DELIVERABLE D3.1

Technical Specification and initial implementation of the protocol and data management web services

GRANT AGREEMENT:	604134
ACRONYM:	eNanoMapper
NAME:	eNanoMapper - A Database and Ontology Framework for Nanomaterials Design and Safety Assessment
PROJECT COORDINATOR:	Douglas Connect GmbH
START DATE OF PROJECT; DURATION:	1 February 2014; 36 months
PARTNER(s) RESPONSIBLE FOR THIS DELIVERABLE:	IDEA, DC, IST, UM
DATE:	31.1.2015
VERSION:	V.1.0

SEVENTH FRAMEWORK PROGRAMME

Call identifier	FP7-NMP-2013-SMALL-7
Document Type	Deliverable Report
WP/Task	WP3/T3.1,T3.2,T3.3
Document ID	eNanoMapper D3.1
Status	Final

Partner Organisations	• Douglas Connect, GmbH (DC) • National Technical University of Athens (NTUA) • In Silico Toxicology (IST) • Ideaconsult (IDEA) • Karolinska Institutet (KI) • VTT Technical Research Centre of Finland (VTT) • European Bioinformatics Institute (EMBL-EBI) • Maastricht University (UM)
Authors	Authors: Philip Doganis, Bengt Fadeel, Roland Grafström, Janna Hastings, Markus Hegi, Nina Jeliazkova, Vedrin Jeliazkov, Cristian Munteanu, Haralambos Sarimveis, Bart Smeets, Georgia Tsiliki, David Vorgrimmler, Egon Willighagen Reviewed by Barry Hardy (DC)
Purpose of the Document	Report on the Technical Specification and initial implementation of the protocol and data management web services
Document History	Enter document modification: 1. Table of Contents, 03/11/2014 2. First draft, 10/11/2014 3. Second draft 08/12/2014 4. Third draft, 27/01/2015 5. Final Version, 31/01/2015

TABLE OF CONTENTS

TABLE OF FIGURES

GLOSSARY

Abbreviation / acronym	Description
API	Application Programming Interface
OECD HT	The OECD Harmonised Templates are standard data formats for reporting studies done on chemicals to determine their properties or effects on human health and the environment.
OpenAM	OpenAM is an open source access management, entitlements and federation server platform.
OpenTox	OpenTox is a predictive toxicology framework with a unified access to toxicological data, (Q)SAR models and supporting information developed under the grant agreement FP7-HEALTH-2007-A- 200787
REST	Representational state transfer (REST) is an abstraction of the architecture of the World Wide Web; more precisely, REST is an architectural style consisting of a coordinated set of architectural constraints applied to components, connectors, and data elements, within a distributed hypermedia system. REST ignores the details of component implementation and protocol syntax in order to focus on the roles of components, the constraints upon their interaction with other components, and their interpretation of significant data elements.
ToxBank	ToxBank (developed under grant agreement FP7-HEALTH-2010-Alternative-Testing- 267042) establishes a dedicated web-based warehouse for toxicity data management and modelling, a 'gold standards' compound database and repository of selected test compounds, and a reference resource for cells, cell lines and tissues of relevance for *in vitro* systemic toxicity research carried out across the FP7 HEALTH.2010.4.2.9 Alternative Testing Strategies SEURAT program.

1. EXECUTIVE SUMMARY

We reviewed the current state of the art of availability and requirements for nanomaterial databases, through reviewing the literature, online resources, engaging in discussions with the NanoSafety Cluster working groups and using the results of a community-based requirement analysis. The eNanoMapper data architecture was developed and consists of a set of web services, providing access to experimental protocols and data, search service and modules, facilitating linking and data transfer between third party databases. This design is expected to facilitate adding new services of any kind, for example supporting different data types and data analysis. The technical solution is able to support the required data types, queries, and annotations, and can enable user friendly applications. The resulting eNanoMapper data access architecture was developed based on the OpenTox framework and Application Programming Interfaces and extensions to them. An eNanoMapper prototype supporting the inclusion of data and protocols was developed and deployed, based on a number of public datasets relevant to Nano Safety.

TODO

2. INTRODUCTION

2.1 EXISTING DATABASES

Several databases, relevant for ENM toxicity assessment, exist. They list nanomaterials and a variety of their properties, or products, containing nanomaterials: EC JRC's NanoHub (www.napira.eu), ModNanoTox database (IST participation), NanoWiki (maintained by an eNanoMapper partner), NanoMaterialRegistry (www.nanomaterialregistry.org), Nanoparticle Information Library NIL (nanoparticlelibrary.net), Nanomaterial-Biological Interactions Knowledgebase (nbi.oregonstate.edu), caNanoLab (cananolab.nci.nih.gov/caNanoLab/), InterNano (http://www.internano.org/), Nano-EHS Database Analysis Tool (http://icon.rice.edu/report.cfm) , nanoHUB (https://nanohub.org/resources/databases), NanoTechnology Characterisation Laboratory (http://ncl.cancer.gov/), the DaNa Knowledge Base (nanopartikel.info), and NanoWerks Nanomaterial Database (www.nanowerk.com/).

The EU NanoSafety Cluster alone (www.nanosafetycluster.eu) has many projects with database generating activities, transferring information on topics including aquatic bioactivity (e.g. ENNSATOX), exposure- dose-response data (e.g. ENPRA), human and ecology hazard (e.g. NanoImpactNet), life cycle assessment (e.g. NanoFATE), and bio-distribution (e.g. ENPRA, HINAMOX). Most of these cover specific aspects of nanotoxicity and lack sufficient content for data analysis and model building. Various supporting community efforts exist (e.g. QualityNano, NHECD, NANOfutures), but for this purpose it is necessary to further implement facilities for linking, processing and retrieval of data from existing ENMs databases and also for the addition of new data originating e.g. from the NanoSafety Cluster.

Product databases have recently been reviewed in a study[1] identifying three databases with nano-specific products[2] and two general products databases. This study 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 products 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[3]. While product databases are very important in the context of protecting consumer health and the environment, they usually contain insufficient characterization and properties of the nanomaterials for use in research. Approaches and feasibility of supporting product databases are discussed in a recent document by JRC[4].

Reviews of emerging databases and analysis tools in nanoinformatics have started to appear in the literature [11] , including not only nano materials specific databases and tools, but also generic toxicology databases, molecular modelling and image processing tools.
There are a number of nanomaterials entries in widely used chemical databases and toxicogenomics databases[5], as well as in the REACH registration database. Below we provide several summaries, the full list of relevant databases can be found in Table I of [6].

PubChem: 117M deposited substances, 74M standardized compounds, 650K assays[7], 200M test results on substances. No specific support for nanomaterials (no description of the material composition;

carbon nanotube assays found under methane compound entry), however a number of substances can be inferred to be nanomaterials: (>200 fullerenes; metal oxides; silver nanoparticles; colloidal gold nanoparticles, etc.).

ChEMBL19: 1.4M chemical compounds; 1.1M assays; 6 assay types[8]. 5K different activity types; 12.8M activity values. The assay protocol is not explicitly given, instead the assays are annotated by literature reference (57K entries) and targets (26 target types, including single proteins, protein families, cell lines, tissues, organisms). Nanomaterials are not explicitly supported, however, a number of fullerenes can be found, as well as biological data for metal oxides (e.g. CHEMBL1201136 for titanium dioxide).

ArrayExpress: European public repository of gene expression data. Nanomaterials: carbon nanotubes, quantum dots, graphene oxide, zinc oxide, silver and gold nanoparticles [9].

Gene Expression Omnibus (GEO): US public repository of gene expression data. More than 1000 hits found in GEO data sets, when searching for term "nanomaterial". www.ncbi.nlm.nih.gov/geo/.

Chemical Effects in Biological Systems: Repository for public data from US National Toxicology Program. Contains a large number of diverse experiments. No specific support for nanomaterials, but includes assay data on metal oxides and fullerenes. cebs.niehs.nih.gov

Comparative Toxicogenomics Database: Curated chemical, gene and disease connections and tools to analyse chemicals, genes and gene signatures (over 10000 compounds). Does not explicitly support nanomaterials, but includes nanomaterial related data. However, the annotations are not always precise (e.g. the term titanium dioxide is considered equivalent with rutile, anatase, brookite, nano-TiO2 (C009495)).

AcTor: US online warehouse of publicly available chemical toxicity data. Does not explicitly support nanomaterials, but includes data for e.g. fullerenes. Silver nanoparticles are annotated as synonym for the generic entry for silver; carbon nanotubes are included as a synonym of graphite; anatase, rutile and titanium oxide are listed as synonyms. actor.epa.gov

The ECHA Dissemination site provides information on registered chemical substances under REACH. The data shown is compiled from joint or individual submissions for a substance. The REACH dossiers format is compliant with the OECD Harmonized Templates (OHTs), with mandatory sections of substance identification and substance composition. While the OHTs do not explicitly support detailed description of nanomaterials (apart from denoting whether the substance is a nanomaterial or not), its data model allows to describe different manifestations of the same chemical composition, and, for example, distinguish between different crystal structures of titanium dioxide (CAS 13463-67-7), anatase (1317-70-0), and rutile (1317-80-2) forms. echa.europa.eu/en/informationon-chemicals/registered-substances

NanoMiner: Over 600 human samples exposed to nanoparticles[10]. All the samples have been annotated, pre-processed and normalized to enable users to utilize the database systematically across the different experimental setups and platforms. Pre-computed analysis results are saved in the database to facilitate visualization and statistical analyses. It allows the user to: 1) search for and plot expression profiles, 2) cluster the samples within the datasets, 3) search for differentially expressed genes across several datasets, 4) analyse enriched KEGG-pathways and GO classes for the detected genes.

2.1.1 DATABASE REVIEW

We have performed an exhaustive search for existing nano-related databases in the first quarter of 2014 and identified 104 potential data sources. A subset of 34 was publicly available online on the Internet.

Most of these sources do not provide machine readable data (eighteen consist of simple web pages (HTML), ten contain PDF documents). Excel tables are available from three sources; database dumps (e.g. MySQL) are available from another three sources. One source provides data in ISA-Tab-Nano format, one in IUCLID5 format and one is based on Semantic MediaWiki. Programmatic access through a publicly available API can be implemented for only four of the sources. Only one source makes a distinction between raw and processed data and provides access to both types of data.

Given the above findings, it becomes clear that nano-related data is relatively abundant, but also quite dispersed across many different sources. Combining data from various sources is hampered by the lack of programmatic access in most cases and the absence (or infrequent use) of suitable domain ontologies.

2.1.2 DATA INTEGRATION APPROACHES

The selection of the database integration strategy depends on the answer to the following two main questions: first, how to establish the correspondence between entities in different databases; and second, how to integrate query results? The integration approaches can be classified based on the number of data models used (single or multiple) and storage (single or multiple). Matching entities from different data sources could be done either by converting the data sources into a common data model, or keeping the data models distinct and establishing equivalence between the separate entities. Equivalence in the latter case can be established through the use of shared ontology identifiers to annotate the data. A particular challenge in matching entities from different sources across either entity matching approach is the need to define the identity criteria for entities of that type. This challenge is partly addressed for small molecules due to the widespread adoption of database-independent unique structure-based identifiers such as the InChI, but for nanomaterials this remains an open challenge. All query integration approaches require entity matching. It could be done during the conversion to a common data model (index time merging) or on the fly (query time merging, federated search), as well as through a hybrid approach. Technology-wise, Open PHACTS uses a single triple store and provides an Application Programming Interface (API), but not a generic query interface (SPARQL Protocol and RDF Query Language)[12]; Toxygates combines accessing metadata through remote SPARQL and NoSQL solution for storing and querying 800M data points for 170 chemicals [13]; EBI's RDF platform uses triple stores and a federated SPARQL endpoint [14]. OpenTox is a distributed system with a common API and data model and a number of different independent implementations, featuring diverse backend solutions and programming languages[15].

2.2 NANO MATERIAL DESCRIPTION

This section first summarises the different requirements for the eNanoMapper framework, posed by the nanotechnology community; and provides an overview of the existing formats and frameworks, suitable for ENM description.

