is the most similar (in terms, e.g., of Euclidean distance) to the observation: this node is known as the Best Matching Unit (BMU).
4. Weight vectors of nodes found within the radius of the neighborhood of the BMU are updated to be similar to the BMU vector. The closer a node is to the BMU, the more the weights are altered. The function used to compute the radius ensures it diminishes that at each iteration, in such a way that it starts covering the whole lattice and corresponds to a single node (the BMU) at the final step. Ideally, average distance between nodes in the lattice and dataset

sample(s) represented by that node decrease at each iteration, eventually reaching a plateau.

5. Repeat starting from step 2 for N iterations or until no significant change in the weight vectors is observed.

Once the SOM have been trained, it is possible to investigate the distribution of each descriptor across the SOM by means of heatmaps, and the comparison of these heatmaps provide insights about relationships between descriptors.

Another useful visualization is the so-called U-Matrix, which shows the distance between each node and its neighbors: large distances indicates dissimilarity among the nodes, and thus can be viewed as boundaries between clusters of nodes. Indeed, after training a SOM is typical to apply clustering algorithms (described in section 6.3.2) to nodes of the lattice, categorizing the original dataset accordingly. Ideally, the clusters derived in such a way are contiguous when drawn with different colors on the lattice, but it may happen that it is not the case. Contiguousness can be ensured by imposing, during clustering, the nodes to be both similar in weight vectors and close to each other in the lattice.

Alternatively, it is possible to guarantee classes to be contiguous by using Supervised SOMs (Melssen et al. 2006), where each node is associated, in addition to its weight vector, to a vector representing specific properties of interest. In this way the SOM learns at the same time relations in the descriptors (X space) and in the desired outcome (Y space), plus the correlation between the two spaces.

SOMs analysis followed by clustering analysis have been adopted as a tool to analyze toxicity-related cell signaling pathways for Metal and Metal Oxide Nanoparticles at different exposure times (Rallo et al. 2010). Supervised SOMs, on the other hand, have been used to explore experimental and simulated crystal structures via powder diffraction patterns, highlighting structure-property relations and demonstrating in such a scenario a more interpretability of the results with respect to their classical counterparts (Willighagen et al. 2007).

6.4 Supervised chemoinformatics techniques for filling data gaps

There are three groups of data filling: (Quantitative) Structure-Activity Relationship methods, trend analysis and read-across (Table 1). They are based on different assumptions and require different minimal number of data points (here: nanoparticles in a group for which the endpoint value **y** has been measured).

Table 1: Nanochemoinformatic methods of data filling

Method	Assumption	Description	Minimal number of data points
QSAR	Mathematical model:	Mathematical model that was not developed as part of the category formation process. The validity of the (Q)SARs	> 15

	y = f(**X**)	should be assessed according to 5 OECD (Q)SAR validation principles.	
Trend analysis	Trend in **y**	When some chemicals in a category have measured values of the endpoint (**y**) and a consistent trend is observed, missing values can be estimated by simple scaling from the measured values to fill in the data gaps.	> 3
Read across	Similarity in **X**	Endpoint value (**y**) for "source chemical" is used to predict the same endpoint for "target chemical".	1-6

6.4.1 Quantitative Structure Activity Relationships (QSAR)

Basics for the (Quantitative) Structure-Activity Relationships ([Q]SAR) approach were formulated for the first time in 1962 by Corwin Hansch and then implemented for designing new chemicals, mainly drugs [Ref]. The original approach was based on defining mathematical dependencies between the variance in molecular structures, encoded by so-called 'molecular descriptors' (e.g. number of particular functional groups, indexes that express topology and branching of a molecule), and the variance in biological activity in a set of compounds. Thus, if one calculated molecular descriptors for a group of similar chemicals and measured activity for a part of this group, the person would easily predict the lacking data from the molecular descriptors and a suitable mathematical model. Dependently on the modelled endpoint (nominal or numerical), the modelling is classified as qualitative or quantitative and abbreviated as SAR or QSAR [Ref].

Later on, when the problem of potential risk related to the use of new chemicals was raised, (Q)SAR methods found many applications in hazard assessment procedures. Examples of SAR and QSAR models developed for predicting various toxicity and ecotoxicity endpoints can be found in the literature [Ref]. Finally, as the application of (Q)SAR can reduce animal testing, which is particularly important in relation to the application of 3R principles (Replacement, Reduction, Refinement of animal testing) (Russel and Burch, 1959), those techniques have been recommended as valuable alternative methods in Article 13 of the EU REACH regulation[1]. The international co-operation among the OECD member countries on (Q)SARs started in 1990. The OECD principles for the validation of (Q)SAR models were released in 2004, and a guidance document was published in 2007.

In 2009[2] the groups of Jerzy Leszczynski (US) and Tomasz Puzyn (EU) jointly proposed to apply the QSAR methodology for predicting toxicity of nanomaterials (Nano-QSAR). The proof-of-the-concept – the first Nano-QSAR developed for toxicity of 17 metal oxides nanoparticles to *E. coli* bacteria – was published by the authors two years later.[3] At the same time, Andre Nel (US) and his collaborators from EU proposed to employ QSAR-like methods for High Throughput Screening to assess nanomaterial safety [Refs]. In parallel, the groups of Yoram Cohen (US) and Robert Rallo (EU) published the first classification Nano-SAR model [Ref] and proposed using self-organizing map analysis for assessing toxicity-related cell signaling pathways [Ref]. Those works were performed for metals and metal oxides as well. Methodology of Nano-(Q)SAR was further developed during next years including new descriptors, methods and models.[4-19]

It is widely accepted that Nano-QSAR models can significantly support current efforts in grouping (i.e. categorization) of nanomaterials and data gap filling within the established groups. There is a number of recently proposed grouping schemes for nanomaterials, for example the ones worked out by the ECETOC Nano Force Group (DF4NANO) [Arts et al. 2015], by the Dutch National Institute for Public Health and the Environment (RIVM) [Seller et al 2015] and by FP7 MARINA research project [Oomen et al. 2015].

QSARs developed for classic chemicals help identifying the direct influence of the structure on the modelled property. As such, the model indicates, which structural features are mainly responsible for the observed property or toxicity. In case of nanomaterials, it might be impossible to go directly from the structure to toxicity, since an additional level of information should be considered. In this context, Nano-QSAR models so-called "global models" can be applied for justifying particular grouping criteria. This means, the properties of higher levels (i.e. stability) might be expressed as a combination of properties from lower lever (i.e. chemical identity) plus the influence of the system (external conditions, e.g. pH). Thus, human toxicity and ecotoxicity can be expressed as a combination of intrinsic and extrinsic properties of nanomaterials. In such a way the hypotheses formulated *a priori* for particular grouping criteria can be verified.

When the grouping criteria for engineered nanomaterials are finally accepted, the efforts of the modelers should be put on developing so-called "local models" – the models capable predicting properties of nanoparticles within the identified groups (categories). In effect, the existing data gaps can be filled with using of scientifically justified methodology. However, only the results from appropriately validated models should be accepted. Well-known universal OECD principles on the validation of QSARs [Ref] provide the conditions that must be fulfilled to accept the model (and the predicted results) to be used for the regulatory purpose. These are:

1. Clearly defined endpoint;
2. Unambiguous algorithm;
3. Defined applicability domain;
4. Provided appropriate measures of goodness-of-fit, robustness and predictive ability;
5. Mechanistic interpretation, if possible.

It should be noted that the condition no. 4. implies that the model must be externally validated that means the validation should be performed with using nanoparticles not previously used for developing the model. Detailed interpretation of the five OECD principles for newly developed Nano-QSARs was widely discussed between the modelers and the summary was presented in Puzyn et al. [Ref]

In the previous contributions the application of Nano-QSAR models was limited rather to simple materials and simple cases, where usually *in vitro* toxicity endpoint was strongly related to one or two simple structural properties of materials that do not depend on the external conditions (i.e. intrinsic properties). In the further perspective, additional work is needed to obtain fully functional models.

First, the models must include information on the structure dynamically changing dependently on the external conditions. This may require including additional "dimensionality" in the set of descriptors. Moreover, pure probabilistic approach that QSAR is now, may be supported by deterministic component, i.e. QSAR equations may be augmented by equations derived based on physical principles.
Second, majority of the existing Nano-QSAR models was developed for nanomaterials build from the only one type of molecules (e.g. uncoated metal oxides nanoparticles) [Ref] or from two types, but one remained unchanged in the set (e.g. nanoparticles having the same core, but differing by coating) [Ref]. Therefore, there is a need to develop new structural descriptors for chemical materials varying by more than one chemical species at the same time.

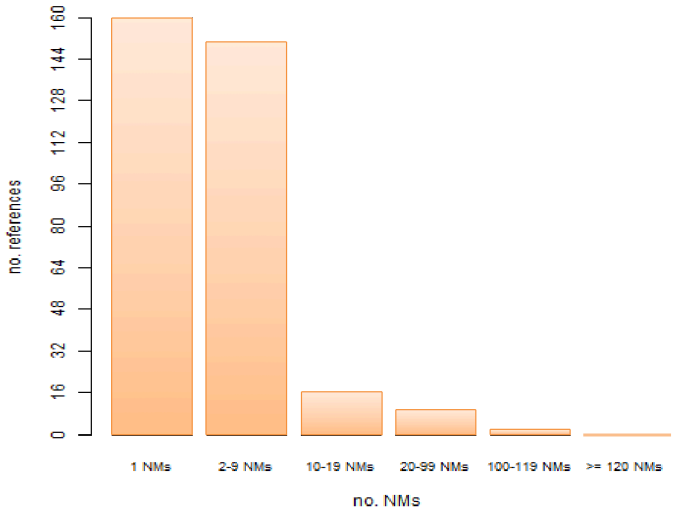


Figure 3: The number of literature references presenting experimental toxicity data vs. the number of nanomaterials (NMs) studied in these references (unpublished data from NanoPUZZLES project)

Third, the development of QSARs requires experimental data measured for sufficient number of materials varying by the structure and being representative for the whole general population of materials of given type (e.g. 50 ZnO nanoparticles differing in size, coating, crystal structure etc., representative for the whole space of possible ZnO nanoparticles variants). Moreover, data for all of them should be obtained by using the same experimental protocol. As it was concluded in various EU projects (refer to Figure 3), when analyzing literature, there are very rare cases, when such relatively large single database is available. Therefore, a possibility and limitations of merging the endpoints (data fusion) at higher ontological levels will be explored. For instance, could the endpoints: "percent apoptotic cells" (BAO_0002006) and "percent dead cells" (BAO_0002046) be merged into a single endpoint "percent cytotoxicity" (BAO_0000006)? Data fusion should be possible at least in the qualitative manner (translation of the numerical values into a nominal scale, e.g. "acceptable level of cytotoxicity" or "unacceptable level"). In effect, the size of available data sets would be extended. However, both (i) the development of detailed ontology and (ii) the studies of the influence of data fusion on the predictive ability are required.

Finally, as described in section 6.3 of this roadmap nanobioinformatics offers a variety of tools for better understanding Modes of Action (MoA) and deriving Adverse Outcome Pathways (AOPs) of engineering nanomaterials. On the other hand, Nano-QSAR can serve as a predictive tool for various endpoints. Thus, further work on the integration of both methodologies would result in increasing efficiency of both.

Nano-QSAR model should be well explained from mechanistic point. This is important, otherwise, it is only a mathematic statistical analysis.

In the hybrid methodology (Nano-QSAR/system biology) technique the QSAR component may serve for predicting the molecular key initiating event. Moreover, omics data may be considered as descriptors for QSAR studies.

Fourches et al (2010) demonstrated the use of QNAR modelling in predicting biological activity and cellular uptake of metal nanoparticles. In a first case, a structural characterisation of the NPs was used to define the molecular descriptors in the modelling exercise. The used molecular descriptors included structural descriptors such as type of metal core and experimental descriptors such as size, R1 and R2 relaxivities representing the magnetic properties and zeta potential reflecting the magnitude of electric charge on the NP surface. In addition, in a second case study modelling cellular uptake, 150 chemical descriptors of the surface-modifying organic molecules were calculated and were used as molecular descriptors in building models for cellular uptake of nanoparticles with the same core structure. This proof-of-concept study illustrated the feasibility of QNAR modelling, but also demonstrates that small variations in nanomaterial properties can drastically influence the biological activity and that modelling these effects remains challenging and will require high quality and large experimental datasets that will allow sufficiently robust modelling approaches (Fourches et al 2010).

6.4.2 Trend analysis

Trend analysis was first proposed by Brown for detecting nonrandom process trends.[20] He computed a “tracking signal” which is defined as the sum of the forecasting errors divided by the Mean Absolute Deviation. This approach was further improved by Trigg and Cembrowskl et al.[22] Trend analysis was firstly applied in filling the data gap for “quantitative endpoints” of chemical toxicology studies in March 2008 with the release of the OECD (Q)SAR Toolbox. According to the toolbox, methods based on trend analysis are applicable for filling data gaps in groups (categories) of chemicals, where clear systematic trend is the endpoint values is observed. That is to say Trend analysis is a method of predicting toxicity of a chemical by analyzing toxicity trends (increase, decrease, or constant) of tested chemicals. For example, in case of classic chemicals category containing compounds with a common functional group and an increasing chain length, the chain length affects the values of the octanol/water partition coefficient, which in turn may affect bioavailability and hence toxicity, both mammalian and aquatic.

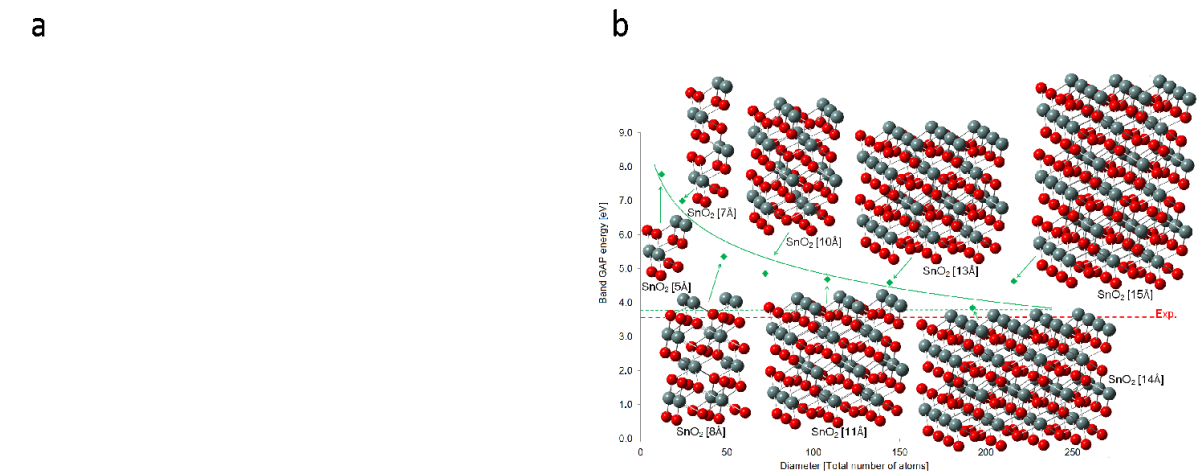


Figure 4: Two types of trends in phys/chem properties observed for nanoparticles when size is increasing [Ref]

Trend analysis techniques for nanomaterials have not been extensively used yet. However, it may serve for estimating size-dependent properties. As it was demonstrated by Gajewicz et al. [Ref] nanomaterials phys/chem properties may change either linearly within the entire range of sizes (Figure 4a) or change up to reaching so-called "saturation point" and then remain unchanged with the increasing size (Figure 4b). In both cases the property of interest can be easily interpolated, which is preferred in a regulatory context or – what is more challenging – extrapolated from the existing trend. From Puzyn et al.[23] research, we can conclude that the cytotoxicity was exponentially increased with the increasing of Enthalpy of formation of a gaseous cation (ΔH_{me+}) of metal oxide nanoparticles (Figure 5). Besides, Mu et al. [24] found that the cytotoxicity was exponential increased with the polarization force parameters (Z/r) of metal oxide nanoparticles (Figure 5).

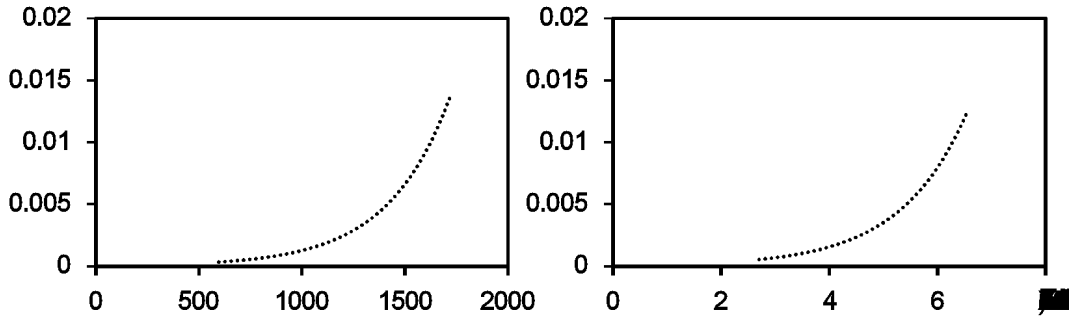


Figure 5: Two types of trends in cytotoxicity observed for metal oxide nanoparticles when ΔH_{me+} and Z/r is increasing respectively.

In a further perspective, it would be very practical to group the properties of nanoparticles according to the presented types of trends. Moreover, trend analysis might be tested to predict not only size-dependent, but also system-dependent properties, when the monotonically changing conditions causes monotonical changes in the properties of nanomaterials.

6.4.3 Read-across

When there is no visible trend in the defined category and the number of data points is too small for developing regular Nano-QSAR, either qualitative or quantitative read-across technique might be applied. Read-across is based on similarities between nanomaterials; the predicted endpoint value for "source chemical" is used to predict the same endpoint for sufficiently similar "target chemical" (Figure 6).

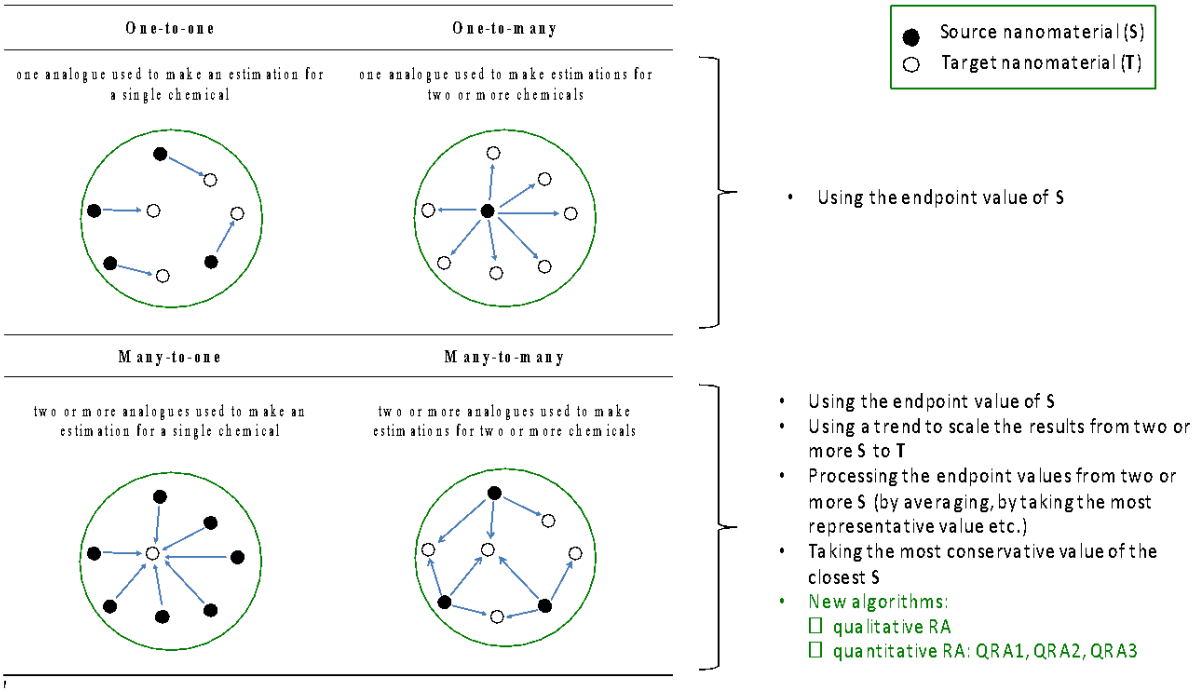


Figure 6: Schemes and currently available algorithms of read-across [Ref]

Based on the assumption that similar chemicals with structural and/or functional similarities have similar physico-chemical, toxicological, and ecotoxicological properties, read across can be applied to predict the unknown endpoint information (e.g. toxicity) for the 'target chemical(s)' with the known toxicity from the 'source chemical(s)'.[25] To identify the chemical similarities, the following two steps can be performed. Firstly, chemicals were represented as feature vectors of chemical properties either by binary or holographic fingerprints. Secondly, the similarity of chemicals can be quantified by various distances, i.e. Hamming, Euclidean, Cosine, Mahalanobis, Tanimoto distance, or linear or nonlinear relationships of the features.

In some cases, the read across approaches provide only the qualitative information and may be used to demonstrate the presence or absence of a property/activity under consideration. In contrast, various different approaches can be applied for quantitative prediction of the endpoint of interest, which are made by applying selected approximation type. For the similar source compounds in the established group, one can

use average, most conservative, mode, and median value. When the compounds' property related to the structural differences within the category follows a linear trend or regular pattern, interpolation or extrapolation from the empirical data for a given endpoint can be performed to fill in the data gaps.

Read-across can be performed in one of the four schemes: one-to-one, one-to-many, many-to-one and many-to-many. In the first two cases, the using of the endpoint value for source nanoparticle as the estimated value of the target nanoparticle is the only possible "algorithm" of read across. However, when read-across is based on more source nanoparticles, once can apply averaging, taking the most conservative value from the source nanomaterials etc. Puzyn et al. established a quantitative read across approach for nanomaterials (Nano-QRA) based on one-point-slope, two-point formula, or the equation of a plane passing through three points. The predictive capacity of Nano-QRA approach is better than other read across methods with different types of approximation in terms of both predictive power and reliability of predictions.[25] Recently, more sophisticated algorithms of qualitative and quantitative read-across were proposed by Gajewicz et al.[26] He proposed quantitative read across approach based on distance weight *k*-nearest neighbor algorithm ($QRA_{k\text{-}NN}$) for toxicity assessment of metal oxide nanoparticles, which displayed predominant prediction accuracy in both training and external validation.[26] These studies provide opportunities to broaden the application of read across method for filling empirical data gaps when adequate nanotoxicity data is not available.

Although read across possesses several advantages, i.e. easy to interpret and implement, applicable in modeling qualitative and quantitative toxicity endpoints, and flexible descriptors and similarity measures for expressing similarity between chemicals, the techniques of read-across have not been sufficiently standardized yet. In effect, very often the results of estimations with read-across are too 'expert-dependent' – may vary dependently on personal experience of expert conducting the study. This is important from the regulatory perspective, because it does not guarantee reliability and repeatability of the results. Moreover, statistical similarity measures cannot provide the information of toxicity mechanisms. Therefore, within some regulatory frameworks (e.g. REACH) bridging studies must be conducted in order to remove areas of uncertainty and prove similarities between the source and target chemicals. For example as a bare minimum physico-chemical measures must be known for both source and target, and the (eco)toxicological bridging studies will then be chosen based on the strategy and the endpoint needing to be fulfilled. In addition, complex similarity measures need complicated model interpretation. Furthermore, in the case of inadequate analog chemicals or conflicting toxicity profiles of analogs, the read across is inapplicable or inaccurate. Therefore, the development of novel read across algorithms that can provide reliable predications of the unknown data without further experiments is of great importance.

Further developments in this area should include design of novel and suitable numerical algorithms for read-across that will be useful in the context of filling data gaps. The feasibility and predictive ability of newly developed read-across algorithms should be verified and validated. Therefore, it would be very practical to establish the principles for the validation of read-across approaches by means of suitable case-studies (i.e. using external data obtained from regulatory (eco)toxicity tests). Furthermore, the

recommendations on existing read-across approaches, which are the most relevant for filling data gaps for nanomaterials, should be delivered. In a further perspective, the acceptable and sufficiently standardized algorithm(s) should be implemented into the user-friendly software (e.g. OECD QSAR Toolbox).

