DELIVERABLE D1.2

Use Cases and Test Suite

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:	Maastricht University (UM)
DATE:	31.3.2015
VERSION:	V1.0

Call identifier	FP7-NMP-2013-SMALL-7
Document Type	Deliverable Report
WP/Task	WP1 Community outreach / Task 1.2 Definition of Personas & Use Cases, Task 1.3 Systems Design
Document ID	Deliverable 1.2
Status	Final Version

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) • University of Maastricht (UM)
Authors	E.L. Willighagen, B. Smeets, L. Rieswijk Reviewed by Barry Hardy
Purpose of the Document	To report the (intermediate) outcomes of Task 1.2 and 1.3.
Document History	V0.1. First draft: 2014-11-10 V0.2. Second draft: 2015-02-08 V0.3. Third draft: 2015-02-21 V0.4. Fourth draft: 2015-02-23 V0.5. Fifth draft: 2015-03-02 Final version: 2015-03-31

TABLE OF CONTENTS

TABLE OF FIGURES

GLOSSARY

Abbreviation / acronym	Description
WP1	**W**ork **P**ackage **1**
ISA-TAB	'**I**nvestigation' (the project context), '**S**tudy' (a unit of research) and '**A**ssay' (analytical measurement) general-purpose **Tab**ular format
ENMs	**E**ngineered **N**ano **M**aterials
EU	**E**uropean **U**nion
QSAR	**Q**uantitative **S**tructure-**A**ctivity **R**elationship
XLS	e**X**ce**L** **S**preadsheet
URI	**U**niform **R**esource **I**dentifier
QC	**Q**uality **C**ontrol
OECD	**O**rganisation for **E**conomic **C**o-operation and **D**evelopment
DB	**D**ata**b**ase
DA	**D**orothy **A**mber
EL	**E**ddy **L**agrange
DS	**D**agmar **S**ommerschool
PI	**P**eter **I**ngold
SAB	**S**cientific **A**dvisory **B**oard
DoW	**D**escription **o**f **W**ork
SVN	**S**ub**V**ersio**N**
RO	**R**esearch **O**bject
enm	**e**Nano**M**apper

1. EXECUTIVE SUMMARY

WP1 (Community Outreach) focuses on the communication between our project and the wider nanotechnology community. Its ultimate goal is to establish various tools which would enable collaboration with members of the nanotechnology community and to receive feedback on our project. This includes our involvement in the NanoSafety Cluster Database Working Group, a Requirements Analysis, and collaborations with other NanoSafety Cluster projects. The goals are to collaboratively develop interoperability documents together with definitions of personas and use cases that capture the community needs and working practices. Analysis of these requirements results in a global systems design, detailed descriptions of typical use cases, and a series of standardised test cases by which the system implementation can be tested. A major objective too is the development of a sustainability plan, capturing the current and future needs of the community and how they will be met.

This deliverable reports on the outcomes of the WP1 task on Definition of personas and use cases (**T1.2)**. There is a strong dependency on the WP1 Task on Requirements Analysis (**T1.1)**.

In this report we present our progress as of the end of Year 1 of the project with respect to the development and implementation of use cases to support the nanosafety community. A Requirements Analysis was performed and based on interviews held with people from the NanoSafety Cluster community, we were able to identify four representative personas. In parallel, a number of use cases have been described that are currently undergoing review so as to prioritize them for implementation in year 2 and year 3 of eNanoMapper. Furthermore, we have described already implemented and upcoming testing approaches for these uses cases and the technological solutions used in these use cases. The exact testing approach will be tuned for each use case, depending on the development model of the underlying components.

We are currently internally and externally prioritizing these use cases based on a variety of criteria outlined in this report. During the end of year 1 annual eNanoMapper meeting in Sofia, the consortium identified which use cases will be implemented as part of tasks in the DoW other than the task about implementing use cases derived from the community, as those described in this deliverable report.

2. INTRODUCTION

2.1. OVERVIEW OF WP1 (COMMUNITY OUTREACH)

WP1 focuses on the communication between our project and the wider nanotechnology community. It establishes various mechanisms we use to ensure collaboration with and feedback on our project. This includes our involvement in the NanoSafety Cluster Database Working Group, a Requirements Analysis, and collaborations with other NanoSafety Cluster projects.

The goals are to collaboratively develop interoperability documents together with definitions of personas and use cases that capture the community needs and working practices. Analysis of these requirements results in a global systems design, detailed descriptions of typical use cases, and a series of standardised test cases by which the system implementation can be tested. A major objective too is the development of a sustainability plan, capturing the current and future needs of the community and how they will be met.

2.2 PERSONAS

Task **T1.2** establishes a list of personas and use cases, reported in this deliverable report, **D1.2**: a small set of unique personas will be defined that reflect the requirements analysis in **T1.1**. These personas are social descriptions of different working users with typical needs and working habits. The personas will be selected to serve as reflecting our user base, and future work will primarily aim to serve their needs. This ensures that solutions are not tuned only to one group of particular users from the community and remain generic. The personas will serve the important role of communicating user needs and solutions and will guide the design, development of the eNanoMapper ontology and platform, and be used as initial sketches of the target audience for the sustainability plan (**T1.9**).

2.3 USE CASES

This task also defined a list of Use Cases based on information gathering through virtual and face-to-face interactions where user interviews, work practices and artifacts were collected (**T1.1**) and organised based on a user-centered conceptual design methodology.

3. PERSONAS

Goal of personas is to define a selection of archetypical users that reflect a unique part of our user base. These personas do not really exist, but we do as if they do: they have a face, a name, and a story. The approach we take here is outlined in Chapter 9 of the Rapid Contextual Design book [1]. Based on the interviews held as part of the Requirements Analysis task (**T1.1**), four personas have been identified and defined based on the notes taken during those interviews.

3.1 DEVELOPMENT

Initial personas were taken from the ToxBank project. These were updated to the recent state of affairs based on the Community Outreach work in **WP1**. This was just a rough starting point, used to categorize the users found in the Requirements Analysis task. This led to six personas: *Dorothy Amber*, a data analyst, *Eddy Lagrange*, who works in an experimental lab, *Angela Paracelsus*, an ambitious postdoc, *Olivia Sanger*, a PhD student with lack of access to data, *David Sommerschool*, a data scientist dealing more with data exchange than statistics, and *Peter Ingold*, a principle investigator. Detailed descriptions follow in the Section 3.2.

These personas led to the following classification of users as identified in the Requirements Analysis (see also **D1.1**).

- Dorothy Amber: U01, U04, U05, U12, U16, U19, U30, and U31
- Eddy Langrange: U03, U10, U13, U23, U25, U26, U27
- Angela Paracelsus: U02
- Olivia Sanger: U09, and U17
- David Sommerschool: U21, U22, and U24
- Peter Ingold: U07, U08, U11, U28, and U29

These results show that our first year Requirements Analysis did not cover all personas equally well, or, equivalently, not all personas cover our user base well.

To further understand our user base, we looked at the similarities between users. Because the Requirements Analysis classified all user notes into categories, used for the affinity diagram reported in **D1.1**, we can use this information to plot a network of users and their notes. This is depicted in Figure 1 where blue nodes are the notes, and the nodes with other colours are users, where each colour indicates one of the above personas. U06 did not have enough user notes to be classified, and is coloured white. The figure shows that there is no strong pattern among users of the same persona. This could partly be caused by how the interviews were held, which were more focused on their views on the NanoSafety research field, rather than a focus on their daily activities which the personas attempt to capture.

Furthermore, because the Angela Paracelsus and Olivia Sanger personas did not classify many users, the users of those two personas were re-classified among other personas. To restore the gender balance among the personas, David was renamed to Dagmar.

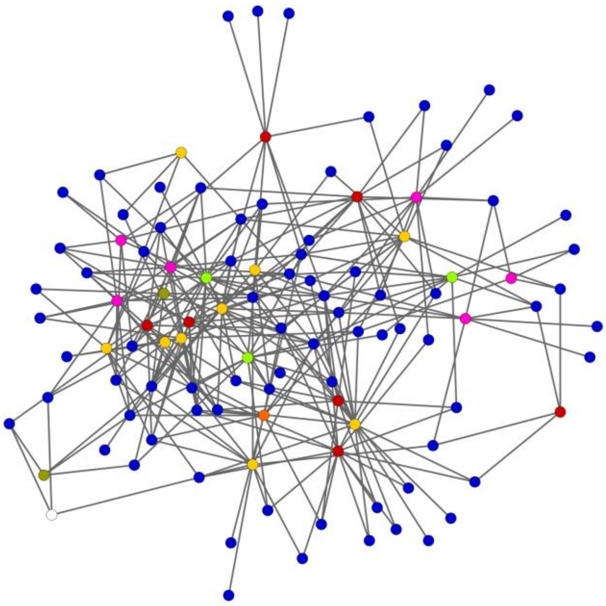

Figure 1 Network of user notes (blue) and users, colored by persona. Angela Paracelsus: orange, Dorothy Amber: yellow, Dagmar Sommeschool: green, Eddy Lagrange: red, Olivia Sanger: green-brown, Peter Ingold: pink). One user (U06) could not be classified and is colored white. The network was created with Bioclipse [2] *from content in a Google Drive spreadsheet and visualized with Cytoscape* [3].

