DELIVERABLE D5.4

User application for importing NanoSafety Cluster data

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:	KI
DATE:	30.7.2015
VERSION:	V2

A Database and Ontology Framework for Nanomaterials Design and Safety Assessment

Call identifier	FP7-NMP-2013-SMALL-7
Document Type	Deliverable Report
WP/Task	WP5 – Tasks 5.3 and 5.6
Document ID	eNanoMapper D5.4
Status	Final

Partner Organisations	• Douglas Connect, GmbH (DC) • National Technical University of Athens (NTUA) • In Silico Toxicology (IST) • Ideaconsult (IDEA) • Karolinska Institutet (KI) • Technical Research Centre of Finland (VTT) • European Bioinformatics Institute (EMBL-EBI) • Maastricht University (UM) • Misvik Biology (MB)
Authors	Nina Jeliazkova (IDEA), Nikolay Kochev (IDEA), Vedrin Jeliazkov (IDEA) Egon Willighagen (UM), Lucian Farcal (DC) Final review and editing by Barry Hardy (DC)
Purpose of the Document	To report the outcomes of Task 5.6 on importing NanoSafety Cluster data
Document History	First draft 2015-04-29 Second draft 2015-07-29 Final version 2015-07-30

TABLE OF CONTENTS

TABLE OF FIGURES

GLOSSARY

Abbreviation / acronym	Description
CC	Creative Commons
EC	European Commission
ENM	eNanoMapper – A Database and Ontology Framework for Nanomaterials Design and Safety Assessment
ENPRA	'Risk Assessment Of Engineered Nanoparticles'
JRC	Joint Research Centre
JSON	JavaScript Object Notation
NANoPUZZLES	'Nanoparticles safety collaboration platform'
NANoREG	'A common European approach to the regulatory testing of Manufactured Nanomaterials'
NM	Nanomaterial
NSC	NanoSafety Cluster
MARINA	'Managing Risks of Nanomaterials'
MODENA	'Modelling Nanomaterial Toxicity'
ModNanoTox	'Modelling nanoparticle toxicity: principles, methods, novel approaches'
OWL	Web Ontology Language

1. SUMMARY

The aim of this work is to analyze and choose technical options to implement application and user interfaces for the import of available data from various EU Nanosafety Cluster (EU NSC) projects. Applications are subjected to integrated testing (Task 5.2) and, depending on licenses, data are released for user testing and evaluation (Task 1.4).

eNanoMapper has been in contact with various NanoSafety Cluster projects for importing data. Most of this data is expected to be provided as spreadsheets for which a general framework has been developed. We engaged with the following EU Nanosafety Cluster projects:

1. With **NANoREG (WP6)** we are working on exporting data from their literature capturing database into eNanoMapper and back again, and annotating fields and content with the eNanoMapper ontology along the way. NANoREG also provided confidential Excel templates used to gather data within **NANoREG WP2, WP4 and WP5**.

2. **ModNanoTox** was provided as MS Excel spreadsheet files, containing 86 assays with more than 100 different endpoints affecting 45 species. We developed a dedicated command line application, built on top of the configurable spreadsheet data parser, which enables parsing the entire ModNanoTox data and importing into the database. The annotation of fields and content with the eNanoMapper ontology is ongoing.

3. **COST Action MODENA** agreed to share data with eNanoMapper and provided the Excel files to eNanoMapper. The data was collected for training purposes. This data was first reviewed by MODENA and provided to eNanoMapper as an Excel spreadsheet, containing information on the nanomaterials (NM) identification, characterization, assay performed, test conditions and a few toxicity parameters (ECs). Additionally, MODENA provided Excel templates used within the project to gather literature data (one Excel file per material). These individual templates are then summarized into the spreadsheet for training purposes described above. The MODENA templates are similar (but not exactly the same) to MARINA templates.

4. **MARINA** project has published an exhaustive paper on the *in vitro* assessment using representative JRC nanomaterials. Data to be imported refers to the *in vitro* toxicity of six NMs and was evaluated in 12 cellular models representing 6 different target organs/systems (immune system, respiratory system, gastrointestinal system, reproductive organs, kidney and embryonic tissues). The toxicity assessment was conducted using 10 different assays for cytotoxicity, embryotoxicity, epithelial integrity, cytokine secretion and oxidative stress. In addition, **MARINA** participant provided confidential Excel templates used to gather the data within the projects, covering variety of physicochemical characterisation experiments and bioassays.

5. **NANoPUZZLES** project provided confidential data templates, designed to resemble ISA-TAB-Nano structure. NANoPUZZLES plans to release the data with CC-BY license after the end of the project.

6. Exhaustive data on physico-chemical properties of **representative nanomaterial**s used in various EU Nanosafety Cluster projects (e.g. ENPRA, MARINA, and NANoREG) are publicly available via the EC-JRC reports, (https://ec.europa.eu/jrc/en/publications).

2. INTRODUCTION

Deliverable 5.4 describe the achievements within the Task 5.6 (User application for importing NanoSafety Cluster data). Its main objective is to analyze and choose technical options, implement application and create user interfaces for importing nanosafety data from various sources. Accordingly, the information provided by different NSC projects is imported in the eNanoMapper database. Initially, a few NSC projects (e.g. NANOSOLUTIONS) or the EU nano database (JRC NanoHUB) were mentioned as possible sources of data to create such an application. However, in the first 18 months of the project, eNanoMapper initiated collaborations with different other projects within the EU NanoSafety Cluster that could provide multiple sets of data for this purpose: e.g. with NANoREG, ModNanoTox, MARINA and NANoPUZZLES and MODENA. The use of different sources implied also the use of different templates which were developed by individual projects.

Therefore, we present below the eNanoMapper approach for importing data, the user applications for data import and the associated testing system.

3. ENANOMAPPER IMPORT

The majority of the EU NanoSafety Cluster projects use spreadsheet templates as the preferred approach for data entry. Therefore, we developed a configurable spreadsheet parser facilitating user friendly data preparation and upload.

3.1 STRUCTURE OF THE IMPORTER

The importer is written in Java as a file reader extending AMBIT's raw reader, "IRawReader" [2]. The API allows the eNanoMapper platform, which is implemented by the IDEA partner in eNanoMapper as an extension of the AMBIT platform, to read these files into the system, and implements mapping of the spreadsheet content to internal concepts, which is outlined in the next section.

The source code of this reader is found on GitHub at https://github.com/enanomapper/nmdataparser with the name GenericExcelParser in the Java package net.enanomapper.parser. This project was initiated with the goal of parsing NanoSafety cluster templates, but also allows import of chemical substances composition and associated experimental data. The project has a Maven build file that details how the software is compiled. Minimal testing is also provided by this nmdataparser project, in the net.enanomapper.parser.test package. This JUnit-based test suite tests that it extracts the basic structure and that measurements and references are read into the eNanoMapper system.

An extensive JUnit test suite, which includes actual import into a test database and verifying the database content against expected values is available via private repository https://bitbucket.org/vedina/parserconfig. This repository is not publicly available as it contains a mix of public and confidential files provided by the NanoSafety cluster.