It is worth mentioning that the proposed algorithms of read-across are universal that means enable to fill the data gaps within categories defined by using of any criteria and grouping (categorization) system to be applied.

References:

1. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerming the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, and amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. In Official Journal of the European Union, L 396/1, 2006.

2. Puzyn, T.; Leszczynska, D.; Leszczynski, J., Toward the Development of "Nano-QSARs": Advances and Challenges. *Small* **2009,** *5*, (22), 2494-2509.

3. Puzyn, T.; Rasulev, B.; Gajewicz, A.; Hu, X. K.; Dasari, T. P.; Michalkova, A.; Hwang, H. M.; Toropov, A.; Leszczynska, D.; Leszczynski, J., Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nature Nanotechnology* **2011,** *6*, (3), 175-178.

4. Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Gini, G.; Puzyn, T.; Leszczynska, D.; Leszczynski, J., Novel application of the CORAL software to model cytotoxicity of metal oxide nanoparticles to bacteria Escherichia coli. *Chemosphere* **2012,** *89*, (9), 1098-1102.

5. Toropov, A. A.; Toropova, A. P.; Puzyn, T.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J., QSAR as a random event: Modeling of nanoparticles uptake in PaCa2 cancer cells. *Chemosphere* **2013,** *92*, (1), 31-37.

6. Kar, S.; Gajewicz, A.; Puzyn, T.; Roy, K.; Leszczynski, J., Periodic table-based descriptors to encode cytotoxicity profile of metal oxide nanoparticles: a mechanistic QSTR approach. *Ecotoxicology and environmental safety* **2014,** *107*, 162-9.

7. Lubinski, L.; Urbaszek, P.; Gajewicz, A.; Cronin, M. T. D.; Enoch, S. J.; Madden, J. C.; Leszczynska, D.; Leszczynski, J.; Puzyn, T., Evaluation criteria for the quality of published experimental data on nanomaterials and their usefulness for QSAR modelling. *Sar Qsar Environ Res* **2013,** *24*, (12), 995-1008.

8. Toropova, A. P.; Toropov, A. A.; Puzyn, T.; Benfenati, E.; Leszczynska, D.; Leszczynski, J., Optimal descriptor as a translator of eclectic information into the prediction of thermal conductivity of micro-electro-mechanical systems. *J Math Chem* **2013,** *51*, (8), 2230-2237.

9. Kar, S.; Gajewicz, A.; Puzyn, T.; Roy, K., Nano-quantitative structure-activity relationship modeling using easily computable and interpretable descriptors for uptake of magnetofluorescent engineered nanoparticles in pancreatic cancer cells. *Toxicol in Vitro* **2014,** *28*, (4), 600-6.

10. Odziomek, K.; Ushizima, D.; Puzyn, T.; Haranczyk, M., Toward quantitative structure-activity relationship (QSAR) models for nanoparticles. *Abstr Pap Am Chem S* **2014,** *248*.

11. Sizochenko, N.; Rasulev, B.; Gajewicz, A.; Kuz'min, V.; Puzyn, T.; Leszczynski, J., From basic physics to mechanisms of toxicity: the "liquid drop" approach applied to develop predictive classification models for toxicity of metal oxide nanoparticles. *Nanoscale* **2014,** *6*, (22), 13986-93.

12. Toropova, A. P.; Toropov, A. A.; Benfenati, E.; Puzyn, T.; Leszczynska, D.; Leszczynski, J., Optimal descriptor as a translator of eclectic information into the prediction of membrane damage: the case of a group of ZnO and TiO2 nanoparticles. *Ecotoxicology and environmental safety* **2014,** *108*, 203-9.

13. Ambure, P.; Aher, R. B.; Gajewicz, A.; Puzyn, T.; Roy, K., "NanoBRIDGES" software: Open access tools to perform QSAR and nano-QSAR modeling. *Chemometr Intell Lab* **2015,** *147*, 1-13.

14. Gajewicz, A.; Schaeublin, N.; Rasulev, B.; Hussain, S.; Leszczynska, D.; Puzyn, T.; Leszczynski, J., Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: hints from nano-QSAR studies. *Nanotoxicology* **2015,** *9*, (3), 313-25.

15. Mikolajczyk, A.; Gajewicz, A.; Rasulev, B.; Schaeublin, N.; Maurer-Gardner, E.; Hussain, S.; Leszczynski, J.; Puzyn, T., Zeta Potential for Metal Oxide Nanoparticles: A Predictive Model Developed by a Nano-Quantitative Structure-Property Relationship Approach. *Chem Mater* **2015,** *27*, (7), 2400-2407.

16. Mikolajczyk, A.; Pinto, H. P.; Gajewicz, A.; Puzyn, T.; Leszczynski, J., Ab Initio Studies of Anatase TiO2 (101) Surface-supported Au-8 Clusters. *Current topics in medicinal chemistry* **2015,** *15*, (18), 1859-1867.

17. Sikorska, C.; Puzyn, T., The performance of selected semi-empirical and DFT methods in studying C-60 fullerene derivatives. *Nanotechnology* **2015,** *26*, (45).

18. Sizochenko, N.; Rasulev, B.; Gajewicz, A.; Mokshyna, E.; Kuz'min, V. E.; Leszczynski, J.; Puzyn, T., Causal inference methods to assist in mechanistic interpretation of classification nano-SAR models. *Rsc Adv* **2015,** *5*, (95), 77739-77745.

19. Tantra, R.; Oksel, C.; Puzyn, T.; Wang, J.; Robinson, K. N.; Wang, X. Z.; Ma, C. Y.; Wilkins, T., Nano(Q)SAR: Challenges, pitfalls and perspectives. *Nanotoxicology* **2015,** *9*, (5), 636-642.

20. Brown, R. G., *Smoothing, Forecasting and Prediction of Discrete Time Series*. Prentice-Hall: Englewood-Cliffs, N. J., 1962.

21. Trigg, D., Monitoring a forecasting system. *J. Oper. Res. Soc.* **1964,** *15*, (3), 271-274.

22. Cembrowski, G. S.; Westgard, J. O.; Eggert, A. A.; Toren, E. C., Trend detection in control data: optimization and interpretation of Trigg's technique for trend analysis. *Clin. Chem.* **1975,** *21*, (10), 1396-1405.

23. Puzyn, T.; Rasulev, B.; Gajewicz, A.; Hu, X.; Dasari, T. P.; Michalkova, A.; Hwang, H.-M.; Toropov, A.; Leszczynska, D.; Leszczynski, J., Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nat. Nanotechnol.* **2011,** *6*, (3), 175-178.

24. Mu, Y.; Wu, F.; Zhao, Q.; Ji, R.; Qie, Y.; Zhou, Y.; Hu, Y.; Pang, C.; Hristozov, D.; Giesy, J. P.; Xing, B., Predicting toxic potencies of metal oxide nanoparticles by means of nano-QSARs. *Nanotoxicology* **2016**, 1-25.

25. Gajewicz, A.; Jagiello, K.; Cronin, M. T. D.; Leszczynski, J.; Puzyn, T., Addressing a bottle neck for regulation of nanomaterials: quantitative read-across (Nano-QRA) algorithm for cases when only limited data is available. *Environmental Science: Nano* **2017,** *4*, (2), 346-358.

26. Gajewicz, A., What if the number of nanotoxicity data is too small for developing predictive Nano-QSAR models? An alternative read-across based approach for filling data gaps. *Nanoscale* **2017,** *9*, (24), 8435-8448.

Benfenati E, Toropov AA, Toropova AP, Manganaro A, Gonella Diaza R (2011) CORAL software: QSAR for anticancer agents. Chem Biol Drug Des 77: 471-476

Bigdeli A, Hormozi-Nezhad MR, Jalali-Heravi M, Abedini MR, Sharif-Bakhtiar F (2014) Towards defining new nano-descriptors: extracting morphological features from transmission electron microscopy images. RSC Advances 4: 60135-60143

Bishop CM (2006) Pattern Recognition and Machine Learning. Springer.

Dekkers S, Ooman AG, Bleeker EAJ, Vandebriel RJ, Micheletti C, Cabellos J, Janer G, Fuentes N, Vazquez-Campos S, Borges T, Silva MJ, Prina-Mello A, Movia D, Nesslany F, Ribeiro AR, Leite PE, Groenewold M, Cassee FR, Sips AJAM, Dijkzeul A, van Teunenbroek T, Wijnhoven SWP (2016) Towards a nanospecific approach for risk assessment. Regulatory Toxicology and Pharmacology 80: 46-59

Gajewicz A, Puzyn T, Rasulev B, Leszczynska D, Leszczynski J (2011) Metal oxide nanoparticles: size-dependence of quantum-mechanical properties. Nanoscience & Nanotechnology – Asia 1: 53-58

Gajewicz A, Schaeublin N, Rasulev B, Hussain S, Leszczynska D, Puzyn T, Leszczynski J (2015) Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies. Nanotoxicology 9: 313-325

Jones DE, Ghandehari H, Facelli JC (2016) A review of the applications of datmining and machine learning for the prediction of biomedical properties of nanoparticles. Computer Methods and Programs in Biomedicine 132: 93-103

Kar S, Gajewicz A, Puzyn T, Roy K, Leszczynski J (2014) Periodic table-based descriptors to encode cytotoxicity profile of metal oxide nanoparticles: a mechanistic QSTR approach. Ecotoxicology and Environmental Safety 107: 162-169

Lam CW, James JT, McCluskey R, Hunter RL (2004) Pulmonary toxicity of single-wall carbon nanotubes in mice 7 and 90 days after intratracheal instillation. Toxicological Sciences 77: 126-134

Liu R, Zhang HY, Ji ZX, Rallo R, Xia T, Chang CH, Nel A, Cohen Y (2013) Development of structure-activity relationship for metal oxide nanoparticles. Nanoscale 5: 5644-5653

Lynch I, Weiss C, Valsami-Jones E (2014) A strategy for grouping of nanomaterials based on key physico-chemical descriptors as a basis for safe-by-design NMs. Nano Today 9: 266-270

Mills KC, Murry D, Guzan KA, Ostraat ML (2014) Nanomaterial registry: database that captures the minimal information about nanomaterial physico-chemical characteristics. Journal of Nanoparticle Research 16: 2219

Nanda KK (2012) Liquid-drop model for the surface energy of nanoparticles. Physics Letters A 376: 1647-1649

NRC (2009) Review of federal strategy for nanotechnology-related environmental, health, and safety research. Washington, DC: The National Academies Press

Pathakoti K, Huang MJ, Watts JD, He X, Hwang HM (2014) Using experimental data of *Escherichia coli* to develop a QSAR model for pedicting the photo-induced cytotoxicity of metal oxide nanoparticles. J Photochem Photobiol B 130: 234-240
Puzyn T, Leszczynska D, Leszczynski J (2009) Toward the development of "Nano-QSARs": advances and challenges. Small 5: 2494-2509
Puzyn T, Rasulev B, Gajewicz A, Hu X, Dasari TP, Michalkova A, Hwang HM, Toropov A, Leszczynska D, Leszczynski J (2011) Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. Nature nanotechnology 6: 175-178
Sayes CM, Smith AP, Ivanov IV (2014) A framework for grouping nanoparticles based on their measurable characteristics. International Journal of Nanomedicine 8: 45-56
Sizochenko N, Rasulev B, Gajewicz A, Mokshyna E, Kuz'min VE, Leszczynski J, Puzyn T (2015) Causal inference methods to assist in mechanistic interpretation of classification nano-SAR models. RSC Advances 5: 77739-77745
Stone V, Pozzi-Mucelli S, Tran L, Aschberger K, Sabella S, Vogel U, Poland C, Balharry D, Fernandes T, Gottardo S, Hankin S, Hartl MGJ, Hartmann N, Hristozov D, Hund-Rinke K, Johnston H, Marcomi A, Panzer O, Roncato D, Saber AT, Wallin H, Scott-Fordsmand JJ (2014) ITS-NANO – Prioritising nanosafety research to develop a stakeholder driven intelligent testing strategy. Particle and Fibre Toxicology 11: 9
Toropova AP, Toropov AA, Maksudov SKh (2006) QSPR modelling mineral crystal lattice energy by optimal descriptors of the graph of atomic orbitals. Chemical physics Letters 428: 183-186
Toropov AA, Toropova AP, Benfenati E, Leszczynska D, Leszczynski J (2010) SMILES-based optimal descriptors: QSAR analysis of fulleren-based HIV-1 PR inhibitors by means of balance of correlations. J Comput Chem 31: 381-392
Toropov AA, Rallo R, Toropova AP (2015) Use of QUASI-SMILES and Monte Carlo optimization to develop quantitative feature property/activity relationships (QFPR/QFAR) for nanomaterials. Current Topics in Medicinal Chemistry 15: 1837-1844
Toropova AP, Toropov AA, Manganelli S, Leaone C, Baderna D, Benfenati E, Fanelli R (2016) Quasi-SMILES as a tool to utilize eclectic data for predicting the behavior of nanomaterials. NanoImpact 1: 60-64
Wang XZ, Yang Y, Li R, Mcguinnes C, Adamson J, Megson IL, Donaldson K (2014) Principal component and causal analysis of structural and acute in vitro toxicity data for nanoparticles. Nanotoxicology 8: 465-476
Ying J, Zhang T, Tang M (2015) Metal oxide nanomaterial QNAR models: available structural descriptors and understanding of toxicity mechanisms. Nanomaterials 5: 1620-1637

7. Data Analysis: Modelling the properties, interactions and fate of nanomaterials

Pietro Asinari[1], Vladimir Lobaskin[2], Thomas Puzyn[3]

[1] Politecnico di Torino, Torino, Italy
[2] University College Dublin, Dublin, Ireland
[3] University of Gdansk, Gdansk, Poland

7.1 Introduction to Materials Modelling

In recent decades, computer simulations have become an indispensable instrument in studies of materials. Now simulations involving hundreds of thousands of atoms on a microsecond time scale become routine while state-of-the-art simulations correspond to one or two order larger size- and time scale (Palmer et al., 2015). Molecular simulations are also becoming an intrinsic part of the applied research, such as drug design, nanotechnologies and nanomedicine, providing possibilities for screening of different compounds, with a perspective of in silico construction of molecules and materials with desired specific properties. Among areas of actual interest is investigation of bionano interface, which is driven by applications of nanomaterials in medicine, food, and cosmetics (Brancolini et al., 2012; Ding et al., 2013;Khan et al., 2013) as well as prediction of toxicity. Although molecular simulations cannot imitate biological events leading to toxicity, they can provide a framework for systematic evaluation of interactions of NMs with biomolecules. Understanding of these interactions and bionano interface structure is crucial for achieving a better control over the surface activity, for developing safety regulations, and reducing the associated health risks.

More generally, physics and chemistry-based materials modelling can serve as a source of additional information about the NMs (e.g. intrinsic and extrinsic properties) where it cannot be measured or is unknown for some reason. Moreover, it can provide a time and cost-effective alternative to experimental measurements of materials' properties. Finally, materials modelling offers a possibility to predict the material's functionality or activity even before it is produced to prevent the appearance of properties of concern, and thus enables development of materials that are safe by design.

7.2 Use of computational models to compute NM properties

The implementation of modelling in the nanomaterial domain is a relatively recent. direction of research. Most published works focus on prediction of nanoparticle cellular uptake, cytotoxicity, molecular loading, molecular release, nanoparticle adherence, nanoparticle size, and polydispersity (Jones et al., 2016). Several studies show very reasonable predictions;, however, most of these models focus on specific types of nanoparticles only and rely on the use of very limited datasets, making the generalization of the models very challenging, given the complexity of the nanomaterial world. Seria interesante tener en cuenta la computación cuántica.

7.2.1 Intrinsic properties

In what regards the chemical composition and intrinsic properties of nanomaterials, several software programs (e.g. Adriana. Code, Dragon, Molcomm-Z and PaDEL-Descriptor) are available and can be used to calculate relevant descriptors on chemical structures. Some descriptors can be extracted directly from results from quantum-mechanical calculations (examples). These calculations can be very computationally-intensive and time consuming. Time and cost of calculations can be reduced by selecting the appropriate level of theory for geometry optimization, but this can go at the cost of the predictive ability of the model. Using simplified, semi-empirical

methods (Recife Model 1, Parametrization Model 6, etc.) it is possible to calculate the molecular parameters for molecules in a short time (Puzyn et al., 2009). However, for structures that are largely different from the structures used for parametrization, the results will not suffice and may lead to incorrect description of the structure. For untypical molecules it is better to use ab initio or Density Functional Theory methods which require more computational resources. This situation also applies for nanomaterials, because they are no longer simple molecular compounds and the implementation of higher levels of theory in the ab initio formalism is recommended (Puzyn et al., 2009). Fortunately, literature indicates that the most significant size-dependent changes of some physicochemical properties of nanoparticles are observed below 5 nm, whereas the changes for sizes between 15 and 90 nm can be neglected. In addition, Gajewicz et al. (2011) showed that for metal oxide clusters several molecular descriptors change with the size of the clusters. The physicochemical properties either change (i) linearly with size or (ii) up to reaching "saturation point", in which the properties have constant values characteristic for the bulk material. This implies that it is possible to estimate the properties of a given nanoparticle by performing calculations for a series of much smaller molecular clusters and then fitting an appropriate function (Gajewicz et al., 2011).Las nuevas propiedades no solamente están vinculadas al tamaño y la geometría, sino a los fenómenos de transporte que se correlaciona con las cuasipartículas dígase fonones, plasmon y pienso que este aporte contribuye de manera sustancialTheoretical descriptors involve quantum chemical or molecular simulation methods to derive molecular properties, but nanomaterials may have their own special properties, e.g. for metal oxide nanomaterials the crystal structure is important (Ying et al., 2015). Different types of theoretical descriptors are discerned: (i) constitutional properties such as periodic table-based descriptors (molecular weight, cation charge, metal electronegativity, etc.) which are easy to obtain (Kar et al., 2014) and (ii) electronic properties (regarding metal oxide NPs) such as band gap and valence gap energy, ΔHMe+ or the molar heat capacity. From a quantum chemistry viewpoint nanoparticles are large systems, which complicates the necessary calculations at the proper level of theory and other approaches are needed to determine the proper structural descriptors for nano-QSARs (Puzyn et al., 2009). These quantum-chemical properties can be calculated using several software programs, e.g. Puzyn et al (2011) established a model to describe the cytotoxicity of metal oxide NP to E. coli, calculating 12 descriptors at the semi-empirical level of the theory using the PM6 method implemented in the MOPAC software. The enthalpy of formation of gaseous cation with the same oxidation state as the metal-oxide structure, ΔHMe+, was shown an efficient descriptor of the chemical stability of metal oxide and their cytotoxicity. Other descriptors that have been calculated for metal oxide nanoparticles include molar heat capacity, average of the alpha and beta lowest unoccupied molecular orbital (LUMO) energies (Pathakoti et al., 2014) and the atomization energy, atomic mass, conduction band energy, ionization energy and electronegativity (Liu et al., 2013). The calculation of these descriptors is computationally demanding.

Other approaches to derive structural descriptors have been described in the literature (i) Glotzer and Solomon (2007) proposed a system of eight orthogonal "dimensions" (surface coverage; aspect ratio, faceting, pattern quantization, branching, chemical ordering, shape gradient and roughness) to measure the structural similarities between various nanostructures. How to quantify these eight dimensions still needs to be solved. (ii) The chemical composition can also be expressed by simple constitutional

descriptors (e.g. atomic numbers) or by a single descriptor based on correlation weights derived from molecular graph or atomic orbitals theory (Toropova et al., 2006). Based on these theories, another approach that has been implemented in nano-QSAR model development makes use of the CORAL software (Benfenati et al., 2011). Based on SMILES, optimal descriptors can be defined and correlated with endpoints such as cytotoxicity of metal oxide nanoparticles (Toropova et al., 2012) or binding affinity of fullerene derivatives to HIV-1 protease (Toropov et al., 2010). However, for general implementation of nano-QSAR models this method of representation of the structure is unfeasible because of the complexity of the molecular architecture. Therefore in a next evolution, the chemical information was integrated with additional heterogeneous (eclectic) data, such as size, concentration, irradiation, porosity, etc. (Toropov et al., 2015). Building on the SMILES notation, additional SMILES-like sequences of symbols that codify the physicochemical and biochemical conditions of chemicals and nanomaterials in biological systems have been introduced and termed quasi-SMILES notation. These can then be used to calculate optimal descriptors and applied in nano-QSAR modelling (Toropov et al., 2015; Toropova et al., 2016). (iii) Simplex representation of molecular structure (SiRMS) are a 2D level generated two-, tri-, and tetra-atomic molecular fragments for which descriptors can be derived (Sizochenko et al., 2015). (iv) The Liquid Drop Model has been described as a novel approach to represent the supramolecular structure of nanoparticles (Sizochenko et al., 2014). The main idea behind this approach is to use a combination of simple descriptors which reflect nanoparticles' structure for the different levels of organization: from a single metal oxide molecule (i.e. chemical structure) to a supramolecular ensemble of molecules (i.e. nanoparticle size). LDM has for example been described to determine the surface energy of nanoparticles (Nanda, 2012). Using the LDM extensive quantum-mechanical calculations can be avoided. (v) QSAR-perturbation approach in which a moving average approach was applied to the data in order to generate new descriptors that reflect their relative importance in the model (Luan et al., 2014)

7.2.2 Extrinsic properties

The environmental fate and biological activity of a nanomaterial can be influenced modified by the medium, which can affect its surface charge, surface reactivity, and surface composition (coating) and even lead to a change in the particle's core composition . Therefore, a set of extrinsic characteristics should complement the standard description. The typical quantities used with nanomaterials include:

- hydration energy, heats of immersion, contact angle for water
- surface charge density at different pH values and salt concentrations
- nanomaterial dissolution rate
- binding energies for essential biomolecules or adsorbates molecular groups

Atomistic simulation, both classical and ab initio, and mean-field theories (Poisson-Boltzmann theory) can be used to evaluate these properties for NM at realistic conditions. Hydration energy (per unit area) or heat of immersion, or contact angle can be used to characterise the degree of hydrophobicity of the material. For example, atomistic molecular dynamics simulations can evaluate the adsorption energies of water molecules at the nanomaterial surface. Hydration free energies of the dissolved material molecules can be computed to predict the NM dissolution rates, using methodology

developed for prediction of free energy of solvation (Jämbeck et al, 2014). The charge and hydration energies of nanomaterials should generally be calculated at relevant temperatures (i.e. room or body temperature, 293 K, or 310 K, respectively and salt concentrations (pure water, physiological concentrations 100 mmol/L to 150 mmol/L) and pH values from 3 to 7, reflecting the condition in the lab or in different compartments of living organisms. For calculation of surface charge at different pH and salt concentrations, one can use the methods based on Poisson-Boltzmann mean field equation that includes charge regulation (Behrens et al., 1999;Behrens et al., 1999a).

7.3 Use of material models for support risk assessment

Modelling nanotoxicity is about predicting the risk due to the use of NM. Risk is defined as the probability that exposure to a hazard will lead to a negative consequence for the cell fate, or more simply, Risk = Hazard × Exposure. Hence modelling, in addition to hazard models, should include exposure models. Exposure models are intended to predict how NM evolve in the environment, including aggregation, and hence may harm human health and/or wildlife. Exposure models are intended to predict how NM evolve in the environment, including aggregation, and hence may harm human health and/or wildlife. NM exposure effects can be based on whole animal evaluations, cellular-level evaluations, or molecular-level evaluations. For example, whole animal evaluations could provide screening-level measurement using species of rat, mouse, zebrafish, and other animal models; cellular-level evaluations could have measures of different types of cell death; and molecular-level evaluations could include global gene expression, gene localization, and function (Harper et al. 2011).

7.4 Challenges: Multiscale modelling of bionano interface

In view of importance of the interactions at the bionano interface for initiation of AOPs and for systemic distribution of NMs, the NM characteristics directly addressing the interactions between NM and biomolecules are most informative. Although they may be not completely independent from the basic properties of the NM, as expressed by their intrinsic descriptors, a systematic evaluation of the descriptors for interactions may make predictive models much more compact and robust. Examples of such descriptors are: content of NM protein corona composition, adsorption enthalpy for an amino acid, lipid molecule, or a protein on the NM surface, hydrophobicity, production of ROS. All of these require a modelling of the NM in realistic environments.

