ACEnano
Analytical and Characterisation Excellence

Analytical and Characterisation Excellence in nanomaterial risk assessment: A tiered approach

Grant Agreement No 720952

Deliverable Report 4.5

Deliverable	D4.5 Initial Data Management Plan
Work Package	WP4: Linking ontology and methodology
Delivery date	M6 - 30 June 2017
Lead Beneficiary	University of Birmingham (UoB)
Nature of Deliverable	Report
Dissemination Level	Confidential (CO)

Submitted by	Iseult Lynch (UoB), Lucian Farcal and Thomas Exner (DC)
Revised by	Wolfgang Unger (BAM), Martin Tanner (TOFWERK)
Approved by	Eva Valsami-Jones (UoB)

Table of contents

Abbreviations

AFM	Atomic Force Microscopy
API	Application Programming Interface
BAM	Bundesanstalt fuer Materialforschung und Pruefung / Federal Institute for Materials Research and Testing
BET	Brunauer, Emmett & Teller method
CEN	The European Committee for Standardization
DC	Douglas Connect
DMP	Data management plan
EUON	European Union Observatory for Nanomaterials
FFF	Field-flow Fractionation
FTIR	Fourier Transform infra-red (spectroscopy)
ICP-MS	Inductively-coupled plasma mass spectrometry
ISO	The International Organization for Standardization
JRC	Joint Research Centre (of the European Commission)
NIST	National Institute of Standards and Technology (USA)
NM	Nanomaterial(s)
NMR	Nuclear Magnetic Resonance
NTA	Nanoparticle Tracking Analysis
OECD	Organisation for Economic Co-operation and Development
ORD	Open Research Data (Pilot)
REST	Representational State Transfer (web services)
ROS	Reactive Oxygen Species
SOP	Standard Operating Procedure
TEM	Transmission Electron Microscopy
UoB	University of Birmingham
WP	Work Package (of the ACEnano project)

Summary

This report describes the initial version of the data management plan (DMP) for the ACEnano collaborative research project. The current DMP covers the general aspects of the ACEnano data management approaches, based on the FAIR guidelines. More specific and clearly-defined measures will be added in parallel to the integration of methods, their data output specifications, their Standard Operating Procedures (SOPs) and specific data sources into the knowledge warehouse, which will follow the time plan enforced by the case study requirements on data availability.

1. Introduction

The European Commission 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 re-use of research data generated by Horizon 2020 projects and takes into account the need to balance openness and protection of scientific information, commercialisation and Intellectual Property Rights (IPR), privacy concerns, security as well as data management and preservation questions [1]. Open data is data that is free to access, re-use, repurpose, and redistribute. The Open Research Data 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 Projects starting from January 2017 are by default part of the Open Data Pilot [2].

To help optimise the potential for future sharing and re-use of data, the ACEnano Data Management Plan (DMP) helps the partners to consider any problems or challenges that may be encountered and helps them to identify ways to overcome these. This DMP is a “living” document outlining how the research data collected or generated within ACEnano will be handled during and after the research project. It follows the Guidelines on FAIR Data Management in Horizon 2020 [1] and is based around the resources available to the project partners in a realistic way, 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 (https://dmponline.dcc.ac.uk/plans/17617) (see Figure 1) [3]. Here, the concepts for the description of data sets, as well as data sharing and archiving approaches adopted in the DMP, are described, followed by the DMP itself in relevant version of the time of this writing. Since the ACEnano methods themselves, their hyphenation and communication interfaces, and integrated data processing templates are under development, which will define the integrated data sources, the current DMP covers the general aspects of ACEnano data management. More specific and clearly-defined measures will be added in parallel to the data integration aspects.

Data management is at the heart of ACEnano, which aims to create a “conceptual toolbox” including a tiered approach to cost efficient NMs analysis that will facilitate decision-making in choice of techniques and Standard Operating Procedures (SOPs), linked to a characterisation ontology framework for NMs grouping and risk assessment. Thus, ACEnano will initiate activities to support data collection, management, interpretation and delivery to a data warehouse for safe use & storage, with the initial processes outlined in this deliverable report.

Figure 1. DMP tool interface used to create the ACEnano data management plan [3].

DATA SET DESCRIPTION

This section refers to what kind of data will the project generate and collect, and to whom might they be useful later. The data set refers to:

- The data and metadata needed to validate results in scientific publications, 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:

- What is the data about?
- Who created it and why?
- In what forms it is available?
- Standards applied.

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 Re-usable).

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 is also important to assure the availability of data for a long period beyond the lifetime of a project. However, this does not refer only to storage in a research data repository, but also to consider the usability of the data. ACEnano is working closely with the OpenRiskNet e-infrastructure (coordinated by Douglas Connect) and the newly starting e-infrastructure for nanoinformatics, NanoCommons, also coordinated by UoB, to develop sustainable databasing solutions for the data from EU projects. ACEnano data will be used as one of the test cases for the e-infrastructure projects, thus ensuring its long-term preservation and accessibility. Project data will also be available via the UoB BEAR data archive.

2. Data Management Plan

2.1. 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 re-used (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.*

2.1.1 Purpose of the data collection and its relation to the objectives of the project

Central to the risk assessment of nanomaterials (NMs) is the need for reliable and reproducible characterisation of NMs, which has presented a significant challenge to date given their extreme diversity and dynamic nature, particularly in complex environments, such as within different biological, environmental and technological compartments. ACEnano is developing two key data-management actions to resolve this:

1) the development of a holistic analytical framework for reproducible NM characterisation, spanning from initial needs assessment through method selection to data interpretation and storage; and

2) the embedding of this framework in an operational, linked-up ontology (“common language”) and data framework to allow identification of causal relationships between NMs properties, be they intrinsic, extrinsic or calculated, and biological, (eco)toxicological and health impacts.

