# Metadata stewardship in nanosafety research: learning from the past, preparing for an “on-the-fly” FAIR future

Exner, Thomas E.; Papadiamantis, Anastasios G.; Melagraki, Georgia; Amos, Jaleesia D.; Bossa, Nathan; Gakis, Georgios P.; Charitidis, Costas A.; Cornelis, Geert; Costa, Anna L.; Doganis, Philip; Farcal, Lucian; Friedrichs, Steffi; Furxhi, Irini; Klaessig, Frederick C.; Lobaskin, Vladimir; Maier, Dieter; Rumble, John; Sarimveis, Haralambos; Suarez-Merino, Blanca; Vázquez, Socorro

*DOI:*
10.3389/fphy.2023.1233879

frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 17 August 2023
DOI 10.3389/fphy.2023.1233879



OPEN ACCESS

EDITED BY
Sungsook Ahn,
Consultant, Bloomington, United States

REVIEWED BY
Jef Hooyberghs,
University of Hasselt, Belgium
Sonja Herres-Pawlis,
RWTH Aachen University, Germany

*CORRESPONDENCE
Thomas E. Exner,
thomas.exner@sevenpastnine.com
Iseult Lynch,
i.lynch@bham.ac.uk

RECEIVED 02 June 2023
ACCEPTED 28 July 2023
PUBLISHED 17 August 2023

CITATION
Exner TE, Papadiamantis AG, Melagraki G, Amos JD, Bossa N, Gakis GP, Charitidis CA, Cornelis G, Costa AL, Doganis P, Farcal L, Friedrichs S, Furxhi I, Klaessig FC, Lobaskin V, Maier D, Rumble J, Sarimveis H, Suarez-Merino B, Vázquez S, Wiesner MR, Afantitis A and Lynch I (2023), Metadata stewardship in nanosafety research: learning from the past, preparing for an "on-the-fly" FAIR future.
*Front. Phys.* 11:1233879.
doi: 10.3389/fphy.2023.1233879



# Metadata stewardship in nanosafety research: learning from the past, preparing for an "on-the-fly" FAIR future

Thomas E. Exner[1]*, Anastasios G. Papadiamantis[2,3], Georgia Melagraki[4], Jaleesia D. Amos[5], Nathan Bossa[6,7], Georgios P. Gakis[8], Costas A. Charitidis[8], Geert Cornelis[9], Anna L. Costa[10], Philip Doganis[11], Lucian Farcal[12], Steffi Friedrichs[13], Irini Furxhi[14], Frederick C. Klaessig[15], Vladimir Lobaskin[16], Dieter Maier[17], John Rumble[18,19], Haralambos Sarimveis[11], Blanca Suarez-Merino[7], Socorro Vázquez[6], Mark R. Wiesner[5], Antreas Afantitis[3,20] and Iseult Lynch[2,20]*

[1]Seven Past Nine GmbH, Schopfheim, Germany, [2]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, United Kingdom, [3]Nanoinformatics Department, NovaMechanics Ltd., Nicosia, Cyprus, [4]Hellenic Military Academy, Vari, Greece, [5]Center for the Environmental Implications of NanoTechnology CEINT, Duke University, Durham, NC, United States, [6]LEITAT Technological Center, Terrassa, Barcelona, Spain, [7]TEMAS Solutions GmbH, Hausen, Switzerland, [8]Research Lab of Advanced, Composite, Nano Materials & Nanotechnology (R-NanoLab), School of Chemical Engineering, National Technical University of Athens, Athens, Greece, [9]Department of Soil and Environment, Swedish University of Agricultural Sciences, Uppsala, Sweden, [10]CNR-ISMCC, Institute of Science, Technology and Sustainability for Ceramics, National Research Council of Italy, Faenza, Italy, [11]School of Chemical Engineering, National Technical University of Athens, Athens, Greece, [12]European Commission, Joint Research Centre (JRC), Ispra, Italy, [13]AcumenIST SRL, Etterbeek, Belgium, [14]Transgero Limited, Limerick, Ireland, [15]Pennsylvania Bio Nano Systems, Doylestown, PA, United States, [16]School of Physics, University College Dublin, Dublin, Ireland, [17]Labvantage—Biomax GmbH, Planegg, Germany, [18]R&R Data Services, Gaithersburg, MD, United States, [19]CODATA-VAMAS Working Group on Nanomaterials, Paris, France, [20]Entelos Institute, Larnaca, Cyprus

**Introduction:** Significant progress has been made in terms of best practice in research data management for nanosafety. Some of the underlying approaches to date are, however, overly focussed on the needs of specific research projects or aligned to a single data repository, and this "silo" approach is hampering their general adoption by the broader research community and individual labs.

**Methods:** State-of-the-art data/knowledge collection, curation management FAIrification, and sharing solutions applied in the nanosafety field are reviewed focusing on unique features, which should be generalised and integrated into a functional FAIRification ecosystem that addresses the needs of both data generators and data (re)users.

**Results:** The development of data capture templates has focussed on standardised single-endpoint Test Guidelines, which does not reflect the complexity of real laboratory processes, where multiple assays are interlinked into an overall study, and where non-standardised assays are developed to address novel research questions and probe mechanistic processes to generate the basis for read-across from one nanomaterial to another. By focussing on the needs of data providers and data users, we identify how existing tools and approaches can be re-framed to enable "on-the-fly" (meta) data definition, data capture, curation and FAIRification, that are

sufficiently flexible to address the complexity in nanosafety research, yet harmonised enough to facilitate integration of datasets from different sources generated for different research purposes. By mapping the available tools for nanomaterials safety research (including nanomaterials characterisation, nonstandard (mechanistic-focussed) methods, measurement principles and experimental setup, environmental fate and requirements from new research foci such as safe and sustainable by design), a strategy for integration and bridging between silos is presented. The NanoCommons KnowledgeBase has shown how data from different sources can be integrated into a one-stop shop for searching, browsing and accessing data (without copying), and thus how to break the boundaries between data silos.

**Discussion:** The next steps are to generalise the approach by defining a process to build consensus (meta)data standards, develop solutions to make (meta)data more machine actionable (on the fly ontology development) and establish a distributed FAIR data ecosystem maintained by the community beyond specific projects. Since other multidisciplinary domains might also struggle with data silofication, the learnings presented here may be transferrable to facilitate data sharing within other communities and support harmonization of approaches across disciplines to prepare the ground for cross-domain interoperability.

