



**The European Nanotechnology Community Informatics Platform: Bridging data and disciplinary gaps for industry and regulators**

Grant Agreement No 731032

**Deliverable Report 10.1**

<b>Deliverable</b>	D10.1 Initial draft of data management plan (Open data pilot)
<b>Work Package</b>	WP10: NA4 - Integration & Sustainability
<b>Delivery date</b>	M6 - 30 June 2018
<b>Lead Beneficiary</b>	Universiteit Maastricht (UM)
<b>Nature of Deliverable</b>	Report
<b>Dissemination Level</b>	Public (PU)
<b>License</b>	Creative Commons Attribution 4.0 International

**Submitted by** Anastasios Papadiamantis (UoB), Lucien Farcas (DC) and Egon Willighagen (UM)

**Revised by** Iseult Lynch (UoB) and Thomas Exner (DC)

**Approved by** Iseult Lynch (UoB)



This project has received funding from European Union Horizon 2020 Programme (H2020) under grant agreement n° 731032.

---

## Table of contents

Summary	4
Introduction	7
Data set description	7
Data sharing	8
Archiving and preservation	8
1. Data summary	11
1.1 Purpose of the data collection/generation	11
1.2 Relation to the objectives of the project	11
1.3 Types and formats of data	12
1.3.1 Types of Data	12
1.3.2 Formats of Data	13
1.4 Reuse of data	15
1.5 Origin of the data	18
1.6 Expected size of the data	18
1.7 Utility of data and models	19
2. FAIR data	21
2.1 Making data findable, including provisions for metadata	21
2.1.1 Making data discoverable, including provisions for metadata	21
2.1.2 Identifiers and naming conventions	22
2.1.2.1 Disseminate testing material identifiers	22
2.1.2.2 Use IRIs or Compact Identifiers in dissemination	22
2.2 Making data openly accessible	23
<b>2.2.1 Open Data</b>	23
<b>2.2.2 Free Access</b>	23
2.3 Making data interoperable	24
2.3.1 Supported data exchange formats	25
2.4 Increase data reuse (through clarifying licenses)	26
3. Allocation of resources	27
4. Data security	29
5. Ethical aspects	29
Final remarks	31
References	33

---

Appendices	34
Appendix A: RDM Copyright, License, and Waiver Clearance Form	34
Appendix B: NanoCommons Customized RDM Online Plan	35

---

## Summary

This deliverable presents the initial data management plan (DMP) for the starting community research infrastructure NanoCommons. It is a living document that will be updated over the course of the project, with the final plan to be delivered at the end of the project. We anticipate that at least two other versions of the plan will be developed, to accommodate the fast-changing field of Open and FAIR (Findable, Accessible, Interoperable and Reusable) data in Europe. The initial DMP covers a significant part of the life cycle of research data, but cannot cover its life after the end of this project. It covers the initial process of thinking about how research data will be captured and handled during the research, networking and transnational activities (JRA, NA and TA, respectively), the follow up process of making it Open and FAIR, and finally the long-term deposition of the data (and models and other outputs), to ensure a life after the end of the NanoCommons project. Note that there are dedicated activities developing a sustainability plan for the complete set of NanoCommons tools/services and approaches (i.e. the research infrastructure itself, including the NanoCommons Data Warehouse (data repository), which will ensure long term access under the FAIR principles and will be made available to current and future nanosafety projects and other stakeholders. The NanoCommons Data Warehouse may also be used as an application for an Advanced Community, and application for establishment of a European Research Infrastructure Consortium (ERIC), via Deliverable report D10.3 “NanoCommons Sustainability Plan”. The NanoCommons data management strategy follows the description-of-action (DoA), (Section 29.3, page 49) which prescribes access to all research data produced in the project, including that produced with Users of the TA facilities. Therefore, this deliverable will also describe suggestions for how the project partners and TA users can make data available in both Open and FAIR formats. Note that the NanoCommons DMP builds on DMP’s from related projects such as OpenRiskNet and adapts it for other types of data, such as that generated by Transnational Access (TA) Users.

### List of abbreviations

APIs - Application Programming Interface  
CEN - European Standardisation Organisation  
ChEBI - Chemical Entities of Biological Interest  
CLW - Copyright, License, Waiver  
DoA - Description of Action  
DMP - Data Management Plan  
EC - European Commission  
ECHA - European Chemicals Agency  
EFSA - European Food Safety Authority  
EFO - The Experimental Factor Ontology  
ENM - Engineered Nanomaterial  
EMA - European Medicines Agency  
EOSC- EU Open Science Cloud  
ERIC - European Research Infrastructure Consortium  
EUON - European Union Observatory for Nanomaterials  
FAIR - Findable, Accessible, Interoperable, Reusable  
GDPR - General Data Protection Regulation  
HGNC - Human Genome Nomenclature Committee  
IPR - Intellectual Property Rights  
ISO - International Standards Organisation  
JRA - Joint Research Activities  
JRC - Joint Research Center  
MoU - Memorandum of Understanding  
MPATH - Mouse Pathology Ontology  
NA - Networking Activities  
NCBI - National Center for Biotechnology Information  
NCIT - National Cancer Institute Thesaurus  
NEMs - New and Emerging Materials  
NM - Nanomaterial  
OAE - Ontology of Adverse Events  
OBI - Ontology for Biomedical Investigations  
OECD - Organisation for Economic Cooperation and Development  
ORD - Open Research Data Pilot  
PATO - Phenotype and Trait Ontology  
QA - Quality Assurance  
QC - Quality Control

---

## D10.1 Initial draft of Data management plan (Open data pilot)

---

QSAR - Quantitative Structure-Activity Relationships

R&D - Research and Development

RDM - Research Data Management

RDMO - Research Data Management Online (Plan)

RDS - Research Datastore

SME - Small to Medium Enterprise

SOP - Standard Operating Procedure

TA - Transnational Access

UO - Units of Measurement Ontology

WWTP - Working Party on Manufactured Nanomaterials

---

## Introduction

The European Commission (EC) is running a flexible pilot under Horizon 2020 called the Open Research Data Pilot (ORD Pilot). The ORD pilot aims to improve and maximise access to, and reuse of research data generated by Horizon 2020 projects and ask projects to think about the totality of their data in all forms, taking into account the need to provide open scientific information, undertake commercialisation activities and Intellectual Property Rights (IPR) management, address data privacy concerns and ensure data security, as well as data management and preservation questions [1]. Open data is data for which everyone has the rights to access, reuse, repurpose, and redistribute. The ORD Pilot aims to make the research data generated by selected Horizon 2020 projects accessible with as few restrictions as possible, while at the same time protecting sensitive data from inappropriate access [2]. Projects starting from January 2017 are by default part of the ORD Pilot, including the Research infrastructures and e-Infrastructures, and as such are required to develop a Data Management Plan (DMP).

To help optimise the potential for future sharing and reuse of data, the NanoCommons DMP helps the project partners and Transnational Access (TA) Users to consider any problems or challenges that may be encountered in making their data Open and FAIR and helps them to identify ways to overcome these. This DMP is a “living” document outlining how the research data collected or generated will be handled during and after the NanoCommons project. It follows the Guidelines on FAIR Data Management in Horizon 2020 [1] and is based around the realistic resources available to the project partners and TA Users taking the current knowledge into account. The ongoing activities to keep the DMP up to date will follow an online, distributed approach as outlined in the Guidelines for creating an online DMP [3]. As part of our community building and sharing of best practices approach, the NanoCommons initial DMP will be made publicly available to other nanosafety and nanoinformatics projects to reuse, and all updates will be announced widely to the relevant communities (including OpenRiskNet).

## Data set description

This section refers to what kinds of data NanoCommons will collect and/or generate, and to whom these data might be useful later. The data set refers to:

- The data and metadata needed to validate results in scientific publications;
- The data and metadata needed to develop and validate the predictive *in silico* models for nanosafety, via the Joint Research Activities (JRA);
- Data and metadata generated by Users through TA activities; and,
- Other curated and/or raw data and metadata that may be required for validation purposes or with reuse value.

Further, these questions are addressed in order to determine the potential reuse value of the data:

- What is the data about?
  - Who created it and why?
  - In which forms it is available?
  - What (if any) standards were applied in generating the data?
-

The metadata provided with the datasets answers such questions to enable data to be found and understood, ideally according to the particular standards applied. Finally, the metadata, documentation and standards will help in making the data FAIR (Findable, Accessible, Interoperable, and Reusable) [4-6].

## Data sharing

According to the ORD Pilot programme, by default as much of the resulting data as possible should be archived as Open Access. Therefore, legitimate reasons for not sharing resulting data should be explained in the DMP. However, data protection or IPR agreements should not be compromised in any way, and data sharing should be done responsibly. Therefore, the DMP describes any ethical or legal issues that can have an impact on data sharing.

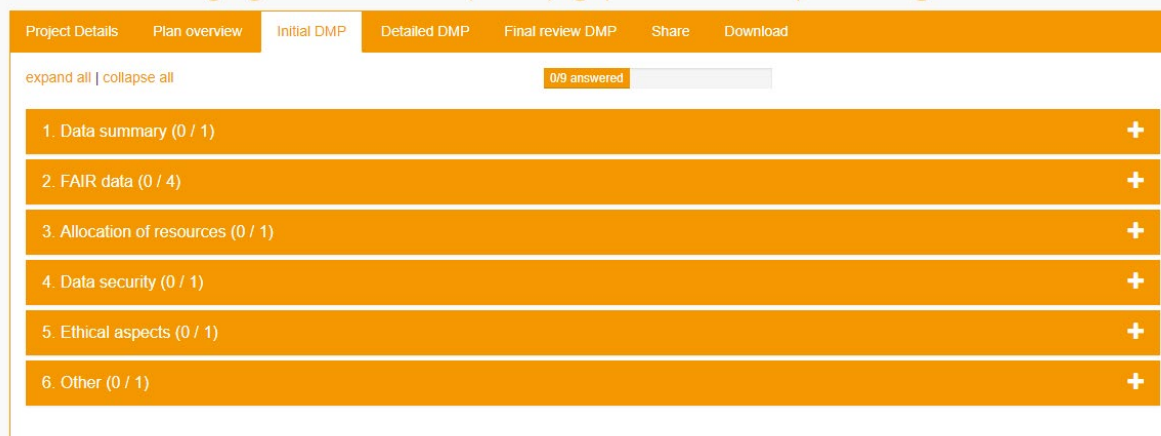
## Archiving and preservation

To ensure that publicly funded research outputs can have a positive impact on future research, for policy development, and for societal change it is important to assure the availability of data for a long period beyond the lifetime of the project. However, this does not refer only to storage in a research data repository, but also the need to consider the (re-)usability of the data. One of the main goals of the research infrastructure being created by the NanoCommons project is to harmonise nanosafety data and make it interoperable and sustainable, as a means to facilitate *in silico* nanosafety assessment thereby reducing the need for costly and ethically questionable *in vivo* studies while supporting the route to market of nano-enabled products. Therefore, the project has a special obligation to preserve both the software tools and any code produced to perform specific *in silico* analyses and the data underpinning these models and tools. This will also require a high degree of clarity about any proprietary or open source tools that will be needed to validate and use the preserved data. The general ethos of NanoCommons is that all datasets imported into the NanoCommons Knowledge Base using funds provided by NanoCommons will be made both Open and FAIR. The agreement with TA Users, previous projects who will provide datasets, and any other data sources utilised in the delivery of the NanoCommons research infrastructure will include specifications on the Open and FAIR provisions, as per Appendix A.

