ProSafe

Grant Agreement Number 646325

Deliverable D 3.4

Report on available database linking tools

Due date of deliverable: *2016/12/31*

Actual submission date: *2017/05/02*

Author(s) and company:	Hugues Crutzen (JRC)
Work package/task:	WP3 / Task 3.3
Document status:	draft / final
Confidentiality:	confidential / restricted / public
Key words:	nanoEHS data database link tools

DOCUMENT HISTORY

Version	*Date*	*Reason of change*
1	15/12/2016	Initial version
2	02/05/2017	Version submitted to MC
3	14/07/2017	Project Office harmonized lay-out

Lead beneficiary for this deliverable: Joint Research Centre, JRC

Owner(s) of this document	
Owner of the content	JRC

Table of Content

1 Description of task

From DoW, amended 22.09.2016

Task 3.3: Linking databases - Start M9, end M24
Task Leader: JRC Partners: TEMAS, IOM

In addition to T3.2, which coordinates the establishment of a nanoEHS community-agreed nanosafety data management system, this task brings NanoEHS community stakeholders together to find adequate IT tools to link and exploit the existing databases. The possible use and development of conversion modules, 'parsers', 'APIs' and other IT software modules will be reviewed in collaboration with specialists from other EU projects, in particular NANoREG, eNanoMapper and the Nanosafety Cluster, by whom those IT tools are to be developed. Care shall be taken to supporting the development or adaptation of IT tools that are user-friendly. This PROSAFE work coordinates the definition of IT tools to become elements of a tool box for regulators, to be developed by NANoREG.

2 Description of work & main achievements

2.1 Summary

The end goal of this ProSafe task on support to linking data, and of any similar initiative in the nanoEHS arena, is of particular importance to scientists and regulator: to "*answer scientific questions that require data from two or more sets of data or ontology sources*"[1].

Critical pending issues and requirements to achieve the goal have been identified in collaboration with ProSafe T3.1.

The initial ambition of ProSafe WP3 partners to achieve a 'linkage' of at least two datasets has been reasonably fulfilled, thanks to the strategic IT-oriented role of FP7 eNanoMapper (eNM) and the agreements between FP7 NANoREG and to other H2020 projects NanoReg2 and caliBrate, just before that large FP7 initiative came to an end.

ProSafe recommends the European Commission to duly consider ways to further integrate the work promoted by ProSafe on data management and linking data/datasets, in particular as carried out by FP7 eNanoMapper, into upcoming strategic nanoEHS R&I funding, and to link this appropriately to the burning issue of data sustainability and curation.

The recommendations of this deliverable (section 2.4) serve also as input to the aspects related to data management in the ProSafe White Paper. They are also valid points worth transferring into the European nanosafety informatics (nanoinfomatics) roadmap being developed under the auspices of the EU NanoSafety Cluster.

The successful implementation of shared/linked nanoEHS data is intimately related to the openness of access to those data. Unless at least publicly-funded datasets of projects are (entirely) open, the positive effects on the nanoEHS community of stakeholders of actions geared towards sharing/linking clearly will remain quite limited.

[1] See conclusions of Egon Willighagen, Micha Rautenberg, Denis Gebele, Penny Nymark, Pekka Kohonen, Nina Jealiazkova, Barry Hardy. Deliverable Report D3.3 Modules and services for linking and integration with third party databases. (Zenodo, 2016). doi:10.5281/zenodo.375813

2.2 Background of the task

Task 3.3 was envisaged as combining of the results of the other two tasks of ProSafe WP3: T3.1 and T3.2.

Task 3.1 delivered a completed mapping of available and reachable datasets in the wide nanoEHS community (deliverable D3.1).
Task 3.2 was divided into two sub-tasks, dealing with different aspects concerning the creation of a database management system: i) *ISA-TAB-NANO as the backbone for a common database (T3.2.1)* and ii) *Minimum requirements in ontology and naming conventions (T3.2.2)*. The two subtasks are related to two deliverables: i) D3.2 – *ISA-TAB-NANO database system established and adopted within the Nanosafety Cluster*, and *ii)* D3.3 – *Minimal ontology and naming convention for nanosafety data.*

The general aim of task 3.3 is the support to the development of tools that effectively allow the linking of suitable databases identified in D3.1, using the 'system logic' established in D3.2 and D3.3. As discussed during ProSafe WP3 meetings and at the second ProSafe Consortium meeting, the intent of this Coordination and Support Action (CSA), in T3.3, was to show the feasibility of the linking databases concept in at least one case, i.e. connecting at least two datasets.

