DELIVERABLE REPORT D5.6

D5.6 User Application for Conformance to Reporting and Curation Standards

Deliverable D5.6

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:	EMBL-EBI
DATE:	14.12.2016
VERSION:	V.1.0

Call identifier	FP7-NMP-2013-SMALL-7
Document Type	Deliverable Report
WP/Task	WP5 - Task 5.8
Document ID	eNanoMapper D5.6
Status	Final

Partner Organisations	●Douglas Connect, GmbH (DC) ●National Technical University of Athens (NTUA) ●In Silico Toxicology (IST) ●Ideaconsult (IDEA) ●Karolinska Institutet (KI) ●European Bioinformatics Institute (EMBL-EBI) ●Maastricht University (UM) ●Misvik Biology (MB)
Authors	Linda Rieswijk (UM) Jiakang Chang (EMBL-EBI) Egon Willighagen (UM) Penny Nymark (MB) Pekka Kohonen (MB)
Purpose of the Document	[T5.8]
Document History	1.First draft, 17/11/2016 2.Final version, 14/12/2016

TABLE OF CONTENTS

TABLE OF FIGURES

GLOSSARY

Abbreviation / acronym	Description
ENM	Engineered Nanomaterial
CEINT	Center for the Environmental Implications of NanoTechnology
CEINT NIKC	Center for the Environmental Implications of NanoTechnology NanoInformatics Knowledge Commons
ChEBI	Chemical Entities of Biological Interest
EC	European Commission
JRC	Joint Research Centre
MIAN	Minimal Information about Nanomaterials
NDCI	Nanomaterial Data Curation Initiative
NIH	National Institutes of Health
NIL	Nanoparticle Information Library
NNI	National Nanotechnology Initiative
OWL	Web Ontology Language
PCCs	Physico-chemical characteristics
RDF	Resource Description Framework
RIVM	National Institute for Public Health and the Environment in the Netherlands
SPARQL	SPARQL Protocol and RDF Query Language
UDS	Uniform Description System

1. EXECUTIVE SUMMARY

The eNanoMapper project aims to build an ontology and database to collate and describe data relevant for "safe by design" engineered nanomaterial development. Within the nanosafety domain several projects have initiated the development of standards for data reporting and curation. With this deliverable we show how the ontology can be used to test the data completeness within the eNanoMapper database, according to these approved standards, which is tested with the use of SPARQL queries. To translate completeness expectations to searches, we used the SPARQL query language to query a combination of the eNanoMapper ontology and an export of the data in http://data.enanomapper.net/ (see also D3.3). The use of such SPARQL queries for automated quality assurance and completeness testing is now being explored in collaboration with the caLIBRAte project. Moreover, caLIBRAte is developing input parameter criteria for risk assessment tools to be calibrated within the project and the SPARQL queries could be used to query for the specific data needs of the tools.

2. INTRODUCTION

Nanomaterials are materials in which the spatial dimensions have at least one size in the 1 to 100 nm range. In addition to the wide diversity of natural nanomaterials available, advances in chemical synthesis techniques have led to an explosion in the number of engineered nanomaterials (ENMs) in recent years. Materials with structures in the nanoscale range often have unique optical, electronic, and mechanical properties, and as a result ENMs are being developed to meet specific application needs in diverse domains across the engineering and biomedical sciences (e.g. drug delivery). However, accompanying the proliferation of nanomaterials is a challenging race to understand and predict their possibly detrimental effects on human health and the environment.

The correct curation of nanomaterial data into databases allows for the development of improved methods for risk assessment of nanomaterials. Data completeness and quality are important aspects in order to serve their intended purpose. Assessing data for their completeness and quality, especially within the nanosafety domain, is particularly difficult considering its highly multidisciplinary nature. To facilitate this task several recommendations have been provided for the development of minimal reporting standards by the NDCI (Nanomaterial Data Curation Initiative). Examples of the minimal reporting standards can be found within the Uniform Description System (UDS) by CODATA/VAMAS Joint Working Group or as presented within the Nanomaterial Registry (i.e. Minimal Information about Nanomaterials).

By making use of these minimal reporting standards, data completeness and quality of the data in the eNanoMapper database can be tested by verifying if the required information is provided for one ENM or a collection of ENMs. From these tests, ratios can be calculated indicating to what extent the information is complete. This is basically the same approach taken by the Nanomaterial Registry, though it goes a step further and uses this information to calculate compliance scores and levels.

This deliverable describes the work from T5.8 which shows an approach that combines the eNanoMapper ontology and the structure of the data in the database server. The data is combined by taking advantage of the Resource Description Framework (RDF) representations of both the ontology (in the RDF-based Web Ontology Language (OWL) format) and a linked data export of the data from the data.enanomapper.net database (see D3.3). Using the semantic web query language SPARQL, tests can be integrated. We will show how this approach can seamlessly use information from the ontology and from the database.

3. CONFORMANCE TO REPORTING AND CURATION STANDARDS

3.1 EXISTING REPORTING AND CURATION STANDARDS

3.1.1 NANOMATERIAL DATA CURATION INITIATIVE

The Nanomaterial Data Curation Initiative (NDCI) is a project of the National Cancer Informatics Program Nanotechnology Working Group (NCIP NanoWG) which is an attempt to explore the critical aspect of data curation within the development of informatics approaches to understanding nanomaterial behaviour. The NDCI has published a series of papers (Hendren, 2015) on nanomaterial data curation with the purpose to:

1. present and evaluate the current state of nanomaterial data curation across the field on multiple specific data curation topics;
2. propose ways to leverage and advance progress for both individual efforts and the nanomaterial data community as a whole; and to
3. provide opportunities for similar publication series on the details of the interactive needs and workflows of data customers, data creators, and data analysts.