2.2.1 API REQUIREMENTS

The eNanoMapper API must be able to capture the physical and chemical identity of ENMs, including the notion of mixtures and the resulting size distribution, differences in amount of surface modification, manufacturing conditions, batch effects, etc. It must also capture the biological identities (e.g. toxicity pathways, effects of ENM coronas, modes-of- action), interactions (cell lines, assays), and a wide variety of measurements. A number of analytic techniques have been adopted and developed to characterise nanomaterials physicochemical properties, including the commonly used dynamic light scattering to measure the particle size distribution and zeta potentiometry to estimate the pH- dependent surface charge. However, with expanding insight into the factors determining toxicity, this list is growing increasingly long. The need for validated *in-vitro* tests has been advocated since 2006[16]. It is proposed to

extend the list of endpoints for hazard identification to include cell uptake, cell viability, oxidative stress, inflammation, fibrosis, immunotoxicity, cardiovascular toxicity, ventilation rate, gill pathologies, mucus secretion, brain pathology, and animal behaviour. The EU guidance document lists the main known effects from experimental studies[17]. High throughput omics data and kinetics[18] are becoming of increasing importance in nanomaterials assessment.

Another API requirement is to be able to represent studies and results of the studies of the toxicology or biological interference of the nanomaterial, in addition to an accurate physicochemical characterisation. The API should enable linking to the corresponding protocols and data sources, where available. Clear visualization of nanomaterials that goes beyond just structural formulae should be available, in order to make the data less abstract for nano-inexperienced biologists.

The API should allow the representation of data and facts compatible with regulatory expectations and (inter) national standards. This usually translates into a set of available study summaries (rarely raw data) for a given ENM. Including links to product databases could also be considered (e.g. whether the nanomaterial occurs in nature, whether it is emitted by cars or if it is present in certain food sources, as well as known therapies in which the nanomaterial is used).

The modelling community presents a different requirement: the data analysis methods usually require a "spreadsheet" or matrix view of data for multiple ENMs. The experimental data in the public datasets is usually not in a form appropriate for modelling. Standardisation in these sources is specific to each database. Even in curated collections the preparation of data for modelling is not a straightforward exercise (e.g. the experimental values can be merged in many different ways into a matrix, depending on which experimental protocols and conditions are considered similar; also there could be multiple values due to replicates or similar experiments). The API should allow adding information based on the outcomes of the predictive toxicology models, including biological role of the ENM, clearance, accumulation, and pathway information (e.g. WikiPathways entries[19]). The modelling API is tightly integrated with data API, and is subject of a separate deliverable D4.1.

2.2.2 ISA-TAB/ISA-TAB NANO

The data models behind existing formats, specifications and systems are described in the following sections.

ISA-TAB/ISA-TAB NANO

ISA-Tab (isatab.sourceforge.net) is a universal text format, based on minimal information standards and is increasingly being adopted as a base standard for representing experiments, including toxicity studies, physicochemical analysis, high throughput and high content datasets from omics (e.g. proteomics, transcriptomics, metabolomics) experiments[20]. The ISA-Tab structure is composed of Investigation, Study, and Assay files. The Investigation is a high level concept to link related studies; the Study is the central ISA-Tab unit containing information on the subject under study (e.g. biological sample or nanomaterial), its characteristics and any treatments applied. The Assays comprise the tests performed either on material taken from the subject or on the whole initial subject, which produce qualitative or quantitative measurements. ISA-Tab is designed to capture the experimental graph by explicitly specifying the protocols used to transform one entity into another (e.g. cell culture, chemical compound or nanomaterial). This ensures that the information on materials origin (provenance) and subsequent processing is required and retained in the database. The ISA-Tab format specifies a limited number of mandatory fields, describing the minimum metadata about an experiment. Different types of experiments are accommodated by a flexible template mechanism, and a specific tool (ISAValidator) is available to validate the input data, according to the defined template. Specific templates could be defined to reflect particular experimental techniques and assays. As the format defines the metadata

only, the data itself could be provided by experiment- specific files (e.g. spreadsheets or Affymetrix CEL files or images) or remote resources. Almost all fields (e.g. materials and experimental factors) can be annotated or defined by ontologies, which gives another dimension of rigorous validation and adherence to specific vocabularies.

The default approach for representation of chemical compounds in ISA-Tab is an ontology entry, which typically points to a single chemical structure. This is insufficient for describing substances of complex composition and nanomaterials, and is the reason for the introduction of a material file in ISATab-Nano[21]. The material file contains information about material name, synthesis, intended application, constituents, linkage and size. The latest ISA-Tab- Nano 1.2 specification recommends using the material file only for nominal characteristics, and to describe the experimentally determined characteristics in regular ISA-Tab assay files. The Material Linkage column is replaced by the Material Constituent column with the linkage type specified in a separate column. The specification recommends using separate material files for materials with different chemical composition or physical characteristics.

OECD HT

The OECD Harmonized Templates (OHTs) are structured (XML) data formats for reporting safety-related studies on chemical substances. The OHTs and the supporting IT tool (IUCLID5, iuclid.eu) are used in a regulatory context, for preparation of substance dossiers for REACH and for other regulatory frameworks operating in Europe; as well as by the JRC NanoHub database. The substance identification section is compliant to "ECHA guidance for identification and naming of substances under REACH and CLP" and requires specification of detailed chemical composition (including impurities and additives), concentrations of each constituent (typical and measured), 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 is 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[a]. 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. 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. Specific guidance for NM registration under REACH has been issued

[a] http://www.oecd.org/chemicalsafety/testing/oecdguidelinesforthetestingofchemicals.htm

by ECHA. A safety data sheet needs to be prepared for all substances, including nanomaterials classified as hazardous[a].

ONTOLOGIES (BIOASSAY, NPO)

The eNanoMapper strategy to adopt and extend ontologies in support of data integration has recently been described in[22] and is the subject of the eNanoMapper report "Initial Ontology Release" (D2.3). NanoParticle Ontology and Bioassay Ontology are among the ontologies considered to reuse.

CODATA UDS

The CODATA group published a working draft on its "Uniform Description System for Materials at the Nanoscale"[23], a result of consolidating the knowledge from a variety of user communities and experts. The UDS considers several aspects of ENMs: chemical composition (atomic composition, molecular composition, chemical moieties, including percentages, chemical identifiers such as CAS and InChI), characterisation (shape, size, physical structure number of layers, shells, crystallographic structure, surface description), intensive properties (melting point, conductivity) and interaction properties; as well as documentation of production and post-production of ENMs. The terms of Uniqueness and Equivalence of ENMs are defined.

NANOSAFETY CLUSTER APPROACHES

As a result of the interactions with the NanoSafety cluster and NSC Database working groups, the eNanoMapper consortium has received a large set of custom spreadsheet templates, without or with only sample data. It is evident, that the Excel templates are the preferred approach for data preparation of the majority of the NanoSafety Cluster projects. One project claims using ISA-TAB-Nano, but in fact they use again custom Excel templates and automatically convert them to ISA-TAB-Nano.

- NanoPuzzles provided a set of Excel templates for a large set of physicochemical and toxicological assays (zeta potential, surface charge, solubility, size (TEM, DLS), reactivity (rate of change), porosity, dissolution rate, crystallinity, agglomeration_size, adsorption, *in_vivo* genotoxicity, *in vivo* cytotoxicity, *in vivo* cell viability). The structure roughly (but not exactly) follows the ISA-TAB –Nano. The NanoPuzzles project developed a script to generate ISA-TAB-Nano files from these templates. There is a feature request for ISA-TAB-Nano parser.
- NanoREG provided a set of Excel templates (TEM: size, shape, surface charge, organ burden, DNA damage *in vivo* (comet assays), cell counts, mRNA expression *in vitro*, micronucleus assay in bone marrow cells, genotoxicity (comet, micronucleus), immunotox, ocular irritation , cell viability, chromosomal damage *in vitro* (micronucleus assay in BEAS 2B cells), protein secretion *in vitro*. The templates are organized by the measurement technology (one Excel template per protocol) and are not ISA-TAB compliant. The NanoREG templates are designed for ease of use by human operators. The eNanoMapper is expected to provide convertors for different formats.
- ModNanoTox has the data retrieved from literature structured (variety of assays) in an Excel file with four spreadsheets (Study details, Particle details, Assay details, Study outcomes).

We have started the development of a configurable parser[b], that enables the import and conversion of the data stored in a supported set of spreadsheet templates, accommodating different row-based, column-based or mixed organisation of the data.

[a] http://publications.jrc.ec.europa.eu/repository/bitstream/111111111/31575/1/reqno_jrc88931_considerations_information_needs_nm_consumer_products_online.pdf

[b] https://github.com/enanomapper/nmdataparser

RAW DATA

The NanoSafety cluster templates typically do not include raw data files. However, supporting raw data files (including microscopy images) is an important requirement, enabling the modelling WP to provide services to process the raw data files and derive descriptors and summaries. For this purpose we follow the ISA-TAB strategy of providing links to the raw data files. Extensions of the API to support links to the raw data files are under development.

For the physical storage of raw data files there are different solutions (ftp servers, cloud storage, data repositories as OpenAire, scientific images management, etc.), which will be evaluated during the next reporting period. There are a number of open source (BisQue[a], The Open Microscopy Environment [b], Euro-Bioimaging [c]) and commercial systems (Columbus Image data storage [d], QUARTZ PCI[e]) for managing scientific images. Assessment and selection of the solution for supporting raw data files is planned for the next reporting period.

2.3 PROTOCOLS AND REFERENCE MATERIALS

The test guidelines for NM characterisation encompass large number of protocols ranging from standardized (by ISO[24], OECD, ASTM, CEN) ; not yet well established in the scientific community or research protocols under development[25]. Requirements on measurements for the implementation of NM definition are published in[26]. The individual measurement methods are summarized in a recent JRC Reference Report EUR 2540442[27].

A number of NanoSafety cluster projects are involved in developing and validating protocols (NanoGenoTox [f], NanoREG, QNANO [g], OECD, ECHA[h], NCL[i]). Upon recommendation by the NSC Hazard assessment working group [j] chair, we have considered for the initial implementation the collections of protocols maintained by NanoImpactNet (http://www.nanoimpactnet.eu/index.php?page=Researchprotocols) and DaNa projects http://www.nanoobjects.info , http://iai-dana.iai.fzk.de/en/nanoinfo/methods/992-standard-operating-procedures.

For the assessment or the applicability of methods and new methods development and validation a list of reference materials has been proposed[28]. The representative manufactured nanomaterials hosted by JRC-IHCP[29] includes: Fullerenes (C60), Single-walled carbon nanotubes (SWCNTs), Multi-walled carbon nanotubes (MWCNTs)[30], Silver nanoparticles[31], Iron nanoparticles, Titanium dioxide[32], Aluminium oxide, Cerium oxide[33], Zinc oxide[34], Silicon dioxide[35], Dendrimers, Nanoclays, Gold nanoparticles, Graphene, nanotubes , quantum dots. The NIST standard reference materials can be found at https://www-s.nist.gov/srmors/browseMaterials.cfm?subkey=42&tableid=231

[a] http://bioimage.ucsb.edu
[b] http://www.openmicroscopy.org/site
[c] http://www.eurobioimaging.eu/content-page/about-euro-bioimaging
[d] http://www.perkinelmer.com/pages/020/cellularimaging/products/columbus.xhtml
[e] http://www.quartzimaging.com/microscope-digital-image-acquisition-and-proces
[f] http://www.nanogenotox.eu/index.php?option=com_content&view=article&id=136&Itemid=158
[g] http://www.nanosafetycluster.eu/working-groups/2-hazard-wg/protocols/other-protocols.html
[h] http://echa.europa.eu/documents/10162/13632/appendix_r7a_nanomaterials_en.pdf
[i] http://ncl.cancer.gov/working_assay-cascade.asp
[j] http://www.nanosafetycluster.eu/working-groups/2-hazard-wg.html

3. TECHNICAL SPECIFICATION

The eNanoMapper data architecture is based on existing developments of consortium partners and consists of a set of web services, providing access to experimental protocols and data, search service and modules, facilitating linking and data transfer between third party databases. This design is expected to facilitate adding new services of any kind, for example supporting different data types and data analysis. With the OpenTox API[a] and ToxBank API[b] as a starting point, we have reviewed the current state of the art of availability and requirements for nanomaterial databases, through reviewing the literature, online resources, NanoSafety Cluster working groups and meetings and WP1 requirement analysis. The technical solution is able to support the required data types, queries, annotation and enable user friendly applications. The outcome of this task is a description of the eNanoMapper data access architecture and API.

3.1 ARCHITECTURE

The eNanoMapper architecture has been informed by previous experience in designing and building a predictive toxicology framework for chemicals (OpenTox[36]). eNanoMapper currently adopts the OpenTox framework design, based on the following technological choices (i) the REpresentational State Transfer (REST) software architecture style allowing platform and programming language independence and facilitating the implementation of new data and processing components; (ii) a common information model, supporting ontology annotation; communication through well-defined interfaces ensuring interoperability of the web components; (iii) Authentication and authorisation, allowing defining access policies of REST resources, based on OpenAM. The system architecture consists of a set of web services, providing access to protocols and data, search services, and enabling development of GUI and libraries, offering user-friendly access to the above functionality. The web services, currently developed by partners could run on the same machine, or on geographically dispersed servers, and communicate via the Internet. This design is expected to facilitate adding new services of any kind, for example supporting different data types or search functionality.