The overarching aim of ACEnano is to introduce confidence, adaptability and clarity into NM risk assessment by developing a widely implementable and robust **tiered approach to NM physico-chemical characterisation** that will simplify and facilitate contextual (hazard or exposure) description and its transcription into a reliable NMs grouping framework. This will be achieved by the creation of a conceptual “toolbox” that will facilitate decision-making in the choice of measurement techniques and SOPs, linked to a characterisation ontology framework for grouping and risk assessment and a supporting data management system.

Both the successful completion of the project’s objectives and the future applicability of the methods toolbox and decision tree will rely on the provision of high quality data that will form the evidence basis for our approach and will facilitate subsequent method validation and standardisation (e.g. via OECD, CEN, ISO etc.) based on the pre-validation data compiled herein.

2.1.2 Types and formats of data generated/collected within the ACEnano project

Method development within ACEnano relates to methods at three phases of development, namely:

(1) newly emerging methods or end-points for which no method currently exists and which require **Analytical innovation** (WP1). WP1 represents the apex of the methods pyramid, and focuses on addressing key gaps in characterisation ability and a focus on method automation, integration and miniaturisation, in order to drive down the costs of characterisation via reduced assay duration time, instrument cost, and complexity such that the methods become benchtop and easy to operate, rather than requiring extensive technical support and staff to operate). In terms of technology readiness, methods here start at TRL2 and at the end of the project reach TRL5 or TRL6.

(2) established methods or approaches which have been demonstrated to be applicable to some NMs, but where some **Analytical optimisation** (WP2) is required in order to improve the applicability domains. For example, many of the methods work well for spherical particles in aquatic media, but struggle greatly for irregularly shaped particles in soil and sediment matrices where significant sample clean-up is required, or work *in vitro* but need adaption to be more representative of the *in vivo* situation. These improvements will enable better fit within the ACEnano toolbox, by increasing their applicability and usefulness in “real” samples. In terms of technology readiness, methods here start at TRL4-5 and at the end of the project reach TRL6 and beyond.

and (3) very well established methods, which are ready for **Method benchmarking** (WP3) and harmonisation of the SOPs for use by non-experts, in order to move these towards formal standardisation process (e.g. OCED, CEN, ISO etc.) and to generate and raise general confidence and expertise in the wider applicability of these key NM characterisation methods through their validation, refinement of their applicability domains in relation to intended stakeholder applications, and via development of the ACEnano characterisation decision tree and workflows for intended stakeholders applications.

Data sets collected will thus relate to measures of the following key descriptors:

- complex **NM shapes**, e.g. by combining microscopy (TEM, AFM), and other approaches – the data will be images, plus descriptors derived from these such as circularity, agglomeration etc.
- ***surface area***, e.g. by BET (powders) and NMR (suspensions) – numerical data primarily.
- ***particle size***, measured by a variety of methods, e.g. microscopy (powders), and NTA, FFF (suspensions) – numerical data primarily.

- multiple ***composition*** and ***surface chemistry*** – spectra and chemical maps (e.g. from FTIR, EDS, XPS, AES, SIMS), topography images (e.g. from microscopy) and numerical data (e.g. from ICP-MS).
- ***coating stability / degradation*** – a range of time-resolved datasets, assessing loss of coating under various conditions (e.g. acidic versus alkaline, high/low natural organic matter, high/low ionic strength, microbial degradation etc.).
- ***NM dissolution*** – a range of time-resolved datasets, assessing NM dissolution under different conditions (e.g. different biological and environmental media, different pHs (e.g. 4 to represent digestive tract and 7 to represent the biological milieu)), and combining solution approaches, various detection methods, and newly developed higher-throughput and lower cost assays.
- ***surface reactivity*** – a range of automated on-chip assays are being developed within ACEnano for the most relevant types of reactivity (e.g. ROS generation, photocatalytic activity). This may also be time-resolved datasets requiring understanding of starting, intermediate and final characterisation data from the methods noted above.
- ***hydrophobicity*** - a range of automated binding assays (for example) are being developed within ACEnano to assess hydrophobicity. This may also be time-resolved datasets requiring understanding of starting, intermediate and final characterisation data from the methods noted above.
- ***surface strain*** – bond angles, and data related to binding affinities and will be utilised to derive surface strain parameters. A number of computational parameters may also be utilised here. As per the EU-US Nanoinformatics roadmap 2030 [4], the use of calculated descriptors will be increasingly important in risk assessment of NMs in the future, to reduce the amount of testing required and facilitate read-across from data-rich NMs to data poorer ones.

These analytical datasets will form the basis for (a) NMs grouping for hazard and exposure assessment purposes, and could also enable studying the longer-term fate of NMs following their interactions with living organisms and the environment. In all cases, the datasets will include information on limits of detection, domains of applicability (e.g. range of NMs tested, size / shape / density or other constraints on the NMs to which the methods can be applied), the repeatability[1] and reproducibility[2] of the test method, the uncertainty associated with the measurement, and the inter-operability of the method by other users.

Thus, the range of data generated and stored / shared will include:

1. Empirical data generated from laboratory experiments. These will include a large amount of high definition images (e.g. transmission electron microscopy), and spectroscopic data as well as simpler univariate data (e.g. total elemental concentrations) during dissolution experiments, for example.
2. Repeatability and reproducibility data for the different methods, including the accompanying

[1] Closeness of the agreement between the results of successive measurements of the same measurand carried out under the same conditions of measurement (i.e. with the same observer, the same measuring instrument, used under the same conditions, at the same location when repeated over a short period of time).
[2] Closeness of the agreement between the results of measurements of the same measurand carried out under changed conditions of measurement. The changed conditions may include the principle of measurement, the method of measurement, the observer, the measuring instrument, the reference standard, the location, the conditions of use or the time.

