DELIVERABLE REPORT D3.3

DELIVERABLE REPORT D3.3

Modules and services for linking and integration with third party databases

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:	UM
DATE:	2016-12-12
VERSION:	V1.0

Call identifier	FP7-NMP-2013-SMALL-7
Document Type	Deliverable Report
WP/Task	WP3/T3.4 T3.6
Document ID	eNanoMapper D3.3
Status	Draft

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	Egon Willighagen (UM) Micha Rautenberg, Denis Gebele (IST) Penny Nymark, Pekka Kohonen (MB) Nina Jealiazkova (IDEA) Review by Barry Hardy (DC)
Purpose of the Document	To report on T3.4 and T3.6.
Document History	1.Table of Contents, 2016-07-07 2.Full draft, 2016-11-21 3.Completed version, 2016-12-12

TABLE OF CONTENTS

TABLE OF FIGURES

GLOSSARY

Abbreviation / acronym	Description
COD	Crystallography Open Database
DOI	Digital Object Identifier
FOAF	Friend of a Friend ontology
HTTP	HyperText Transport Protocol
IRI	Internationalized Resource Identifier
OWL	Web Ontology Language
RDF	Resource Description Framework
SPARQL	SPARQL Protocol and RDF Query Language

1. EXECUTIVE SUMMARY

Understanding the biological effects of nanomaterials needs at least insight into the physicochemical identity; recent research has however shown how important the biological identity is in fully understanding the biological mechanisms. This requires, however, interlinking nanomaterial databases with databases from other domains. This deliverable reports on our efforts outlined in Tasks 3.4 and 3.6 to implement the Linked Data ideas to data in the nanosafety community, taking into account this recent guidance document, experimenting with a number of technical solutions to link data. We report on work that lead to the Resource Description Framework (RDF) support of the database, reusing the eNanoMapper ontology, and interlinking with other databases. We show how the RDF can be used and demonstrate its applicability with a few examples. The related deliverable D5.6 to follow is about data completeness and is also based on the output of this work.

2. INTRODUCTION

Understanding the biological effects of nanomaterials is, at the core, based on the physicochemical identity, but recent research has shown how important the biological identity is in fully understanding the biological mechanisms. This requires, however, interlinking nanomaterial databases with databases from other domains [1].

Linked Data is a concept from computing sciences reflecting that data should not only be machine readable, but also linked across independent databases. As such, it provides an implementation of the idea and the requirement outlined in the first paragraph. Formally, Linked Data are data that conform to a number of conventions: it is exposed as semantic web data and resources are identified with IRIs; the IRIs are dereferenceable, which means that a user can use the IRI to retrieve information about that resource using the HTTP standard; the data set links out to other data sets via Internationalized Resource Identifiers (IRIs).

The reason why this approach is needed for nanosafety informatics was recently established by a collaboration of members of the European NanoSafety Cluster's Working Group 4 (WG4) on Databases and the U.S.A. NanoWG. The resulting paper, submitted to a peer-reviewed journal, can be found in Annex A.

This deliverable reports on our efforts outlined in Task 3.6 to implement the Linked Data ideas to data in the nanosafety community, taking into account this recent guidance document, experimenting with a number of technical solutions to link data.

3. LINKED DATA

According to a recent survey (see Annex A), there is a need for linking databases. Rather than a single repository or database that aggregates all information, the nanosafety field is so broad, it is preferential to have targeted databases instead. By linking the databases and ensuring a sufficient level of interoperability at varying levels, researchers are still able to efficiently integrate data before they set out their data analysis.

The survey resulted in a recommendation of interlinked data sets (see Annex A). The eNanoMapper work behind this deliverable was started in Task 3.6 by implementing the suggested approaches and by exploring how well the work went.

Using the data available in the data.enanomapper.net server, and particularly the NanoWiki and Protein Corona data sets (Task 3.4), we selected a subset of external databases to link to. The aim of this linking is to support the development of uses cases.

Before introducing a number of linked databases below, it should be stressed that a key advantage of semantic web approaches is that the ontology is available in the same format. As such, data exported from eNanoMapper is automatically linked to the eNanoMapper ontology and the ontologies embedded in that.

3.1 Linked Nanomaterial Repositories

While the next sections show a few databases which happen to also include nanomaterials, there are a few nanomaterial-specific databases around. The eNanoMapper project currently links to one other nanomaterial database, caNanoLab. The materials in caNanoLab are indexed in the search application at search.data.enanomapper.net and when finding materials from caNanoLab in the search results, only basic information is provided about the found materials. The result list links to caNanoLab, and for the user to get detail, the user follows the link to view that detail in caNanoLab itself.