While the OpenTox framework is intentionally chemical compound centric, eNanoMapper uses an extension, allowing representation of chemical substances with defined composition and experimental data, associated with substances, rather than associated with the chemical structures. The ENMs are considered a special case of substances.

3.2 API

3.2.1 PROTOCOLS

The protocol management service allows uploading metadata and textual description of an experimental protocol (lab or *in-silico*). The protocol access and upload procedures, as well as the metadata serialization in semantic web format, builds upon existing API and open source implementation[c], enhanced according to the specific eNanoMapper requirements. The outcome of this task is a REST web service implementation of the defined Protocol API, as well as a web service for user management and interaction with the OpenAM authentication and authorization solution.

[a] http://opentox.org/dev/apis/api-1.2
[b] http://api.toxbank.net/
[c] https://github.com/enanomapper/toxbank-api-server

3.2.2 DATA

The Nano Particle Ontology (NPO) defines a Nanomaterial (NPO_199) as equivalent class of chemical substance (NPO_1973) and one of (nano-object, nanoparticle, engineered nanomaterial, nanostructured material, nanoparticle formulation). The chemical substance itself is a subclass of chemical entity (NPO_1972). The definition of the terms "substance" and "material" are discussed in[37] ,comparing ISO, REACH and general science definitions of the terms. The REACH definition of a substance encompasses all forms of substances and materials on the market, including nanomaterials; and may have complex composition. The paper notes the OECD HT definition of "reference substances" is very similar to the definition of the term "reference material".

The same publication refers to the "test" and "measurement" terms as the fundamental concepts[37]. The 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[28] by OECD WPMN (Table 1) use the terms and categories from the OECD Harmonized Templates. The NPO distinguishes between endpoint of measurement (e.g. particle size NPO_1694) and assay used to measure the endpoint (e.g. size assay NPO_1912), where the details of the assay could be specified (e.g. uses technique electron microscopy NPO_1428). This structure is generally the same as the one supported by the OECD templates (e.g. in the OECD HT granulometry type of experiment several size-related endpoint can be defined, as well as the equipment used, the protocol and specific conditions). The CODATA UDS also requires specification of how particular property is measured. The ISA-TAB-Nano also allows defining the qualities measured and detailed protocol conditions and instruments. The level of details in the OECD HT, CODATA-UDS, ISA-TAB-Nano and available ontologies differ, which is due to their original focus. Mapping between terms defined in the different sources is an ongoing effort in collaboration with WP2 and NanoSafety cluster WG4.

Table 1. Endpoints recommended by OECD WPMN and relation to CODATA UDS and ISA-TAB-Nano

#	Endpoints agreed by the OECD WPMN	OECD HT (XML schema)	CODATA UDS	ISA-TAB-Nano
Nanomaterial Information / Identification				
1	Nano material name	SUBSTANCE	Information Category/ General Identifiers	Material file
2	CAS number	SUBSTANCE	Information Category/ Characterisation/Chemical composition	Material file
3	Structural formula/molecular structure	SUBSTANCE	Information Category/ Characterisation/Chemical composition	Material file
4	Composition of NM being tested (incl. degree of purity, known impurities or additives)	SUBSTANCE	Information Category/ Characterisation/Chemical composition; Information Category/ Characterisation/Physical structure	Material file
5	Basic Morphology	GI_GENERAL_INFORM	Information Category/ Characterisation/Physical structure	Material file

6	Description of surface chemistry (e.g. coating or modification)	SUBSTANCE SURFACE_CHEMISTRY	Information Category/ Characterisation/Surface description	Material file (nominal) Assay file (measured)
7	Major commercial uses	PRODUCT_TYPE_USE, DIRECTIONS_FOR_USE		
8	Known catalytic activity	CATALYTIC_ACTIVITY	Information Category/ Characterisation/Intensive properties	Assay file
9	Method of production (e.g. precipitation, gas phase)	SUBSTANCE	Information Category/ Production, Specification	Study file (protocol)
Physical-chemical Properties and Material Characterization				
10	Agglomeration / aggregation	AGGLOMERATION_AGGREGATION	Information Category/ Characterisation/Physical structure	Assay file, Data file
11	Water solubility	PC_WATER_SOL	Information Category/ Characterisation/Interaction	Assay file, Data file
12	Crystalline phase	CRYSTALLINE_PHASE	Information Category/ Characterisation/ Crystallographic structure	Assay file, Data file
13	Dustiness	DUSTINESS		Assay file, Data file
14	Crystallite size	CRYSTALLITE_AND_GRAIN_SIZE	Information Category/ Characterisation/ Crystallographic structure	Assay file, Data file
15	Representative TEM picture(s)	ATTACHMENTDOCUMENT		Assay file, Data file
15	Particle size distribution	PC_GRANULOMETRY	Information Category/ Characterisation/ Size	Assay file, Data file
17	Specific surface area	SPECIFIC_SURFACE_AREA	Information Category/ Characterisation/ Surface description	Assay file, Data file
18	Zeta potential (surface charge)	ZETA_POTENTIAL_SECTION	Information Category/ Characterisation/Interaction	Assay file, Data file
19	Surface chemistry (where appropriate)	SURFACE_CHEMISTRY	Information Category/ Characterisation/Surface description	Assay file, Data file
20	Photo-catalytic activity	PHOTOCATALYTIC_ACTIVITY	Information Category/ Characterisation/Interaction	Assay file, Data file
21	Pour density	POUR_DENSITY	Information Category/ Characterisation/Intensive properties	Assay file, Data file
22	Porosity	POROSITY	Information Category/ Characterisation/Intensive properties	Assay file, Data file

23	Octanol-water partition coefficient, where relevant	PC_PARTITION	Information Category/ Characterisation/Intensive properties	Assay file, Data file
24	Redox potential		Information Category/ Characterisation/Interactions	Assay file, Data file
25	Radical formation	RADICAL_FORMATION_POTENTIAL	Information Category/ Characterisation/Interactions	Assay file, Data file
26	Other relevant information (where available)	PC_OTHER		Assay file, Data file
Environmental Fate				
27	Dispersion stability in water			Assay file, Data file
	Biotic degradability			
28	- Ready biodegradability	TO_BIODEG_WATER_SCREEN	Information Category/ Characterisation/Interactions	Assay file, Data file
29	- Simulation testing on ultimate degradation in surface water	TO_BIODEG_WATER_SIM	Information Category/ Characterisation/Interactions	Assay file, Data file
30	- Soil simulation testing		Information Category/ Characterisation/Interactions	Assay file, Data file
31	Sediment simulation testing		Information Category/ Characterisation/Interactions	Assay file, Data file
32	- Sewage treatment simulation testing		Information Category/ Characterisation/Interactions	Assay file, Data file
33	Identification of degradation product(s)	EN_MAIN_DEGRADATION	Information Category/ Characterisation/Interactions	Assay file, Data file
34	Further testing of degradation product(s) as required		Information Category/ Characterisation/Interactions	Assay file, Data file
	Abiotic degradability and fate			
35	- Hydrolysis, for surface modified nanomaterials	TO_HYDROLYSIS	Information Category/ Characterisation/Interactions	Assay file, Data file
36	Adsorption- desorption	EN_ADSORPTION	Information Category/ Characterisation/Interactions	Assay file, Data file
37	Adsorption to soil or sediment	EN_STABILITY_IN_SOIL	Information Category/ Characterisation/Interactions	Assay file, Data file
38	Bioaccumulation potential	EN_BIOACCUMULATION	Information Category/ Characterisation/Interactions	Assay file, Data file

39	Bioaccumulation in sediment	EN_BIOACCU_TERR	Information Category/ Characterisation/Interactions	Assay file, Data file
Environmental Toxicology				
40	Effects on pelagic species (short/ long term)	EC_FISHTOX, EC_CHRONFISHTOX, EC_DAPHNIATOX, EC_CHRONDAPHNIATOX, EC_ALGAETOX	Information Category/ Characterisation/Interactions	Assay file, Data file
41	Effects on sediment species (short/ long term)	EC_SEDIMENTDWELLINGTOX	Information Category/ Characterisation/Interactions	Assay file, Data file
42	Effects on soil species (short/ long term)	EC_SOILDWELLINGTOX	Information Category/ Characterisation/Interactions	Assay file, Data file
43	Effect on terrestrial species	EC_PLANTTOX, EC_HONEYBEESTOX	Information Category/ Characterisation/Interactions	Assay file, Data file
44	Effect on micro-organisms	EC_SOIL_MICRO_TOX, EC_BACTOX	Information Category/ Characterisation/Interactions	Assay file, Data file
45	Other relevant information		Information Category/ Characterisation/Interactions	Assay file, Data file
Mammalian Toxicology				
46	Pharmacokinetics (ADME)		Information Category/ Characterisation/Interactions	Assay file, Data file
47	Acute Toxicity	TO_ACUTE_ORAL, TO_ACUTE_INHAL, TO_ACUTE_DERMAL	Information Category/ Characterisation/Interactions	Assay file, Data file
48	Repeated dose toxicity	TO_REPEATED_ORAL, TO_REPEATED_INHAL TO_REPEATED_DERMAL	Information Category/ Characterisation/Interactions	Assay file, Data file
49	Chronic toxicity	TO_CARCINOGENICITY, TO_SENSITIZATION	Information Category/ Characterisation/Interactions	Assay file, Data file
50	Reproductive toxicity	TO_REPRODUCTION	Information Category/ Characterisation/Interactions	Assay file, Data file
51	Developmental toxicity	TO_DEVELOPMENTAL	Information Category/ Characterisation/Interactions	Assay file, Data file
52	Genetic toxicity	TO_GENETIC_IN_VITRO,	Information Category/ Characterisation/Interacti	Assay file, Data file

		TO_GENETIC_IN_VIVO	ons	
53	Experience with human exposure	TO_EXPOSURE_OTHER	Information Category/ Characterisation/Interactions	Assay file, Data file
54	Other relevant test data	TO_OTHER	Information Category/ Characterisation/Interactions	Assay file, Data file
Material Safety				
55	Flammability	PC_AUTO_FLAMM, PC_FLAMM	Information Category/ Characterisation/Interactions	Assay file, Data file
56	Explosivity		Information Category/ Characterisation/Interactions	Assay file, Data file
57	Incompatibility		Information Category/ Characterisation/Interactions	Assay file, Data file

To summarise, the most important data objects, necessary to represent nanomaterials and NM characterisation are the **substance** with its **composition**, and a data object, able to represent a **test method**, its application to the substance under specific conditions and the **measurements** obtained as a result of this process. Therefore, the objects supported by the API are substances[a] (as a superclass of nanomaterials), protocols, protocol endpoints, conditions, protocol applications and measurements.

The API extends the original compound-centric dataset concept to allow datasets of nanomaterials. The OpenTox infrastructure contains all major statistical and machine learning (ML) algorithms for the development of regression, classification or clustering models, as well as chemoinformatic algorithms, such as structure optimisation and descriptor calculation. A ML algorithm is made available as a web resource and a model is created by sending a HTTP POST to the algorithm URI, with specified dataset URI and modelling parameters, where relevant. The model is again a web resource, and another HTTP POST to the model URI can be used to launch prediction of a specified dataset of chemical structures or materials.

The API offers access to a variety of searches by combination of measurement endpoints (e.g. all ENMs with size between 50 and 60 nm and having genotoxicity data) and is tightly integrated with a chemical structure search. This allows searching for the component of a material using a chemical structure, and highlighting its function as a core, coating or functionalisation. The searching can be used for many applications, one of which being NanoQSAR modelling.