ProSafe and this task started shortly after the launch of the FP7 project eNanoMapper and linked with it to raise the awareness about the need to develop an appropriate nanomaterials / nanoEHS data management system to enable in Europe. ProSafe, being a CSA and not a research and innovation action, did not intend to devote efforts to the technical / IT development of linking. This was, and is still being done, by other EU-funded projects such as eNanoMapper and H2020 OpenRiskNet.

2.3 Description of the work carried out and results

Thanks to the database mapping work of T3.1 by IOM and JRC, some important findings were obtained (D3.1, section 2.9):

"*iv. Even for databases that could be readily identified, it is clear that very few are currently sufficiently compatible in formats to promptly allow data availability and exchange in a harmonised way.* […] *Whilst these issues have been acknowledged for several years, there are few projects to date that have adopted unified or shared approaches to nanoEHS data management*" (p.50),

"*v. Of the ongoing NSC-related projects just seven, not including eNanoMapper, are identified as having some potential for data exchange* […] *(table 10). eNM indeed may be the 'bridge' to exchange or link data in the short term via ongoing collaborations, and this is also highlighted in that project DoW*" (p.50), and

"*iv. Table 12* […] *provides a summary of* […] *gaps and outstanding issues* [in nanoEHS data management and] *shows several positive actions towards harmonisation that are well underway and making significant inroads*", (p.51).

Those two key tables 10 and 12 from D3.1 are copied hereafter.

Table 10 from ProSafe D3.1: "*active projects with data-linkage potential for ProSafe.*"

Project Name	Time frame	Comments on potential for interaction with ProSafe
NANOFASE	2015 – 2019	Not predominately Tox but possibly some; will have Ecotox data
NANOMILE	2013 – Mar 2017	Database plans not clear. Likely to collaborate with eNanoMapper (eNM) on Phys-chem and Toxdata loading to eNM database. (Will also have omics)
NANoREG	2013 – 2017	NANoREG database and JRC ISA-TAB-Nano templates[2] used. Ongoing collaboration and synergy with eNM. Public release of NANoREG data in 2017; transfer into eNM database instance
NanoReg2	2015 – 2018	Will use databases from NANoREG, supplemented with additional aspects; will also use the eNanoMapper data model and data management tools to consume/upload collected datasets that are likely based upon the NANoREG adapted ISA-TAB-Nano and possible older data (ENPRA, etc.) using MARINA-type templates
NANOSOLUTIONS	2013 – Mar 1017	Data being assembled. Possible collaboration with eNM on testing the parsing of physicochemical and tox data and ISA-TAB-Nano transfers (also will have omics data that may be tested in collaboration).
SUN	2013 – Mar 2017	Data being assembled. Possible collaboration with eNM on testing the parsing of physicochemical and tox data and ISA-TAB-Nano transfers (also will have omics data that may be tested in collaboration)
DaNa	Ongoing non EU	A German government sponsored database, not FP7 or H2020. Appears potentially to have some datasets of interest for possible exploitation.

Table 12 from ProSafe D3.1: "*summary of key issues and requirements, and current initiatives in the nanoEHS Landscape, as evidenced through the EU NSC.*"