The major challenge here is the need to use multiscale models for the characterisation of interactions. The relevant systems sizes of several nanometers are too large for direct atomistic simulation, so a coarse-grain description is required, which would be able to preserve information about the interaction specificity. In addition to this, the number of relevant molecules involved in the interactions with NM can be enormous, so the corona composition (i.e. list of proteins) as such may be an impractical property to be used for predictions. Each nanoparticle immersed in plasma may have its own unique corona (Dobrovolskaya et al, 2014). In comparison to this, protein abundances in the corona may reflect the properties of the NM that determine its propensity to bind certain type of

molecule. Therefore, one should aim for statistical descriptors of the proteins interacting with the NM.

In contrast to NMs, the development of descriptors for biomolecules is relatively straightforward due to their chemical uniformity, e.g. the same amino acids present in all proteins or nucleic acids in all DNA. For proteins, the simplest descriptors can be constructed using their amino acid (AA) sequence. These can include counts of amino acids of different types, net charge or total mass. Already this characterization is very rich and capable of predicting complex events at the bionano interface (Walkey et al., 2014;Kamath et al., 2015). Moreover, obtaining descriptors from AA sequences can be done by using a wide range of software tools such as the EMBOSS PepStats tool (Rice et al., 2000). More advanced descriptors for proteins can be built by analyzing their structure. In some cases, starting with the AA sequence of the protein the 3D structure of the molecule can be retrieved from the Protein Data Bank and then used to construct the descriptors. When the structure is not available, one can then use a structure prediction software. There are multiple automated tools available for this task, such as i-Tasser (Roy et al., 2010). Using the measured or predicted 3D structure of the protein, several advanced descriptors can be calculated. Lopez et al. developed a one-bead-per-amino acid (united atom – UA) model of globular proteins, which is suitable for this purpose (Lopez et al., 2015; Lopez et al., 2017). Some examples of advanced descriptors that can be calculated include protein globule dimensions (radius of gyration and hydrodynamic radius), aspect ratio, dipole moment, rotational inertia, dielectric constant, hydrophobicity, surface charge at different pH and salt concentrations. In addition, protein charge at different pH can be calculated using the Poisson-Boltzmann cell model with charge regulation as reported by Barroso da Silva et al. (Barroso da Silva et al., 2017).

For proteins, an evaluation of interaction properties requires an assumption about the protein structure at the conditions of interest. With the known 3D structure of the protein and the nanomaterial, bionano interaction descriptors can be systematically calculated based on how the proteins adsorb onto the surface of the NMs. While a calculation of the precise conformation of adsorbed molecules and a careful evaluation of ensemble averages is definitely a challenging task, several relevant quantities can be calculated using a simplified approach. To make the problem tractable, one can make two major approximations: assume additivity of the interactions between the building blocks of the biomolecule and the NM and neglect the change of conformation for adsorbed molecules. While these assumptions prevent one from obtaining accurate adsorption energies, they allow for a uniform screening of thousands of molecules and ranking them based on how strongly they will attach to the surface of the NM. This ranking represents a statistical measure of the content of the biomolecular corona and constitutes a unique fingerprint of a NP. Using the united atom protein model (Lopez et al., 2015), one can compute preferred adsorbed orientation and evaluate mean adsorption energy at different conditions. Moreover, using the same bottom-up construction approach, one can engineer an ultra-coarse-grained model (united amino acid - UAA) that closely reproduces the total protein-protein pairwise interaction energy profiles obtained in the united atom model. In the UAA model, one would typically need between 5 and 30 united-amino acid beads to capture the geometry and reproduce the adsorption characteristics of the original protein. This second coarse-graining can be

based on the mass distribution in the complete protein and can be optimized by tuning the protein diffusion coefficients to those obtained using UA model. The UAA model would be then suitable for modelling competitive protein adsorption and formation of protein corona (Poggio et al., 2017).

7.5 Challenges: Missing predictive models for some descriptors

In the mechanistic toxicity assessment paradigm, the NM properties should be related to the molecular and biological modes of action of the material. Such an approach is proposed, in particular, in the H2020 SmartNanoTox project. Then, the attention is focussed on the Molecular Initiating Events of the AOPs, triggered by the NMs interaction with the biological tissue. Where such MIEs are known, a calculation of the relevant descriptors is essential. Among the known candidate MIEs one can name production of ROS, cellular uptake, cell association, or lysosomal damage. ROS production and oxidative stress are known to be correlated with the conduction band gap for metal oxide NMs (Burrello et al., 2010;Ying et al., 2015). The models proposed in these latter works use reactivity descriptors to build the energy band structure of oxide nanoparticles and predicts their ability to induce an oxidative stress by comparing the redox potentials of relevant intracellular reactions with the oxides' electronic energy structure. At the same time, the descriptors for interactions of NMs with lipids, lung or cell membrane, or receptor proteins are missing. Supposedly, they can be constructed based on molecular interaction descriptors, using the multiscale methodology as described above, and hydrophobicity descriptors.

Another obviously missing property is NM dissolution rate, which is associated with (metal) ion release. Dissolution can be an important factor understanding the cellular response to a range of different NMs and has the potential to become a key component of a screening process for categorizing NMs with common hazard potential based on their potential to release ionic species. Several approaches to this problem are taken by SmartNanoTox project: (i) comparisons of bond energies with solvation energies for a given ion/atom/molecule (ii) kinetic models to assess the timescale of any dissolution (iii) biased MD simulations of free energy barriers to dissolution of NPs including surface reconstruction and change on contact with water, (iv) where appropriate direct MD studies of spontaneous dissolution and the influence of surface ligands and coronas. If successful, these approaches will lead to a molecular understanding of the relevant mechanisms of hazard and tractable predictive models for different nanoparticle/ligand/water systems. In addition, catalytic activity of NMs can be assessed in the first instance by calculating frontier orbitals for given NP systems by density functional theory and correlating them with experimental data to provide tractable expressions for use in assessing toxicological activity.

From the point of release, the state of the NM can change in many respects both before and after the contact with biological tissues. The affected properties may include oxidation, adsorption of foreign material from the atmosphere, waters or soil, partial removal of the engineered coating. The relevant descriptors are: time after release, temperature, coating quality (percentage of coverage), amount of pollutants.

7.6 Challenges: Coupling and linking models for predicting biological events

The ability of the NM to dissociate and produce reactive species, to affect the conformation of vital biomolecules, or interfere in metabolic or reproductive processes determines the NM's ability to cause hazardous effects. From a biological point of view, this can be explained as inducing MIEs leading to initiation of an AO. NM properties profoundly affect the molecular processes at the bionano interface. Thus, detailed characterization of the NM after initial contact with organism at different stages of the systemic transport can provide molecular level descriptors for "mechanism-aware" toxicity prediction schemes. Materials modelling along with experimental NM characterization after the contact can be used to develop the relevant NM descriptors. At the first level, such descriptors would include characterization of the interfacial NM contact with biomolecules in terms of binding energies of biomolecule elements (amino acids, lipid headgroups, etc.). Such descriptors should be organized in a bionano interactions database, which will be used for prediction of the NM corona formation including characterization of the corona outer surface, and prediction of likelihood of the particular hazardous effects. To finally develop the mechanism-aware QSARs one should perform systematic analysis of the NM-induced pathways and map the NM physicochemical properties to the MIE and thus to the specific AO for any NM. This approach is described in detail in Chapter 8. The overall assessment scheme thus will combine materials modelling, systems biology, *in vivo* and *in vitro* studies.

References

Behrens SH, Borkovec M (1999) Electrostatic Interaction of Colloidal Surfaces with Variable Charge. J. Phys. Chem. B 103: 2918.

Behrens SH, Borkovec M (1999a) Exact Poisson-Boltzmann solution for the interaction of dissimilar charge-regulating surfaces. Phys. Rev. E 60: 7040.

Benfenati E, Toropov AA, Toropova AP, Manganaro A, Gonella Diaza R (2011) CORAL software: QSAR for anticancer agents. Chem Biol Drug Des 77: 471-476

Brancolini G, Kokh DB, Calzolai L, Wade R, and Corni S (2012) Docking of Ubiquitin to Gold Nanoparticles. ACS Nano 6: 9863

Bigdeli A, Hormozi-Nezhad MR, Jalali-Heravi M, Abedini MR, Sharif-Bakhtiar F (2014) Towards defining new nano-descriptors: extracting morphological features from transmission electron microscopy images. RSC Advances 4: 60135-60143

Brandt E, Lyubartsev A (2015) Systematic Optimization of a Force Field for Classical Simulations of TiO2–Water Interfaces. J Phys. Chem. C 119: 18110-18125.

Brandt E, Lyubartsev A (2015a) Molecular Dynamics Simulations of Adsorption of Amino Acid Side Chain Analogues and a Titanium Binding Peptide on the TiO2 (100) Surface. J. Phys. Chem. C 119: 18126-18139.

Burello E, Worth AP (2011) A theoretical framework for predicting the oxidative stress potential of oxide nanoparticles. Nanotoxicology 5: 228-235

Barroso da Silva FL, Dias LG (2017) Development of constant-pH simulation methods in implicit solvent and applications in biomolecular systems. Biophys. Rev. (in press) DOI 10.1007/s12551-017-0311-5

Dekkers S, Ooman AG, Bleeker EAJ, Vandebriel RJ, Micheletti C, Cabellos J, Janer G, Fuentes N, Vazquez-Campos S, Borges T, Silva MJ, Prina-Mello A, Movia D, Nesslany F, Ribeiro AR, Leite PE, Groenewold M, Cassee FR, Sips AJAM, Dijkzeul A, van Teunenbroek T, Wijnhoven SWP (2016) Towards a nanospecific approach for risk assessment. Regulatory Toxicology and Pharmacology 80: 46-59

Ding F, Radic S, Chen R, Chen P, Geitner N, Brown J, and Ke P (2013) Direct observation of a single nanoparticle–ubiquitin corona formation. Nanoscale 5: 9162

Dobrovolskaia MA, Neun BW, Man S, Ye X, Hansen M, Patri AK, Crist RM (2014) Protein corona composition does not accurately predict hematocompatibility of colloidal gold nanoparticles. Nanomedicine: Nanotechnology, Biology and Medicine 10: 1453-1463

Gajewicz A, Puzyn T, Rasulev B, Leszczynska D, Leszczynski J (2011) Metal oxide nanoparticles: size-dependence of quantum-mechanical properties. Nanoscience & Nanotechnology – Asia 1: 53-58

Gajewicz A, Schaeublin N, Rasulev B, Hussain S, Leszczynska D, Puzyn T, Leszczynski J (2015) Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies. Nanotoxicology 9: 313-325

Harper SL, Dahl JA, Maddux BLS, Tanguay RL, Hutchison JE (2008) Proactively designing nanomaterials to enhance performance and minimise hazard. Int J Nanotechnol. 5: 124–142.

Jämbeck JPM, Lyubartsev AP (2014) J. Phys. Chem. 118: 3793-3804.

Jones DE, Ghandehari H, Facelli JC (2016) A review of the applications of datmining and machine learning for the prediction of biomedical properties of nanoparticles. Computer Methods and Programs in Biomedicine 132: 93-103

Kamath P, Fernandez A, Giralt F, Rallo R (2015) Predicting Cell Association of Surface-Modified Nanoparticles Using Protein Corona Structure - Activity Relationships (PCSAR). Curr. Top. Med. Chem. 15: 1930.

Kar S, Gajewicz A, Puzyn T, Roy K, Leszczynski J (2014) Periodic table-based descriptors to encode cytotoxicity profile of metal oxide nanoparticles: a mechanistic QSTR approach. Ecotoxicology and Environmental Safety 107: 162-169

Khan S, Gupta A, Nandi C (2013) Controlling the Fate of Protein Corona by Tuning Surface Properties of Nanoparticles. J. Phys. Chem. Lett. 4: 3747

Lam CW, James JT, McCluskey R, Hunter RL (2004) Pulmonary toxicity of single-wall carbon nanotubes in mice 7 and 90 days after intratracheal instillation. Toxicological Sciences 77: 126-134

Liu R, Zhang HY, Ji ZX, Rallo R, Xia T, Chang CH, Nel A, Cohen Y (2013) Development of structure-activity relationship for metal oxide nanoparticles. Nanoscale 5: 5644-5653

Lopez H, Lobaskin V, (2015) Coarse-grained model of adsorption of blood plasma proteins onto nanoparticles, J. Chem. Phys. 143: 243138

Lopez H, Brandt EG, Mirzoev A, Zhurkin D, Lyubartsev A, Lobaskin V (2017) Multiscale modelling of bionano interface. Adv Exp Med Biol. 947: 173-206

Lynch I, Weiss C, Valsami-Jones E (2014) A strategy for grouping of nanomaterials based on key physico-chemical descriptors as a basis for safe-by-design NMs. Nano Today 9: 266-270

Mills KC, Murry D, Guzan KA, Ostraat ML (2014) Nanomaterial registry: database that captures the minimal information about nanomaterial physico-chemical characteristics. Journal of Nanoparticle Research 16: 2219

Nanda KK (2012) Liquid-drop model for the surface energy of nanoparticles. Physics Letters A 376: 1647-1649

NRC (2009) Review of federal strategy for nanotechnology-related environmental, health, and safety research. Washington, DC: The National Academies Press

Palmer JC, Debenedetti PG (2015) Recent Advances in Molecular Simulations: A chemical Engineering Perspective. AIChE J. 61: 370–383.

Pathakoti K, Huang MJ, Watts JD, He X, Hwang HM (2014) Using experimental data of Escherichia coli to develop a QSAR model for predicting the photo-induced cytotoxicity of metal oxide nanoparticles. J Photochem Photobiol B 130: 234-240

Poggio S., Power D, Lopez H, Lobaskin V (2017) Bionano interactions: A key to mechanistic understanding of nanoparticle toxicity. (in press)

Puzyn T, Leszczynska D, Leszczynski J (2009) Toward the development of “Nano-QSARs”: advances and challenges. Small 5: 2494-2509

Puzyn T, Rasulev B, Gajewicz A, Hu X, Dasari TP, Michalkova A, Hwang HM, Toropov A, Leszczynska D, Leszczynski J (2011) Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. Nature nanotechnology 6: 175-178

Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16: 276-277

Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. 725-738

Sayes CM, Smith AP, Ivanov IV (2014) A framework for grouping nanoparticles based on their measurable characteristics. International Journal of Nanomedicine 8: 45-56

Sizochenko N, Rasulev B, Gajewicz A, Mokshyna E, Kuz’min VE, Leszczynski J, Puzyn T (2015) Causal inference methods to assist in mechanistic interpretation of classification nano-SAR models. RSC Advances 5: 77739-77745

Stone V, Pozzi-Mucelli S, Tran L, Aschberger K, Sabella S, Vogel U, Poland C, Balharry D, Fernandes T, Gottardo S, Hankin S, Hartl MGJ, Hartmann N, Hristozov D, Hund-Rinke K, Johnston H, Marcomi A, Panzer O, Roncato D, Saber AT, Wallin H, Scott-Fordsmand JJ (2014) ITS-NANO – Prioritising nanosafety research to develop a stakeholder driven intelligent testing strategy. Particle and Fibre Toxicology 11: 9

Toropova AP, Toropov AA, Maksudov SKh (2006) QSPR modelling mineral crystal lattice energy by optimal descriptors of the graph of atomic orbitals. Chemical physics Letters 428: 183-186

Toropov AA, Toropova AP, Benfenati E, Leszczynska D, Leszczynski J (2010) SMILES-based optimal descriptors: QSAR analysis of fulleren-based HIV-1 PR inhibitors by means of balance of correlations. J Comput Chem 31: 381-392

Toropov AA, Rallo R, Toropova AP (2015) Use of QUASI-SMILES and Monte Carlo optimization to develop quantitative feature property/activity relationships (QFPR/QFAR) for nanomaterials. Current Topics in Medicinal Chemistry 15: 1837-1844

Toropova AP, Toropov AA, Manganelli S, Leaone C, Baderna D, Benfenati E, Fanelli R (2016) Quasi-SMILES as a tool to utilize eclectic data for predicting the behavior of nanomaterials. NanoImpact 1: 60-64

Walkey CD, Olsen JB, Song F, Liu R, Guo H, Olsen DWH, Cohen Y, Emili A, Chan WCW (2014) Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles. ACS Nano 8: 2439–2455.

Wang XZ, Yang Y, Li R, Mcguinnes C, Adamson J, Megson IL, Donaldson K (2014) Principal component and causal analysis of structural and acute in vitro toxicity data for nanoparticles. Nanotoxicology 8: 465-476

Ying J, Zhang T, Tang M (2015) Metal oxide nanomaterial QNAR models: available structural descriptors and understanding of toxicity mechanisms. Nanomaterials 5: 1620-1637

8. Data Analysis: Nanobioinformatics

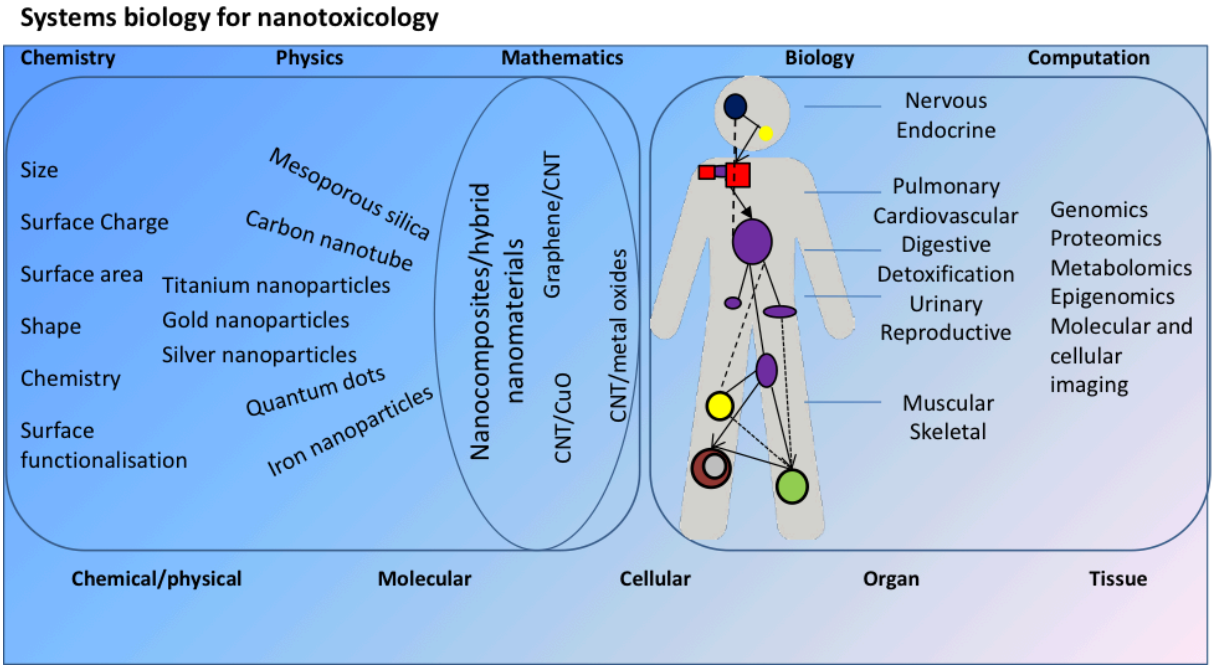
Sabina Halappanavar[2,3], Penny Nymark[4,5], Roland Grafström[4,5]

2. Environmental Health Science and Research Bureau, Health Canada, Ottawa, Canada
3. Department of Biology, University of Ottawa, Ottawa, Canada
4 Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
5 Misvik Biology, Turku, Finland

Conventional human health risk assessment (HHRA) approaches, on which the chemical regulatory system is founded, involve chronic or subchronic animal exposures and the targeted analyses of adverse effects such as cancerous tumours or non-cancer effects of regulatory importance. However, these assays are time and cost-intensive and require the prior knowledge of mode of action of a toxicant. Moreover, most of the chronic exposure models use maximum tolerate dose and thus lack broader application. The pace at which technology is evolving, new substances or chemicals are being regularly added to the market, which require rapid screening for their safety. For most part, the type of toxicity induced by novel substances is not known, and due to the time and cost burden associated with the conventional testing, timely screening of novel chemicals for the potential hazard is not possible. Thus, newer approaches that significantly reduce time and cost required to complete the assessment of a chemical for its potential toxicity, yet providing comprehensive understanding of the underlying mode-of-action of the toxicity are constantly being sought.

A comprehensive understanding of the toxicity induced by nanomaterials will require a comprehensive appreciation of material physics and chemistry along with their

anticipated behaviour at various levels of biological organisation including molecular, cellular, organ, and tissue levels as shown in the Figure 8-1(modified from ref. 8-1). Integration of the information derived from these various levels using statistical, mathematical and bioinformatics tools is the key to understanding the overall complexity of the biological responses induced by this novel class of materials and for their effective regulation (ref. 8-1, 8-2).



With the advent of novel molecular techniques, biological data is being generated at a phenomenal pace. Sophisticated tools collectively known as 'omics' that can generate exhaustive inventories of molecular entities such as genes (transcriptomics), proteins (proteomics), small biomolecules (metabolomics), and biological networks (bioinformatics) in normal homeostasis condition and how these entities change under stress or during a disease process have been developed. Genome-scale sequencing tools have resulted in a renaissance of big data enabling visualisation of genetic landscape that is perturbed following a substance exposure. Consequently, the need for machines/computers that can enable handling, organisation and curation of large datasets has become inevitable. Mathematical models and statistical algorithms have been developed to understand how the various molecular entities interact with one another and their relationship with the observed phenotype, i.e. cellular toxicity or disease process.

Figure 8-1 shows various types of data that are used in bioinformatics or systems biology approaches, the 'omics' platforms available for genome-wide profiling and how integration of the various layers of omics data can enhance understanding and appreciation of the biology at action during normal and disease states in an organism, enabling holistic understanding (systems level) of the perturbed system. In general, the omics data can be categorised into three individual categories: components, interactions and functional states data (ref. 8-3). Components data provide individual catalogues of molecular entities such as genes, proteins, lipids, metabolites etc. that are differentially

expressed. Interactions data provide details on how these individual entities interact within a biological space and functional state data incorporates data from all 'omics' platforms and interactions data to reveal the cellular state or phenotype of an organism following a challenge.

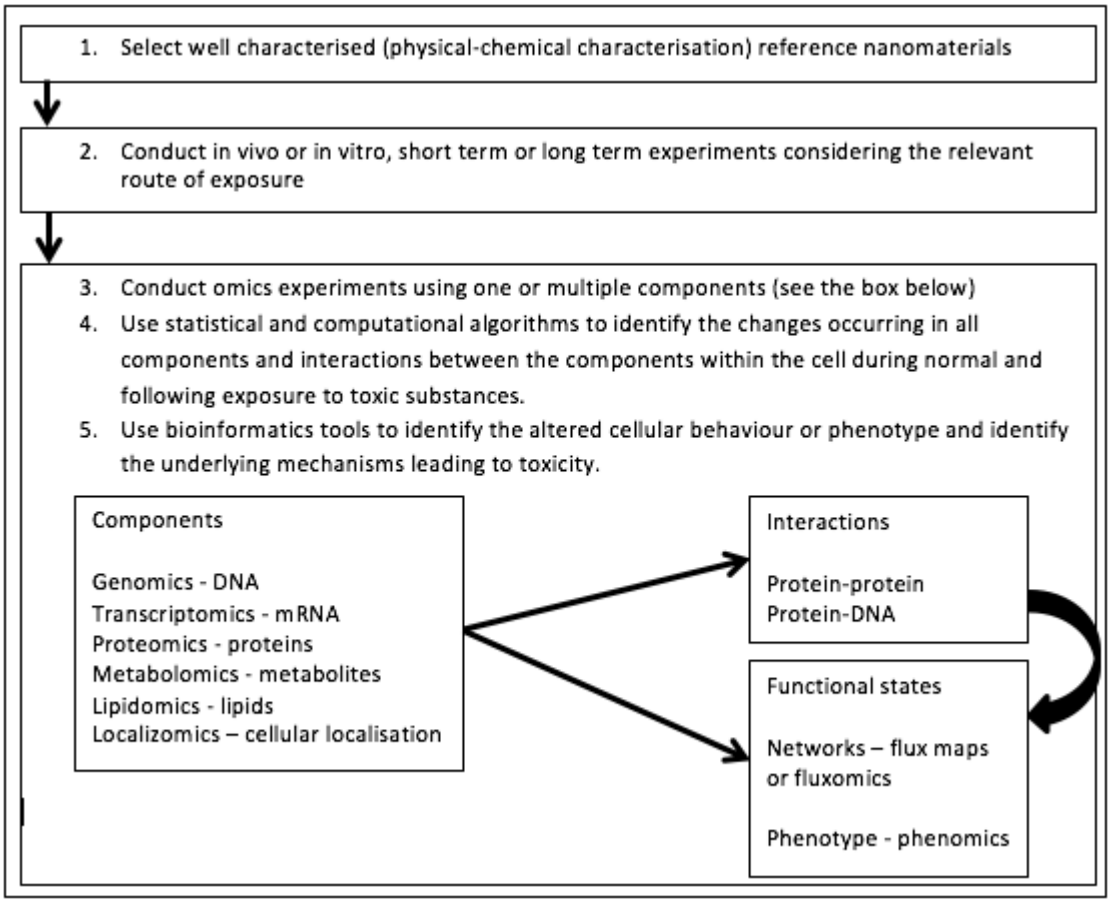
Table 8-1 (modified from Ref. 8-4) lists various omics platforms available and a brief explanation of the type of data that they generate.