3.2 MAPPING OF THE INPUT DATA TO THE ENANOMAPPER DATA STRUCTURE

The parser enables import of the data, stored in the supported set of spreadsheet templates, and accommodates different row-based, column-based or mixed organizations of the data. The parser configuration is defined in a separate JSON file, mapping the custom spreadsheet structure into the internal eNanoMapper storage components: **Substance**, **Protocol**, **Measurement**, **Parameters,** and **Conditions**. The JSON configuration syntax includes a set of keywords, specifying different strategies for reading the data from one or several sheets, as well as allowing combination of the excel structures (sheets, rows, columns, blocks of cells and cells) into the eNanoMapper data model. The parser code, the JSON syntax, documentation and example files are available at https://github.com/enanomapper/nmdataparser/. The mapping enables a uniform approach towards import, storage and searching of the ENM physicochemical measurements and biological assay results. While the parser itself is open source, the configuration files may not be, thus not revealing the organisation of confidential data templates. The parser is currently being used to parse ModNanoTox, MARINA and MODENA spreadsheets and confidential templates from EU NanoSafety Cluster projects. Maps of the confidential spreadsheet templates are available on request, in compliance with the agreements between the corresponding projects. More formats will be supported as needed for indexing data from different sources. The development of ISA-Tab-Nano import and export is ongoing.

3.3 PARSING DATA FROM FP7-NANOREG

3.3.1. DATA DESCRIPTION

NANoREG WP2 has provided several spreadsheet templates to eNanoMapper (see Annex). The templates contain no data or only data for demonstration purposes. NANoREG reportedly develops these templates in order to gather experimental data generated in the project. NANoREG has a separate activity for collecting literature data, which uses a custom relational database schema. The description of this schema has been made available by WP6, allowing a collaboration with eNanoMapper to map the schema table fields to the eNanoMapper ontology (UM). This process is ongoing and which will lead to a mapping report that compares for each mapping the information from the NANoREG database schema with the label and description extracted from the eNanoMapper ontology. Figure 1 shows an example of a draft of the report that was discussed in July. The input is a shared Google spreadsheet with the field names from the NANoREG database (along with indicators about the status of that field in the schema: release state (final or not) and importance) and a column with the ontology identifier. The ENM ontology label and description is programmatically retrieved from the ontology's OWL file. NANoREG WP6 anticipates that the actual data will be openly available in the future, but currently it is not. Therefore, the remainder of this section focuses on the data from WP2.

CAS

    Essential: YES
    Final: YES

    CAS registry number (CHEMINF)
    Description: Identifier used by the Chemical Abstracts Service database.

    ☐ Good ☐ Needs discussion ☐ Bad
    Comments:

Figure 1: Section of the mapping report linking the NANoREG WP6 database schema for nanotoxicity literature to the eNanoMapper ontology.

The header of each mapping section is the field name in the database schema, which is followed with some metadata (see text). Using the URI of the mapped term, programmatically the label and description are looked up in the ontology and depicted. The following boxes give room for NANoREG to review the mapping, in preparation for further discussion.

3.3.2. MAPPING OF INPUT DATA TO ENANOMAPPER DATA INFRASTRUCTURE

The mapping to the eNanoMapper data model is performed via JSON configuration file. Annotation of the particle and assay types is typically introduced as additional column(s) or in the configuration files. Further annotation for data values will be done once imported in the database.

3.3.3. DATA CLEAN UP

The NM-Series identifier is written in inconsistent (with or without dashes, spaces, additional information, *e.g.* NM200 instead of NM-200). We add additional column with the correct spelling.

3.3.4. RESULTING DATASET IN ENANOMAPPER

See Annex I.

3.4. PARSING DATA FROM FP7-MODNANOTOX

3.4.1. DATA DESCRIPTION

The ModNanoTox EU FP7 project (http://www.birmingham.ac.uk/generic/modnanotox/index.aspx) has produced a survey and selection of relevant physicochemical properties to use towards building a range of descriptors of engineered nanoparticles (mainly metal-based) and their potential toxicity. This dataset nicely demonstrates the complexity of the nanosafety domain. The ModNanoTox database provides physicochemical descriptors and toxic activities of nanoparticles from several studies. The database version from August 2013 includes 86 assays with more than 100 different endpoints affecting 45 species.

Unfortunately, only a few nanoparticles (usually fewer than three) have been tested for each endpoint. Physicochemical descriptors for the characterisation of nanoparticles are incomplete as well (about 75 % missing values). The two most comprehensive species in the dataset are *Daphnia magna* (water flea) and *Danio rerio* (zebrafish), with 34 and 14 assays each. The best represented endpoint for *Daphnia* is *Mortality*, and we were able to extract about forty *LC50* and sixty *% survival*' data entries. In both cases the number of measured nanoparticle properties was very low. Most studies report only two to four different nanoparticle properties (descriptors) and the descriptor types are very inconsistent (overall 36 different descriptors, which results in very sparse matrices with a high number of missing values).

The ModNanoTox data set was provided as an MS-Excel spreadsheet file. It consists of four sheets describing respectively (i) investigation study details; (ii) particle details and physicochemical properties; (iii) assay protocol description and (iv) assay measurement outcomes. The information in all sheets is organized as a sequence of dynamic blocks of data, each one containing a variable number of rows.

3.4.2. MAPPING OF INPUT DATA TO ENANOMAPPER DATA INFRASTRUCTURE

The configurable spreadsheet parser described was extended to support the recognition of blocks and the synchronization between blocks within the four sheets. The next step is to divide the data in each block into groups and sub-groups and match them across the sheets. This last operation is implemented by a dedicated command line application, built on top of the configurable data parser and allowing parsing of the entire ModNanoTox complex organisation into the internal eNanoMapper data model.

3.4.3. DATA CLEAN UP

- Added identifiers, allowing to automatically match the content between Sheet3 (Assay details) with Sheet4 (Study outcomes)
- Fixed misspelled reference code
- Fixed mismatched particle codes between Sheet 2(Particle details) and Sheet 4(Study outcomes)
- Clean up of endpoint names and units, which are stored in several columns in different fashion for different records (Sheet 2 and Sheet 4)
- Assay annotation with ontology entries

3.4.4. RESULTING DATASET IN ENANOMAPPER

The resulting data set loaded into eNanoMapper is described in this section. We here show a few screenshots of aspects of the ModNanoTox data as exposed by the AMBIT platform.
A material is stored in the database as a substance, and as substances are in AMBIT, it has general information, a composition and properties. The toxicity data is shown in Eco Tox and Tox tabs. A screenshot of a gold nanoparticle displaying the ecotoxicity data is shown next.
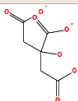
Search ▼   Nanomaterials ▼   OpenTox ▼   Demo ▼   Help ▼

Home › All assessments › This assessment › Versions ›

| Assessment identifier | Collect structures | Endpoint data used | Assessment details | Report |

Title ❶ :	ModNanoTox database
Maintainer ❶ :	eNanoMapper
Purpose ❶ :	The ModNanoTox EU FP7 project has produced a survey and selection of relevant physicochemical properties to use towards building a range of descriptors of engineered nanoparticles (mainly metal-based) and their potential toxicity.
Version ❶ :	1
Version start date ❶ :	26.07.2015
Version last modified on ❶ :	26.07.2015
Status ❶ :	◉ Draft ◯ Published ◯ Archived
Source ❶ :	http://www.birmingham.ac.uk/generic/modnanotox/index.aspx
Source URL ❶ :	ModNanoTox
License or Waiver ❶ :	http://www.birmingham.ac.uk/generic/modnanotox
Rights holder ❶ :	ModNanoTox
Rating ❶ :	★ ★ ★ ★ ★ ☆ ☆ ☆ ☆ ☆
Identifier ❶ :	00000000-0000-0000-0000-000000000004