One of the papers published by the NDCI focusses on how to assess the completeness and quality of (curated) nanomaterial data (Robinson et al., 2015). Within this paper, several data resources were used in order to provide sufficient guidelines, including caNanoLab (Lijowski and Michal, 2010), DaNa Knowledge Base (Marquardt C and Kühnel D, 2013), Center for the Environmental Implications of NanoTechnology (CEINT) NanoInformatics Knowledge Commons (CEINT NIKC), NanoNext Database on the Environmental Fate and Effects of Nanomaterials – developed by the National Institute for Public Health and the Environment in the Netherlands (RIVM) (Slooff W, 2016) , European Commission (EC) Joint Research Centre (JRC) NANOHub Database, MOD-ENP-TOX Datasets, NanoMILE Knowledge-base, ModNanoTox Datasets, and Nanoparticle Information Library (NIL)). Several recommendations were made with regards to (meta) data provision (Table 1) and also computational solutions (Table 2) which will facilitate data completeness and quality. Further details about these recommendations are presented in the supplementary files of the paper from the NDCI (Robinson et al., 2015).

Table 1: Recommendations for (meta) data quality and completeness as provided by the Nanomaterial Data Curation Initiative (NDCI)

Recommendations	Brief description + importance
Physico-chemical characteristics	Details on in-house determination (incl. biologically relevant exposure conditions).
Temporal measurements	Temporal metadata from any experimental study.
Assessment of artefacts	(Meta) data directed towards assessment of possible artefacts from any biological study.
Experimental errors and uncertainty	(Meta) data related to experimental errors and uncertainty from any experimental study.
Impurities	Data identifying (biologically significant) impurities from any experimental study.
Manufacturer ID	Manufacturer supplied IDs should be provided when trying to integrate data from different experimental studies.
Measurement details	Sufficient metadata should be provided to precisely identify any measured data from experimental studies.
Provenance	Provenance metadata are essential for all curation efforts.
Surface composition and structure morphology	Data regarding the surface composition and structure/morphology are important when reporting data from any experimental study.

Table 2: Computational recommendations to support evaluation of the completeness and quality of curated nanomaterial data

Recommendations	Brief description + importance
Development of computational tools	Automation of computational tools to assess data completeness and quality should be considered.
Standard data exchange templates based on ISA-TAB-Nano specification	Scenario-specific templates.
Provision of completeness and quality scores by nanomaterial data resources	Customisable scoring systems which allow for the selection of criteria upon which the degree of data completeness (in terms of fitness for purpose), or quality, is defined and provide the decision process and justification involved in this.

3.1.2 Nanomaterial Registry

A growing amount of data on engineered nanomaterials has led to the development of the National Institutes of Health (NIH)-funded public tool, Nanomaterial Registry by the US-based National Nanotechnology Initiative (NNI) (https://www.nanomaterialregistry.org/) which is a digital nanotechnology data and information infrastructure to support effective data sharing, collaboration, and innovation across disciplines and applications within the nanomaterial community. that has been developed in support of this goal.

The Nanomaterial Registry properly archives curated nanomaterial data and has made them available to the community. As multidisciplinary nanomaterial data are archived, they are transformed into information via specific data curation and structured presentation. One goal of the Registry is that researchers can use this information in downstream analyses to elicit the discovery of emergent trends and data gaps. Through this mechanism, the Registry will support the continuum of understanding in nanotechnology translating data to knowledge.

3.1.2.1 Minimal Information about Nanomaterials (MIAN)

The Nanomaterial Registry has developed certain Minimal Information about Nanomaterials (MIAN) standards for physico-chemical characteristics and for biological and environmental interaction studies, which are used to calculate the compliance level as well as to determine nanomaterial record similarity which may facilitate read-across or QSAR modelling.

Physico-chemical characteristics (PCCs)

The MIAN for physico-chemical characteristics covers the most descriptive aspects of a nanomaterial or more specifically the characteristics that determine its interaction with the biological and environmental systems. In addition to the characterisation data, the Nanomaterial Registry also captures, the protocols, parameters, and metadata associated with each measurement which can significantly influence the eventually obtained results. Furthermore, data with regards to best practice is collected which may be used to evaluate to quality of the characterization (e.g. raw data available, proper controls used, instrument calibration, use of replicates, measurement protocol, citation protocol, modifications protocol). Table 3 presents the main MIAN for physico-chemical characteristics according to the Nanomaterial Registry.

Table 3: The MIAN for physico-chemical characteristics according to Nanomaterial Registry

Physico-chemical characteristics	
aggregation/agglomeration state	surface charge
shape	surface reactivity
particle size	stability
size distribution	solubility
surface area	surface chemistry
composition	purity

Biological and environmental interaction studies

In addition to MIAN for physico-chemical characteristics, the Nanomaterial Registry has also created MIAN for studies performed in order to obtain biological and environmental interactions of nanomaterials. The minimal information for these types of studies ensures that the data from these experiments are curated in such a way that crucial information in the study is summarized on, for example, how the experiment was performed and what the main conclusions were. This will assist the user in obtaining a better interpretation in relation to its own data (Table 4).

