DELIVERABLE REPORT D2.2

Report on Ontology Content Types and Existing Community Efforts

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:	22.9.2014
VERSION:	V.1.1.

SEVENTH FRAMEWORK PROGRAMME

Call identifier	FP7-NMP-2013-SMALL-7
Document Type	Deliverable Report
WP/Task	2 - 2
Document ID	eNanoMapper D2.2.
Status	Submitted

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) • University of Maastricht (UM)
Authors	Janna Hastings, Gareth Owen, Egon Willighagen Reviewed by Barry Hardy (DC)
Purpose of the Document	To report on results of surveying the content types and existing ontology and community efforts that are required for eNanoMapper ontology development
Document History	1.First draft, 23/07/2014 2.Internal draft for review 31/07/2014 3.Internal draft for review (2), 22/09/2014 4.Completed version, 30/09/2014

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY 6
2. INTRODUCTION 7
3. CONTENT TYPES 8
3.1. NANOPARTICLE TYPES 8
3.2. ENM PHYSICO-CHEMICAL CHARACTERIZATION 9
3.3. BIOLOGICAL CHARACTERISATION 9
3.4. ENVIRONMENTAL CHARACTERISATION 10
3.5. EXPERIMENTAL MEASUREMENTS AND PROTOCOLS 10
3.6. NANOMATERIAL LIFECYCLE 11
3.7. KNOWN SAFETY INFORMATION 11
3.8. LITERATURE 11
4. EXISTING COMMUNITY EFFORTS 12
4.1 ONTOLOGIES 12
4.1.1. NPO *12*
4.1.2. ChEBI *13*
4.1.3. CHEMINF *14*
4.1.4. OBI and BAO *15*
4.1.5. Biological Ontologies *17*
4.1.6. ENVO *17*
4.1.7. InterNano Nano-Manufacturing Taxonomy *17*
4.2. STRUCTURED FILE FORMATS 18
4.2.1. OECD Harmonized Templates *18*
4.2.2. ISA-TAB NANO *18*
4.3. OVERLAPS 19
4.4. GAPS 19
5. COMMUNITY NEEDS 21
5.1 EFFORTS THAT NEED IRI-FICATION 21
5.2 SPECIFICATIONS 21
5. CONCLUSION 23
6. BIBLIOGRAPHY 24

TABLE OF FIGURES

Figure 1: An example of a nanoparticle type in ChEBI. 14
Figure 2: An extract of the CHEMINF property hierarchy. 15
Figure 3: An extract of the assay classification in OBI. 16
Figure 4: An extract of the assay classification in BAO. 16
Figure 5: A selection of imported biological entities in BAO. 17
Figure 6 A subset of the InterNano Nano-Manufacturing Taxonomy illustrated in BioPortal. 18

GLOSSARY

Abbreviation / acronym	Description
WP	Work Package
OWL	Web Ontology Language
MIREOT	Minimum Information to Reference an External Ontology Term
OBI	The Ontology for Biomedical Investigations
BAO	BioAssay Ontology
NPO	NanoParticle Ontology
ChEBI	Chemical Entities of Biological Interest
GO	Gene Ontology
ENVO	Environment Ontology
CHEMINF	Chemical Information Ontology
PRO	Protein Ontology
PR	Protein Ontology (alternative abbreviation)
CL	Cell Ontology
IRI	International Resource Identifier

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. Work Package 2 of this effort will develop and disseminate a comprehensive ontology for the nanosafety domain, encompassing nanomaterials and all information relating to their characterization, as well as relevant experimental paradigms, biological interactions and safety information. This deliverable report describes the primary content areas within the nanomaterial safety domain which need to be covered by the ontology and comprehensively surveys pre-existing ontology and vocabulary efforts for coverage of those content areas. Gaps and challenging content areas are identified.

2. INTRODUCTION

Nanomaterials are materials in which the units have at least one dimension sized in the 1-100nm 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 eNanoMapper project (www.enanomapper.net) is creating a pan-European computational infrastructure for toxicological data management for ENMs, based on semantic web standards and ontologies. eNanoMapper aims to develop a comprehensive ontology and annotated database for the nanosafety domain to address the challenge of supporting the unified annotation of nanomaterials and their relevant biological properties, experimental model systems (e.g. cell lines), conditions, protocols, and data about their environmental impact. Rather than starting afresh, the developing ontology will build on existing work, integrating existing ontologies in a flexible pipeline. The establishment of a universal standardisation schema and infrastructure for nanomaterials safety assessment is a key project goal, which will catalyze collaboration, integrated analysis, and discoveries from data organised within a knowledge-based framework. This framework will support the discovery of nanomaterial properties responsible for toxicity, and the identification of toxicity pathways and nano-bio interactions from linked datasets, ontologies, 'omics data and external data sources.

Ontologies are structured controlled vocabularies enhanced with explicit formal relationships between entities in support of advanced automated reasoning for inference and error detection. **Work Package 2** of the eNanoMapper project focuses on the development and dissemination of a comprehensive ontology for the nanosafety domain, encompassing nanomaterials and all information relating to their characterization, as well as relevant experimental paradigms, biological interactions and safety information.

To be sure that the ontology attains a good coverage of the entities that need to be annotated in the nanosafety domain, we have undertaken to perform a comprehensive survey of the content types that are involved, including detailing their structure and attributes. Furthermore, we have surveyed and evaluated pre-existing ontology and terminology efforts that are relevant to one or more of the identified content areas. Based on these evaluations we have gathered some notes about duplication across resources and about gap and challenge areas. This deliverable reports on the outcomes of these exercises.

3. CONTENT TYPES

The comprehensive suite of ontologies developed by eNanoMapper will cover the following broad content areas (as addressed by the use cases):

1. A categorisation of nanoparticle classes based on their properties, constituency and shape, including the use of axioms to achieve polyhierarchical classification.
2. Physicochemical properties for ENM characterisation.
3. A biological characterisation that describes the ENM-specific interactions with, for example, proteins to form a corona (see Section 3.5).
4. Environmental characterization
5. Experimental design and encoding for experiments in which nanosafety is assessed.
6. The full nanomaterial lifecycle including manufacturing and environmental decay or accumulation.
7. Known safety information about ENMs.
8. Primary literature (e.g. journal articles) and patents

3.1. NANOPARTICLE TYPES

Nanoparticles are classified primarily on the basis of their primary constituent, e.g. silica, carbon, titanium dioxide, gold or silver, nanoclay, etc., and their shape. Examples of nanoparticle classes already available in public ontologies (see Section 4 for the details of the ontologies) include 'chitosan nanoparticle' (NPO:261), 'spherical nanoparticle' (NPO:1551), 'gold nanoparticle' (NPO:401), 'core-shell silica nanoparticle' (NPO:1572) and 'citrate-coated silver nanoparticle' (CHEBI:82778). Classes of nanoparticle may also be functionally defined, for example, 'fluorescent silica nanoparticle' (NPO:1553) and 'long circulating nanoparticle' (NPO:1591).