The structure of the NanoCommons DMP follows the Horizon 2020 ORD Pilot instructions on how to create a DMP. Therefore, additional to this document, the plan is stored and updated online using the recommended DMP tool (Figure 1) available at [dmponline.dcc.ac.uk/](https://dmponline.dcc.ac.uk/) (see also Appendix B).



## NanoCommons - The European Nanotechnology Community Informatics Platform: Bridging data and disciplinary gaps for industry and regulators



**Figure 1.** Screenshot of the tool used to create the NanoCommons DMP.<sup>1</sup>

### The NanoCommons Research Data Management Plan

The Data Management Plan (DMP) is essential to ensure the capture of return-on-investment (RoI) for the resources put in by the European Union, via the EC's H2020 research framework. An RDM strategy does not describe the data, but it describes the processes of the management of the data throughout its entire lifecycle. The NanoCommons ethos is that Data should be both FAIR and Open, and all our activities will expressly utilising formats that support this. NanoCommons will build on the lessons and experience from the NanoREG project, which utilised a General Assembly meeting to decide on an open licence for specific outputs if the Consortium Agreement did not specify the reuse conditions explicitly.

Before we continue, we should define to what kind of data this DMP applies. Section 1.3 provides more detail of the types of data to be imported into or generated within NanoCommons, but the starting point is basically any data produced in the NanoCommons project, which includes, but is not limited to, experimentally measured data and computed data via modelling or statistics, including that produced by TA Users as part of their access provision.

#### Expectations

The NanoCommons DoA describes the goal of this deliverable (due at Month 6):

*A specific strategy to provide open access to data will be developed according to the recommendations of the [Horizon2020 ORD Pilot](#). The strategy will include the development of initial and final FAIR DMPs (M6 and M48 respectively), which will outline and support implementation of procedures to ensure full access to data incorporated within NanoCommons.*

<sup>1</sup> <https://dmponline.dcc.ac.uk/plans/26512>

However, a second section of our DoA is essential, and is part of our agreement (contract) with the EC, which describes in more detail how NanoCommons is expected to provide access to third-parties of data produced within NanoCommons (emphasize ours):

### *29.3 Open access to research data*

*Regarding the digital research data generated in the action ('data'), the beneficiaries must:*

- a. **deposit** [it] in[to] a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and **disseminate** — **free of charge** for any user — the following:*
  - i. the data, including associated metadata, needed to validate the results **presented in scientific publications** as soon as possible;*
  - ii. other data, including associated metadata, as specified and within the **deadlines** laid down in the 'data management plan';*
- b. provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and — where possible — provide the tools and instruments themselves).*

*This does not change the **obligation to protect results** in Article 27, the confidentiality obligations in Article 36, the security obligations in Article 37 or the obligations to protect personal data in Article 39, all of which still apply. As an exception, the beneficiaries do not have to ensure open access to specific parts of their research data if the achievement of the action's main objective, as described in Annex 1, would be jeopardised by making those specific parts of the research data openly accessible. In this case, the DMP **must contain the reasons for not giving access**.*

These formal expectations set aside, the project has quite some liberty in setting our own expectations, and these may even be different from one partner to another. All partners have their legal obligations and the NanoCommons project is a collaboration into which we all invest and from which all partners will harvest knowledge and data. As such, enabling each other is our best bet to result in a successful innovation project that has **maximum impact** for partners and for the wider nanosafety research and nanoinformatics communities. Clear mutual expectations are an essential component of that.

## 1. Data summary

**Summary of the data addressing the following issues:**

- *State the purpose of the data collection/generation*
- *Explain the relation to the objectives of the project*
- *Specify the types and formats of data generated/collected*
- *Specify if existing data is being reused (if any)*
- *Specify the origin of the data*
- *State the expected size of the data (if known)*
- *Outline the data utility: to whom will it be useful*

### 1.1 Purpose of the data collection/generation

The data collected/generated within the NanoCommons Project aims to provide a harmonised and streamlined workflow covering all aspects of the data lifecycle, which includes experimental planning, data acquisition, manipulation and analysis, data storage and Open and FAIR access when possible (for further information please refer to Appendix B). The collected/generated data will be used to meet the project's objectives in developing and openly providing the *in silico* nanosafety modelling and nanoinformatics tools needed by the nanosafety scientific community for the production of high quality, harmonised and sustainable data. To achieve these objectives the NanoCommons project aims to implement appropriate standardised experimental procedures and data curation techniques, to promote ontology harmonisation and produce datasets aligned with the highest standards of scientific quality within the project.

Thus, the data collected/generated within NanoCommons will be available in the correct format and readily available to be implemented and used for analysis via the tools developed within the project. At the same time, data harmonisation will allow data interoperability and the combination of different datasets (from a range of sources, as long as the conditions for data sharing are compatible with the needs of NanoCommons - see Appendix A for details), where applicable and following agreement of the respective owners, that will allow more complex analyses to be performed and can also promote cross-project/field cooperation and translational research.

### 1.2 Relation to the objectives of the project

NanoCommons is driven by the European nanosafety, nanomedicine and emerging materials research and regulatory communities need for an e-infrastructure providing a standardised, reproducible and interoperable way to access all available data, knowledge and analysis and facilitate the application of nanoinformatics and modelling tools that have been adapted and verified as suitable for application to nanomaterials.

More than 10 years of nanosafety research has delivered tangible insights into the key science and policy required for the development of safe nano-enabled products. However, this knowledge has yet to be systematised, or made "FAIR", in a manner that allows:

- modellers to develop predictive frameworks and assess their domains of applicability,
- industry to utilise the data, models and tools for safe-by-design strategies or as supporting evidence for use in regulatory dossiers,
- regulators to compare one form to another or make estimations of data requirements for New and Emerging Materials (NEMs) based on shared properties with NMs, or
- educators to utilise in teaching toxicology, ecotoxicology, environmental fate and modelling of the behaviour of ENMs.

To address this gap, NanoCommons will create an openly accessible e-infrastructure serving the current and future (unmet) needs of the key research communities and pivotal industrial users and regulators.

NanoCommons will facilitate:

1. the efficient collection, integration, curation and maintenance of existing data and methods along with development and optimisation of the tools and user interfaces to interrogate them (JRA),
2. the provision of access to the data, methods and tools collected or produced under the project, along with expert guidance in their use and in experimental design and workflows to harmonise data quality into the future (TA), and
3. community building including bridging disciplinary gaps (e.g. toxicology and ecotoxicology, experimental and modelling), promoting best practice in data quality (e.g. Quality Assurance (QA) audits, Independent Experimental Data audits), and development of User case studies demonstrating the capability of the NanoCommons infrastructure to address real stakeholder challenges in partnership with industry & regulators (NA).

NanoCommons is designed to integrate the Knowledge infrastructure for risk assessment of novel and emerging materials on a European and international scale, and provide (remote) access to data, data mining, modelling and risk assessment tools to all European researchers, from academia and industry, as well as regulators, ensuring their optimal use and joint development.

## 1.3 Types and formats of data

### 1.3.1 Types of Data

An important distinction should be made here between various forms of data to be generated and utilized within NanoCommons:

1. Raw or experimental data
  2. Derived (processed or computed) data
  3. Data associated with formal publications (literature curated data) (This is typically summarised data, from which the original data cannot really be reconstructed).
  4. Interactive data, whereby the data is put in the context of other data. For example, integrated with other data sets, etc.
-

The first type of data is basically what comes directly from the instruments utilised in the experimental assays or the computational calculations. Ideally this follows the design of a pre-registration, in which the design of the experiment is published before the experiment is performed (see *e.g.* the Center of Open Science Pre-registration Challenge: [cos.io/prereg/](https://cos.io/prereg/)). The second type is data derived from the raw data, and aimed at making comparisons between experimental conditions, drawing conclusions, and other kinds of use and reuse of the data. This data may be stored as spreadsheets, but also in many other formats.

The third and fourth types of data are more about presentation of the data: the third kind is the presentation in formal journal applications and, for example, presented in tables as PDFs, spreadsheets, or data files. The fourth kind is particularly interesting and important to our project: data must be interactively available to enable reuse. For example, our modelling tasks depend on data to be available in a FAIR way (see below).

The Open Science expectations around data basically apply to all four kinds of data. However, it is clear that different solutions are needed for the different kinds. This is one reason why writing a clear DMP is non-trivial. The use of electronic notebooks (for both experimental and computational workflows) will ensure that data of types 1 and 2 above will be collected in a harmonised, ontology-linked, and database-compatible manner from the outset, thereby integrating data management with data generation, rather than data management being an add-on activity after associated with, for example, publication requirements for datasets to be deposited in appropriate databases. Free and commercial tools for keeping an electronic notebook are both available.

### **Data Life Cycle**

The Data Life Cycle includes the entire process through which data is generated, acquired, analysed, manipulated and made available through publications and/or data repositories along with the metadata produced during the entire process. The fourth kind of data presented in the previous paragraph can also be seen as an overview of the life cycle of data. However, it should be noted that the availability of data does not constitute the start of the cycle: instead, the design of an experiment is a more appropriate start - see Appendix B for further details of the data life cycle as utilised in NanoCommons.

### **1.3.2 Formats of Data**

The NanoCommons project aims to mainly use the following 3 data formats, which are broadly accepted by different subgroups of the nanoinformatics community. Each has advantages and disadvantages, which NanoCommons aims to overcome through suitable modifications, that will be documented in full detail within the subsequent deliverables and published metadata. In any case, the nature of the NanoCommons project makes it possible that more data formats will be implemented to facilitate the different needs of the e-infrastructure users, and will be added to subsequent versions of this document as the needs arise.

## 1. ISA (-tab or -json)

- a. The advantage of use of the ISA type templates is the flexibility they demonstrate with respect to creation and design for any type of experiments and the addition / subtraction of columns to fit all experimental needs (methods, descriptors etc.).
- b. Disadvantages of the ISA- templates include the lack of description of the file formats for both data and metadata. They are also not necessarily linked with specific ontologies and thus the naming of the columns is not regulated and protocolled. At the same time, only the file names, and not the file types, are generally available. The latter disadvantage could be potentially overcome with the use of an interoperability layer within the data management infrastructure, which will describe the dataset using a specific ontology and taxonomy.
- c. The ISA-(TAB/JSON)-nano add-on is currently not as developed [7], but is the started norm in the nano-field, although is rarely applied to the letter. ISA- (TAB/JSON)-nano is being promoted by the US National Cancer Institute and the eNanoMapper data portal.