SOPs, data capture templates, and analysis of uncertainty and domain of applicability.
3. Ontological terms, decision trees and other supporting data and metadata to enable the full implementation of the ACEnano analytical tool box.
4. Inputs and parameters from model runs through the decision tree and ACEnano toolbox, and the resulting outputs.
5. Empirical relationships and derived parameters underpinning the decision tree and results generation templates.
6. Summary data underpinning reports/deliverables and publications produced as part of project.

As file formats, we are looking mainly at two options advertised as standards for the nanosafety field, since they follow different philosophies with their strengths and weaknesses:

1. **ISA (-tab or -json) including the nano extension (ISA-Tab-Nano)**

- The advantage is their flexibility. It is very easy to create files for new experiments and calculations and add new columns if needed for updated methods;
- The disadvantage is that there is no description of the file formats (data + metadata). Even if they use ontologies for the data entries, the naming of the columns is not regulated or protocolled. Also, the data file format is not specified (only the file names are given). This disadvantage can be circumvented by the interoperability layer, which will be used to describe the data schema using controlled vocabulary.
- The “nano”add-on for ISA-tab-Nano is less well developed [5], but there is quite some consensus around it in the nano field, and it is strongly promoted by both the US National Cancer Institute (NCI) and the eNanoMapper data portal.

2. **NanoREG templates**

- The advantage is the well-defined layout of the files that were developed by the JRC within NanoREG to be compliant with NMs regulation [6], which definitely simplifies the interoperability (this is however also a disadvantage);
- they cover many (but by no means all) of the characterisation end-points of relevance to ACEnano, and thus offer a good starting point to build upon.
- Note however that there are mistakes in several, that need to be corrected, but full documentation of any changes made will be included in the ACEnano versions of the templates.

In both cases, only the strict usage of ontologies for this data description can guarantee that the information is easily understood by the user or automatically transferred between services.

2.1.3 Reuse of data

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 re-usable and extensible tools, and to make use of existing characterisation data, including from real industrial NMs in products, where possible.

Where possible, NMs utilised in other projects (e.g. NanoMILE, NanoFASE) are also being utilised in ACEnano, to enhance the characterisation datasets available. However, given the method development and method optimisation goals of ACEnano, utilisation of reference materials is also required (as per Table 3 in the ACEnano Description of Action which summarises the range of NMs for testing, including commercially available NMs, legacy materials and novel, not yet commercial 2nd generation NMs). SOPs developed within NanoMILE and NanoFASE for characterisation of NMs in biological and environmental matrices are being re-used where possible, and any changes documented as part of the metadata and domains of applicability sections of the SOPs.

Table 1 below summarises the datasets that are potentially being re-used within the context of ACEnano.

Table 1: Existing data generated through other Nano Safety Cluster (NSC) projects and other appropriate projects will be used by the ACEnano project through collaborative efforts or agreements with the relevant project beneficiaries. This data will in most cases be for use to benchmark the ACEnano SOPs and newly developed innovative methods, and as test cases to parameterise the ACEnano decision tree and analytical toolbox approach. 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 ACEnano**
NanoMILE	Characterisation data and calculated descriptors for up to >100 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 methods, as well as for benchmarking of the newly developed assays.
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 NMs in different environmental compartments. This data will be utilised to parameterise the decision trees for appropriate characterisation for NMs in different application scenarios.
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 ACEnano decision tree to select appropriate characterisation methods, extraction and/or dispersion protocols, etc. based on where the specific NM releases from a specific application are likely to occur.
NanoDefine	Project focussed on development of	Developed a decision tree to help

	methods/ standards supporting implementation of the EC definition of NMs.	SMEs answer the question as to whether their NM remains "nano" according to the EC definition. This decision tree will be built into the ACEnano analytical toolbox.
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 method benchmarking and reproducibility assessment.
eNanoMapper	Ontology terms, database structure for nanosafety	ACEnano 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.
EC4SafeNano	Country-specific data on characterisation services available in member states (service providers) via survey of user needs and services	The characterisation service provider data will be linked to the ACEnano decision tree and analytical toolbox (i.e. where to go once the appropriate tests have been identified).

2.1.4 Origin of the data

On data sources, the original license of data usage will be considered, also for the version integrated into the ACEnano decision tree and analytical toolbox (via parameterisation of the decision tree and/or benchmarking the repeatability of the ACEnano methods). Therefore, an authentication and authorisation service will be integrated in ACEnano, with Douglas Connect handling any licenses[3].

Thus, we anticipate that data integrated into ACEnano toolbox, decision tree, ontology or other outputs will include:

- Data and models generated and owned by ACEnano Partners, which will be assigned an Creative Common's licence allowing re-use for all purposes as long as cited appropriately as far as possible, or re-use for non-commercial purposes under specific, well-justified cases as approved by the coordinator under the terms of the Consortium Agreement;
- Open Source data and models, used the license mentioned by the owners;
- 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.

[3] https://creativecommons.org/share-your-work/licensing-types-examples/

2.1.5 Expected size of the data

It is anticipated that the amount of specialist data that will be required and produced by partners, both in terms of primary data (characterisation and analytical data), databases to support the decision trees, and the ontological terms across the whole consortia will be in the region of **10s of Terabytes**. To facilitate access to the data, both during the project and beyond its lifetime, a process will be implemented to ensure that all partners store their data in the projects' centrally managed datastore, i.e. in the ACEnano knowledge warehouse as it is generated, to ensure that it is easily FINDable. Re-use means also in ways the original data provider never thought about and, thus, data might become important for reasons not foreseeable by the data provider. While the ACEnano partners are aware that this puts some additional work on each partner, we have experience with implementing such a strategy (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), such that we can streamline and simplify the process as much as possible for partners. I absolutely believe that it is better to store data centrally, which finally shows to be not useful to anybody else, than to search for data stored locally, when somebody needs it.