Search ▼   Nanomaterials ▼   OpenTox ▼   Demo ▼   Help ▼

🏠 ⟩ Substance ⟩ XLSX-ba0dcfe6-41b3-3cc7-a800-e5b538852d8c ⟩ Study

| IUC Substance | Composition | Eco Tox (2) | P-Chem (3) | Tox (1) |

Filter…   Expand all Collapse all

AgNP

6.1.1 Short-term toxicity to fish (1)

test medium	Organism	Meas. Conc.	exposure	Endpoint		effect	N	p	Guideline	Owner	UUID
Carbon filtered water + "Instant Ocean Synthetic Sea Salt"	Danio rerio	0 µg l-1	24 hrs-96hrs	Tail Detachment and presence of edema	-	Development and Morta	30	N/A	Manual counting / Non linear regression / Probit analysis	ModNanoTox	XLSX-b8c21d4...
		0 µg l-1	96 h	Mortality	0	Mortality	30	N/A			
		10 µg l-1	96 h	Mortality	0	Mortality	30	N/A			
		25 µg l-1	96 h	Mortality	0	Mortality	30	N/A			
		50 µg l-1	96 h	Mortality	0	Mortality	30	N/A			
		100 µg l-1	96 h	Mortality	20.033	Mortality	30	N/A			
		125 µg l-1	96 h	Mortality	43.514	Mortality	30	N/A			
		Total Ag	96 h	LC50	128.54 µg l-1	Mortality	30	N/A			
		Total Ag	96 h	EC50	128.4 µg l-1	Mortality	30	N/A			
		0 µg l-1	96 h	Hatching Rate	100	Hatching	30	N/A			
		10 µg l-1	96 h	Hatching Rate	100	Hatching	30	N/A			
		25 µg l-1	96 h	Hatching Rate	100	Hatching	30	N/A			
		50 µg l-1	96 h	Hatching Rate	100	Hatching	30	N/A			
		100 µg l-1	96 h	Hatching Rate	70.07	Hatching	30	N/A			
		125 µg l-1	96 h	Hatching Rate	33.053	Hatching	30	N/A			

Showing 1 study(s) (1 to 1)   ◀ Previous Next ▶

6.1.3 Short-term toxicity to aquatic invertebrates (1)

test medium	Organism	Meas. Conc.	exposure	Endpoint		effect	Calculation Details	N	p	Guideline	Owner	UUID
ASTM	Daphnia magna	Total Ag	24 h	LC50	11.41	Immobilization		5	N/A	Manual counting / Non linear regression / Probit analysis	ModNanoTox	XL...
		Total Ag	48 h	LC50	11.02	Immobilization		5	N/A			
		Total Ag	24 h	LC50	81.84	Immobilization		5	N/A			
		Total Ag	48 h	LC50	72	Immobilization		5	N/A			
		Total Ag	24 h	EC50	10.2 µg l-1	Immobilization	Non-lin	5	N/A			
		0 µg l-1	24 h	% Feeding Rates	99.59 %	Feeding Inhibition (cells/ml/Daph		6	>0.05			
		5 µg l-1	24 h	% Feeding Rates	87.69 %	Feeding Inhibition (cells/ml/Daph		6	>0.05			
		10 µg l-1	24 h	% Feeding Rates	71.24 %	Feeding Inhibition (cells/ml/Daph		6	<0.05			
		20 µg l-1	24 h	EC50	32.43 %	Feeding Inhibition (cells/ml/Daph		6	<0.05			
		Total Ag	24 h	Reproduction Response	13.64 µg l-1	Feeding Inhibition (cells/ml/Daph	Non-lin	6	N/A			
		0 µg l-1	21 d	Mean number of neonates	96.1 % difference between th	Reproduction (neonates produces/fe	Total nu	10	0.004			
		0 µg l-1	21 d	Mean number of neonates	104.07	Reproduction (neonates produces/fe		10	>0.05			
		0.5 µg l-1	21 d		110	Reproduction (neonates produces/fe		10	>0.05			
		1 µg l-1	21 d	Mean number of neonates	57.59	Reproduction (neonates produces/fe		10	<0.05			
		2.5 µg l-1	21 d	Mean number of neonates	56.3	Reproduction (neonates produces/fe		10	<0.05			
		Total Ag	21 d	EC50	1 µg l-1	Reproduction (neonates produces/fe	Non-lin	10	N/A			
		Total Ag	21 d	Growth Response	-	Growth Response	Diffe	10	>0.05			

Showing 1 study(s) (1 to 1)   ◀ Previous Next ▶

The chemical structures are assigned (where relevant), allowing to show all NMs with a specified coating (next figure, showing NM with citrate coating).

A summary table based on the current annotation of assays and NM types shows the best studied particle is NPO_1892 (gold nanoparticle) followed by NPO_1486 (TiO2). The numbers in the cells is the number of result values for each NM type / assay type.

A Database and Ontology Framework for Nanomaterials Design and Safety Assessment

Search ▾   Nanomaterials ▾   OpenTox ▾   Demo ▾   Help ▾

Home 〉 All assessments 〉 This assessment 〉 Versions 〉

| Assessment identifier | Collect structures | Endpoint data used | Assessment details | Report |

| - 7 - | | | | citrate | HAS_COATING | | S |

Showing from **1** to **10** in pages of [10 ▾] substances ◀ Previous Next ▶

Filter...

	Substance Name	Substance UUID	Substance Type	Public name	Reference substance UUID	Owner	Info	Contained in as
☑ - 1 -	MNT49-Ag NP	XLSX-c3...	NPO_1892	Ag NP	MNTX-97...	In house	Shape = spherical Sheet2-Row# = 766 Size = 41.6	coating ❶
☑ - 2 -	MNT31-citrate-AgNPs	XLSX-c6...	NPO_1892	citrate-AgNPs	MNTX-97...	In house	Name / Batch No. = Not Given Shape = spherical Sheet2-Row# = 448 Size = 10.0	coating ❶
☑ - 3 -	MNT13-cit-Ag	XLSX-80...	NPO_1892	cit-Ag	MNTX-97...	In house	Sheet2-Row# = 285 Size = 17.0	coating ❶
☑ - 4 -	MNT1-AgNP	XLSX-a8...	NPO_1892	AgNP	MNTX-97...	In house	Name / Batch No. = NA Sheet2-Row# = 3 Size = 30.0	coating ❶
☑ - 5 -	MNT21-AgNP	XLSX-d1...	NPO_1892	AgNP	MNTX-97...	In house	Sheet2-Row# = 336 Size = 15 ± 6	coating ❶
☑ - 6 -	MNT43-Nano-Ag 20 nm	XLSX-4c...	NPO_1892	Nano-Ag 20 nm	MNTX-97...	Ted Pella Inc., California, USA	Sheet2-Row# = 610 Size = 20.0	coating ❶
☑ - 7 -	MNT70-C50	XLSX-ab...	NPO_1892	C50	MNTX-97...	Ted Pella Inc. (USA)	Sheet2-Row# = 1158 Size = 50.0	coating ❶
☑ - 8 -	MNT38-AgNPCIT	XLSX-e7f...	NPO_1892	AgNPCIT	MNTX-97...	In house	Sheet2-Row# = 544 Size = 12.0	coating ❶
☑ - 9 -	MNT41-AgNP	XLSX-34...	NPO_1892	AgNP	MNTX-97...	Stanford Materials Corp (Aliso Viejo, California)	Sheet2-Row# = 586 Size = 30-50	coating ❶
☑ - 10 -	MNT70-C20	XLSX-c9...	NPO_1892	C20	MNTX-97...	Ted Pella Inc. (USA)	Sheet2-Row# = 1147 Size = 20.0	coating ❶