3.2 Linked Biology Databases

Four databases with biological content are currently linked to and from the eNanoMapper database available at data.enanomapper.net. One is the ChEMBL database with information about interaction of chemicals with biological entities (targets), mostly proteins and protein complexes. Some of these compounds, like fullerene, are classified as nanomaterials. Four C_{60} derivatives in ChEMBL have been entered into NanoWiki and are part of the data.enanomapper.net content.

The eNanoMapper database is linked to two additional databases with transcriptomics experimental data, the ArrayExpress (www.ebi.ac.uk/arrayexpress/) and the Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo/) databases. The ArrayExpress and GEO databases contain experiments that study the transcriptional changes due to exposure to a nanomaterial. Eleven nanomaterials are provided with links to ArrayExpress, reflecting nine different assays. For the GEO, three nanomaterials have been provided with links in the eNanoMapper database, studied with four different assays. Physicochemical properties of these nanomaterials are captured in the eNanoMapper database.

These three databases describe the biological properties of nanomaterials and constitute links to these databases. For example, ChEMBL links to UniProt for protein information, and ArrayExpress and GEO include probe identifiers which can be linked to gene databases uses in BridgeDb (see Section 4.2). This way, by linking out to one database, taking advantage of the Linked Data cloud, it effectively links to many other databases too.

However, eNanoMapper in itself also has information about biological properties. For example, the Protein Corona dataset describes which proteins bind to the tested nanomaterials (silver and gold cores, coated with a series of coating components). These proteins are annotated to the UniProt database, linking data.enanomapper.net to UniProt.

3.3 Linked Chemical Databases

The Crystallography Open Database (COD) is a database of crystal structures. A number of the crystal structures cover nanoparticles, which are apparently well-defined enough to measure the chemical structure. We captured eight nanoparticles found in the COD with a chemical formula similar to C270.75H371Cd18Cl8N18O105.917Tb6, which is quite uncommon and very detailed, compared to chemical formulas for many metal oxides which merely capture the rough ratio of elements such as TiO_2 for titanium dioxide

3.4 Linked Vendor Databases

To demonstrate the possibility of linking vendor databases, we have manually extracted minimal physical-chemical characterization provided on the vendor website. In this trial we have selected Sigma-Aldrich as the vendor and selected 10 nanomaterials. In most cases chemical composition, size, and melting point were provided. Size is absent for the C_{60} buckyball product. For some nanomaterials we also extracted other properties like HOMO and LUMO information.

3.5 Other Linked Databases

A database that does not fit in any of the aforementioned categories is CrossRef, which hosts literature reference information and hands out the Digital Object Identifier (DOI) for journal articles, book chapters, software, and data.

4. LINK SETS

Link Sets are a mechanism proposed by the Open PHACTS project allowing the linking of data sets. Links between data sets are needed because Open PHACTS does not modify externally provided data sets before it gets used in their knowledge base. Instead, link sets indicate how the various data sets are linked together. That allows the system to still be queried over all data sets by modularizing the approach. Practically, this means that data integration is further decoupled from data analysis and data aggregation.

Better still these link sets can be turned on and off on demand. This idea is formalized in Open PHACTS as "scientific lenses" and describes how and when those link sets should be used. For example, some data analysis might be aiming to equate all nano-titanium dioxides, because you are interested in overall safety. However, when you wish to study fine-grained effects due to size differences, even TiO_2 nanoparticles of 20 and 25 nm may need to be distinguished. Similarly, a "synthesis batch" effect can be turned on or off, depending on whether this is relevant to the research question.

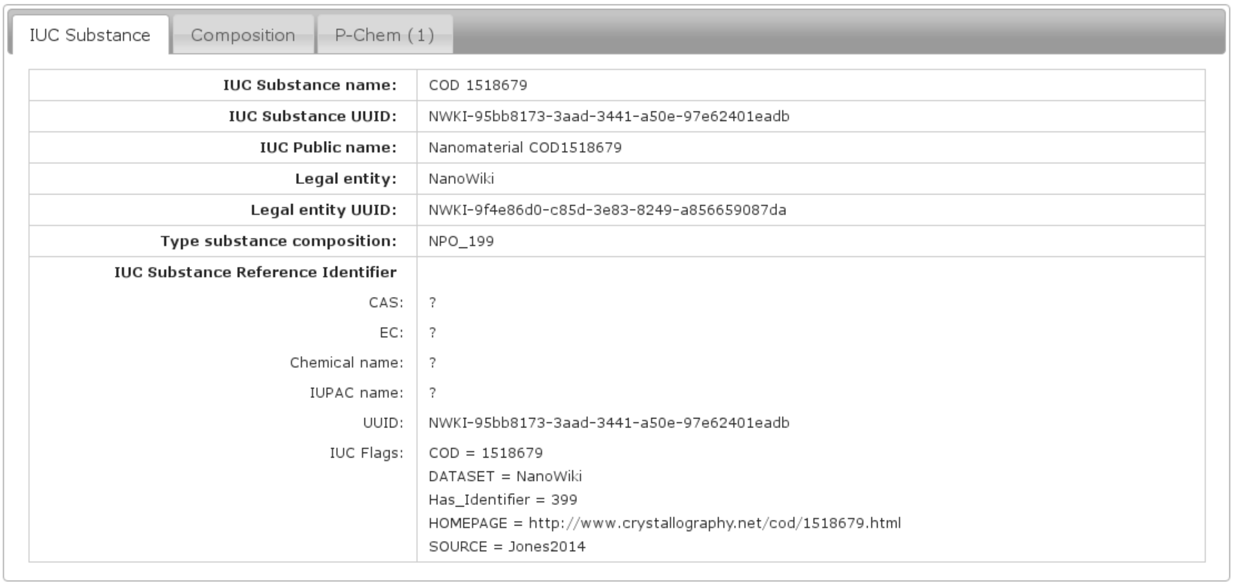
In collaboration with the NanoSafety Cluster WG4 on databases (task 2016-1), we have developed a number of link sets, summarized below. The full list of link sets mentioned in this table are available in Annex B.