The data to be made directly publicly accessible is the description of the methods themselves, along with the SOPs, analysis of uncertainty, domain of applicability and benchmarking against other relevant methods. Guidance on the use of the decision tree, appropriate sample preparation and dispersion for different methods / products etc. will also be developed, linked to ECHA's European Union Observatory for Nanomaterials (EUON) and guidance for deriving selecting appropriate characterisation tools for various NMs in certain media or processes. The size of this data may be comparatively limited and thus easier to make interoperable and re-usable. In addition, full characterisation data for each NMs in all media in which they were tested will be made available to the community to support interpretation of hazard and effect data generated in other projects.

A key aspect of the ACEnano approach is alignment with, and utilisation of the resources being developed in the OpenRiskNet e-infrastructure, coordinated by ACEnano partner Douglas Connect and in which UoB is a partner. A key aspect of OpenRiskNet, which will also be adopted by ACEnano, 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 ACEnano project. 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, ACEnano will agree with the owners of the data as to whether a local copy can be generated or how best to integrate the data into the ACEnano project.

2.1.6 Utility of ACEnano data and models

The ACEnano decision tree and analytical toolbox are being developed as tools for industry (especially SMEs) and regulators, but they will clearly also be of use to the research community through the benchmarking of various methods, with different measurement principles, for the same end-points. The data are being collected in a manner that will also maximise the potential for subsequent standardisation of new approaches via OECD, CEN, ISO etc. thus further supporting the utilisation by industry and regulators. The decision trees will be designed to allow users to progress stepwise through different classical problems that key industrial and regulatory stakeholders may expect to address at different point during the development, registration and project stewardship of nanotechnology products. These may cover all aspects relevant to NMs from initial identification of material as "nano" according to EU (or other) definitions by tiering analysis of different method for size and size distribution analysis (in powders and different environmental and biological media) to understanding function NM changes (dissolution, coating swapping/degradation that may affect fate and hazard through EM, spectroscopy and elemental analysis.

The key outputs of relevance to industry and regulators will thus be:

- The increased and better understood application domain of both the "Established" and "Developing techniques" through improved sample handling, preparation / clean-up that allows analysis in previous inaccessible media types while maintaining a high fidelity for the key NM descriptors to be addressed (e.g. coating stability, charge, aggregation state etc.).
- A thorough Benchmarking of possible alternative technical options that is context specific to both the list of key descriptors relevant for the situation, the NM of interest and the media involved. The so-developed decision tree to guide users through choosing the most cost effective chain of options from the ACEnano Toolbox will deliver a "fit for purpose" characterisation to support any given NM risk assessment.

2.2 FAIR DATA

2.2.1 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.*

All ACEnano-generated data sets, including those related to method development/optimisation, will be placed into the ACEnano Knowledge Warehouse on a continuous basis utilising the agreed ACEnano templates (derived and updated from NanoREG initially, and developed for methods where no templates exist as yet). The data management will follow the best practice approaches developed in the ToxBank and eNanoMapper projects as well as the new developments planned in the OpenRiskNet e-infrastructure project.

Data management will rely on a rigorous authentication and authorisation (A&A) management system. The data provider will grant relevant authorisation on a case-by-case basis regarding datasets and data users following the 3 concepts:

- confidentiality (ensuring that information is not accessed by unauthorised persons),

- authentication (ensuring that the users of critical functions are the persons they claim to be), and
- integrity (ensuring that external occurrences cannot alter the information in such away to be not detected by authorised persons).

If no legal or contractual reason demands for a different approach, datasets will be shared with the consortium first and fully open access will be granted after a short data embargo period to allow for publication of the scientific results. ACEnano fully commits to the findable, accessible, interoperable and reusable (FAIR) data management paradigm.

Harmonisation and data sharing especially with other EU NanoSafety Cluster (NSC) projects are of highest priority. The eNanoMapper data warehouse is accepted as one of the repositories to be used by NSC projects and a long-term sustainability plan is being developed at the moment, including onward development and integration with the CEINT Nanoinformatics Knowledge Commons database through the NanoCommons research infrastructure for nanoinformatics, coordinated by UoB (starting January 2018).

ACEnano will guarantee the FAIR usage of the produced data by adopting file standards like the ISA-TAB nano specifications and updates of these as well as the usage of the recommended vocabulary and ontologies like the eNanoMapper ontology, the BioAssay ontology, Ontology of Adverse Effects, to name just a few, besides the new NMs characterisation and exposure ontology being developed in the ACEnano project. A common information model supporting ontology annotation and update functionality will be provided to keep the data consistent with ongoing developments e.g. in the exposure ontology as well as new versions of standard file formats.

Similar to the data management system implemented within eNanoMapper, the data warehouse should provide support for upload, export and search options for NMs and experimental data, e.g. through Representational State Transfer (REST) web services API and a user-friendly web browser interface. All these measures will facilitate data exchange and re-use between researchers, institutions, organisations and countries especially regarding re-combinations with different datasets from different origins.

ACEnano will also benefit from OpenRiskNet e-infrastructure, which is integrating existing data sources and making them more easily findable, accessible and interoperable. This will be based, on one hand, on the metadata provided by the data sources and, on the other hand, on the interoperability layer, which will harmonise these metadata into service descriptions and data schemata, which can be queried through the OpenRiskNet discovery service.