Table 4: The MIAN for biological and environmental interaction studies according to Nanomaterial Registry

Biological interaction studies	
test subject	exposure summary
general study details	endpoint
Environmental interaction studies	
test media	exposure summary
general study details	endpoint

3.1.2.2 Calculation of compliance according to the Nanomaterial Registry

During the curation process of newly archived data, the Nanomaterial Registry uses a specific method to assess and communicate the quality of the data. An algorithm was developed which can calculate the compliance level of the data, taking the quantity and quality of physico-chemical characteristics measurements into account. After the calculation has been completed, the compliance levels are represented as medals in the Registry. For the physico-chemical characteristics a compliance level is provided per category (presented in Table 3). This compliance level then represents a calculated value of the record's adherence to community standards in characterizing and reporting nanomaterial properties. These different records are then awarded one out of four medals based on compliance level scores for each physico-chemical characteristic (from best characterized to least): Gold, Silver, Bronze, and Merit.

For a calculation of the qualitative score, different aspects are taken into account according to their respective degree of specificity. The formula that the algorithm applies is presented in Figure 1. The explanation of the different weighting factors within the formula is presented in Table 5. More information can be found on the Nanomaterial Registry website.

$$CL_{PCC} = \sum_{Measurements} \left\{ \frac{\sum_G \left[\frac{M_G * P_G}{W_G} \right]}{\sum_G M_G} \right\}$$

Figure 1: The formula needed to calculate the compliance level of physico-chemical characteristics as developed by the Nanomaterial Registry

Table 5: Explanation of the weighting factors for physico-chemical characteristics data compliance

Physico-chemical characteristics	**Weighting factors**
Group (G)	Physico-chemical data are divided over several groups e.g. Measurement Type, Technique, Instrument, and several meta-data groups (uncertainty, replicates, etc.)
Element	Each group consists of elements; data which are categorized into a given group are called that group's elements. For the group "Measurement Type", "Mean Hydrodynamic Diameter" is an example of an element.

Multiplier (M)	Each group has an association with it an integer value (i.e. multiplier), which defines the importance of the group to the compliance level score.
Points (P)	The integer, which is associated with each element of each group, is called its point's value.
Weight (W)	For each group within a physico-chemical characteristic, the weight is the product of the group multiplier and the maximum point's value assigned to any element of that group.
Measurement	A measurement is a single characterization of a particle. It may include data for any or all of the groups for a given physico-chemical characteristic.

Results of this calculation could be taken into account e.g., in the read-across tools currently implemented in the eNanoMapper database (nano-lazar), giving a higher weight to better quality measurements. This approach has been used before with positive results (Willighagen et al., 2011).

3.1.1.3 Nanomaterial similarity

Within the Nanomaterial Registry it is also possible to find alike nanomaterials based on similarity between two nanomaterials. This similarity is defined by the Nanomaterial Register as follows:

1. If "SHAPE" is defined for both nanomaterials, it must be equal, regardless of all other data.
2. If the nanomaterials have the same Instance of Characterization for "SIZE" and the Instance of Characterization is not "As Processed," the nanomaterials are a 30% match.
    a. If there is no "SIZE" information, look for "AGRREGATION/AGGLOMERATION STATE".
    b. If the nanomaterials have the Instance of Characterization "As Processed" for "SIZE", and the Techniques for characterization are the same (e.g. DLS), the nanomaterials are only 22.5% similar.
3. If the "SIZE" values are within 10%, those two nanomaterials are an additional 15% match.

   a. But if the "SIZE" values are within 25%, those two nanomaterials are only an additional 10% match.

4. If both nanomaterials have the same Material Type for their most outward chemistry, they are an additional 25% similar.
5. If Isoelectric Point value is within 10% and the nanomaterials were characterized in the same way, another 15% similarity can be added.
   a. But if Isoelectric Point value is only within 25% and the nanomaterials were characterized in the same way, only another 10% similarity can be added.

For such similarity to be able to be calculated, the information required for the calculation must be provided. As such, even this scheme defines a level of *completeness*.

3.1.3 Uniform Description System for Materials on the Nanoscale (UDS)

The CODATA/VAMAS Joint Working Group, chaired by John Rumble, developed the Uniform Description System (UDS) for Materials on the Nanoscale. Within this international working group several representatives from virtually every scientific and technical discipline involved in the development and use of nanomaterials, including physics, chemistry, materials science, pharmacology, toxicology, medicine, ecology, environmental science, nutrition, food science, crystallography, engineering, and others have joined forces. The framework is divided in four "Major Information Categories Used to Describe a Nanomaterial". The structure of the UDS is presented in Figure 2. The UDS provides standards for describing specifically the nanomaterial itself and its production process.