4.1 Link Sets enriching databases

The link sets given in the next sections are finding their way into the RDF but also on the webpages of the demonstration data server at http://data.enanomapper.net/ and the search engine at http://search.data.enanomapper.net/.

4.1.1 Homepages

Not all other databases support the semantic web formats, or have individual web pages for individual nanomaterials. For these cases we have adopted the FOAF ontology [2], allowing nanomaterials to have links to remote (database) webpages. The ability to *deep link* to such pages makes manual information retrieval by scholars easier.



IUC Substance name:	COD 1518679
IUC Substance UUID:	NWKI-95bb8173-3aad-3441-a50e-97e62401eadb
IUC Public name:	Nanomaterial COD1518679
Legal entity:	NanoWiki
Legal entity UUID:	NWKI-9f4e86d0-c85d-3e83-8249-a856659087da
Type substance composition:	NPO_199
IUC Substance Reference Identifier	
CAS:	?
EC:	?
Chemical name:	?
IUPAC name:	?
UUID:	NWKI-95bb8173-3aad-3441-a50e-97e62401eadb

IUC Flags:	COD = 1518679
	DATASET = NanoWiki
	Has_Identifier = 399
	HOMEPAGE = http://www.crystallography.net/cod/1518679.html
	SOURCE = Jones2014

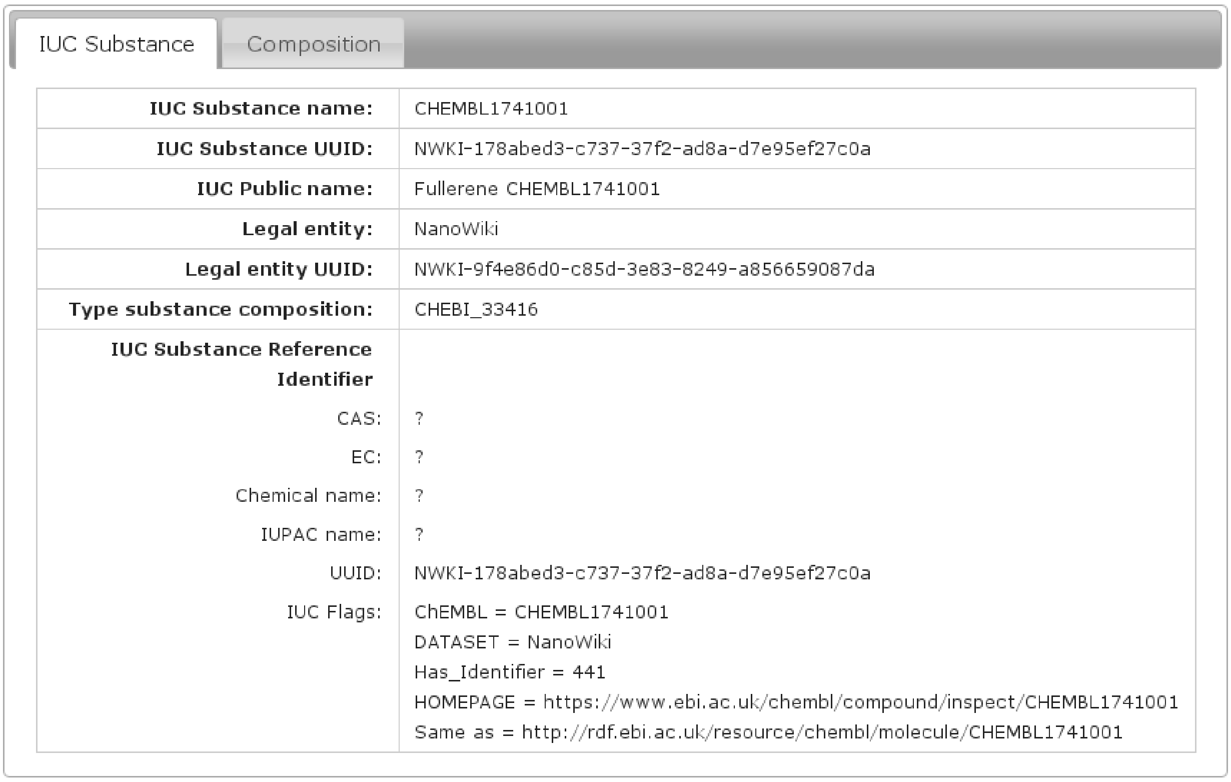
Fig. 1: Screenshot of nanomaterials in the Crystallography Open Database with a link to the homepage of that nanomaterial in that database.