KEYWORDS

**data management along the data lifecycle, nanosafety, FAIRification ecosystem, data-users perspective, data-providers perspective, reporting standards, machine-actionable metadata**

# 1 Introduction

Most researchers, research funders and stakeholders agree on the value of data ("data is the new oil: Like oil, data is valuable, but if unrefined it cannot really be used", a quote attributed to mathematician Clive Humby). However, implementing processes to refine data and make it Findable, Accessible, Interoperable and Re-useable (FAIR) in accordance with the FAIR data principles [1], is challenging in practice, requiring a mix of technological solutions and community-defined and agreed standards. This has more recently been visualised via the FAIR Hourglass, which distinguishes between "FAIRification" of research data (by which data captured using localised or domain-specific practices are transformed into formats that follow open standards for interoperability, opening the door to machine-processing), and FAIR Orchestration in which the FAIR-ready data is put into action by exposing them to software applications and services that can perform operations on them [2].

The nanosafety research community has been an early adopter of the FAIR principles (even before these were actually established), with efforts underway to support nanosafety data management, sharing and re-use as far back as 2008 when the Data Working group (initially WG4 and later WG-F) of the nanosafety cluster was established [3]. Early efforts in ontology development for nanomaterials included the Nanoparticle Ontology [4] which was subsequently integrated with other nanosafety-related concepts such as exposure and toxicology via the eNanoMapper ontology [5] which applied an approach of splicing across existing ontologies to re-use terms where appropriate and add new terms only where these had not been previously defined. However, given the diversity of domains that converge in nanosafety, including materials design and characterisation, toxicology, exposure and risk assessment, establishing data sharing based on harmonised and standardised ontologies and (meta)data schemas across the complete nanosafety community has proven to be challenging in practice. True progress will require the joint efforts of all players in the data management cycle (data creators, analysts, curators, managers, customers as well as data stewards and data shepherds) [6] and an ecosystem of tools and supports for all stages of the data lifecycle based on machine-actionable metadata.

While most researchers see the need for the implementation of open and FAIR data principles, the transition from theory to practice is still very much ongoing. Data management and FAIRness regularly show up as a topic for discussion in round tables targeting strategies to reduce barriers. For example, the NanoRisk Governance conference—the final meeting of the 3 nano-risk governance projects—in January 2023 had a session on reducing barriers to data sharing. Similarly, the EU's working document on "Supporting and connecting policymaking in the Member States with scientific research" [7] underlines the ongoing struggle. As a consequence, data produced in national and EU-funded research projects still remains largely fragmented or inaccessible and is captured and stored using a multitude of (often incompatible and/or non-machine actionable) data formats and shared only within project consortia. This is not, or at least not solely, due to a lack of good data management concepts and proposed solutions such as data curation workflows, but rather is due to a lack of community adoption of these as a consequence of a lack of follow-on funding to further embed them into the community and optimise the services for other (non-project specific) users. Indeed multiple EU- and US-funded projects have developed data warehouses that are adopted by the community to greater or lesser extents (e.g., NanoSafety data Interface (https://search.data.enanomapper.net/), Nanoinformatics Knowledge Commons [8], NanoPharos (https://db.nanopharos.eu/), NanoCommons Knowledge Base (https://ssl.biomax.de/nanocommons/cgi/login_bioxm_portal.cgi), see [9] for further examples) and important steps towards establishing a common nanosafety e-infrastructure have been made to further break down (or connect) data silos [6,9–14].

Given the amount of investment and progress to date, we present here a novel proposition: that instead of developing completely new infrastructure solutions that magically "solve" the issues of making research data re-useable, the needs of the data generators and data users/re-users can be met by reframing the existing tools and

approaches to more closely match these perspectives. To demonstrate this, we generated hypotheses to explain the slow uptake of existing solutions into data management best practice from the data generator and data user perspectives, and use these hypotheses to show how the existing concepts and services can be repurposed to restructure data workflows and bring them into line with the identified needs of the nanosafety (and beyond) community.

Hypothesis 1—Existing solutions need to be re-designed to be both generic and customisable in order to address the broadest set of data provider's needs (the data provider perspective): Most data curation templates and data upload interfaces available currently were designed in the context of very specific projects with their specific goals and/or are based on (meta)data models of specific databases/data warehouses. This makes it very hard to implement/apply these in the standard data workflows of research institutions or individual laboratories (existing or to be established). They are, on one hand, not flexible enough to cover all of the information needed for intra-lab reporting and quality control. On the other hand, work performed for different projects has to be reported in different formats requiring provision of different metadata and, thus, a lab involved in multiple projects cannot commit to one single solution. Even more general reporting formats like ISA-TAB and its ISA-TAB-nano derivative are not constructed in a way that they follow the experimental workflows and do not fit naturally into current laboratory practices. Therefore, data management is seen as something separated from the normal lab practice, adds to the already immense workload and, thus, is avoided whenever possible or postponed to the last minute rather than being something that is integral to the data generation process.

Hypothesis 2—Existing solutions can be made interoperable through recording of rich metadata and a deeper understanding of the concept of data re-use (the data user perspective): It is definitely impossible to predict every future reuse of data and build a single solution that provides the data in the optimal format for each potential re-use. However, what is shared and what is considered sufficient metadata is often decided with one specific application or one specific stakeholder group in mind and then implemented in a bespoke data solution without considering further reuse, exploitation, or publication. Even if the (meta)data schema might be optimal for this application, important metadata is often missing and the chosen structure makes it almost impossible to go beyond this primary use. That different tasks need the same data presented differently can be showcased by the fact that the NanoCommons consortium provides two data warehouses as part of its infrastructure (and supports many more): the NanoCommons Knowledge Base and the nanoPharos database for modelling. The first has a material-focused data model, which is well suited for risk assessment and safe-by-design applications since it gives easy access to all information available for the specific material at hand. The second is optimised to provide training and test datasets for developing nanoinformatics models and tools. Here the goal is not to have all information for a single material but information for one or more specific endpoint(s) for as many materials as possible. As described below in more detail, it is possible to convert one representation to the other but automated solutions for this are only slowly emerging and often data provenance trails [15] are destroyed during the process rendering independent data quality control almost impossible and highly impractical. Thus, publicly available data is only available in one or the other format depending on what it was used for in the primary application or, if it is stored in multiple sources, users might not even be aware that they use duplications of the same data. It is then often seen as easier (or even necessary) to go back and re-curate the data again from the primary literature (see, e.g., [16]) either because it is not found in the data sources frequently used by the user and/or curation status and metadata such as provenance, internal and external quality control performed and how datasets are updated when new versions of the data become available (i.e., dataset versioning) is not always clear. Here also, solutions are emerging (e.g., Data Version Control, https://dvc.org/doc/use-cases/versioning-data-and-models), but are not yet routinely integrated into nanosafety data warehouses.