Omics Platforms	
Genomics	Genome is the 'blue print' that holds information on the structure and function of an organism that is encoded in the DNA (genetic material), organised in subunits of individual genes. Genomics is the study of this blue print - genes and the interaction between them. Variations in gene sequences due to mutations can influence the organisms' response to a stressor and alter its susceptibility to diseases.
Transcriptomics	The transcriptomics is the study of the complete set of RNA transcripts produced by the genome at a given time during development, normal homeostasis or disease states. Transcriptome is highly sensitive to the changing internal and external environment and thus, transcriptomic changes accurately reflect the organisms' response to endogenous and extrinsic stimuli.
Proteomics	The proteins are functional units of genes. The proteomics is the study of the full set of proteins encoded by a genome enabling their identification and quantification during normal homeostatic and following exposures to stressors. The proteome helps understand the functional impact of altered transcriptome linking the gene expression changes to a phenotype (Phenome).
Metabolomics	Metabolomics is the study of metabolites (low molecular weight) present in biological fluids, cells and tissues. Altered levels of metabolites are good indicators of altered physiological states following exposures to stressors and thus, are used as sensitive markers of exposure and/or effects in biomonitoring and surveillance studies.
Epigenics	Epigenetics is the study of changes in gene expression that are not the consequence of changes in DNA sequence. It is the study of chromatin and the effects of RNA interference on transcription. Chemical modifications to DNA or DNA-associated proteins involved in DNA packaging (chromatin) are one of the epigenetic mechanisms and methylation of DNA is one of the epigenetic endpoints commonly studied. Epigenetic changes are heritable, and are

	influenced by the environmental processes, environmental exposures.
Microbiome	The term 'microbiome' refers to a group of microorganisms in a given environment. The study of taxonomic and functional changes to the composition of the microbiome and its impact on human health and disease is a rapidly evolving field in toxicology. Multi-omic technologies and advances in the computational and bioinformatics tools are playing an important role in advances in this field.

However, considering the ever-growing list of nanomaterials and the next generation hybrid nanomaterials appearing on the market, the comprehensive testing 'omics' tools are not sustainable. Thus, a strategy involving few representative or reference classes of nanomaterials of diverse physical –chemical properties should be queried in an organised and systematic manner using the 'omics' tools in Figure 8-2.

Figure 8-2: Experimental work flow and the information generated



The resulting data can then be used to inform various components of human health risk assessment process including (ref. 8-5),

1. To identify hazard induced by toxic substances, thereby informing mechanisms-of-action or mode of action.
2. To build adverse outcome pathways identifying causally linked molecular changes that result in disease development.
3. To support the design and development of targeted mechanisms-based in vitro assays that eventually form the basis of predictive toxicology tools.
4. Identification of candidate markers of exposure or effects that can inform biomonitoring and surveillance activities
5. To identify critical effect levels – derivation of transcriptomics/pathways-driven point of departure using dose-response modelling.
6. To support weight of evidence (for data poor chemicals, omics data can be used to link the exposure to an effect).
7. To build gene/protein signatures that can be used to classify group of chemicals based on their genomic response.
8. To prioritise substances that may need further toxicity assessment by other methods.

Transcriptomics – a case study in bioinformatics

Of all the tools, gene expression profiling or transcriptomics (measures changes in the coding or non-coding RNA in cells or tissues following exposure to a substance) tools have been the most advanced. Due to the mature microarray and sequencing technologies, the broad annotation level of genes, and the availability of statistical software for reliable and reproducible analyses of the large data generated, transcriptomics is extensively applied to identify chemicals' mode of action. In the context of nanomaterials, a combination of gene and protein expression profiling and bioinformatic analyses have been applied to elucidate the mechanisms by which nanomaterials induce pulmonary toxicity at an occupationally relevant dose (ref. 8-6, 8-7, 8-8); to identify potential biomarkers of pulmonary effects induced by nanomaterials (ref. 8-9, 8-10, 8-11); characterize repercussions of local inflammation (lungs) on other secondary tissues (e.g., heart and liver) following nanomaterial exposure; and validate the relevance of in vitro data to predicting in vivo responses to NM exposure (ref. 8-12, 8-13, 8-14). Moreover, a database of toxicity fingerprints that are specific to lung diseases (ref. 8-15, 8-16) and computational tools that can be used to predict the toxicity of new ENMs that have yet to undergo experimental testing (ref. 8-15, 8-16) have been developed. More recently, Labib et al.(ref. 8-17) demonstrated how transcriptomics data can be used in an adverse outcome pathway framework to identify the most relevant pathways or networks of interest to a disease, and strategies that can be used to calculate pathway dose-response that can be then used for calculating critical effect levels. In addition, predictive tools developed based on chemical toxicity are worth attention, since toxicological responses can be expected to be comparable on a mechanistic level. For example, an omics-based description of toxicological responses that broadly captures and accurately predicts liver toxicity on both cellular and organismal level was recently described (ref. 8-18). The so called Predictive Toxicogenomics Space covers several toxicity-associated mechanisms such as oxidative stress, cell cycle disturbances, DNA damage response and mitochondrial dysfunction, commonly also associated with ENM (reviewed in ref. 8-19). In another study, a framework for predicting the hazards associated with complex mixtures of

chemicals using single-chemical transcriptomics data was established (ref. 8-20). Thus, applicability of transcriptomics not only to identify the subtle biological effects induced by low doses of nanomaterials very early after the exposure but also in risk characterisation of nanomaterials has been well demonstrated.

Although regulatory acceptance of transcriptomics data is not yet achieved, a lot of efforts are being made to harmonise the protocols and data analyses methods. Guidance documents and development of standards are being established. A committee for the 'application of genomics to mechanisms-based risk assessment' is established by the ILSI/HESI. OECD has established Molecular Screening and Toxicogenomics advisory group and have initiated efforts to harmonise genomics approaches for risk assessment. The European Chemicals Agency have also initiated discussion among academia, regulators and industry on the implementation of new approach methodologies (NAMs) into regulations such as REACH (ref. 8-21). However, for now, the data can be effectively used to inform chemicals' mode of action, identify important events relevant to disease progression and in the development of mechanisms-based High throughput screening in vitro assays that are predictive of in vivo responses. Moreover, for data poor substances such as nanomaterials, the data can be used as weight of evidence, and for screening or prioritising nanomaterials for further testing.

8.2 Challenges moving forward

While a tremendous progress has been made in the area of transcriptomics, several challenges lie ahead. Prior to its routine inclusion in safety testing of substances and acceptance in regulatory science, standard operating protocols have to be developed; data reporting and data analysis standards have to be established, quality check and quality control standards have to be defined, analysis algorithms have to be developed and standardised, and internationally harmonised guidelines have to be developed. The regulatory acceptance criteria have to be developed and areas of regulatory applications have to be identified. Appropriate training courses to analyse and interpret transcriptomics data in a consistent manner have to be established. In addition, appropriate data management strategies are a fundamental requirement for efficient nanobioinformatics. Databases for storing omics data in standardized formats are available and provide access to ENM-associated omics data. However, metadata and associated toxicological and physico-chemical data requires ENM-specific databases capable of linking to the external omics databases. An example of such a database is the eNanoMapper database (ref. 8-22). This will enable linked and annotated (using ontologies as outlined in Section 5 of this report) buildup of transcriptomics data for reference substances, useful in further nanobioinformatics modeling approaches.

Other challenges involve data, tools, software and model sharing. Although some published datasets are deposited in the public repositories and are accessible, the reporting formats for ENM and their associated toxicity and physico-chemical data are not standardised for uptake and analysis by other researchers. Transcriptomics is one of the extensively tested and applied genome-wide profiling tools, although standards are yet to be developed for data analysis and data representation. Transcriptome profiling can involve different microarray platforms and based on the statistical algorithms used, the interpretation of the data can vary from laboratory to laboratory. Thus consistency,

reproducibility and reliability are the major issues that need to be tackled and may be addressed to some extent within the nanosafety community by the establishment of consistently tested reference ENM data sets.

8Application of other 'omics' data to nanotoxicology

Although, due to the methodological limitations and large diversity of proteins and metabolites within the biological samples, not applied as extensively as transcriptomics, data derived from other 'omics' platforms such as, proteomics and lipidomics have been used to gain understanding of the underlying mechanisms of nanomaterial induced toxicity. Multi-omics approach involving lipidomics, proteomics and transcriptomics was applied to derive an understanding of carbon nanotube induced toxicity (ref. 8-23, 8-24, 8-25). A redox proteomics approach was proposed as first tier screening method for prioritisation of nanomaterials for further testing (ref. 8-26). Thus, each omics platform will provide a unique perspective of the changing phenotype, and development and validation of tools that aid in managing, processing and integration of multi-platform data towards biologically meaningful interpretation of the observed changes will be the key.

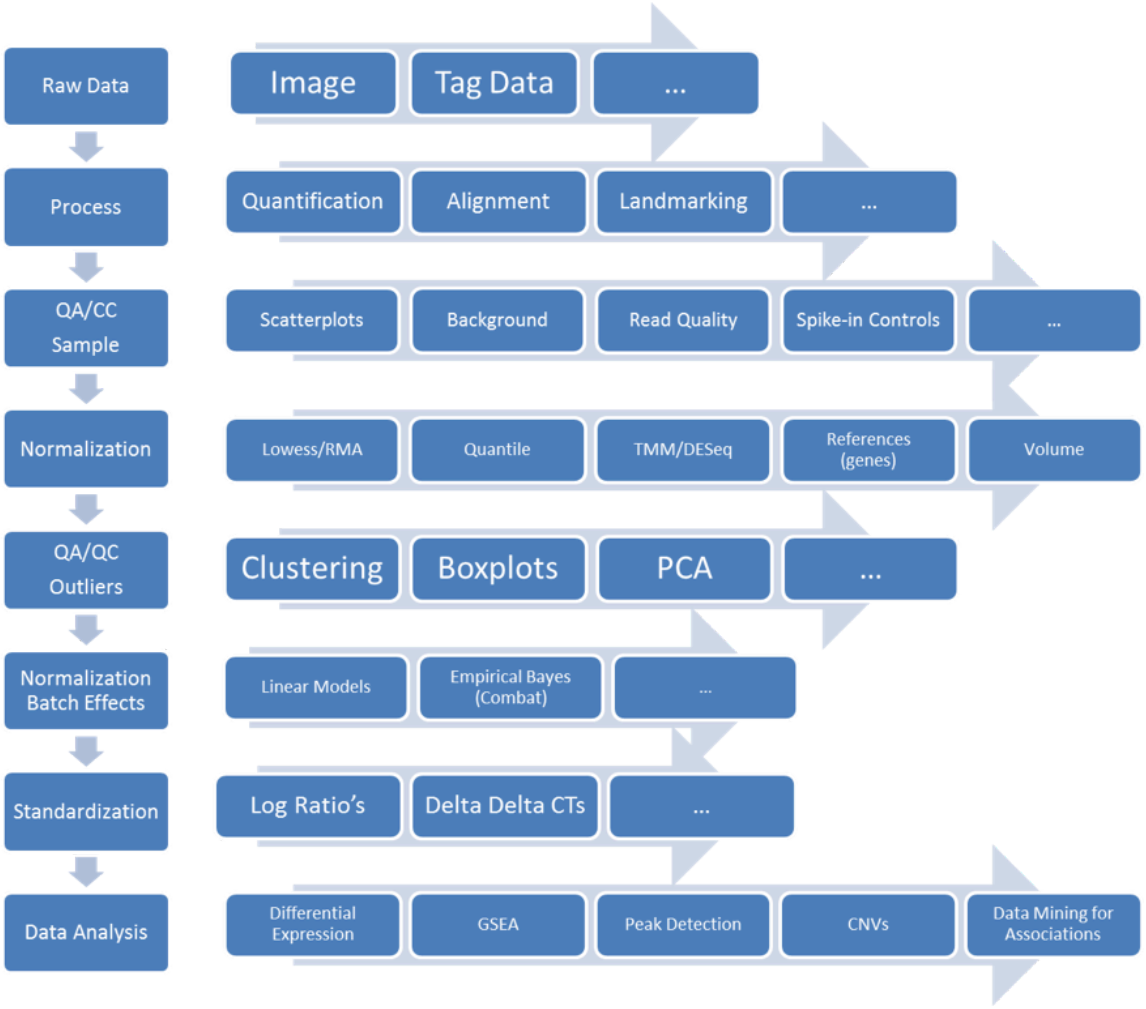
OMICS DATA ANALYSIS METHODS

Dario Greco[1], Andrew Williams[2], Sabina Halappanavar[2,3], Penny Nymark[4,5], Pekka Kohonen[4,5]

Dario Greco affiliation?
2. Environmental Health Science and Research Bureau, Health Canada, Ottawa, Canada
3. Department of Biology, University of Ottawa, Ottawa, Canada
4 Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
5 Misvik Biology, Turku, Finland

As stated above, the key to obtaining biologically relevant results from the microarray studies is the stringent and accurate analysis of large and complex datasets using appropriate statistical and bioinformatics methods. The Fig 8-3 shows the steps involved in analysing the 'omics' data in general.

Figure 8.3 Flow chart of data analysis



For many omic technologies and platforms several analytical steps are conceptually common. First, the raw data files must be read into the software environment, the quality of the raw data needs to be evaluated in order to ensure that technically suboptimal data points are excluded. Next, the data preprocessing, consisting mainly of normalization and batch effect evaluation and correction are carried out. Primary normalization and data filtering for factors contributing to variation such as differences in dye incorporation, hybridization efficiencies, etc. within arrays and across arrays will enable identification of differentially expressed genes or proteins. Handling batch effects successfully is largely accepted to be a crucial aspect of omics data analysis, but is unfortunately still neglected and poorly documented in many published studies (ref. 8-27, 8-28, 8-29) . However, as current microarray and RNA-seq platforms have a relatively good level of technical reproducibility, the largest sources of bias in experiments tends to be the biological material itself (ref. 8-30). Known biases such as, cell culture growth batches can be modelled as long as a balanced experimental design has been employed, e.g., using the limma linear modelling or general linear modelling framework. Since omics experiments are derived from complex protocols consisting of multiple steps, the probability to introduce unwanted bias, which is not otherwise corrected by data normalization, remains high. Several normalization methods are available and the choice of one over the others depends on intrinsic properties of the omics technology used and on the experimental design. The scientific community has largely converged on the use of methods and tools implemented in the R programming

language as it is free and publicly available. Bioconductor provides tools for the analysis of high-content genomic data and is open source and open development (www.bioconductor.org). A few of the widely used normalization methods include, locally weighted scatterplot smoothing (LOWESS) or data-driven LOWESS, and robust multiarray analysis (RMA).

Typically, the identification of the responding molecular species to a specific exposure is carried out by using univariate statistical methods that aim at testing each molecular feature in the data set individually (ref. 8-31). Upon the definition of likelihood (usually p-values) and magnitude (fold changes) of the molecular alterations, the features that are significantly responding to a given exposure are identified and lists of e.g. differentially expressed genes (in the case of transcriptomics) are compiled. In transcriptomics data analysis, a number of methods have been proposed, of which linear models followed by eBayes testing gained enormous popularity (ref. 8-32) Since microarray analysis involves multiple comparisons, false positives are very common and thus, tests such as the moderated t-tests were developed specifically for microarray analysis. The p-values from the statistical test are then adjusted either using the false discovery rate (FDR) correction to minimize the number of false positives or by controlling the Family-wise error rate (FWER) for example with Bonferroni correction. A false discovery rate adjusted p-value of less than 0.05 and a fold- change cut-off of 1.5 in either direction are routinely applied to the microarray datasets. The resulting stringent list of differentially expressed genes or proteins is then queried to identify altered functional pathways. Advanced statistical techniques such as hierarchical clustering, K-means clustering, self-organising maps enable identification of similar expression patterns across the samples, signatures specific to a class of chemicals, tissue or a cell type or a phenotype. The various statistical methodologies used to analyse the big data are summarised in Section 6.

In toxicogenomics, efforts establishing reproducible data analysis frameworks that are communicable to regulators are currently being established. The MAQC consortium accessed the technical performance and application of ‘omics technologies for clinical application and safety assessment have been investigated. The consortium completed three projects evaluating the performance of microarrays, genome-wide association studies and RNA-sequencing, with particular reference to the reproducibility of transcriptomics data, between-experiment concordance, within-laboratory repeatability, and cross-platform reproducibility. The results from these studies indicate that using a p-value and a fold change threshold and subsequently sorting by the fold-change to identify the most prominent differentially expressed genes enhanced reproducibility of the results while balancing the sensitivity and specificity. The work of the consortium has advanced microarray and RNA-seq analytical pipelines that can be leveraged for developing data analysis frameworks and best practices (ref. 8-33). However, it should be also considered that, given the complex nature of the molecular interactions, multivariate analysis could help highlighting additional sets of molecular features that might not be strongly associated to exposure effect when considered independently (ref. 8-34, 8-35, 8-36). In this sense, multivariate approaches relying on machine learning algorithms can also aid the finding of molecular biomarkers with toxicity predictive value to be further implemented in high-throughput targeted assays.

The primary readout of omics experiments usually consists of lists of molecular features significantly altered due to an exposure. To further facilitate the interpretation of these results, the molecules (genes, proteins, or metabolites) are mapped onto existing pathway databases and gene ontologies. Eventually, the goal is to anchor the expression changes at the gene or protein levels to the observed phenotype in an organism. A single gene or protein may be involved in multiple functions and therefore identifying isolated groups of genes or proteins that are differentially expressed may not be sufficient to understand the perturbed biology. Software tools for the systematic annotation of gene interactions derived from the literature are available. Classification systems such as gene ontology tools help identify categories of molecules that are altered following exposure. Kyoto Encyclopedia of Genes and Genomes, Gene Microarray Pathway Profiler, Ingenuity Pathway Analysis or WikiPathways tools can be used to identify pathways and functions that are perturbed following exposure to substances in experimental models. Although these literature-based tools often provide network representations of co-citation relationships, they are not really providing any regulatory gene network inference capability.

The statistical evaluation of the pathway and ontology over-representation is usually performed either by a hypergeometric test or a Kolmogorov-Smirnov test. Many tools are freely available online for carrying out this task, which is typically performed by uploading, for instance, a list of differentially expressed genes onto a web service and retrieving lists of significantly enriched biological themes. It should be noted that these services do not always include updated version of the pathways and ontologies definitions, risking introduction of bias in the outcome (ref. 8-37) . A robust approach that considers the complexity of biology and avoids testing isolated genes for significance is gene set enrichment analysis (GSEA). The method determines whether a priori defined sets of genes, such as pathways or gene ontologies, are statistically over-represented in relation to genes outside the pathway when compared to an exposure control (ref. 8-38, 8-39). These methods can be assumed to allow better comparison between diverse omics data sets (ref. 8-40, 8-41). Furthermore, the results are then useful for omics-based scoring methods, which can be used for predictive modelling (ref. 8-42, 8-43). As stated early in the section, omics data can be used to construct AOPs (ref. 8-20) and mechanistic descriptions of key events are being incorporated within a broader biological / toxicological context. GSEA using toxicity-predictive gene sets can be used to evaluate quantitatively such key events.

In recent years, multi-omics approaches have been used in a number of biomedical fields. The aim in this type of analyses is to portray a more comprehensive landscape of a biological state of interest by interrogating multiple molecular compartments from the same biological system. Computational methods specifically addressing multi-omics modeling have been proposed (ref. 8-44, 8-45, 8-46), but this approach is still under-used in nanotoxicology with only a few studies on multi-walled carbon nanotubes (ref. 8-6, 8-47, 8-48, 8-49).

Omics analysis is normally referred to as a high-content analysis, where few samples are tested for a high number of parameters (e.g. genes) and is relatively slow and costly. However, reduced sets of toxicity-associated genes can be assayed at higher throughput

and lower cost, e.g., Luminex® or more recently TempO-seq (RASL-seq) targeted RNA sequencing technology (ref. 8-50). To the benefit of the nanoinformatics community, high-throughput transcriptomics platforms are in development, e.g., in the LINCS and the Tox21 Phase III projects, and enable rapid gene profiling experiments with both several doses and biological replicates using multiple models of 800–1500 genes (reviewed in ref. 8-51). Although, NM effects analyzed using traditional microarrays, such as Agilent or Affymetrix GeneChips®, form the basis for most existing gene profiling analyses of NMs and provide reference values for recent next-generation sequencing and future generation of HTS data from selected toxicity-reflective gene sets.

There is also a clear need to develop new technologies and incorporate novel data streams for human health risk assessment. For example, applying toxicogenomics to characterize the biological responses to exposures to nanomaterials and evaluate possible dose-response relationships (ref. 8-17, 8-52, 8-53). Software such as BMDExpress provides an opportunity to conduct such analyses (ref. 8-54). Benchmark dose analysis along with multivariate technics such as GSEA (ref. 41) to derive the most sensitive enriched pathway as well as the overall median BMD value for key gene members of significantly enriched pathways, provide good estimates of the most sensitive apical endpoint benchmark dose (ref. 8-55, 8-56).

References Section 8

8-1: Halappanavar et al, Wiley Interdiscip Rev Nanomed Nanobiotechnol. 2017 Mar 15; doi: 10.1002/wnan.1465.
8-2: Riebeling C et al., Adv Exp Med Biol. 2017;947:143-171
8-3: Andrew Joyce et al Nature reviews 2006
8-4: DeBord GD et al., American Journal of Epidemiology. Vol. 184, No. 4
8-5: Sauer et al, Regul Toxicol Pharmacol. 2017 Sep 18. pii: S0273-2300(17)30292-1.
8-6: Halappanavar S, et al., Environ Mol Mutagen. 2015 Mar;56(2):245-64
8-7: Feliu N et al., ACS Nano. 2015 Jan 27;9(1):146-63.
8-8: Costa P et al.,2017 Journal of applied toxicology
8-9: Halappanavar S, et al., Environ Mol Mutagen. 2011 Jul;52(6):425-39. doi: 8-10: 10.1002/em.20639. Epub 2011 Jan 21
8-10: Saber AT et al., Wiley Interdiscip Rev Nanomed Nanobiotechnol. 2014 Nov-Dec;6(6):517-31
8-11: Guo NL, et al., J Toxicol Environ Heal Part A. 2012;75:1129–53
8-12: Nymark P et al, Nanotoxicology. 2015;9(5):624-35.
8-13: Jackson P, Hougaard KS, Vogel U, Wu D, Casavant L, Williams A, Wade M, Yauk CL, Wallin H, Halappanavar S. Exposure of pregnant mice to carbon black by intratracheal instillation: toxicogenomic effects in dams and offspring. Mutat Res. 2012 Jun 14;745(1-2):73-83. doi: 10.1016/j.mrgentox.2011.09.018. Epub 2011 Oct 6
8-14: Kinaret P et al., ACS Nano. 2017 Apr 25;11(4):3786-3796
8-15: Nikota J, Williams A, Yauk CL, Wallin H, Vogel U, Halappanavar S*. Meta-analysis of transcriptomic responses as a means to identify pulmonary disease outcomes for engineered nanomaterials. Part Fibre Toxicol. 2016 May 11;13(1):25.