NanoEHS Landscape issue or development need	Current initiatives or *outstanding action*
No standardised language, i.e. ontology or controlled vocabulary, to classify and describe data yet available to be applied as standard	Developments ongoing in both US and EU. eNanoMapper ontology under continuing development and promotion/training to FP7 / H2020 projects and beyond. Linked intimately with use of ISA-TAB-Nano format. Development of ontology extensions for nano-exposure data with NSC WG3 & NECID. *Not necessarily easy to use for average data generator: help and guidance and very good integration with end user tools will be required.*
A lack of standardised data across the nano-EHS data domains. Diverse and poorly described formats and layouts. None or poorly described meta-data / documentation; needs expansion for further data types.	To be pursued by consistent use of ontology in data definitions and ISA-TAB-nano format as basis of data capture format. *Extend to other data types or develop analogous alternatives*
Need for data capture front-ends and data collection templates. For the end user needs to be lightweight, intuitive and easy to use. ISA-TAB-nano format alone not easy to use and a potential	eNanoMapper providing for parsing of other templates into a NSC-wide DB instance (at present at eNM), and underlying ISA-TAB-nano format. See for instance data from MARINA templates.

[2] Totaro S. et al; Data logging templates for the environmental, health and safety assessment of nanomaterials; EUR 28137 EN; doi:10.2787/505397; January 2017

barrier to uptake.	Follow the NANoREG example and build on its open access and free templates, based on a simplified, user-friendly adaptation of the ISA-TAB-Nano format to record *experimental* data. *Need to develop other front ends with shared standards for a variety of templates for different data types and situations.* *For occupational exposure, pursue the eNanoMapper-NECID work on ontology mapping.* *Explore the appropriateness of connecting to ontology and front end data capture via eNanoMapper APIs* *Disconnected remote/off-line data capture modes needed.* *End user training and education requirements. Resources, materials, guidance, data templates and templates for data planning and workflows.*
Wide diversity in database construction and use: structurally and functionally. Database being re-invented.	eNanoMapper database instance to provide a consistent NSC-level model; to date for physicochemical and (eco)tox data. *Possible extensions for other data types domains.* *Making this available and understood. Mediating updates and consistency if DBs localised / distributed*. *Make DBs/data models, workflows descriptions and models, technology-agnostic, i.e. not dependant on particular DB brand or technology.*
No consistent, well-organised, shared approaches or accepted discipline applied widely for planning and managing data collection, processing and management of the data. No coherent approach to good nano-EHS Data Management Planning (DMP) and practice. This is needed for Open Data standards.	Conceptually being addressed, in examining / providing resources that can be shared, by the NSKI (EU), the CEINT Nanoinformatics Knowledge Commons (US) and the US-EU CoR. *Need to develop standardised approaches, resources, (templates, design guides etc) with training and guidance the can be adopted for use consistently throughout the community.*
Difficulty in discovering, obtaining (permission), accessing and retrieving data, even if it was "properly" databased. Inability to exploit or combine valuable EU-funded dataset for mutual and public benefits. Need to meet new "Open Data" requirements. Currently no accepted route or resources for long-term data availability and sustainability. Data not submitted (recorded) for IPR reasons.	NSC to assess sustainability and funding. *Possibly establish new dedicated Sustainability WG to investigate and consult on strategy and resourcing.* *Adoption of standards and consistency to be enhanced, as well as the ability to locate data in a well-resourced (sustainable) repository.* *Need for a data sharing charter, possibly in '*Research Data Alliance (RDA)[3]*-style.* *Enhancement of implementation of data as part of peer-review process.*
A lack of interoperability and re-usability of ontologies, datasets, DBs, IT tools and data management procedures.	eNanoMapper to help standardise. *Consistency and compatibility required in databases implementation and data collection formats. Involve ontologies, APIs and possible exchange formats.* *Data managers and end user (data collectors, curators, generators) training and education requirements. Resources, materials, guidance across the whole lifecycle for each and every data domain*

[3] https://rd-alliance.org/about-rda/who-rda.html

The ProSafe partners of T3.3, in particular JRC, have worked on streamlining nanoEHS experimental data recording within the large FP7 project NANoREG (ISA-TAB-Nano logic, see D3.2), and pushed for its recognition in the EU NanoSafety Cluster (NSC), in collaboration with eNanoMapper (ontology mapping, see D3.3). The discussions within NANoREG and in the NSC, with the backing of ProSafe, also aimed at ensuring the possibility of linking the NANoREG dataset with existing and future initiatives, with an eye on long-term sustainability.