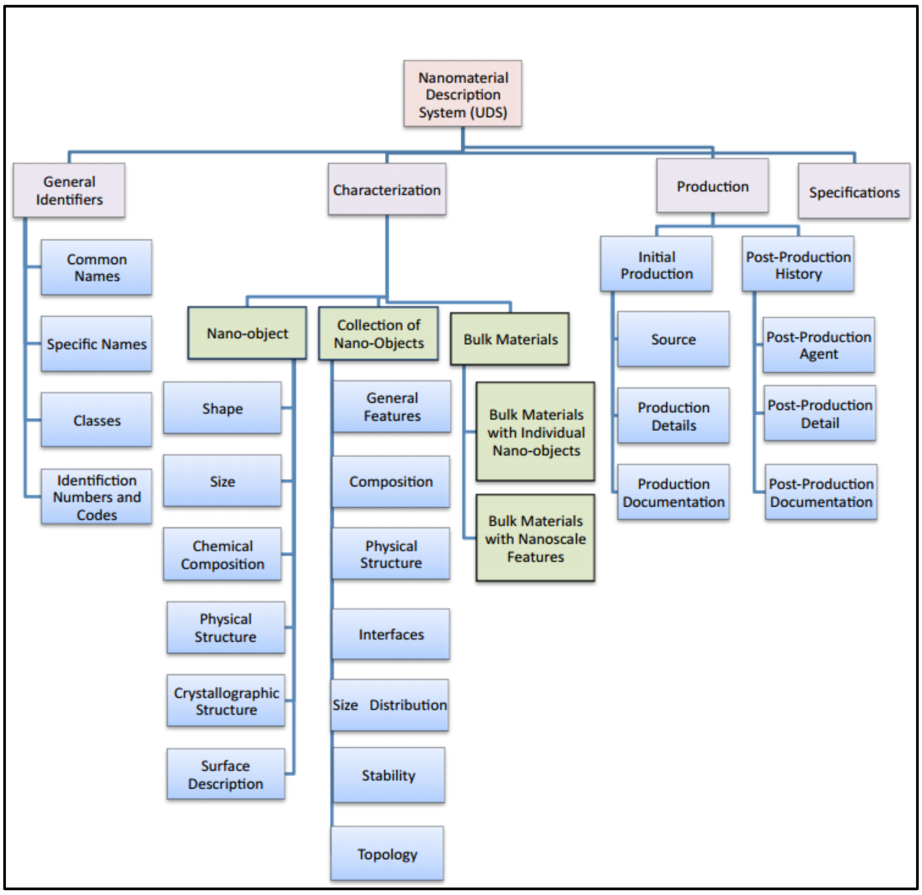


Figure 2: The structure of the Uniform Description System (UDS) for Materials on the Nanoscale as developed by CODATA/VAMAS Joint Working Group (adapted from UDS for Materials on the Nanoscale, v1.0, 1 February 2015)

4. TESTING CONFORMANCE

The comprehensive eNM ontology is developed for the nanosafety domain, encompassing nanomaterials and all information related to their characterisation. The ontology includes and defines common vocabulary terms for use in nanosafety research with the explicit representation of the relationships between different entities, thereby providing effective means for the standardisation and integration. It also includes information describing relevant experimental paradigms, biological interactions and safety indications, in order to support the full scope of relevant research in the determination and characterisation of novel ENMs.

We have developed a pipeline for ontology curation including reuse and integration of community-developed external ontologies.

4.1 Mapping standards to the eNanoMapper ontology

4.1.1 NDCI

Table 6: NDCI annotation terms mapping results

NDCI Recommendations	Annotations of eNanoMapper ontology
Physico-chemical characteristics	http://purl.enanomapper.org/onto/ENM_9000015
Temporal measurements	http://www.ebi.ac.uk/efo/EFO_0000719
Assessment of artefacts	http://purl.enanomapper.org/onto/external/iao-slim.owl (iao_slimmed) http://purl.enanomapper.org/onto/internal/iao-ext.owl (iao_ext ontology)
Experimental errors and uncertainty	http://semanticscience.org/resource/SIO_000769 (uncertainty value)
Impurities	http://purl.obolibrary.org/obo/UO_0000193 (purity percentage)
Manufacturer ID	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C104504 (batch Number)

Measurement details	http://purl.obolibrary.org/obo/IAO_0000100 (data set) http://purl.enanomapper.org/onto/ENM_8000021 (data type) http://purl.obolibrary.org/obo/OBCS_0000011 (measurement scale) http://purl.bioontology.org/ontology/npo#NPO_1779 (datum value) http://purl.bioontology.org/ontology/npo#NPO_1680 (parameter) http://www.bioassayontology.org/bao#BAO_0000179 (result) etc.
Provenance	http://purl.obolibrary.org/obo/core#provenance_notes
Surface composition and structure morphology	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C53414 (composition)

4.1.2 Nanomaterial Registry

Physico-chemical characteristics (PCCs)

Table 7: The physico-chemical characteristics MIAN mapping results

Physico-chemical characteristics	
PPCs MIAN standards	**eNM ontology entity URIs**
Aggregation	http://purl.bioontology.org/ontology/npo#NPO_1967
Shape	http://purl.bioontology.org/ontology/npo#NPO_274
Particle size	http://purl.bioontology.org/ontology/npo#NPO_1694
Size distribution	http://purl.bioontology.org/ontology/npo#NPO_1697
Surface area	http://purl.bioontology.org/ontology/npo#NPO_1235
Composition	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C53414

Surface charge	http://purl.bioontology.org/ontology/npo#NPO_1812
Surface reactivity	-
Stability	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C54072
Solubility	http://purl.obolibrary.org/obo/PATO_0001536
Surface chemistry	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C64351
Purity	http://purl.obolibrary.org/obo/UO_0000193

4.1.3 UDS

In the Spring 2016 version of the NSC Newsletter the mapping of descriptive terms for nanomaterials, from the UDS (Uniform Description System, CODATA-VAMAS Working Group On the Description of Nanomaterials. (2015)) to the corresponding eNanoMapper ontology terms, as performed by the Maastricht University team, was presented. The respective terms were retrieved from the Uniform Description System for Materials on the Nanoscale v1.0. Zenodo. 10.5281/zenodo.20688). The starting point of this action was the current UDS document provided by John Rumble, discussed in detail during an UDS/eNanoMapper workshop in May 2015 (codata.org/blog/2015/08/07).