Other databases that use the homepage link from the FOAF ontology include the Sigma-Aldrich demonstration data. These links can be recreated from the eNanoMapper RDF exports with SPARQL CONSTRUCT queries like the following for Sigma-Aldrich:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
CONSTRUCT {
  ?enmMaterial foaf:page ?linkout
} WHERE {
  ?enmMaterial foaf:page ?linkout .
  FILTER (regex(str(?linkout), "sigmaaldrich.com"))
}
```

4.1.2 Semantic equivalence

If the linked database does provide semantic access to their data, then a semantic link can be made. An example of this is the use of the *sameAs* relation from the OWL ontology. An example of such linking is provided with the links to the ChEMBL database [3]. Here, eNanoMapper collected a small set of fullerene-based materials (buckyballs, fullerenols) frequently classified as nanomaterials. For this set, both links to homepages and *sameAs* relations are provided.



IUC Substance name:	CHEMBL1741001
IUC Substance UUID:	NWKI-178abed3-c737-37f2-ad8a-d7e95ef27c0a
IUC Public name:	Fullerene CHEMBL1741001
Legal entity:	NanoWiki
Legal entity UUID:	NWKI-9f4e86d0-c85d-3e83-8249-a856659087da
Type substance composition:	CHEBI_33416
IUC Substance Reference Identifier	
CAS:	?
EC:	?
Chemical name:	?
IUPAC name:	?
UUID:	NWKI-178abed3-c737-37f2-ad8a-d7e95ef27c0a

IUC Flags:	ChEMBL = CHEMBL1741001
	DATASET = NanoWiki
	Has_Identifier = 441
	HOMEPAGE = https://www.ebi.ac.uk/chembl/compound/inspect/CHEMBL1741001
	Same as = http://rdf.ebi.ac.uk/resource/chembl/molecule/CHEMBL1741001

Fig. 2: Screenshot of a nanomaterial listed in the ChEMBL database with both a linked homepage and a semantic link using the *sameAs* relation from the OWL ontology.

These links can be recreated from the eNanoMapper RDF exports with SPARQL CONSTRUCT queries like the following for ChEMBL:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
construct {
  ?enmMaterial owl:sameAs ?linkout
} where {
  ?enmMaterial owl:sameAs ?linkout .
  FILTER (regex(str(?linkout), "chembl"))
}
```

4.2 BridgeDb - Dynamic Linking

Database linking for genes, proteins, and metabolites is well established. We opted for the adoption of in BridgeDb framework [4] developed by the UM partner.

4.2.1 Data Sources

BridgeDb captures the following information about databases: what kind of entities it captures, the format of the identifier, the kind of entities captured in that database, and patterns of webpage URLs for those entities. A full overview is given in Table X with examples for the Ensembl gene database.

```
Ensembl     En    http://www.ensembl.org/
      http://ensemblgenomes.org/id/$id ENSG00000139618  gene       1
      urn:miriam:ensembl     ^ENS[A-Z]*[FPTG]\d{11}$     Ensembl
```

If available, information about this database identifier in the EMBL-EBI MIRIAM Registry is included too.

We previously reported about the development of documented BridgeDb web services that can be used to map identifiers, for genes, proteins, metabolites, and biological interactions. An anticipated goal was to provide mappings for nanomaterial identifiers between databases too, but the currently available link sets are not large enough in size (number of mappings) to warrant investing time in this.

However, a task is ongoing with the NanoSafety Cluster WG4 on Databases to continue to develop such databases. For example, the caNanoLab data is now available in the search interface at http://search.data.enanomapper.net/, but when loaded in data.enanomapper.net too, then a link set can be exported and loaded into the BridgeDb API.

Therefore, current data sources include the main databases for gene, RNA, and protein databases (Ensembl, NCBI Gene, UniProt, etc.) and for metabolites (HMDB, ChEBI, PubChem Compound, CAS registry, and ChemSpider).

5. LINKED DATA GENERATION

In order to support a fully Linked Data cloud for nanosafety related knowledge, we have extended AMBIT to expose data in the warehouse in various semantic web RDF serialization formats. The RDF output takes advantage of the eNanoMapper ontology [5] and the ontologies it reuses.