8-16: Andrew Williams and Sabina Halappanavar. Application of biclustering of gene expression data and gene set enrichment analysis methods to identify potentially disease causing nanomaterials. Beilstein Journal of Nanotechnology – Invited contribution to a special edition on Nano Bioinformatics. 2015. Beilstein J Nanotechnol. 2015 Dec 21;6:2438-48.
8-17: Labib S et al., Part Fibre Toxicol. 2016 Mar 15;13:15
8-18: Kohonen P, Parkkinen JA, Willighagen EL, Ceder R, Wennerberg K, Kaski S, Grafström RC.,Nat Commun. 2017 Jul 3;8:15932. doi: 10.1038/ncomms15932.
8-19: Nymark P, Kohonen P, Hongisto V, Grafström RC. TOXIC AND GENOMIC INFLUENCES OF INHALED NANOMATERIALS AS A BASIS FOR PREDICTING ADVERSE OUTCOME. AnnalsATS. 2017. In press
8-20: Labib S et al., Arch Toxicol. 2017 Jul;91(7):2599-2616
8-21: ECHA (European Chemicals Agency). 2016. Topical Scientific Workshop on New Approach Methodologies in Regulatory Science. 19pp. European Chemicals Agency, Helsinki, Finland. Available from: http://echa.europa.eu/documents/10162/22049802/tsws_background_document_en.pdf
8-22: Jeliazkova N, Chomenidis C, Doganis P, Fadeel B, Grafström R, Hardy B, Hastings J, Hegi M, Jeliazkov V, Kochev N, Kohonen P, Munteanu CR, Sarimveis H, Smeets B, Sopasakis P, Tsiliki G, Vorgrimmler D, Willighagen E.
Beilstein J Nanotechnol. 2015 Jul 27;6:1609-34. doi: 10.3762/bjnano.6.165. eCollection 2015.
8-23: Shvedova et al 2012 Toxicol Appl Pharmacol. 2012 Jun 1;261(2):121-33
8-24: Tyurina YY et al., ACS Nano. 2011 Sep 27;5(9):7342-53
8-25: Teeguarden JG et al., Toxicol Sci. 2011 Mar;120(1):123-35.
8-26: Riebeling C et al., Toxicol Appl Pharmacol. 2016 May 15;299:24-9.
8-27: Chen C et al., PLoS One. 2011 Feb 28;6(2):e17238
8-28: Leek JT et al., Nat Rev Genet. 2010 Oct;11(10):733-9.
8-29: Goh WWB et al., Trends Biotechnol. 2017 Jun;35(6):498-507
8-30: Hansen KD, Wu Z, Irizarry RA, Leek JT.. Nat Biotechnol. 2011 Jul 11;29(7):572-3
8-31: Saeys Y et al., Bioinformatics. 2007 Oct 1;23(19):2507-17.
8-32: Ritchie ME et al., Nucleic Acids Res. 2015 Apr 20;43(7):e47
8-33: Sauer UG et al., Regul Toxicol Pharmacol. 2017 Sep 18. pii: S0273-2300(17)30292-1
8-34: Inza I et al., Artif Intell Med. 2004 Jun;31(2):91-103.
8-35: Kursa MB BMC Bioinformatics. 2014 Jan 13;15:8.
8-36: Abeel T et al., Bioinformatics. 2010 Feb 1;26(3):392-8.
8-37: Wadi L et al., Nat Methods. 2016 Aug 30;13(9):705-6.
8-38: Williams A and Halappanavar S. Beilstein J Nanotechnol. 2015 Dec 21;6:2438-48.
8-39: Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl Acad.Sci. USA 102, 15545–15550 (2005).
8-40: Rahmatallah, Y. et al. Brief Bioinform. 17, 393-407, (2016)
8-41: Williams A and Halappanavar S. Beilstein J Nanotechnol. 2015 Dec 21;6:2438-48.
8-42: Grafstrom RC, Nymark P, Hongisto V, Spjuth O,Ceder R, Willighagen E, Hardy B, Kaski S,Kohonen P. Altern Lab Anim 2015, 43:325–332.
8-43: Kohonen P, Parkkinen JA, Willighagen EL, Ceder R, Wennerberg K, Kaski S, Grafström RC.Nat Commun. 2017 Jul 3;8:15932.

8-44: FISCH KM et al., Bioinformatics. 2015 Jun 1;31(11):1724-8
8-45: Meng C et al., BMC Bioinformatics. 2014 May 29;15:162
8-46: Yang Z and Michailidis G Bioinformatics. 2016 Jan 1;32(1):1-8
8-47: Nymark P, Wijshoff P, Cavill R, van Herwijnen M, Coonen ML, Claessen S, Catalán J, Norppa H, Kleinjans JC, Briedé JJ. Nanotoxicology. 2015;9(5):624-35. doi: 10.3109/17435390.2015.1017022.
8-48: Dymacek J, Snyder-Talkington BN, Porter DW, Mercer RR, Wolfarth MG, Castranova V, Qian Y, Guo NL.Toxicol Sci. 2015 Mar;144(1):51-64. doi: 10.1093/toxsci/kfu262.
8-49: Snyder-Talkington BN, Dong C, Porter DW, Ducatman B, Wolfarth MG, Andrew M, Battelli L, Raese R, Castranova V, Guo NL, Qian Y.J Toxicol Environ Health A. 2016;79(8):352-66. doi: 10.1080/15287394.2016.1159635
8-50: Grimm FA et al, Green Chem. 2016 Aug 21;18(16):4407-4419.
8-51: Collins AR, et al, Wiley Interdiscip Rev Nanomed Nanobiotechnol. 2017 Jan;9(1).
8-52: Moffat I et al., Crit Rev Toxicol. 2015 Jan;45(1):1-43
8-53: Chepelev NL et al., Crit Rev Toxicol. 2015 Jan;45(1):44-52.
8-54: Yang L et al., BMC Genomics. 2007 Oct 25;8:387.
8-55: Dean JL et al., Toxicol Sci. 2017 May 1;157(1):85-99
8-56: Farmahin R et al., Arch Toxicol. 2017 May;91(5):2045-2065

9. The community: Overview of Stakeholders

Andrea Haase[1] Kai Paul4
[1] German Federal Institute for Risk Assessment, Germany
[2] School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, B15 2TT Birmingham, United Kingdom
[3]Greendecision Srl.
4 Blue Frog Scientific Limited, Quantum House, 91 George Street, Edinburgh, EH2 3ES, United Kingdom

Different nanoinformatics stakeholders may be identified and described via different approaches. One approach is based on the data life cycle (Figure 7) as described by Harper et al. (2013).

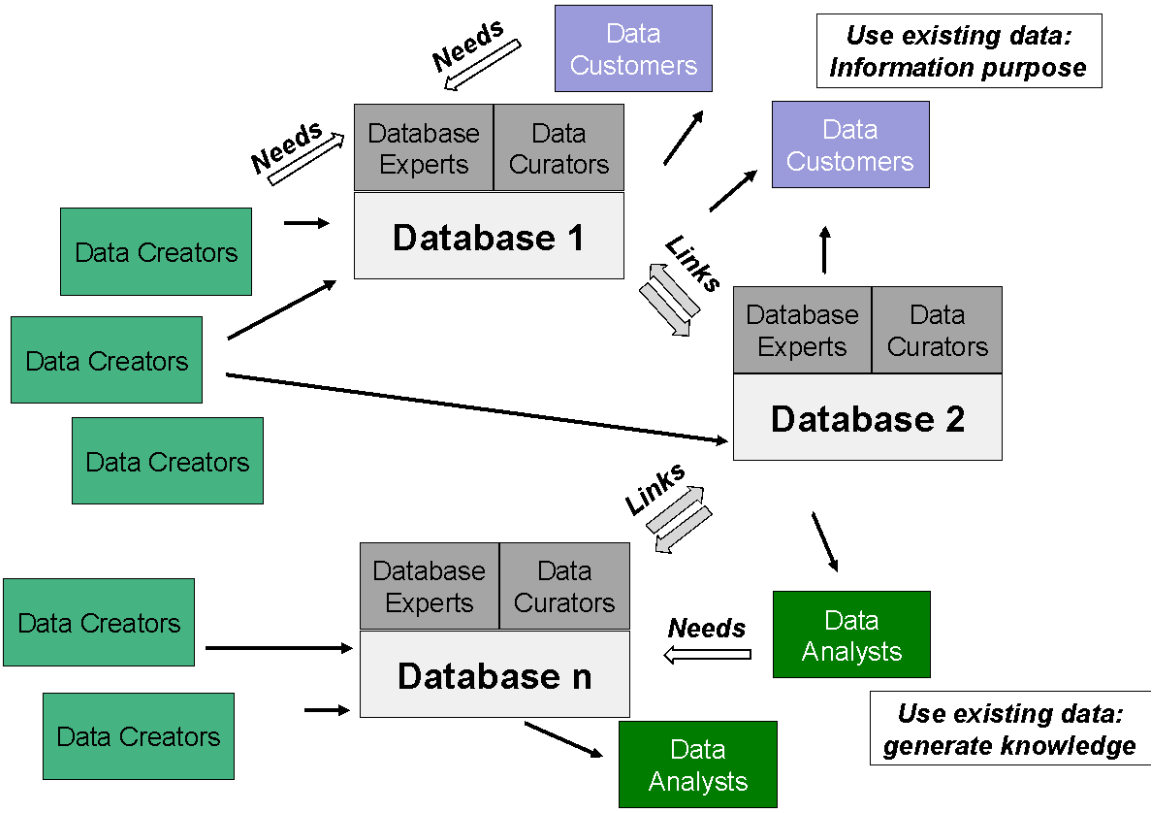


Figure 7: Overview of nanoinformatics stakeholders according to the data life cycle

The data life cycle starts with the generation of raw experimental data by different independent researchers or research groups (= Data Creators in Figure 7). Typically this data is processed, analyzed, and published by those groups. Unfortunately, and despite long ongoing discussions, in most cases the raw and also the full processed datasets are not published alongside the scientific publication. Some other scientific fields like protein crystallography or proteomics, in contrast, require that the primary data be stored in a database as a prerequisite for any peer-reviewed publication. In these fields there is a long tradition of depositing data in publicly accessible databases and accordingly knowledge that it is not only generated by research groups that create new experimental data but also by research groups re-analyzing existing data in data repositories.

In the field of nanoEHS, however, *in silico* toxicologists (=Data Analysts in Figure 7) that aim to derive computational models from primary data often first need to extract the details from the published literature in order to render the data usable for computational analysis and predictive modeling. Although data extraction is possible from publications, and can even be facilitated by computational means, this approach is still limited. Typically it will result in loss of data as publications usually highlight certain data in a study that fits the message of the authors. In addition, the authors usually depict mean or median values only, the whole set of experimental results is only rarely included. No effect data or data that does not demonstrate the sought after effects are often not published at all. It is well known and widely acknowledged that in particular

no-effect data are very important for regulatory decision-making but also they are important for the advancement of nanoEHS science in general.

Storing all nanoEHS data in federated, interoperable data repositories would allow for inter-laboratory comparisons and support the definition of the errors and variability within and between studies. It would also serve a range of other purposes such as supporting the establishment of nanomaterial grouping approaches, facilitating the generation of various *in silico* models, enabling meta-analysis of data etc. Overall there would be plenty of benefits starting from the level of the individual researcher up to the scientific, regulatory and industrial communities, as summarized in Figure 8.

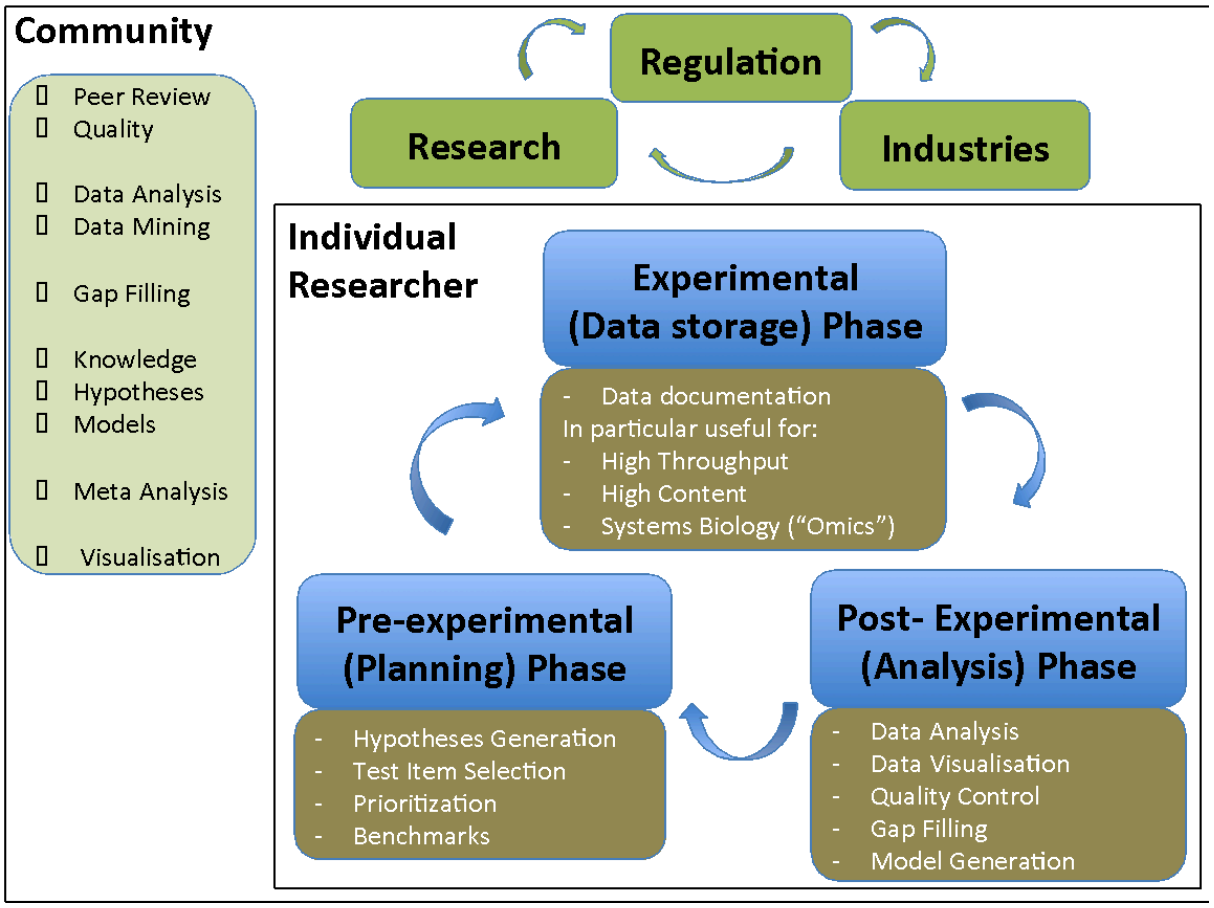


Figure 8: Impact of nanoinformatics for various stakeholders

Looking into the various stakeholders from the perspective of academia, industry and regulators one may assume that each has different needs and main objectives, as summarized in Table 2.

Table 2: Overview of (Nano)informatics needs from the perspectives of the different stakeholders

Perspective of Academia (Research users based in non-industrial research settings such as universities or research centers)	**Perspective of Industry** (Manufacturers and downstream users of nanomaterials, insurers, contract research organisations, regulatory consultancies etc.)	**Perspective of Regulators** (Government or international bodies producing regulations, policies or

		recommendations for sfae use of nanotechnologies*)*
Create new experimental data and deposit data in an access-controlled manner (at least until published)	Use existing data for justification of waiving of individual tests	Use existing data for analysis of plausibility
Secure experimental data by uploading into databases such that they can be assessed and re-used (by the same groups, by consortia, by the public)	Use existing data to fill data gaps for regulatory purposes (e.g. risk assessment and management).	Use existing data to "verify" reliability of data obtained with non-standard test guidelines
Benchmark own data by comparison with data obtained in other groups ("peer review")	Use existing data for interpretation of safety assessment results	Use existing data to better understand justifications in QSAR models
Use data for design of new experiments/ experimental studies, for compound selection etc.	Use existing data to design new materials with specific properties (i.e. safer products that retain their quality/performance)	Use existing data for substance prioritization
Use data for model building	Use data to establish health and safety procedures to protect workers, consumers, and the environment	Use data to build weight of evidence arguments
Use available datasets and modelling software, ideally under open license		Require that the data are generated according to standards (e.g. ISO, OECD) and regulatory demands (e.g. REACH.)
Lab data becoming	GLP & OECD TG	for use in Regulatory

	Interested party		
GOAL	**Academia**	**Industry**	**Regulator**
Use data for design of new experiments/ experimental studies, for compound selection etc.	insert tick or cross any required supporting text	insert tick or cross any required supporting text	insert tick or cross any required supporting text

Use existing data for substance prioritization			

[KP1]It may be possible to make the goals more concise.

For example, researchers want to benefit from new hypothesis that can drive new research. Industry might be more interested to derive information about a new material in an early development phase to learn whether the material properties are useful for the specific product needs and to get early warning signs of possible hazards and risks of the material. Regulators, finally, would appreciate linkages between specific material properties and hazards that they then may feed into specific regulatory actions.

Each of the stakeholders has their own specific needs and objectives. Most likely there will be no single one fit-for-all-purpose database. However, there might be common data elements that would be useful for field-specific purposes as well as serving the dual role of being useful for predictive modeling and establishing structure-property relationships.

One of the most important elements in further developing the field of nanoinformatics is starting and enhancing the dialogue between the different stakeholders such that they become aware of the needs of other stakeholders. As nanoscience in general but also nanoEHS is highly interdisciplinary, nanoinformatics can only mature if all the stakeholders actively participate in this process.

10. The community: Impact on stakeholders

Danail Hristozov[1], Andrea Haase[2] , Nina Jeliazkova[3], Iseult Lynch[4], Kai Paul[5]

[1]Greendecision Srl.
[2] German Federal Institute for Risk Assessment, Germany
[3] Idea consult Ltd, Bulgaria
[4] School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, B15 2TT Birmingham, United Kingdom
[5] Blue Frog Scientific Limited, Quantum House, 91 George Street, Edinburgh, EH2 3ES, United Kingdom

To be able to predict the properties, interactions and/or the adverse (eco)toxicological effects of the nanomaterials, it is fundamental to have access to high quality (meta)data. The many nanosafety projects have cost hundreds of millions of euros to generate a huge amount of relevant physicochemical, toxicokinetics, fate, exposure and (eco)toxicity data in over a decade of research. However, this information is only accessible via disparate and heterogeneous sources, offering different types of information in different formats (e.g. technical reports, excel sheets, data inventories, knowledge bases, scientific

publications). The only way to make a good use of this enormous volume of available data for the EHS assessment of nanomaterials is to curate them, link them to relevant modelling tools and make those accessible to their potential users from the research community, industry and regulation by means of user-friendly interfaces.

10.1 Impact on the Research Community

The challenging goal of developing and implementing a global nanoinformatics infrastructure will have a significant impact on research cohesion and international collaboration as it will require coordinated cooperation among EU and US scientific projects, centers and institutions to avoid overlaps, strengthen complementarities and create synergies that would eventually bring us closer to the goal. In addition, it will inevitably have a huge impact on the international efforts for harmonisation and standardisation of ontologies and data representation and sharing specifications, which will facilitate research across several domains, including chemistry, biology, toxicology, environmental science and others.

The modelling community will especially benefit from ready-to-use curated datasets, spanning endpoints of regulatory importance, and from open source and/or open access modelling components, developed in collaboration with experts from the respective scientific domains. This will allow comparison between different modelling approaches, which will ultimately lead to advancement in the nanoinformatics research area. The inclusion of data quality and completeness criteria, including information of what is technically and analytically feasible by the experimental setups will be a unique asset towards increasing the trust and validity of the model results. The modelling community will also benefit from interoperability of data and modelling components, allowing dynamic retrieval and analysis of data, beyond the static datasets.

10.2 Impact on Industry

Nanoinformatics can have a significant impact on industry as it can optimise testing for risk assessment under REACH, and also under e.g. the Biocides, Cosmetics, and Foods Directives. It can lower the cost not only of regulatory compliance but also of R&D&I for developing new products. There is already a large and growing market for data-driven modelling solutions that, once implemented, can increase confidence in nanotechnology to encourage innovation across several sectors, including but not limited to electronics, construction, packaging, food, energy, health care, automotive.

Indeed, key fundamental properties of nanomaterials may highlight their usefulness for certain objectives, e.g. pest control, medicine delivery or remediation. These databases may highlight more efficacious technologies than those that already exist, allowing for their replacement. The information may highlight undesirable properties such as excessive toxicity. Therefore, access to large and usable databases can steer innovation not only towards new nanotechnological advances but also safer implementation of these technologies, through SbD strategies.

Once a product is fully developed by industry, in order to gain full access to a market, they must first ensure they comply with the relevant legislation for that particular market. Informatics is often an integral part of the process from, *in silico* and materials model development and prediction of properties and effects for SbD to read across for regulatory risk assessment. Under REACH (Article 13, Article 25 & Annex VII-X), for example, it states that animal testing should only be conducted as a last resort, and that a registrant must exhaust all other forms of data acquisition before animal testing. This includes data from the literature, *in vitro* or non-testing (e.g. QSAR) approaches and analogues, and can be achieved only if the available data is accessible in a way that is consistent with regulatory requirements.

The correct storing and indexing of data is integral for the most successful outcome of literature collation. For example, it is integral that the data curation processes only allow good quality data to be held within the database and that data stored is easily searchable. At the very basic level any data placed within this domain must have key available chemical identifiers these include the composition of the chemical but also such identifiers as source, CAS and EC No. Searching is further complicated by nanomaterials as the same identifiers may be used for nanomaterials with different chemical and biological properties. As a minimum, to ensure the successful use of literature the aforementioned identifiers must be combined with key charateristics such as size and surface coating. Beyond this, key terms must be linked to the literature or database for which it resides, for example "lethality", "EC50", "rat", "fate", "monitoring", "environmental concentrations", "fate/landscape modelling" and "toxicological endpoint QSARs". This allows for an ease of searching for the most appropriate data that can be used in a regulatory context. Not limited to these terms, there must also be the use of key ontologies that are nanospecific. However, for the successful implementation of such phrases, the definitions must first be established and have wide acceptance across many varying areas of nanotechnology application.

Creating a database containing a wide spread of data quality may lead to misinterpretation or over weighting of studies that do not have fundamental prerequisites to be considered as reliable by users who may not be experts in the area of nanomaterial science. Therefore, a key goal within this field to make the data immediately usable is to reach community agreement on what makes a good quality study. This of course will also be field dependent (i.e. a good quality toxicology study will have different required attributes than a good quality environmental fate study). In this instance, a general set of quality criteria should be decided upon and from these further subsets relevant to the endpoint or use group. Studies that do not reach a critical quality score may not be suitable for the database, or the score can be indicated so as to aid the regulatory expert. This can be seen on the ECHA dissemination website, where the Klimisch Score (Klimisch, 1997) is extensively used. Although industry wish to be proactive and responsible they will be reluctant to test if they can not ensure the longevity of their investment in relation to regulatory compliance. Furthermore, any

~~data generated by industry under currently unsuitable guidelines may be wasted and will not advance the knowledge of the nano-community. It is critical that clear instruction and legislation now comes to light in order to ensure innovation, sustainability and safe production of nanomaterial technologies. The is currently no prescriptive legislation within the EU which would ensure, if followed, regulatory compliance for industry. For instance there are several highlighted nanomaterial physical chemical properties within IUCLID, but there is currently no consensus on which are the most important, and no legal obligation to supply these. It is unfeasible to request every single property for cost, time and relevance purposes. Nanoinformatics can aid in many areas of dossier preparation allowing a responsible, time- and cost-effective release to market. Therefore, with such uncertainty it becomes difficult to devise a registration (and testing) strategy.~~

~~Ensuring the correct curation will allow companies cost effective strategies for getting to the market. Whereby, literature data may be used to fill data gaps. The data used are often at the digression of the expert applying them, but ensuring easy access in a familiar and easy to use database will ensure data generated across many disciplines will be available and usable for this purpose. This will not only reduce cost, but the unnecessary use of animal testing and wasted materials.~~

In the domain of regulatory compliance, nanoinformatics can facilitate the applicability of standard tests as data curation and access to the wider community can help to reach consensus on testing methodologies and implementation of regulations. Currently with a lack of clarity on how nanomaterials will be regulated and tested, in part created by lack of access to quality data, it becomes difficult for companies to justify running full testing or regulatory programs or, much worse, investing in possible nanotechnology innovations. The data from nanoinformatics exercises will enable sound decisions to be made, and aid companies to comply with regulations.