The task 3.3 partners have attended several meetings where the issues related to data management and linking data were debated. Examples of the meetings are:

- CODATA-VAMAS workshop in July 2015 on the ontology-related "Uniform Description System for Materials on the Nanoscale",
- Monthly TC calls of the NanoSafety Cluster WG4,
- Several TC calls of the US NanoWG,
- DG RTD-eNanoMapper workshop on 25 January 2016 in Brussels,
- Meetings of the US-EU CoR on Databases and Ontologies,
- GuideNANO-organised event in Leiden: "*Coordination Meeting: Tools/ Frameworks and Databases in Nanosafety*", 16 September 2015,
- A WP3 session on data management at the ProSafe CM3 in Dessau, February 2016,
- Informal EU-US meeting at CM3 on databases and ontologies with representatives of CEINT, eNanoMapper and CEREGE.

T3.3, thanks to the input from T3.1, looked at the landscape of existing attempts by various initiatives to link datasets or extracting data from them, using *ad hoc* IT tools (so-called parsers). As reported in table 12 of D3.1 (see above), the landscape is rather 'empty' and badly organised.

It became evident that having an initiative at the core of the EU NSC, or the wider nanoEHS nanoinformatics arena, and adequately funded with public or public/private money to look at those data/datasets linking issues with a community-oriented service mentality and higher-level nanoinformatics expertise, as eNanoMapper has been doing, is a wise choice.

Ad-hoc tools, such as parsers or semi-automatic applications that convert and transpose data from one dataset to another, can be developed, but a broad-view approach in support to the NSC was adopted by eNanoMapper, as part of its DoW. As explained in the published deliverable D3.1[4], "*the eNanoMapper data architecture was developed and consists of a set of web services, providing access to experimental protocols and data, search service and modules, facilitating linking and data transfer between third party databases*."

eNanoMapper has unfortunately ended, but has managed to set up a database[5] – with possible 'database instances' for p*r*ojects –, using a harmonised data structuring approach and having powerful IT capabilities (using their developed ontologies for nanoEHS and advocating the use of semantic web approaches) to link existing data or datasets.

[4] Philip Doganis, Bengt Fadeel, Roland Grafström, Janna Hastings, Markus Hegi, Nina Jeliazkova, Vedrin Jeliazkov, Cristian Munteanu, Haralambos Sarimveis, Bart Smeets, Georgia Tsiliki, David Vorgrimmler, Egon Willighagen, Barry Hardy. Deliverable Report D3.1 Technical Specification and initial implementation of the protocol and data management web services. (Zenodo, 2015). doi:10.5281/zenodo.375637

[5] Jeliazkova, N.; Chomenidis, C.; Doganis, P.; Fadeel, B.; Grafström, R.; Hardy, B.; Hastings, J.; Hegi, M.; Jeliazkov, V.; Kochev, N.; Kohonen, P.; Munteanu, C. R.; Sarimveis, H.; Smeets, B.; Sopasakis, P.; Tsiliki, G.; Vorgrimmler, D.; Willighagen, E. Beilstein J. Nanotechnol. 2015, 6, 1609–1634. doi:10.3762/bjnano.6.165

eNanoMapper's work on linked data is described in their published deliverable D3.3[6]. It reports on work that led to the Resource Description Framework (RDF) support of the database, reusing the eNanoMapper ontology, and interlinking with other databases.

The eNanoMapper database, now containing (part of) the NANoREG dataset has been linked to a certain extent to the US cancer Nanotechnology Laboratory (caNanoLab) portal[7]. eNanoMapper has also basically demonstrated the possibility to link vendor databases, extracting a minimum set of NM physicochemical characteristics, using a test case based on Sigma-Aldrich NMs.

It is worth noting that preliminary bilateral agreements between NANoREG on the one side and NanoReg2 and caliBrate on the other have been struck in early 2017 to transfer (part of) the NANoREG dataset, with the help of partners that were involved in eNanoMapper.