Nanoparticles may be simple (e.g. nanodot with a particular composition) or complex, in that they may be composed of several layers and their surfaces may be heterogeneously functionalised with attached groups of any composition. The NPO includes a few general classes for these complex particle types including 'surface functionalized nanoparticle' (NPO:1881); however, more specific detailed classes of types of functionalization are not yet present. The molecular composition of the nanoparticle includes a specification of the constituent groups and atoms together with their bonding arrangement. Constituent groups are described in ChEBI. When describing the molecular composition of nanoparticles it may be necessary to distinguish the molecular composition of specific parts of the nanoparticle, e.g. the surface, core, linkage etc. The NPO includes classes for the different parts of the nanoparticle e.g. 'silica core' (NPO:1865).

Nanoparticles are also commonly described by their dimensionality and shape. 'Dimensionality' describes the number of dimensions of the particle that are within the 'nanoscale' (i.e. between 1 and 100 nm). Thus quantum dots, hollow spheres and free nanoparticles, in which all three dimensions are in the nanoscale, are described as three-dimensional (some sources use the term 'zero-dimensional'). Analogously, nanorods, nanotubes, nanowires and nanofibres, which have two dimensions in the nanoscale, are known as two-dimensional, while thin films or surface coatings, which have only one dimension in the nanoscale, are classed as one-dimensional.

Nanoparticles come in different shapes, providing another useful descriptor for classification purposes. Thus two-dimensional nanoparticles may occur as rods, helices, zig-zags, or belts, whilst three-dimensional nanoparticles may be conical, cylindrical, ellipsoidal, elliptical, polyhedral, spherical, etc.

Both the dimensionality and the shape of nanoparticles can be important factors in determining the toxicity of nanoparticles, their cellular uptake etc. In addition, nanoparticles have other relevant material or mechanical properties such as being soft or hard (stiff).

The eNanoMapper project is planning collaboration with the FutureNanoNeeds project which has been tasked with the development of systems of classification for novel nanomaterials which could be multi-component (while existing classification systems tend to assume homogeneity to a greater extent). There is a body of assumptions in the existing classification systems, e.g. that gold nanoparticles will be spherical, which are challenged by novel work. To this end, a workshop may be jointly organized between eNanoMapper and FutureNanoNeeds at a future date. For example, many of the special properties of nanoparticles can be altered or lost due to interactions between particles. Accordingly, nanocomposites (in which the active nanoparticles are separated from each other by being distributed in a second phase) are now widely used. Neither the NPO nor ChEBI currently contains descriptions of nanocomposites with their mediums. For such composite nanomaterials, the 'second phase' in which the nanoparticles are distributed may itself be classed as three-, two-, or one-dimensional.

3.2. ENM PHYSICO-CHEMICAL CHARACTERIZATION

The sorts of physicochemical properties that are used in the characterisation of nanomaterials include the state of dispersion, aggregation and agglomeration of the nanomaterial, the size (and size distribution) of the particle, the specific surface area and porosity, the surface composition and reactivity (a measure of the extent to which the surface atoms of the nanomaterial can induce the production of reactive oxygen species), and the purity (and impurities).

More specific physicochemical characterization measurements include:

- Solubility and dispersability in different media including water
- Zeta potential is especially important to predicting the aggregation and agglomeration behavior of particles, and may be measured over a pH range
- Diffusion and gravimetric deposition rates, which will affect the dispersal and exposure time for substances leaking into the environment.

The materials, rather than the particles, are described by the particle size distribution and the remainder of the medium in which the particles are contained. Impurities can play a crucial role in determining the safety or toxicity of nanomaterials, so should be quantified and described.

3.3. BIOLOGICAL CHARACTERISATION

Through interactions with biological systems, nanoparticles may become covered with biological material, particularly proteins in the blood, and lipids in the pulmonary system. This is referred to as the "corona" of the nanoparticle and it strongly depends on the exposure medium (e.g. bovine serum) and the duration of exposure. The corona may be composed of a monolayer or multiple layers, and the proteins may be denatured by their adsorption to the nanoparticle, creating entirely novel biomolecular entities with unknown reactivities.

The corona affects biodistribution and cellular uptake of the nanoparticle and may also cause some toxic effects. In short, exposure to any biological medium changes the external nature of the nanoparticle and thus its biological effects. It is therefore very important that all data points include metadata to describe the history of each sample and to control carefully for exposure to biological material.

Similarly to protein-protein interactions, nanoparticle-protein interactions are characterized by binding affinity, stoichiometry, and kinetic properties.

In addition to proteins, nanoparticles can also bind to and interact with DNA, or interact with whole cells. Cell association (i.e. binding and uptake to cells of a given type) is a measurement of special importance for its relevance to inflammatory responses, biodistribution, and toxicity *in vivo*.

3.4. ENVIRONMENTAL CHARACTERISATION

Nanoparticles may be released into the environment throughout their lifecycle, including their initial synthesis, incorporation into a product, use by consumers, and disposal, so subjecting workers, consumers and the environment to potential exposure. A variety of methods and measurements may be used in order to assess exposure, including the use of particle number, particle mass, and surface area detection devices.

When describing the environmental hazards and assessments of the environmental impact of nanomaterials, once they get 'into the wild', it will also be necessary to refer to a wide range of different ecosystems environment types and locations.