Nanoinformatics can lead to the development of predictive models to facilitate both SbD strategies in the early stages of innovation and more cost-efficient regulatory risk assessment once the nano-enabled products are ready to go to the market. These could be for a particular endpoint (i.e. a QSAR for determining possible mutagenicity) or for exposure modelling. Models that can predict varying endpoints are used to tailor the testing strategy and to highlight any potential issues with particular methodologies/test guidelines which need to be addressed before testing can commence. In extreme cases model predictions may direct a company totally away from further pursuing a substance. When testing becomes technically infeasible these models can be used instead of standard testing. They can also be used as weight of evidence arguments when compiling dossiers that use read-across or trend analysis. In addition, exposure models can be used to predict both environmental and human exposure in a more cost and time-effective manner than the use of monitoring programmes, and sometimes they are

the only alternatives as it is not possible to monitor the release of a substance which is not yet on the market, and it is infeasible to expect that every nanomaterial must be monitored. ~~The derived exposure levels for each scenario (workers/ human) or compartment (environmental) can be compared to the derived critical values for toxicological responses (e.g. predicted no effect concentrations and derived no effect levels) and quantitatively assess risk. These predictive models can therefore show that product release and exposure are at safe levels, or can highlight the need for further risk management measures and operational condition considerations. This allows the responsible production and use of the material. However, it is not known if certain model parameters not nano-specific, such as release factors, are still appropriate for nanomaterials. Release factors are often worst-case estimated fractions of to the environment or during a worker activity, and the applicability of such factors must be addressed by industry and regulatory bodies. Based on knowledge from nanoinformatic exercises at general life-cycle stages it will be possible to produce a framework where there is a further element of realism so as not to be restrictive to the technology but also to ensure protection.~~

~~Nanoinfromatics will be key to reaching consensus on modes of action and adverse outcome pathways for particular nanomaterials. Information toward adverse outcome pathways can be used for the inception of new tests which can be used instead of animal testing. This information allows the use of targeted *in vitro*, *in chemico* and *in silico* tests at varying points of the AOP in order to define if a hazard will present. These tests reduce animal testing and are more time and cost-effective. These are not just a goal in the nanotechnology community, but globally for chemicals across all stakeholders. For instance, one recent development for the chemical industries and REACH is the implementation of non-animal test methods for skin sensitization (ECHA Chapters on, Guidance on Information Requirements and Chemical Safety Assessment R7a-c: Endpoint Specific Guidance, 2017). For some consumer products such methodologies form the backbone of the safety assessment and legislative processes as consumer products do not allow any animal testing to ensure their responsible release into market. Therefore, *in vitro* data, read-across to chemicals registered under other legislations (e.g. REACH; although data protection laws must be considered in their use by third-parties) and data from QSARs are used. To substantiate the application of read-across the similarities are further drawn by the use of QSARs and *in silico* assessments. Currently during tier 1 and tier 2 environmental safety assessments QSARs are exclusively used as is basic fate (landscape) modelling (Salvito et al., 2002). Only at tier 3 is experimental data used but still alongside landscape modelling. Without such tools, an informative and responsible safety assessment for consumer products becomes difficult, if not impossible, leading to a cessation of development in extreme cases. Further when selecting an analogue it is highly useful, and sometimes required to ensure the success of testing, that some properties can be predicted using QSARs.~~

Companies, especially SMEs, with limited resources for health and safety management are expected to benefit greatly from an interoperable nanosafety data and modelling infrastructure. Its implementation through the existing risk assessment and management tools (e.g. SUNDS) can have a significant practical value for both industries and regulators since it would make it possible to integrate technical data about the risks, benefits and costs of nanomaterials into sustainability portfolios to make informed decisions about how to address their safer production, downstream use and end-of-life treatment. It can also aid industries in making decisions on whether to invest in developing new nanotechnology products or to select conventional alternatives.

Such a nanoinformatics platform that combines data curation and modelling capabilities with user-friendly interfaces would be particularly interesting for SMEs as it would enable them to more readily perform regulatory EHS assessments and select options for safer product design. This can reduce their R&D&I costs and can enable them to more effectively compete with larger industries. Moreover, the application of high-quality curated data will reduce uncertainty in risk assessment and will improve risk communication, which will contribute to more positive market interpretation of their products and to better business cases.

Regulatory consultancies:

The pressure to assist companies in making technically challenging decisions about safety of their products has increased proportionally to the evolution of regulations. While there are a small number of consultancies providing nanoEHS assessment support to businesses, those can be are limited in terms of the data sources and analytical tools they can use. The nanoinformatics infrastructure could be used by these consultants or directly by the businesses for risk analysis and/or R&D&I decision making. The same also applies to researchers working in academia who design, develop or use nanomaterials or nano-enabled products. In addition, regulators in a variety of sectors (e.g. consumer products, cosmetics, food, medicines, chemicals and substances) also require data sources and modelling tools to do their own safety assessments. Similarly, standardisation bodies (e.g. OECD, ISO) would greatly benefit from annotating data originating from existing standards as this this would allow them to further improve those or to better adapt them to nanomaterials.

10.3 Impact on Regulatory Agencies

The nanoinformatics data and modelling infrastructure will further enable the safety assessment of nanomaterials. This will provide regulators with access to curated data and enhanced prediction capacity at moderate costs, to inform hazard and exposure modelling for risk analysis. Moreover, data may be immediately used for the advancement of regulation. Tests that are trialed and used in academia or by industry can highlight possible deficiencies, and therefore show where adaptations or new test methods need to be devised. The data may also highlight the need for different assessment factors, safety factors or methods for determining key critical values under certain legislative of frameworks. This will ensure that the implementation of the regulation is responsible and protective. This can only be established once a comprehensive data set has been realized. In this sense nanoinformatics can aid in the progression and iterative processes of regulation. The legislation, and guidance around it (i.e. testing guidance and practical guides), will give industry confidence in following

the regulatory framework in order to achieve compliance. Further, when obligations are part of a legal framework (which is outwith the current legislation) it will ensure that industry must comply with the provision of critical data in order to properly assess and address nanomaterial risk. However, under many frameworks such as REACH, there is simply guidance on best practices which do not form any particular part of the legislation. The decision on what encompasses "critical data" must be agreed upon by all stakeholders using sound scientific justification but also not breach competition laws to ensure a fair market. Currently, for instance, there are several highlighted nanomaterial physical chemical properties within IUCLID, however, there is currently no consensus on which are the most important, and no legal obligation to supply this. It is unfeasible to request every single property for cost, time and relevance purposes. Nanoinformatics can aid in many areas of dossier preparation allowing a responsible, time- and cost-effective release to market.

The nanoinformatics data, when properly realized can also aid in the creation, implementation and validation of new testing methods. These can be for screening purposes or to highlight non-threshold and threshold effects. The data will be key in developing intelligent alternative testing strategies such as *in vitro*, *in silico* and *in chemico* methodologies. Reducing cost, time and use of animals. The data may not only be useful for nanomaterials but regulation on a wider scale with a key goal in appropriate use of these alternative methods to reduce the need to experimental testing on animals. Further, data on higher tier test may be used as proof-of-concept for the alternative testing strategies.

The data can be collated in a comprehensive repository such as the EUON, and this database can be used by tools such as the OECD QSAR toolbox, allowing read-across, data collation and trend analysis to be more easily realized for data-gap filling. The data can also be used to generate reliable predictive models for exposure assessment, and highlight the key properties which affect the fate of the materials so that the release of the materials can be properly partitioned. The models may also be used to generate screening data for substances of concern. In combination, the raw data from the studies and models, may be used to screen and highlight nanomaterials of concern. Here, when substances are highlighted, then can be moved into current frameworks within the legislation, such as compliance checks (i.e. check the dossier for sound scientific justifictions and the correct implementations of risk management measures) or placed on to the relevant lists for further action, such as the CoRAP (community rolling action plan) list or list of substances of very high concern (SVHC). There is also currently a lack of differently tiered models, for instance, Tier 1 exposure model requires little chemical and activity specific data, thus being the most unrealistic and conservative. Such models are quick and cost-effective and can be used in situations where little risk is expected. Tier 2-3 modelling requires much more information, but are more realistic and less restrictive. Such tiering is useful for cost and time-effectiveness, but also to ensure optimal realism and protectiveness.

Nanoinfromatics will be key to reaching consensus on modes of action and adverse outcome pathways for particular nanomaterials. Information toward adverse outcome pathways can be used for the inception of new tests which can be used instead of animal testing. This information allows the use of targeted *in vitro*, *in chemico* and *in silico* tests

at varying points of the AOP in order to define if a hazard will present. These tests reduce animal testing and are more time and cost-effective. These are not just a goal in the nanotechnology community, but globally for chemicals across all stakeholders. For instance, one recent development for the chemical industries and REACH is the implementation of non-animal test methods for skin sensitization (ECHA Chapters on, Guidance on Information Requirements and Chemical Safety Assessment R7a-c: Endpoint Specific Guidance, 2017). For some consumer products such methodologies form the backbone of the safety assessment and legislative processes as consumer products do not allow any animal testing to ensure their responsible release into market. Therefore, *in vitro* data, read-across to chemicals registered under other legislations (e.g. REACH; although data protection laws must be considered in their use by third-parties) and data from QSARs are used. To substantiate the application of read-across the similarities are further drawn by the use of QSARs and *in silico* assessments. Currently during tier 1 and tier 2 environmental safety assessments QSARs are exclusively used as is basic fate modelling (Salvito et al., 2002). Only at tier 3 is experimental data used but still alongside landscape modelling. Without such tools, an informative and responsible safety assessment for consumer products becomes difficult, if not impossible, leading to a cessation of development in extreme cases. Further when selecting an analogue it is highly useful, and sometimes required to ensure the success of testing, that some properties can be predicted using QSARs.

Having a comprehensive repository of nanomaterials is essential for the successful implementation of read-across. It must be usable but also tailorable to the users needs and able to search the particular material or property of interest. It is also important for the researching community to reach a consensus on the most pertinent properties which should be compared during read-across strategies to prove that the nanomaterials will behave in a similar manner, i.e. size, surface coating etc. This will form the core of a read-across argument and the required bridging studies. Nanoinformatics will allow such a consensus to be reached, and for guidance to be dispensed in such documents as ECHA's read-across assessment framework (2017).

All these strategies form the backbone of more streamlined and cost-effective test strategy and dossier preparation. Along the entire process they allow more informed decisions to be made and also the reduction of testing on animals. The wide use in industry will also lead to the rapid advancement of such methodologies with the success and failures of their implementation dictating the future path of research or research focus.

Further, nanoinformatics from varying stakeholders, in particular industry, from monitoring studies of nanomaterial concentrations in the environment and during worker activities will aid in the inception of nanomaterial specific release factors should they be required. Here, there must be effort industry in creating these initial monitoring programmes, using the information to present sound scientific justification for any

adaptation to release factors to be used. Based on knowledge and site specific averages, for the general life-cycle stages which could be found on a nanoinformatics database, it will be possible to produce a framework where there is a further element of realism so as not to be restrictive to the technology but also to ensure regulatory protection. A similar framework was set up by CEFIC and such end goals specific environmental release factors were achieved for varying industries. Nanoinformatics may also help highlight the most appropriate risk management measures and operational conditions, as well as their effectiveness.

Overview of existing Databases and nanoEHS database Projects

Andrea Haase[1], Iseult Lynch[2], Nina Jeliazkova[3]

[1] German Federal Institute for Risk Assessment, Germany
[2] University of Birmingham, UK
[3] Ideaconsult Ltd, Sofia, Bulgaria

The following general, i.e. not nano-specific, databases could be of interest for nanoEHS (Table 3) and may provide some important general approaches.

Table 3: Overview of general (i.e. not nano-specific) databases

Name	Link	Description
eChemPortal	https://www.echemportal.org/echemportal/index.action	Global Portal to Information on Chemical Substances (includes information on Physico-chemical properties, ecotoxicity, environmental fate and behaviour, toxicity)
ChEMBL	https://www.ebi.ac.uk/chembl/	manually curated chemical database of bioactive molecules with drug-like properties, contains compound bioactivity data (e.g. Ki, Kd, IC50, and EC50)
ChEBI	https://www.ebi.ac.uk/chebi/	a freely available dictionary of molecular entities focused on 'small' chemical compounds
ChemSpider	http://www.chemspider.com/	a free chemical structure database providing text and structure search access to over 58 million structures

PubChem	https://pubchem.ncbi.nlm.nih.gov/	Free database of chemical molecules, consists of three dynamically growing primary databases. - Compounds (82 million entries) - Substances (198 million entries) - BioAssay (1.1 million entries)
DrugBank	https://www.drugbank.ca/	unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information
ToxNet	https://toxnet.nlm.nih.gov/	group of databases covering chemicals and drugs, diseases and the environment, environmental health, occupational safety and health, poisoning, risk assessment and regulations, and toxicology
ToxBank	http://toxbank.net/	central data warehouse for toxicity data management and modelling, includes a "gold standards" compound database, a repository of selected test compounds, a reference resource for cells, cell lines and tissues of relevance for in vitro systemic toxicity research
ToxCast	https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data	EPA's most updated, publicly available high-throughput toxicity data on thousands of chemicals
ToxRefDB	http://actor.epa.gov/toxrefdb	provides detailed chemical toxicity data
ECHA DB	https://echa.europa.eu/information-on-chemicals/registered-substances	Provides information on substances registered with ECHA
Array Express	https://www.ebi.ac.uk/arrayexpress/	Functional genomics data
TG-GATES	http://toxico.nibiohn.go.jp/english/	Toxicogenomics data
Gene Expression Omnibus	https://www.ncbi.nlm.nih.gov/geo/	High Throughput Expression Data
Organism specific databases	http://www.wormbase.org/#012-34-5, http://wfleabase.org/database/	Genomic data for the various species

This section highlights an important difference between the US and the EU in terms of approaches. Over the last 10 years or so, the US had a concerted effort on nanoEHS with 3 large-scale centres of excellence. (CEINT at Duke University, UC CEIN at UCLA and

more recently CNN at Harvard) and one needs to visit the respective websites for detailed informatics information. By contrast, the EU has funded over 50 nanosafety-related projects each ranging from 2-4 years in duration. Somewhat confusingly, both the project and the outputs from the project often carry the project name in the EU context, so datasets are referred to as the NanoX project dataset, and the NanoY project visualisation tools etc. There is strong incentivisation for tools / approaches / ontologies developed in one projects to be carried forward into subsequent projects, but an agreed naming convention for these co-developed hybrid-products has yet to be agreed. This is an important issue for the EU nanoinformatics community to resolve sooner rather than later in terms of making real progress and enhancing clarify for international collaborators.

Within the OpenRiskNet (www.openrisknet.org), a project funded under the Horizon 2020 EINFRA-22-2016 Programme (project ID: 731075) an open e-infrastructure will be delivered, providing resources and services to a variety of communities requiring risk assessment, including chemicals, cosmetic ingredients, therapeutic agents and nanomaterials. OpenRiskNet is working with a network of partners, organized within an Associated Partners Programme. One of the OpenRiskNet case studies will address specific needs identified by the nanosafety community. The case study will be defined based on project partners' experience in NanoEHS projects and activities within NanoSafety Cluster (NSC) working groups and task forces. Interactions with nanosafety projects have already been established in order to identify the key questions to be addressed, and where the OpenRiskNet infrastructure could be deployed and tested. OpenRiskNet will support the sustainability and further development of the eNanoMapper infrastructure supporting NSC needs. It offers the potential to incorporate data and tools developed within the NSC within the broader European scientific infrastructure and to combine them with resources developed within other areas such as chemical safety assessment.

More specifically addressing the informatics needs of the nanosafety community, the Horizon2020 project NanoCommons (project ID: 731032) will establish a nanoinformatics platform to convert the nanoEHS scientific discoveries into legislative frameworks and industrial applications, through concerted efforts to integrate, consolidate, annotate and facilitate access to the disparate datasets.drive best practice and ensure maximum access to data and tools. Networking Activities will span community needs assessment through development of demonstration case studies (e.g. exemplar regulatory dossiers). Joint Research Activities will integrate existing resources and organise efficient curation, preservation and facilitate access to data/models. Transnational Access will focus on standardisation of data generation workflows across the disparate communities and establishment of a common access procedure for access to the data and the modelling and risk prediction/management tools. NanoCommons will integrate across EU and US approaches to nanosafety data management and includes efforts to ensure sustainability of the nanosafety knowledge infrastructure through an advanced infrastructure and eventual integration into the EU Observatory for NanoMaterials (EUON, https://euon.echa.europa.eu/).

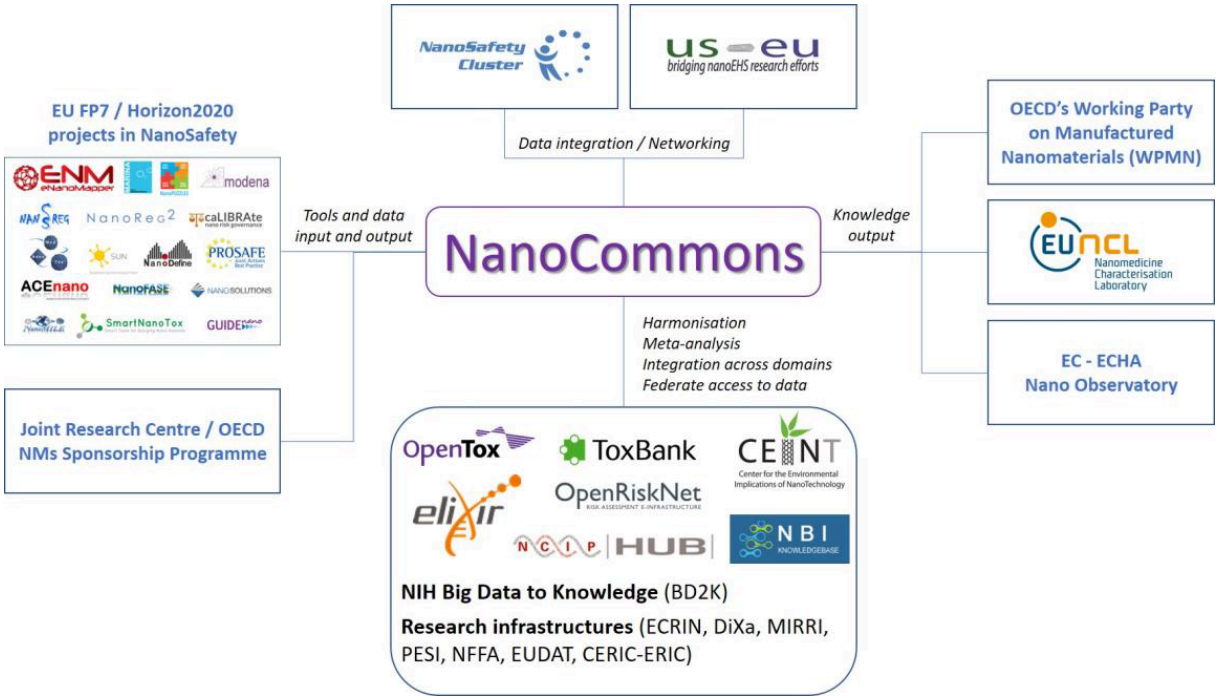


Figure X: Schematic illustration of the positioning of NanoCommons and how it will provide an integrating platform for the nanosafety knowledge community in Europe and internationally.

Appendix 1 provides a brief overview of some the recently finished or currently running projects whose main efforts were targeted towards developed of databases, or which developed large scale data sets or models for interrogating datssts. it is not intended as a complete overview, as projects contributed text voluntarily, rather than been systematically added. Table 4 provides an overview of the main databases and datasets specifically developed for nanoEHS.

Table 4: Overview on nano-specific databases:

Name	Link	EU/ US	Freely accessible/ Registration	description
eNanoMapper	http://search.data.enanomapper.net/	EU	partly	Contains primary research data from various finished nanoEHS projects and from literature
NanoHub	https://nanohub.org/			
DaNa	http://www.nanopartikel.info/	D	Freely accessible	Contains information for the general public and for researchers, SOPs
OCHEM	http://ochem.eu	EU	Freely accessible	Contains experimental data on nano and non-nano materials, allows to generate new models on the basis of a

				wealth of descriptors of various kind, allows for proper model evaluation, and allows to store models either privately or publically.
NECID	http://www.necid.eu	EU		Focus on exposure data
NanoDatabank	http://www.cein.ucla.edu/new/p192.php?pageID=408	US	Not currently?	Contains over 1000 uploaded investigations from CEIN as well as external investigators. Hazard focussed.
Nanomaterial-Biological Interactions Knowledge base	http://nbi.oregonstate.edu/	US	Freely Accessible	Contains over 200 *in vivo* toxicological assessments of nanomaterials in the embryonic zebrafish model. Includes nanomaterial characterization, mortality, and 21 morbidity endpoints such as morphological malformations, behavioral abnormalities and disrupted physiological function.
NanoMILE	https://ssl.biomax.de/nanomile/cgi/login_bioxm_portal.cgi	EU	Registration required	Contains characterisation data and HTS toxicity data for 120 NMs, with detailed mechanistic, omics and ecotox data for a sub-set. Supplemented with literature data in places, and used as basis for QSAR development.
ModNanoTox	ADD LINK	EU	Freely Available	Curated database on ecotox data, focussed mainly on silver, spanning 2007-2017. Currently integrating into CEINT's NIKC database and already available via eNanoMapper database.

10.2 Modelling Projects

The following table gives on overview of the most important modelling projects.

Table 5: Overview on modelling projects

Name	Link	EU/ US	Finished?	Short description
NanoPUZZLES	http://nanopuzzles.eu/	EU	yes	
ModENPTox	http://fys.kuleuven.be/apps/modenptox/	EU	yes	
PreNanoTox	http://prenanotox.tau.ac.il/	EU	yes	
MembraneNanoPart	http://www.membranenanopart.eu/	EU	yes	Multiscale modelling of NM-membrane and NM-protein interactions.
MODERN	http://modern-fp7.biocenit.cat/	EU	yes	
eNanoMapper	http://www.enanomapper.net/	EU	yes	
COST TD1204 MODENA	http://www.modena-cost.eu/	EU	yes	
SmartNanoTox	http://www.smartnanotox.eu/	EU	ongoing	Bionano interactions models and database. AOPs for pulmonary exposure, pathway modelling, nano- and bioinformatics-based mechanism-aware prediction tools.
UC CEIN	http://www.cein.ucla.edu/new/p10.php?pageID=170	US	ongoing	*In silico* data transformation and decision-making tools are involved in data processing to provide hazard ranking, exposure modeling, risk profiling, and construction of nano-SARs. These research activities are combined with

				educational programs that

10.3 NanoEHS projects generating large-scale datasets

Table 6 gives on overview on other important and interesting projects that are providing large-scale data sets relevant to nanoEHS.