The description of the capabilities of a database and the data schema will allow for:

- Accessing specific search functionality, and
- Identify the data fields to be searched (e.g. where information on biological assays are stored);
- Finding the best file format for data exchange;
- Understanding all the data and tools, with transparent access to metadata describing the experimental setup or computational approaches.

It is the intention of the ACEnano project partners to make datasets generated through the work of the ACEnano project readily discoverable through linking them to the relevant outputs (SOPs, publications, deliverables, or benchmarking reports) of the project. Another key tool that will be used is the OpenAIRE facility in which metadata records for the datasets included in publications generated by ACEnano efforts will be published. Other datasets might also be published here if deemed suitable, or via FigShare for example. Through the use of OpenAIRE we will ensure that search keywords, version numbers, and as much of the further metadata as possible will be captured (see Table 2).

Table 2. List of metadata required by the OpenAIRE. Each field is either mandatory (M), mandatory when applicable (MA), recommended (R) or optional (O) Field Requirement Identifier M

Field	Requirement
Identifier	M
Creator	M
Title	M
Publisher	M
PublicationYear	M
Subject	R
Contributor	MA/O
Date	M
Language	R
ResourceType	R
AlternateIdentifier	O
RelatedIdentifier	MA
Size	O
Format	O
Version	O
Rights	MA
Description	MA
GeoLocation	O

2.2.2 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*

It is the intention of the ACEnano consortium that all data will be made accessible where technically possible, as re-use may be in ways that cannot currently be foreseen. Open access will be facilitated through the centralised data storage platform, and may be supplemented through linkages to specialised data collections such as e.g. Electron Microscope image libraries which may be systematically stored within "in house" systems).

The summary data from the characterisation of the ACEnano library of NMs, and the benchmarking data on the domains of applicability and repeatability and reproducibility of each method will be made openly available as part of ACEnano decision tree and analytical toolbox. These datasets will be available in readily accessible formats, for example comma separated value (CSV) files and where reasonable linked to DOIs in open repositories (e.g. the OpenAIRE or the UoB Open Research Archive). An embargo period may be necessary for some data sets until supporting scientific papers are published.

Where data links to published papers, both the paper and the data will be included in appropriate repositories, e.g. the OpenAIRE or NERC Open Research Archive, or into the OpenRiskNet or NanoCommons Knowledge bases.

It is expected that all data access will be credited through citations of DOIs where direct open access is provided, and where DOIs are not practicable access will be provided through contact to the generating research group and access tracked through personally having to hand the data over, and most likely in these cases as part of a collaboration agreement.

A list of datasets that cannot be shared or only be shared under restrictions will be added to this document as they are identified. Currently no such datasets are identified.

Open Standards to be applied with ACEnano include:

- Data and models will be stored and served using well-developed and widely applied standards and technologies that promote data reuse and integration, such as JSON-LD, RDF and related semantic web technologies;
- ACEnano resources will be aligned with activities of OpenRiskNet and the OpenTox community and collaborate with OpenRiskNet in developing open standards for nanomaterials characterisation in order to support predictive toxicology;
- ACEnano study data and metadata description using standards such as ISA-Tab-Nano, already in use in a number of developing nanomaterials repositories and the ISA format is widely utilise also in omics and toxicogenomic resources (e.g. diXa), which will further simplify integration of the NMs characterisation datasets with existing or emerging hazard datasets for these NMs;
- The ACEnano analytical toolbox and decision tree will be encoded using suitable open standards (e.g. QMRF, BEL, SBML) and annotated according to the appropriate minimal information standards (MIRIAM) for dissemination through appropriate repositories (e.g. BioModels).

Thus, ACEnano does not propose to create new file standards but rather to employ the existing approaches as to define a core set of information, on which the scientific community agrees that they are important to document, but which can also be modified and extended if necessary for a specific application. For defining this core set, regulatory files formats like **OECD harmonised templates (OECD HT)** and **Standard for Exchange of Nonclinical Data (SEND)** will be considered. Even if these file formats are too limited and do not have the flexibility to be used outside regulatory purposes, the ACEnano Net partners developing the guidelines will include all information needed for these reports in the templates if possible.

Given the highly dynamic nature of NMs, and their ageing and evolution during transport, storage, handling and following dispersion [7], it has been suggested that information on provenance of the NMs may be a route to improved reproducibility of results [8, 9]. Additionally, the need for time-resolved characterisation data is important, and will be implemented into the ACEnano decision tree also, in terms of underlying outliers – is the outlier a result of sample ageing over time, or a result of slightly different dispersion quality, for example. To help develop this into the ACEnano platform, aspects of the **PROV** metadata standard [10] will also be considered, which provides Provenance information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness. The PROV Family of Documents defines a model, corresponding serialisations and other supporting definitions to enable the inter-operable interchange of provenance information in heterogeneous environments such as the Web [10].

2.2.3 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?*

ACEnano will utilise OpenRiskNet's interoperability layer, which will provide data schemata that describe the format of the data using a controlled vocabulary:

- Metadata standards and data documentation approach will consider the existing standards that can be consolidated and the equivalent data that can be retrieved independent of the file format;
- Enhancement of existing NM ontologies (Nanoparticle Ontology, eNanoMapper etc.) through supplementing with sample Provenance terms, NMs ageing term, NMs characterisation terms etc. as required.

Additionally, we will provide guidelines and training on the usage of standard file formats and ontologies in the context of ACEnano, both for project consortium members, and for the wider stakeholder community:

- Best-practice examples like diXa and ToxBank can be used to create templates for data storage and sharing;
- File formats like ISA-Tab (or current improvements of these including ISA-Tab-Nano) are well suitable since data schemata for different endpoints are already available and they are flexible enough to also accompany modifications needed due to changed and enhanced experimental and computational protocols.