We believe that solutions to address all issues outlined in the hypotheses above have been proposed and introduced in some form in the nanosafety community. However, they suffer from the same shortcomings as identified above for nanosafety data: they are fragmented and trapped in incompatible and overly specific silos. In the following sections, we will present a subset of these solutions, for which we see high potential for extracting them from their silos and integrating them into overarching data management workflows aligned with standard laboratory practices and, in this way, for establishing an on-the-fly data management practice that increases the FAIRness and machine-actionability of the resulting data and thus its potential for re(use). Following such a concept, the data is collected and curated in a form useable for the intra-lab processing and analysis (primary use) but also directly for sharing and reuse, drastically reducing the workload required to complete the harmonisation and FAIRification tasks. Data stewards and data shepherds [6] can then use these workflows to customise them for specific settings while still keeping them harmonised and interoperable. Note that this subset represents our personal preferences and it is not meant to represent a full review of everything existing or to disqualify other solutions not listed here. The selected approaches are also only presented to the detail needed to understand their benefits and what they provide to the overall integrated and harmonised nanosafety data management concept. More information can be obtained from the original literature introducing the approaches (references provided) and, in the constantly updated and extended NanoCommons User Guidance Handbook, which is developed as a one-stop knowledge resource around nanosafety data and nanoinformatics. Another important information source on nanomaterial/nanosafety data management and FAIRification is the nanoHUB, while an emerging source is the GO FAIR AdvancedNano Implementation Network [17].

## 2 (Meta)data reporting templates and minimal reporting guidelines

The term reuse is a complicated concept and the exact meaning of "reuse of research data" (and what is needed to be able to do so) does not seem to be fixed yet, varying between disciplines and individuals with no common standard applied as yet [18]. Efforts to evaluate criteria that distinguish reuse from other related actions and to find a definition that makes reuse measurable and understandable to a broader audience have:

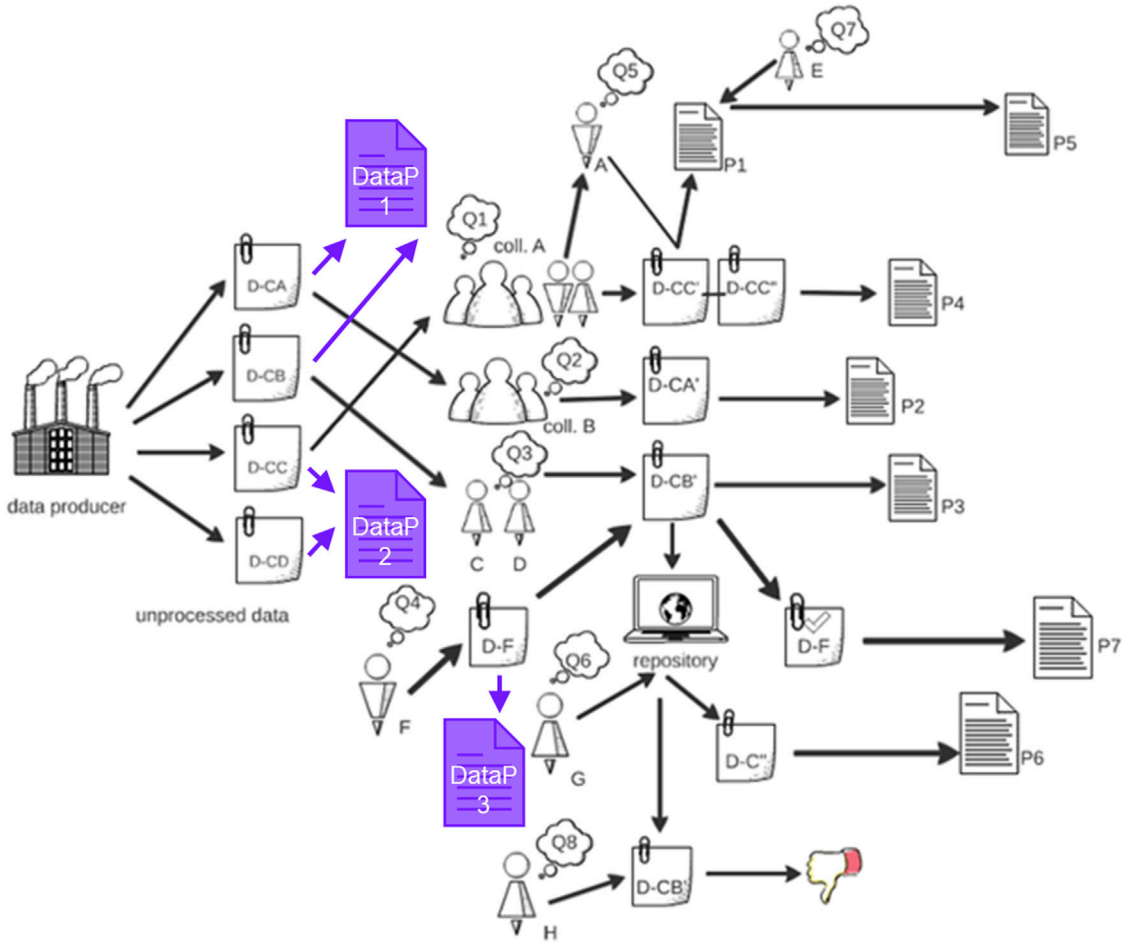


FIGURE 1
Schematic illustration of the concept of reusable research and the complexity of (re-)use scenarios: D are various datasets (D-), which are used alone and/or integrated with other datasets and used by different individuals and collaboration teams to address a range of questions (Q) leading to a set of publications (P). Importantly, characteristics like "character of data," "user," "purpose" and "time" are no longer solid pillars, and the process of data use and re-use is no longer linear as these now occur in parallel and progress at different rates. A key concern of data generators has always been that their findings or datasets will be "scooped" by others if they share the original data before their primary research findings are published. One way to (at least partially) mitigate this, as shown in the modified version of the original figure, is inclusion of one or more data-focussed papers (DataP). These describe the dataset itself and how it was generated as well as give an example of the sort of questions the dataset can be utilised to answer without providing a full interpretation of what the data mean in the specific context in which it was generated. Adapted from [18].

1. Distinguished the three variables: research question, research data and research method to enable a clear differentiation between reuse and related concepts. According to this view, the reuse of data enforces the usage of the same data with a different method and a different research question in mind [19]; or
2. Focussed on identifying the characteristics of reuse by examining the etymology of the term and analysis of the current discourse, leading to a range of reuse scenarios that show the complexity of today's data-driven research landscape (Figure 1), and underlining that there is *no reason to distinguish use and reuse.* In this re-conception, (re)use is defined as the use of any research resource regardless of when it is used, the purpose, the characteristics of the data and its user. This reflects the fact that the research landscape is no longer linear, but rather a complex and dynamic landscape where characteristics like "character of data", "the user", "purpose" and "time" are no longer solid pillars, and thus use and re-use are interchangeable [18].