3.2.3 API DOCUMENTATION

The REST API is documented using the Swagger[b] specification for documenting REST web services and available via Swagger-UI at http://enanomapper.github.io/API/

[a] https://github.com/opentox-api/api-specification/issues/3
[b] http://swagger.io/

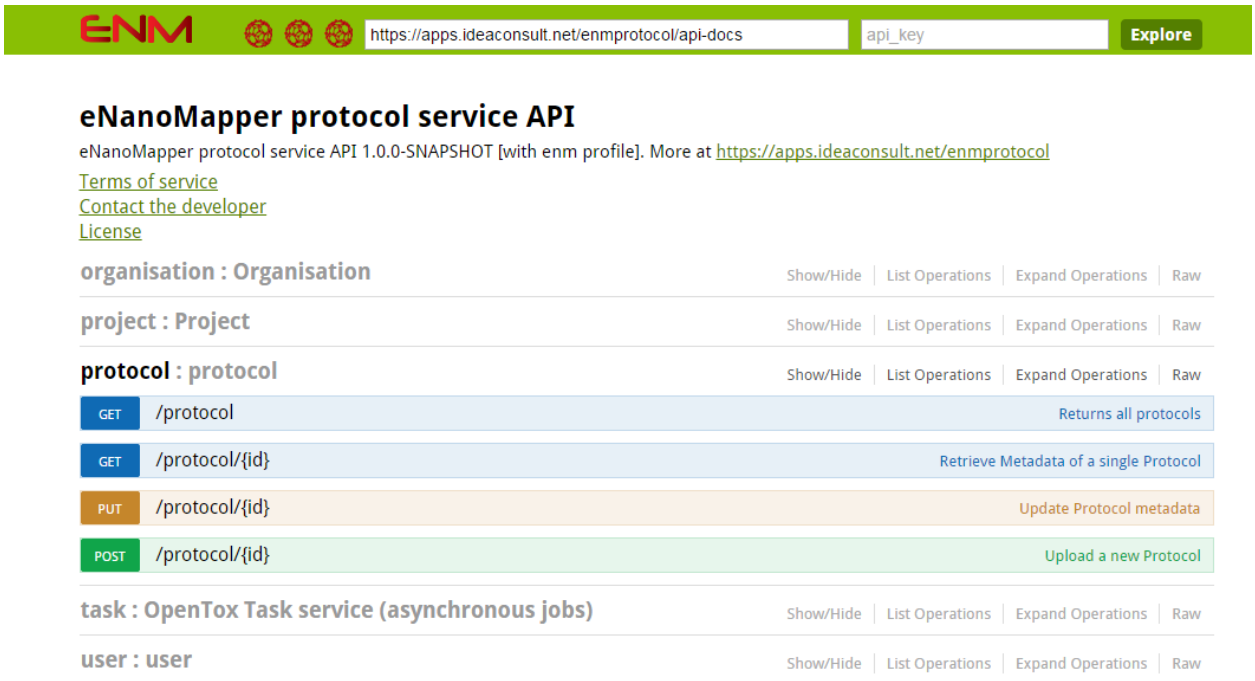


Figure 1. Protocol service API documentation

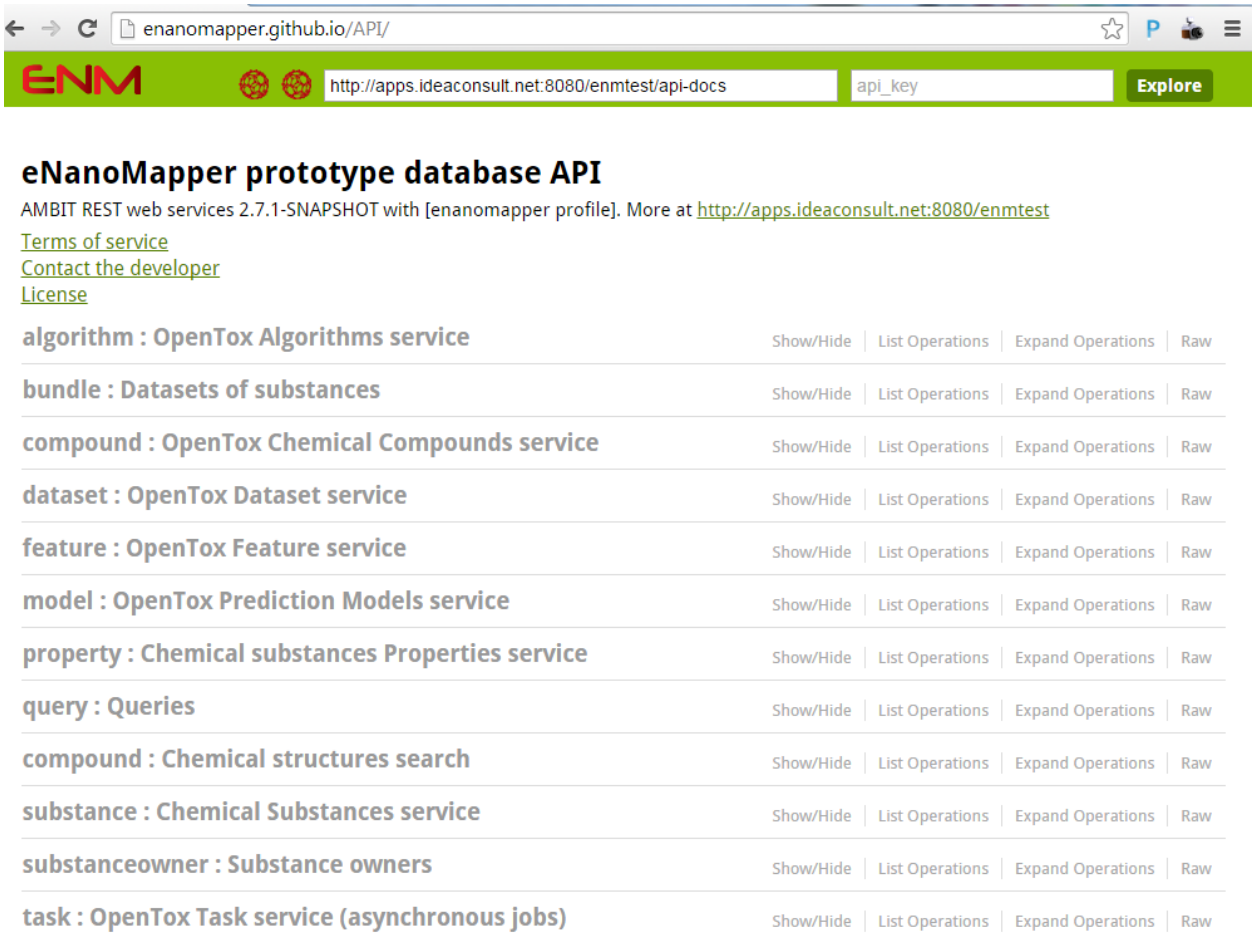


Figure 2. Data service API documentation

SUBSTANCE

The substance resource supports assigning a nanomaterial type, chemical composition with relevant concentration and constituent's role, as well as links to OpenTox Compound resources for specifying the chemical structure.

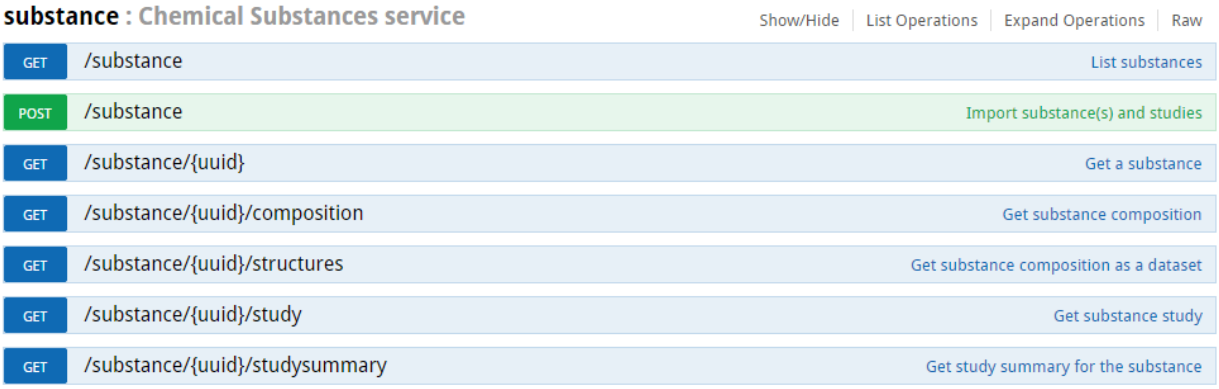


Figure 3. Substance API documentation

PROPERTIES

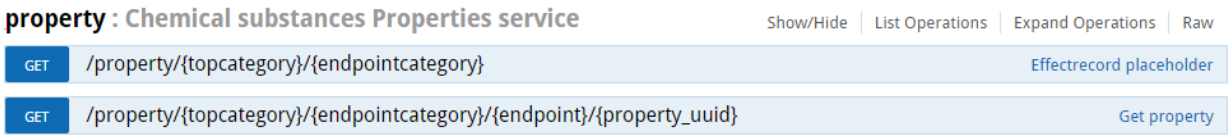


Figure 4. Endpoints API documentation

Retrieve and search substances

substance : Chemical Substances service

Show/Hide | List Operations | Expand Operations | Raw

GET /substance — List substances

Implementation Notes

Returns a list of substances, according to the search criteria

Response Class

Model | Model Schema

http://jsonschema.opentox.org/ {

URI (string, *optional*),
composition (array, *optional*),
externalIdentifiers (array[object], *optional*),
format (string, *optional*),
i5uuid (string): Unique identifier for the substance,
name (string, *optional*): Name of the substance (company specific),
ownerName (string, *optional*): Name of the substance owner (company producing the substance),
ownerUUID (string, *optional*): Unique identifier for the substance owner (company producing the substance),
publicname (string, *optional*),
referenceSubstance (object, *optional*): Reference substance,
substanceType (string) = ['Existing Chemical' or 'UVCB' or 'mono constituent substance' or 'multi constituent substance' or 'nanomaterial' or 'nanoparticle']: Substance type

}

Parameters

Parameter	Value	Description	Parameter Type	Data Type
search	formaldehyde	Search parameter	query	string
type	name (default)	Query type	query	string
compound_uri		If type=related finds all substances containing this compound; if type=reference - finds all substances with this compound as reference structure	query	string
page	0	Staring page	query	int
pagesize	10	Page size	query	int

Response Messages

HTTP Status Code	Reason	Response Model
200	OK. Substance(s) found	
404	Substances not found	
403	Forbidden	
401	Not Authorized	
405	Method not allowed	
500	Internal server error	
501	Not implemented	
503	Service unavailable	

Try it out!

Figure 5. Substance API details

Retrieve and search substances

substance : Chemical Substances service

Show/Hide | List Operations | Expand Operations | Raw

GET	/substance	List substances
POST	/substance	Import substance(s) and studies
GET	/substance/{uuid}	Get a substance
GET	/substance/{uuid}/composition	Get substance composition
GET	/substance/{uuid}/structures	Get substance composition as a dataset
GET	/substance/{uuid}/study	Get substance study

Implementation Notes

Substance study

Response Class

Model | Model Schema

http://jsonschema.net {
citation (object, *optional*),
effects (array[Effect]),
interpretation (object, *optional*),
owner (object, *optional*),
parameters (object, *optional*),
protocol (object),
reliability (object),
uuid (string): Unique identifier for the study document
}

Response Content Type: application/json

Parameters

Parameter	Value	Description	Parameter Type	Data Type
uuid	IUC4-efdb21bb-e79f-3286-a988-b6f6944d3734	**Substance UUID**	path	string
top		Top endpoint category	query	string
category		Endpoint category (The value in the protocol.category.code field)	query	string
property		Property UUID	query	string
property_uri		Property URI http://apps.ideaconsult.net:8080/enmtest/property/{UUID}, see Property service	query	string
page	0	Staring page	query	int
pagesize	10	Page size	query	int

Response Messages

HTTP Status Code	Reason	Response Model
200	OK	
400	Invalid substance identifier	
404	Substance not found	
403	Forbidden	
401	Not Authorized	
405	Method not allowed	
500	Internal server error	
501	Not implemented	
503	Service unavailable	

Try it out!

Figure 6. Substance study API documentation

Dataset of substances

substanceowner : Substance owners Show/Hide | List Operations | Expand Operations | Raw

GET	/substanceowner	List substance owners
GET	/substanceowner/{uuid}	Get a substance owner
DELETE	/substanceowner/{uuid}	Delete all substance by a substance owner
GET	/substanceowner/{uuid}/dataset	Get structures and study data of a substance owner as a Dataset

Implementation Notes

Returns a dataset, containing all structures with study data. See OpenTox Dataset service. Uses Property resources instead of Feature resources.

Response Class

Model | Model Schema

Dataset

Response Content Type application/json

Parameters

Parameter	Value	Description	Parameter Type	Data Type
uuid	IUC4-44BF02D8-47C5-385D-B203-9A8F315911CB	**Substance owner UUID**	path	string
page	0	Staring page	query	int
pagesize	10	Page size	query	int

Response Messages

HTTP Status Code	Reason	Response Model
200	OK	
400	Invalid substance owner identifier	
404	Substance owner not found	
403	Forbidden	
401	Not Authorized	
405	Method not allowed	
500	Internal server error	
501	Not implemented	
503	Service unavailable	

Try it out!