We extracted 301 terms which are used to describe nanoparticles including terms about properties, production details, measured values, and stability information. An example of a group of extracted terms is listed in Table 8. For the mapping process it is necessary to consider not only the term itself but also its position in the classification system and description. For example, depending on the description, the term “concentration” could lead to a measured value with the property “concentration”, to given product information or an assay description. The term ”nanoplate” listed in Table 8 e.g. was therefore recognized as “material entity” and, in detail, as a “1 dimensional nano-object”. If a term or it’s synonym was not found in the eNanoMapper ontology it was added to the list of terms to be added to the next ontology version. With the 3rd release of the ENM ontology 38 terms from UDS were added and now there are 212 terms of UDS matching to the ontology. For a complete list of annotations see Appendix A.

Table 8: Example for UDS terms and description (UDS, page 12, table 2).

Examples of Formal Classes Approved by ISO TC 229	
Term	**Description**
Nanoparticle	Nano--object with all three external dimensions in the nanoscale

Nanoplate	Nano--object with one external dimension in the nanoscale and the two other external dimensions significantly larger
Nanofibre	Nano--object with two similar external dimensions in the nanoscale and the third dimension significantly larger
Nanotube	Hollow nanofibre
Nanorod	Solid nanofibre
Nanowire	Electrically conducting or semi--conducting nanofibre

4.2 USING SPARQL TO TEST FOR DATA COMPLETENESS

SPARQL is a query language which can be used for searching data in a Resource Description Framework (RDF) format. In the RDF format, data is stored within a graph database. In a graph database, information is compiled according to a series of statements which are also called triples. As the name triple already indicates, RDF statements are normally broken down into three constituent parts: the subject, predicate, and object of the statement. All the possible triples can be stored and retrieved using a triple store. The retrieval of the triples from an RDF format can be done using a query language such as SPARQL.

The advantage of RDF is that it not only holds data, but that the eNanoMapper ontology is in fact represented in Web Ontology Language (OWL), which is also using the RDF. Consequently, the lines between schema and data disappear, something one can take advantage of, as we explored in this deliverable.

With the available data in the RDF format (see D3.3) and the ontology in the OWL format, specific SPARQL queries may be developed to retrieve, for example, information on the physico-chemical characteristics as mentioned by any of the aforementioned data completeness schemes, and used to test the compliance with those schemes.

4.2.2 PHYSICAL CHEMICAL PROPERTIES

Several schemes require, at a minimum, size information to be provided. However, size information can be available in various forms with a variety of endpoints. Size information, for example, can use a variety of ontology annotations, depending on what specifically was measured. See the eNanoMapper subtree on the right side of Figure 3.

Using the ontology approach, the query can take advantage of the hierarchy of the ontology and one only needs to identify the top-most ontology terms (such as “size” in the right tree) and use that in the query to type the assay:

```
BIND (npo:NPO_1697 as ?propertyType)
{ ?assay a ?propertyType . }
UNION
{ ?assay a [ rdfs:subClassOf+ ?propertyType ] }
```

In this example the top node is NPO's term for "particle size". The *rdfs:subClassOf* predicate reflects the hierarchical structuring of the eNanoMapper ontology as expressed in OWL.

4.2.1 NANOMATERIAL REGISTRY

Using the ontological mappings in Section 4.1, SPARQL queries can be defined to test if a particular nanomaterial has a piece of expected information available. For example, to test if 'particle size' (*NPO_1694*) information is available for some nanomaterial (e.g. *ex:NWKI-002f5129-d46a-39c7-8f26-5626aec2174e*), we can run the following SPARQL query:

```
prefix obo: <http://purl.obolibrary.org/obo/>
prefix bao: <http://www.bioassayontology.org/bao#>
prefix sso: <http://semanticscience.org/resource/>
prefix npo: <http://purl.bioontology.org/ontology/npo#>
prefix ex: <http://localhost/ambit2/substance/>
```

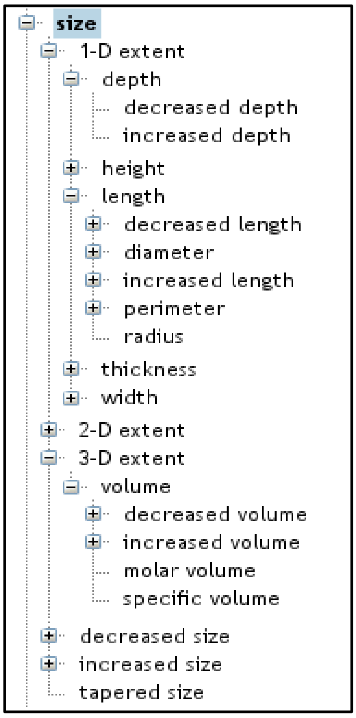


Figure 3: Physical Chemical Property 'size'

```
select distinct ?substance ?type ?title ?value ?unit where {
 BIND (ex:NWKI-002f5129-d46a-39c7-8f26-5626aec2174e as ?substance)
 BIND (npo:NPO_1694 as ?propertyType)
 { ?assay a ?propertyType . }
 UNION
 { ?assay a [ rdfs:subClassOf+ ?propertyType ] }
 ?substance a obo:CHEBI_59999 ;
  obo:BFO_0000056 ?mgroup .
 ?mgroup obo:OBI_0000299 ?endpoint .
 ?endpoint sso:has-value ?value ;
      sso:has-unit ?unit .
 ?assay a bao:BAO_0000015, ?type ;
  bao:BAO_0000209 ?mgroup ;
  dc:title ?title .
 FILTER (?type != bao:BAO_0000015)
} ORDER BY ASC(?substance)
```