5.1 STRUCTURE OF THE OUTPUT RDF

The structure follows approaches in the CHEMINF, NPO, and BAO ontologies. For example, a nanomaterial is encoded as (in the Turtle format):

```
substance:NFYS16-M12 a obo:CHEBI_59999 ;
    rdfs:label "NM-400" ;
    npo:has_part substance:NFYS16-M12_core ;
    obo:BFO_0000056 mgroup:NWKI-a017c3a0-a ;
    dcterms:source substance:NFYS16 ;
    dcterms:type enm:ENM_9000081 .

substance:NFYS16-M12_core
    a npo:NPO_1597 ;
    af:CAS+Registry+Number "1314-13-2" ;
    af:ECEINECS "215-222-5" ;
    af:NamesIUPAC+name "ZnO" ;
    sso:CHEMINF_000200 substance:NFYS16-M12_core_smiles .

substance:NFYS16-M12_core_smiles
    a sso:CHEMINF_000018 ;
    sso:SIO_000300 "O=[Zn]" .
```

An example of a physical chemical property:

```
as:NWKI-a017c3a0-a
    a npo:NPO_1694 , bao:BAO_0000015 ;
    dc:title "Primary Particle Size" ;
    bao:BAO_0000209 mgroup:NWKI-a017c3a0-a ;
    bao:BAO_0002846 ap:00000000 .

mgroup:NWKI-a017c3a0-a
```

```
    a bao:BAO_0000040 ;
    obo:OBI_0000299 ep:IDca109b2d .

ep:IDca109b2d
    a bao:BAO_0000179 ;
    rdfs:label "PARTICLE SIZE" ;
    obo:IAO_0000136 substance:NFYS16-M12 ;
    obo:STATO_0000035 "70.0-90.0"^^xsd:string ;
    sso:has-unit "nm" .
```

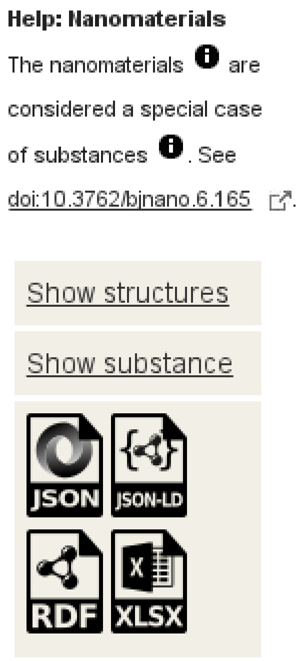
A full description is currently not yet available, but guidance documentation is being prepared.

5.2 Implementation

The implementation is written as a Java class, called the *SubstanceRDFReporter*, and is part of the AMBIT software, discussed earlier.

5.3 Integration in the AMBIT web interface

RDF can be downloaded from the data.enanomapper.net instance using the download options. A format icon for RDF is provided. Furthermore, the RDF is also available at an API level. Visitors of the database can click the icon (screenshot to the right) to download the data in a linked data format.



Help: Nanomaterials

The nanomaterials ❶ are considered a special case of substances ❶. See doi:10.3762/bjnano.6.165 .

6. SPARQL END POINT

A SPARQL query interface for eNanoMapper data and ontologies has been set up at https://sparql.enanomapper.net/. Recently, due to technical reasons, it was necessary to implement a completely new setup of the SPARQL endpoint infrastructure (see earlier deliverables). Performance reasons and the end of support for the 4store project, required us to migrate from the RDF triple-store 4store to a Virtuoso[1] backend. To get the best integration in our existing application framework and network environment we customized a Docker Virtuoso image to provide an out-of-the box triple-store solution, public available at https://hub.docker.com/r/insilicotox/ist-enm-virtuoso/. Dockerized applications simplify installation and eliminate problems caused by third party library dependencies. Virtuoso has full SPARQL (https://www.w3.org/TR/sparql11-query/) language support. Unlike with the 4store triple-store, subqueries are fully supported in Virtuoso. Also an increased variety of output formats are now supported (see Annex C).

To sustain service and data security we developed a ruby-wrapper, handling the requests from and to the triple-store. This wrapper is also part of the updated eNanoMapper ontology viewer application (fully described in deliverable D5.7).

6.1 Example Queries

This section lists a few example SPARQL queries that implement some data aggregation to support a typical nanosafety research-related task.