For each of the remaining four personas the Requirements Analysis user notes were used to find quotes, roles and goals.

3.2 RESULTS

This section lists the accepted personas (four) and personas not currently backed with a significant number of users from our Requirements Analysis (two).

3.2.1 ACCEPTED PERSONAS

DOROTHY AMBER

data analyst

Description

The computer specialist where everyone dumps their data to do some quick analysis and to get some fancy graphics. She increasingly uses more advanced bioinformatics tools, but mostly spends hour after hour normalizing data, making it fit together and into her tools. Loves to be first author on a major paper, but repeatedly runs into being limited to second hand data. Worse, lack of access to the original RAW data hinders detailed analysis, blocking true innovation.

Quotes

- *"We created templates for biologists to provide us with the data we need for NanoQSAR modeling"*
- *"We started with a script in R, but now rewrote that in Python"*
- *"There is a need for people who can validate data. That seems to be a challenge."*

Goals

- To uncover underlying statistical proof for structure-activity relations

Roles

- (statistical) analyst

EDDY LAGRANGE

experimental lab worker

Description

Performs lab work and deals with lots of experimental data, which he manually copies on USB sticks from devices in the lab to computers elsewhere. Does not do a lot of data analysis himself and barely knows how to open an ISA-TAB template in Excel. Word he does know, and writes his thesis in this system, together with the Mendeley plugin. Would die for a system to make it easier to publish his experimental work.

Quotes

- *"We only use in vitro studies. We didn't do in vivo studies."*
- *"We are one of the labs that develop new nanomaterials."*
- *"We would like a system that allows us to share negative data too."*

Goals

- Study the biological effects of ENMs using experimental approaches
- Write research papers

Roles

- (experimental) researcher

DAGMAR SOMMERSCHOOL

data scientist

Description

Dagmar has a degree in computing sciences and works in an EU project where she is responsible for data sharing within the consortium. She designed a nice system using her favorite database system and a matching, user- friendly interface. The data still has to arrive.

Quotes

- *"Our biggest concern is how to ensure that biologists do not make a mess of the data in their spreadsheets."*
- *"We aim at five to ten reviewers for each paper, in our collection of 1000 papers for this project."*
- *"There is not a standard for reporting on nanomaterial characterizations"*

Goals

- Make public data available to local researchers
- Normalize and otherwise integrate data from different sources

Roles

- Data manager

PETER INGOLD

principal investigator

Description

Peter is a tenured professor and principal investigator on several large EU projects. No longer does any research himself or writes papers himself, an occasional review or editorial set aside. Deals with management and acquisition and leaves practical work and education to his post-docs. Only wants to know if colleagues can make the deliverable deadlines.

Quotes

- *"Our new application idea is the evaluation of nanomaterials as remediation possibility for contaminated sites"*
- *"The system will not be free, but commercial; we need resources to support software development"*
- *"I lead a WP which is about developing tools to inform industry about how to handle the safe production of ENMs"*

Goals

- Grant acquisition
- Run a research lab

Roles

- Research Manager

3.2.2 IGNORED PERSONAS

ANGELA PARACELSUS

ambitious post-doc

Description

Post-doc who enjoys her research but also wants to get a good grip of what happens around her for things that can strengthen her work and possibly lead to a tenure track position. She joined the NanoSafety Cluster working groups to learn about other research, and is eager to show her research to others. She finds high quality research important and that proper credit is given to (her) contributions.

OLIVIA SANGER

open scientist

Description

PhD-student at a prestigious European research institute, with limited access to literature, interfering with her doing research. Searching literature and databases is an important aspect of her work. As a result, she got interested in the Open Science community, and has already convinced some fellow students in the lab to start use Figshare and publish in PLoS ONE.

3.3 DISCUSSION

As indicated, the selected personas are based on the interviews held as part of **T1.1**. This implies that the personas are restricted to the users interviewed: because we have not interviewed legislators, journalists, citizen scientists, information technologists, those users are not really well represented by these personas. We are aware of this limitation, and are discussing if the current personas sufficiently supporting the use cases and whether we want to also target other personas later in the project.

4. USE CASES

4.1 USE CASES DERIVED FROM THE REQUIREMENT ANALYSIS

Based on the requirement analysis, the following use cases were extracted: *1. Write protocol; 2. Design protocol; 3. Rank paper; 4. Upload protocol; 5. Upload dataset; 6. Design study; 7. Search protocol; 8. Search dataset; 8. Search ENM; 9. Download protocol; 10. Download dataset; 11. Extract data from paper; 12. Prepare data for modeling; 13. ISA-TAB creation; 14. Build QSAR; 15. Validate ISA-TAB; 16. Validate protocol; 17. Validate model; 18. Find ISA-TAB template; 19. Create template ISA-TAB; 20. Create template (XLS etc); 21. Map nanomaterial to biological pathway; 22. Search for perturbed gene; 23. Search for all knowledge of nanomaterial X; 24. Search for all knowledge of impurities; 25. Calculate precaution for X; 26. Copy open facts to generate it; 27. Find/ link related data sources; 28. Find data with similar patterns; 29. Save my data in format X; 30. Publish my data with my paper; 31. Find people with similar X; 32. Who else with protocol X; 33. What URI to use for X; 34. Give me all names for ontology term A; 35. Create NSL dictionary; 36. Harmonize Terminology; 37. Win Nobel prize / Automatic Knowledge discovery tool; 38. Find ontological contradiction; 39. Comp aided ENM design; 40. Define safe-by design; 41. Integrate data for risk assessment; 42. Verify data against regular requirements; 43. Convert from one template to another, e.g from Modclust to OECD; 44. Find producer of eNM C; 45. Find QC data of eNM C; 46. User alert for new info on D (gene, nano tag); 47. Find nanoparticles used in product; 48. Format data to be used by C; 49. Map nano material found DB 'A' to DB 'B'; 50. Find all DB's with data of Paper X; 51. Write up tutorials on use cases; 52. Register and get access (public); 53. Map existing used schemas to ontology IDs (e.g. OECD harmonized templates); 54. Annotation of data with ontology IDs (e.g. experimental data in ISA-Tab); and 55. Experimental design.*

Full descriptions of a selection of these uses cases are provided in Appendix A. Out of the aforementioned 55 use cases, descriptions have been developed for 31 of them so far. Each description indicates which actor it is targeted against, the scope of the use case, a brief description, a trigger that starts of the use case, the basic flow of the use case, and the expected outcome of the use case. The latter is also the test condition for a future implementation of the use case to decide if the use case has been successfully implemented.

4.2 THE USE CASES SHORT LIST TO BE PRIORITIZED

The eNanoMapper consortium described the following short list of use cases on the short list to be prioritized and for each the matching persona(s) are given (DA=Dorothy Amber, EL=Eddy Lagrange, DS=Dagmar Sommerschool, PI=Peter Ingold):

Use Case	**DA**	**EL**	**DS**	**PI**
02. Design protocol		X		X
04. Upload protocol		X	X	
05. Upload dataset	X	X	X	X
06. Design study	X			X
07. Search protocol		X		X

08. Search dataset	X	X	X	
08. Search ENM			X	X
10. Download dataset			X	
13. ISA-TAB creation	X	X	X	
14. Build QSAR	X			
17. Validate model	X	X	X	
20. Create template (xls etc)		X	X	
21. Map nanomaterial to biological pathway	X			
23. Search for all knowledge of nanomaterial X	X	X	X	X
27. Find/ link related data sources	X	X	X	X
29. Save my data in format X	X	X	X	
33. What URI to use for X	X		X	
34. Give me all names for ontology term A		X	X	
36. Harmonize Terminology			X	
38. Find ontological contradiction			X	
39. Comp aided ENM design	X			
44. Find producer of eNM C		X		
45. Find QC data of eNM C	X	X		
49. Map nano material found in DB 'A' to an entry in DB 'B'	X		X	
50. Find all DB's with data of Paper X	X		X	
52. Register and get access (public)	X	X	X	X
53. Map existing used schemas to ontology IDs (e.g. OECD harmonized templates)			X	
54. Annotation of data with ontology IDs (e.g. experimental data in ISA-Tab)		X	X	
55. Experimental design	X	X		X

4.3 PRIORITIZATION

The eNanoMapper project only has 16 man months set aside dedicated for use case development. At the same time, with the number of use cases already identified in the first year and expected to increase and possible change in the next two years, a selection of use cases need to be identified that will be implemented in the task in **WP5** dedicated to making these use cases reality. A further observation is that these use cases can also be developed as part of other eNanoMapper tasks, for various reasons. For example, some use cases may result in a marketable product to support the sustainability track, and other use cases may overlap with scientific interests of eNanoMapper or associate partners and implementation could be funded with other resources. That too, makes it important to identify which use cases **WP5** will work on, to reduce duplication of effort.