## 2. NIKC

- a. The advantages of the NIKC template, which is an extension of the ISA-tab nano, are its dynamic nature and design flexibility, being able to facilitate all aspects of nanosafety research (e.g. nanomaterials characterization including transformations and ageing, exposure, environmental as well as human hazard and risk assessment, regulatory), while accepting the attachment of images. The currently developed NIKC/NanoCommons ontology, which builds on the eNanoMapper ontology, also ensures harmonised data curation, the use of specific nomenclature and the ability to combine available datasets, test for experimental and data gaps and increase data quality. The NIKC *instance* approach is also a big step forward in conceptually representing nanomaterials data as it recognises the dynamic nature of nanomaterials and their interaction with their surroundings. NIKC considers the nanomaterial as single entity with its surrounding environment “demanding” the characterisation of both to be complete. It also offers the possibility to create a data tree that includes experimental workflows, from different aspects of nanosafety research that can link back to the same nanomaterial (when applicable) allowing data interoperability and translational research. Finally, another novelty of the NIKC templates is the use of both protocols and instrument settings / types as data points. This allows both the identification of data mismatches based on slight experimental deviations, as well as potential differences between descriptors being measured using different instruments.
- b. The disadvantages of the NIKC templates are that are complicated and time consuming to design and build. They also span at several different tabs that require the logging of a large amount of information. As a result, they need extended training from experienced curators that are also tasked to QC the produced templates. Some of the issues could be circumvented with the development and use of an automated system of template creation, based on a questionnaire system and having a pre-defined set of nanomaterials characterisation template. Additionally, as more and

more study templates are developed, reuse will be possible with potentially minor tweaks.

### 3. ToxML

- a. The advantage of the ToxML data format is the robust design of the files, which also simplifies interoperability. The latter, in certain cases, can also be considered as a disadvantage.
- b. The drawbacks for the use of ToxML is the difficulty to implement changes, which have to be proposed and then reviewed and accepted by the supervisory board making the process extremely time consuming. As a result, the NanoCommons consortium should make provisions for the implementation of non-standard entries and add the necessary tools to update the entries when the required standard covering them is made available.

In all cases, the usage of a common ontology for data description and curation can guarantee that the both the data and produced metadata can be understood by both users and machines and can be automatically transferred between services.

## 1.4 Reuse of data

NanoCommons aims to use the data collected/generated through the project, as well as high quality curated peer-reviewed published data. In terms of tools, the project will identify and use already existing tools (modelling, Quantitative Structure-Activity Relationships (QSARs), Omics etc.) and will work to improve them further using the data collected/generated through the project. The NanoCommons project also aims to harmonise, combine and analyse existing datasets to improve and refine scientific findings and conclusions, promote cross-project/field collaboration and promote translational research.

For data templates, SOPs and decision trees, we will use, as much as possible, existing data, software tools, open and readily available to all partners. We will aim to produce reusable and extensible tools, and to make use of existing nanomaterials safety and characterisation data, where possible.

SOPs developed within NanoMILE and NanoFASE for characterisation of NMs in biological and environmental matrices are being reused where possible, and any changes documented as part of the metadata and domains of applicability sections of the SOPs. For example, in the development of SciNote online notebooks, we are utilising existing lab data and lab protocols.

Table 1 below summarises the datasets that are potentially being reused within the context of NanoCommons.



**Table 1:** Existing data generated through other NanoSafety Cluster (NSC) projects and other appropriate projects will be used by the NanoCommons project through collaborative efforts or agreements with the relevant project beneficiaries. This data will in most cases be for use to demonstrate the value to users of data sharing (open and FAIR data) and to facilitate the application of the range of modelling tools being developed. They will also be utilised for analysis of reproducibility of the overall method via comparison of different measurement approaches.

Project	Description of Data	Purpose in NanoCommons
NanoMILE	Characterisation data and calculated descriptors for up to >75 different NMs via the NanoMILE NMs library. Data identifying ENM property factors important for hazard, and representative ENMs that range in those properties.	To provide baseline characterisation data for a wide panel of NMs, that will be used to parameterise the domain of applicability of the various computational models, as well as for benchmarking of the newly developed models and tools.
NanoFASE	Fate process relevant data (e.g. separation between sludge and effluent in WWTPs) and characterisation of the transformations of NMs in the different environmental compartments. Data on microbial degradation of commonly applied coating materials.	Considerable method development underway, specifically for functional assays to characterise the transformed states of nanomaterials in different environmental compartments. This data will be utilised to demonstrate the validity and utility of the NIKC instances approach, and how the concept expands dramatically the scale of the data available, but also the predictions that can be made utilising the data.
NanoPolyTox, GUIDEnano, SUN	Release methods and data for NMs enabled product types especially of worked case studies.	Release data will be utilised to support the NanoCommons risk assessment tools (models) and will provide methods, SOPs and datasets for benchmarking the computational tools. Any of these projects could also service as Case Study for data importation into the NanoCommons data warehouse, allowing all required steps to be fully



		documented and described for subsequent data uploaders.
NIKC	Characterisation data of nanomaterials through their entire lifecycle in conjunction with characterisation of the surrounding environment. Anticipated use scenarios, matrix, concentration in products, system characteristics (environmental, biological, laboratory, etc.). Exposure and hazard measurements, calculations, and estimates and metadata associated with each of these, including bibliometrics, protocols, equipment, temporal and spatial descriptors, etc.	The currently curated hazard and exposure data will be utilised to support the NanoCommons hazard and exposure modelling. Over time the expansion of the NIKC towards all aspects of nanosafety research (e.g. fate, risk assessment, regulatory), will broaden data utilisation and interoperability and will also promote translational research. The use of experimental protocols and instruments as data points, will also allow the development of data QC tools to test for data completeness and for protocol optimisation and standardisation.
NANoREG	Characterisation data for the JRC repository test NMs, and the templates for data collection for selected characterisation end-points.	The characterisation data are deposited in the eNanoMapper database, and an MoU will be sought to access the characterisation data as part of the Model benchmarking and Domain of applicability assessment.
eNanoMapper	Ontology terms, database structure for nanosafety.	NanoCommons will utilise the existing ontology and extend it with a range of characterisation terms to support the implementation of the decision tree and analytical toolbox.

<p>Projects funded under NMBP-14-2018 (nanoinformatics call)</p>	<p>The projects funded under this call will be specifically developing nanoinformatics models, and will also require a knowledge warehouse, APIs, and large-scale datasets, so we will provide access to NanoCommons developments to facilitate further reuse.</p>	<p>By facilitating access to and reuse of datasets that it has compiled, curated, integrated and aligned, NanoCommons will be serving the community needs, and more reuse builds a stronger case for subsequent application for funding as an Advanced Research Community.</p>
--	--	--

## 1.5 Origin of the data

The data sources and offered tools through the project will take into account the original licences for the versions integrated into the NanoCommons infrastructure, if applicable. For data and tools created directly through the project, a respective licensing system will be developed. Similarly, any commercial, open source, or freeware software requiring registration and licensing will be handled in a similar way. Both licensing systems will run as a single entity through an authentication and authorisation service run through DC and/or Biomax [8], taking into account European GDPR law.

It is anticipated that the data and tools produced/developed/integrated into the NanoCommons infrastructure, along with any supporting metadata, documentation and source code where applicable, will include:

1. Data, models and tools developed and owned by NanoCommons Partners will be assigned with a Creative Commons licence, allowing their full and free reuse, modification, and redistribution, where applicable, as long as their origin is cited appropriately as far as possible. When only non-commercial reuse is allowed, then a specific, well-justified case needs to be submitted and approved by the coordinator under the terms of the Consortium Agreement;
2. Open Source data, tools and models, used the license mentioned by the owners;
3. Data from third parties, and not yet available in existing open databases used under the conditions specified by the data owner and included in a formal agreement.

## 1.6 Expected size of the data

The data generated/collected through the project's open calls, and produced through partner and user collaboration, will be in the region of **10s to 100s Terabytes** and will consist of raw, analytical and metadata, and the databases to support the project's actions. Data access, during and beyond the project's life cycle, will be facilitated through processes that will ensure that all data will be stored in the project's centrally managed datastore, i.e. the NanoCommons knowledgebase to be handled by Biomax, which will also make it easily FINDable. The data will be backed up using a number of online accessible mirrors, hosted by different partners, to ensure continuous online time, access and security. Such a process will also ensure future data reusability, even in cases unforeseen from the original data owners/providers. Although such practices put extra effort on project partners, it is considered to be highly significant avoiding the need and difficulty to search for data stored locally when needed.

Current partners experience with the implementation of such strategies (e.g. DC are going through this process in the EU-ToxRisk project at the moment, while UoB are back implementing this approach for all our previous project data), will prove essential to streamline and simplify the process as much as possible for other partners.

In the case when data or tools available online from external sources will be required accessibility will be achieved through its original source, along with the implementation of a harmonisation system layer to ensure data and tool interoperability with that of the NanoCommons knowledge base. Thus, no additional storage capacity will be needed saving a significant amount of storage, maintenance and access cost. In the case when data or tools are not publicly available or do not comply with the FAIR principles, the consortium will negotiate, if possible, with the data/tools owners for the data to be transferred to standard data repositories or if the existing solution should be improved within the framework of the associated partner programme.

The data to be made directly publicly accessible is the description of the models and tools themselves, along with the SOPs, analysis of uncertainty, domain of applicability and benchmarking against other relevant methods / models / tools. Guidance on the use of the different models and approaches, appropriate input data and how to generate the appropriate input data, and data capture templates will also be developed and integrated. The size of this data may be comparatively limited and thus easier to make interoperable and reusable.

A key aspect of the NanoCommons approach is the alignment with, and utilisation of the resources being developed in the [OpenRiskNet e-infrastructure](#), coordinated by NanoCommons partner DC and in which several NanoCommons partners are also involved (UoB, UM, are partners, Biomax is an associate partner). A key aspect of OpenRiskNet, which will also be adopted by NanoCommons, is not to combine data from different sources into one data warehouse but rather to access the data from its original source and use the interoperability layer added to the data services to harmonise them. In this way, no additional capacity for data storage is needed for data external to the NanoCommons. However, some of the data considered for integration is not yet available in open-accessible databases or these don't comply with the FAIR principles. In such cases, NanoCommons will work with the data owner (who can apply to become a NanoCommons User) to agree the term of access and reuse. Only in cases where the terms comply with NanoCommons policies (FAIR and Open) will budget from the project be expended to integrate and communicate with the dataset. Appendix A outlines our current RDM Copyright, License, and Waiver Clearance Form, which all users and data integrators will have to sign prior to putting their data into the NanoCommons system.