A *curl* example for retrieving ontology endpoint entries in RDF XML format:

```
curl -vG -H "Accept:application/rdf+xml" --data-urlencode \\
        query@query.rq https://sparql.enanomapper.net
```

Where the SPARQL resides in a file with the name *query.rq* and this content:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?root_name ?parent_name ?child_name
WHERE {
```

[1] Virtuoso SPARQL Query Service: http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSSparqlProtocol

```
   VALUES ?root_name {'endpoint'}
   ?root rdfs:label ?root_name .
   ?child rdfs:subClassOf{,5} ?root .
   ?child rdfs:subClassOf ?parent .
   ?child rdfs:label ?child_name .
   ?parent rdfs:label ?parent_name .
}
```

6.1.1 Research Questions

We previously outlined a short list of research questions. Using the Linked Data and the RDF in the SPARQL endpoint, we can define queries that aggregate data to support answering the research questions. Two research questions have been selected.

6.1.1.1 Which metal oxide nanoparticles are genotoxic?

One research question is supported by a query that lists all genotoxicity information for all metal oxides. This is conveniently supported by the eNanoMapper ontology. A SPARQL query can take advantage of the hierarchical organization in the ontology, which classifies specific metal oxides, such as titanium dioxide, as a “metal oxide” (NPO_1541). Similarly, the eNanoMapper ontology classifies the end points, and the below query will find biological data annotated with both “genotoxicity assay” (BAO_0002167) and “DNA damage assay” (ENM_9000017). Thus, aggregating all information available to answer the question stated above. Such questions are highly relevant to the risk assessment community, where tools for assessing the risks associated with particular nanomaterials are posed.

```
prefix bao:   <http://www.bioassayontology.org/bao#>
prefix obo:   <http://purl.obolibrary.org/obo/>
prefix sso:   <http://semanticscience.org/resource/>
prefix dcterm: <http://purl.org/dc/terms/>
prefix npo: <http://purl.bioontology.org/ontology/npo#>

SELECT DISTINCT ?title ?protocolTitle ?substanceLabel
                ?typeLabel ?value ?unit WHERE {
  { ?assay a [ rdfs:subClassOf+ bao:BAO_0002167 ] . }
  UNION
  { ?assay a  bao:BAO_0002167 . }
  ?assay dc:title ?title ;
    bao:BAO_0000209 ?mgroup ;
    bao:BAO_0002846 ?protocol .
  ?protocol dc:title ?protocolTitle .
  ?mgroup obo:OBI_0000299 ?endpoint .
  ?endpoint obo:IAO_0000136 ?substance ;
            sso:has-unit ?unit ;
            sso:has-value ?value .
  ?substance rdfs:label ?substanceLabel ;
             dcterm:type ?type .
  OPTIONAL { ?type rdfs:label ?typeLabel }
  { ?substance dcterm:type npo:NPO_1541 }
  UNION
  { ?substance dcterm:type [ rdfs:subClassOf+ npo:NPO_1541 ] }
}
```

6.1.1.2 All carbon nanotubes with a zeta potential < 0.0 mV

A second research question is answered by aggregating a list of all carbon nanotubes with a negative zeta potential. Here too, we take advantage of ontological hierarchies and query for "carbon nanotube" (NPO_606) which will also find "single", "double", and "multi-wall" subtypes and even "JRCNM04000a" (ENM_9000080), all subclasses in the ontology.

```
prefix bao:    <http://www.bioassayontology.org/bao#>
prefix obo:    <http://purl.obolibrary.org/obo/>
prefix sso:    <http://semanticscience.org/resource/>
prefix dcterm: <http://purl.org/dc/terms/>
prefix npo:    <http://purl.bioontology.org/ontology/npo#>
prefix enm:    <http://purl.enanomapper.net/>