ACEnano partner Douglas Connect coordinated the eNanoMapper project and ACEnano partners have been actively involved in NanoSafety Cluster initiatives discussing the role of harmonisation and ontology development for data to be used in multiple modelling and assessment tools applied to nano exposure and safety. ACEnano will continue to monitor and engage in the parts of such initiatives that relate to NMs characterisation, including ageing, transformation and environmental fate (and hazard) as well as provenance terminologies and the metadata ontologies use to describe datasets produced by ACEnano will map as readily as possible into the data formats and ontologies initially developed by the continuation efforts of eNanoMapper, as well as the newly funded NanoCommons project (coordinated by UoB). This process will be facilitated by working with the US CEINT to further develop Environmental fate ontology for NMs using the aforementioned ISA-TAB-Nano format (now co-lead by CEINT and UoB) to ensure a systematic approach to the structure and format of datasets.

As a starting point, all NMs (partner produced or externally sourced) to be used in ACEnano will be entered and tracked using the Biomax system developed in the NanoMILE platform. This means appropriate characterisation data is recorded and that direct links between initially synthesised particles and their “in product”, released or aged speciation variants will exist and can be further connected to their subsequent characterisation and fate data resulting from experiments (Figure 2). Data sets on characterisation and environmental fate will be in accordance with the ISA-TAB-Nano format to facilitate sharing of data among project partners and researchers, organisations and initiatives external to the project. Additionally, the integrated and harmonised user interfaces for data analysis across the ACEnano partner instrument manufacturers will be linked to the ACEnano ontologies and to the metadata standards.

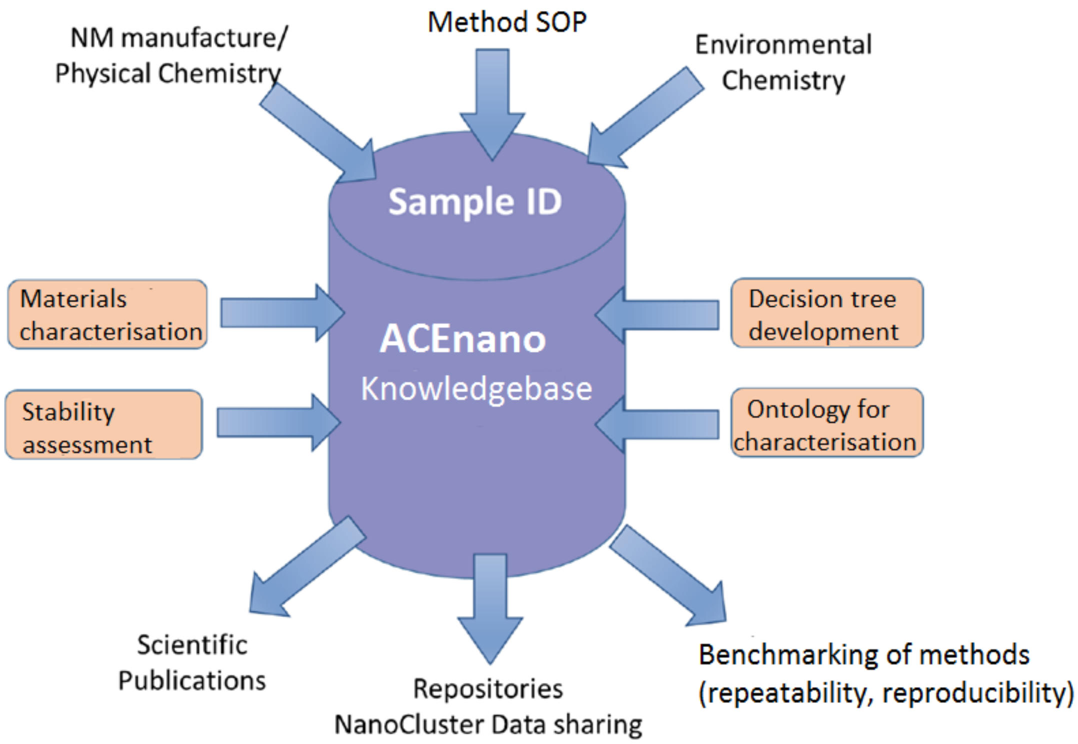

Figure 2. Schematic illustrating the types of data to be held in the ACEnano database and the input data needed.

2.2.4 Increase data re-use (through clarifying licenses)

- *Specify how the data will be licenced to permit the widest reuse possible*

- *Specify when the data will be made available for re-use. 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 useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why*
- *Describe data quality assurance processes*
- *Specify the length of time for which the data will remain re-usable*

It is anticipated that within the lifetime of the ACEnano project some data will be released as part of dissemination activities (e.g. supporting data for papers, protocols or methods papers or benchmarking data). This will be done in collaboration with the relevant partners for each dataset (e.g. method / assay development partners, round-robin testing partners, and/or partners involved in the comparison with / benchmarking against other relevant methods) and will be managed to ensure that there is not any significant reduction in IPR potential. For this purpose a dissemination review process has been implemented in the ACEnano project (Figure 3). Initially, requests to disseminate data will be reviewed by UoB and PROD (WP6 Dissemination leader) with additional advice from the ACEnano project IPR advisor. If there are potential IPR issues to be resolved prior to release of data then further discussions will be held among the partners involved in generating the data, and where such IPR issues are not identified the data is opened up to dissemination through publication, presentation and open access.

It is envisaged that most of the re-usable data will be available to project partners and 3rd parties through open access scientific outputs and associated data, or data and parameters related to the use of the open access ACEnano decision tree, with all of the underlying data sets being available through the ACEnano knowledge warehouse. It is the project consortium's intention to leave this information and data available in as long-term repositories (see 2.2 above) as can be achieved by the end of the project to ensure the legacy and long-term continuation of the projects impact.