3.5. EXPERIMENTAL MEASUREMENTS AND PROTOCOLS

In addition to the various measurement *outcomes* (physicochemical and biological characterization and properties) discussed above, the measurement *techniques and tools* also need to be standardized in the eNanoMapper ontology. These include (Mu *et al.*, 2014):

- Transmission electron microscopy, scanning electron microscopy and atomic force microscopy provide information about the nanoparticle morphology.
- Crystallographic methods can be used to determine the shapes of particles.
- Dynamic light scattering (DLS) provides information on the hydrodynamic radii of nanoparticles in solution.
- Surface charge properties are determined with zeta-potential measurements.
- Chemical composition is revealed by auger electron spectroscopy, x-ray photoelectron spectroscopy, time-of-flight mass spectrometry and elemental analyses.
- Surface ligands and adsorbed molecules are identified with magic angle spinning nuclear magnetic resonance, liquid chromatography mass spectroscopy (LC-MS) and Fourier-transform infrared spectroscopy. Surface-enhanced Raman spectroscopy may also be used.
- Size exclusion or thin layer chromatography can identify nanoparticle-bound lipid molecules.
- Binding of surfactant molecules onto the surfaces of nanoparticles may alter their surface plasma resonance absorption and can also be determined using UV-vis absorption spectroscopy.
- Surface pressure-area isotherm measurements can be used to study the properties of lipid monolayers in the presence of nanoparticles.
- Differential scanning calorimetry (DSC) and isothermal titration calorimetry (ITC) can be used to measure thermodynamic changes in supported membranes or liposomes.
- Steady-state and time-resolved fluorescence spectroscopy are used to study nanoparticle-protein binding affinities, complex formation, and binding-induced protein conformational changes.
- Stepwise photobleaching has also been used to characterize nanoparticle-protein interactions.
- Proteins bound to a nanoparticle surface may be identified by 2D polyacrylamide gel electrophoresis.
- The adsorption and desorption processes of nanoparticle-DNA complexes can be measured using cyclic voltammetry.
- Cellular uptake can be monitored using X-ray fluorescence microscopy to determine the chemical element distribution of nanoparticles in cells.

- Magneto-photoacoustic imaging can be used to differentiate membrane-adhered from endocytosed nanoparticles in a cell.
- Atomic force microscopy measures the force between nanoparticles and the cell surface in cell association.

The details of the experimental methods that may be captured in the ontologies may include links to instruments used in the measurements (see existing implementation in NPO discussed below). Furthermore, the explicit and detailed protocols for the above measurements should ideally be captured in formalized ontology annotations in just as much detail as they are described in the experimental methods sections of high-quality publications, i.e. in sufficient detail to allow reproducibility. The ontology, however, can only supply a formalization of the vocabulary to be used in such descriptions. The enforcement of the minimum level of detail required when annotating data of a given type (about a particular experiment) needs to be done via alternative methods combined with the use of the ontology as knowledgebase and vocabulary. For example, experimental templates can suggest which fields need to be filled for various different types of experiment, and Minimum Information guidelines could be created to use as checklists for automatic quality-checking of data. (See the OECD Harmonized Templates as discussed below.)

3.6. NANOMATERIAL LIFECYCLE

The full 'cradle to grave' nanomaterial lifecycle from synthesis through use to recycling, disintegration or environmental accumulation needs to be described in the ontology.

While coverage of this aspect of nanomaterials is poor in existing ontologies, the NPO contains a few relevant classes, including 'biodegradable nanoparticle' (NPO:836) as a class of nanoparticle type. The definition in NPO refers to the property (quality) of being 'biodegradable' (NPO:191). NPO also contains a small number of manufacturing-relevant classes under 'material synthesis technique' (NPO:1921).

3.7. KNOWN SAFETY INFORMATION

The ontology should support the rapid retrieval of relevant safety information given a particular class of ENMs and a particular biological context. While the safety data itself will be included in the eNanoMapper database (not the ontology), the ontology needs to include classes for different types of toxicological endpoint as well as the experiments that are conducted to evaluate and assess toxicity in different systems.

Regulatory language used to describe safety hazard classes is also required in the context of the ontology (to enable organizing and searching the known information from the literature).

3.8. LITERATURE

Currently, the primary literature is a major source of information and several NanoSafety Cluster projects have ongoing efforts to extract information from literature, both manually and computationally (text mining). Capturing the provenance (type, title, author, date of creation, source, etc.) of this literature is needed and has several possible use cases in the community.

Patent resources are also an important source of information on novel nanomaterials in development. We will extend our provenance information to also annotate patent sources, and harness patent content mining systems such as SureChembl (https://www.surechembl.org/search/) where possible.

4. EXISTING COMMUNITY EFFORTS

4.1 ONTOLOGIES

Rather than 're-inventing the wheel' and thus causing further fragmentation of data annotation, the eNanoMapper project re-uses existing ontologies and vocabularies that have been created for ENMs.

The following external ontologies have been identified as already in part covering the nanosafety-relevant content areas outlined above:

1. The NanoParticle Ontology (NPO, Thomas et al., 2011) is the most comprehensive ontology for nanomaterials that has been created to date. It includes a classification of nanomaterial types based on particle composition, properties and shape. It also includes classes for properties relevant for describing nanomaterials and for experiments used to characterize nanomaterials. It is available at http://bioportal.bioontology.org/ontologies/NPO.
2. ChEBI, the ontology of Chemical Entities of Biological Interest (Hastings et al., 2013) includes the molecular groups and chemical classes that are needed to describe the chemical composition of nanomaterials. ChEBI also contains a small nanomaterial classification in its 'chemical substance' branch, not fully overlapping with NPO. ChEBI can be found at http://www.ebi.ac.uk/chebi/.
3. The chemical information ontology (CHEMINF) includes chemical qualities and descriptors, both calculated and measured (Hastings et al., 2011). This ontology is already the standard for chemical property representation in Open PHACTS (Williams et al., 2012), and nanomaterial-relevant descriptors will be added to it -- http://code.google.com/p/semanticchemistry/.
4. For experiments assessing the safety of nanomaterials, the Ontology for Biomedical Investigations (OBI, Brinkman et al., 2010) available at http://obi-ontology.org/, and the BioAssay Ontology (BAO, Vempati et al., 2012) available at http://bioassayontology.org/ may both be relevant, although nanomaterial-specific content is sparse, they do include some classes not already present in NPO.
5. For biological characterization, several ontologies are relevant: the Gene Ontology (GO, GO Consortium, 2000) available at http://amigo.geneontology.org/, the Protein Ontology (PRO, Natale et al., 2011) available at http://pir.georgetown.edu/pro/pro.shtml, the Cell Ontology (CL, available at https://code.google.com/p/cell-ontology/) and others, including anatomy and medical classifications systems.
6. For environmental characterization, the Environment Ontology (ENVO, Buttigieg et al., 2013) is relevant. Available at http://purl.bioontology.org/ontology/ENVO.
7. For nanomaterial manufacturing, the InterNano Nano-Manufacturing Taxonomy provides vocabulary, using numeric identifiers. It is available at http://purl.bioontology.org/ontology/InterNano. However, it does not contain any definitions, nor relationships other than subsumption (is_a).

4.1.1. NPO

The NPO was created out of the need to standardize data description in cancer nanotechnology research and enable searching and integration of diverse experimental reports. The NPO uses the Basic Formal Ontology as upper level and is developed in the Web Ontology Language (OWL). It refers to multiple external ontologies including ChEBI. As of the last release (2011-12-08), the NPO contains 1904 classes.