For the purpose of this paper, we adopt the non-linear approach and consider the data management requirements to be broadly the same irrespective of whether researchers are undertaking FAIRification to support first analysis of their own data or (re-) use of data of/by others. Even if this sounds contradictory at first, we do this exactly since we acknowledge that the incentives and need for FAIRification are seen as being lower for data providers, but effective data sharing can only work with their commitment. FAIR is, in our opinion, stressing too much the re-use. Harmonisation and interoperability of data is supporting the first time use of data since the data provider can use existing interoperable tools in their analysis and, in this way, save time by avoiding additional data wrangling to just get the data into the format needed by these tools. Additionally, achieving the objective of Safe-and-Sustainable-

by-Design of advanced materials is not possible by individual experiments, studies or research groups alone. Thus, the people processing and analysing data in the broader context of the project are not those who generated the data in the lab, and thus all areas of FAIR become highly relevant to make the data useable in the first instance before even considering re-usability. Therefore, we hope that this paper will foster further "change of mindset" to recognise the inherent value of data FAIRification for each individual, each research group, each project and finally for the scientific community more generally. Thus, when analysing the two hypotheses presented above, and how the existing data management tools can be modified or applied to address both the data generator and data user needs more effectively the resulting ecosystem should reflect the non-linear nature of the scientific process and that the two perspectives are broadly aligned.

Hypothesis 1 claims that existing data curation and reporting templates are too strongly influenced by the application and stakeholder group they were designed for to be universally applicable in the on-the-fly data collection settings of individual labs. Similarly, hypothesis 2 stresses that the often limited amount of available metadata is insufficient to support data re(use) in other areas. However, this does not mean that standardised reporting templates do not fulfil an important role in the data management process and that the work of the past and ongoing projects were not essential to reach the current level of awareness across the community and the establishment of good data management practices within the specific projects and leading to the availability of FAIRer data from these projects. The learnings from projects that have used the existing standardised templates show the importance of guiding data providers on the (meta)data to be provided to describe an experiment in enough detail to allow evaluation of the derived conclusions by other researchers as well as by risk assessors and regulators. This is further underlined by the proclamation of the reproducibility crisis [20], which states that more than 70% of 1,500 researchers (most likely self-selected as researchers were aware of concerns about reproducibility) have tried and failed to reproduce another scientist's experiments using the information provided in scientific publications.

Before starting with the description of individual templates, we stress that the focus during template development needs to be put on the data providers, i.e., the data models implemented in these templates should not be defined by what a specific data solution needs but should be customised to provide support in selecting the most appropriate ways to describe the outcomes of studies and experiments according to the consensus of the data producers. Only then will the templates be intuitive enough to be filled in directly by the experimentalist and responsibility for data documentation can be shared between the multiple parties, working together to execute the specific experimental workflow representing individual or combinations of assays that constitute the full study.

The NANoREG project templates [21] and the corresponding database running on the eNanoMapper data warehouse, are made for standard methods used in regulatory settings, which are well defined in the OECD Guidance Documents and standard Test Guidelines [22]. In these cases, the requirements on metadata completeness are less demanding compared to new emerging techniques since information on the specific experimental setting and protocols is already available (although not in a machine-actionable format) from the Test Guidelines [22]. A similar approach was adopted by the GRACIOUS and Gov4Nano projects for data quality evaluation (e.g., [23,24]). In contrast, projects in which new methods have been/are developed and optimised, which look at complex materials or nano-enabled products, or follow nanomaterial's fate in complex (biological) environments (e.g., H2020 projects ACEnano, NanoFASE, RiskGONE and ASINA) require high detail in metadata description. Such requirements are directly reflected in the design and implementation of their data reporting templates and respective data warehouses. RiskGONE has been championing the Template Wizard as a means to support the research community in the co-creation (by data generators together with data management experts) of data collection templates [25].

Complexity in nanosafety research comes from multiple sources, all of which need to be covered by metadata. These include:

1. Nanomaterial characterisation (section 2.1): Nanomaterials have very different properties compared to their bulk counterparts and unambiguous characterisation of the material in the relevant dispersion medium before (and ideally during and after) starting the experiment is of uttermost importance (e.g., [26,27]). What characteristics have to be reported is still a topic of debate but different standards and minimum reporting requirements have been released, and the ECHA nanoforms guidance specifies key characteristics required for regulatory purposes. Characterisation becomes even more important, and more challenging, when the material is embedded into a product matrix.
2. Non-standard (mechanistic-focussed) methods (Section 2.2): The unique features of nanomaterials but also the general trend to replace existing *in vivo* tests by alternative *in vitro* and *in silico* methods make it necessary to develop new assays and apply them without a standard Test Guideline being in place. Since specific experimental settings and parameters sometimes have a large influence on the obtained results and might, in extreme cases, make the results questionable (e.g., if the nanomaterials interfere with the assay for example, [28], or if the test conditions inadvertently remove key proteins from the system (e.g., [29])), the exact protocols with all of the parameters utilised have to be reported, preferably for the complete development and optimisation process as well as the final protocol, to define the potential ranges of the parameters and the overall applicability domain of the experiment.
3. Measurement principles and experimental setup (Section 2.3): As an addition to the previous point, the new methods might also use completely new measurement principles and experimental setups including new biological *in vitro* cell models, model organisms but also computational/*in silico* models, which need to be described in more detail compared to commonly used experiments. Harmonised approaches for reporting of characterisation assays (CHADA), models (MODA) and biological organisms (BODA) are in various stages of development and application, as described in detail below.
4. The environments role in nanomaterials fate (Section 2.4): It is well known, and demonstrated, that the nanomaterial constitution and its properties are dependent on the current environment and the different life-cycle stages the particles