## 1.7 Utility of data and models

FAIR data are at the core of NanoCommons vision and philosophy. As a result, the project aims to make the data, tools and services produced, developed and offered by the project accessible and beneficial to all stakeholder of the nanosafety community (researchers, modellers, regulators, industry and especially SMEs). To achieve that, the project's outputs need to have the appropriate metadata and unique identifiers, which will make them "Findable" and raw data and metadata should be stored in a data repository in formats that are Accessible and understandable by both humans and machines. In parallel, the data and metadata formats need to be harmonised to make them

---

Interoperable and Reusable through a straightforward tiered licensing system. For this reason, NanoCommons will create streamlined and user-friendly data management process, addressing the whole data life cycle and is provided to the community (Users) which will provide the offered services under a complete package of easily accessible, standardised and harmonised services in order to be able to produce high quality and complete scientific results and conclusions.

The offered services can be used in a way to maximise the potential of subsequent use in the standardisation of novel approaches, via OECD, CEN, ISO, *safer-by-design* etc. further supporting the utilisation by academics, industry and regulators. The harmonised interoperability between academia and industry can provide a wide amount of data allowing innovative R&D and computational activities and lowering the barriers of real innovation resulting in new safer and secure products, processes and services. The close cooperation with the 3<sup>rd</sup> pillar of nanosafety, regulators, is also key to promoting the wider acceptance of the produced workflows, tools and services.

Potential beneficiaries of the data, tools and the NanoCommons infrastructure:

- ❖ Academics, at all levels, working in all fields of nanosafety research and the wider toxicity community. The services offered can help them uncover underlying research patterns and reach new scientific conclusions.
- ❖ Regulatory agencies (e.g. ECHA, EMA, EFSA) and policy makers.
- ❖ SMEs that do not have the resources or the knowledge to develop and use *in-house* tools for *safer-by-design* approaches and risk assessment requirements.
- ❖ Industry and the R&D community, which can use the offered services to address the '3Rs' principles and let them design novel and safer experimental approaches.
- ❖ Consumers, through the interoperability of all of the above that will offer them new safer products containing nanomaterials.

## 2. FAIR data

The FAIR principles refer to a number of features that data, software, etc. should have to maximize their value and societal impact [9]. They are grouped into four categories, as given before. Each of the four aspects of the principles will see a different way it is implemented for that kind of data. For example, in some cases, *raw data* may not be findable to people outside NanoCommons until the primary publications are completed, and interoperability mostly applies to metadata. For *data associated with formal publications* it must be publicly available and interoperable at a very high level to benefit the community. Similarly, when computational approaches are taken into account, FAIR raw data is essential for any beneficial scientific effects.

How these principles are implemented, how they are used, is totally up to the user. They have been defined quite broadly so that apply to different kinds of scientific output. This has led to confusion how to make your data FAIR. In fact, it is not a black-and-white situation, but there are many shades of grey. The point is that data should be as FAIR as possible. This in turn suggests there is a scale of FAIR-ness, and metrics have been proposed [4].

### 2.1 Making data findable, including provisions for metadata

- *Outline the discoverability of data (metadata provision)*
- *Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?*
- *Outline naming conventions used*
- *Outline the approach towards search keyword*
- *Outline the approach for clear versioning*
- *Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how*

#### 2.1.1 Making data discoverable, including provisions for metadata

This principle prescribes that the output must be findable. That is, effort must be made to make sure people can find the data. To be findable, the principles specify:

- *F1. (meta)data are assigned a globally unique and persistent identifier*
- *F2. data are described with rich metadata (defined by R1 below, in section 2.4)*
- *F3. metadata clearly and explicitly include the identifier of the data it describes*
- *F4. (meta)data are registered or indexed in a searchable resource*

To meet these principles the project aims to use data repositories to store the data generated/collected by the Project and the tools developed from the Consortium. This will allow the datasets and tools to be easier to find and access. The repositories will also include an appropriate licensing system to allow a layered access to the available data based on the accessibility decided by the data and/or tools owners. Both datasets and tools will be complemented with the appropriate metadata (and source code in the case of software tools whenever possible) and unique repository and/or DOI identifiers to allow users to easily query and reference. The Project also aims to establish

collaborations with existing and under development data repositories and link them together through a joint query system. To achieve this, necessary harmonisation between the repository outputs needs to take place to allow the queried results to be returned in a consistent manner, which will also promote interoperability.

As a result, NanoCommons aims to develop a data and tools repository system which will allow:

1. accessing a query system with a filtering functionality;
2. to identify the specific dataset's/tool's licensing and access information, desired data fields to be searched (e.g. ENM characterisation parameters, toxicity results, tools source code);
3. a harmonised data exchange file format; and,
4. full access to the necessary metadata, which will describe the experimental setup, the analytical tools used and a summary of the research outputs.

To achieve that harmonised output, NanoCommons will use a joint workflow system implemented in online lab-books using the NanoCommons ontology, which will allow continuous sharing of protocols, data and tools between project partners and users. This will allow the data collected/generated and the tools developed through NanoCommons to be stored in data repositories using a common format and will also make the metadata creation faster. In all cases, the created metadata will be assigned with a unique identifier and version number so it can be easily findable and citable from potential users.

### 2.1.2 Identifiers and naming conventions

Of particular importance are the use of identifiers here. As many aspects of the experiments should be linked to identifiers, including but not limited to: ENMs, cell lines, bioassays, species, genes, proteins, and metabolites. Identifiers should come from internationally recognized databases or resources whenever possible. For example, gene identifiers should be ideally from Ensembl or NCBI Gene, or from the HGNC for human genes. For proteins UniProt would be a good resource. For nanomaterials, the proper JRC identifiers (JRCNMxxx) must be used (the NM-xxx type identifiers only apply for very old batches). If nanomaterials do not have a globally unique identifier, the NanoCommons project can provide unique identifiers. Ontology identifiers are first class supported identifiers.

#### 2.1.2.1 Disseminate testing material identifiers

Similar to the idea of pre-registration of clinical trials, each project should disseminate the identifiers they have chosen to represent the nanomaterials selected to be tested or used in that project. These identifiers must be publicly announced as soon as a material is selected, widely distributed within the project, disseminated to other projects, and consistently used in all reporting, experimental protocols, results, metadata, etc.

#### 2.1.2.2 Use IRIs or Compact Identifiers in dissemination

To ensure clear and explicit semantics, when using identifiers in dissemination, either full identifiers should be used, such as Internationalized Resource Identifiers (IRIs) or, alternatively, Compact Identifiers [10].

---

## 2.2 Making data openly accessible

- *Specify which data will be made openly available? If some data is kept closed provide rationale for doing so*
- *Specify how the data will be made available*
- *Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?*
- *Specify where the data and associated metadata, documentation and code are deposited*
- *Specify how access will be provided in case there are any restrictions*

This principle prescribes that data must be accompanied with metadata that explains how people get access to the data. This does not imply they always get access (generally they do), but if they do, how. To be accessible, the principles specify:

- *A1. (meta)data are retrievable by their identifier using a standardized communications protocol*
  - *A1.1 the protocol is open, free, and universally implementable*
  - *A1.2 the protocol allows for an authentication and authorization procedure, where necessary*
- *A2. metadata are accessible, even when the data are no longer available.*

### 2.2.1 Open Data

The aforementioned section from our DoA briefly outlines the gist of the H2020 Open Data pilot: we give access to third parties to our data. In fact, the section is quite close to the core rights involved in Open Science, the rights to use/reuse, modify, and reshare knowledge. At the same time, it does set some reasonable limits: for example, where there are good reasons to not make data available (e.g. because of privacy aspect of human data), this is acceptable but must be explicitly reasoned for, and this must be documented. However, this guidance leaves plenty of room for us to implement our approach, and this leaves plenty of options and decision to be made.

### 2.2.2 Free Access

One controversial aspect of the current implementation of Open Access is to provide free access, free as in without cost to the user. Of course, hosting data is not without cost, but NanoCommons aims to achieve the sustainability and accessibility of data and tools beyond the project's lifetime.

To place the NanoCommons resources on a sustainable footing for the longer term, a business plan to secure future access and iterative development of the overall platform and set of associated infrastructure facilities will be developed. NanoCommons is foreseen to be an infrastructure for all. Nevertheless, its sustainability will depend on how to bring in the resources to run and maintain this infrastructure beyond the lifetime of the project. Therefore, the project will focus on how to gain value added for the different stakeholder groups, hence, defining the information flow directions as well as the streams of resources. Additionally, the implementation of the sustainable concept will be supported by testing of this business plan during the runtime of the project. Further shaping and



modifications according to the feedback gained during the sustainability-testing-phase is envisaged which will lead to a sustainable integration of NanoCommons infrastructure.

In the long run, NanoCommons envisages transferring its resources into either the [EU Open Science Cloud](#) (EOSC), which is under implementation following the adoption of the Implementation Roadmap for the European Science Cloud (Staff Working Document SWD(2018) 83) on 14 March 2018 or ECHA's [European Observatory for Nanomaterials](#) (EUON).

But generally, it is accepted that hosting raw can be covered in kind by existing solutions:

1. Institutional repositories (e.g. UoB [BEAR Hosting Services](#))
2. European repositories (e.g. [ZENODO](#), part of OpenAIRE)
3. Commercial repositories (e.g. [Figshare](#) or [Mendeley Data](#))

Guidance on selecting repositories can be found in [Science Europe's Practical Guide to the International Alignment of Research Data Management](#).

## 2.3 Making data interoperable

- *Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability*
- *Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?*

This principle prescribes that data must be sufficiently annotated such that people can understand what the data means. This means a clear data format must be used, values must have units, columns must have clear annotation what was measured, etc. This is also where ontologies (vocabularies) typically come in. To be interoperable, the principles specify:

- *I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.*
- *I2. (meta)data use vocabularies that follow FAIR principles*
- *I3. (meta)data include qualified references to other (meta)data*

The NanoCommons objectives are to make the data generated/collected and the tools developed during the project's lifecycle as interoperable as possible and to promote dataset combination and translational research. This means that the acquired data needs to be captured using a harmonised approach. To achieve that the project will use common data curation templates based on the ISA-TAB extended file format, which is an accepted ASTM standard (ASTM International E2909-13), are dynamic and flexible and can be modified accordingly to accommodate all User and project needs. The use of best practice examples (e.g. [diXa](#) and [ToxBank](#)) may be used for template creation, as well as the NIKC data curation template, which is especially versatile, treats experimental protocols and instruments as data points and can accommodate all fields of nanosafety research.