GET	/substanceowner/{uuid}/structure	Get structures of a substance owner as a Dataset
GET	/substanceowner/{uuid}/substance	Get all substances of the substance owner

bundle : Datasets of substances Show/Hide | List Operations | Expand Operations | Raw

GET	/bundle/{id}	Get a bundle
GET	/bundle/{id}/metadata	Get metadata for a bundle
GET	/bundle/{id}/substance	Get a list of all substances in a dataset

Figure 7. Datasets of substances API documentation

4. INITIAL IMPLEMENTATION

4.1 PROTOCOL WEB SERVICES

4.1.1 IMPLEMENTATION

The source code is available at GitHub and is based on a fork of ToxBank Protocol service [a], with subsequent updates and customisations http://apps.ideaconsult.net:8080/enmprotocol .
The Swagger-ui documentation of the REST API is available at enanomapper.github.io/API/ .

4.1.1.1 TECHNOLOGY

MySQL, RestLet, Java , RDF, JSON, Ajax. The API does not mandate particular storage technology.

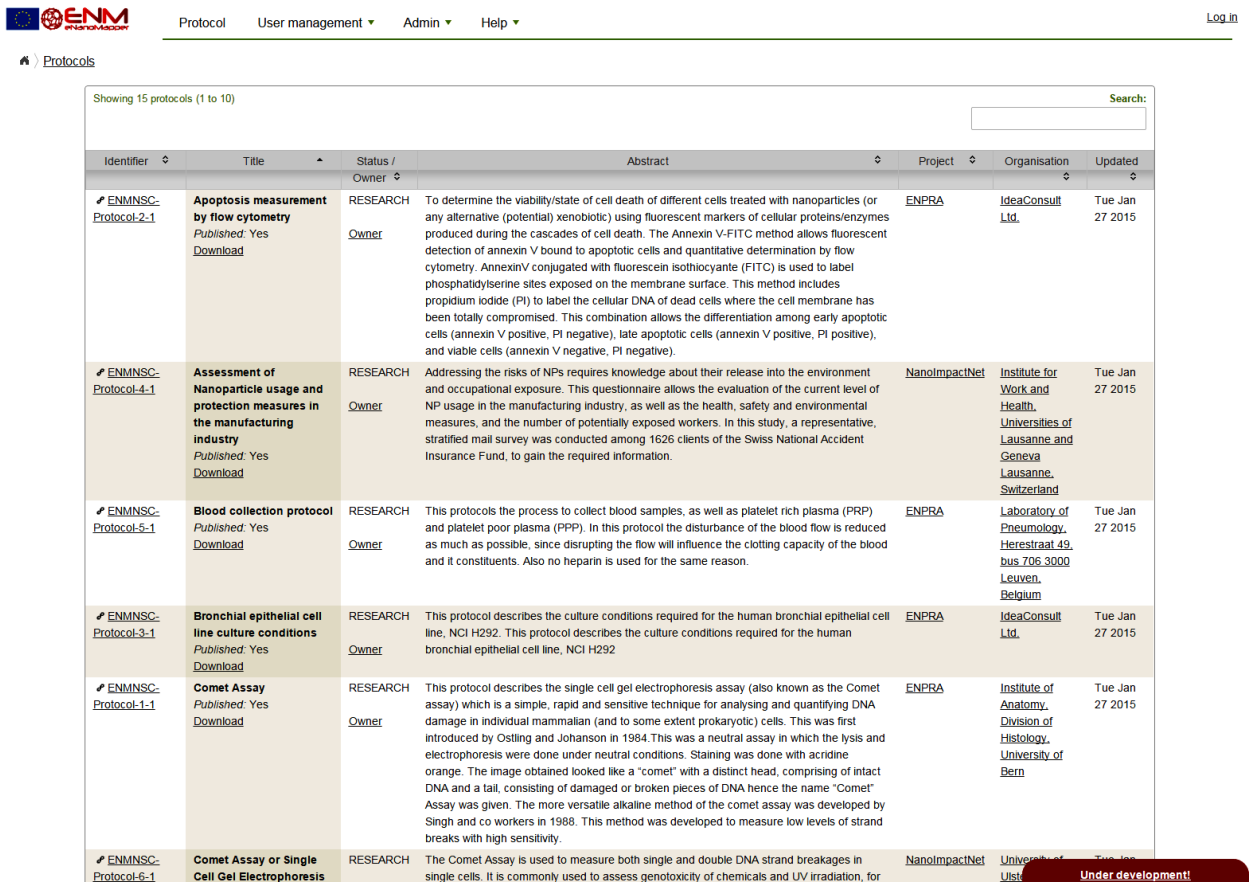
4.1.2 CONTENT

Upon recommendation by the NSC Hazard assessment working group [b] chair, we have considered for the initial implementation the collections of protocols maintained by NanoImpactNet (http://www.nanoimpactnet.eu/index.php?page=Researchprotocols) and DaNa projects http://www.nanoobjects.info , http://iai-dana.iai.fzk.de/en/nanoinfo/methods/992-standard-operating-procedures.

The protocol service instance includes 15 publicly available documents (PDF files with associated metadata) (Figure 7), 2 Standard Operating Procedures, 12 Research protocols and one of the documents, the NANOMMUNE "QUALITY HANDBOOK STANDARD PROCEDURES FOR NANOPARTICLE TESTING" includes 86 protocols in the following categories:

- Material Production and characterisation (16)
- Functionalisation and coating (23)
- Material characterisation (7)
- Dispersion (1)
- *In vitro* toxicity testing (6)
- Viability assays (6)
- Functionality and inflammation (17)
- *In vivo* Toxicity Testing (8)
- Transcriptomics (2)

[a] https://github.com/enanomapper/toxbank-api-server
[b] http://www.nanosafetycluster.eu/working-groups/2-hazard-wg.html



Identifier	Title	Status / Owner	Abstract	Project	Organisation	Updated
ENMNSC-Protocol-2-1	**Apoptosis measurement by flow cytometry** *Published:* Yes Download	RESEARCH Owner	To determine the viability/state of cell death of different cells treated with nanoparticles (or any alternative (potential) xenobiotic) using fluorescent markers of cellular proteins/enzymes produced during the cascades of cell death. The Annexin V-FITC method allows fluorescent detection of annexin V bound to apoptotic cells and quantitative determination by flow cytometry. AnnexinV conjugated with fluorescein isothiocyante (FITC) is used to label phosphatidylserine sites exposed on the membrane surface. This method includes propidium iodide (PI) to label the cellular DNA of dead cells where the cell membrane has been totally compromised. This combination allows the differentiation among early apoptotic cells (annexin V positive, PI negative), late apoptotic cells (annexin V positive, PI positive), and viable cells (annexin V negative, PI negative).	ENPRA	IdeaConsult Ltd.	Tue Jan 27 2015
ENMNSC-Protocol-4-1	**Assessment of Nanoparticle usage and protection measures in the manufacturing industry** *Published:* Yes Download	RESEARCH Owner	Addressing the risks of NPs requires knowledge about their release into the environment and occupational exposure. This questionnaire allows the evaluation of the current level of NP usage in the manufacturing industry, as well as the health, safety and environmental measures, and the number of potentially exposed workers. In this study, a representative, stratified mail survey was conducted among 1626 clients of the Swiss National Accident Insurance Fund, to gain the required information.	NanoImpactNet	Institute for Work and Health, Universities of Lausanne and Geneva Lausanne, Switzerland	Tue Jan 27 2015
ENMNSC-Protocol-5-1	**Blood collection protocol** *Published:* Yes Download	RESEARCH Owner	This protocols the process to collect blood samples, as well as platelet rich plasma (PRP) and platelet poor plasma (PPP). In this protocol the disturbance of the blood flow is reduced as much as possible, since disrupting the flow will influence the clotting capacity of the blood and it constituents. Also no heparin is used for the same reason.	ENPRA	Laboratory of Pneumology, Herestraat 49, bus 706 3000 Leuven, Belgium	Tue Jan 27 2015
ENMNSC-Protocol-3-1	**Bronchial epithelial cell line culture conditions** *Published:* Yes Download	RESEARCH Owner	This protocol describes the culture conditions required for the human bronchial epithelial cell line, NCI H292. This protocol describes the culture conditions required for the human bronchial epithelial cell line, NCI H292	ENPRA	IdeaConsult Ltd.	Tue Jan 27 2015
ENMNSC-Protocol-1-1	**Comet Assay** *Published:* Yes Download	RESEARCH Owner	This protocol describes the single cell gel electrophoresis assay (also known as the Comet assay) which is a simple, rapid and sensitive technique for analysing and quantifying DNA damage in individual mammalian (and to some extent prokaryotic) cells. This was first introduced by Ostling and Johanson in 1984.This was a neutral assay in which the lysis and electrophoresis were done under neutral conditions. Staining was done with acridine orange. The image obtained looked like a "comet" with a distinct head, comprising of intact DNA and a tail, consisting of damaged or broken pieces of DNA hence the name "Comet" Assay was given. The more versatile alkaline method of the comet assay was developed by Singh and co workers in 1988. This method was developed to measure low levels of strand breaks with high sensitivity.	ENPRA	Institute of Anatomy, Division of Histology, University of Bern	Tue Jan 27 2015
ENMNSC-Protocol-6-1	**Comet Assay or Single Cell Gel Electrophoresis**	RESEARCH	The Comet Assay is used to measure both single and double DNA strand breakages in single cells. It is commonly used to assess genotoxicity of chemicals and UV irradiation, for	NanoImpactNet	Univ... Ulst...	Tue Jan

Figure 7. Protocol service screenshot

4.2 DATA MANAGEMENT WEB SERVICES

4.2.1 IMPLEMENTATION

The eNanoMapper prototype database[38] provides support for upload and search for nanomaterials and experimental data through a REST web services API (enanomapper.github.io/API/) and a web browser interface as implemented by AMBIT web services[39], and is populated with content provided by project partners.

4.2.1.1 TECHNOLOGY

MySQL, RestLet, Java, JSON, Ajax. The API does not mandate particular storage technology.

4.2.1.2 DATA MODEL

The data model of the prototype follows designs of various proposals in this domain. For example, ISA-Tab is a very elegant approach, achieving universality by explicitly describing all steps in an experiment and recording details of the input and output nodes (i.e. how samples are processed). ISA-Tab alone defines only the metadata of the experiment, and requires further standardisation of the data files. The eNanoMapper consortium has experience in converting ISATab to a linked resources (RDF) format (using the toxbank.github.com/isa2rdf/ tool) and maintaining a semantic web database (a triple store) with a searchable interface (SPARQL queries)[40]. While ISA-Tab ensures all experimental details are retained, the chemical compound or ENM is hidden in the step of the experimental graph, and such a data model is usually less convenient for preparing and querying the data and applying predictive modelling. Building on previous experience and taking into account the observation that the majority of NanoSafety Cluster projects prefer to prepare their experimental data using custom spreadsheet templates, we take a pragmatic approach, representing measurements by a data model which is inspired by, but simpler than ISA-Tab.

We still find it useful to use terminology borrowed from the ISA-Tab programming model, namely protocol application. A protocol application explicitly describes a single step of the experimental graph, the application of a particular protocol with its specific parameters to the source material and the corresponding results (be it a sample or data readout). For the purposes of ENM database integration, the source material is always a chemical substance (ENM) with its composition and linkage, while the result is a set of measurements, annotated with the relevant endpoint and experiment conditions. The measurement can be specified by a value, range of values, error measure and units. This model directly supports the OHT data model, and is very similar to the measurement group concept in BAO, as well as encompassing the measurement value concept in CODATA UDS. In order to support raw data, we decided to extend the measurement value beyond scalar values and include links to measurement artefacts, such as image and raw data files, similarly to ISA-Tabs approach. The ability to describe derived measurements, by linking measurement groups, as supported by BAO and implied in UDS, is currently being considered, in order to support the modelling activities in eNanoMapper. An outline of the data model of the current prototype is given in Figure 8.