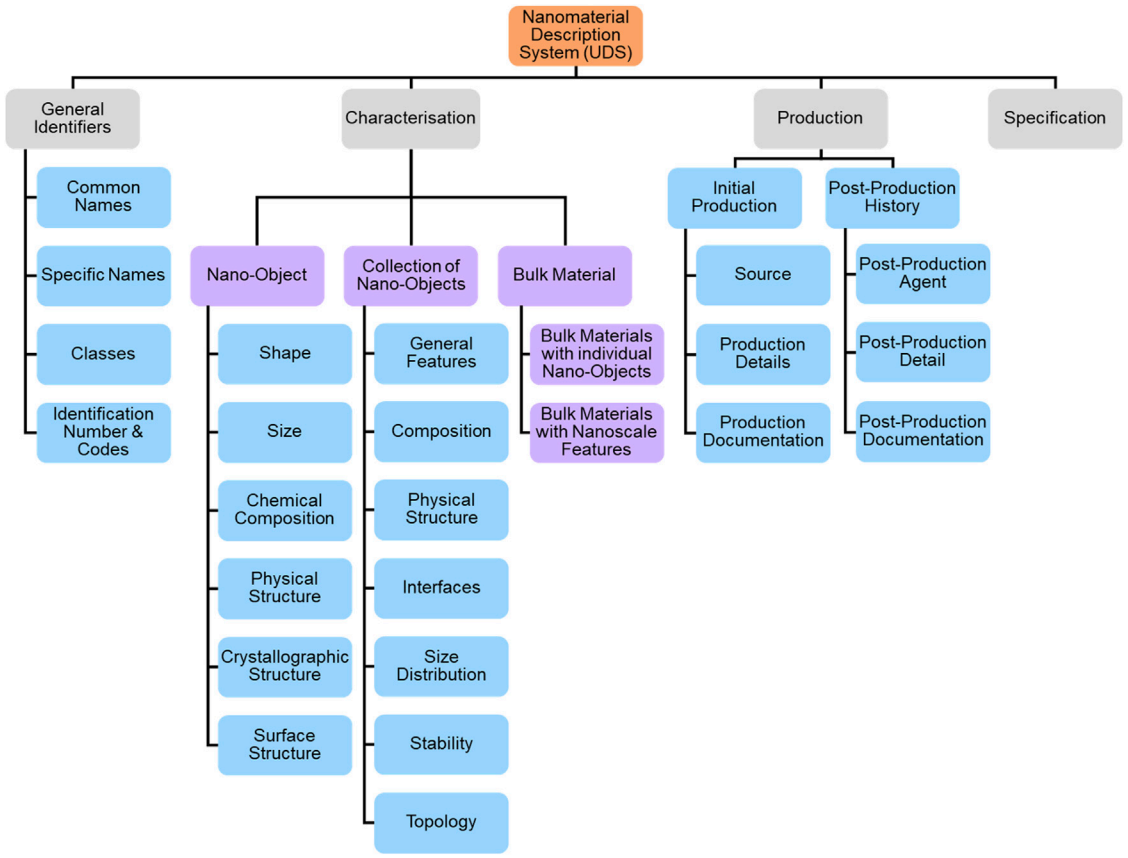


FIGURE 2
Organigram of nanomaterials characteristics covered by the UDS, including the pre- and post-production history, also known as provenance information. Redrawn from [39].

already went through [30–32]. Therefore, additional information preferably in the form of metadata has to be provided and the data management has to be flexible enough to offer the ability to map complex experiments e.g., from nanomaterial environmental exposure and fate experiments.

5. Meeting requirements from new research foci (Section 2.5): Even if the first four points are highly interlinked, they were still addressed separately as can be seen by the fact that the examples used in the following subsections are all taken from different projects. However, the emerging research focus on safe-and-sustainable-by design of new (advanced) materials, integrating physicochemical characterisation, hazard, exposure and fate with life-cycle assessment and social-economical impact, will require an even stronger harmonisation and interoperability of (meta)data across multiple disciplines and, thus, even stronger guidance on how data is reported and annotated to allow central processing and analysis.

## 2.1 Nanomaterial characteristics

Completeness of a dataset covers two different aspects. Firstly, the information requirements have to be established. For nanomaterials and more specifically nanosafety, this is foremost the physicochemical characteristics of the material necessary to uniquely identify the material and define its state at the beginning of the investigation and the required toxicological endpoints [33], which depends on the applied regulation as well as other settings such as, e.g., the manufactured or imported amount of the material in REACH [34]. Secondly, the full sets of (meta)data have to be defined for each experiment used to determine a physicochemical or toxicology endpoint, which have to be provided to understand and evaluate the provided data. This first subsection concentrates on the information requirements for physicochemical characterisation of nanomaterials while the following subsections will then look at the metadata to be provided for specific assays.

Different European and international agencies request different information, even if the way to generate the data is standardised by the OECD in the form of Test Guidelines [22]. Additionally, problems such as establishing the uniqueness of a nanomaterial or evaluating the equivalency of two nanomaterials to the desired level to allow data integration are not limited to regulatory settings. Therefore, global guidelines should be applied and continuously updated to reflect the state-of-the-art and guide the best practices for the physicochemical characterisation of materials irrespective of the intended use of the data. These could form the basis for more harmonised regulations supporting, e.g., the EU chemicals strategy

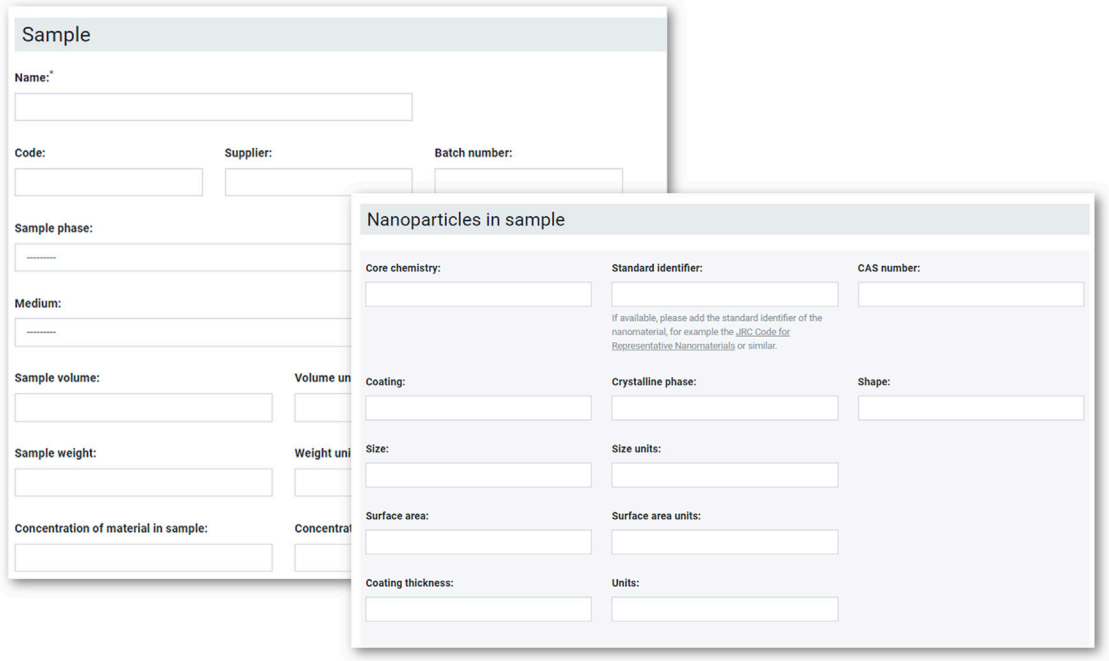


FIGURE 3
Composite figure of two screengrabs, showing examples of a web interface to an ACEnano questionnaire asking for metadata used for description of a sample analysed.