Row Labels	ENM_9000006	NPO_1373	NPO_1384	NPO_1486	NPO_1541	NPO_1542	NPO_1548	NPO_1892	NPO_401 NPO_1892	Grand Total
⊟ ECOTOX	7	6		230		12	1	607	1	864
EC_ALGAETOX				13				67		80
EC_BACTOX								63		63
EC_CHRONDAPHNIATOX				17				15		32
EC_CHRONFISHTOX				2				55		57
EC_DAPHNIATOX	1			146		12	1	160	1	321
EC_FISHTOX				46				122		168
EC_HONEYBEESTOX	6	6		6						18
EC_SOILDWELLINGTOX								125		125
⊟ ENV FATE								51		51
TO_BIODEG_WATER_SCREEN								51		51
⊟ P-CHEM	5	13	5	85	6	14	3	590	7	728
AGGLOMERATION_AGGREGATION								7		7
ASPECT_RATIO_SHAPE								9		9
CRYSTALLINE_PHASE				20	1			11		32
CRYSTALLITE_AND_GRAIN_SIZE								7		7
PC_GRANULOMETRY	2	5	1	36	1	10	2	390	6	453
PC_UNKNOWN				5	2			22		29
PC_WATER_SOL			1					62	1	64
SPECIFIC_SURFACE_AREA	2	5	1	16	1	4		11		40
SURFACE_CHEMISTRY								6		6
ZETA_POTENTIAL	1	3	2	8	1		1	65		81
⊟ TOX	16	79	2	85				1214		1396
BAO_0000451								168		168
BAO_0002084								24		24
BAO_0002733								8		8
BAO_0003009		63						42		105
CHMO_0000234								140		140
CHMO_0000239								103		103
CHMO_0000287				4				58		62
CHMO_0000538			2					133		135
CSEO_00001191								6		6
MMO_0000368								14		14
NPO_1773								7		7
OBI_0302736	8	8		8				23		47
TO_REPRODUCTION	2	2		2				11		17
UNKNOWN_TOXICITY	6	6		71				477		560
Grand Total	28	98	7	400	6	26	4	2462	8	3039

3.5. PARSING DATA FROM COST ACTION MODENA

3.5.1. DATA DESCRIPTION

The data provided by COST Action MODENA have been assembled from different laboratories using a variety of methods. Further details are given in the Annex due to the confidential status of the data at the moment of submitting this report.

3.5.2. MAPPING OF INPUT DATA TO ENANOMAPPER DATA INFRASTRUCTURE

The data is organised in a tabular fashion, suitable for modelling. It contains repeated rows of nanomaterials, in order to include EC values for several toxicology assays, and related entries of physicochemical characterisation data for each material. The mapping of the input data to eNanoMapper data model is straightforward through a JSON configuration. The data values from the repeated rows are added to the same material.

3.5.3. DATA CLEAN UP

The data file provides only minimal level of experimental details (cell type, size measuring protocol specified as "TEM" in the relevant column header, dispersion protocol specified by a single word (e.g. 'Bath', 'Vortexing'), without referring to specific protocol. The nanomaterial identifiers use a combination between chemical formulas, JRC NM-SERIES identifiers (where relevant) and (optionally) free text

referring to the NM manufacturer (e.g. 'Sigma'). We have added a new column to specify the exact NM-Series identifier (e.g. NM-111) in a consistent format (the current observation is that each project has its own version of writing NM-series identifiers).

The composition is described in a dedicated column 'Coating' by free text (e.g. 'naked, 80% anatase 20% rutile', 'naked, rutile with minor anatase', 'organic','Peptide') . The content of this column is difficult to parse automatically and will be subject of manual annotation. We have also added additional column with annotation of the NM type from NPO / eNanoMapper ontology (e.g. ENM_9000006 for CeO2).

Overall, the data is structured in a form appropriate for modelling, without providing sufficient experiment details or details about the chemical composition. The MODENA project has provided also a multi worksheet Excel template for gathering experimental data (see Annex for details), which might be better suited for importing data in the eNanoMapper database. The lack of details about the chemical composition makes it difficult to use chemical descriptors.

3.5.4. RESULTING DATASET IN ENANOMAPPER

Screenshots are provided in the Annex I.

3.6. PARSING DATA FROM FP7-MARINA

3.6.1. DATA DESCRIPTION

The *in vitro* data in MARINA project were collected using Excel templates including tabs on test conditions, raw data, test results and test summary. The in vitro toxicity assessment was performed by different laboratories for six representative oxide NMs (provided by the EC-JRC Nanomaterials): two TiO2 NMs with different surface chemistry–hydrophilic (NM-103) and hydrophobic (NM-104), two forms of ZnO–uncoated (NM-110) and coated with triethoxycaprylylsilane (NM-111) and two SiO2NMs produced by two different manufacturing techniques–precipitated (NM-200) and pyrogenic (NM-203). The testing was performed in twelve cellular models representing six different target organs/systems (immune system, respiratory system, gastrointestinal system, reproductive organs, kidney and embryonic tissues). The toxicity assessment was conducted using ten different assays for cytotoxicity, embryotoxicity, epithelial integrity, cytokine secretion and oxidative stress. However this set of data provides a rich source of information for modeling. Data included:

- General description
- SOP used
- Endpoint
- Test item
- Dispersion type
- Cell model
- Culture conditions
- Exposure time
- Raw data
- Test results
- Test summary