The ordering of use cases by priority will be based on rankings from four types of criteria as outlined in the next subsections.

4.2.1 SCIENTIFIC ADVISORY BOARD INPUT

The Scientific Advisory Board (SAB) will be asked to provide input in this prioritization effort. SAB members will give a subset of at least 10 use cases a score reflecting how important that use case is to them. These scores will be used to order the use cases and identify the most important ones. The criteria SAB members use are not predefined and it is left to the SAB member to decide why they find one use case more important than another. They will be asked to provide comments.

4.2.2 FEASIBILITY GIVEN ENANOMAPPER TECHNOLOGIES

Based on the currently available and soon ready components developed in WP2, WP3, and WP4, the feasibility of implementing this use case will be summarized. This will go hand in hand with the development of a number of “paper prototypes” which give an idea of how the product will work, but fully on paper, based on the Conceptual Design model. For each use case the prototyped functionality is linked to existing technologies and a brief gap analysis is made to anticipate the feasibility of the full project.

4.2.3 AVAILABILITY OF USERS IN THE NSC COMMUNITY

A further criterion that we will want to use is if there are people in the NanoSafety Cluster community (and to a lesser extend beyond this community) that can help validate the outcome of the implementation. For example, are people available that can perform test scripts and provide us with their evaluations (see also the next section)?

4.2.4 RELEVANCE TO THE SUSTAINABILITY TASK

A fourth criterion is related to the sustainability track of eNanoMapper, as outlined in **T1.9**. For example, use cases that can support a revenue-generated product and as such can support the sustainability of the whole project will be ranked higher based on this criterion.

4.4 PRIORITIZED USE CASES

This list does not exist yet. The prioritization process is still in progress.

5. TEST SUITE

This section describes the testing approach for the use cases. While **T1.4** aims at development the specific testing plans for the selected use cases, we here outline the general approach. Similar to our approach for validating other components of the eNanoMapper platform (software modules, ontology components, etc.), we apply a general scheme involving the following order of steps in the release process, where a release here reflects to the delivery of a solution to a particular use case.

5.1 DEVELOPMENT CYCLE

The development model used by the partners integrates testing at various levels, and the modular design makes this easier than with monolithic systems. Part of this is reflected by the early adoption of issue tracking systems and an automated build system (see **D5.1**). Still, we have further formalized the full development process, with a focus on the testing. One aspect is that testing of modules needs to be complemented by testing of the integrated system.

As written in the Description of Work (DoW) of this project, we will see a release soon-release often approach, common in Open Source projects. This will help enable early adopters to follow our progress and use (parts of) our solutions. This implies that we will see many releases, at various levels of maturity. We have not formalized (yet) how we will label the releases, but common schemes include (development, production), (unstable, stable), and (alpha, beta, release candidate, release). Important here is that it is communicated what users can expect.

It is also established in a series of project meetings held in autumn 2014 that each maturity level has different levels of testing, starting with very basic “unit” testing to more high level testing, e.g. with test scripts that verify if some task can be successfully completed. Each use case description has a brief “measurable outcome” description that is a test at this higher level. Additionally, each release level has an associated, clearly defined period and level of support. More unstable releases have less formal support, for example.

The full development, test, release cycle, each leading to a release of any of the maturity levels, is established to look like the following, in which the milestones refer to time points in this cycle:

1. Milestone 1: Development Cycle: decide type of next release (and thus goals and intermediate expectations)
    a. alpha: Open and aimed at enm developers and users and external developers; support only until the first next release
    b. beta: Open and aimed at enm and *selected* external users; support only until the first next release
    c. public: Open and aimed at the general audience;
2. Milestone 2: Run validation before official release (in order of successfulness; where applicable):
    a. automated testing
    b. enm-internal review testing and technical assessment,
    c. enm-internal domain and acceptance testing
    d. including testing against enm use cases (formal and from DoW)

3. Milestone 4: Report measurable outcomes including the revision that was tested (SVN: version; git: hashtag) and a pointer to the running installation or specific file
4. Milestone 3: Review validation results
5. Milestone 4: Release X: tagging development repository with a version or date and announce release of RO via one or more official channels
6. Milestone 5: Post-release: aggregate community feedback (in parallel with the next cycle) including testing by selected domain experts. Channels:
    a. public: public issue trackers (e.g. GitHub)
    b. confidential: private issue tracker (e.g. Bugzilla, webform, email address)

5.1 TESTING OR UNDERLYING TECHNOLOGIES

Because the success of a use case also very much depends on the accurate implementation of the underlying technologies (ontology, software modules), the testing plan of a use case also refers to the testing of the integrated technologies and not just to the modules. Therefore, a stacked set of tests is needed, starting from low level unit tests to test underlying technology, up to high-level testing that ensures that the underlying technology is not just cool, but that it indeed also solves real world problems.
We have identified a number of "research objects" (RO) and for each indicate type of RO what the currently anticipated testing approach is, how often that testing is performed, and whom or what will do that testing:

Research Object	Testing Approach	Frequency	Who / What
all/any	1. has suitable license/copyrights	1. before start of implementation of RO 2. double check at the end	1. developers 2. release review team
ontology	1. using domain knowledge, check coverage of ontology 2. check organization of ontology (hierarchy) is sensible 3. (automatic) check all ontology terms are defined 4. and then check those definitions are agreeable to domain experts 5. (automatic) check no contradictions in ontology	1. 6-monthly 2. 6- monthly 3. daily 4. 6- monthly 5. daily	1. domain expert 2. domain expert and WP3 3. can be automated (Jenkins) 4. domain expert 5. can be automated (Jenkins)
use case	1. Verify the test scripts, matching them to Use Cases, paper prototypes and tech specs.	1. monthly	1. developers, release review team
API	1. unit tests by developers	1. daily	1. can be automated

	2. test scripts using machine readable API specification 3. test scripts using more than one service	2. daily 3. daily	(Jenkins) 2. can be automated (Jenkins) 3. can be automated if testing scenarios are defined
data convertor (import/export)	1. unit tests for parsing supported set of input formats 2. unit tests for generating supported set if output formats 3. check if data objects of interest to NSC can be supported 4. check if ontology annotation exist for terms in each format	1. daily 2. daily 3. 6-monthly 4. daily	1. can be automated (Jenkins) 2. can be automated (Jenkins) 3. domain experts 4. existence check can be automated, domain expert to verify
Modeling Algorithm Services	1. Develop nQSAR models for literature data present in the database (a) Test scripts to verify integration with APIs and database (b) Test scripts to verify model building (c) Test scripts on 2. Algorithms in terms of time	1. 6-monthly 2. 6-monthly	1. domain experts 2. developers
Descriptor Calculation Services	1. Generate and validate descriptors using literature data (for example images) 2. Test scripts for descriptor calculation services using literature data (for example images)	1. 6-monthly 2. 6-monthly	1. domain experts 2. developers
Models	1. Evaluate performance of models over their lifecycle: 2. Validate models 3. Evaluate performance of models in database to the ones reported in literature 4. Test models services: 5. Test scripts to verify	1. 6-monthly 2. 6-monthly	1. domain experts 2. developers

	model storage and recall 6. Test scripts to verify integration with validation services 7. Test scripts to check consistency of predictions with information found in the literature		
UI components	1. automatic 2. manual	1. daily 2. 6-monthly	1. scripts 2. domain experts
Database , Model, AA services	1. response time	1. hourly	1. smokeping-like tests (automatic)
Database Implementation	1. check if data objects of NSC interest can be imported / exported 2. check if the types of searches supported are sufficient 3. check if ontology annotations are missing	1. daily 2. 6-monthly 3. 6-monthly	1. can be automated 2. domain expert 3. domain expert

6. CONCLUSION

This report presents our progress with respect to the development of implementations of use cases to support the nanosafety community. Based on interviews held with people from the NanoSafety Cluster community, we have identified four personas, capturing unique needs for each. In parallel, a number of use cases have been described that are currently undergoing review as to prioritize them for implementation in year 2 and year 3 of eNanoMapper. Furthermore, we have described already implemented and upcoming testing approaches for these uses cases and the technological solutions used in these use cases. The exact testing approach will be tuned for each use case, depending on the development model of the underlying components.

The current state is that the use cases still need a completed prioritization. We are behind schedule with this, due to unforeseen circumstances. During the end of Year 1 eNanoMapper annual meeting in Sofia, the consortium identified which use cases will be implemented as part of tasks in the DoW other than the task about implementing use cases derived from the community, as those described in this deliverable.

BIBLIOGRAPHY

[1] K. Holtzblatt, J. B. Wendell, and S. Wood, *Rapid Contextual Design: A How-to Guide to Key Techniques for User-Centered Design (Interactive Technologies)*. Rapid Contextual Design, 2004.

[2] O. Spjuth, J. Alvarsson, A. Berg, M. Eklund, S. Kuhn, C. Mäsak, G. Torrance, J. Wagener, E. L. Willighagen, C. Steinbeck, and J. E. S. Wikberg, "Bioclipse 2: a scriptable integration platform for the life sciences.," *BMC Bioinformatics*, vol. 10, p. 397, 2009.