2.4 Evaluation and conclusions

This ProSafe work in WP3 to find a way to deploy a streamlined data management strategy by combining ISA-TAB-Nano with adequate nanoEHS ontology and, eventually, linking data/datasets has proven useful and has attracted the attention of and stimulated on-going collaboration in the NSC and with the US, in particular with Duke University, which has been linked to ProSafe in WP1. Also the US cancer Nanotechnology Laboratory (caNanoLab) portal have been connected to the eNanoMapper database.

The existence of an IT-oriented project, such as eNanoMapper, to support the EU NSC on nanoEHS data management proved to be crucial. R&I projects on their own have a hard time in 'getting data organised' and thinking of linking/sharing data before the end of the action.

ProSafe, as action intimately related to NANoREG, successfully managed to stimulate the sharing/linking of at least two NSC projects administratively, via the Coordinators (open access, non-disclosure agreements, etc.), and technically, thanks to eNanoMapper and its database for the EU NSC. Links through the eNanoMapper database system have been established and grounds have been prepared for data exchange between NANoREG, NanoReg2 and caliBrate.

ProSafe recommends the European Commission to duly consider ways to further integrate the work promoted by ProSafe on data management, into upcoming strategic nanoEHS R&I funding, and to link this appropriately to the burning issue of data sustainability and curation. The buy-in of stakeholders is essential.

The findings and recommendations of this deliverable serve also as input to the aspects related to data management in the ProSafe White Paper.

They are also valid points worth transferring into the European nanosafety informatics (nanoinfomatics) roadmap being developed under the auspices of the EU NanoSafety Cluster.

The successful implementation of shared/linked nanoEHS data is intimately related to the openness of access to those data. Unless funding authorities (in case of public funding) push for opening up the datasets of projects, even before (part of) those datasets are transferred to a suitable location for sustainable long-term curation, the positive effects on the nanoEHS community of stakeholders of actions geared towards sharing/linking clearly will remain quite limited.

This ProSafe T3.3 work, and WP3 in general, have been a good example of cross-project collaboration: Coordination and Support (ProSafe) with Research and Innovation (eNanoMapper).

[6] Egon Willighagen, Micha Rautenberg, Denis Gebele, Penny Nymark, Pekka Kohonen, Nina Jealiazkova, Barry Hardy. Deliverable Report D3.3 Modules and services for linking and integration with third party databases. (Zenodo, 2016). doi:10.5281/zenodo.375813

[7] https://cananolab.nci.nih.gov/caNanoLab/#/

3 Deviations from the work plan

The work in T3.3 started about a year later than foreseen, since the efforts of its leader were first focused on T3.2 and T3.1. T3.3 had to wait for results from those tasks to be able to start. However, this delay did not impact the rest of the project.

4 Performance of the partners

The partners performed adequately.

5 References / Selected sources of information (optional)

Egon Willighagen, Micha Rautenberg, Denis Gebele, Penny Nymark, Pekka Kohonen, Nina Jealiazkova, Barry Hardy. Deliverable Report D3.3 Modules and services for linking and integration with third party databases. (Zenodo, 2016). doi:10.5281/zenodo.375813

ProSafe deliverables D3.1, D3.2 and D3.3

Philip Doganis, Bengt Fadeel, Roland Grafström, Janna Hastings, Markus Hegi, Nina Jeliazkova, Vedrin Jeliazkov, Cristian Munteanu, Haralambos Sarimveis, Bart Smeets, Georgia Tsiliki, David Vorgrimmler, Egon Willighagen, Barry Hardy. Deliverable Report D3.1 Technical Specification and initial implementation of the protocol and data management web services. (Zenodo, 2015). doi:10.5281/zenodo.375637

The eNanoMapper databadse: Jeliazkova, N.; Chomenidis, C.; Doganis, P.; Fadeel, B.; Grafström, R.; Hardy, B.; Hastings, J.; Hegi, M.; Jeliazkov, V.; Kochev, N.; Kohonen, P.; Munteanu, C. R.; Sarimveis, H.; Smeets, B.; Sopasakis, P.; Tsiliki, G.; Vorgrimmler, D.; Willighagen, E. Beilstein J. Nanotechnol. 2015, 6, 1609–1634. doi:10.3762/bjnano.6.165

-oOo-