SELECT DISTINCT ?substanceLabel ?typeLabel ?endpoint
  ?value ?unit ?condLabel ?condvalue ?condunit
WHERE {
  ?endpoint obo:IAO_0000136 ?substance ;
            sso:has-unit ?unit ;
            sso:has-value ?value ;
            rdfs:label ?epLabel .
  OPTIONAL {
    ?endpoint enm:has-condition ?condition .
    ?condition rdfs:label ?condLabel .
    OPTIONAL { ?condition sso:has-unit ?condunit }
    OPTIONAL { ?condition sso:has-value ?condvalue }
  }
  ?substance rdfs:label ?substanceLabel ;
             dcterm:type ?type .
  ?type rdfs:label ?typeLabel .
  BIND("ZETA POTENTIAL" AS ?epLabel)
  { ?substance dcterm:type npo:NPO_606 }
  UNION
  { ?substance dcterm:type [ rdfs:subClassOf+ npo:NPO_606 ] }
  FILTER (?value < 0)
} ORDER BY ASC(?substance)
```

6.1.2 Conformance testing

Conformance testing can be performed in order to see if data are sufficiently complete and that they can be considered to *conform* some standard, i.e. meeting the requirements of that standard. For example, we can test if data are complete [6]. However, practically, data can be represented in many different ways. Size information of the primary particle can be a single average size, a size range, or even just an upper size limit. Using a relational database approach, this can be hard to implement, but with a semantic approach this can be simplified to just asking if the required type of information is present, and take advantage of the ontological hierarchy. Thus, it is unnecessary to test for all kinds of particle size information, but just query for any kind of primary particle size information. Additionally, with the data linked to the eNanoMapper ontology, this becomes trivial. This use case is described in detail in D5.6. The approach basically comes down to formulating each completeness requirement as a SPARQL query, and if no results are returned for some requirement query, it means that the requirements are not met. D5.6 further outlines how this can then be used to calculate a completeness score of the type as, for example, used by the NanoMaterial Registry.

6.2 RDFIO tools for collaborative RDF editing via Semantic MediaWiki

As a collaborative development, user community researchers at the Uppsala University, Sweden, created a suite of tools with the name RDFIO. RDFIO is capable of importing of RDF data into Semantic MediaWiki, and exporting it again in the same RDF format (the data being expressed in the same ontology). In a manuscript under preparation it was shown how the developed functionality enables a number of usage scenarios where the interoperability of SMW and the wider Semantic Web is leveraged. The enabled usage scenarios include the following; *i)* Bootstrapping a non-trivial wiki structure from existing RDF data, *ii)* Round-tripping of RDF into and out of SMW, for community collaboration of the data while in SMW, and *iii)* Creating mash-ups of existing, automatically imported data and manually created presentations of this data. The functionality of the RDFIO is relatively simple to execute: the user can make use of “RDF import forms” (paste RDF code) or alternatively can run any SPARQL CONSTRUCT query and import results into an RDFIO instance for browsing and editing. The RDFIO was used in this way (use case iii) to edit eNanoMapper database data included in the ongoing NanoWiki Semantic MediaWiki project (https://figshare.com/articles/NanoWiki_4/4141593). The RDFIO project is further described in https://www.mediawiki.org/wiki/Extension:RDFIO. Code for the project is being made available at: https://github.com/rdfio/RDFIO.

7. CONCLUSION

This deliverable discusses the principles of Linked Data and how these approaches are used in the eNanoMapper project to answer two key use cases: answering scientific questions that require data from two or more sets of data or ontology sources; and, testing for conformance. The latter is reported in detail in D5.6 report, taking advantage of approaches described in this deliverable. The results demonstrate how an approach using the semantic query language SPARQL offers the flexibility needed to answer a range of questions taking advantage of the hierarchy of the ontology terms used in the data expressed in the RDF format.

It should also be noted that this work will continue in the next months in the tasks leading to D3.4, together with the NanoSafety Cluster WG4, and together with other NanoSafety Cluster projects, such as caLIBRAte. The latter has shown interest in reusing the approaches outlined for data completeness testing.

8. BIBLIOGRAPHY

1. Samwald, M., Jentzsch, A., Bouton, C., Kallesoe, C., Willighagen, E., Hajagos, J., Marshall, M., Prud'hommeaux, E., Hassanzadeh, O., Pichler, E., Stephens, S., May 2011. Linked open drug data for pharmaceutical research and development. Journal of Cheminformatics 3 (1), 19+. URL http://dx.doi.org/10.1186/1758-2946-3-19
2. Graves, M., Constabaris, A., Brickley, D., Apr. 2007. FOAF: Connecting people on the semantic web. Cataloging & Classification Quarterly 43 (3-4), 191-202. URL http://dx.doi.org/10.1300/j104v43n03_10
3. Willighagen, E. L., Waagmeester, A., Spjuth, O., Ansell, P., Williams, A. J., Tkachenko, V., Hastings, J., Chen, B., Wild, D. J., May 2013. The ChEMBL database as linked open data. Journal of Cheminformatics 5 (1), 23+. URL http://dx.doi.org/10.1186/1758-2946-5-23
4. Van Iersel, M. P., Pico, A. R., Kelder, T., Gao, J., Ho, I., Hanspers, K., Conklin, B. R., Evelo, C. T., Jan. 2010. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. BMC bioinformatics 11 (1), 5+. URL http://dx.doi.org/10.1186/1471-2105-11-5
5. Hastings, J., Jeliazkova, N., Owen, G., Tsiliki, G., Munteanu, C. R., Steinbeck, C., Willighagen, E., Mar. 2015. eNanoMapper: harnessing ontologies to enable data integration for nanomaterial risk assessment. Journal of Biomedical Semantics 6 (1), 10+. URL http://dx.doi.org/10.1186/s13326-015-0005-5
6. Marchese Robinson, R. L., Lynch, I., Peijnenburg, W., Rumble, J., Klaessig, F., Marquardt, C., Rauscher, H., Puzyn, T., Purian, R., Åberg, C., Karcher, S., Vriens, H., Hoet, P., Hoover, M. D., Hendren, C. O., Harper, S. L., April 2016. How should the completeness and quality of curated nanomaterial data be evaluated? Nanoscale 8 (19), 9919-9943. URL http://dx.doi.org/10.1039/c5nr08944a

ANNEXES

ANNEX A: PAPER: INTEGRATION AMONG DATABASES AND DATA SETS TO SUPPORT PRODUCTIVE NANOTECHNOLOGY: CHALLENGES AND RECOMMENDATIONS

See below

ANNEX B: LINK SETS

Sigma Aldrich