Another important aspect of data harmonisation is the use of a vocabulary (i.e. ontology) that will employ common agreed definitions for the terms used by all aspects of nanosafety research and will

---



allow both qualitative and quantitative data combination and reusability. As a result, the mid to long-term plan of NanoCommons is to create a single nano-wide ontology implementing already existing and established ontologies (e.g. [OBI](#), [ChEBI](#), [PATO](#), [UO](#), [NCIT](#), [EFO](#), [OAE](#), [eTOX](#), [eNanoMapper](#), [MPATH](#), etc; see Hastings *et al.*, 2015) [11], which will be linked and integrated to the NanoCommons ontology. This common vocabulary will be used for both the data and metadata curation irrespective of the file format the data is being stored.

Data and tools storage and accessibility is also going to be addressed through the use of multiple backups and mirrors that will be accessible through the NanoCommons APIs. These will also include an appropriate query and interoperability system that will allow the retrieval and combination of publicly available and disparate datasets, without the need for individual search and subsequent combination. The implementation of tools within the NanoCommons Knowledge Commons and databases will also allow the analysis of the retrieved combined datasets along with the appropriate quality control checks to identify whether further analysis is possible and meaningful.

In any case, NanoCommons will provide the necessary training resources and information in a simple and understandable way, targeted to non-informatics experts, and aims to produce online training courses that will allow further data and metadata exploitation. The interoperability work will be based on already available APIs like the [OpenTox](#), which was developed to cover the field of QSAR-based predictive toxicology (for chemicals / small molecules rather than nanomaterials) and has multiple functionalities (e.g. dataset generation, model building, prediction, validation), [Open PHACTS APIs](#) that are able to handle data collection and sharing and open database access APIs like [diXa](#), [ToxBank](#), [EGA](#), and [PubChem](#).

### 2.3.1 Supported data exchange formats

It is recommended to discuss with NanoCommons when deciding the proper data exchange format. It is essential to keep the data life cycle in mind here. For long term data storage, a format is preferred that is self-contained and self-explanatory. For short term data exchange, for example, to get data indexed in a search engine more application-oriented format may be more appropriate. These formats do not exclude each other, but are rather complementary. In all cases it is essential to explore if conversion or export to some format causes data loss.

Most database come with other supported formats and exploring the formats in which data will be shared is essential part of a DMP. For data hosted with eNanoMapper technology, various formats are supported. These include a custom Resource Description Framework format (for which a tutorial is being developed), and various spreadsheet templates (IOM templates [12] and JRC Templates [13]. The latter two take advantage of a spreadsheet annotation tool developed in eNanoMapper, called [nmdataparser](#). This approach has been used to make NANoREG data available to the NanoSafety Cluster community.

For long term data storage, the ISA formats may be of interest. Options are then the original ISA-Tab format, including the ISA-Tab-Nano extension, and the newer JavaScript Object Notation (JSON)-based ISA formats, for which also a nanomaterial extension has been developed [14].

Of course, for the many omics data types, domain specific formats are recommended, along with deposition of Open Data in domain repositories with those formats.

---

## 2.4 Increase data reuse (through clarifying licenses)

- *Specify how the data will be licensed to permit the widest reuse possible*
- *Specify when the data will be made available for reuse. If applicable, specify why and for what period a data embargo is needed*
- *Specify whether the data produced and/or used in the project is usable by third parties, in particular after the end of the project? If the reuse of some data is restricted, explain why*
- *Describe data quality assurance processes*
- *Specify the length of time for which the data will remain reusable*

This principle basically outlines that community practices for data sharing and data science should be followed. To be reusable, the principle specifies:

- *R1. meta(data) are richly described with a plurality of accurate and relevant attributes*
  - *R1.1. (meta)data are released with a clear and accessible data usage license*
  - *R1.2. (meta)data are associated with detailed provenance*
  - *R1.3. (meta)data meet domain-relevant community standards*

Data reusability is in the core of NanoCommons objectives. The data generated/collected/analysed and the tools developed through NanoCommons are going to be implemented with the appropriate metadata and where possible will be made publicly available through a tiered licensing system, which will allow data accessibility in collaboration always with the data owners and their exploitation needs. In many cases, a data embargo may need to be imposed. Examples of such cases may include pending publications, PhD candidates, ongoing EU and national projects requiring confidential data etc. Embargos can be lifted when all relevant data exploitation by the data owners has been completed. In a similar manner, any data, processes or tools source code originating from the project may be embargoed by the consortium until potential IP requirements and publications have been completed.

NanoCommons envisages to also link to its database high quality data from other publicly available sources spanning the full spectrum of nanosafety research. To achieve that the Project needs to harmonise the data as much as possible, develop an appropriate query system and make it available to third parties under the produced licensing system. Such an action means that the data and tools under offer needs to be systematically checked using strict QA processes.

QA processes for data generated/collected by NanoCommons will be automatically implemented into the experimental workflows developed prior to any offered service and will be made available to all partners using online lab-books. The checks will include the evaluation of the experimental protocols used or to be used, reproducibility, the tools source code and abilities and the presence of potential gaps in datasets.

A similar approach needs to be implemented for pre-existing datasets to be imported into the NanoCommons database. In these cases, and especially for the curation of already published peer-reviewed data, a manual curation process will most likely have to be implemented that will also mean that the curators will decide on the level of data quality. For other datasets, tools performing automated QA analysis and (pre)processing (e.g. [arrayanalysis.org/](http://arrayanalysis.org/) and [github.com/BiGCAT-UM](https://github.com/BiGCAT-UM)) and QA and analysis of sequencing data pipelines developed by the Consortium (e.g. Maastricht's University RNA seq and MeDIP seq) will be implemented and made available for use.

### 3. Allocation of resources

*Explain the allocation of resources, addressing the following issues:*

- *Estimate the costs for making your data FAIR. Describe how you intend to cover these costs*
- *Clearly identify responsibilities for data management in your project*
- *Describe costs and potential value of long-term preservation*

The Project already runs a survey that tries to monitor the awareness of the nanosafety community on FAIR data and aims to promote it through targeted workshops. NanoCommons aims to use data and tools sources, which are already FAIR and will demand no extra cost for the project. In the case where a non-FAIR source is required to cover the project's needs the relevant partners will assume the cost of implementation and Users may be supported through the NanoCommons periodic calls.

It is the intention to minimise the costs by using free-to-use data repositories and dissemination facilities to achieve GREEN open access, for example OpenAIRE, the UoB and NERC Open Research Archives etc. Some partners have foreseen and budgeted funds to get a few high impact papers into GOLD open access journals where the increased impact from this would warrant the expense.

The responsibility for collation and entry into the NanoCommons knowledge warehouse / database platform of the partner datasets lies with the individual partners. Their interlinking with the ontology the models and tools that are being developed for offering through TA will be facilitated (and resourced) via WPs 1-4 (JRA). All partners developing tools have additional budgets for integration of their tools into NanoCommons for provision of TA.

The responsibility and budget for developing and managing the web based structure (including aiding partners with data entry and uploading) of the NanoCommons infrastructure is held by DC, Biomax and NovaM, and D4.1 will describe the overall vision for the infrastructure (hardware, software, integration etc.) and the responsibilities of each partner towards the delivery of the NanoCommons research infrastructure.

#### **Transnational Access (TA) to Knowledge Management and computational Tools**

Within NanoCommons, funded access will be provided via short (1 month) or longer term (3-6 month) projects to support users with one of the following activities:

- Knowledge Base and Data Mining
- Processing and Analysis services
- Predictive nanoToxicology
- Risk assessment visualisation and reporting
- Nanoinformatics workflows for nanosafety experiments / best practice workflows.

All TA will be offered via **centralised calls** for User Access run on a 6-monthly basis. Calls for access will include micro-projects (up to 1-month Access), mini-projects (up to 6 months of access) and workflow Access to include site-visits to expert labs and support in establishment of nanoinformatics workflows to underpin excellence in experimental design. A pre-requisite for TA projects related to TA modality will be that data utilised will be made open access and FAIR via NanoCommons research infrastructure via the TA funding held by the partners to support the TA Users.

## D10.1 Initial draft of Data management plan (Open data pilot)

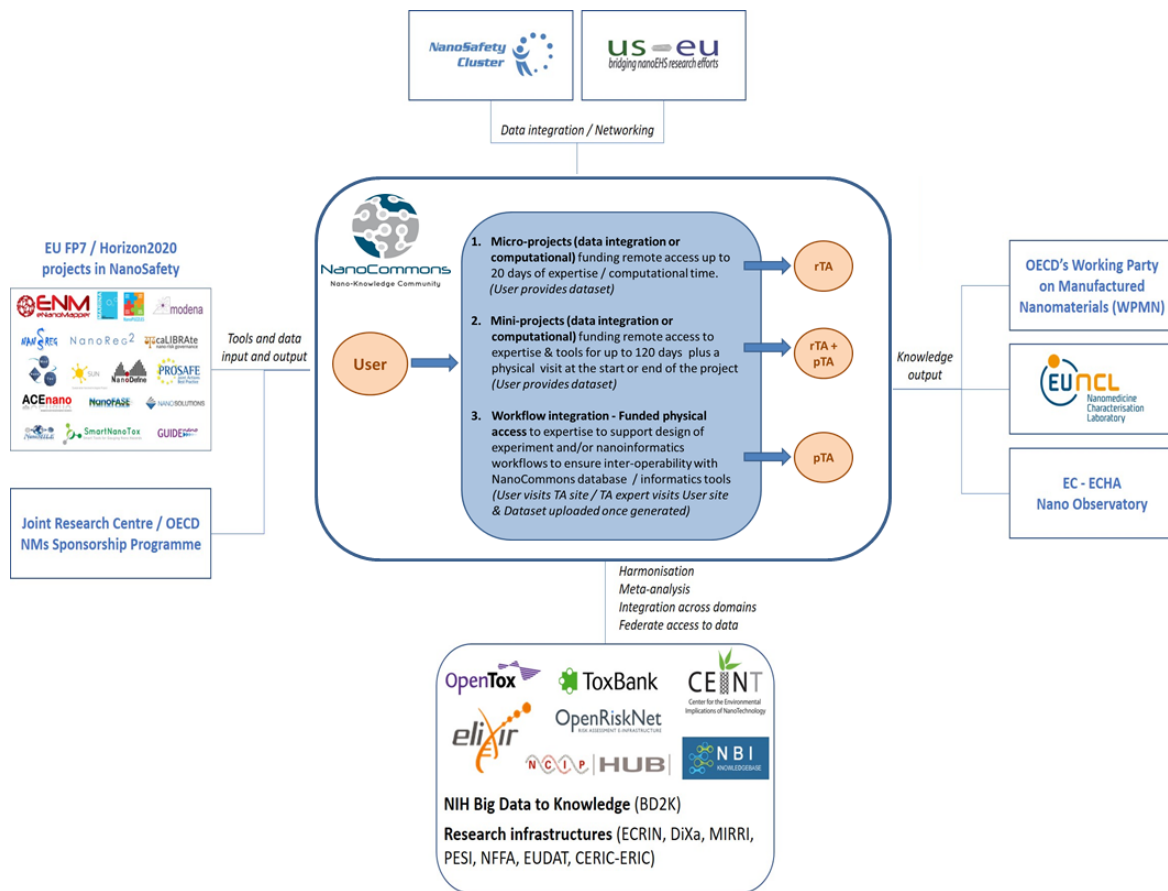


Figure 2. Flow of data in the NanoCommons project.