It covers the following content areas:

- Experiment types of relevance in nanomaterial characterization, e.g. ‘dynamic light scattering’;
- Experimental methods for synthesis, e.g. ‘solvent displacement method’;
- Type of chemical components of a nanoparticle formulation which include the nanoparticle, active chemical constituents of the nanoparticle, and functionalizing components;
- Molecular structure of the chemical components (e.g., atom, element, compound, liposome, micelle, etc.) – in some cases imported from ChEBI;
- Biochemical role or function of these chemical components (e.g., anticancer drug, surface modifying agent, MRI contrast agent, spacer, etc.);
- Type of nanoparticle based on its structure, function or chemical composition (e.g.,quantum dot, solid lipid nanoparticle, iron oxide nanoparticle, biodegradable nanoparticle, nanotube, gold nanocantilever, etc.);
- Chemical linkages between chemical components (e.g., amide linkage, disulfide linkage, encapsulation, etc.);
- Physical locations of chemical components within a nanoparticle (e.g., core, surface, etc.);
- Nanoparticle shape (e.g. spherical, cylindrical, etc.);
- Physical state of a formulation (e.g., emulsion, hydrogel, etc.);
- Physical, chemical, or functional properties of chemical constituents and functionalizing agents (e.g., organic, hydrophilic, magnetic, etc.);
- Applications in cancer diagnosis, therapy, and treatment (e.g., chemotherapy, diagnostic imaging, detection of cancer cells, etc.);
- Underlying mechanisms guiding the design for the formulation (e.g., endocytosis, active targeting, etc.);
- Type of stimulus (e.g., magnetic field, ultrasound, pH change, etc.) for activating the function of nanoparticles , and the response to that stimulus (e.g., drug release from nanoparticle in response to magnetic field, heat generation from nanoparticle in response to infrared light, etc.).
-

The NPO contains a sophisticated axiomatization for some parts of the ontology, embedding detailed domain knowledge of relevance for nanomaterial research into logical definitions captured in the ontology. This is described in detail in (Thomas et al., 2011).

For example:

- the class ‘intracellular fluid in a tumor cell’ is defined as “'intracellular fluid' and (contained_in some 'intracellular space of tumor cell') and (has_quality some 'tumor intracellular pH')”;
- the class ‘carbohydrate-coated nanoparticle’ is defined as “nanoparticle and (has_component_part some (carbohydrate and (has_role some 'nanoparticle surface modifying role')))”;
- the class ‘fluorescence imaging contrast agent’ is defined as “'optical imaging contrast agent' and (has_application some 'fluorescence imaging') and (has_property some fluorescent)”
- the class ‘elimination rate constant’ is linked to the class ‘pharmacokinetics study’ using the relationship ‘parameter determined from’.

4.1.2. CHEBI

ChEBI is both a chemical database and a chemical ontology. It offers a wide range of useful chemical information including chemical structures and properties, citations to the literature and both a structure-based and a role (activity)-based ontology classification. As of the last release (1 September 2014) there are 40,536 classes in ChEBI.

The bulk of the chemicals annotated in ChEBI are organic molecular entities of biological interest as metabolites or agents that can intervene in biological processes. The most relevant for the eNanoMapper project are the functional groups and atoms which are used to describe the composition and functionalization of nanoparticles. However, in addition, ChEBI does contain a small 'chemical substance' branch with a small nanoparticle classification, primarily around types of nanoparticle based on their chemical constitution, some of which do not also appear in the NPO (thus are not superfluous).

For example, ChEBI contains 'palladium-gold nanoparticle' (CHEBI:52523) which is defined as a gold nanoparticle covered with a thin coat of palladium atoms. Figure 1 shows the structural classification of this entity within the ChEBI ontology, showing the parent classes including 'nanoparticle' and 'polyatomic entity'.

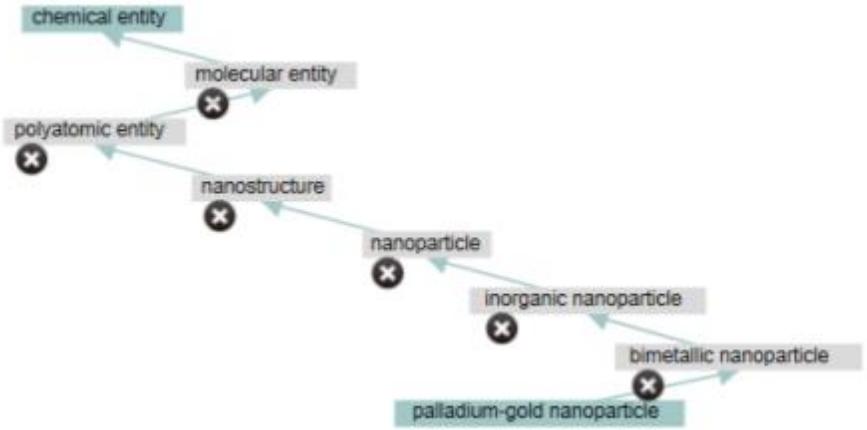

Figure 1: An example of a nanoparticle type in ChEBI.

While ChEBI is not developed natively in OWL, it is available in an OBO and an OWL exported format.

4.1.3. CHEMINF

CHEMINF is an ontology of chemical information entities – descriptors and other chemically relevant data items, designed to support data sharing and standardization of cheminformatics data in the context of the Semantic Web. It is developed natively in the OWL language and stored on the publicly available repository server in Google Code (http://code.google.com/p/semanticchemistry/). As of the last release (14 August 2014), the ontology contains 723 classes.

Many cheminformatics descriptors, such as molecular mass, pKa and so on, are straightforwardly relevant also to data about nanomaterials. Figure 2 shows an extract of the CHEMINF property hierarchy in the Protégé ontology editing tool.