[3] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A software Environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, pp. 2498–2504, 2003.

www.eNanoMapper.net

APPENDIX A

2. Design Protocol

Primary Actor: Senior Researcher

Scope: Work out a protocol, which produces identical results.

Brief: A researcher needs to design a new protocol. After having reviewed the protocols available, the researcher draws out a new one that is tested and then published.

Trigger: Need to unify a procedure and make it reproducible.

Basic Flow:

1. A researcher with good knowledge of the area decides to design a new protocol.
2. Explore: Find templates and similar protocols on eNanoMapper protocol service, as well as via google and other resources.
3. The researcher reviews current protocols through examining publications and SOPs.
4. Critical areas that need attention and/or changes are identified.
5. A new protocol is drawn out in detail.
6. Test the protocol with (junior) researchers, possibly in different labs. Make sure all results are identical.
7. Refine protocol, and go back to (7), until results are identical.
8. Upload and publish protocol > *Use case 4*

Completion condition: The protocol is defined in enough detail, so that it can be reproduced by any lab worker with identical results.

4. Upload Protocol

Primary Actor: Protocol Designer / Senior Researcher, Software Integrator / Programmer

Scope: To upload one or more protocol(s)

Brief: To upload a protocol, with the following objectives:

- Easily accessible - Found by search engines
- Accessible via API (machine readable)
- Unique ID for references.

Trigger: *Use Case 2 completed*

Basic Flow:

1. **Researcher:**
    a. Go to eNanoMapper or another protocol service and consult on compatible formats for upload.
    b. Upload the protocol via the web form provided.
    c. Review the uploaded protocol and retrieve the URI for reference, i.e. on publication, research report Deliverable, communication with fellow researchers.
2. **Programmer:**
    a. Consult the API of eNanoMapper or another protocol service on compatible formats for upload, i.e. text/uri-list; application/rdf+xml; text/n3.
    eg: http://api.toxbank.net/index.php/Protocol#POST:_Upload_a_new_Protocol
    b. Upload the protocol(s) via curl commands.
        i. ***example:***
        curl -X POST -H 'subjectid:TOKEN' -H 'Content-Type:multipart/form-data' -F 'filename=@FILE_NAME' \
        -d 'title=VALUE' -d 'anabstract=VALUE' -d 'author_uri=VALUE' -d 'author_uri=VALUE' -d 'author_uri=VALUE' \
        -d 'keywords=VALUE' -d 'summarySearchable=VALUE' -d 'project_uri=Project URI' -d 'organisation_uri=Organisation URI'
        -d 'user_uri=Owner User URI' http://enanomapper.org:8080/protocol (example address not valid)
    c. Retrieve the metadata of the protocol(s) for reference, i.e. on publication, research report Deliverable, communication with fellow researchers.
        i. ***example:***
        curl -X GET -H 'Accept:application/rdf+xml' -H 'subjectid:TOKEN'
        http://enanomapper.org:8080/protocol/example_protocol1 (example address not valid)

Completion condition: Protocol(s) are online, appear in search & have URIs.

5. Upload Dataset

Primary Actor:

- Producer of experimental data (principal investigators, laboratory personnel)
- Producer of derived data
- Owner of data (company, research facility, public regulatory organization, industrial associations)
- User of data (modeler that has found someone else's data in literature)

Scope: Share (public/limited), make data usable by others, process, model, commercialization

Brief: Datasets may be uploaded by:

- Producers of data, who have made experiments themselves
- Scientists that publish derived data from pre existing datasets, including intermediate datasets that are produced in preprocessing stages
- Owners of data that want to make the data available for calculations or dissemination
- Users of data that want to use/publish a dataset found in literature

Trigger:

- Publish the data and make available to others (incl. obtaining the URI/doi)
- Publish Supporting Information for a publication
- Make additional calculations
- Commercialize

Input data = the dataset and the associated license

Basic Flow:

1. Prepare data
    a. Gather data
        i. Make experiments
        ii. Preprocess data
        iii. Submit publication (where applicable)
        iv. Collect from literature
    b. Prepare data in format accepted for uploading
    c. Include meta information (i.e. ISA-Tab Nano)
2. Upload data
3. Validate the dataset
4. Define licensing of dataset (Public, Limited, Commercial) and assign access rights to groups/people
5. Visualize the dataset
6. Obtain URI and DOI
7. Disseminate URI and DOI (incl. submitting the link to the Journal)

Completion condition: **needs to be added**

6. Design Study

Primary Actor: Senior Researcher (academic or research institution or at a company).

Scope: To create a high level design for a study, to test the validity of a hypothesis.

Brief: The process of designing a study to expand knowledge about the safety of an engineered nanomaterial (or group of engineered nanomaterials).

Trigger: Can be academic or commercial:

- **Academic**: Gap of knowledge about the behaviour of an engineered nanomaterial or group of engineered nanomaterials in a scientifically relevant area.
- **Commercial**: Need to verify the suitability of an engineered nanomaterial(s) for use in product(s). The initiator may be product development, legal, marketing or another unit.

Basic Flow:

1. Formulate a hypothesis
2. Identify key elements of the study
    a. Choose engineered nanomaterial to study
    b. Study literature / find similar cases -> 11
    c. Networking: find people, groups, organizations -> 31
    d. Choose the endpoint(s)
3. Determine a high level plan for the study (Use Case 23)
    a. Outline experimental factors (organism, cell lines, phys/chem parameters ...)
    b. Find existing protocols / methods / assays -> 7
    c. Determine statistical models -> 14 / 21
4. Data generation planning (if needed and not available) -> 21
    a. Choose the engineered nanomaterials
    b. Find and choose engineered nanomaterial supplier or suppliers, or synthesize
    c. Find reagents / kits / ... for protocol
    d. Determine data to be generated and their format
5. Implementation planning
    a. In/outsourcing decisions
    b. Team planning
    c. Time planning

Completion condition: Descriptions of engineered nanomaterials, endpoints, experimental conditions and implementation plan in enough detail to start the study.

7. Search for Protocol

Primary Actor: Study Designer, Researcher, Lab personnel, Regulator.

Scope: Find protocols, compare protocols, evaluate protocols, find studies that followed a protocol.

Brief: Find all available protocols for an endpoint, a engineered nanomaterial or a engineered nanomaterial-endpoint combination. Find all the work that has been done with a specific protocol.

Trigger:

- Decision about the basic protocols to be used by the study designer.
- Search for updates, details etc of protocols by research and lab personnel.
- Evaluation of existing protocols by regulators or to design a new protocol.

Basic Flow:

1. The researcher goes to google or directly to eNanoMapper landing page and enters the name of the procedure and/or a few keywords of it.
    a. Besides other resources, an eNanoMapper landing page shows in the google results and follows the link.
2. He/she has three choices to find all eNanoMapper supported internal and external resources:
    a. Browse a hierarchy
    b. Search by keyword
    c. Advanced search provides him/her with a search form and/or language to find protocols according to criteria (i.e. an endpoint, an engineered nanomaterial or a class of engineered nanomaterials, author, popularity …)
3. The researcher finds all protocols applicable to his/her purpose. Together with the protocol, he/she finds additional useful meta-information about creators, usage (by whom), popularity (number of downloads) etc.
4. He/she selects suitable protocols for his/her procedure.

Completion condition: Selection of applicable protocols or expansion or expansion of knowledge on protocols available for a procedure.

8.1. Search an engineered nanomaterial

Primary Actor: **needs to be identified**

Scope: The primary actor wants to find an engineered nanomaterial via structured search, supported by ontologies.

Brief: **needs to be added**

Trigger: The free text search was deemed insufficient and is not subject of this use case.

Basic Flow

1. Search engineered nanomaterial (by identifier, name, formula, etc.) – e.g. ZnO
    a. Show known categories (multiple)
    b. Click on each category
    c. Show range/scale/possible values for the category (and what range the hit is in) - e.g. zeta potential
    d. Filter - show all data of same type, but allow filtering / changing the scope to all engineered nanomaterials or filter by some parameter
    e. Visualise : showing if the engineered nanomaterial is in XX% top values; show real data (e.g. download or web browser UI; spider plots; charts; interactive exploration)

2. Search engineered nanomaterial by property . Purpose: provide information if a property X has influence / is important for property (activity) Y; supports experimental design by guiding the selection of engineered nanomaterial to test.
    a. Search by property (as in NPO hierarchy)
    b. Show all engineered nanomaterials in (selected) categories #3 & #6 (examples) which have high values of the selected property A

Completion condition: Procedure:

- Review the data model and the API and decide if extensions are needed to support the structured search.
- Implement API.
- Implement user interface.
- Select data for demonstration (possible VTT generated HTS data).
- Import the data.

8.2. Search Dataset

Primary Actor: Researcher or Developer, Application (e.g. web service)

Scope: Establish online access to eNanoMapper data with search functionality which supports Authorisation and Authentication (A&A).