```
<http://localhost/ambit2/substance/NWKI-4f1a3729-4a88-31bc-abca-7eb6503a73d4>   foaf:page
        <http://www.sigmaaldrich.com/catalog/product/aldrich/420816>
<http://localhost/ambit2/substance/NWKI-ed06c39e-c7f2-3173-bd74-35b8257544b5>   foaf:page
        <http://www.sigmaaldrich.com/catalog/product/aldrich/420840>
<http://localhost/ambit2/substance/NWKI-c93a144a-cd6d-3fe0-ab86-13adb168ccb1>   foaf:page
        <http://www.sigmaaldrich.com/catalog/product/aldrich/421553>
<http://localhost/ambit2/substance/NWKI-02dc913f-f977-3ae2-ad43-ae0ec2afd177>   foaf:page
        <http://www.sigmaaldrich.com/catalog/product/aldrich/572500>
<http://localhost/ambit2/substance/NWKI-67b71fb2-85c1-3130-bd4e-84da8bc05fc9>   foaf:page
        <http://www.sigmaaldrich.com/catalog/product/aldrich/637238>
<http://localhost/ambit2/substance/NWKI-ea48c37d-840a-30d2-9740-1198a7e2c3b6>  foaf:page
        <http://www.sigmaaldrich.com/catalog/product/aldrich/718483>
<http://localhost/ambit2/substance/NWKI-d6376fe0-5362-3f76-b7b1-75f9985417c1>   foaf:page
        <http://www.sigmaaldrich.com/catalog/product/aldrich/379646>
<http://localhost/ambit2/substance/NWKI-3a035bdb-def7-35c6-bb4c-6df4f10c45d5>   foaf:page
        <http://www.sigmaaldrich.com/catalog/product/aldrich/482951>
<http://localhost/ambit2/substance/NWKI-6e6a5fec-23c2-3fff-8522-a22689091196>   foaf:page
        <http://www.sigmaaldrich.com/catalog/product/aldrich/637246>
<http://localhost/ambit2/substance/NWKI-2af8ed34-029d-3644-8327-23a596cfae0d>   foaf:page
        <http://www.sigmaaldrich.com/catalog/product/aldrich/748161>
```

Crystallography Open Database

```
<http://localhost/ambit2/substance/NWKI-bb2f1a47-0260-3271-b9c3-edb850d15be9>  foaf:page
        <http://www.crystallography.net/cod/1517077.html>
```

```
<http://localhost/ambit2/substance/NWKI-dd085452-8fd2-3025-a64f-2784fdf17879>    foaf:page
        <http://www.crystallography.net/cod/1517080.html>
<http://localhost/ambit2/substance/NWKI-04962f10-34c8-3118-953e-d40d7244209c>  foaf:page
        <http://www.crystallography.net/cod/1518678.html>
<http://localhost/ambit2/substance/NWKI-95bb8173-3aad-3441-a50e-97e62401eadb> foaf:page
        <http://www.crystallography.net/cod/1518679.html>
<http://localhost/ambit2/substance/NWKI-7c51d17d-bb0a-34e8-8e5a-f1bc2dad935f>   foaf:page
        <http://www.crystallography.net/cod/1518680.html>
<http://localhost/ambit2/substance/NWKI-f90f7f33-091c-3d54-b313-213083c0e052>    foaf:page
        <http://www.crystallography.net/cod/1517078.html>
<http://localhost/ambit2/substance/NWKI-6ede15ea-8037-379f-933e-7b947e155c6a>  foaf:page
        <http://www.crystallography.net/cod/1517079.html>
<http://localhost/ambit2/substance/NWKI-2e056907-7ac9-3b13-a599-ee10f2548d78>  foaf:page
        <http://www.crystallography.net/cod/2100388.html>
```

ChEMBL

Annex C: SPARQL Endpoint Supported Output Formats

The new Virtuoso RDF triple store backend comes with a huge variety of output formats for SPARQL queries:

- "application/json"
- "application/ld+json"
- “application/microdata+json"
- "application/odata+json"
- "application/rdf+json"
- "application/rdf+xml"
- "application/sparql-results+json"
- "application/sparql-results+xml"
- "application/turtle"
- "application/vnd.ms-excel"
- “application/x-turtle"
- "text/csv"
- "text/html"
- "text/n3"
- "text/ntriples"
- "text/plain"
- "text/rdf+n3"
- "text/rdf+ttl"
- "text/rdf+turtle"
- "text/tab-separated-values"
- "text/turtle"