4.3 Evaluation

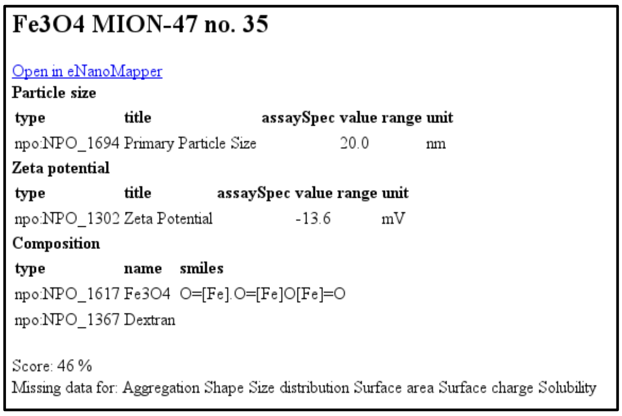
To experiment with the above proposed approach, we tried to re-implement the MIAN scoring function used by the Nanomaterial Registry. A script was developed that implements the ideas of this scoring function. It tests if the required information is available, and based on the amount of such information, it calculates a score. The more information is available, the higher the score.

4.3.1 Methods

The script was implemented as a Bioclipse (Sjuth et al., 2009) script (see github.com/egonw/completeness). The mappings from Section 4.1 are used in combination with SPARQL queries as such discussed in section 4.2. The script tests for all materials in each data set made available in an eNanoMapper SPARQL endpoint (see D3.3) if information is provided. As discussed, the SPARQL queries take into account the ontological relation between terms. Therefore, if the aim is to find out if the size information is provided, this can be as a general term of “particle size” and also a more specific term like “diameter”.

4.3.2 Results

The script tests for the availability of data for nine properties: NPO_1967, NPO_274, NPO_1694, NPO_1697, NPO_1235, NPO_1812, and NPO_1302 from the NanoParticle Ontology, and these two from other ontologies: PATO_0001536, and C53414 (see the aforementioned mappings). An HTML report is generated reporting the properties found and an indication of which are missing.


Fe3O4 MION-47 no. 35

Open in eNanoMapper

Particle size

type	title	assaySpec	value	range	unit
npo:NPO_1694	Primary Particle Size		20.0		nm

Zeta potential

type	title	assaySpec	value	range	unit
npo:NPO_1302	Zeta Potential		-13.6		mV

Composition

type	name	smiles
npo:NPO_1617	Fe3O4	O=[Fe].O=[Fe]O[Fe]=O
npo:NPO_1367	Dextran	

Score: 46 %

Missing data for: Aggregation Shape Size distribution Surface area Surface charge Solubility

Figure 4: Validation results for extracted data about an iron nanomaterial from a PNAS publication by Shaw et al., 2008

The script basically iterates over data sets in a SPARQL endpoint and for each studies all materials in that data set. For each material it checks if the properties are given. Because each property may need a different SPARQL query, the template states:

```
"C53414" : [
 label : "Composition",
 score : 1.0,
iri : "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C53414"
query : "compositionCheck"
]
```

The result matrix returned by the SPARQL query reflects the differences in the queries, as can be seen in Figure 4. A simple additive scheme now results in a score represented as a percentage from [0,100]. A histogram can then visualise how sparse a particular matrix is. This is shown in Figure 5 for the NanoWiki. It must be stressed that the current scores also largely reflect the ontology annotation. The scores must not be seen as quality scores, and are merely a score of how much results are machine readable (at least at this moment).

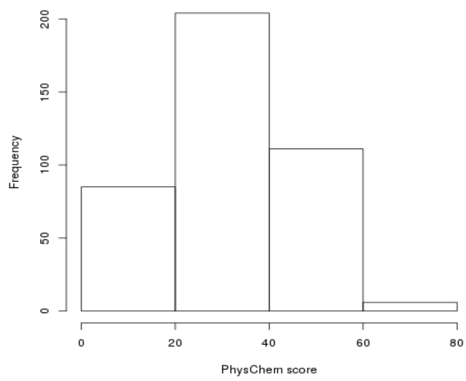


Figure 5: Distribution of scores of data completeness of 406 NanoWiki materials. Being mostly data extracted from literature, and focussing on testing nanoQSAR hypothesis, scores are generally low.

4.3.3 Discussion

The current approach does not manage to fully implement the MIAN scoring function. Partially this is due to information we cannot find: the functions in the paper and on the website of the Nanomaterial Registry seem different, but mostly the weights are missing. Furthermore, the MIAN approach turns out to require many SPARQL queries, as the current eNanoMapper RDF structures information differently, normalizing the structure to the ontologies on which the eNanoMapper ontology is based. This is not a technical limitation, and merely means that the generalization that this approach introduces, stresses the importance of well-structured and well-annotated data and ontology. In short, it does not reduce the amount of data curation needed, and supports the idea that researchers must set aside a clear amount of funding for research data management.

The lack of weights may be overcome by introducing a custom scheme. Still, the overall scores are low. There are various reasons that can be discussed here. First, the data set it was tested with is data (manually) extracted from literature and literature is generally short on physical-chemical characterization, as journals have not yet adopted community standards for minimal reporting standards. Second, the used NanoWiki data set is only sparsely annotated with ontology terms. In fact, in the end, part of the work leading to this deliverable involved actually further improving the annotation. There are currently no NanoSafety Cluster data sets that have a sufficient level of ontological annotation, and the NANoREG data may be the first to fulfil the requirements to really be a proper data set with which to test this approach.