3.6.2. MAPPING OF INPUT DATA TO ENANOMAPPER DATA INFRASTRUCTURE

The mapping of the Excel file is configured in a JSON file (below).

```json
{
  "TEMPLATE_INFO": {
    "NAME": "MARINA",
    "VERSION": "Basic (based on 'Test results')",
    "TYPE": 1
  },
  "DATA_ACCESS": {
    "ITERATION": "ROW_SINGLE",
    "SHEET_INDEX": 1,      "START_ROW": 2,      "END_ROW": 2,
    "START_HEADER_ROW": 1,      "END_HEADER_ROW": 1,
    "RECOGNITION": "BY_INDEX",
    "VARIABLES": {
      "Assay_endpoint": {
        "ITERATION": "ABSOLUTE_LOCATION",   "SHEET_INDEX": 1,
        "COLUMN_INDEX": "B",           "ROW_INDEX": 16
      },
      "Outcome_metric": {
        "ITERATION": "JSON_VALUE",          "JSON_VALUE": "%"
      },
      "C": {
        "IS_ARRAY": true,           "TRIM_ARRAY": true,
        "ITERATION": "ABSOLUTE_LOCATION",
        "SHEET_INDEX": 1,           "COLUMN_INDEX": "B",
        "COLUMN_INDICES": [
          2, 3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20
        ],
        "ROW_INDEX": 43
      },
      "TimeDesignations": {
        "IS_ARRAY": true,           "TRIM_ARRAY": true,
        "ITERATION": "ABSOLUTE_LOCATION",
        "SHEET_INDEX": 1,           "COLUMN_INDEX": "B",
        "COLUMN_INDICES": [
          2, 3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20
        ],
        "ROW_INDEX": 60
      },
      "TimePoints": {
        "IS_ARRAY": true,           "TRIM_ARRAY": true,
        "ITERATION": "ABSOLUTE_LOCATION",          "SHEET_INDEX": 1,           "COLUMN_INDEX": "B",
        "COLUMN_INDICES": [
          2, 3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20
        ],
        "ROW_INDEX": 61
      },
      "Replicates": {
        "ITERATION": "JSON_VALUE",    "JSON_VALUE": 3
      }
    },
    "VARIABLE_MAPPINGS": [
      { "NAME": "Time",
        "KEYS_VARIABLE": "TimeDesignations",   "VALUES_VARIABLE": "TimePoints"
      }
    ]
  },
  "PARALLEL_SHEETS": [
    {
      "ITERATION": "ROW_SINGLE",
      "SHEET_INDEX": 3,       "START_ROW": 2,
      "START_HEADER_ROW": 1,          "END_HEADER_ROW": 1,
```

```
      "ALLOW_EMPTY": true,        "RECOGNITION": "BY_INDEX"
   }
],
"SUBSTANCE_RECORD": {
   "PUBLIC_NAME": {
      "ITERATION": "ABSOLUTE_LOCATION",        "SHEET_INDEX": 1,
      "COLUMN_INDEX": "B",        "ROW_INDEX": 27
   },
   "COMPANY_UUID": {
      "ITERATION": "ABSOLUTE_LOCATION",        "SHEET_INDEX": 1,
      "COLUMN_INDEX": "B",        "ROW_INDEX": 29
   },
   "COMPANY_NAME": {
      "ITERATION": "ABSOLUTE_LOCATION",        "SHEET_INDEX": 1,
      "COLUMN_INDEX": "B",        "ROW_INDEX": 29
   },
   "OWNER_NAME": {
      "ITERATION": "JSON_VALUE",        "JSON_VALUE": "JRC Nanomaterials Repository"
   },
   "COMPOSITION": [
      {
         "STRUCTURE_RELATION": "HAS_CORE",
         "PROPERTIES": {
            "NAME": {
               "ITERATION": "ABSOLUTE_LOCATION",            "SHEET_INDEX": 1,
               "COLUMN_INDEX": "C",            "ROW_INDEX": 27
            },
            "CASRN": {
               "ITERATION": "ABSOLUTE_LOCATION",            "SHEET_INDEX": 1,
               "COLUMN_INDEX": "C",            "ROW_INDEX": 28
            }
         },
         "PROPORTION": {
            "TYPICAL_VALUE": {
               "ITERATION": "ABSOLUTE_LOCATION",            "SHEET_INDEX": 1,
               "COLUMN_INDEX": "B",            "ROW_INDEX": 31
            },
            "TYPICAL_UNIT": {
               "ITERATION": "ABSOLUTE_LOCATION",            "SHEET_INDEX": 1,
               "COLUMN_INDEX": "C",            "ROW_INDEX": 31
            }
         }
      }
   ]
},
"PROTOCOL_APPLICATIONS": [
   {
      "PROTOCOL_TOP_CATEGORY": {
         "ITERATION": "JSON_VALUE",        "JSON_VALUE": "TOX"
      },
      "PROTOCOL_CATEGORY_CODE": {
         "ITERATION": "JSON_VALUE",        "JSON_VALUE": "NPO_1709_SECTION"
      },
      "PROTOCOL_GUIDELINE": {
         "guideline1": {
            "ITERATION": "ABSOLUTE_LOCATION",            "SHEET_INDEX": 1,
            "COLUMN_INDEX": "B",            "ROW_INDEX": 13
         }
      },
      "PROTOCOL_ENDPOINT": {
```

```
    "ITERATION": "ABSOLUTE_LOCATION",          "SHEET_INDEX": 1,
    "COLUMN_INDEX": "B",          "ROW_INDEX": 14
  },
  "RELIABILITY_STUDY_RESULT_TYPE": {
    "ITERATION": "JSON_VALUE",          "JSON_VALUE": "experimental result"
  },
  "CITATION_TITLE": {
    "ITERATION": "JSON_VALUE",
    "JSON_VALUE": "http://dx.doi.org/10.1371/journal.pone.0127174"
  },
  "CITATION_YEAR": {
    "ITERATION": "JSON_VALUE",          "JSON_VALUE": "2015"
  },
  "CITATION_OWNER": {
    "ITERATION": "ABSOLUTE_LOCATION",          "SHEET_INDEX": 1,
    "COLUMN_INDEX": "B",          "ROW_INDEX": 9
  },
  "PARAMETERS": {
    "Dispersion protocol": {
      "ITERATION": "ABSOLUTE_LOCATION",          "SHEET_INDEX": 1,
      "COLUMN_INDEX": "B",          "ROW_INDEX": 35
    },
    "Dispersion agent": {
      "ITERATION": "ABSOLUTE_LOCATION",          "SHEET_INDEX": 1,
      "COLUMN_INDEX": "B",          "ROW_INDEX": 38
    },
    "Aids used to disperse/Sonification": {
      "ITERATION": "ABSOLUTE_LOCATION",          "SHEET_INDEX": 1,
      "COLUMN_INDEX": "C",          "ROW_INDEX": 40
    },
    "Aids used to disperse/Vortexing": {
      "ITERATION": "ABSOLUTE_LOCATION",          "SHEET_INDEX": 1,
      "COLUMN_INDEX": "E",          "ROW_INDEX": 40
    },
    "Aids used to disperse/Stirring": {
      "ITERATION": "ABSOLUTE_LOCATION",          "SHEET_INDEX": 1,
      "COLUMN_INDEX": "G",          "ROW_INDEX": 40
    },
    "Cell line/Type - Short name": {
      "ITERATION": "ABSOLUTE_LOCATION",          "SHEET_INDEX": 1,
      "COLUMN_INDEX": "B",          "ROW_INDEX": 47
    },
    "Cell line/Type - Full name": {
      "ITERATION": "ABSOLUTE_LOCATION",          "SHEET_INDEX": 1,
      "COLUMN_INDEX": "B",          "ROW_INDEX": 48
    },
    "Cell line/Type - Supplier": {
      "ITERATION": "ABSOLUTE_LOCATION",          "SHEET_INDEX": 1,
      "COLUMN_INDEX": "B",          "ROW_INDEX": 49
    },
    "Cell culture conditions - Medium": {
      "ITERATION": "ABSOLUTE_LOCATION",          "SHEET_INDEX": 1,
      "COLUMN_INDEX": "B",          "ROW_INDEX": 53
    },
    "Cell culture conditions - Serum": {
      "ITERATION": "ABSOLUTE_LOCATION",          "SHEET_INDEX": 1,
      "COLUMN_INDEX": "B",          "ROW_INDEX": 54
    },
    "Cell culture conditions - Serum concentration in growth medium": {
      "ITERATION": "ABSOLUTE_LOCATION",          "SHEET_INDEX": 1,
```

```
              "COLUMN_INDEX": "B",              "ROW_INDEX": 55
        },
        "Cell culture conditions - Serum concentration in treatment medium": {
            "ITERATION": "ABSOLUTE_LOCATION",              "SHEET_INDEX": 1,
            "COLUMN_INDEX": "B",              "ROW_INDEX": 56
        }
      },
      "EFFECTS_BLOCK": {
        "LOCATION": {
            "ITERATION": "ABSOLUTE_LOCATION",              "IS_ARRAY": true,
            "TRIM_ARRAY": true,              "SHEET_INDEX": 3,
            "COLUMN_INDEX": "B",              "ROW_INDEX": 4
        },
        "ROW_SUBBLOCKS": "= TimePoints.size()",
        "COLUMN_SUBBLOCKS": "= Replicates",
        "SUBBLOCK_SIZE_ROWS": "= C.size() + 4",
        "SUBBLOCK_SIZE_COLUMNS": 6,
        "VALUE_GROUPS": [
          {
            "NAME": "=Assay_endpoint",
            "UNIT": "=Outcome_metric",
            "START_COLUMN": 5,              "END_COLUMN": 5,
            "START_ROW": 3,              "END_ROW": "=3 - 1 + C.size()",
            "PARAMETERS": [
              {
                "NAME": "Time point",
                "ASSIGN": "ASSIGN_TO_SUBBLOCK",
                "COLUMN_POS": 1,                      "ROW_POS": 2,
                "MAPPING": "Time",                      "UNIT": "h"
              },
              {
                "NAME": "Replicate",
                "ASSIGN": "ASSIGN_TO_SUBBLOCK",
                "COLUMN_POS": 1,                      "ROW_POS": 1
              },
              {
                "NAME": "Concentration",
                "ASSIGN": "ASSIGN_TO_VALUE",
                "COLUMN_POS": -3,                      "ROW_POS": 0,
                "UNIT": "µg/ml"
              }
            ]
          }
        ]
      }
    }
  ]
}
```
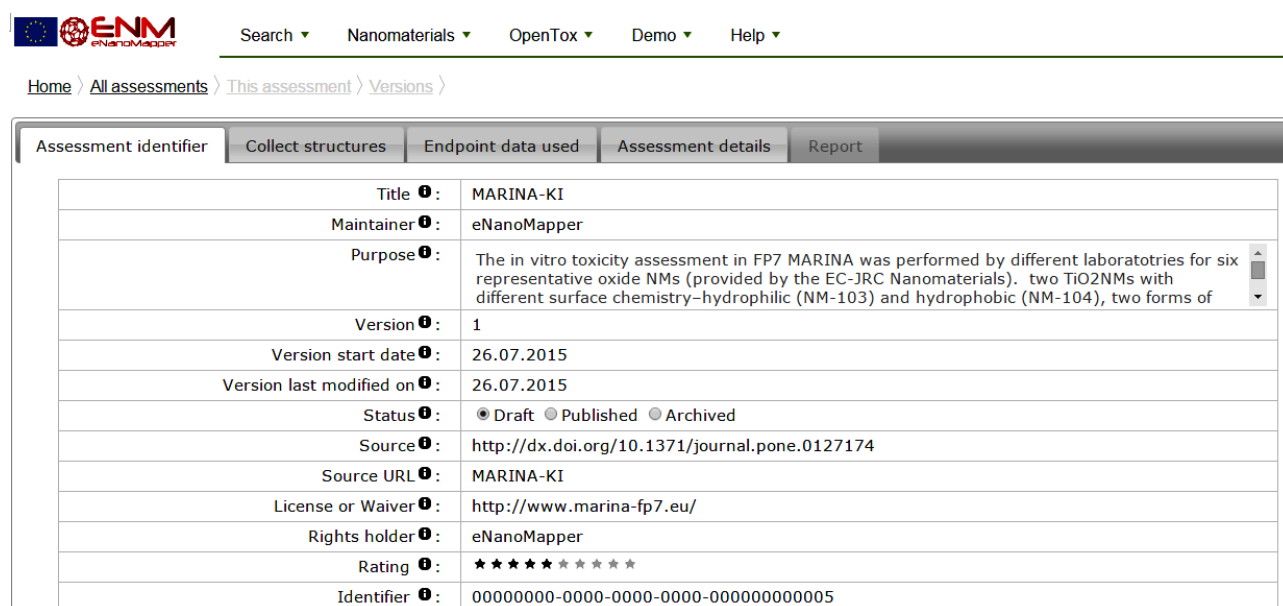
3.6.3. DATA CLEAN UP