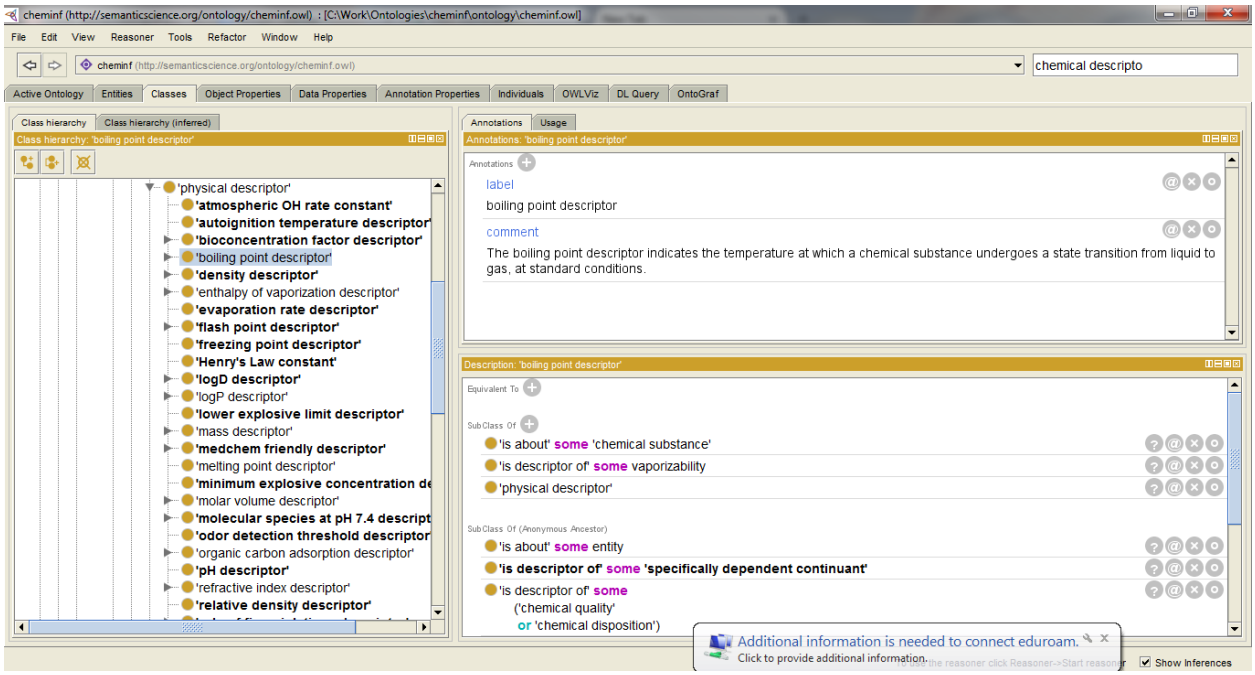

Figure 2: An extract of the CHEMINF property hierarchy.

4.1.4. OBI AND BAO

There are two ontologies in the biomedical domain that include assay descriptions: OBI and BAO. While OBI is broad in scope and has the involvement of many different communities in its development, BAO has arisen from a more specific targeted need in chemical biology data annotation. Both are developed in OWL. OBI is located at http://obi-ontology.org/ and BAO at http://bioassayontology.org/. OBI has 2,797 classes as of the last release (2014-08-18) and BAO has 3,337 (2014-04-25).

Figure 3 shows an extract of the assay classification in OBI.

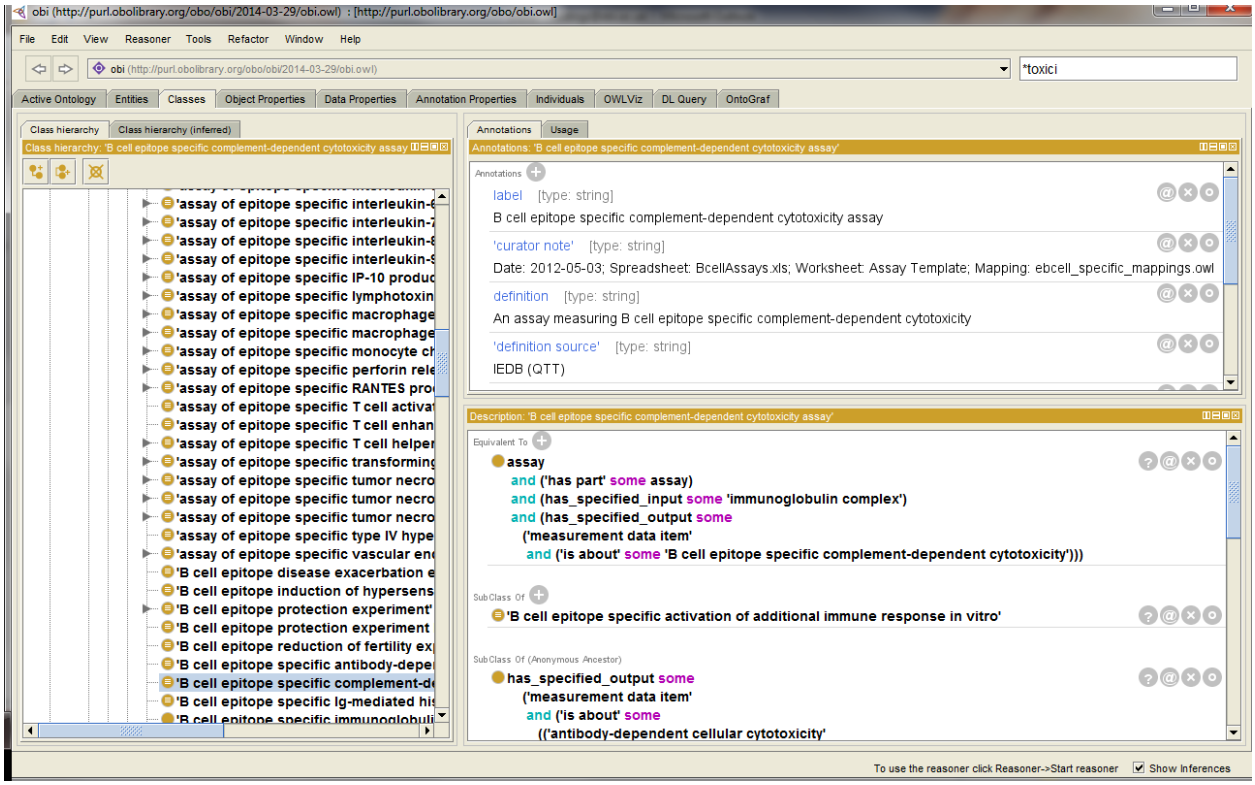

Figure 3: An extract of the assay classification in OBI.

Figure 4 shows an extract of the assay classification in BAO.