We also note that the use of the ontological structure may be a good alternative to calculate weights for contributions to the final score. For example, the path length between the top level property being tested for, say, "particle size" and "diameter" reflects that the latter is more precise, exactly what MIAN models in their score.

4.3.4 Evaluation Conclusion

Taking these discussion points into account, the experiment shows that the approach can indeed implement the idea of a scoring function that reflects the "quality" of the data. The idea of expressing contributions to the score as a series of repeated data queries formalized in SPARQL queries, abstracts the idea and links to calculation of the common language eNanoMapper is developing. That improves the interoperability of the score calculation and makes it easier to change it to particular needs. The latter may show critical information when the score is used as weights in statistical modelling, which is known to improve pattern recognition as we showed with the ChEMBL data. But unlike with the ChEMBL data example, this data is not tabularly structured and is much easier to annotate. Nanosafety data is much less available at this moment and much more diverse in nature, making ontological annotation needed of many more resources.

5. CONCLUSION

The comprehensive curation of nanomaterial data into databases allows for the development of improved methods for risk assessment of nanomaterials. In this report, we have assessed data for their completeness and quality within the nanosafety domain. Several minimal reporting standards have been introduced and applied to the eNanoMapper ontology-based approach, which indicated to what extent the information is complete, as well as the reliability of computational models given the input ENMs:

1) Developed a pipeline for ontology curation including reuse and integration of community-developed external ontologies. Mapped minimal reporting standards terms to eNanoMapper classes, which can test data annotation and completeness of the data.
2) The ontology also has been used to support data extraction, i.e. automated conformance testing of data. The data was also combined by taking advantage of the RDF and linked data export of the data from the database. The automated conformance was tested by using the semantic web query language SPARQL, and shows that tests can be integrated.

6. BIBLIOGRAPHY

CODATA-VAMAS Working Group On the Description of Nanomaterials. (2015). Uniform Description System for Materials on the Nanoscale v1.0. Zenodo. 10.5281/zenodo.20688

Hendren, C. O., Powers, C. M., Hoover, M. D., & Harper, S. L. (2015). The Nanomaterial Data Curation Initiative: A collaborative approach to assessing, evaluating, and advancing the state of the field. *Beilstein journal of nanotechnology*, *6*(1), 1752-1762.

Lijowski, M.(2010, 11). CaNanoLab – A Tool To Benefit Biomedical Nanomaterials Research. *Nature Precedings*. doi:10.1038/npre.2010.5175.1

Marquardt, C., D. Kühnel, V. Richter, H. F. Krug, B. Mathes, C. Steinbach, and K. Nau. "Latest Research Results on the Effects of Nanomaterials on Humans and the Environment: DaNa – Knowledge Base Nanomaterials." *Journal of Physics: Conference Series* 429 (04, 2013): 012060. doi:10.1088/1742-6596/429/1/012060.

Robinson, R. L. M., Lynch, I., Peijnenburg, W., Rumble, J., Klaessig, F., Marquardt, C., ... & Karcher, S. (2016). How should the completeness and quality of curated nanomaterial data be evaluated?. *Nanoscale*, *8*(19), 9919-9943.

Shaw, S. Y., Westly, E. C., Pittet, M. J., Subramanian, A., Schreiber, S. L., Weissleder, R., May 2008. Perturbational profiling of nanomaterial biologic activity. Proceedings of the National Academy of Sciences 105 (21), 7387-7392.

Slooff W. (2016). Ecotoxicological Effect Assessment: Deriving Maximum Tolerable Concentrations (MTC) from Single-species Toxicity Data.". Web

Spjuth, Ola, et al. "Bioclipse 2: A scriptable integration platform for the life sciences." *BMC bioinformatics* 10.1 (2009): 397.

Willighagen, E. L., Alvarsson, J., Andersson, A., Eklund, M., Lampa, S., Lapins, M., ... & Wikberg, J. E. (2011). Linking the resource description framework to cheminformatics and proteochemometrics. Journal of Biomedical Semantics, 2(1), 1.

ANNEXES

Appendix A: Complete list of UDS mappings to ENM ontology terms

Table 2: List of terms which are identical or synonym to ENM terms and their URIs