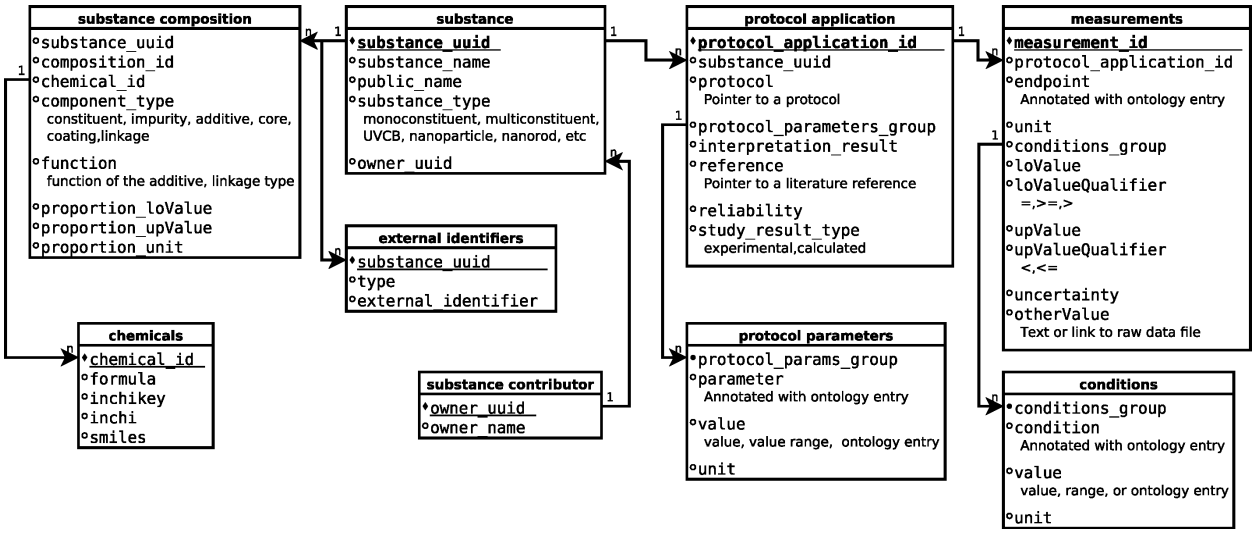


Figure 8. Outline of the data model

The data model is high level and follows the JSON serialisation supported by the API, rather than a database schema. The endpoints supported include the one recommended by OECD WPMN for the Representative manufactured nanomaterials[28] (Table 1). The chemicals table illustrates the link to the chemical structure and does not include all the details of the implementation, supporting chemical structure search. The flexible data model allows supporting variety of assays used (e.g. comet assays, micronucleus assay, Neutral Red uptake, Cell Transformation Assay), while the ontology annotation will be performed in the next reporting period.

4.2.1.3 DEPLOYMENT

A prototype implementation is deployed at http:/apps.ideaconsult.net/enmtest. The test instance was initially deployed in May 2014, and is used on a daily basis by several partners, in order to test and integrate the API between work packages and identify missing functionality and develop further improvements of the API.

The public instance at (http://apps.ideaconsult.net/enanomapper) is intended as an early illustration about the functionality and as means to gather focused feedback.

4.2.2 CONTENT

Two data sets have been made available through the prototype: NanoWiki and a dataset taken from literature that focuses on protein coronas, while a third one, ModNanoTox, is under development.

4.2.2.1 NANOWIKI

NanoWiki was originally developed as an internal knowledgebase of the toxicity of, primarily, metal oxides at the Karolinska Institutet and Maastricht University. Detailed description is provided in in Deliverable 5.3.

4.2.2.2 PROTEIN CORONA

The second demonstration data set, extracted from[41], focuses on the biological identity of ENMs. The authors used the composition of the protein corona fingerprint to predict the cell association of a 105-member library of surface- modified gold nanoparticles. 785 distinct serum proteins were identified by LC-MS/MS, from which 129 were suitable for relative quantification. The relative abundance of each of these proteins on a nanoparticle formulation defines the serum protein fingerprint for that formulation. To determine the extent to which individual proteins within the serum protein fingerprint predict cell

association, the authors developed a series of log-linear models that relate the relative abundance of each adsorbed serum protein to net cell association. Cell association was chosen as the model biological interaction because of its relevance to inflammatory responses, biodistribution, and toxicity *in vivo*. The eNanoMapper prototype described in this paper is able to capture this protein corona, and modelling approaches can extract this data for statistical analysis. This dataset is used as a primary example dataset to support modelling as described in the WP4 deliverable D4.1.

4.2.2.3 MODNANOTOX

A third data set is currently not available from the prototype instance, but being worked on for integration. This data set nicely demonstrates the complexity of the nanosafety domain. The ModNanoTox project (http://www.birmingham.ac.uk/generic/modnanotox/index.aspx) has produced a survey and selection of relevant physicochemical properties to use towards building a range of descriptors of engineered nanoparticles (mainly metal-based) and their potential toxicity. The ModNanoTox database provides physicochemical descriptors and toxic activities of nanoparticles from several studies. The database version from August 2013 includes 86 assays with more than 100 different endpoints affecting 45 species. Unfortunately, only a few nanoparticles (usually less than three) have been tested for each endpoint. Physicochemical descriptors for the characterisation of nanoparticles are incomplete as well (about 75% missing values). The two most comprehensive species in the dataset are Daphnia magna (water flea) and Danio rerio (zebrafish), with 34 and 14 assays each. The best represented endpoint for Daphnia is “Mortality”, and we were able to extract about forty “LC50” and sixty “% survival” data entries. In both cases the number of measured nanoparticle properties was very low. Most studies report only two to four different nanoparticle properties (descriptors) and the descriptor types are very inconsistent (overall 36 different descriptors, which results in very sparse matrices with a high number of missing values).

4.2.2.4 DATA IMPORT AND EXPORT

The data model allows integration of content from a variety of sources, namely OHTs (IUCLID5 .i5z files or direct retrieval of information from IUCLDI5 servers); custom spreadsheet templates (e.g. ModNanoTox); and custom formats, provided by partners (e.g. NanoWiki). ISA-Tab files are converted by compressing the chain of protocols into a single entry, yet retaining all the protocol parameters and recording the material as a substance and the rest of the factors as experimental conditions. The NanoWiki RDF dump (described in D5.3) is converted with a custom parser. The data import is performed by HTTP POST to the substance resource, which translates to a regular web form for file upload. The supported import formats are currently being extended with ISA-TAB-Nano and a large set of custom spreadsheet templates, taking into account the observation that the latter is the preferred approach to a data preparation format of the majority of the NanoSafety Cluster projects. A configurable parser enables import of the data, stored in the supported set of spreadsheet templates, accommodating different row-based, column-based or mixed organization of the data. The configuration metadata is defined in a separate file, mapping the custom spreadsheet structure into the internal eNanoMapper storage components: Substance, Protocol, Measurement, Parameters and Conditions. This enables uniform approach towards import, storage and searching of the ENM physicochemical measurements and biological assay results. https://github.com/enanomapper/nmdataparser

The NMDataParser uses a JSON configuration, which defines a map between spreadsheet fields and the data model. The JSON configuration syntax includes a set of keywords (Table 2) and JSON fields that allow defining different strategies for reading the data from one or several sheets, as well as to internally route and combine the excel structures (sheets, rows, columns, blocks of cells and cells) into the eNanoMapper database organization.

DATA_ACCESS	This section defines the basic parameters for data access and iteration of the primary sheet
ITERATION	Defines the iteration mode. Possible iteration modes are: *ROW_SINGLE, ROW_MULTI_FIXED, ROW_MULTI_DYNAMIC, ABSOLUTE_LOCATION, JSON_VALUE, JSON_REPOSITORY, VARIABLE*
SHEET_INDEX	The primary sheet for iteration
SHEET_NAME	The primary sheet name
START_ROW	The starting row for iteration
START_HEADER_ROW	The first (starting) header row
END_HEADER_ROW	The last (ending) header row
ALLOW_EMPTY	Flag that defines whether empty cells are allowed. Default value is *true*
RECOGNITION	The mode for sheet/column/row recognition. These elements can be recognized by index or by name.
DYNAMIC_ITERATION	Defines how dynamic iteration is performed in mode *ROW_MULTI_DYNAMIC*. Several rows are read at once where the criterion for row group recognition is: *NEXT_NOT_EMPTY* or *NEXT_DIFFERENT_VALUE.*
DYNAMIC_ITERATION_COLUMN_INDEX	The column used for the dynamic iteration.
VARIABLES	Defines an array of excel locations that are read into work variables stored for later used if the reading process
PARALLEL_SHEETS [ ]	This is an array of sections similar to section DATA_ACCESS that define the simultaneous reading of several sheets together with the primary sheet.
SUBSTANCE_RECORD	Section that defines the excel locations for reading of the basic fields of a Substance Record: *COMPANY_NAME, OWNER_NAME, SUBSTANCE_TYPE, OWNER_UUID, COMPANY_UUID, PUBLIC_NAME, ID_SUBSTANCE, COMPOSITION*
PROTOCOL_APPLICATIONS [ ]	This is an array of sections , defining the excel data locations for reading of Protocol Application data. Each section includes following fields: CITATION_TITLE, CITATION_YEAR, CITATION_OWNER, INTERPRETATION_RESULT, INTERPRETATION_CRITERIA, PROTOCOL_GUIDELINE, PARAMETERS (an array of data locations), EFFECTS (an array of sections)
EFFECTS [ ]	This is an array of sections. Each section defines data structures (effect record) for particular measurements and includes following excel data locations: SAMPLE_ID, ENDPOINT, LO_VALUE, UP_VALUE, ERR_VALUE, TEXT_VALUE, VALUE, LO_QUALIFIER, UP_QUALIFIER, ERR_QUALIFIER, UNIT, CONDITIONS (an array of data locations)
REPOSITORY	A JSON structure for defining preconfigured data (e.g. protocol, parameters) to be read directly from the JSON file into the data classes.

Table 2. Keywords of the spreadsheet data parser configuration

Examples with the public Protein Corona dataset are available as test resource at GitHub [a]. While the parser is open source, the configuration files may not be, thus not revealing the organisation of the confidential data templates. The parser is currently being tested on NanoReg and ModNanoTox templates. Maps of the confidential spreadsheet templates are available on request, in compliance with the confidentiality agreements between projects.

[a] https://github.com/enanomapper/nmdataparser/tree/master/src/test/resources/net/enanomapper/parser/csv

More formats will be supported as needed for indexing data from different sources. Currently the study and substance resources in the eNanoMapper prototype database support JSON serialisation. The development of ISA-Tab –Nano and RDF import and export is ongoing and the development will be completed during the next reporting period.

4.2.2.5 SCREENSHOTS

The following screenshots illustrate the nanomaterial components, phys-chem and toxicity data in the current implementation. The user interface is implemented as JavaScript widgets consuming the REST API.

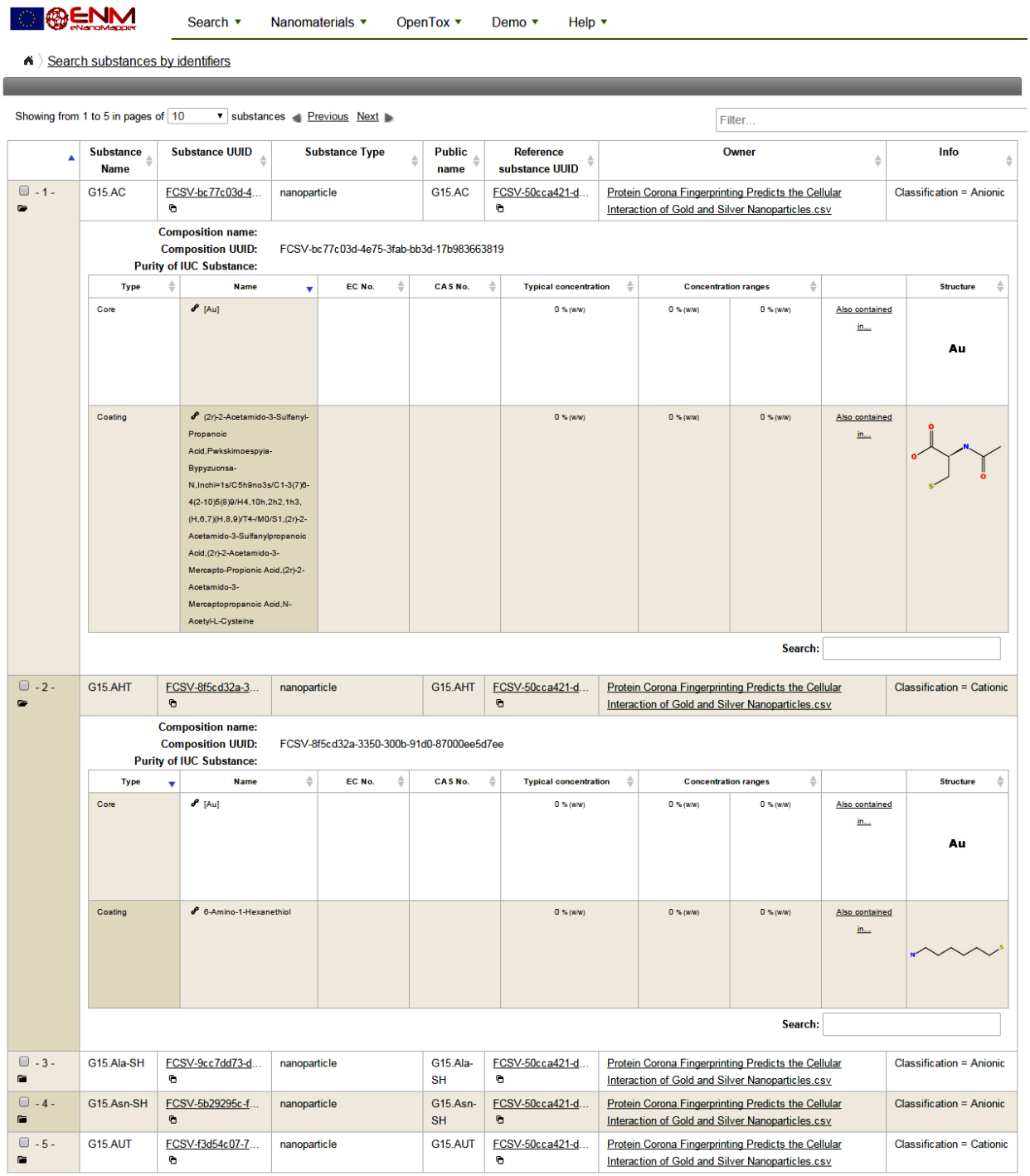


Figure 9. Nanomaterial components (core and coating of gold particles in protein corona dataset)