Table 6: Overview on interesting projects

Name	Link	EU/ US	Finished?	Short description
NanoMILE	*http://nanomile.eu-vri.eu/*	EU	yes	Mechanistic understanding of NNs interactions with living systems and the environment, across their entire life cycle, leading to a framework (approach, experimental protocols, experimental data, QSAR models…) for MNMs classification according to their biological or environmental impacts.
NanoSolutions		EU	yes	
SUN		EU	yes	
ProSafe		EU	yes	
NANoREG		EU	yes	
FutureNanoNeeds				
NANECO	http://ochem.eu	NATO	yes	Development of QSAR models for metallic nanomaterials
NanoToxClass		ERANET	Ongoing	
NanoReg2		EU	Ongoing	
caLIBRAte		EU	Ongoing	
ACEnano	http://www.acenano-project.eu	EU	Ongoing	Development of a holistic analytical framework for reproducible NM characterisation, embedded in an operational ontology ("common language") and data framework to allow

				identification of causal relationships between NMs properties, be they intrinsic, extrinsic or calculated, and biological, (eco)toxicological and health impacts.

11. Milestones and Pilot Projects

Primary Author: Fred Klaessig

Contributors/Reviewers: A. Haase (BfR), Y. Cohen (UCLA), V. Grassian (UCSD), V. Stone (Herriot Watt), U. Vogel (NRCWE), D. Spurgeon (CEH), G. Visser (DSM), A. Falk (BioNanoNet), A. Worth (JRC), D. Winkler (CSIRO), I. Lynch (U. of Birmingham), Marc Williams (U.S. Army), Alan Kennedy (U.S. Army), Lisa Strutz (U.S. Army) and nanoWG participants.

11.1 Introduction

Other sections of the Nanoinformatics 2030 Roadmap are very inclusive regarding concepts and collaborations that advance the goals outlined in Section 3. In suggesting milestones and pilot studies, however, we are placing some boundaries on expectations. Informatics and ontologies require a disciplined attention to definitions, controlled vocabularies, well-defined data sets and metadata, etc. Consequently, we wish to be explicit here regarding the steps taken in crafting the NanoInformatics Roadmap's milestones and pilot projects: this Introduction provides context; the three Perspectives describe challenges facing the scientific fields in achieving the Roadmap's goals; and the Commentary connects this work to related EU Roadmaps. The resulting milestones and suggested pilot studies are provided as tables.

The milestones are listed according to near, mid- and far time horizons together with the scientific fields expected to contribute most to that specific topic. The early, or near-term, objectives identify a base set of activities; the mid-term objectives measure progress; and the far-term goals anticipate regulatory requirements if the resulting tools are to be accepted by risk assessment professionals.

The overarching strategy involves a progression of predictive computational models, each specific either to a topic (property, species, biological response) or to a stage in a nanomaterial's life cycle and each having utility to risk assessment. A modularized approach allows for flexibility in using available data, in judging model accuracy and in addressing regulatory requirements. Two visualizations are used to offset the flexibility regarding models. The Particle Description can be used to align phys-chem properties to specific particle regions (e.g. Core , shell, hydration layer etc.) and compositions. The

Particle Journey can be used to align models to stages in the nanomaterial's life cycle or to laboratory tests (e.g. membrane/biological barrier contact, internalisation, biodistribution / sub-cellular localisation, site of action, mode of action, transformations, clearance mechanisms etc.).

The milestones address three recognized challenges facing nanoinformatics and predictive computational models: (1). limited data sets; (2). limited data access due to proprietary, intellectual property or legal restrictions combined with the lack of long-term support for a nano-data repository and for data curation for acceptable recall and precision to retrieve data from appropriate repositories; and (3). regulatory requirements for harmonized test methods conducted according to GLP. In response, the milestones (a) encourage data generation through collaborations, surrogate test methods, newer screening techniques, while (b) recognizing that progress will be uneven and (c) suggesting that read-across and related data-filling techniques (QSARs & trend analyses) are the means for introducing the fruits of this work into the regulatory process.

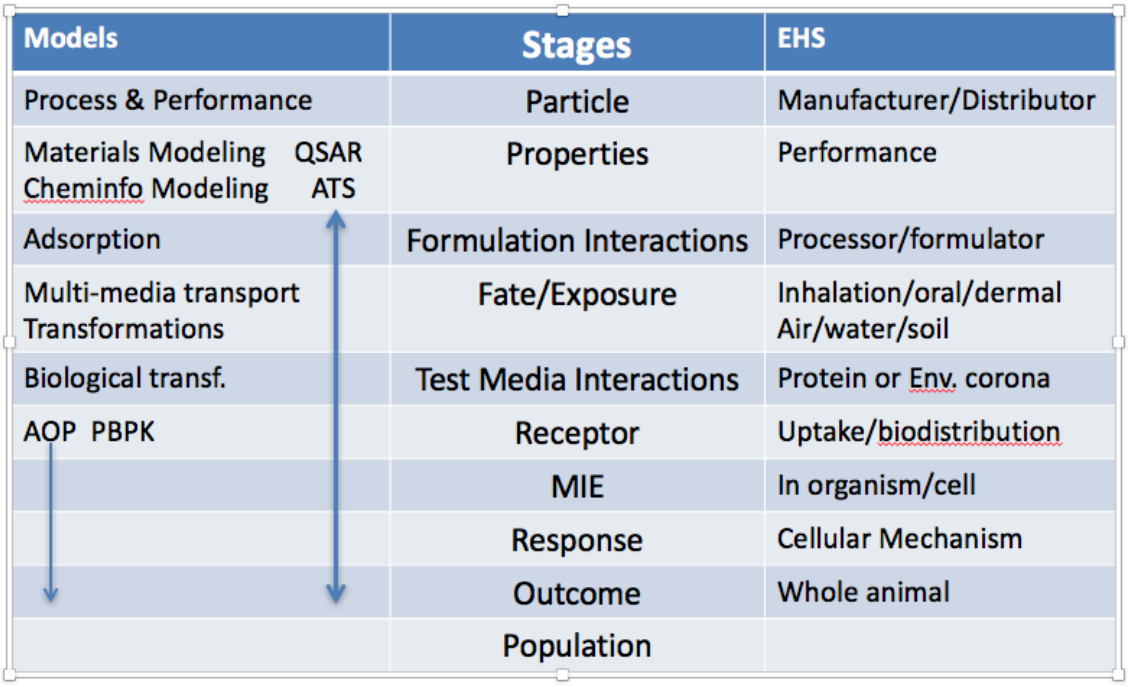
The reader is reminded that the background to the individual milestones and their sources were provided in Section 4, Introduction, and the citations are: the Nanoinformatics 2020 Roadmap (1,2); the COST sponsored workshop in Maastricht (3,4); and a 2014 NSF-sponsored workshop (5)..

11.2 Perspectives for Toxicological Milestones

The Nanoinformatics 2030 Roadmap responds to two aspects of toxicology and related biological sciences (ecotoxicology, medicine, physiology, systems biology). Firstly, there is hypothesis-driven research conducted against a backdrop of bioinformatics, assay development, alternative test strategies, adverse outcome pathways (AOPs), introduction of new capabilities with 'omics' and so on. Secondly, there is the manner in which toxicology is practiced in a regulatory context, i.e. an insistence on harmonized test methods conducted according to GLP. This insistence is substantive, reflecting societal considerations of public health, statutory language and legal precedent that are embodied in regulatory agency procedures.

While the distinctions between hypothesis-driven research and regulatory practice may be well known to many in the toxicological sciences, researchers in the physical and computational sciences are generally less aware of the distinctions and their importance for how research is utilisable (or not) by regulators. Accordingly, the Roadmap 'co-locates' computational models with the stages found in a material's life cycle as in Table 12-1: the middle column lists the life cycle stages through to the point of sampling where laboratory test protocols prevail (abiotic, mesocosm, *in vitro* or *in vivo*); the left-hand column aligns computational models to those stages and laboratory tests; and the right-hand column identifies the likely user of the model's estimates (manufacturer, processor, formulator) or the associated risk assessment concept.

Table 11-1: Overview on how different models relate to LCA stages



Alternative test strategies (ATS) and adverse outcome pathways (AOP) are examples of hypothesis-driven research. Neither is utilized currently for chemicals by regulators, as they have not yet undergone validation as outlined by the OECD (6). In general, regulatory expectations of reliability and relevance, such as expressed in the Klimisch score (7), favor established assays from EPA or the OECD conducted according to GLP.

Risk assessment professionals may estimate a property/biological activity when chemical substances are grouped and tools, such as QSAR/QSPRs, trend analysis or read-across default rules, are used for filling property/biological activity data gaps. Read-across can also be used for estimating effects across species.

Data-filling techniques (QSARs, trend analyses and read-across) have been considered for nanomaterials (8, 9) and are a potential means for introducing new approaches (ATS, AOP and computational methods) to the regulatory process. Procedures for grouping chemical substances remain to be established, but we can anticipate that similarity in toxicokinetics will be a critical selection factor. In Table 12.1, toxicokinetics (incorporated into PBPK modeling) includes uptake, biodistribution, and receptor interactions at the Molecular Initiating Event (MIE).

The criteria regulators will consider necessary for model acceptance will become increasingly visible with future progress (see the FDA's guidance (10) on PBPK models as a current example). The milestones alert the reader to such matters through phrases such as 'credible AOPs', 'validation requirements', and 'regulatory endorsement' but don't necessarily give guidance on how to achieve these.

11.3 Perspectives for Physico-Chemical Milestones

While several nanoEHS disciplines describe chemical substances using simple chemical formulae for molecular identities, e.g. TiO_2, these fields differ when touching upon physicochemical properties. The Chemical Abstract Services does not index TiO_2 information according to volume or shape. Yet, in early 2017, the EPA with 'nanoscale form' and ECHA with 'nanoform' decided to differentiate particles with identical core compositions using size, shape and surface chemistry/coating distinctions (11, 12).

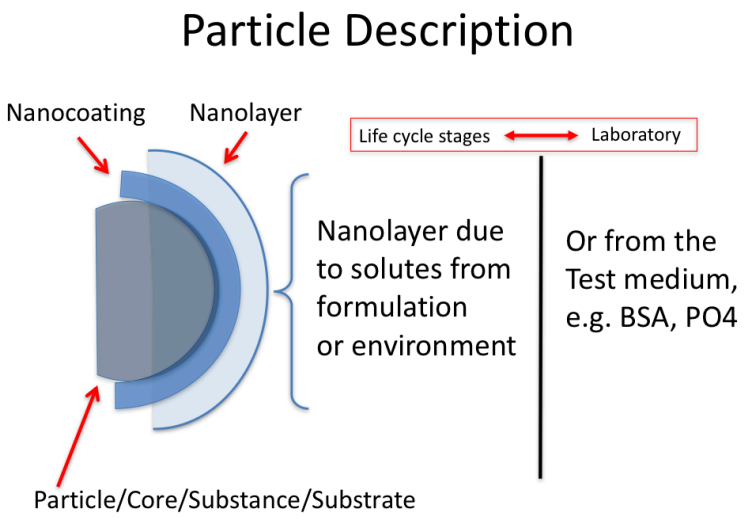
In materials science, a phase of uniform composition that is in equilibrium with other phases through the phase rule defines the molecular identity, which was one justification for not considering size (volume) when indexing information. However, the physico-chemical properties often considered meaningful to toxicological studies are non-equilibrium functions, perhaps steady-state or those emphasizing kinetic pathways, which reflect the non-equilibrium nature of nanomaterials. Using the EPA ruling (11) as an example: dissolution is kinetics (solubility is equilibrium); zeta potential reflects coatings and adsorbed species (not the core composition); dispersion stability may involve steric or electrostatic factors; and surface reactivity is re-phrased to be biology, "..the degree to which the nanoscale material will react with biological systems." Surface reactivity essentially encompasses the nano-bio interface.

There are complicating factors regarding molecular identity. For organic molecules, the molecular entity in the solid and in solution is essentially the same covalently bonded molecule. For inorganic materials, metals or metal oxides, the molecular identity in the solid may encompass ionic or metallic bonding and may not be the species found in solution. The experience gained with QSAR/QSPRs for drug discovery may not be translatable to metal oxide toxicity. The second complicating factor is the dual nature of the particle (13): acting as a particle for dispersal, biodistribution and cell entry and acting as a chemical reservoir for some modes of action (dissolution, drug release, biopersistence).

Returning to equilibrium and steady state distinctions, melting is both a phase transition and a form of dissolution. Melting point depression can be estimated using the Gibbs-Thompson equation, which combines equilibrium thermodynamic concepts with case-specific solid-liquid interface energies. Functional assays (14) involve transport properties, which may be constrained by case-specific macroscopic conditions (flow rate) or surface kinetics. These case-specific considerations will influence the selection of descriptors in models.

To illustrate the potential for distinguishing among identities, Figure 12-1 is a particle visualization, a physical model, utilizing terms defined by ISO TC-229. One recommendation is to assign a physico-chemical property to the localized region and composition likely to govern that phenomenon, e.g. zeta potential with surface layer and shape with particle substrate. The particle description highlights possible sources for a changing nanolayer composition across the life cycle (Table 12-1).

Figure 12-1 Nanomaterial physical model



In the milestones, coatings also include surface layers or protein or other acquired biomolecule coronas that were not present when first manufactured. The first milestone supports a review of data collected from the OECD test study programs such as NANoREG to establish a base case.

One pilot project focuses on dissolution, a common theme to several of the nanoEHS disciplines, and it aims to clarify issues, such as ionic solids not retaining their nominal molecular identity upon dissolution. There is a large body of dissolution data and solubility modeling that may be applicable to nanoscale materials, but may be indexed under other metadata or ontology rules than those used in nanoEHS. Collecting this, and indexing it with nanoEHS terms may unlock additional large datasets for use in model development.

11.4 Perspectives for Modelling Milestones

There is a great diversity in model types, including computational ones. The regulatory framework is itself a model, as it is a simplified representation of a much more complex system. It is a form of decision model that utilizes numerical values for selected variables (production volumes, intended uses, human health and ecotoxicity endpoints). There are variants both broader and narrower (15, 16) that extend beyond statutory requirements. In populating decision models, one may use laboratory generated test results or the numerical estimates from computational models. These in turn can be based on quantum mechanical calculations of molecular bonding or other descriptors examined in Sections 6 & 7.

There are models that utilize thermodynamic concepts, such as dynamic energy budget or Ostwald-Freundlich dissolution (17, 18). For the most part, dispersal models of particle-as-colloid accept the applicability of classical DLVO theory. As discussed in Section 12.3, size-dependent properties imply that the nanomaterial is not at

equilibrium, but rather in a steady state or a kinetically hindered state. This raises significant concerns when a computational estimate of dissolution is incorporated into a decision model or physiologically-based pharmacokinetic/ADME model without considering kinetically hindered dissolution mechanisms (10, 16, 18).

There is also uncertainty regarding the meaning of 'structure' when proposing a computational model for QSPR. Is it the structure of a molecule (bond lengths, angles, functional groups) or is it the particle's external shape influencing those molecular concepts or is it the particle's internal arrangement of surface, coating, surface layer? The same questions about the meaning of 'structure' arise with QSARs.

All models, frameworks and theories are prone to variants of Type III errors, where the question posed extends beyond the model's domain, yet the model returns a result. Basing computational models solely on *in vitro* assay data to predict *in vivo* outcomes raises the prospect of such errors, as does using QSPR or other models to predict properties outside of the domain of the 'training' dataset. Models, like experiments, can be surprisingly robust and can fail as well (19).

Model validation, which is the subject of an OECD guidance document regarding QSARs (20), raises two related issues. Firstly, the subject matter, the QSAR, must have a "defined domain of applicability" and secondly, should have a "mechanistic interpretation (if possible)" that tie the descriptors to the endpoint being predicted. There is also a guidance document on computerized systems, including databases, data approval and periodic review that may be applicable to the data sets used to validate a model (21).

It is not yet known how these guidance documents will be applied to computational models or the underlying datasets. This is one reason for favoring a modularized approach, where each module can be tested against data specific to a target endpoint, thereby enhancing its acceptability in data-filling. Descriptors might be tested using broad datasets extending beyond nanoscale materials, but once accepted then be re-calibrated to a narrower nanoscale material dataset for a regulatory submission.

11.5 Commentary on related EU activities

The European Nano Safety Cluster has published two related documents: the 2016 "Closer to the Market Roadmap" (CTTM) and the 2017 "Regulatory Research Roadmap" (RRR) (22, 23). Additionally, the Joint Research Centre has published a final report for the NanoComput project. Some commentary is appropriate as there are significant overlaps, but with different focal points.

The CTTM emphasis is on assuring workers and consumers that there are procedures, policies and programs in place to reduce uncertainties surrounding nano-enabled products. Integral to the CTTM program is providing "solid operational knowledge (high level of scientific expertise and robust accumulated datasets)" (Recommendations in 22).

A significant overlap occurs in the discussions of two bottlenecks (21, page 30) that also identify the responsible parties for resolving hurdles (basic scientific knowledge, research to support regulation and Nanotechnology Market/CTTM). For "uncertainties in risk assessment and in regulation," the recommendation for regulatory research in the CTTM is to improve & stabilize regulation and to communicate uncertainties. Regarding the "lack of validated methods (toxicological and analytical) for nanosafety assessment," the CTTM recommends developing scientific knowledge via equipment, harmonization, round robins, validation studies and general guidelines on how to standardize nano-specific protocols.

The RRR (23) has a fully integrated risk analysis framework as its objective, while the Naoninformatics Roadmap leverages databases & metadata considerations to expand the use of computational models. In both cases, validation is critical to successful use by regulators.

Both the RRR and Nanoinformatics Roadmap attempt to bring awareness of regulatory requirements forward in time. For the RRR, this is expressed as:, "It should also be noted that while the hexagon diagrams indicate prioritisation, issues situated on the right-hand side (long term and distant future priorities) of each prioritisation diagram need to be considered at an early stage to ensure that any short-term activity generates outputs that will be useful for developing longer-term priorities." The RRR connects high quality data to validated methods, while the Nanoinformatics Roadmap ties quality to the metadata found in either ISA-TAB-nano or ISA-TAB-JSON formats and in the ontology (NPO or eNanoMapper).

The EC's Joint Research Centre has issued a report (24) reviewing current computational models that may be useful to regulatory authorities. It is comprehensive and shares many concepts with this Roadmap, but with a different emphasis. The JRC's advisory role to the Commission leads it to specific recommendations regarding public dissemination, filling knowledge gaps with concrete regulatory applications in mind and developing a one stop hub for databases and models. The Roadmap offers milestones directed at a wider stakeholder group whose activities may contribute useful data for modeling, but leaving applicability to regulatory frameworks as a second validation step.

In the Table listing milestones, the scientific fields most involved in achieving a specific goal along the roadmap are indicated, aligning roughly with the CTTM approach. Additionally, the same color code used with the RRR's hexagons has been added to the Milestone Table to identify those activities that are predominantly data generation, method development and regulatory framework milestones. Relative to the JRC report, the milestones place greater emphasis on read-across exercises as a means to gain feedback on model and dataset acceptability.

References:

(1). de la Iglesia D, Harper S, Hoover MD, Klaessig F, Lippell P, Maddux B, Morse J, Nel A, Rajan K, Reznik-Zellen R, Tuominen MT. Nanoinformatics 2020 Roadmap. Published by the National Nanomanufacturing Network Amherst, MA 01003. DOI: 10.4053/rp001-110413

(2). Maojo, V., 2010, The ACTION-Grid White Paper on Nanoinformatic, International Cooperative Action on Grid Computing and Biomedical Informatics between the European Union, Latin America, the Western Balkans and North Africa

(3). Winkler, D.A., Mombelli, E., Pietroiusti, A., Tran, L., Worth, A., Fadeel, B. and McCall, M.J., 2013. Applying quantitative structure–activity relationship approaches to nanotoxicology: current status and future potential. *Toxicology*, *313*(1), pp.15-23.

(4). Winkler, David A. "Recent advances, and unresolved issues, in the application of computational modelling to the prediction of the biological effects of nanomaterials." *Toxicology and applied pharmacology* 299 (2016): 96-100.

(5). Grassian, V.H., Haes, A.J., Mudunkotuwa, I.A., Demokritou, P., Kane, A.B., Murphy, C.J., Hutchison, J.E., Isaacs, J.A., Jun, Y.S., Karn, B., Khondaker, S.I., Larsen, S.C., Lau, B.L.T., Pettibone, J.M., Sadik, O.A., Saleh, N.B., and Teague, C. 2016. NanoEHS–defining fundamental science needs: no easy feat when the simple itself is complex. *Environmental Science: Nano*, *3*(1), pp.15-27.

(6). OECD Series on Testing and Assessment Number 34, "Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment", ENV/JM/MONO(2005)14.

(7). Klimisch, H-J., M. Andreae, and U. Tillmann, "A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data," Regulatory toxicology and pharmacology, 1997, 25.1: 1-5.

(8) OECD Series on Testing and Assessment Number 194, "Guidance on Grouping of Chemicals, Second Edition", ENV/JM/MONO(2014)4.

(9). OECD Series on the Safety of Manufactured Nanomaterials, "Approaches on Nano Grouping/Equivalence/Read-Across Concepts Based on Physical-Chemical Properties (GERA-PC) for Regulatory Regimes". ENV/JM/MONO(2016)3.

(10). FDA, Physiologically Based Pharmacokinetic Analyses — Format and Content Guidance for Industry, 2016, (accessed May 2017) https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM531207.pdf.

(11). § 704.20 Chemical substances manufactured or processed at the nanoscale, Federal Register, 82(8): 3641- 3655.

(12). European Chemicals Agency, "Guidance on Appendix 4: Recommendations for nanomaterials applicable to the Guidance on Registration…" Draft (Public) Version 1.0, January 2017.

(13). Johnston, J.M., Lowry, M., Beaulieu, S., and Bowles, E. 2010. State-of-the-Science Report on Predictive Models and Modeling Approaches for Characterizing and Evaluating Exposure to Nanomaterials. U.S. Environmental Protection Agency, Office of Research and Development, Athens, GA. EPA/600/R-10/129, September 2010.

(14). Hendren, C.O., Lowry, G.V., Unrine, J.M. and Wiesner, M.R., 2015. “A functional assay-based strategy for nanomaterial risk forecasting”. *Science of the Total Environment*, 2015, 536: 1029-1037.

(15). Igor Linkov, Matthew Bates, Benjamin Trumpa,, Tom Seager, Mark Chappell, Jeffrey Keisler, 2013, Nano Today, 2013, 8.1: 5-10.

(16). Arts JH, Hadi M, Irfan MA, Keene AM, Kreiling R, Lyon D, Maier M, Michel K, Petry T, Sauer UG, Warheit D, 2015, Regulatory Toxicology and Pharmacology. 2015, 71(2): S1-27.

(17). Tin Klanjscek, Erik B. Muller, Roger M. Nisbet, 2016, Journal of Theoretical Biology 404: 361–374.

(18). Lijun Wang and George H. Nancollas, 2009, Dalton Trans.: 2665–2672

(19). Mäki, Uskali, 2005, 'Models are experiments, experiments are models', Journal of Economic Methodology, 12:2, 303 - 315

(20). OECD, Series on Testing and Assessment No. 69, “Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models”. ENV/JM/MONO(2007)2.

(21). OECD, OECD Series on Principles of Good Laboratory Practice and Compliance Monitoring No. 17, “Application of GLP Principles to Computerised Systems”. ENV/JM/M ONO(2016)13.

(22). Falk, Andreas, Christa Schimpel, Andrea Haase, Benoît Hazebrouck, Carlos Fito López, Adriele Prina-Mello, Kai Savolainen, Adriënne Sips, Jesús M. Lopez de Ipiña, Iseult Lynch, Costas Charitidis, Visser Germ, 2016, “Closer to the market Roadmap”-(CTTM)

(23). Vicki Stone, Serli Önlü, Enrico Bergamaschi, David Carlander, Anna Costa, Wilson Engelmann, Antoine Ghanem, Sonja Hartl, Danail Hristozov, Janeck J. Scott-Fordsmand, Keld Alstrup Jensen, Frank von der Kammer, Jacques-Aurelien Sergent, Monita Sharma, Maria Dusinska, Bernd Nowack, Phil Sayre, Ulla Vogel, Martie van Tongeren, Socorro Vázquez-Campos, and Wendel Wohlleben, 2017, Research priorities relevant to development or updating of nano-relevant regulations and guidelines, Nano Safety Cluster Research Regulatory Roadmap.