for sustainability: One substance—one assessment [35]. Such consistent reporting is promoted by the use of minimum information standards, which specify the requirements for data content and provide a structured framework for capturing the information. In 2008, a project on 'Minimum Information for Biological or Biomedical Investigations' (MIBBI) [36] started a reporting tradition, which was almost immediately transferred to nanosafety research through the Minimum Information for Nanomaterial Characterization (MINChar) Initiative [37]. A newer version of a checklist for nanomaterials characterisation is the Minimal Information About Nanomaterials (MIAN) [38], which is used for curating and archiving the nanomaterial information in the Nanomaterial Registry for understanding its biological and environmental implications. MIAN defines the 12 essential physicochemical characteristics, i.e., composition, purity, shape, size, size distribution, surface chemistry, aggregation/agglomeration state, solubility, stability, surface area, surface charge, and surface reactivity as well as relevant techniques. It even provides a defined list of parameters (metadata) vital to the replication of the data produced and thus provides starting points for both aspects of (meta)data completeness discussed above.

While MIAN was driven by the requirements of the specific database, the Uniform Description System for Materials on the Nanoscale (UDS) was developed focusing on (pre-)standardisation needs by the joint CODATA-VAMAS Working Group (WG) on the Description of Nanomaterials shortly after but independently of MIAN [39,40]. Various industry standards have been derived from the UDS including, ASTM E3144-19, E2909-13, E3172-18 and E3206-19. While there are large overlaps of the essential characteristics between MIAN and UDS, the latter separates the characteristics into the properties of a nano-object and of a collection of nano-objects and includes information on the production and post-production history which provides essential provenance information (Figure 2).

The above standards and checklists provide clear criteria for essential characteristics to define a nanomaterial and separate it from similar but still distinct materials with potentially different functional and toxicity profiles, but they are quite unsuitable for fast database search and retrieval of information on the same or similar materials (where similarity depends on the specific use or application). One way to support linking of data collected on the same object are unique identifiers such as the European Union Joint Research Centre (JRC) nanomaterials repository identifiers and the European Registry of Materials (ERN) identifiers, which are in use to separate nanomaterials from bulk chemicals. Even if they uniquely refer to one material, they do not hold information on the material themselves but rely on services providing this information, similar to the Chemical Abstract Service (CAS) registry numbers for chemicals or Digital Object Identifiers (DOIs) as a general identifier. In contrast, chemical line representations, like canonical Simplified Molecular Input Line Entry System (SMILES) and the IUPAC International Chemical Identifier (InChI) have shown that it is possible to efficiently store chemical structural information in a form that can also be used as an identifier for efficient storage and retrieval [41,42]. Recently, an extension of the InChI to cover nanomaterials was proposed [43], with further work ongoing via IUPAC project 2022-001-2-800 - InChI extension for nanomaterials. This "InChI for nano" or NInChI encodes the chemical composition, morphology/shape, size, crystal structure, and chirality. Further discussions on other properties to include are

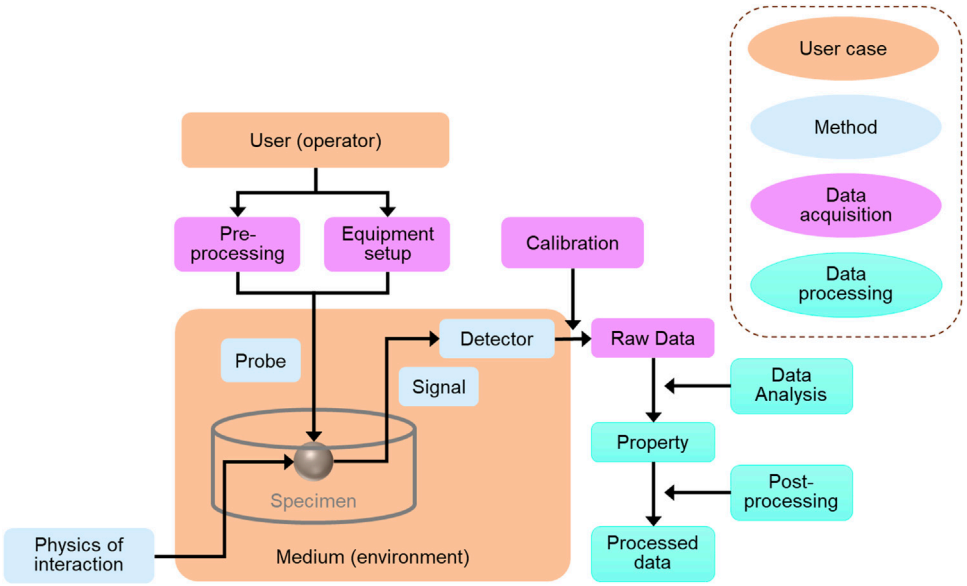


FIGURE 4
Schematic flow-chart providing a visual representation of the structure guiding (meta)data collection in the CHADA format. Reproduced from [53].

ongoing and will be integrated into the first official NInChI standard to be released soon. However, it should also be noted that the NInChI is not meant to cover all information needed to define nanomaterial sameness in every situation but rather it defines groups (or sets) of materials and other (meta)data required to determine if data was generated on nanomaterials similar enough to be combined for use in risk assessment or modelling, for example. Discussions on what additional information might be useful to include as an Auxillary Information file accompanying the NInChI are also underway.

## 2.2 Non-standard methods for characterisation and toxicity assessment

Continuing with examples from physicochemical characterisation of nanomaterials, it is widely understood that many of the techniques for nanomaterials characterisation may impose artefacts that depend on the technique chosen for measuring a specific endpoint, the sample preparation steps, and on the instrument settings, all of which can influence the final results (e.g., [44,45]). Generation of knowledge of sufficient quality, considering these influences and limitations, for new, innovative methods relies on reported (meta)data that allow detailed parameter analysis. Thus, for a general data concept covering methods of different maturity and not limited to already standardised and regulatory validated techniques, the (meta)data schema has to orient towards the methods with the highest information requirements, i.e., it needs to be able to flexibly cover metadata customised to the specific methods and their relevant parameters. The ACEnano project addressed this challenge by adopting a data reporting concept in the form of questionnaires. High-level standardisation was achieved by differentiating between sample preparation, measurement and data treatment protocols as the highest level and then subsections like nanomaterial information, sample description, experimental equipment, and predefined sample preparation steps as the next lower level of the hierarchical (meta)data schema. The user is then guided through workflows to provide the required fields for defining the material, the endpoint, and the measurement technique. The implementation of these workflows as web-based user interfaces for data upload (Figure 3) to the ACEnano Knowledge Warehouse (https://acenano.douglasconnect.com/) is described in more detail below. Such interfaces are more flexible than pre-defined templates in the way that later steps in the workflow depend on earlier answers and follow directly the needs of the data provider.