Data that may need restricting and embargoing (outside the embargo related to publication of papers – see 2.2) will be those related to method and technical development work undertaken between research labs and equipment manufacturers (e.g. Perkin Elmer, Horiba, Sciex, Biolin, CSEM, Postnova, TOFWERK, Vitrocell and Malvern) within ACEnano in relation to the analytical and characterisation techniques and equipment. This data will be released when safe in relation to the IPR of the partners, either through full scientific papers, technical development notes by the manufactures in collaboration with the academic partners, or in the associated data to manuals and protocols, whichever is most appropriate.

As noted also in section 2.2.2, all ACEnano data will be include in the ACEnano knowledge warehouse, as it is generated. The following data quality assurance processes will be applied to the datasets at the point of entry into the ACEnano knowledge warehouse:

- Tools for performing quality control analyses and (pre)processing will be made available in ACEnano, e.g.: https://github.com/BiGCAT-UM.
- Some partners (e.g. UoB, NERC) are developed their own pipelines for quality control and analysis of NMs characterisation data, along the lines of the processes utilised by NeXus (an international standard for the storage and exchange of neutron, x-ray, and muon experiment data) and OME-XML (Open Microscopy Environment XML). Similar approaches will be

integrated for the core methods developed, optimised and benchmarked in ACEnano.

- We also plan to integrate tools for manual curation of datasets.

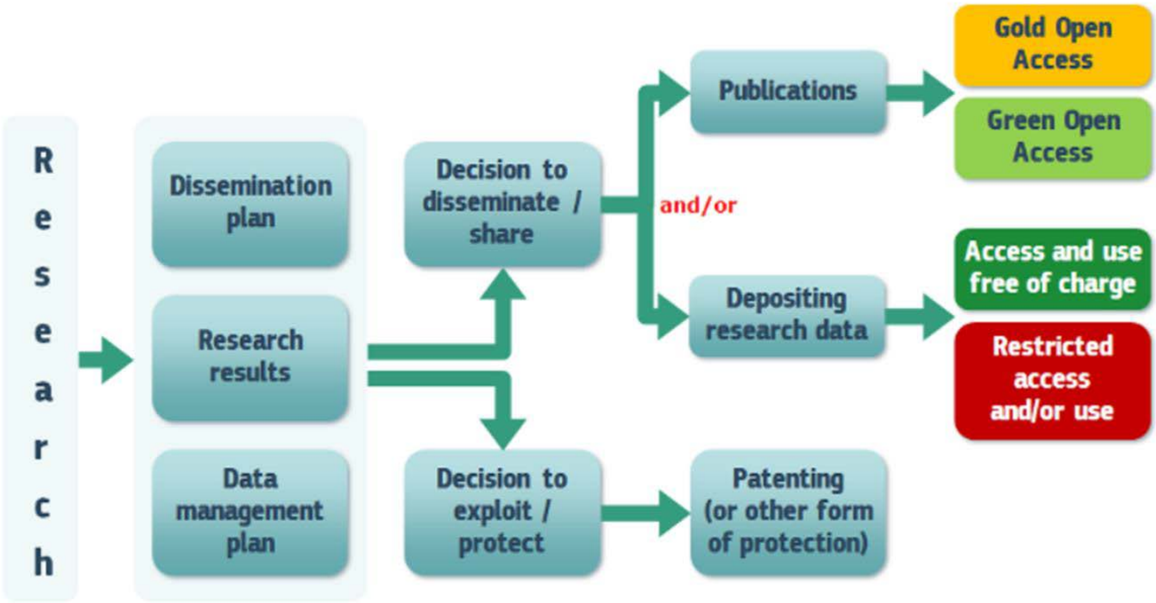

Figure 3: Schematic of decision tree for the dissemination of research results and data from the ACEnano project.

Quality assurance in the data processing, analysis and decision tree tools will be implemented as follows:

- Protocolling of the performed calculations, increasing the reproducibility of the studies, will be supported by the automatic logging and auditing functionalities of modern microservices frameworks as well as the integrated workflow management systems.
- Validation of the services will be enforced by the consortium and appropriate measures of measurement uncertainty should be provided for all methods.

2.3. 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*

At present the costs for making the ACEnano data FAIR are to be confirmed, but will leverage the OpenRiskNet and NanoCommons e-infrastructure platforms. 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 for using limited funds for getting 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 ACEnano knowledge warehouse database platform of the partner datasets lies with the individual partners.   Their interlinking with the decision tree software and inter-comparison for benchmarking of the various ACEnano method and assay SOPs and protocols will be the task and work package leaders for the WP in which the method development, optimisation or benchmarking originated (Figure 4).. The responsibility and budget for developing and managing the web based structure (including aiding partners with data entry and uploading) of the ACEnano decision tree is held by DC, with additional budgets for advising and design help being held by UoB and PROD.