(24). Worth, A., Aschberger, K., Asturiol Bofill, D., Bessems, J., Gerloff, K., Graepel, R., Joossens, E., Lamon, L., Palosaari, T. and Richarz, A., “Evaluation of the availability and applicability of computational approaches in the safety assessment of nanomaterials” , EUR 28617 EN, Publications Office of the European Union, Luxembourg, 2017, ISBN 978-92-79-68708-2, doi:10.2760/248139, JRC106386.

Year	Milestone	Tox.	P-Chem	Models
Near	1). Document benchmark NMs: their biological & physico-chemical data, coatings, manufacturing technique(s), production volumes; primary use patterns.	**X**	**X**	**X**
Near	2). Develop functional assays and NM-descriptors to model environmental changes: confirm where possible with *in situ* instrumentation and relate to pristine NMs, their dissolution, dispersal, homo- and hetero-aggregation		**X**	**X**
Near	3). Develop high throughput methods for measuring NM interactions with plasma proteins (protein coronas) for PBPK modeling of NM distribution in the body.	**X**		
Near	4). Propose data sharing/file transfer, ontology, & terminology criteria for interoperable nanoEHS databases and online modeling services and promote appropriate training programs	**X**	**X**	**X**
Mid	5). Develop surrogate & fast screen assays suitable for tiered testing that align with credible AOPs in order to evaluate NM descriptors for computational model validation	**X**		
Mid	6). Consensus on validated particle descriptors useful for physico-chemical properties and for environmental changes to serve as a basis for modeling biological endpoints			**X**
Mid	7). Identify NP fingerprints (biomarkers, NP property descriptors, functional assays) to allow for NP grouping and with selected OECD TG's *in vitro* endpoints	**X**	**X**	
Mid	8). Clarify computer model validation requirements for regulatory purposes (particle descriptors including coatings; chemical grouping)	**X**		**X**
Mid	9). Establish high throughput *in vitro* protocols for generating large datasets useful for validating model descriptors	**X**		
Far	10). Complete a suite of validated models for environmental fate and effect that are useful & endorsed by regulators for QSAR, trend analysis and read-across purposes			**X**
Far	11). Complete a suite of PBPK models that include ADME and NP-protein corona factors			**X**
Far	12). Develop appropriate assays for identifying the AOP profile for new NP classes and the minimum characterization data set for classifying a new NM to a class	**X**		
Far	13). Regulatory endorsement of *in vitro* predictive models for NMs	**X**		**X**

= Data Generation; = Method; and = Regulatory

Pilot Projects	
Data set availability (schedule and access criteria): ● caNano: accessible for non-confidential data ● Nanomaterial Registry: accessible; limited nanoEHS data ● UC-CEIN (nanoinfo.org) & CEINT: have requirements; ● NANoREG: access in 2017 ● OECD Working Party access awaiting clearances ● Identify other database resources & access criteria ● Data management plans for academic institutions ● Open Science end-point vision.	**Informatics Infrastructure:** ● Instances of Characterization standards at ASTM; ● Extensible particle ontology standard at ASTM; ● ISA-TAB-nano upgrade led by Duke and OSU; ● Incorporation of UDS considerations into standards; ● Revisit error expression, data templates, metadata selection with existing datasets and templates ● Establish a coordination site
Dissolution: ● Clarify industry interest and identify participants; ● Pursue collaboration with Materials Genome Initiative & European Modeling Council; ● Pursue collaboration with Pharmaceutical colleagues regarding drug release experience; ● Clarify regulators requirements for use in read-across; ● Examine nanomaterials aging and transformation implications.	**Informatics literacy:** ● Survey Ph.D. students & Post-docs on informatics acceptance; ● Survey P.I.s on informatics acceptance; ● Incorporate help desk and P.I. proposals from NanoCommons and Oregon State University

13. References

1. McWilliams, A., The Maturing Nanotechnology Market: Products and Applications,

BCCResearch, Editor. Nov. 2016

2. Cientifica White Papers [Internet]. 21 Dec 2009 [cited 7 Jul 2017]. Available: http://www.cientifica.com/research/white-papers/

3. Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0. In: FORCE11 [Internet]. 10 Sep 2014 [cited 7 Jul 2017]. Available: https://www.force11.org/fairprinciples

4. de la Iglesia D, Harper S, Hoover MD, Klaessig F, Lippell P, Maddux B, Morse J, Nel A, Rajan K, Reznik-Zellen R, Tuominen MT. Nanoinformatics 2020 Roadmap. Published by the National Nanomanufacturing Network Amherst, MA 01003. DOI: 10.4053/rp001-110413

5. Maojo, V., 2010, The ACTION-Grid White Paper on Nanoinformatic, International Cooperative Action on Grid Computing and Biomedical Informatics between the European Union, Latin America, the Western Balkans and North Africa

6. Winkler, D.A., Mombelli, E., Pietroiusti, A., Tran, L., Worth, A., Fadeel, B. and McCall, M.J., 2013. Applying quantitative structure–activity relationship approaches to nanotoxicology: current status and future potential. Toxicology, 313(1), pp.15-23.

7. Winkler, David A. "Recent advances, and unresolved issues, in the application of computational modelling to the prediction of the biological effects of nanomaterials." Toxicology and applied pharmacology 299 (2016): 96-100

8. Grassian, V.H., Haes, A.J., Mudunkotuwa, I.A., Demokritou, P., Kane, A.B., Murphy, C.J., Hutchison, J.E., Isaacs, J.A., Jun, Y.S., Karn, B., Khondaker, S.I., Larsen, S.C., Lau, B.L.T., Pettibone, J.M., Sadik, O.A., Saleh, N.B., and Teague, C. 2016. NanoEHS–defining fundamental science needs: no easy feat when the simple itself is complex. Environmental Science: Nano, 3(1), pp.15-27

9. Bañares MA, Haase A, Tran L, Lobaskin V, Oberdörster G, Rallo R, Leszczynski J, Hoet P, Korenstein R, Hardy B, Puzyn T. 2017. CompNanoTox2015: novel perspectives from a European conference on computational nanotoxicology on predictive nanotoxicology. Nanotoxicology. doi: 10.1080/17435390.2017.1371351. [Epub ahead of print]

xxx

Hastings J, Jeliazkova N, Owen G, Tsiliki G, Munteanu CR, Steinbeck C, et al. eNanoMapper: harnessing ontologies to enable data integration for nanomaterial risk assessment. J Biomed Semantics. BioMed Central; 2015;6: 10.

4. Jeliazkova N, Chomenidis C, Doganis P, Fadeel B, Grafström R, Hardy B, et al. The eNanoMapper database for nanomaterial safety information. Beilstein J Nanotechnol. Beilstein-Institut; 2015;6: 1609–1634.

5. Tsiliki G, Munteanu CR, Seoane JA, Fernandez-Lozano C, Sarimveis H, Willighagen EL. RRegrs: an R package for computer-aided model selection with multiple regression models. J Cheminform. Springer International Publishing; 2015;7: 46.

6. Drakakis G, Moledina S, Chomenidis C, Doganis P, Sarimveis H. Decision Trees for Continuous Data and Conditional Mutual Information as a Criterion for Splitting Instances. 2016;

7. Helma C, Rautenberg M, Gebele D. Nano-Lazar: Read across Predictions for Nanoparticle

Toxicities with Calculated and Measured Properties. Front Pharmacol. Frontiers; 2017;8. doi:10.3389/fphar.2017.00377

8. Hendren CO, Lowry GV, Unrine JM, Wiesner MR. A functional assay-based strategy for nanomaterial risk forecasting. Science of The Total Environment. 2015;536: 1029–1037.

9. Thomas DG, Gaheen S, Harper SL, Fritts M, Klaessig F, Hahn-Dantona E, et al. ISA-TAB-Nano: a specification for sharing nanomaterial research data in spreadsheet-based format. BMC Biotechnol. 2013;13: 2.

10. Hendren CO, Powers CM, Hoover MD, Harper SL. The Nanomaterial Data Curation Initiative: A collaborative approach to assessing, evaluating, and advancing the state of the field. Beilstein J Nanotechnol. 2015;6: 1752–1762.

11. Lowry GV, Gregory KB, Apte SC, Lead JR. Transformations of nanomaterials in the environment. Environ Sci Technol. 2012;46: 6893–6899.

12. CEINT NanoInformatics Knowledge Commons (NIKC) | Center for the Environmental Implications of NanoTechnology [Internet]. [cited 7 Jul 2017]. Available: https://ceint.duke.edu/research/nikc

13. de la Iglesia D, Harper S, Hoover M, Klaessig F, Lippell P, Maddux B, et al. Nanoinformatics 2020 Roadmap [Internet]. National Nanomanufacturing Network Amherst; doi:10.4053/rp001-110413

14. Maojo V. The ACTION-Grid White Paper on Nanoinformatic, International Cooperative Action on Grid Computing and Biomedical Informatics between the European Union, Latin America, the Western Balkans and North Africa. 2010.

15. Winkler DA, Mombelli E, Pietroiusti A, Tran L, Worth A, Fadeel B, et al. Applying quantitative structure-activity relationship approaches to nanotoxicology: current status and future potential. Toxicology. 2013;313: 15–23.

16. Winkler DA. Recent advances, and unresolved issues, in the application of computational modelling to the prediction of the biological effects of nanomaterials. Toxicology and Applied Pharmacology. 2016;299: 96–100.

17. Grassian VH, Haes AJ, Mudunkotuwa IA, Demokritou P, Kane AB, Murphy CJ, et al. NanoEHS – defining fundamental science needs: no easy feat when the simple itself is complex. Environ Sci: Nano. The Royal Society of Chemistry; 2016;3: 15–27.

18. OECD Series on Testing and Assessment Number 34, “Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment”, ENV/JM/MONO(2005)14.

19. Klimisch HJ, Andreae M, Tillmann U. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul Toxicol Pharmacol. 1997;25: 1–5.

20. OECD Series on Testing and Assessment Number 194, “Guidance on Grouping of Chemicals, Second Edition”, ENV/JM/MONO(2014)4.

21. OECD Series on the Safety of Manufactured Nanomaterials, “Approaches on Nano Grouping/Equivalence/Read-Across Concepts Based on Physical-Chemical Properties

(GERA-PC) for Regulatory Regimes". ENV/JM/MONO(2016)3.

22. FDA, Physiologically Based Pharmacokinetic Analyses - Format and Content Guidance for Industry [Internet]. 2016. Available: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM531207.pdf

23. European Chemicals Agency Guidance, Appendix 4: Recommendations for nanomaterials applicable to the Guidance on Registration" Draft (Public) Version 1.0, January 2017.

24. Chemical Substances When Manufactured or Processed as Nanoscale Materials; TSCA Reporting and Recordkeeping Requirements. In: Federal Register. 12 Jan 2017.

25. Johnston, J.M., Lowry, M., Beaulieu, S., Bowles E. State-of-the-Science Report on Predictive Models and Modeling Approaches for Characterizing and Evaluating Exposure to Nanomaterials. U.S. Environmental Protection Agency, Office of Research and Development, Athens, GA. EPA/600/R-10/129. 2010.

26. Linkov I, Bates ME, Trump BD, Seager TP, Chappell MA, Keisler JM. For nanotechnology decisions, use decision analysis. Nano Today. 2013;8: 5–10.

27. Arts JHE, Hadi M, Irfan M-A, Keene AM, Kreiling R, Lyon D, et al. A decision-making framework for the grouping and testing of nanomaterials (DF4nanoGrouping). Regul Toxicol Pharmacol. 2015;71: S1–27.

28. Klanjscek T, Muller EB, Nisbet RM. Feedbacks and tipping points in organismal response to oxidative stress. J Theor Biol. 2016;404: 361–374.

29. Wang L, Nancollas GH. Pathways to biomineralization and biodemineralization of calcium phosphates: the thermodynamic and kinetic controls. Dalton Trans. 2009; 2665–2672.

30. Mäki U. Models are experiments, experiments are models. Journal of Economic Methodology. 2005;12: 303–315.

31. OECD, Series on Testing and Assessment No. 69, "Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models". ENV/JM/MONO(2007)2.

32. Falk A, Schimpel C, Haase A, Hazebrouck B, López CF, Prina-Mello A, et al. Research roadmap for nanosafety Part III: Closer to the market (CTTM) [Internet]. 2016. Available: https://www.nanosafetycluster.eu/uploads/files/pdf/CTTM_NSC_Roadmap_final_for_NSC.PDF

33. Stone V, Önlü S, Bergamaschi E, Carlander D, Costa A, Engelmann W, et al. Research priorities relevant to development or updating of nano-relevant regulations and guidelines. European NanoSafety Cluster [Internet]. 2017. Available: https://www.nanosafetycluster.eu/uploads/files/pdf/RRR_Final_version_090317.pdf

34. OECD, Series on Principles of Good Laboratory Practice and Compliance Monitoring No. 17, "Application of GLP Principles to Computerised Systems". ENV/JM/M ONO(2016)13.

35. Worth A, Aschberger K, Bofill DA, Bessems J, Gerloff K, Graepel R, et al. Evaluation of the availability and applicability of computational approaches in the safety assessment of nanomaterials [Internet]. 2017. Available:

http://publications.jrc.ec.europa.eu/repository/bitstream/JRC106386/kjna28617enn.pdf

36. Chomenidis, C., Drakakis, G., Tsiliki, G., Anagnostopoulou, E., Valsamis, A., Doganis, P., … Sarimveis, H. (2017). Jaqpot Quattro: A novel computational web platform for modelling and analysis in nanoinformatics. Journal of Chemical Information and Modeling, acs.jcim.7b00223. https://doi.org/10.1021/acs.jcim.7b00223

Appendix 1: Summary of Database Projects (2010-2017)

The NSC Working Group on Databases together with the caLIBRAte project, distributed a database survey in December 2016. Thirty-two responses were received, from the following projects: Cerasafe, Cosmo, DaNA, eNanoMapper, Handbook of Chemistry and Physics, HSE Nano, Keele University (several projects), Mercury, NanoFate, NanoImpactNet, NanoMILE, NanoPUZZLES, NANoREG, Nanosolutions, Nanovalid, NECID, Neptune, S2NANO, Sanowork, Scaffold, Serenade, SIRENA, SUN, TINE, UK NanoRegister, and VieilleNanos. According to the responses, the majority of types of data and information on nanomaterials collected by the responding projects (multiple answers possible) were on physicochemical characterization (24), *in vitro* toxicity (17), *in vivo* toxicity (17), ecotoxicology (14), human exposure (12), or environmental release/fate (10). Other questions of the survey addressed the main objective(s) of the database, database design and implementation, database availability/accessibility, the use of semantics technology methods, the data collection and curation, the copyright and licensing aspects. The results of the survey will be published on the NanoSafetyCluster website. Further details of selected projects are given below.

A1.1 eNanoMapper

The EU FP7 project eNanoMapper ran from February 2014 to February 2017 and developed a computational framework for nanomaterials toxicological data, which is based on open standards, open source, common languages, and an interoperable design, enabling a more effective and integrated approach to risk assessment. eNanoMapper has created a modular, extensible infrastructure for transparent data sharing, data analysis, and the creation of computational toxicology models, which aims to support data management in the area of nanoEHS and to enable an integrated approach for the risk assessment of nanomaterials. To achieve these, eNanoMapper developed an ontology, a data infrastructure and modelling tools with applicability in risk assessment of nanomaterials. The ontology includes common vocabulary terms used in nanosafety research. The database includes functionalities for data protection, data sharing, data quality assurance, search interfaces for different needs and usages, comparability and cross-talk with other databases (https://search.data.enanomapper.net). A collection of descriptors, computational toxicology models and modelling tools were developed, enabling the use and integration of nanosafety data from various sources [5–7], including web tools: Jaqpot (http://www.jaqpot.org, [36]) which allows online

Modelling (building and validating models), Read-across, Interlaboratory comparison and Experimental Design, while Nano-lazar, available at https://nano-lazar.in-silico.ch/predict, offers online Read across toxicity predictions. The project also provided a rich library of information and documentation (tutorials, webinars, reports and publications) to support and guide the users. In addition, a collection of modelling tools developed within FP7 nano modelling projects was created: http://www.enanomapper.net/nsc-modelling-tools

A1.2 NECID

Under the leadership of IFA (Institute for Occupational Safety and Health of the German Social Accident Insurance) and TNO (TNO – innovation for life) a working group of PEROSH (Partnership for European Research in Occupational Safety and Health) institutes developed and tested a database software called NECID (Nano Exposure and Contextual Information Database). The software supports the user to collect and store data of exposure measurements ofNOAA (Nano-Objects and their Agglomerates and Aggregates). In addition to measurement data of individual instruments the collection and documentation of work conditions, or so called "contextual information", is a focus of this project.

The NECID software includes a nanomaterial specific exposure database, as well as features for data sharing and data assessment. The software runs locally on a computer but also offers a web-based central database for the exchange of information. A key factor for the project is the harmonization of "nano exposure measurements" and their documentation. Therefor NECID uses, as far as possible, a harmonised ontology to enable links to other databases. During the construction of NECID, cooperation and exchange of information to other projects like NANoREG, Marina, caLIBRAte, GUIDEnano were important parts of the work..
After an intensive testing phase within the project a software license for NECID is available to every organization dealing with the challenge of handling NOAA or the risk assessment of these tasks. At the moment the license is free of charge. For further information please contact NECID@DGUV.de or visit the webpage WWW.NECID.eu.

A1.3 SERENADE

CEREGE-Labex SERENADE is the primary contact in Europe for the US database efforts led by CEINT– Duke University with ongoing effort on data management, curation and with the US-nanoinformatics program as to determine a strategic plan for data standardization, templates and guidance documents for data harmonization between Europe and USA. Discussions were also active during the ProSafe –OECD conference in Paris (end of 2016) to link EU and US databases (interoperability, ontology, data exchange formats). The CEINT group works in close collaboration with the EU Nanosafety Cluster Database Group and the EU-US Database CORs (Community of Research) on templates harmonization and especially on the NanoReg templates and format. All partners to share expertise for products stability assessment (simulation of products use), environmental fate study, ecotoxicology, end of life with the ProSAfe project.. and develop common set up, protocols in order to compare data and implement exposure models.

A1.4 GuideNano

A web-based Exposure Scenario Library has been developed within the GUIDENANO project to read-across the exposure scenarios. The library includes contextual information (NMs properties, task description, exposure controls) and measurement data of 200 occupational exposure scenarios covering a wide range of NMs (CNT, CNF, SiO2, ZnO, Ag..). The library can be searched by NM name, life-cycle, source domain, contributing exposure scenario. The ES Library is hosted online and managed by IOM and available using the link http://guidenano.iom-world.co.uk/. GuideNano partners continue to work with eNanoMapper and other members of NSC Working Group to map the ES Library variables with those already available in the eNanoMapper database and to add new terms if necessary with the aim of constructing an exposure ontology and ultimately to make all the exposure data available via the database developed in eNanoMapper.

A1.5 SUN

The SUN project has successfully accomplished the design, implementation and population of a web-based data repository, a searchable operational project database to store and maintain the data generated by the project. An extensive exercise was carried out with SUN project partners to develop data collection templates, procurement, completeness, quality-checked, collation and storage of the scientific project data into a flexible and user friendly operational database. The implemented database provides facilities to search, query and retrieve selected project data-sets. We anticipated sharing and uploading the SUN data to an instance of the "final" eNanoMapper database early on in the project however, data sharing permissions, embargos etc. needs to be formalised with SUN project partners. To advance this, SUN partners are currently involved in further related developments, having been contacted by the NANOREG2 and CaLIBRAte projects, aiming to supply them with final SUN data.

A1.6 NanoInformatics Knowledge Commons (NIKC)

The NanoInformatics Knowledge Commons (NIKC) Database was designed by the Center for Environmental Implications of NanoTechnology (CEINT) to gather engineered nanomaterial exposure and toxicity data into an organizational structure permitting readily accessible data for broader scientific inquiry. The NIKC consists of a database (DB) and associated applications for data entry and data analysis; the DB contains CEINT data as well as data extracted from published literature, and is accessible to CEINT members as well as NIKC collaborator groups in the US and abroad. The NIKC is an important component in realizing the goals of CEINT, which include: elucidating the general principles that determine nanomaterial behavior in the environment; identifying data and metadata necessary to support forecast of exposure potential, bioaccumulation, and bioactivity; and identifying key functional assays [8] that are predictive of measurements of interest.

The NIKC supports development of analytical tools such as the Nano Product Hazard and Exposure Analytical Tool (NanoPHEAT), a custom-built app designed to graphically indicate exposure risk outcomes from products incorporating engineered nanomaterials. CEINT has also adopted management of the community-drive(n) ISA-TAB-Nano project [9], which establishes consistent file-sharing formats for nanomaterial data to enable integration of information even in advance of formally established standard(s) processes. ISA-TAB-Nano was developed by the National Cancer Informatics Program's Nanotechnology Working Group (NCIP NanoWG) and has been adopted and adapted by a number of projects including the EU-wide NANoREG project. CEINT is leading the community-based effort to expand the standardized protocol templates used to develop consistent and comparable data, with particular focus on including critical elements of nanomaterial datasets identified via CEINT's work. These include: transformation and exposure endpoints, inclusion of media parameters within the primary dataset describing nanomaterial characterizations, and functional assay measurements used to predict (exposure and hazard) outcomes of interest.

A1.7 QsarDB

QsarDB has been developed over the course of past decade within several EU funded and national (in Estonia) research initiatives (see www.qsardb.org). It is a general repository solution for organizing, storing, preserving and using QSAR models. It is also designed for accommodating nano-structures and nano-materials. The storage of QSAR models and related data is a complicated issue and available storage solutions have been reviewed recently [REF1]. QsarDB is open and gives freedom to develop model to the developer and allows preserving and efficient reusing of models. What is equally important, it gives an easy access to QSAR models to potential users, providing transparent view to the constituents of the model and allows independent verification. QsarDB consists of several components (e.g. data format, repository and tools). Qsar DataBank data format [REF2] is a format for representing QSAR model information (data and models) in systematic and machine readable way. Qsar DataBank data format is generic and has been also used for Quantitative nano-Structure-Activity Relationships [see example collection of models http://hdl.handle.net/10967/120]. The format is extendable, for example to include further developments for models with nanostructures and nanoparticles. The archives in Qsar DataBank data format can be freely deposited to the QsarDB smart repository [REF3]. The QsarDB smart repository is a practical resource and tool that enables research groups, project teams and institutions to share, present and use Quantitative Structure-Activity Relationships data and models. At the moment, the repository includes over 400 (Q)SAR models, is expanding and developed further.

A1.8 GRACIOUS

The newly funded GRACIOUS H2020 project will continue the efforts of the above projects to establish a data curation system, which will be developed based on the eNanoMapper database and on elements and templates from other relevant nanosafety data inventories such as NANoREG, NanoReg2, DANA 2.0, SUN, MARINA and NanoETox

to allow both the integration of newer data and the use of raw and aggregated data for regulatory risk assessment and Stage-Gate innovation decision making. This data curation system will be designed to allow seamless integration with a variety of modelling tools (ranging from simple rules and theoretical models to complex *in silico* (e.g. Q(n)SP/AR) algorithms) into an interoperable data and modelling 'infrastructure'. This 'infrastructure' will be connected to the GRACIOUS interoperable module for grouping and read-across of nanoforms to deliver to it curated data and computing capabilities. The module will be specifically designed to enable existing user-friendly risk assessment and management software tools (e.g. SUNDS, caLIBRAte SoS) to perform grouping and read across. Its results will be delivered as easy to comprehend dynamic charts and textual reports to facilitate further analysis and/or decision making.

References in this paragraph/chapter:

[REF1] Sild, S.; Piir, G.; Neagu, D.; Maran, U. Storing and using quantitative and qualitative structure–activity relationships in the era of toxicological and chemical data expansion, in Big Data in Predictive Toxicology (series Issues in Toxicology), Eds. D. Neagu, A. Richarz, Royal Society of Chemistry, 2017, in press.

[REF2] Ruusmann, V.; Sild, S.; Maran, U. QSAR DataBank - an approach for the digital organization and archiving of QSAR model information. J. Cheminf. 2014, 6:25. DOI: 10.1186/1758-2946-6-25

[REF3] Ruusmann, V.; Sild, S.; Maran, U. QSAR DataBank repository: open and linked qualitative and quantitative structure–activity relationship models. J. Cheminf. 2015, 7:32. DOI: 10.1186/s13321-015-0082-6

.