- NM codes in provided Excel files are written in a way (e.g."NM-111 (ZnO)") that prevent automatic matching with JRC representative material codes used in JRC repository or in the NANoREG templates (e.g. NM-111). Renamed to e.g. "NM-111".

- The NM code (Row 29, Sheet 1) in the file *MARINA-invitro-WP09-P32-NM111ZnO-HMDM_Cytokines_TNFa.xls* was *Standard MARINA Nanomaterial Code & Name: NM-110 (ZnO)* NM110 , while the expected one is *NM111*. Fixed.

3.6.4. RESULTING DATASET IN ENANOMAPPER

The resulting data set loaded into eNanoMapper is described in this section. We here show a few screenshots of aspects of the MARINA/KI data as exposed by the AMBIT platform

A material is stored in the database as a substance, and as substances are in AMBIT, it has general information, a composition, and properties. The toxicity data contains two assays "ELISA / TNF-α release in culture medium" (BioAssay Ontology BAO_0002993) and lactate Dehydrogenase Assay "Cell viability - cell membrane damage: Lactate dehydrogenase assay (LDH) (CytoTox 96® Non-radioactive cytotoxicity assay - Promega, G1780)" (mapped to NPO_1709). A screenshot of TiO2 (NM-103) displaying the data from the two assays is shown next.

The mapping and the screenshots refer to import of six MARINA data files, provided by partner KI (2 assays). The FP-7 MARINA project provided to eNanoMapper 5 templates (see Annex I for details) for reporting physicochemical and in-vitro data. The structure of the templates is very similar to the ones described and we are able to parse these with appropriate JSON configuration.