Brief:

- A database has to provide an index or list of the datasets. A dataset can be an internal or external resource.
- The data exchange has to work with the A&A system to ensure the protection of confidential information.
- Different ways of access have to be provided (e.g. human and machine based access).

Trigger: Need of data accessibility for users, web services and linked resources.

Related use cases:
5. Upload dataset
10. Download dataset
27. Find/ link related data sources
29. Save my data in format X

Basic Flow

1. **User request** via REST API or Graphical User Interface (GUI) User-Access Control System
    1. Search via GUI with different options,
    2. RESTful webservice. API definition.
2. **Linked Resource** search via a linked resource as linked data
3. **Web services** e.g. search and return list

Completion condition: needs to be added

10. Download Dataset

Primary Actor: Researcher or Developer, Application (e.g. web service)

Scope: Establish online access to eNanoMapper data with download functionality which supports Authorisation and Authentication (A&A).

Brief:

- What kind of data are we dealing with? And will be used in this use case?
  - ISA-tab nano files / DB search (filter) results / tabular study results for experiments / ...
- Which download formats should be provided?
  - RDF, JSON, n-triple, raw file, ...
- The data exchange has to work with the A&A system to ensure the protection of confidential information.
- Different ways of access have to be provided (e.g. human and machine based access).

Trigger: Need of data exchange between users, web services and linked resources.

Related use cases:
5. Upload dataset
8. Search dataset
27. Find/ link related data sources
29. Save my data in format X

Basic Flow

1. **User request** via REST API or Graphical User Interface (GUI)
   User-Access Control System / Download formats (RDF, JSON, n-triple) /
   1. Download via GUI
   2. RESTful webservice. API definition.
2. **Linked Resource** download via a linked resource as linked data
3. **Web services** e.g. modelling services

Completion condition: needs to be added

13. Create ISA-Tab

Primary Actor: Data provider (experimental lab) or Data curator (database)

Scope: Research; Sharing; Data Exchange

Brief: Data providers wish to exchange or share their data by making it available in a database or repository, or database curator wishes to format data in ISA-Tab for optimal processing.
Input: There is some data available and it needs to be described in the ISA-Tab format.
Output: Data formatted in ISA-Tab.

Trigger: There is data available and a decision to make it public and format it for exchange and comparison with other datasets. ISA-Tab is chosen as a metadata-rich format for this purpose.

Basic Flow:

1. Input: There is a file containing experimental data. This is typically in one or more spreadsheets
2. Processing: There are three options for converting data into ISA-Tab format:
   a. Use the tool ISACreator. This is a desktop application which is available at https://github.com/ISA-tools/ISAcreator. The tool manages the creation of ISA-Tab files from spreadsheets with associated metadata creation, but does not (yet) support the ISA-Tab Nano extension for the description of the complexity of the particles in the experiment.
   b. Use a custom script -- requires programming expertise. However, such scripts are available as part of the eNanoMapper database implementation.
   c. Manually. Creating ISA-Tab files manually is possible but is obviously complex and error-prone. This is to be avoided in general except for the smallest of datasets.
3. Metadata: During the process of converting to ISA-Tab, metadata should be associated to each field. The metadata takes the form of ontology annotations. By default, ISACreator allows selection of ontology annotations from ontology aggregation resources such as BioPortal (http://bioportal.bioontology.org). However, in the future we should provide the facility to restrict the annotation selection to the eNanoMapper ontology in the first instance.
4. Template: For repetitive ISA-Tab constructions, it is desirable to have a template which can be re-used in the processing (step 2) of multiple input datasets. See Use Cases 18 & 19.
5. Validation: Once the file has been created it should be validated. See Use Case 17.

Completion condition: When the use case is fulfilled there is a concrete file in ISA-Tab format containing all the data initially targeted for inclusion that validates in all applications using the ISA format.

14. Build QSAR

Primary Actor: Harry, Phillip, Georgia, Cristian, Bart, ….

Scope: Engineered Nanomaterial Descriptors, Toxicology Prediction Models

Brief: The scope is to find the best engineered nanomaterial descriptors, datasets based on specific physical/chemical/biological information and regression method in order to obtain nanoQSAR models that can accurately predict toxicological properties. The best models should be implemented into OpenTox platform. Each step involves a programming task to develop custom tools.

Trigger: The necessity of theoretical methods for engineered nanomaterial toxicology prediction generated this use case. The first scope is to find the information that can describe engineered nanomaterial toxicology. The second one is to describe a biological mechanism of the toxicological effects of engineered nanomaterials and the cellular response. Non-extended data sets for engineered nanomaterials will limit the results.

Basic Flow

1. **Regression** tools
    a. RRegrs repository: https://github.com/muntisa/RRegrs
    b. Pseudo code: https://docs.google.com/document/d/1St4y5kp0dQhB7ahM8o18BzoXL98viK12VZeh80ftQz8/edit
2. **Descriptors**
    a. Development of new engineered nanomaterial descriptors
        i. Single Values of Markov Transition Connectivity Matrix - R - RDMSV repository: https://github.com/muntisa/RDMSV
        ii. Image descriptors (Philip): https://docs.google.com/document/d/1IMo0eJdYaGM3yg_QGD7qMjBut1y9R4Y6j9DYt4AyNYM/edit
        iii. Quantum descriptors (Philip)
        iv. …. **information added?**
    b. Tools for existing descriptors
        i. Markov descriptors for:
            1. Drugs/engineered nanomaterial- RDMMP respository: https://github.com/muntisa/RDMMP
            2. Proteins - to be done
        ii. … **information added?**
    c. ….**information added?**
3. **Data set** generation using different information (engineered nanomaterial, linker and protein descriptors)

a. Corona protein case
    i. Molecular Star Graph descriptor - based methods: https://docs.google.com/document/d/1MmJR2SO9a09Cg41XZD_R1stSAztNqkRW8Epsrxt8VVE/edit
    ii. Checking the paper models:: https://docs.google.com/document/d/1LYuox_W80eBobRyaz_YLCCTPqLLgKSgmqbEzulSAtyl/edit
    iii. Clustering study (Georgia)
    iv. GO study (Georgia)
    v. Pathway analysis (Bart): https://docs.google.com/document/d/13PVBY0kZvUrLuj7TMPea_wzOUPI9Nz9uECYBB55iE64/edit#heading=h.nzhfau90fhw7
b. Fullerene datasets - to be done
c. ….**information added?**

4. **OpenTox** implementations
    a. New tools
        i. Image Processing: http://opentox.ntua.gr/mediawiki/index.php/Image_Processing
        ii. API extension: http://www.google.com/url?q=http%3A%2F%2Fopentox.ntua.gr%2Fmediawiki%2Findex.php%2FAlgorithm_Preprocessing&sa=D&sntz=1&usg=AFQjCNGrSwPjj_NNzs4foiWDh5BGjdyFXQ
        iii. Json representation of a Study Resource, based on current IDEA implementation: http://opentox.ntua.gr/mediawiki/index.php/JSON-Study
        iv. Extensions in RDF model representation: http://opentox.ntua.gr/mediawiki/index.php/RDF-Models
        v.
    b. Best QSAR models - to be done
    c. **information added?**
5. …..**information added?**
6. **information added?**

Completion condition: needs to be added

17. Validate Model

Primary Actor: Georgia, Cristian, Bart

Scope: Establish variability and significance of the model given the question (hypothesis) at hand

Brief: Validating the model can be interpreted in three ways, namely statistically, biologically and experimentally:

- Establish predictiveness of the model on external data (test set) & assign measures of accuracy for model's estimates.
- Then establish the biological significance of the findings given the relevant biological information available (biological databases).
- Repeat biological experiments in vitro/in vivo given the above information.

Trigger: Are the models created by data analysis sound and unbiased? QSAR models required for statistical validation & outcome of predictive model for biological validation.

Basic Flow

1. **Statistical validation** (following Data preparation & Data analysis - see case 14)
    a. Assessment of statistical fit
    b. Predictivity using cross validation fit
    c. Assess data quality
    d. Assess applicability of the model

How to do it? CV, bootstrapping, Y-randomization
What to check? R2, RMSE, Q2

2. **Biological validation** (following **Statistical** validation -see step 1 above)
    a. Pathway analysis (databases: GO, KEGG, NCBI,..)
        i. Pathway description
        ii. Pathway importance
        iii. Pathway integration
    b. Disease ranking
    c. Drug ranking
    d. Graphical representation
    How to do it? Combining knowledge available at the above databases via R or standalone software. What to check? Evidence to support the hypothesis (the state/ disease analyzed)
3. **Experimental** validation
    a. repeat the experiment to cross-check and update results
4. …. **information added?**

Completion condition: needs to be added

20. Create Template (xls etc)

Primary Actor: Researcher (also in literature research), Developer

Scope: Data processing

Brief: Templates for data where ISA-TAB format is not applicable or conversion of existing data to ISA-TAB is not reasonable. Harmonization of existing data to create machine read- and processable data.