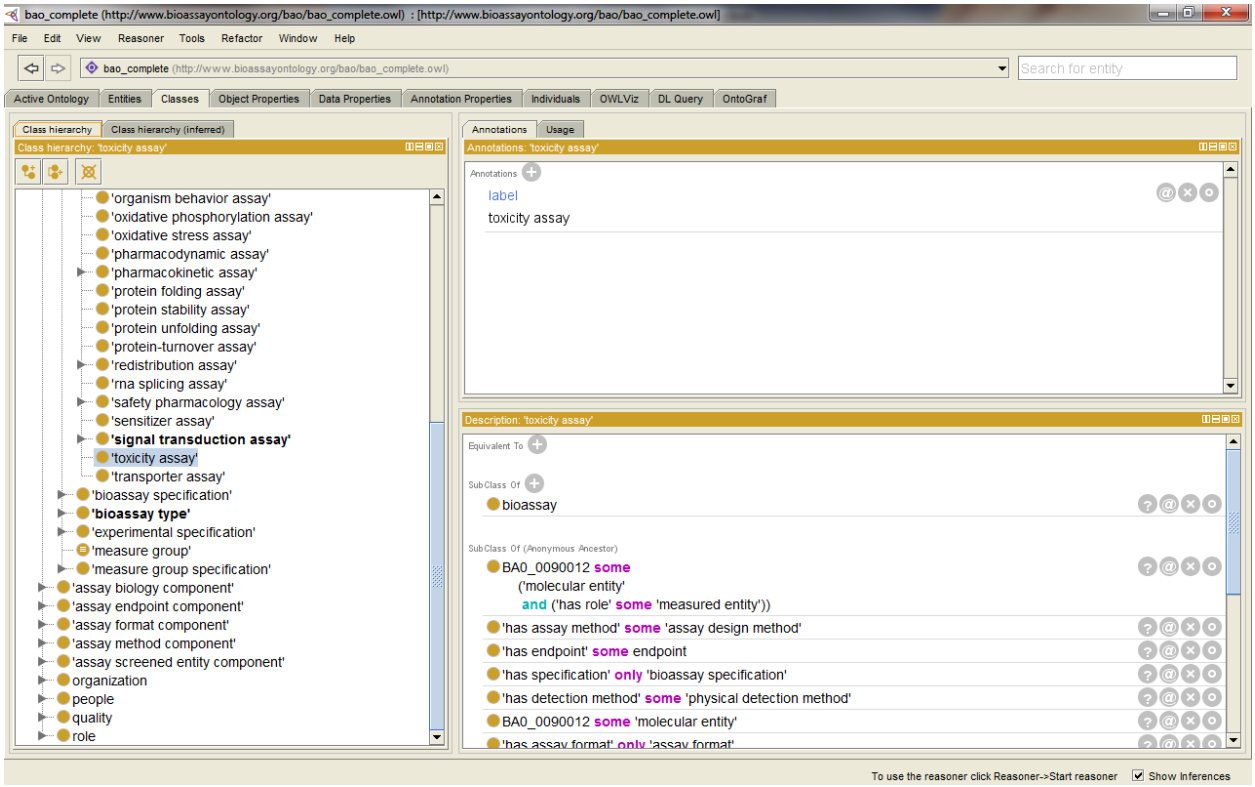

Figure 4: An extract of the assay classification in BAO.

As can be seen in Figures 3 and 4, OBI is much more metadata-rich than BAO, while BAO has a cleaner classification hierarchy. OBI also has a wider diversity of assays represented. However neither OBI nor BAO have any nanomaterial-specific content or assay types.

4.1.5. BIOLOGICAL ONTOLOGIES

Foremost among the existing biological ontologies is the GO, the widely used ontologies for biological processes, molecular functions and cellular components used in gene product annotation. GO is the gold standard for annotation in these three sub-domain areas. It contains 41,694 classes.

For annotation of proteins, PRO may be used. With 83,656 classes, it contains species-neutral grouping classes for all proteins identified in UniProt as well as links to the species-specific UniProt entries.

For annotation of cellular entities, the Cell (CL) ontology may be used (5,901 classes).

Importing biological entities from external ontologies is common practice in assay-related ontologies. For example, BAO has a good selection of imported biological entities for use in annotation of assays against biological endpoints, as illustrated in Figure 5.

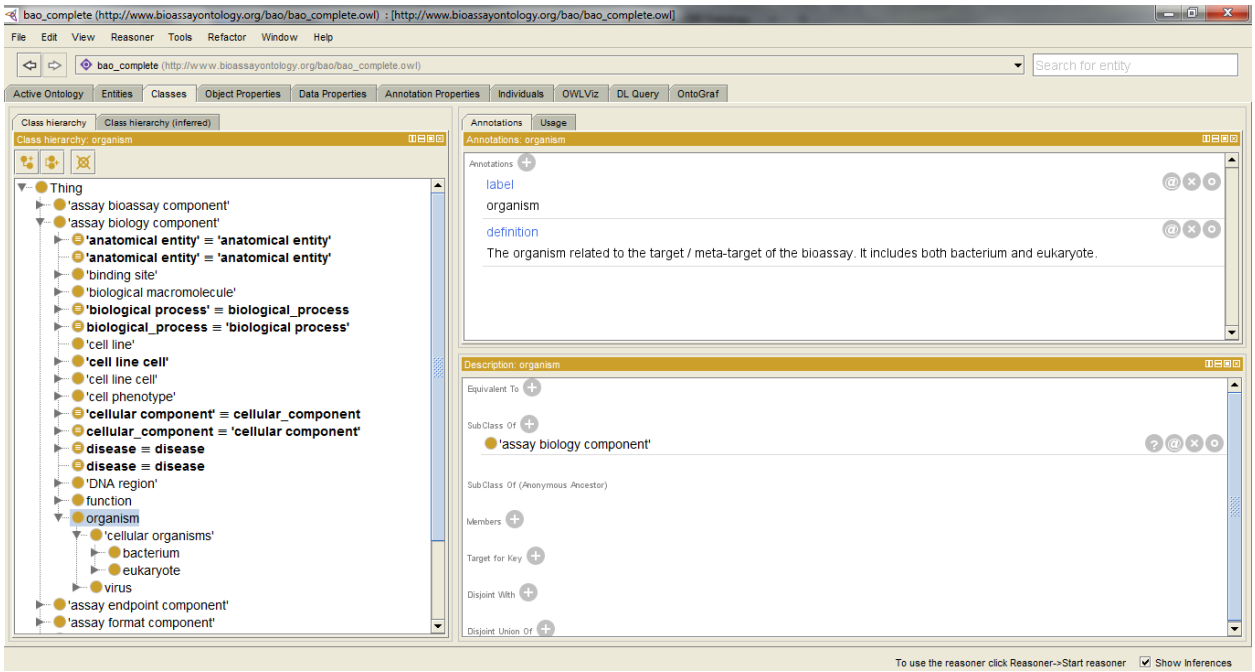

Figure 5: A selection of imported biological entities in BAO.

4.1.6. ENVO

ENVO contains terminology covering a wide range of environments including, for example, marine zones, tidal zones, soil and so on (1,688 classes). It also contains biomes such as desert and grassland; environmental features such as archaeological sites, caves and beaches; and environmental conditions such as arid and subtropical.