Search ▾   Nanomaterials ▾   OpenTox ▾   Demo ▾   Help ▾

| IUC Substance | Composition | **Tox (2)** |

Filter…   Expand all | Collapse all

Titanium Dioxide

NPO_1709 LDH Release Assay (1)   ▾

BAO_0002993 Cytotoxicity Assay (1)   ▴

Reference ▲	Protocol ⇕	Endpoint ⇕	Result ⇕	Concentration ⇕	Time point ⇕	Replicate ⇕	Owner ⇕
		TNF-α (ng/ml)	1.893 ng/ml	0 µg/ml	24 h	Replicate 1	
	ELISA / TNF-α release in culture medium	TNF-α (ng/ml)	19.698 ng/ml	1 µg/ml	24 h	Replicate 1	
		TNF-α (ng/ml)	18.521 ng/ml	5 µg/ml	24 h	Replicate 1	
		TNF-α (ng/ml)	18.028 ng/ml	10 µg/ml	24 h	Replicate 1	
	Aids used to disperse/stirring: Y	TNF-α (ng/ml)	18.415 ng/ml	25 µg/ml	24 h	Replicate 1	
		TNF-α (ng/ml)	19.297 ng/ml	50 µg/ml	24 h	Replicate 1	
	Cell line: HMDM (primary cells)	TNF-α (ng/ml)	20.887 ng/ml	75 µg/ml	24 h	Replicate 1	
	Cell line/type - supplier: ATCC	TNF-α (ng/ml)	22.696 ng/ml	100 µg/ml	24 h	Replicate 1	
		TNF-α (ng/ml)	-0.343 ng/ml	0 µg/ml	24 h	Replicate 2	
	Cell culture conditions - Serum: FBS (Gibco/413105K)	TNF-α (ng/ml)	5.989 ng/ml	1 µg/ml	24 h	Replicate 2	
		TNF-α (ng/ml)	5.77 ng/ml	5 µg/ml	24 h	Replicate 2	
	Cell culture conditions - Serum concentration in growth medium: 0.1	TNF-α (ng/ml)	5.301 ng/ml	10 µg/ml	24 h	Replicate 2	
		TNF-α (ng/ml)	5.16 ng/ml	25 µg/ml	24 h	Replicate 2	
	Cell culture conditions - Serum concentration in treatment medium: 0.1	TNF-α (ng/ml)	5.229 ng/ml	50 µg/ml	24 h	Replicate 2	
		TNF-α (ng/ml)	6.156 ng/ml	75 µg/ml	24 h	Replicate 2	
2015	Cell line/Type - Full name: human monocyte-derived macrophages (HMDM)	TNF-α (ng/ml)	7.281 ng/ml	100 µg/ml	24 h	Replicate 2	P32 - KI
		TNF-α (ng/ml)	2.189 ng/ml	0 µg/ml	24 h	Replicate 3	
	Dispersion agent: water with 0,05% BSA	TNF-α (ng/ml)	2.474 ng/ml	1 µg/ml	24 h	Replicate 3	
		TNF-α (ng/ml)	2.002 ng/ml	5 µg/ml	24 h	Replicate 3	
	Aids used to disperse/vortexing: N	TNF-α (ng/ml)	1.786 ng/ml	10 µg/ml	24 h	Replicate 3	
		TNF-α (ng/ml)	1.57 ng/ml	25 µg/ml	24 h	Replicate 3	
	Aids used to disperse/sonification: N	TNF-α (ng/ml)	1.735 ng/ml	50 µg/ml	24 h	Replicate 3	
		TNF-α (ng/ml)	1.29 ng/ml	75 µg/ml	24 h	Replicate 3	
	Dispersion protocol: NANOGENOTOX Dispersion protocol (http://www.nanogenotox.eu/files/PDF/web%20nanogenotox%20dispersion%20protocol.pdf)	TNF-α (ng/ml)	1.206 ng/ml	100 µg/ml	24 h	Replicate 3	
		TNF-α (ng/ml)	9.502 ng/ml	0 µg/ml	24 h	Replicate 4	
		TNF-α (ng/ml)	8.808 ng/ml	1 µg/ml	24 h	Replicate 4	
		TNF-α (ng/ml)	8.957 ng/ml	5 µg/ml	24 h	Replicate 4	
		TNF-α (ng/ml)	8.306 ng/ml	10 µg/ml	24 h	Replicate 4	
	Cell culture conditions - medium: RPMI (Sigma, RNBB2729	TNF-α (ng/ml)	8.244 ng/ml	25 µg/ml	24 h	Replicate 4	
		TNF-α (ng/ml)	9.507 ng/ml	50 µg/ml	24 h	Replicate 4	
		TNF-α (ng/ml)	14.416 ng/ml	75 µg/ml	24 h	Replicate 4	
		TNF-α (ng/ml)	15.438 ng/ml	100 µg/ml	24 h	Replicate 4	

Showing 1 study(s) (1 to 1)   ◀ Previous Next ▶

3.7. PARSING NMS CHARACTERIZATION DATA

3.7.1. DATA DESCRIPTION

The main physico-chemical parameters described:
- Composition
- Surface chemistry
- Crystal phase
- Crystalline size
- Primary Particle size
- Particle size distribution

- Aggregates/agglomerates size
- Specific surface area
- Zeta potential surface charge
- Solubility
- TEM images

JRC NANOMATERIALS REPOSITORY List of Representative Nanomaterials [*Refs. 6-10*]:
https://ec.europa.eu/jrc/sites/default/files/JRC%20Nanomaterials%20Repository%20-
%20List%20of%20Representative%20Nanomaterials.pdf

3.7.2. MAPPING OF INPUT DATA TO ENANOMAPPER DATA INFRASTRUCTURE

After several unsuccessful attempts to obtain the data from JRC in machine readable format we decided to extract it from the published PDF reports. A closer look revealed that automatic data extraction would be very hard to perform or even unfeasible, due to a high degree of variability in data formats, units, types of experiments, measurement methods and specific conditions, often described in free text. Therefore we began manual extraction of the relevant information in an ad-hoc spreadsheet, which so far includes 46 lines (materials) and 406 columns (properties). After finishing the data extraction phase we plan to perform an analysis of the resulting spreadsheet in order to reduce (if possible) the number of columns by mapping and merging semantically equivalent representations.

We plan to use the same approach, as described above, for importing the resulting ad-hoc spreadsheet in the eNanoMapper database using a configurable Java parser with a specific JSON configuration.

4. APPLICATIONS

The data import can be performed either with the help of command line application (requires direct access to the database) or through REST API.
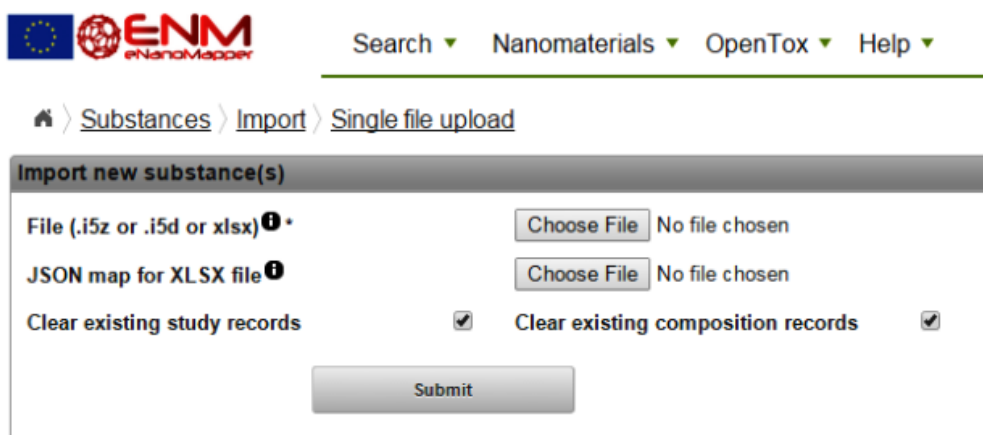
4.1 COMMAND LINE APPLICATION

(code on private repository on BitBucket, uses https://github.com/enanomapper/nmdataparser/ as a dependency)

```
usage: java -jar enmcli-jar-with-dependencies.jar
-c,--config <file>          Config file (DB connection parameters)
-h,--help                   This help
-i,--input <file>           Input file
-j,--json <file>            JSON config file
-m,--clearMeasurements <value>   true|false
-o,--output <file>          Output file
-p,--parser <type>          File parser mode : xlsx|modnanotox
-s,--isSplitRecord <value>       true|false
-t,--clearComposotion <value>    true|false
```

4.2 API AND WEB BASED UPLOAD

The eNanoMapper database REST API supports data upload through HTTP POST to the substance resource (described previously in D3.1). The user interface is a regular web form for file upload (Figure 2). The two checkboxes control whether the composition records and study records for the materials being imported will be cleared, if already in the database. Each material entry in the database is assigned a unique identifier in the form of a UUID. If the input file is *.i5z or *.i5d, the identifiers are the IUCLID5 generated UUIDs already present in these files (e.g. IUC5-5f313d1f-4129-499c-abbe-ac18642e2471). If the input file is a spreadsheet, the JSON configuration defines which field to be used as an identifier and uses the field itself or generates UUID from the specified field (e.g. FCSV-bc77c03d-4e75-3fab-bb3d-17b983663819 indicates the entry imported from CSV file). The parser may be configured to use a custom prefix on import, e.g. "NWKI-" for NanoWiki entries, generating UUID like "NWKI-71060af4-1613-35cf-95ee-2a039be0388a".

Figure 2. Data upload web page of the database system showing support for two file formats

5. INTEGRATED TESTING

5.1. AUTOMATIC TESTING

An extensive JUnit test suite, which includes actual import into a test database and verifying the database content against expected values is available via private repository https://bitbucket.org/vedina/parserconfig. This repository is not publicly available as it contains a mix of public and confidential files provided by the NanoSafety cluster.