Trigger: Need of processable, common or compatible structure for data input into a Data Store or Data Warehouse.

<u>Related use cases:</u>
19. Create template ISA-TAB

Basic Flow

1. Analysis of the data
2. Structure the existing data and add missing information (assay-wise) in Excelsheets
3. Create a Templatefile (JSON) to the Exceldatasheet that maps data fields. Matrix of the information that is in a specific Excelsheet.

Completion condition: needs to be added

21. Map engineered nanomaterial to biological pathway

Primary actor: Student who is asked to update or create pathways in WikiPathways with engineered nanomaterial information.

Scope: The inexperienced student is looking for an identifier that can be added to the pathway to cross-reference to any engineered nanomaterial database (preferably our own, I'm guessing). He also wants to receive information about structural information so he can add a graphical representation of the engineered nanomaterial in the pathway.

Trigger: An academic researcher wants to investigate the effects of engineered nanomaterials on biological systems, for which he/she is planning to use WikiPathways. He/she requires current pathways to be extended with engineered nanomaterial information or he/she wants to add completely new pathways in which engineered nanomaterials are involved. Because he/she is busy himself, he/she asks an inexperienced student to do the job for him/her.

Basic flow:

1.
    a. **Scenario 1:** A pathway already exists in WikiPathways, but recently a paper came out that suggested a role of engineered nanomaterial X in that pathway, which the researcher wants to study. He then wants to add the engineered nanomaterial to the pathway as is described in the paper.
    b. **Scenario 2:** A new biological process is discovered in which engineered nanomaterials play a huge role. Because the new process is not covered at all yet in WikiPathways, an academical researcher wants to add the new pathway including the engineered nanomaterials in question to WikiPathways so he can perform analyses.

Both scenarios now basically lead to step 2, both requiring engineered nanomaterials to be mapped in a pathway on WikiPathways. Scenario 2 might require a bit more biological context as a pathway has to be built from scratch, but it should be well-described in the paper.

2. The researcher of course knows quite a lot about the engineered nanomaterial in question, but he/she has no time for mapping stuff to WikiPathways, thus gets a student to do it for him/her.
3. The student is new to the subject and googles the engineered nanomaterial in question, possibly mistyping the name.
4. eNanoMapper database still pops up on top and somehow looks reliable (I think this is an important step – There is something to google search results that makes them look reliable or not, I'm not sure what it is, but this is something that needs to be considered

– BTW, are the ways in which google shows the result manipulatable by ourselves, or is this 100% determined by google?)

5. Upon click, the student is taken to an overview of the most important properties of the engineered nanomaterial.
6. A nice graphical representation (and structure formula – if possible) of the engineered nanomaterial is shown so the inexperienced student knows what he/she is looking at, making it less abstract. He/she can also use this information to draw a visualization of the engineered nanomaterial in the pathway.
7. The most important thing for the student to have by then is an identifier which he/she can add to the engineered nanomaterial in the pathway, which allows people to be linked back to the eNanoMapper database from inside the pathway.

Completion condition: needs to be added

23. Search for all knowledge of engineered nanomaterial X

Primary Actor: Researcher (industrial or academic), Regulator.

Scope: Find all general, as well as specific information available on one particular engineered nanomaterial – According to its name and/or detailed characteristics.

Brief: Classes of engineered nanomaterials have unique names - Specific engineered nanomaterials may, however often does not have a name, but is characterized through its many properties (characterization properties). A search for engineered nanomaterials cannot be based only on keywords, but has to include specifics like size distribution, dissolvability etc. Further, suggestions and a guided search process may be very useful - Results should be prioritized.

Trigger:

- An industrial researcher considers using an engineered nanomaterial for a product. He investigates the impact on humans and the environment (safety study)
- An academic researcher is integrating an engineered nanomaterial characterization into a cell membrane model.
- A regulator assesses the safety of an engineered nanomaterial, which company X wants to use in a product.

Basic Flow:

1. **Industrial Researcher**
    a. The researcher goes to google and enters the name, as well as a few characteristics of the engineered nanomaterial.
    b. Besides other resources, the eNanoMapper landing page shows in the results. He/she follows this link.
    c. He/she has two choices: (a) browse services and resources (b) search: basic, as well as advanced. Advanced search gives him/her a form for entering characterization details.
    d. He/she finds all data and information provided or linked by eNanoMapper either by browsing or in the search results.
2. **Academic Researcher**
    a. (Discuss with users)
3. **Regulator**
    a. Similar to Case 1

Completion condition: **needs to be added**

27. Search for/ Link related data sources

Primary Actor: Researchers, Automated scripts that search for related data, Legal consultant

Scope: Research state of the art, Experimental design, Compilation of database, Modelling

Brief: The following will search for/link data sources:
- Researchers that investigate the state of the art
- Researchers that design experiments
- Researchers that perform modelling and analytics on experimental data
- Researchers that compile database
- Automated scripts that search for related data

Trigger:
- Investigate the state of the art
- Experimental design
- Create models
- Compile database to use in-house or publish on web/in paper

Basic Flow:
1. Define search target (characteristics of desired data)
    a. Identify specific research question
        i. Group(s) of engineered nanomaterials
        ii. Toxicity aspect investigated
            1. Physical chemistry
            2. Ecotoxicity
            3. Environmental
            4. Toxicity
        iii. Character of data
            1. Experimental
                a. Conditions
                b. Parameters
            2. Modelling
                a. Algorithms
                b. Parameters
    b. Desired format(s)
2. Search for data sources
    a. Scopus
    b. Google Scholar
    c. Google
    d. Presentations
    e. Networking
3. Locate data source
4. Verify data

5. Identify characteristics
    a. Experimental
    b. In separate file/in text
    c. Data formats
6. Download data
7. Use
    a. Evaluate characteristics
        i. Experimental
        ii. Model
    b. Create Linked Data links
    c. Publish Database
    d. Design Experiment

Completion condition: needs to be added

28. Find data with similar patterns

Primary Actor: **needs to be added**

Scope: (this helps with experimental design - to support selection of engineered nanomaterials to test) **needs to be extended**

Brief: **needs to be added**

Trigger: **needs to be added**

Basic Flow:
Search engineered nanomaterial : (ZnO)

1. Show known categories (multiple) (could come from ontology only)
2. Click on each category
3. Show range/scale/possible values for the category (and what range the hit is in) - e.g. zeta potential (required database)
4. Filter - show all data of same type, but allow filtering / changing the scope to all engineered nanomaterials or filter by some parameter (require database)
5. Visualise : this engineered nanomaterial is in 20% top values; show real data in some nice web browser UI; spider plots; charts; interactive exploration

Another point of view

1. Starting from the property (as in NPO hierarchy)
2. How can I evaluate if property X has influence / is important for property (activity) Y (e.g. show all engineered nanomaterials in (selected) categories #3 & #6 (examples) which have high values of property A)

Completion condition: **needs to be added**

29. Save my data in format X

Primary Actor: Researcher (industry/ academic), Regulator, Student, Management/ Administration stuff.

Scope: Input, output, file format, database.

Brief: The primary actor has data in raw or standard file format obtained from various sources (e.g. laboratory experiment, instrumental measurements, scientific publication, data curation, database search, calculation output).

Trigger: He/she needs to store it in a target file format (X):

- To share with the scientific community
- To submit to a regulatory agency
- To upload it to a database
- To use it as a database query
- To use it as input to a software
- To make a predictive model (QSAR)
- To visualize/ represent data
- To curate the data
- To integrate with other data sources
- To deal with management/ administration supervision/activities.

Basic Flow:

1. Review available file formats (e.g. open access, industry standards etc) for the target (e.g. database, software or laboratory).
2. Choose optimal file format suitable for the actor's task. If not available develop your own file format or upgrade an existing one.
3. Choose software tool for formatting if available, if not write tailor-made converter (e.g. custom program or script).
4. Actual conversion/ saving the data in format X., plus documentation or supplementary information.
5. If needed for large data sources, prepare a script for serialization.

Completion condition: needs to be added

33. What URI to use for X

Primary Actor: Ulrich R. Ivory is a chemometrician at an experimental biology group taking care of the NanoQSAR development.

Scope: The user is looking at a data set provided to him in a spreadsheet using one of the NanoSafety Cluster spreadsheet templates.

Brief: The user is looking at one of the columns in the spreadsheet and his/her eye falls on one of the terms and the associated Internationalizable Resource Identifiers (IRI) and recognizes it to be part of the eNanoMapper ontology. He ponders about the exact meaning of that term and why they did not take a related term instead, which seems to him/her to be a better match. Might there be a specific ontological reason that IRI was chosen?

Trigger:

- It starts with a term in some document and a matching a specific ontology term (via an identifier or IRI)
- Both the document and the ontology are available for reference
- The initial objective is to obtain understand why that IRI was chosen and what implications that has

Basic Flow:

1. Take one or more ontology term identifiers or IRIs
2. Look up appropriate ontologies
3. List the labels, definitions and other primary information about the ontology terms
4. Derive ontological implications such as implied classifications and others properties that may be derived via reasoning
5. Optionally, compare two or more IRI with each other, based on the information aggregated in steps 3 and 4

Completion condition: Can the user find the appropriate ontological information, including implied properties?