---

## 4. Data security

- *Address data recovery as well as secure storage and transfer of sensitive data*

The NanoCommons approach on sensitive data security, secure storage, recovery and transfer will follow existing privacy, copyright, *sui generis*, and GDPR laws and will include:

1. Responsible and fully secured management processes for personal data. These will include the anonymisation, encryption, logging of data usage as well as data deletion following usage. Implementation has already begun with the [NanoCommons website](#) having all necessary security protocols and tools (anti-hacking and malware plugins) added to prevent any malicious attacks.
2. The management team and the NanoCommons Consortium will ensure that all partners and users will follow strict ethical guidelines covering all aspects of the project's infrastructure. A *privacy-by-design* approach will be followed and controlled by an independent Data Protection Officer.
3. The presence in the consortium of established software and tools development partners (e.g. Biomax, DC) will ensure data protection with the use of their state-of-the-art firewall capabilities that can also protect the entirety of the project's virtual environment.

Data sharing and transfer among persons or partners will be, where appropriate, third party secure file transfer facilities, such as the NanoCommons knowledge warehouse, Figshare and OpenAIRE and via the internal communication platform. In the longer-term it is anticipated that NanoCommons datasets will be curated for and stored in certified repositories not dependent on the project funds for long term preservation (e.g. the OpenAIRE or UoB Research Data Store), if we are not able to secure our status as an Advanced Research Infrastructure or ERIC.

## 5. Ethical aspects

- *To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former*

The ethical aspects of the project will cover all aspects of the NanoCommons data lifecycle, will be included in detail within Work Package's 21 deliverables and will be strictly followed by all NanoCommons partners and infrastructure users.

No ethical issues arise from the project overall, or the data management aspects, since the focus of the project is on making existing datasets accessible rather than generation of new data. However, a strong effort will be made to ensure that ethical information regarding the datasets integrated into NanoCommons will have as a minimum a statement regarding the ethical approvals in place at the time the data were generated. Where data are generated, for example, through application of the modelling and other tools developed and made available to the community via Transnational Access,

these will be generated within the ethical framework of NanoCommons, and as above will have a clear declaration regarding the ethical approvals associated with any underlying datasets.

Thus, we do not anticipate any ethical or legal issues relating to the datasets generated as part of ACEnano project that would impact on data sharing. No personal data will be collected and so informed consent for data sharing and long-term preservation of such datasets is not required.

## 6. OTHER

- *Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)*

As part of University of Birmingham's commitment to ensuring FAIR and Open data, all research active staff (Postdoctoral fellows, PhD students) are expected to prepare DMPs for their own data, as per the [University's Research Data Management Policy](#). The UoB data management policy defines research data as “*the evidence that underpins the answer to the research question, and can be used to validate findings regardless of its form.*” Thus, data covers quantitative and qualitative statements, raw data from measurements and derived data – either cleaned or extracted from a researcher's primary dataset or derived from an existing source.

A detailed set of guidance on preparation of DMPs available via the UoB DMP site: [intranet.birmingham.ac.uk/as/libraryservices/library/research/rdm/Data-management-plans.aspx](http://intranet.birmingham.ac.uk/as/libraryservices/library/research/rdm/Data-management-plans.aspx)

NanoCommons will utilise the following aspects of the UoB data storage services:

- **Research Data Store (RDS):** The RDS is a central storage service for 'active' research data. It is highly resilient and is hosted in two data centres on campus. Space on the RDS is allocated to projects and managed accordingly. Up to 3TB of storage will be allocated by default to the Project though additional capacity may be purchased.
- **BEAR DataShare:** BEAR DataShare is a file synchronisation and sharing service provided by IT Services. The service allows users to securely save and sync files with colleagues and partners anywhere in the world, from any device. It provides 25 GB storage capacity per user.

---

## Final remarks

### A living DMP

The core of the RDM will be fixed: where our DoA or law outlines specific elements, these elements will be core part of the RDM and be pretty static. That said, over four years, the laws will change, which will affect our plan. Second, to achieve maximal impact of our project, we have to be able to aptly respond to changes in the field.

### Important Laws

European law changes all the time, and while not too fast, NanoCommons spans four years and the data outlives the project for another 10-20 years. For that reason, it is important to realize what laws apply to which kinds of data. Important laws to keep in mind include:

- Copyright laws
- Privacy laws
- GDPR law

Both affect our work. Organizations like the ELIXIR, [Electronic Frontier Foundation](#), and [Creative Commons](#) are very knowledgeable in the field, and so do several members of the European Parliament, in particular [Julia Reda](#). IT journals also provide information, like the German C't with [Welche Änderungen die neue EU-Datenschutz-Regulierung in Deutschland bringen wird](#).

### RDM tools

There are many tools around that support and/or provide guidance, including the following services:

- [DMPOnline](#) (used in this report)
- [Data Stewardship Wizard](#) (ELIXIR-NL)
- [RDMO: Research Data Management Organiser](#)
- [ReDBox Research Data Management Plan \(RDMP\) tool](#)

### RDM courses and guidance

- [H2020 Guidelines on FAIR Data Management in Horizon 2020](#)
- “Data management made simple” [15] with *Twelve tips for writing a data-management plan*
- [ELIXIR Webinar: Requirements in data protection law and the upcoming General Data Protection Regulation \(GDPR\) implementation](#)
- [DataONE Webinar: Data management plans 2.0: Helping you manage your data](#)
- [SpringerNature’s Research Data Support](#)
- [ORCID and Data Privacy in Germany](#)
- “Support Your Data: A Research Data Management Guide for Researchers” [16]
- [Data Protection Officer \(DPO\) Certification course](#) by Maastricht University
- [12 Things to Know About the GDPR and Data Security](#)
- [Complete guide to GDPR compliance](#) by GDPR.eu, a H2020 funded project
- [OpenAIRE: a pillar for Open Science in the EU: a Horizon 2020 GRACIOUS Webinar](#)
- [Ten simple rules for machine-actionable data management plans \(preprint\)](#) by OpenAIRE
- [The \(GA4GH\) Data Use Ontology \(DUO\)](#)
- [Science Europe: Practical Guide to the International Alignment of Research Data Management](#)



---

- [“Data management” in ELIXIR TeSS](#)

A recent paper looked into the effect of RDM in research, which is worth reading too [17]. They report three “key takeaways: (1) Most PIs practice internal data management in order to prevent data loss, to facilitate sharing within the research team, and to seamlessly continue their research during personnel turnover; (2) PIs still have room to grow in understanding specialized concepts such as metadata and policies for use and reuse; (3) PIs may need guidance on practices that facilitate FAIR data, such as using metadata standards, assigning licenses to their data, and publishing in data repositories.” These provide clear guidance of mistakes not to make yourself.

### Our DMP

The previous sections already outlined the context of our DMP. The vision of NanoCommons is to create a commons of knowledge about nanosafety. The stated mission is to provide FAIR and open access to research data. One step towards this is provision of advice and support to Users to best practice in development and implementation of DMPs, through provision of the NanoCommons DMP as a model for other NSC and beyond projects on how to manage their data to be FAIR, and also to provide instructions on how best to link their data into the NanoCommons (or other relevant) infrastructure. This appendix thus helps Users and the wider nanosafety community to:

- Acquire knowledge about DMPs
- Develop their DMP
  - Describe the life cycle of your data
  - Develop where data is stored locally
  - Develop when, how, where data is disseminated
- Discuss their the DMP within NanoCommons to align processes and outputs.

### Planning the data life cycle

Before writing an implementation of a DMP for specific research, it is good to identify the context, particularly when the data life cycle starts and ends for that research project. For example, the data life cycle starts when the experiment is designed: this design determines the amount of data that will be generated, with what experiments, and put requirements on what your electronic lab notebook needs to be able to record (*e.g.* photos of gels or other scanned material?).

Equally important is to know early on how the data will need to be analysed or integrated with other data sources, and what happens with the data after you are done with it. The end of life of data is often not before long after the projects ends. Top class data will be important for the next 50 years.

### Planning the research data management

The DMP presented in this work helps you with planning your research data management. The effort needed for the planning should not be underestimated; it does not need to be very long, but if taken seriously, it allows you to plan everything more efficiently, saving you time in the long run.

One area where time can be saved, is making data FAIR. Particularly making interoperable benefits from planning. For example, ontology annotation of data is easier when the details are fresh in your mind, *e.g.* about the standard operating procedures. In fact, identifying with identifiers or ontology terms which cell lines, buffers, etc, to use, can best be done when the experiment is designed.



---

## References

1. Directorate-General for Research & Innovation, E.C., *Guidelines on FAIR Data Management in Horizon 2020 v3.0*.  
[http://ec.europa.eu/research/participants/data/ref/h2020/grants\\_manual/hi/oa\\_pilot/h2020-hi-oa-data-mgt\\_en.pdf](http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf), 26 July 2016.
2. OpenAire. *OpenAire*. 2018 [20 July 2018]; Available from: <https://www.openaire.eu/>.
3. OpenAire. *How to create a DMP Plan*. 20 July 2018]; Available from: <https://www.openaire.eu/opendatapilot-dmp>.
4. Wilkinson, M.D., et al., *A design framework and exemplar metrics for FAIRness*. *Scientific Data*, 2018. **5**: p. 180118.
5. Carbon, S., et al., *A Measure of Open Data: A Metric and Analysis of Reusable Data Practices in Biomedical Data Resources*. *bioRxiv*, 2018.
6. Corpas, M., et al., *A FAIR guide for data providers to maximise sharing of human genomic data*. *PLOS Computational Biology*, 2018. **14**(3): p. e1005873.
7. Thomas, D.G., et al., *ISA-TAB-Nano: A Specification for Sharing Nanomaterial Research Data in Spreadsheet-based Format*. *BMC Biotechnology*, 2013. **13**(1): p. 2.
8. Creative Commons. *Licensing types*. *Share your work: Licensing types 2018* [cited 2018 20 July 2018]; Available from: <https://creativecommons.org/share-your-work/licensing-types-examples/>.
9. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship*. *Scientific Data*, 2016. **3**: p. 160018.
10. Dürst, M. and M. Suignard, *Internationalized resource identifiers (IRIs)*. 2005, RFC 3987, January.
11. Hastings, J., et al., *eNanoMapper: harnessing ontologies to enable data integration for nanomaterial risk assessment*. *Journal of Biomedical Semantics*, 2015. **6**(1): p. 10.
12. Jeliaskova, N., et al., *Deliverable Report D3.4: ISA-Tab templates for selected set of common bioassays*, in *eNanoMapper*, B. Hardy, Editor. November 2016.
13. Totaro, S., H. Crutzen, and J. Riego Sintes, *Data logging templates for the environmental, health and safety assessment of nanomaterials*. EU Science Hub, 2017.
14. eNanoMapper. *Investigation-Study-Assay (ISA): New material schema for ISA-JSON*. *Nanomaterial characterisation and bioassays data entry templates* [20 July 2018]; Available from: <http://ambit.sourceforge.net/enanomapper/templates/isa.html>.
15. Schiermeier, Q., *Data management made simple*. *Nature*, 2018. **555**(7696): p. 403.
16. Borghi, J.A., et al., *Support Your Data: A Research Data Management Guide for Researchers*. *Research Ideas and Outcomes*, 2018. **4**.
17. Mannheimer, S., *Toward a Better Data Management Plan: The Impact of DMPs on Grant Funded Research Practices*. *Journal of eScience Librarianship*, 2018. **7**(3): p. 5.