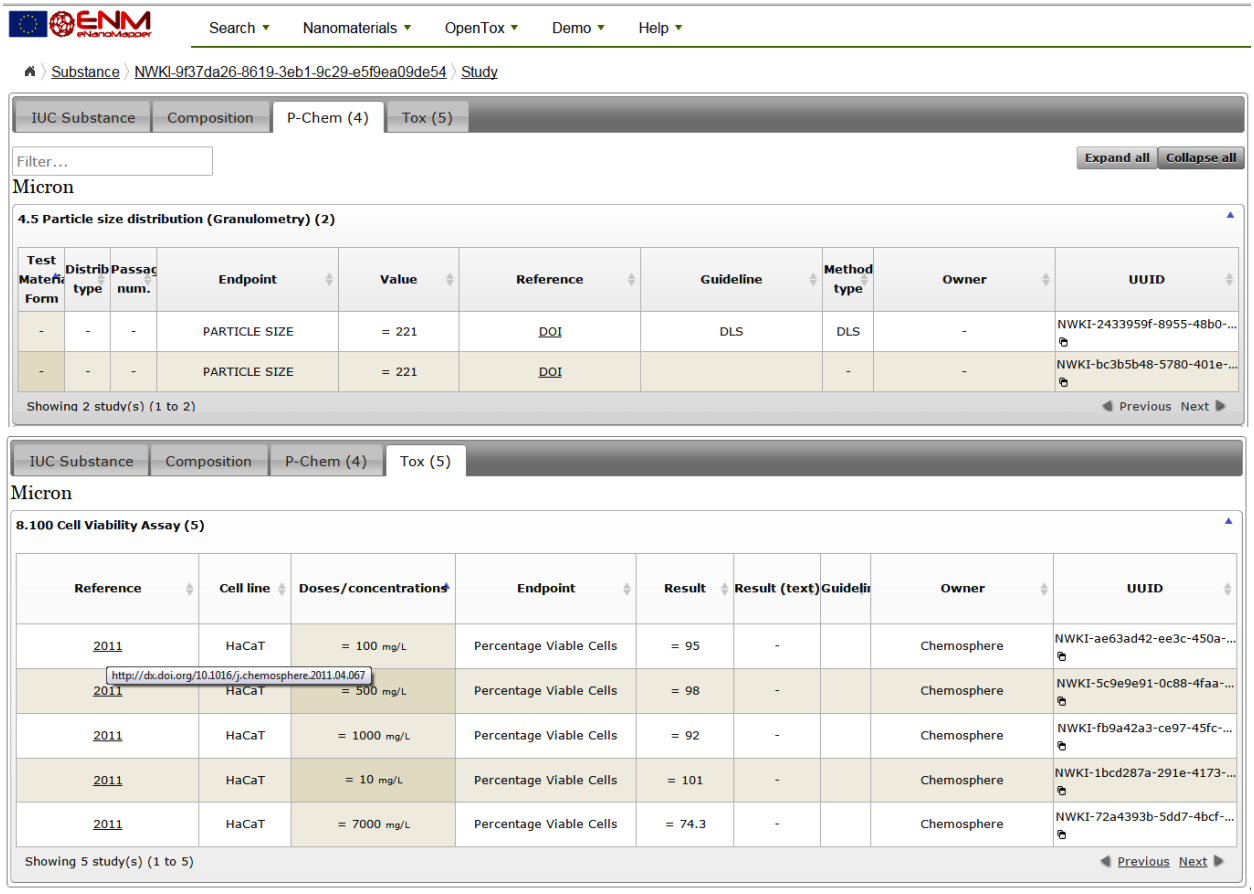


Figure 10. Physicochemical and toxicity data (nanoWiki)

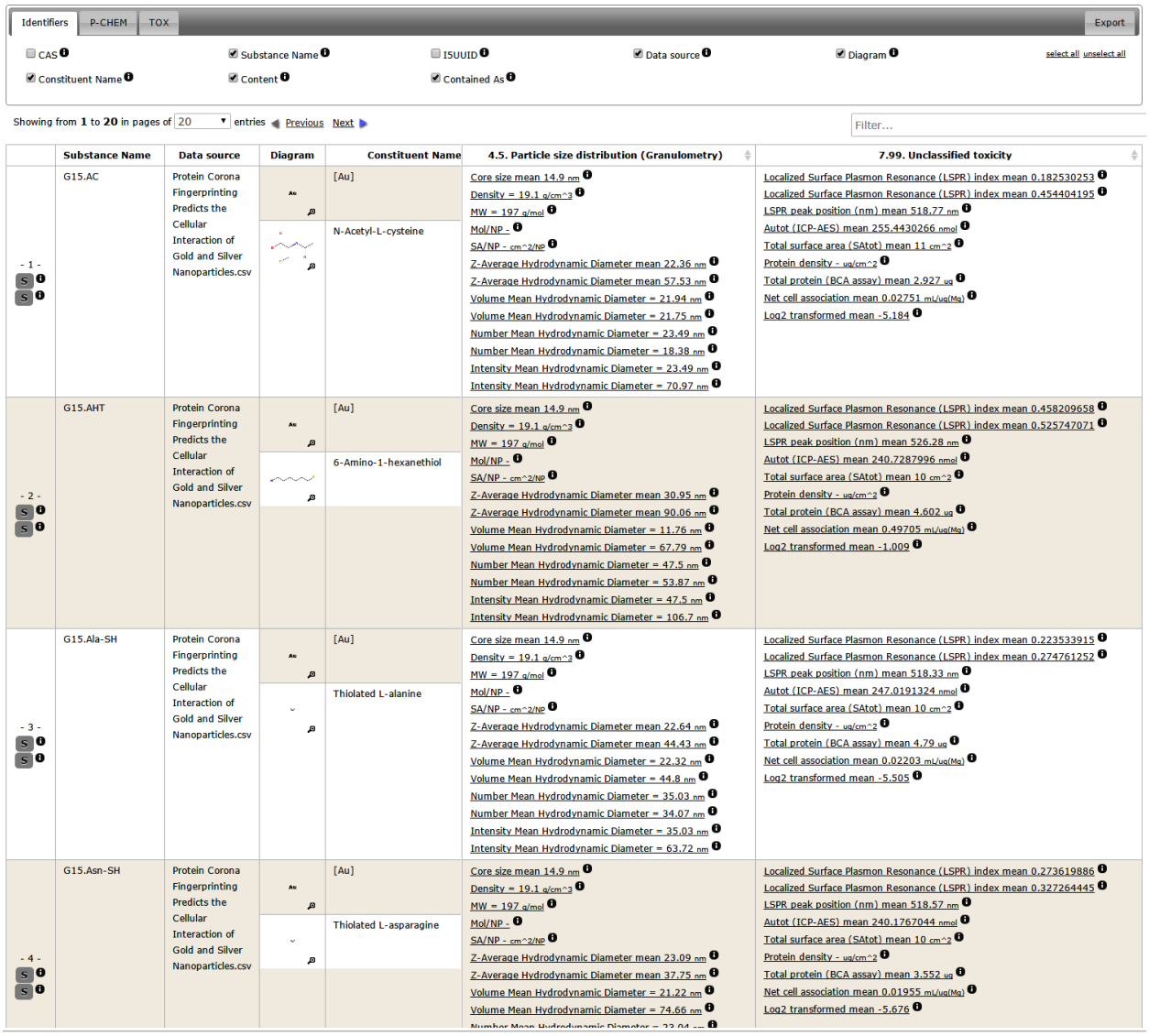


Figure 11. Dataset view (Protein corona)

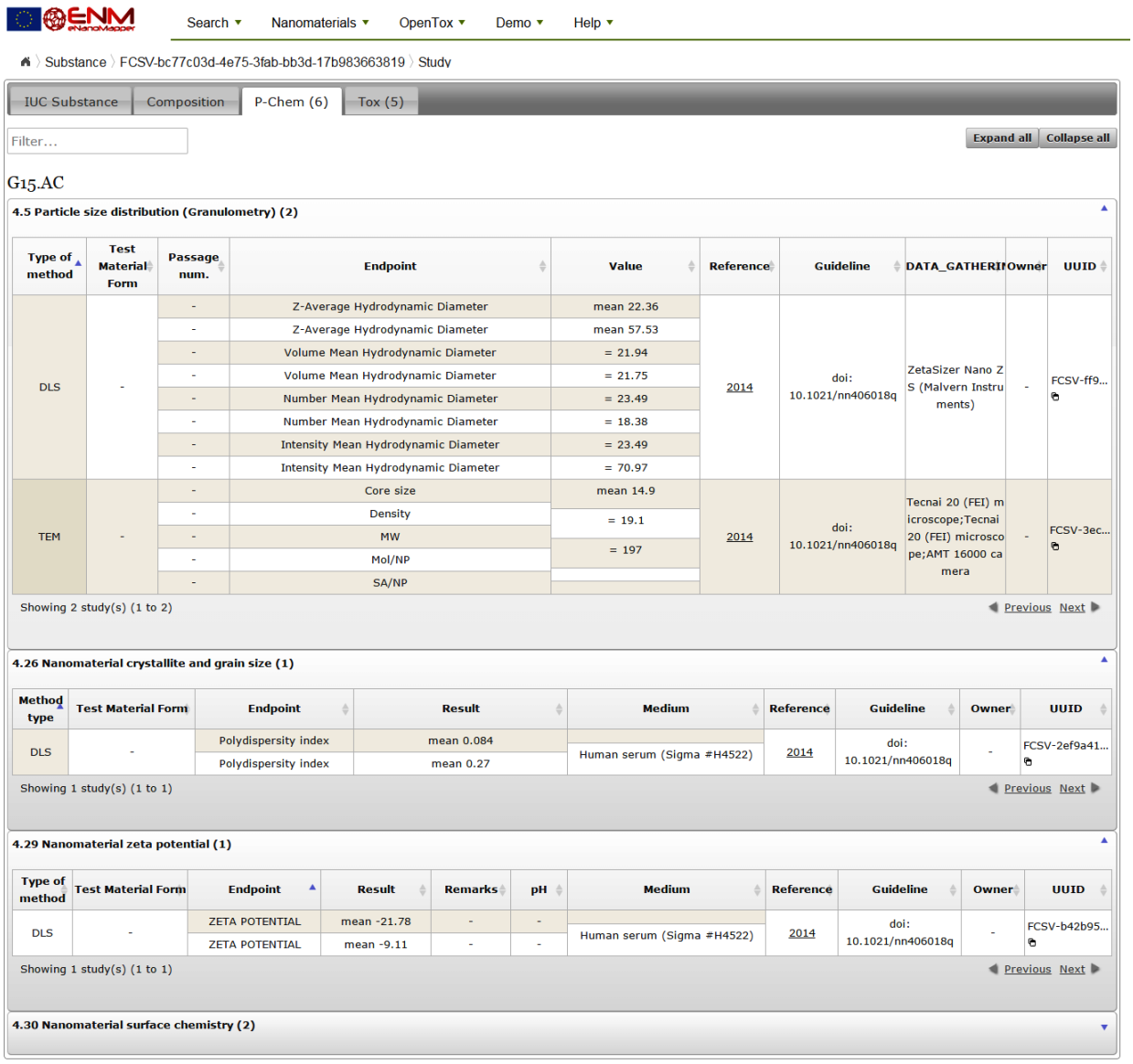


Figure 12. Physico-chemical data (Protein corona dataset)

5. CONCLUSION

We have performed an exhaustive review of existing nano-related data models, databases, and nanomaterial related entries in chemical and toxicogenomic databases. The API with resources supporting substances, protocol and measurements is in line with recent publications in the domain and is able to support variety of tests and endpoints, recommended by OECD WPMN. The annotation with ontology entries is an ongoing collaboration with WP2. The database prototype API implementation relies on existing open source project with a long history. The demonstration data provided by partners illustrates the capability of the API and the implementation to handle diverse information. It has been used for Quantitative Structure-Activity Relationships for nanomaterials (NanoQSAR) modelling. Research is ongoing to extend the OpenTox algorithm and modelling APIs for nanomaterials, allowing

these new models to be exposed with unique URIs suitable for reuse. The REST API with JSON serialisation is the current state of the art in web system development and data integration and enables building graphical summaries of the data, JavaScript widgets, custom user interfaces and programmatic interaction.