34. Names for Ontology Terms

Primary Actor: Ontology user (may be end-user, data provider, data curator, or partner)

Scope: User; Research; Sharing; Data Exchange

Brief: Ontologies act as synonym providers for terminology, while providing semantics-free IDs that can be used in data annotation. The user provides a term and obtains a list of alternative names for that term, one of which may be regarded as 'primary'.

Trigger: User has ID or name for ontology term and the objective of finding alternative names.

Basic Flow:

1. Input: An ontology term. This may be specified as either an ID or a label (text).
2. Processing: There are alternative interfaces to search in the ontology:
    a. In a repository such as BioPortal (http://bioportal.bioontology.org). The user navigates to the ontology and enters the search term in the search box.
    b. In a desktop application such as Protege (http://protege.stanford.edu/). The user should install Protege, open the application (from http://purl.enanomapper.org/onto/enanomapper.owl), and perform the search in the search box.
3. Selection: The closest matches will be displayed in the drop-down list. The user selects one and views the metadata on the screen of whichever tool (from step 2).
4. Validation: The user should check that the associated metadata (synonyms and ID) matches in meaning the input term. This cannot be done automatically.

Completion condition: When the use case is fulfilled there will be a list of alternative names for the ontology term provided as input, and importantly a primary identifier from the ontology that can be used in data annotation. If nothing is found in the ontology for the given search term, it is likely that the user may wish to request that this term be added to the ontology.

36. Harmonize Terminology

Primary Actor: User (may be end-user, data provider, data curator, or partner)

Scope: User; Research; Sharing; Data Exchange

Brief: Terminology differs in use between different labs, projects and data providers. The simple use of different terminology (even though it means the same thing) hinders discoverability and challenges searching. Thus it is desirable to harmonize the terminology by creating a mapping as synonym sets (which can be added to the ontology) and wherever possible adopting a single vocabulary.

Trigger: There are at least two resources with differing terminology used. This input is abstracted into (at least two) list of terms. For simplicity we will assume that mapping proceeds pairwise (i.e. two lists at a time).

Basic Flow:

1. Input: For each term in list A,
2. Processing: Select the closest match in term B.
3. If it is a good match (i.e. the terms are “true synonyms”), continue to next term.
4. If it is only a partial match, choose which of list A or B is broader in meaning, e.g. “assay” vs. “biological assay”. Can the broader term stand as a parent (is-a) to the narrower term? If yes, mark the direction of relatedness.
5. Are they close siblings rather than broader or narrower? E.g. IC50 and IC90. If yes, they cannot be inter substituted but share a parent. Identify the shared parent.
6. If there are no close matches for a term in term A, add to list of gaps.
7. When you are finished with list A, continue with the unmapped terms from B.
8. When finished with both lists, you have a list of mappings which should be added to the ontology as synonyms or relationships.
9. A further step might be to alter the terminology used in the source to be consistent as far as possible.

Completion condition: When the use case is fulfilled there will be a list of mappings between the different input terminologies also available as metadata in the ontology. The mappings may be of the types described above: “good”, “related-parent”, “related-sibling” and there may be a remaining set with no matches which may be targeted for inclusion into the ontology. There are usually some gaps. In severe cases there may be no mappings at all.

38. Find Ontological Contradictions

Primary Actor: Ontology developer, Ontology user (e.g. database developer)

Scope: Development, Curation

Brief: When extending the ontology it may be possible to add new knowledge that contradicts something in the remainder of the ontology. For example, if the ontology contains an axiom claiming that nanoparticles are always smaller than 100nm, but then a particle is added with a larger size. In this case it should be possible to detect the inconsistency in the ontology.

Trigger: New knowledge or content is added to the ontology, possibly as a side effect of data annotation.

Basic Flow:

1. Add term and properties to ontology.
2. Perform ontology reasoning using a reasoner (either programmatically in the OWL API -- http://owlapi.sourceforge.net/ -- or manually in Protege -- http://protege.stanford.edu/).
3. Check for inconsistent classes. Using the OWL API you will have to extract these from the reasoner output programmatically. In Protege they will appear in red and classified beneath the 'Nothing' class.
4. Note that inconsistent classes are evidence of a contradiction, but one contradiction does not necessarily lead to one inconsistent class. In fact, it is more common that a contradiction creates several inconsistent classes.
5. The underlying contradiction must now be diagnosed and repaired. This can be a complex affair but justifications (in the "explanation" plugin in Protege) can help.
6. The contradiction should be fixed by amending the relevant definition or fixing incorrect class data.

Completion condition: A list of inconsistent classes, or a "clean" result showing there are no inconsistencies in the ontology.
The reasoner might crash or never return. In this case try another reasoner. If none will classify your ontology, you will have to diagnose the problem the hard way by iteratively removing content.

39. Computer-aided engineered nanomaterial design

Primary Actor: Researcher (working in academia, research institute or private sector) seeking the design of an engineered nanomaterial that will serve a specific function (i.e. drug transfer, coatings)

Scope: Improve properties of existing engineered nanomaterials/ Substitute existing engineered nanomaterials/ Calculate properties (incl. safety) for new given structures/ Use results to drive engineered nanomaterials design

Brief:
Design of engineered nanomaterials will be carried out by researchers in universities, research institutions or for private companies. They design modifications to existing engineered nanomaterials, design new engineered nanomaterials or assess the behavior of existing ones.

Trigger:

- The goal could be to
    - **improve behavior** of **existing** engineered nanomaterial(s) (reduce toxicity)
    - **substitute** engineered nanomaterial(s) in an **existing function**
    - **seek design** that achieves a **new function**
    - **explore the behavior** of a **new** given engineered nanomaterials (incl. assess safety) to decide whether to keep or redesign

Basic Flow:

1. Identify function for which we want to enhance/reduce
2. Explore substances that achieve similar functionality
3. Identify structures critical to this functionality
4. Design structure: based on existing structures, design new structures that
    a. improve existing engineered nanomaterials
    b. substitute existing engineered nanomaterials
    c. for new function
5. Input structure parameters into eNM in order to calculate the properties of designed engineered nanomaterials (in the case of new nanomaterials users could start here), both in terms of safety and in relation to its end use.
6. Laboratory results (where applicable)
7. If successful, Ok otherwise: Return to Step #2

Completion condition: Finding a design that meets requirements of functionality within limits of safety aspects

42. Verify data against regular requirements

Primary Actor: Simon M. Escher works at a company making outdoor clothes and responsible for the innovation track.

Scope: The user got a fact sheet about a new coating spray for outdoor clothes that would improve the water-resistance and air-permeability of their current products, and, in particular, the products can now be washed without loosing these properties. But he is not sure the engineered nanomaterials in this spray will be acceptable to be used in their products.

Brief: He already looked up a lot of regulation related to this kind of sprays and their engineered nanomaterials in it. The legal department has been looking into the laws, regulations, and guidance documents, but is unsure if enough is known about the engineered nanomaterial to be accepted for regulatory assessment. More specifically, they want to clarify which regulatory rules apply to this material, and thus to their company.

Trigger:

- A new industrial product that may need regulatory assessment

Basic Flow:

1. TBD **needs to be extended**

Completion condition: **needs to be added**

44. Find producers of engineered nanomaterials

Primary Actor: Study Designer, Researcher or Lab Personnel

Scope: Finding possible producers for an engineered nanomaterial (Special case of 23)

Brief: In order to conduct a study, certain quantities of engineered nanomaterial must be available, in a very well characterized form. If the study aims for the safety of an engineered nanomaterial used in a commercial product, the engineered nanomaterial must be produced in exactly the same way by the same provider. Objective of this use case is, to find all information for possible producers of engineered nanomaterials in order to select one.

Trigger: Decision to use an engineered nanomaterial either for academic purposes or within a product. Need to find the best producer for the material.

Basic Flow:

1. Via keyword search, or via a keyword hierarchy ontology, he specifies the engineered nanomaterial.
2. The researcher finds a list of producers for a engineered nanomateriall. He also finds lists of producers of related engineered nanomaterials.
3. Some basic associated information may be online. Contact information of producer is shown.

Completion condition: One or more producers that meet Primary Actor's criteria have been found.

45. Find quality control data of engineered nanomaterials

Primary Actor: Study Designer, Researcher or Lab Personnel

Scope: Finding quality control data for one particular engineered nanomaterial and/or similar engineered nanomaterial. (Follow up case of 44)

Brief: An engineered nanomaterial has been chosen and candidate producers are being reviewed:

- To select one or more as suppliers for new experiments.
    - Main objective: scientific quality / reproducibility of new experiments.
- To decide whether to accept the experiments with engineered nanomaterial can be used/joined (validity of data, usefulness for scientific conclusions etc).
- To assess the safety of a nanomaterial for use in a product

Trigger:

- Decision on which engineered nanomaterial to use in a study
- Decision to use data for modelling, scientific conclusions etc
- Decision to use an engineered nanomaterial in a product

Basic Flow:

1. *(continues from Use Case 44)*
2. We request all quality control information from the vendor.
3. Via eNanoMapper, nanomaterialregistry.org, google and other resources, we try to find publications and data sources that use the same engineered nanomaterial in order to use the measurements from experiments as independent quality control data to verify the vendor's data.
4. A decision on whether quality control is sufficient is made.