The questionnaires can, thus, be seen as an extension of the checklist idea in the way that the user, additionally to knowing what to report, can now use the web forms to also fill in the (meta)data. It is important to note that the (meta)data requested by the questionnaire is information, which is currently often only available from protocols, standard operating procedures (SOPs) or Test Guidelines (depending on the degree of standardisation of the method). Having this information in the form of structured metadata, potentially in addition to the free text protocols, allows their automatic integration into the processing and analysis steps and, in this way, enables better exploration and understanding of their influence on the results. However, one question not completely answered in the ACEnano interfaces is how data management can be flexible enough to allow the creation of workflows for novel techniques using new, innovative methods but still provide a harmonised and interoperable outcome across multiple providers of the same or similar type of data. For this, existing expert knowledge needs to be collected to define a (meta)data schema composed of mandatory, recommended and optional (meta) data fields via community consensus. The (meta)data schema has then to be made available in a form allowing reuse by other projects and continuous refinement to stay up-to-date with the scientific and technical advances as well as the increased demands coming from new research directions like Safe-and-Sustainable-by-Design. ACEnano started a (meta)data schema for the technologies covered by the project

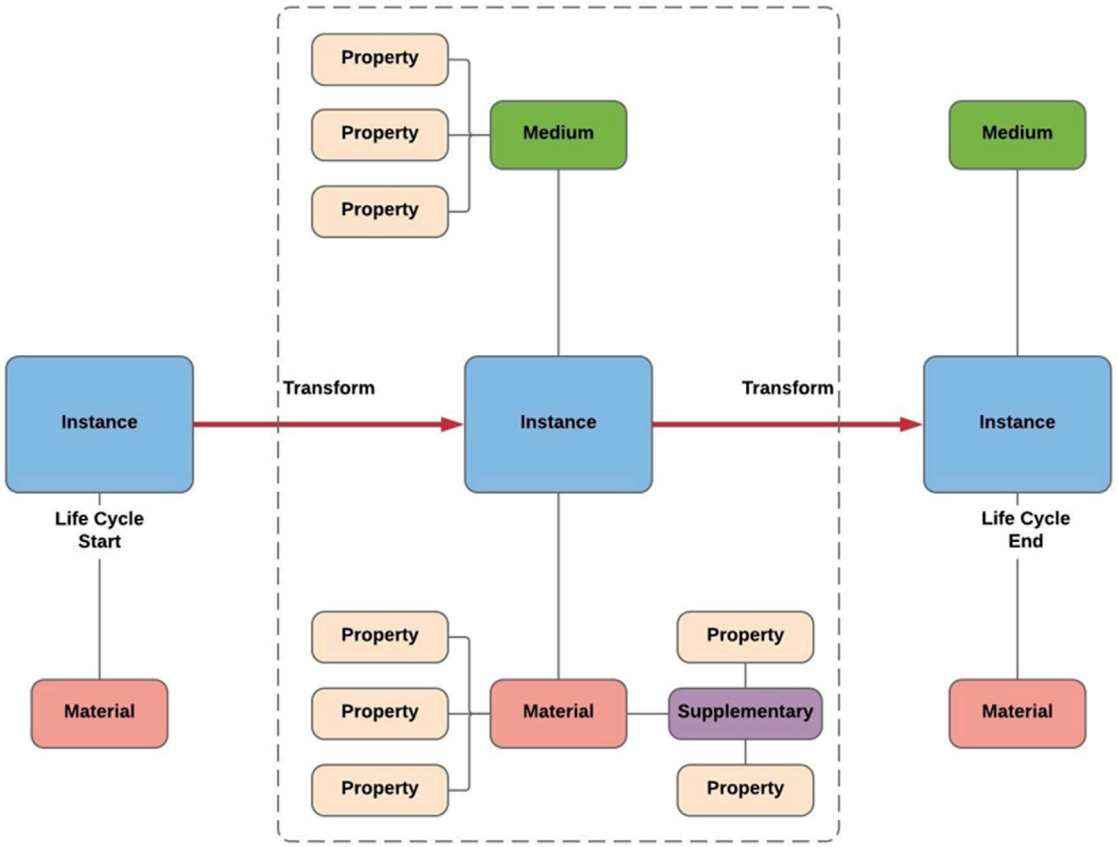


FIGURE 5
Illustration of the experimental instance, which contains the monitoring and analysis of both the material and the environment in which it exists, allowing easy visualisation of where transformations of the nanomaterials properties may occur. Reproduced from [8] with permission from Elsevier. NOTE: Colours are intrinsic properties of an instance map.

in the form of an internal database of fields. The ASINA project, described in Section 2.5, has followed a similar approach starting from the NanoFASE templates.

## 2.3 Models, measurement principles and experimental setup

The increasing diversity and complexity of the materials under investigation, and the underpinning protocols and their variations, leads to the extended demands for (meta)data coverage described in the previous two subsections. Additionally, the novelty of emerging technologies the novelty of the emerging technologies to be more nano-specific or to avoid *in vivo* testing requires very detailed descriptions of the models used (experimental or theoretical) and the underlying theory and prediction/simulation/measurement principles. Only in this way, enough confidence in the results can be generated to get acceptance from other researchers or even risk assessors and regulators before the technique is fully standardised and validated, which might not even be fully possible for complex computational models [11,46].