5.2. MANUAL VALIDATION

The correctness of the data is manually verified by the team compiling the data as well as by the data provider, although full validation is only possible for small datasets. Once the data is online, a mechanism for notification of possible errors would be a reasonable approach for supporting continuous curation. Summary views of the data will help in this respect and are considered as part of user interface being developed (D5.7, due M24).

The overall process is iterative, involving subsequent steps of data import, identifying discrepancies, introducing solutions (e.g. annotations) and new tests. A persistent challenge is the inconsistent use of the naming of different entries, not only endpoints, species, cell lines and various experimental parameters, but even nanomaterial identifiers , which were invented for the reason of being standardized (next screenshot - "NM-110 (ZnO)", "ZnO NM-110", "NM110") "JRC2011 NM110" all refer to the same NM-110 material from the set of JRC Representative Manufactured nanomaterials).

6. CONCLUSION

We have analyzed a number of spreadsheet data templates, provided to eNanoMapper by NanoSafety Cluster projects, and developed a configurable Java parser, enabling the data import into eNanoMapper database. The spreadsheets are the tool of choice for data preparation in the NanoSafety Cluster and facilitate the data sharing by avoiding extra work from the data providers. Our previous experience and the perceived complexity of ISA-TAB and ISA-TAB-Nano also informed this choice. In addition, we are planning in the near future to support import and export from / into standardized formats like ISA-TAB-Nano using the same internal data model.

The Java parser configuration is defined in a separate JSON file, mapping the custom spreadsheet structure into the internal eNanoMapper storage components. This approach allows to import data from a number of spreadsheets, which differ not only by column names, but have different row-based, column-based or mixed organizations of the data. The parser maps the spreadsheet into the components of the eNanoMapper internal data model [3, 5]. The database and import templates are designed to require that the test protocol be specified for every data entry, along with number of experimental details. The upload can be done via standalone application and web interface, uploading both the Excel file and the JSON configuration via web form.

While working with NanoSafety Cluster input files, we performed only minimal data clean up and not data evaluation and are currently not assigning any data quality scores. The reasons for this approach are explained in [5]. While there is a common agreement on the importance of data curation, there is no well-established common understanding of how it should be performed. Approaches range from simple data cleaning to the entire spectrum of data-related activities including evaluation, on-going data management, and added value provisioning through analytic tools. The focus of this deliverable is on the parsing and importing of NanoSafety Cluster data into the eNanoMapper database, allowing a unified approach to storage and querying of NM related data. The amount of data clean up differ for the files provided by the different projects and are described in Section 3. By far the most extensive cleanup was performed for the ModNanoTox Excel spreadsheet.

The ontology annotation [4] through automatic tools and community involvement might help to overcome some of the challenges. Next steps include providing tutorials and possibly sand-box installations of the database to facilitate data upload testing of user generated files.

7. BIBLIOGRAPHY

1. Farcal L, Torres Andón F, Di Cristo L, Rotoli BM, Bussolati O, Bergamaschi E, et al. (2015) Comprehensive *In Vitro* Toxicity Testing of a Panel of Representative Oxide Nanomaterials: First Steps towards an Intelligent Testing Strategy. PLoS ONE 10(5): e0127174. doi:10.1371/journal.pone.0127174
2. N. Jeliazkova and V. Jeliazkov, "AMBIT RESTful web services: an implementation of the OpenTox application programming interface" J. Cheminform., vol. 3, no. 1, p. 18, Jan. 2011.
3. N. Jeliazkova, V. Jeliazkov, E. Willighagen, B. Smeets, C. Munteanu, B. Fadeel, R. Grafström, P. Kohonen, H. Sarimveis, G. Tsiliki, P. Doganis, D. Vorgrimmler, and J. Hastings, "The first eNanoMapper prototype: a substance database to support safe-by-design," in IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2014.
4. J. Hastings, N. Jeliazkova, G. Owen, G. Tsiliki, C. R. Munteanu, C. Steinbeck, and E. Willighagen, "eNanoMapper: harnessing ontologies to enable data integration for nanomaterial risk assessment," J. Biomed. Semantics, vol. 6, no. 1, 2015.
5. N. Jeliazkova, C. Chomenidis, P. Doganis, B. Fadeel, R. Grafström, B. Hardy, J. Hastings, M. Hegi, V. Jeliazkov, N. Kochev, P. Kohonen, C. R. Munteanu, H. Sarimveis, B. Smeets, P. Sopasakis, G. Tsiliki, D. Vorgrimmler, and E. Willighagen, "The eNanoMapper database for nanomaterial safety information," Beilstein J. Nanotechnol., vol. 6, pp. 1609–1634, Jul. 2015.
6. SINGH Charanjeet; FRIEDRICHS Steffi; CECCONE Giacomo; GIBSON Peter; JENSEN Keld Astrup; LEVIN Marcus; GOENAGA INFANTE Heidi; CARLANDER David; RASMUSSEN Kirsten, Cerium Dioxide, NM-211, NM-212, NM-213. Characterisation and test item preparation, 2014, http://publications.jrc.ec.europa.eu/repository/handle/JRC89825
7. RASMUSSEN Kirsten; MAST Jan; DE TEMMERMAN Pieter-Jan; VERLEYSEN Eveline; WAEGENEERS Nadia; VAN STEEN Frederic; PIZZOLON Jean Christophe; DE TEMMERMAN Ludwig; VAN DOREN Elke; JENSEN Keld Astrup; et al., Multi-walled Carbon Nanotubes, NM-400, NM-401, NM-402, NM-403: Characterisation and Physico-Chemical Properties, 2014, http://publications.jrc.ec.europa.eu/repository/handle/JRC91205
8. RASMUSSEN Kirsten; MECH AGNIESZKA; MAST Jan; DE TEMMERMAN Pieter-Jan; WAEGENEERS Nadia; VAN STEEN Frederic; PIZZOLON Jean Christophe; DE TEMMERMAN Ludwig; VAN DOREN Elke; JENSEN Keld Astrup; BIRKEDAL Renie; et al., Synthetic Amorphous Silicon Dioxide (NM-200, NM-201, NM-202, NM-203, NM-204): Characterisation and Physico-Chemical Properties, 2013, http://publications.jrc.ec.europa.eu/repository/handle/JRC83506
9. RASMUSSEN Kirsten; MAST Jan; DE TEMMERMAN Pieter-Jan; VERLEYSEN Eveline; WAEGENEERS Nadia; VAN STEEN Frederic; PIZZOLON Jean Christophe; DE TEMMERMAN Ludwig; VAN DOREN Elke; JENSEN Keld Astrup; et al., Titanium Dioxide, NM-100, NM-101, NM-102, NM-103, NM-104, NM-105: Characterisation and Physico-Chemical Properties, 2014, http://publications.jrc.ec.europa.eu/repository/handle/JRC86291
10. SINGH Charanjeet; FRIEDRICHS Steffi; LEVIN Michael; BIRKEDAL Rene; JENSEN Keld Astrup; POJANA Giulio; WOHLLEBEN Wendel; SCHULTE Stefan; WIENCH Karin; TURNEY T; et al., NM-Series of Representative Manufactured Nanomaterials - Zinc Oxide NM-110, NM-111, NM-112, NM-113: Characterisation and Test Item Preparation, 2011, http://publications.jrc.ec.europa.eu/repository/handle/JRC64075

8. ANNEXES

ANNEX I. OVERVIEW OF NANOSAFETY CLUSTER TEMPLATES AND DATA FILES

(Confidential)

ANNEX II. STRUCTURE OF TEMPLATES AND DATA FILES

(Confidential)