4.1.7. INTERNANO NANO-MANUFACTURING TAXONOMY

The Nano-Manufacturing taxonomy arising from the InterNano project is available in the OWL language in BioPortal, as illustrated in Figure 6. It includes branches such as application areas for nanotechnology, health and safety, nanomanufacturing and characterization processes, a classification of nanomaterial

types, and social and economic impacts. The taxonomy is used for intelligent searching and to organize content on the internano.org website.

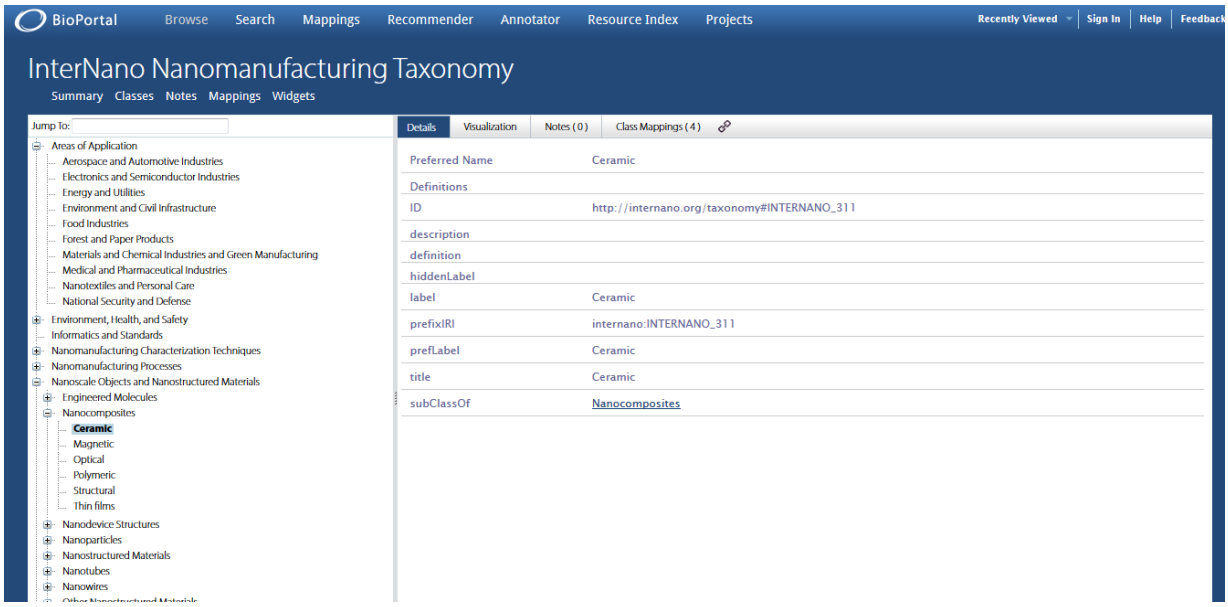

Figure 6 A subset of the InterNano Nano-Manufacturing Taxonomy illustrated in BioPortal.

While a very useful resource in terms of terminology, the taxonomy does not include much additional metadata, such as synonyms or definitions. It also does not include further ontological content such as additional relationships.

4.2. STRUCTURED FILE FORMATS

4.2.1. OECD HARMONIZED TEMPLATES

The OECD Harmonized Templates are structured (XML) data formats for reporting safety-related chemical studies. They contain vocabularies in the form of picklists for some of the specified fields, and documented guidance material. Nanomaterials are considered as a special type of chemical and as such are represented in several of the templates and associated picklists. There are also several templates that are specific to nanomaterial assessment over and above the templates that apply to any chemicals. These fall under the header “additional physico-chemical properties of nanomaterials” and include agglomeration/aggregation, crystalline phase, aspect ratio/shape, dustiness, porosity, pour density and radical formation potential. The full set of OECD templates is available from http://www.oecd.org/ehs/templates/templates.htm.

4.2.2. ISA-TAB NANO

ISA-TAB is a commonly used format for representing experimental data in structured tab-separated files. There are three file types: investigation, study and assay. The investigation file contains the broad reference information about the project in the context of which the biological experiment has been performed, such as the point of contact and publications, and includes reference details for the other files. The study file includes all the information about the sample being tested, and the investigation file records the raw data of the assay (or specifies the files for the raw data for non-spreadsheet, e.g. image data). The ISA-TAB Nano specification enhances the “pure” ISA-TAB specification with support, in the form of an additional file type “material”, for describing nanomaterials. While the study file enables

description of samples of biological origin, the material file enables description of samples of non-biological origin, whether nanoscale or not. It enables the description of complex nanomaterial formulations including chemical components, functionalizing agents, and medium of suspension.

4.3. OVERLAPS

Many of the ontologies that have been identified for incorporation into the eNanoMapper ontology suite partially overlap in their content. When overlapping ontologies are used in a vocabulary suggestion tool, it might result in the user being shown duplicate suggestions, and even if the tool is sophisticated enough to filter the duplicates out of view, there is a risk of conflicting definitions and content distribution across different ID spaces. For example, based on exact label matching only, the overlap between the ChEBI ontology (which itself has 38,735 classes as of April 2014) and the NPO (1,903 classes) is 395. This is a small but nevertheless significant number of exactly shared labels. Most of these are groups, atoms or chemical classes that are included in NPO so as to support description of nanomaterial characterization. Some, but not all, of these are cross-referenced to ChEBI via an additional annotation 'dbXref' in the NPO OWL file. Other overlapping classes derive from the fledgling nanoparticle classification that is included in ChEBI. For this branch of NPO, there are no cross-references annotated to ChEBI (and neither does ChEBI annotate cross-references to NPO). Some of the overlap arises from drug classes that are included in the NPO, e.g. thalidomide and tamoxifen, assumedly because the NPO was designed for cancer nanotechnology research and these are cancer drugs.

The OBO Foundry (Smith et al., 2007) recommends collaboration to resolve overlap between neighbouring ontologies in situations such as these. A strategy that suggests always favouring one ontology over another is not possible, since for groups and chemical classes the ChEBI IDs are preferred, while for nanoparticle classes it is the NPO IDs.

For another example, between BAO and NPO there are 37 overlapping labels. These include abstract classes such as 'physical quality', 'shape', 'size'; and role classes such as 'solvent', 'dihydrofolate reductase inhibitor' and 'fluorochrome'. Note that label sharing in itself isn't a problem unless the IDs are different. If the MIREOT strategy is followed (Courtot et al., 2011), the IDs and definitions will be exactly the same, which presents no problem for data annotation. This is the case for the bulk of the overlap between BAO and ChEBI, which with 696 shared labels would otherwise be very challenging.