The next steps include provision of RDF serialisation of the resources; support for multiple data formats on import and export; support for multiple search interfaces (including ones based on semantic technologies); public release of the data services; and improvements of the API and the implementation, based on the feedback and with close collaboration with other eNanoMapper WP and NSC working groups.

6. BIBLIOGRAPHY

(1) Wijnhoven, S. W. P. *Development of an inventory for consumer products containing nanomaterials*; 2010.

(2) Woodrow Wilson database. Consumer Products An inventory of nanotechnology-based consumer products currently on the market. http://www.nanotechproject.org/inventories/consumer/ (accessed Oct 15, 2012).

(3) Bleeker, E. A. J. *Interpretation and implications of the European Commission Recommendation on the definition of nanomaterial*; 2012.

(4) Aschberger, K.; Rauscher, H.; Crutzen, H.; Rasmussen, K.; Christensen, F. M.; Sokull-Klüttgen, B.; Stamm, H. *Considerations on information needs for nanomaterials in consumer products*; 2014.

(5) Hendrickx, D. M.; Boyles, R. R.; Kleinjans, J. C. S.; Dearry, A. Workshop report: Identifying opportunities for global integration of toxicogenomics databases, 26-27 June 2013, Research Triangle Park, NC, USA. *Arch. Toxicol.* **2014**, *88*, 2323–2332 DOI: 10.1007/s00204-014-1387-3.

(6) Kohonen, P.; Ceder, R.; Smit, I.; Hongisto, V.; Myatt, G.; Hardy, B.; Spjuth, O.; Grafström, R. Cancer Biology, Toxicology and Alternative Methods Development Go Hand-in-Hand. *Basic Clin. Pharmacol. Toxicol.* **2014**, *115*, 50–58 DOI: 10.1111/bcpt.12257.

(7) Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H. PubChem BioAssay: 2014 update. *Nucleic Acids Res.* **2014**, *42*, D1075–D1082 DOI: 10.1093/nar/gkt978.

(8) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090 DOI: 10.1093/nar/gkt1031.

(9) Rustici, G.; Kolesnikov, N.; Brandizi, M.; Burdett, T.; Dylag, M.; Emam, I.; Farne, A.; Hastings, E.; Ison, J.; Keays, M.; et al. ArrayExpress update--trends in database growth and links to data analysis tools. *Nucleic Acids Res.* **2013**, *41*, D987–D990 DOI: 10.1093/nar/gks1174.

(10) Kong, L.; Tuomela, S.; Hahne, L.; Ahlfors, H.; Yli-Harja, O.; Fadeel, B.; Lahesmaa, R.; Autio, R. NanoMiner - integrative human transcriptomics data resource for nanoparticle research. *PLoS One* **2013**, *8*, e68414 DOI: 10.1371/journal.pone.0068414.

(11) Panneerselvam, S.; Choi, S. Nanoinformatics: emerging databases and available tools. *Int. J. Mol. Sci.* **2014**, *15*, 7158–7182 DOI: 10.3390/ijms15057158.

(12) Williams, A. J.; Harland, L.; Groth, P.; Pettifer, S.; Chichester, C.; Willighagen, E. L.; Evelo, C. T.; Blomberg, N.; Ecker, G.; Goble, C.; et al. Open PHACTS: semantic interoperability for drug discovery. *Drug Discov. Today* **2012** DOI: 10.1016/j.drudis.2012.05.016.

(13) Nyström-Persson, J.; Igarashi, Y.; Ito, M.; Morita, M.; Nakatsu, N.; Yamada, H.; Mizuguchi, K. Toxygates: interactive toxicity analysis on a hybrid microarray and linked data platform. *Bioinformatics* **2013**, *29*, 3080–3086 DOI: 10.1093/bioinformatics/btt531.

(14) Jupp, S.; Malone, J.; Bolleman, J.; Brandizi, M.; Davies, M.; Garcia, L.; Gaulton, A.; Gehant, S.; Laibe, C.; Redaschi, N.; et al. The EBI RDF platform: linked open data for the life sciences. *Bioinformatics* **2014**, *30*, 1338–1339 DOI: 10.1093/bioinformatics/btt765.

(15) Hardy, B.; Douglas, N.; Helma, C.; Rautenberg, M.; Jeliazkova, N.; Jeliazkov, V.; Nikolova, I.; Benigni, R.; Tcheremenskaia, O.; Kramer, S.; et al. Collaborative development of predictive toxicology applications. *J. Cheminform.* **2010**, *2*, 7.

(16) Maynard, A. D.; Aitken, R. J.; Butz, T.; Colvin, V.; Donaldson, K.; Oberdörster, G.; Philbert, M. A.; Ryan, J.; Seaton, A.; Stone, V.; et al. Safe handling of nanotechnology. *Nature* **2006**, *444*, 267–269 DOI: 10.1038/444267a.

(17) European Commission. *Types and uses of nanomaterials, including safety aspects*; Brussels, 2012.

(18) Cárdenas, W. H. Z.; Mamani, J. B.; Sibov, T. T.; Caous, C. A.; Amaro, E.; Gamarra, L. F. Particokinetics: computational analysis of the superparamagnetic iron oxide nanoparticles deposition process. *Int. J. Nanomedicine* **2012**, *7*, 2699–2712 DOI: 10.2147/IJN.S30074.

(19) Kelder, T.; van Iersel, M. P.; Hanspers, K.; Kutmon, M.; Conklin, B. R.; Evelo, C. T.; Pico, A. R. WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.* **2012**, *40*, D1301–D1307 DOI: 10.1093/nar/gkr1074.

(20) Sansone, S.-A.; Rocca-Serra, P.; Field, D.; Maguire, E.; Taylor, C.; Hofmann, O.; Fang, H.; Neumann, S.; Tong, W.; Amaral-Zettler, L.; et al. Toward interoperable bioscience data. *Nat. Genet.* **2012**, *44*, 121–126 DOI: 10.1038/ng.1054.

(21) 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.; et al. ISA-TAB-Nano: A Specification for Sharing Nanomaterial Research Data in Spreadsheet-based Format. *BMC Biotechnol.* **2013**, *13*, 2 DOI: 10.1186/1472-6750-13-2.

(22) Hastings, J.; Willighagen, E.; Owen, G.; Jeliazkova, N.; The eNanoMapper Consortium; Steinbeck, C. eNanoMapper: Opportunities and challenges in using ontologies to enable data integration for nanomaterial risk assessment. In *Proceedings of the ISMB Bio-Ontologies SIG meeting, Boston, USA, July 11-12 2014*; 2014.

(23) CODATA-VAMAS Working Group on the Description of Nanomaterials. *Uniform Description System for Materials on the Nanoscale*; 2014.

(24) ISO standards Nanotechnologies http://www.iso.org/iso/home/store/catalogue_tc/catalogue_tc_browse.htm?commid=381983&published=on&includesc=true (accessed Dec 8, 2014).

(25) Aschberger, K.; Rauscher, H.; Crutzen, H.; Rasmussen, K.; Christensen, F. M.; Sokull-Klüttgen, B.; Stamm, H. *Considerations on information needs for nanomaterials in consumer products*; 2014.

(26) JRC. *Requirements on measurements for the implementation of the European Commission definition of the term "nanomaterial";* 2012.

(27) Rauscher, H.; Roebben, G. *Towards a review of the EC Recommendation for a definition of the term "nanomaterial". Part 1: Compilation of information concerning the experience with the definition*; 2014.

(28) *Series on the Safety of Manufactured Nanomaterials No.27. LIST OF MANUFACTURED NANOMATERIALS AND LIST OF ENDPOINTS FOR PHASE ONE OF THE SPONSORSHIP PROGRAMME FOR THE TESTING OF MANUFACTURED NANOMATERIALS: REVISION*; 2010.

(29) JRC. *Towards a review of the EC Recommendation for a definition of the term "nanomaterial";* 2014.

(30) *Multi-walled Carbon Nanotubes, NM-400, NM-401, NM-402, NM-403: Characterisation and Physico-Chemical Properties*.

(31) *NM-300 Silver Characterisation, Stability, Homogeneity*.

(32) Rasmussen, K.; Mast, J.; De Temmerman, P.-J.; Verleysen, E.; Waegeneers, N.; Van Steen, F.; Pizzolon, J. C.; De Temmerman, L.; Van Doren, E.; Jensen, K. A.; et al. *Titanium Dioxide, NM-100, NM-101, NM-102, NM-103, NM-104, NM-105: Characterisation and Physico-Chemical Properties*; Institute for Health and Consumer Protection, 2014.

(33) Charanjeet Singh; Friedrichs, S.; Ceccone, G.; Gibson, N.; Jensen, K. A.; Levin, M.; Infante, H. G.; Carlander, D.; Rasmussen, K. *Cerium Dioxide, NM-211, NM-212, NM-213. Characterisation and test item preparation*; 2014.

(34) *Zinc Oxide NM-110, NM-111, NM-112, NM-113: Characterisation and Test Item Preparation*.

(35) *Synthetic Amorphous Silicon Dioxide (NM-200, NM-201, NM-202, NM-203, NM-204): Characterisation and Physico-Chemical Properties*.

(36) Hardy, B.; Douglas, N.; Helma, C.; Rautenberg, M.; Jeliazkova, N.; Jeliazkov, V.; Nikolova, I.; Benigni, R.; Tcheremenskaia, O.; Kramer, S.; et al. Collaborative development of predictive toxicology applications. *J. Cheminform.* **2010**, *2*, 7.

(37) Roebben, G.; Rasmussen, K.; Kestens, V.; Linsinger, T. P. J.; Rauscher, H.; Emons, H.; Stamm, H. Reference materials and representative test materials: the nanotechnology case. *J. Nanoparticle Res.* **2013**, *15*, 1455 DOI: 10.1007/s11051-013-1455-2.

(38) Jeliazkova, N.; Jeliazkov, V.; Willighagen, E.; Smeets, B.; Munteanu, C.; Fadeel, B.; Grafström, R.; Kohonen, P.; Sarimveis, H.; Tsiliki, G.; et al. The first eNanoMapper prototype: a substance database to support safe-by-design. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; Belfast, 2014.

(39) Jeliazkova, N.; Jeliazkov, V. AMBIT RESTful web services: an implementation of the OpenTox application programming interface. *J. Cheminform.* **2011**, *3*, 18 DOI: 10.1186/1758-2946-3-18.

(40) Kohonen, P.; Benfenati, E.; Bower, D.; Ceder, R.; Crump, M.; Cross, K.; Grafström, R. C.; Healy, L.; Helma, C.; Jeliazkova, N.; et al. The ToxBank Data Warehouse: Supporting the Replacement of *In Vivo* Repeated Dose Systemic Toxicity Testing. *Mol. Inform.* **2013**, *32*, 47–63 DOI: 10.1002/minf.201200114.

(41) Walkey, C. D.; Olsen, J. B.; Song, F.; Liu, R.; Guo, H.; Olsen, D. W. H.; Cohen, Y.; Emili, A.; Chan, W. C. W. Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles. *ACS Nano* **2014**, *8*, 2439–2455 DOI: 10.1021/nn406018q.