## Appendices

### Appendix A: RDM Copyright, License, and Waiver Clearance Form

Who are the copyright owners (names + email addresses)?	
Under what conditions will the data be available to the consortium?	
Under what conditions will the data be available to the rest of the world?	

The NanoCommons RDM Copyright, License, and Waiver (CLW) Clearance Form will be used in cases where data needs to be transferred, stored and analysed through the NanoCommons infrastructure via the project's TA's. In such cases NanoCommons will establish a minimum set of requirements for the data to be applicable to be included to the infrastructure, which are based on the FAIR data principles.

If a CC license is or was used, then it suffices to list the name of the license. In that case, the answer to the second and third question may, in fact, be the same. In any other case, data providers will have to agree to the NanoCommons terms and conditions regarding the optimum use of data and licensing. Such actions may include a local instance of part or the entire dataset to be used/analysed to be stored in NanoCommons servers, so that it can facilitate experimental workflows and is readily available for use by the NanoCommons experts. At the same time, it would be desirable for any raw data and metadata produced/generated to be made available in accordance with FAIR principles and accessible through the NanoCommons tiered licensing system. If that is not possible for raw data (e.g. pending publications), then a descriptive sub-data set and the metadata should be made available under the FAIR principles. Finally, the NanoCommons consortium should attempt to make all fully user exploited raw data and metadata publicly accessible to promote data interoperability and translational research. In such cases, the NanoCommons consortium will make sure that any use of data and metadata will be properly used and cited by external users, through the appropriate actions (e.g. DOI assignment) and terms and conditions.

## Appendix B: NanoCommons Customized RDM Online Plan

The NanoCommons customised RDMO Plan will cover the entirety of the data lifecycle (Figure A1), which contains 6 steps, i.e. experimental planning, data acquisition, data manipulation, data analysis, data storage and online access using Open data and FAIR principles whenever possible. Each step of the data lifecycle aims to provide the NanoCommons partners, TA Users and the wider nanosafety and nanoinformatics communities with all the necessary tools for the production, analysis and handling of high-quality data.

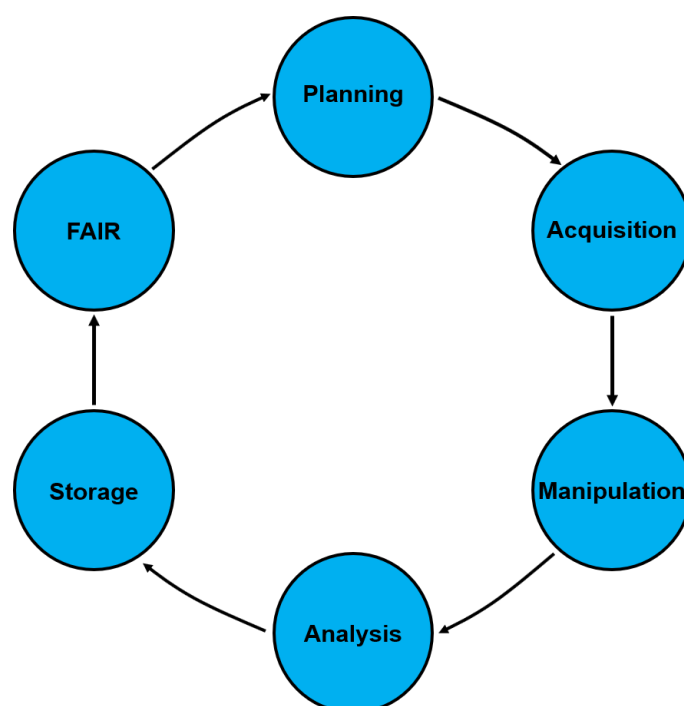


Figure A1. The complete data lifecycle.

### Step 1: Experimental planning

The first step of the data lifecycle is the experimental planning, during which the identification of research endpoints takes place, the number of samples and time points, and the necessary workflow and experimental assays and/or modelling tools are identified. During this step the whole experimental workflow is designed, the detailed experimental and/or analytical protocols identified, what is recorded and in what detail, and the appropriate data curation templates created. This process will take place irrespective of the nature of the needed work (experimental and/or theoretical), as it will customise the workflow to the specific needs of individual Users.

The specified workflow will be implemented into the SciNote online lab-book (Figure A2, [scinote.net](https://scinote.net)), which will include all the separate steps of the experimental and/or theoretical workflow. SciNote was the preferred online lab-book due to a combination of user friendliness, dynamic and versatile nature and significant capabilities, although other options are also commercially available. SciNote offers the capability of implementation of a desired data master curation template for experimental data to be

stored and extracted, while providing also the opportunity for distinct smaller scale templates, which can be protocol specific. It also offers the capability of analytical protocols implementation, metadata creation and reporting. It also allows the design and assignment of experimental workflows at various levels ranging from an individual researcher to a consortium wide scale, while being readily available both through local (LAN) and wide (WAN) networks based on specific needs.

Each workflow step will be linked to the necessary experimental and/or computational detailed protocols (Figure A3), the results page (Figure A4) and the master data curation template (Figure A5), which will be possible to automatically extract and upload to the desired data repository. The complete experimental or computational stepwise workflow will be accessible by the TA Users and the NanoCommons partners that will be responsible for the successful completion of the specific task, and will allow the streamlined online transfer of data between partners facilitating the prompt completion of the entire workflow.

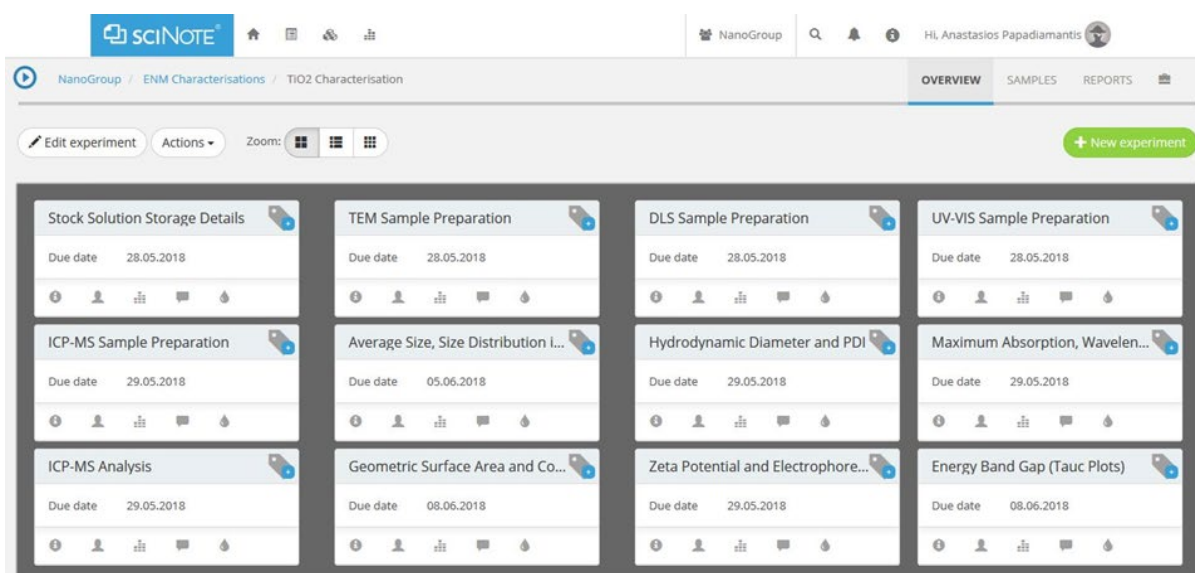
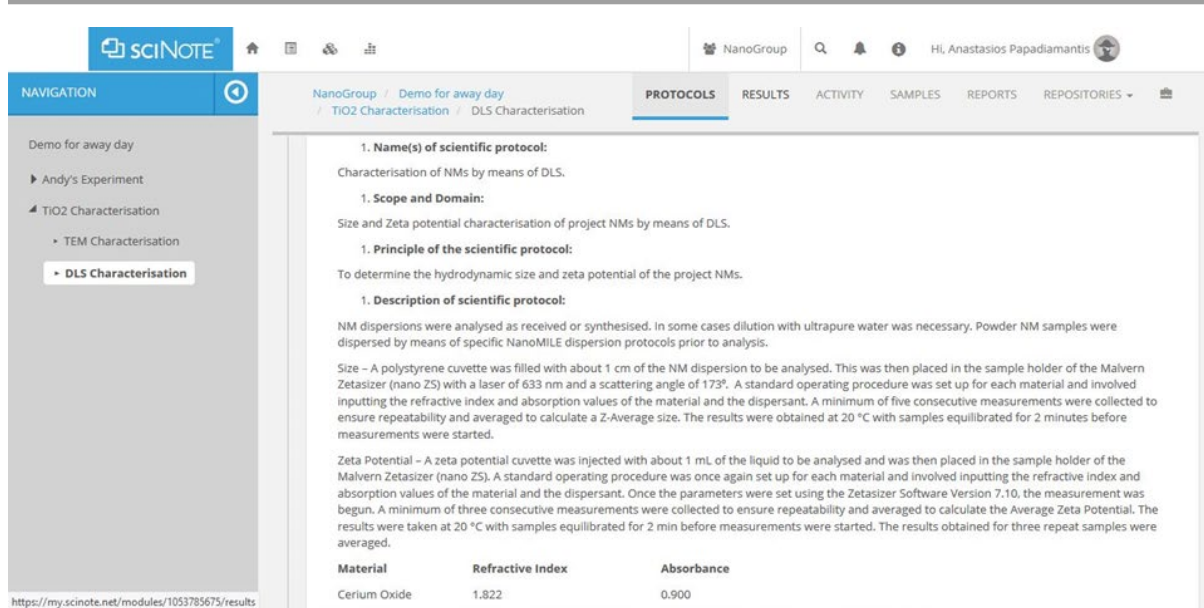


Figure A2. Experimental workflow for the physicochemical characterisation of an engineered nanomaterial (ENM) using the SciNote online lab-book.