Completion condition: The quality control data of engineered nanomaterials have identified the producers that can be selected for a purpose.

49. Mapping engineered nanomaterial in database A to B

Primary Actor: The primary actor for this use case can vary, as this use case seem to be very general: given a particular engineered nanomaterial in some database, is this engineered nanomaterial listed in other databases too.

Scope: The users starts with an engineered nanomaterial in database A and wants to find the same engineered nanomaterial in database B based on some equivalence criterion.

Brief: The outcome of this task is a statement if there is an exact or close match. If one database already links to the other database, this statement has already been made explicitly. If not, a similarity search may be needed. This search can have multiple reasons, but a general similarity search is outside the scope: instead of looking for the closest match, the use case wants to establish a statement that two engineered nanomaterials are actually the same. However, a similarity search may assist the user in making that claim.

Trigger: A engineered nanomaterial in database A, with a given set of physicochemical properties

Basic Flow:

1. Record the physicochemical properties in database A
2. Use these properties to search engineered nanomaterials with the same characteristics (this does not imply an integrated, automated process; supporting manual searching is acceptable)
3. Compare the properties of the engineered nanomaterial from database A with search hits in database B
4. Establish a claim if there database B has the same engineered nanomaterial as found in database A

Completion condition: The final outcome is a claim about the equivalence of two engineered nanomaterials, one from database A, one from database B. This claim can be positive and negative. Additionally, the outcome of this use case is a list of search hits based on the properties of the engineered nanomaterial in the first database. This list must present at least one engineered nanomaterial from database B, possibly a dissimilar one, but at least it being the closest match. Furthermore, this list must enable the ability of the user to compare the found materials with the search material.

50. Finding Databases with “Supplementary Information”

Primary Actor: Bjoern Koudwater

Scope: The user wants to compare his own research results with those in a recent ACS Nano paper, but does not want to extract the data points from that hard to read Figure 4.

Brief: The user wants to know if data from a particular paper can be downloaded from online databases. He doesn't want to try to figure out what the data behind a particular figure is. He already checked the paper's Supplementary Information but that is a human-oriented PDF too and doesn't read well in Excel. He also emailed the author but got no response. But surely some of the data must be deposited in public databases too, under the H2020 Open Data pilot??

Trigger:

- The user reads a paper and wants to compare the data with his/her own experimental results
- The user starts looking in online and/or public databases to see if those have data from that paper and tries the title and the DOI

Basic Flow:

1. Have a research paper
2. Search the paper in a central indexing server
3. List all databases that have data from that paper
4. Filter the list based on the type of data, or based on a Figure number

Completion condition: Various outcomes can be measured, including the number of indexed databases, the number of papers from which data is included in all databases, and the time it takes for the user to download the data from that paper.

52. Register and get access (public)

Primary Actor: Researcher, Regulator, any Person

Scope: Establish a GUI where users can create or apply for an account.

Brief: A random person has to be able to register at the eNanoMapper website, verified via eMail address. eNanoMapper may invite users to access the resources or a GUI.

Trigger: Need of controlled access to resources. Enable persons to become eNanoMapper (public) user.

Basic Flow:

1. User wants access
    a. Person registers on GUI for an account
    b. Optional: eNanoMapper approves the account
    c. New user accepts its account (email) and sets a password
        i. Use username and password to login
        ii. Use username and password (to receive token) to get access to resources
2. eNanoMapper invites user
    a. eNanoMapper creates account or invites user to visit registration form

Completion condition: **needs to be added**

53. Interterminology Mapping

Primary Actor: Richard DeDerde is responsible for making their data platform scheme interoperable with the eNanoMapper ontology.

Scope: The user has a list of terms coming from their scheme to capture data used in their platform, such as “POUR DENSITY”, “DATA_GATHERING_INSTRUMENTS”, “Method name”, etc.

Brief: The user wants to make their data platform interoperable with the eNanoMapper ontology. In order to do that they take the names of the fields in their data scheme, together with their knowledge about the data contained in their database, and look up the best terms in the ontology to associate with that field name. Examples of field names that need to be mapped to ontology terms include domain-specific terminology such as “pour density” and more generic data organisation terminology such as “method name” or “comment”.

Trigger:

- It starts with a list of field names from a data schema targeted for interoperability.
- The initial objective is to obtain “community” ontology identifiers for every field in the scheme.
    - Best is one identifier per field
    - In many cases it is not as straightforward and meta-data about the mapping is also needed e.g. comments, different levels of accuracy.
- The input data is:
    - A liist of fields (may include other contextual information)
    - The ontology with labels, definitions and identifiers
    - Knowledge about the data scheme and the data contained in it

Basic Flow:

1. Look at term from field list
2. Search in ontology for similar term
3. Use knowledge to select best match <some magic happens>
4. Capture result in spreadsheet again as ID (URI) along with any comments about the selection process
5. Repeat until all terms in the field list are done
6. Make a report showing the mappings, including:
    a. Original field with label and definition
    b. Mapped ID with label and definition
    c. Ontological implications following the ontology mapped to

Completion condition: The case is signed off when Richard is happy about the report. Acceptance of the mapping is based on user opinion and may not ontological be (fully) correct: the aim is a “best match”, not the ultimate match.

54. Ontology Data Annotation

Primary Actor: Mary Carlisle is an experimental biologist who ran a number of cell line experiments and wants to ensure others understand her data.

Scope: The user has data and metadata resulting from and describing her experiments. Experimental details include the cell line she used, annotation of experimental conditions (serum, etc), and measurements include units.

Brief: The user wants to make their data interoperable with other data from the NanoSafety Cluster. She already used a community-provided template using the eNanoMapper ontology but the data she enters also contains domain terminology. This too, she wants associated with ontology terms as to make her results as easily findable as possible. She knows from literature that data access will (statistically) increase the citation to the corresponding research paper. Examples of data that needs to be mapped to ontology terms include the HUVEC cell line, the non-SI units she has been using (she's from an English city in the North-East of England), etc.

Trigger:

- It starts with experimental data, possibly directly from a measurement device
- The initial objective is to obtain “community” ontology identifiers for every term used in the data
    - Best is one identifier per field
    - In some cases it is not as straightforward and meta-data about the mapping may be provided, e.g. comments, different levels of accuracy.
- The input data is:
    - List of data units, experimental features, factor values, etc
    - The ontology with labels, definitions and identifiers
    - Knowledge about the data scheme and the data contained in it

Basic Flow:

1. Look at term from the data
2. Search in ontology for similar term
3. Use knowledge to select best match <some magic happens>
4. Capture result in spreadsheet again as ID (URI) along with any comments about the selection process
5. Repeat until all terms in the data are done

Completion condition: The case is signed off when Mary is happy about the report and is able to find her data back using the annotations she added.

55. Experimental Design

Primary Actor: Biologists working on Nanoparticle Biosynthesis

Scope: Investigate how engineered nanoparticle properties are influenced by the variation of specific process parameters.

Brief: Biologists in a pharmaceutical company are working on the optimal design of a specific engineered nanoparticle which has been extensively investigated as a promising drug delivery system for controlling the release of therapeutic agents. They need to investigate how the different properties of engineered nanoparticles may vary and what their optimal set are. They need to define the statistical methodology, the types or experiments that will take place, and how their results will be validated.

Trigger: A particular experimental factor is under study because of special laboratory interest. Scientists need to estimate how its behaviour is influenced by other parameters, and what is their optimal set for the problem under study.

Basic Flow:

1. Define the problem, e.g. determine the combined effect of some variable on the extracellular biosynthesis of silver nanoparticles.
2. Decide which are the variables/ parameters/ factors that are of special interest and are directly involved in the study (quantitative, qualitative or both). Explicitly define the experimental responses that need to be optimized.
3. Decide on the different levels/ experimental conditions that the above variables/ parameters/ factors will be studied.
4. Decide on the statistical methodology used and the details of the design, i.e. how many runs of the experiment with how many replicates for each set of conditions. E.g. composite experimental design, ANOVA results with F/t test p-values (coefficient of determination) being the measures looked upon for statistical significance.
5. Run the experiments.
6. Decide which software to use & run the statistical calculations.
7. Illustrate results, e.g. Pareto charts to illustrate variables' degree of significance, 3D surface plots to illustrate the relationship between response and experimental levels of the variables utilized in the study.
8. Experimental verification of the model & compare with the predicted data.

Completion condition: Satisfactory/ acceptable experimental verification.

As a reference, see: http://www.ncbi.nlm.nih.gov/pubmed/24390838