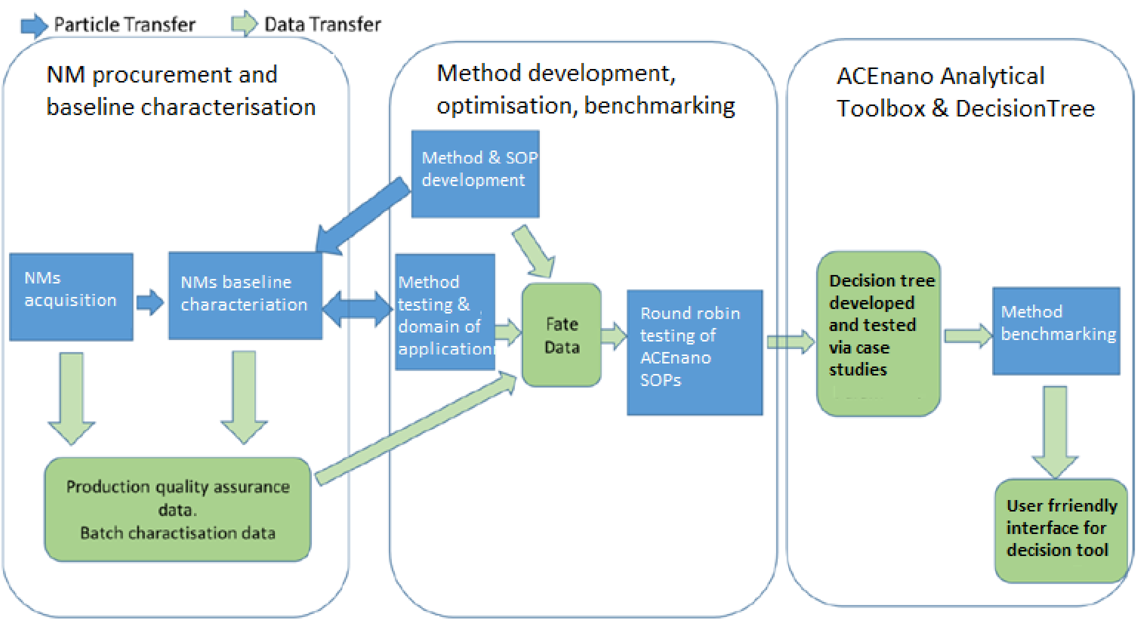

Figure 4. Flow of NMs and data in the ACEnano project.

2.4. DATA SECURITY

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

In its raw format and while held by the ACEnano project partners, it is expected that each partner will take normal precautions for data security, both in terms of data recovery as well as secure storage and transfer of sensitive data. For most of the partners this is dealt with in terms of institutional serves under regular offsite backup, and the main risk elements are standalone analytical machines not linked to networks, and data held and developed by the smaller partners. Here the recommendation is that files are regularly transferred to secure or alternative offsite facilities, either via mobile storage media or “cloud storage” facilities. For example, UoB will offer all partners access to BEAR DataShare, its file synchronisation and sharing service that allows users to securely save and sync files with colleagues and partners anywhere in the world, from any device. It provides 25GB storage capacity per user initially, which can be increased as needed for data-heavy partners (e.g. chemical images, TEM etc.).

Data sharing and transfer among persons or partners will be, where appropriate, third party secure file transfer facilities, such as the ACEnano knowledge warehouse, Figshare and OpenAIRE and via the internal communication platform. In the longer-term it is anticipated that NanoFASE 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).

2.5. 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*

No ethical issues arise from the project overall, or the data management aspects, since the focus of the project, and the data collection, are on the analytical performance of the methods and their benchmarking against one another, as well as on the characterisation datasets resulting from the application of the various methodologies to the library of ACEnano test nanomaterials, including representative test materials from the JRC repository, and certified reference standard materials from various suppliers, including JRC, NIST and BAM.

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.

2.6. 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 "t*he 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: https://intranet.birmingham.ac.uk/as/libraryservices/library/research/rdm/Data-management-plans.aspx

The ACEnano project will utilise the following aspects of the UoB data storage services for the UoB and partner-generated datasets, respectively:

- 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 25GB storage capacity per user.

3. References

1. Guidelines on FAIR Data Management in Horizon 2020, Version 3.0, [Internet]. Available: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
2. OpenAIRE - OpenAIRE [Internet].. Available: https://www.openaire.eu/
3. New H2020 DMP guidelines: http://www.dcc.ac.uk/news/new-h2020-dmp-guidelines
4. EU-US Nanoinformatics 2030 Roadmap (note only the consultation version link is available at the time of submission of this deliverable report. Available at: https://www.nanosafetycluster.eu/Nanoinformatics2030 Accessed 02/01/2018.
5. ISA-Tab-Nano, https://wiki.nci.nih.gov/display/icr/isa-tab-nano and Thomas DG, Gaheen S, Harper SL, Fritts M, Klaessig F, Hahn-Dantona E, Paik D, Pan S, Stafford GA, Freund ET, Klemm JD, Baker NA. ISA-TAB-Nano: A Specification for Sharing Nanomaterial Research Data in Spreadsheet-based Format. BMC Biotechnology, 2013, 13:2. https://doi.org/10.1186/1472-6750-13-2
6. NanoREG templates (phys-chem, in vivo and in vitro toxicology. http://www.nanoreg.eu/media-and-downloads/templates/269-templates-for-experimental-data-logging. Accessed 02/01/2018.
7. Izak-Neu E, Huk A, Reidy B, Uggerud H, Vadset M, Eiden S, Voetz M, Himley M, Duschl A, Dusinska M, Lynch I. Impact of Storage Conditions and Storage Time on Silver: Nanoparticles' Physicochemical Properties and Implications for their Biological Effects. RSC Advances, 2015, 5, 84172-84185.
8. Baer DR, Munusamy P, Thrall BD. Provenance information as a tool for addressing engineered nanoparticle reproducibility challenges. Biointerphases, 2016, 11, 04B401. https://doi.org/10.1116/1.4964867
9. ISO/PRF 20579-4 Surface chemical analysis -- Guidelines to sample handling, preparation and mounting -- Part 4: Reporting information related to the history, preparation, handling and mounting of nano-objects prior to surface analysis, 2017, ISO, Geneva, Switzerland.
10. https://www.w3.org/TR/prov-overview/ Accessed 02/01/2018.