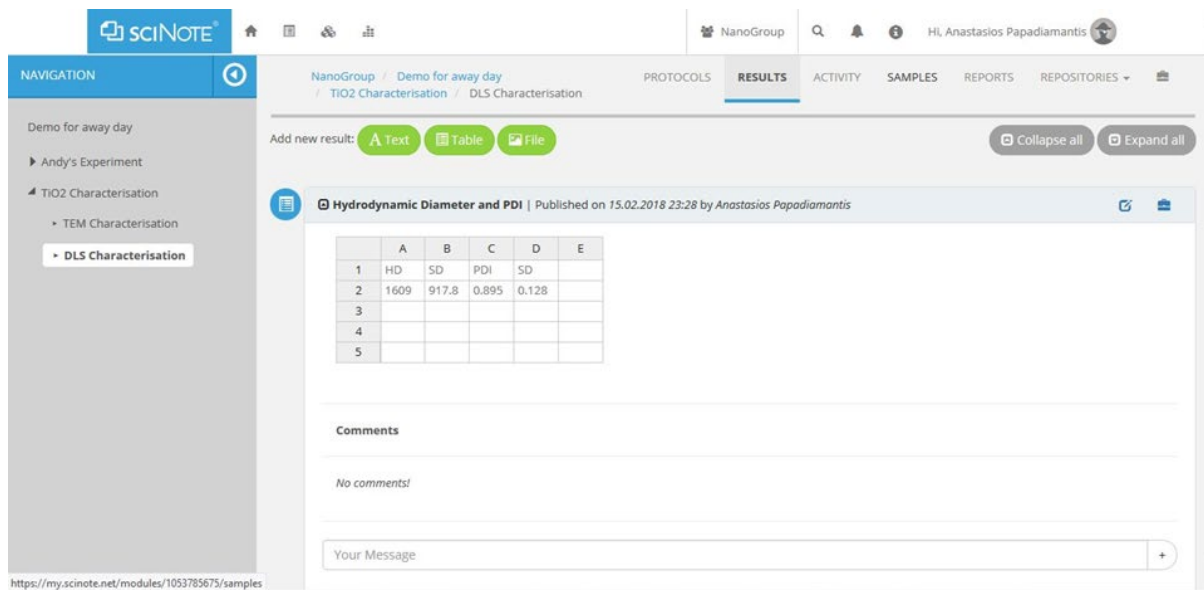
## D10.1 Initial draft of Data management plan (Open data pilot)



The screenshot shows the SciNote interface with a navigation sidebar on the left and a main content area. The main content area displays a detailed protocol for DLS Characterisation. The protocol includes sections for Name(s) of scientific protocol, Scope and Domain, Principle of the scientific protocol, and Description of scientific protocol. The description is divided into two parts: Size and Zeta Potential. The Size section describes the use of a polystyrene cuvette and a Malvern Zetasizer (nano ZS) with a laser of 633 nm and a scattering angle of 173°. The Zeta Potential section describes the use of a zeta potential cuvette and a Malvern Zetasizer (nano ZS) with a standard operating procedure. Below the text, there is a table with three columns: Material, Refractive Index, and Absorbance. The table contains one row of data for Cerium Oxide.

Material	Refractive Index	Absorbance
Cerium Oxide	1.822	0.900

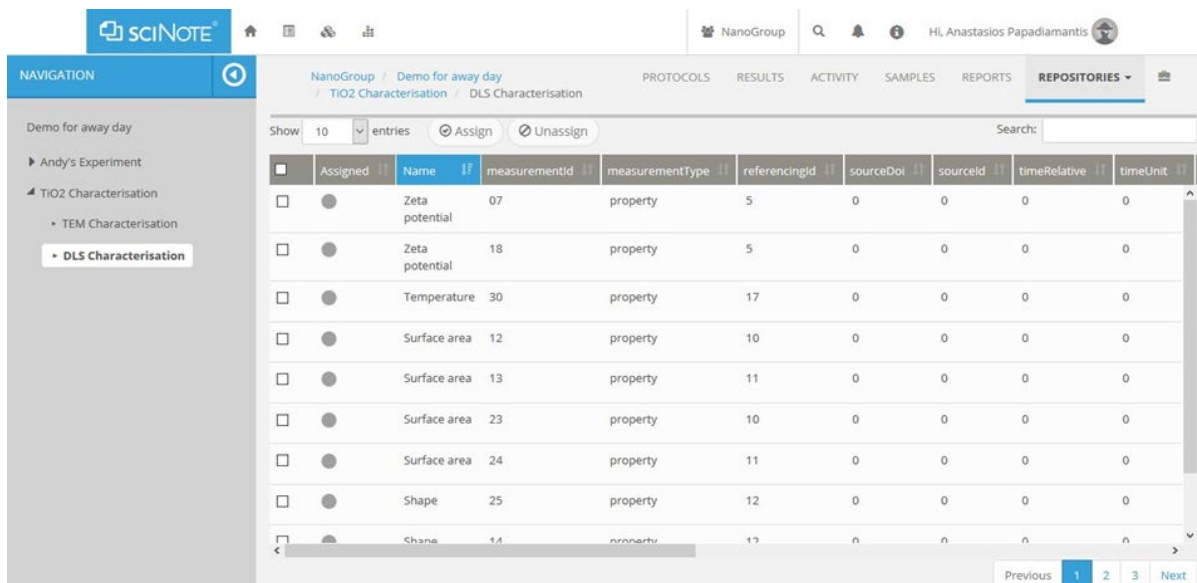
Figure A3. All experimental/computational workflow steps will be directly linked to detailed protocols.



The screenshot shows the SciNote interface with a navigation sidebar on the left and a main content area. The main content area displays a results page for Hydrodynamic Diameter and PDI. The results are presented in a table with five columns: A, B, C, D, and E. The table contains five rows of data. Below the table, there is a section for Comments, which is currently empty. At the bottom of the page, there is a text input field for a message.

	A	B	C	D	E
1	HD	SD	PDI	SD	
2	1609	917.8	0.895	0.128	
3					
4					
5					

Figure A4. The results page will be where individual datasets acquired or calculated will be stored.



The screenshot shows the SciNote interface with a navigation sidebar on the left and a main data table. The sidebar includes 'Demo for away day', 'Andy's Experiment', 'TiO2 Characterisation', 'TEM Characterisation', and 'DLS Characterisation'. The main table is titled 'NanoGroup / Demo for away day / TiO2 Characterisation / DLS Characterisation'. It has a search bar and 'Assign'/'Unassign' buttons. The table columns are: Assigned, Name, measurementId, measurementType, referencingId, sourceDOI, sourceId, timeRelative, and timeUnit. The data rows are as follows:

Assigned	Name	measurementId	measurementType	referencingId	sourceDOI	sourceId	timeRelative	timeUnit
<input type="checkbox"/>	Zeta potential	07	property	5	0	0	0	0
<input type="checkbox"/>	Zeta potential	18	property	5	0	0	0	0
<input type="checkbox"/>	Temperature	30	property	17	0	0	0	0
<input type="checkbox"/>	Surface area	12	property	10	0	0	0	0
<input type="checkbox"/>	Surface area	13	property	11	0	0	0	0
<input type="checkbox"/>	Surface area	23	property	10	0	0	0	0
<input type="checkbox"/>	Surface area	24	property	11	0	0	0	0
<input type="checkbox"/>	Shape	25	property	12	0	0	0	0
<input type="checkbox"/>	Shape	14	property	12	0	0	0	0

Figure A5. The complete set of acquired/computed data will be automatically transferred to a curation template, which will be possible to automatically extract and send for uploading to the NanoCommons and any other linked data repository.

### Step 2: Data acquisition and data curation

During the second step the actual data acquisition and simultaneous curation takes place along with data digitisation using either directly online lab-books described above or through the offline acquisition/analysis and subsequent transfer to the online lab-books and the specific data curation template.

### Step 3: Data manipulation and data curation

The third step of the DMP will focus on data manipulation, i.e. the process of data cleansing and control of the dataset's quality and completeness. During this stage the dataset completeness will be evaluated, potential gaps and/or outliers identified and any repetition or supplemental data acquisition needed will take place. When the dataset is completed it will be stored into a data repository, compatible with EUON, and any tools needed for the subsequent analysis will be implemented and linked to it. The tools developed and on offer from the NanoCommons Project are based on an ongoing survey ([www.surveymonkey.co.uk/r/PK2KXWW](http://www.surveymonkey.co.uk/r/PK2KXWW)) regarding the needs and desires of the NanoSafety Community, and beyond.

### Step 4: Data analysis

This step consists of all potential analytical or computational work that needs to take place to the produced dataset, in order to extract meaningful data interpretation and research outputs. The analysis will be based on the needed tools provided via the NanoCommons platform (e.g. QSARs, Omics, modelling, visualisation) and will be fully integrated with the NanoCommons ontology (with any missing terms or concepts added in draft form in real time, which will then be checked, amended

and integrated into the final ontology once it is confirmed by partner MU that there is no equivalent term already in place), and will be implemented based on individual needs, within the NanoCommons KnowledgeBase platform, leading to peer-reviewed open access publication(s).

#### Step 5: **Data Storage**

Following the completion of the analytical or computational procedure, the produced data will be backed up and stored and the necessary metadata will be created. The metadata will be used to inform potential future users regarding the ownership of the datasets and user rights, information on retrieval and analytical descriptions of the respective datasets. During this step any necessary documentation describing the above will also be created and DOI numbers will be assigned to each dataset. DOIs will be used to assist with publications (submitting raw datasets to peer-reviewed journals in a digital form), referencing of the respective datasets by future users, to ensure that the original data owners are acknowledged and to facilitate the transition from closed to FAIR and Open data.

#### Step 6: **FAIR Data**

The final step in the NanoCommons online data management processes will be the transformation of the data collected/generated/analysed through the project as FAIR and Open. NanoCommons is dedicated to FAIR data, but also acknowledges the need for the protection of the original data owners until all desired data exploitation and publications have been completed. In that sense, the NanoCommons Consortium will be in close contact with TA Users and project partners and will try to ensure that all data will eventually become FAIR and, when possible, Open. As noted in Appendix A, a prerequisite for securing of TA funding and technical support via NanoCommons is agreement that data imported and generated as a result of JRA, NA and TA activities will be made Open and FAIR. For that, copyright and IP owners must be clearly defined.