URI	term
http://purl.obolibrary.org/obo/ENVO_00010505	aerosol
http://purl.bioontology.org/ontology/npo#NPO_1365	aspect ratio
http://purl.obolibrary.org/obo/OBI_0000070	assay
http://www.bioassayontology.org/bao#BAO_0000525	assay comment
http://www.bioassayontology.org/bao#BAO_0000520	assay description
http://www.bioassayontology.org/bao#BAO_0000522	assay narrative
http://www.bioassayontology.org/bao#BAO_0000523	assay protocol
http://www.bioassayontology.org/bao#BAO_0000521	assay title
http://semanticscience.org/resource/CHEMINF_000446	cas registry number
http://semanticscience.org/resource/CHEMINF_000231	charge density descriptor
http://semanticscience.org/resource/CHEMINF_000131	charge descriptor
http://purl.obolibrary.org/obo/UO_0000219	charge unit
http://purl.bioontology.org/ontology/npo#NPO_1497	chemical component
http://purl.bioontology.org/ontology/npo#NPO_1494	chemical component in nanoparticle formulation
http://semanticscience.org/resource/CHEMINF_000464	chemical database identifier
http://semanticscience.org/resource/CHEMINF_000123	chemical descriptor
http://purl.bioontology.org/ontology/npo#NPO_1890	coat component
http://www.bioassayontology.org/bao#BAO_0000180	concentration endpoint
http://purl.obolibrary.org/obo/UO_0000189	count unit
http://purl.bioontology.org/ontology/npo#NPO_1512	crystalline state
http://semanticscience.org/resource/CHEMINF_000060	dimensional extend descriptor
http://www.bioassayontology.org/bao#BAO_0000179	endpoint/result
http://purl.bioontology.org/ontology/npo#NPO_1909	engineered nanomaterial
http://purl.bioontology.org/ontology/npo#NPO_471	entrapment
http://purl.bioontology.org/ontology/npo#NPO_1828	entrapped component
http://purl.bioontology.org/ontology/npo#NPO_1597	fiat material part
http://purl.bioontology.org/ontology/npo#NPO_1616	functionalization of nanoparticle
http://semanticscience.org/resource/CHEMINF_000093	geometric descriptor
http://purl.obolibrary.org/obo/UO_0000152	half life

http://purl.bioontology.org/ontology/npo#NPO_1436	instrument
http://semanticscience.org/resource/CHEMINF_000107	iupac name
http://purl.obolibrary.org/obo/UO_0000001	length unit
http://purl.obolibrary.org/obo/RO_0001015	location of
http://purl.bioontology.org/ontology/npo#NPO_1822	mean particle size
http://purl.bioontology.org/ontology/npo#NPO_1798	measure of variability
http://purl.bioontology.org/ontology/npo#NPO_1805	measured value
http://purl.bioontology.org/ontology/npo#NPO_1495	molecular component in nanoparticle
http://semanticscience.org/resource/CHEMINF_000043	molecular entity name
http://semanticscience.org/resource/CHEMINF_000042	molecular formula
http://purl.bioontology.org/ontology/npo#NPO_1018	nano-object
http://purl.bioontology.org/ontology/npo#NPO_199	nanomaterial
http://purl.bioontology.org/ontology/npo#NPO_1404	nanoparticle sample
http://purl.bioontology.org/ontology/npo#NPO_122	nanorod
http://purl.bioontology.org/ontology/npo#NPO_636	nanosphere
http://purl.bioontology.org/ontology/npo#NPO_1910	nanostructured material
http://purl.bioontology.org/ontology/npo#NPO_126	nanotube
http://purl.bioontology.org/ontology/npo#NPO_383	nanowire
http://purl.bioontology.org/ontology/npo#NPO_1539	particle diameter
http://purl.obolibrary.org/obo/UO_0000187	percent
http://www.bioassayontology.org/bao#BAO_0002132	percent purity
http://semanticscience.org/resource/CHEMINF_000025	physical features
http://purl.bioontology.org/ontology/npo#NPO_888	physical state
http://semanticscience.org/resource/CHEMINF_000044	preferred name
http://purl.obolibrary.org/obo/UO_0000109	pressure unit
http://purl.obolibrary.org/obo/BFO_0000015	process
http://purl.obolibrary.org/obo/OBI_0000272	protocol
http://purl.obolibrary.org/obo/BFO_0000019	quality
http://purl.bioontology.org/ontology/npo#NPO_589	quantum dot
http://purl.obolibrary.org/obo/OBCS_0000076	range
http://www.bioassayontology.org/bao#BAO_0000709	research institute
http://www.bioassayontology.org/bao#BAO_0000179	result
http://purl.bioontology.org/ontology/npo#NPO_274	shape
http://purl.bioontology.org/ontology/npo#NPO_760	shell
http://purl.bioontology.org/ontology/npo#NPO_1889	shell component
http://purl.bioontology.org/ontology/npo#NPO_1697	size distribution
http://purl.bioontology.org/ontology/npo#NPO_1807	size value
http://purl.bioontology.org/ontology/npo#NPO_888	state
http://semanticscience.org/resource/CHEMINF_000085	structural descriptor
http://www.bioassayontology.org/bao#BAO_0000384	summary content

http://semanticscience.org/resource/CHEMINF_000229	surface area descriptor
http://purl.bioontology.org/ontology/npo#NPO_1962	surface coating
http://purl.bioontology.org/ontology/npo#NPO_1883	surface functionalization of nanoparticle
http://purl.bioontology.org/ontology/npo#NPO_1944	synthesis part
http://www.bioassayontology.org/bao#BAO_0000585	temperature endpoint
http://purl.bioontology.org/ontology/npo#NPO_1806	temperature value
http://semanticscience.org/resource/CHEMINF_000092	topological descriptor
http://purl.obolibrary.org/obo/UO_0000000	unit
http://purl.obolibrary.org/obo/UO_0000095	volume endpoint

Table 3: Terms from UDS to be added to the ENM ontology

URI	**term**
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C79889	authority
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C104504	batch number
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C53414	composition
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C75947	decay
http://purl.bioontology.org/ontology/MESH/D004282	documentation
http://purl.bioontology.org/ontology/npo#NPO_1414	emulsifier
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C62381	engineered nanoparticle
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C93591	manufacture date
http://purl.bioontology.org/ontology/SNOMEDCT/255508009	medium
http://semanticscience.org/resource/SIO_000116	name
http://purl.obolibrary.org/obo/CHMO_0002855	percentage yield
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C54072	stability
http://purl.bioontology.org/ontology/SNOMEDCT/22303008	version