The strategy for managing duplication in the combined ontology resource is to systematically prune (i.e. remove) duplicated content as part of the import process for ontologies that are re-used. In each case, a primary provider for the content domain is selected. For example, for nanoparticle classification it is the NPO, for small molecules such as drugs it is ChEBI, for biological assays it is OBI etc. Duplicated content in other ontologies that are imported are then removed by the custom import script.

4.4. GAPS

In general, nanotechnology-specific assays, properties and related materials are sparsely represented in existing ontologies and thus need to be added to the existing assay and chemical information ontologies.

Several of the content areas that are relevant for nanomaterial safety assessment are thus far not covered by any existing publicly available ontology. One such gap is the nanomaterial lifecycle, from manufacturing through to environmental and biological impact. The InterNano Nanomanufacturing Taxonomy forms a good starting point for terminology in this area, but that terminology needs to be defined and further annotated. Known safety information is another gap. Database efforts such as the

OECD Database on Research into the safety of manufactured nanomaterials and the related OECD templates may serve as starting points here.

The eNanoMapper strategy for plugging these gaps will be to recruit expert groups in the relevant domain areas who are willing to collaborate in the development of new ontologies to meet the need in each gap area. This will be done during the ontology development phase of the project, which has already started, with the first initial release of the ontology due Month 12 and subsequent revision iterations on a regular basis.

5. COMMUNITY NEEDS

Before and during the first half-year of the eNanoMapper project, various related projects and efforts have been identified that need "IRI-fication" (i.e. the provision of an ontology-backed identifier) of their used terminology. In addition, several documentation resources have been created at a national, European, and pan-European level describing, for example, regulatory specifications of relevance in the nanosafety domain. We acknowledge that these resources need IRI-fication too, which will facilitate matching the content of datasets with the required annotations identified in regulatory specifications and other documentation resources.

5.1 EFFORTS THAT NEED IRI-FICATION

Table 1 shows a selection of the projects that eNanoMapper has been interacting with, for which a clear need has arisen for the provision of ontology-backed identifiers for the terminology that is in use. This table reflects the current state of the interactions for ontology mapping (where mappings are already in progress), while it is anticipated that through the outreach activities of WP1 the need for such mappings will be extended to many additional projects in the future.

For each, the primary goal of the effort is given, as well as the domains which are covered.

Table 1: Projects and efforts for which generated data needs conversion to ontology identifiers

PROJECT NAME	PRIMARY GOAL	DOMAINS COVERED
NANoREG	Literature extraction	Literature, nanomaterials, bio assays
NSC Database WG	Data quality annotation	Nanomaterials (physical, chemical properties), bio assays
CODATA/VAMAS, NANoREG	Nanomaterial characterization	Nanomaterials (physical, chemical, interactions)
NSC NanoQSAR cluster	Literature extraction	Literature, nanomaterials, bio assays
NanoSolutions	Safety classification of ENMs	Nanomaterial safety and biological interactions
MARINA	Integrative ENM safety assessment	ENM health and environmental safety, physicochemical properties, bio assays
NanoPuzzles	Algorithmic computational modelling of ENMs	inorganic and carbon ENMs, structure and physical / biological characterization, toxicity
ModNanoTox	Modelling of ENMs for toxicity prediction	nanomaterials (physical, chemical, interactions), toxicity, bioaccumulation

5.2 SPECIFICATIONS

Table 2 shows several documents that describe information that needs to be captured in ontological format and thereby mapped to data resources. Again, this list may grow during the lifetime of the project.

Table 2: Documentation resources that include terminology that needs to be mapped to ontology identifiers

RESOURCE NAME	PRIMARY GOAL	DOMAINS COVERED
OECD Harmonized templates	Various	Various
EU Nanomaterial definition	Nanomaterial characterization	Nanomaterials (size particularly)
Guideline for the Danish Inventory of Nanoproducts		

5. CONCLUSION

In this report we have surveyed the content areas that need to be described by the eNanoMapper ontologies. Areas of overlap and gaps have been identified. The strategy for managing the areas of overlap is to systematically prune duplicated content from the overarching set of ontologies in combined use. The strategy for managing the gaps is to work with ontology providers in the community in these areas in order to develop the missing resources.

6. BIBLIOGRAPHY

Brinkman, R.R., Courtot, M., Derom, D., Fostel, J.M., et al. (2010) Modeling biomedical experimental processes with OBI. J Biomed. Semantics 22;1 Suppl. 1:S7.

Buttigieg, P.L., Morrison, N., Smith, B., Mungall, C.J., Lewis, S.E. and the ENVO Consortium (2013), The environment ontology: contextualising biological and biomedical entities. Journal of Biomedical Semantics 4:43.

Gene Ontology Consortium, The (2000), Gene ontology: tool for the unification of biology. Nat Genet. 25(1):25-9.

Hardy, B., Douglas, N., Helma, C., Rautenberg, M., et al. (2010). Collaborative Development of Predictive Toxicology Applications, J. Cheminf. 2:7.

Hastings, J., Chepelev, L., Willighagen, E., Adams, N., Steinbeck, C. and Dumontier, M. (2011) The Chemical Information Ontology: provenance and disambiguation for chemical data on the biological semantic web. PLoS ONE, 6(10): e25513.

Hastings J., de Matos P., Dekker A., Ennis M., Harsha B., Kale N., Muthukrishnan V., Owen G., Turner S., Williams M. and Steinbeck C. (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res. 41:D456-63.

Mu, Q., Jiang, G., Chen, L., Zhou, H., Fourches, D., Tropsha, A. and Yan, B. (2014) Chemical Basis of Interactions Between Engineered Nanoparticles and Biological Systems. *Chemical Reviews* 114 (15), 7740–7781.

Natale, D.A., Arighi, C.N., Barker, W.C. et al. (2011) The Protein Ontology: a structured representation of protein forms and complexes. Nucleic Acids Res. 39(Database issue): D539–D545.

Smith, B., Ashburner, M., Rosse, C., Bard, C., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, Nat. Biotechnol. 25, 1251–1255.

Thomas, D.G., Pappu, R.V., and Baker, N.A. (2011) NanoParticle Ontology for Cancer Nanotechnology Research. J Biomed In-form. 44(1): 59–74.

Vempati, U.D., Przydzial, M.J., Chung, C., Abeyruwan, S., et al. (2012) Formalization, annotation and analysis of diverse drug and probe screening assay datasets using the BioAssay Ontology (BAO) PLoS One 7(11):e49198.

Williams, A.J., Harland, L., Groth, P., Pettifer, S. et al. (2012), Open PHACTS: semantic interoperability for drug discovery. Drug Discovery Today, Vol. 17, No. 21-22, pp. 1188–1198.