In the regulatory area, the first encounter of guidelines requesting inclusion of descriptions of the used models and algorithms as part of the submitted data was a reaction to the low acceptance rates of *in silico* and especially for Quantitative Structure-Activity Relationship (QSAR) predictions. In 2004, the "OECD Principles for the validation, for regulatory purposes, of (Quantitative) Structure-Activity Relationships Models" were published (OECD, 204) [49]. Specific guidelines followed, such as the "Practical guide: How to use and report (Q)SARs" [48] and the "Read-Across Assessment Framework (RAAF)" [49], which make the use of the QSAR Model Reporting Format (QMRF) and QSAR Prediction Reporting Format (QPRF) for reporting on QSAR models mandatory [50]. The information in the QMRF is structured according to the OECD validation principles [47], and includes a defined endpoint (what is predicted), an unambiguous algorithm (a reproducible description of how predictions were produced), a defined domain of applicability (a metric of how applicable the model is to each prediction made), appropriate measures of goodness-of–fit, robustness and predictivity (overall assessment of how good the model is) and a mechanistic interpretation, if possible (an explanation of how numerical results reflect changing biological functions). A modified version of the QMRF from June 2021 allows the reporting of other *in silico* models (e.g., Structure Activity Relationship models, expert systems, etc.) within skin sensitisation research and more specifically as defined in the Guideline on Defined Approaches on Skin Sensitisation in TG 497 Annex (OECD 497) [51].

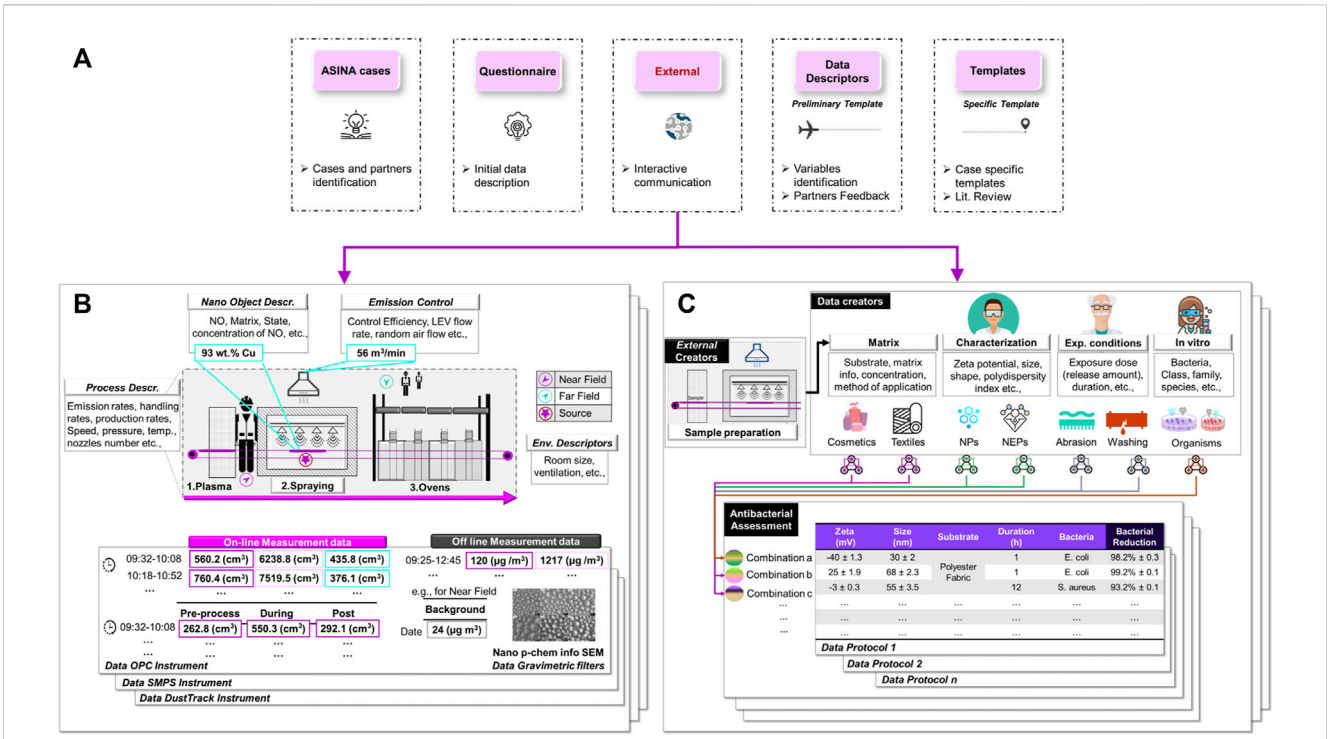


FIGURE 6
Composite figure illustrating **(A)** the roadmap followed in ASINA to generate final templates to be FAIRified for fit-on-purpose specific case studies, and its visualized application in **(B)** the exposure and emission data demonstrating a snapshot of an extensive exposure assessment, and **(C)** the antimicrobial assessments for the products produced in the manufacturing line depicted in **(B)**. Adapted from [56–58].

QMRF were specifically designed for QSAR and are thus not optimal for other types of modelling like *ab initio* and physics-based approaches. Thus, an alternative was proposed by the materials modelling community and is strongly promoted by the European Materials Modelling Council (EMMC). This MOdelling DAta (MODA) reporting template guides modellers/users towards a complete high-level documentation of materials models that starts with end-user cases, i.e., the specific question to be answered, and includes all of the relevant computational details on the solving and post-processing methods that are required for the model reproduction and curation, as well as interfacing with other models [52]. MODA models are classified by the entity of the material being modelled (electron, atom, mesoscopic material entity, continuum volume entity), its quantities (properties of a material or phenomenon), the physics equation of the phenomenon being assessed, its material relations, its solving methods and parameters, the input preprocessing data, and the post-processing data. Concepts similar to MODA have recently been developed for the description of workflows for data-centric models used in various applications [53] and, in this way, become direct competitors of QMRF.

In a similar way to MODA, the concept of materials CHAracterisation DAta (CHADA) was introduced to be applied as a building block for user studies concerning complex material characterization cases accompanied by interactions of modelling and process workflows [54]. It addresses the same concerns as the ACEnano concept above, that characterisation data can vary significantly, due to differences in the sample treatment, characterization methods used, the equipment setup and calibration, and the characterization conditions [55]. Similar to MODA, CHADA provides a systematic structure consisting of user-case, data inputs regarding the sample (material type, sample treatment, dimensions, medium), the characterization method (the physical basis of the method, probe, detector, equipment, the calibration, setup and conditions used), the data acquisition, and data post-processing [54] as depicted in Figure 4. An equivalent reporting process for Biological Organisms (BODA) is under development currently and will be standardised via a CEN Workshop Agreement in due course.

All three reporting formats, QMRF, MODA and CHADA, described in this section have the advantage that detailed information on the technique/model, measuring principle/theoretical background is provided to the reader as part of the overall dataset, rather than being provided as separately stored documents as in the assay-specific templates discussed before, e.g., the ACEnano Techniques and Endpoints Catalogue. On this basis, the usefulness of the methodology, model or measurement as well as the expected quality and confidence can be evaluated and, in principle, the results can be reproduced and the methods adopted to similar questions without the need to consolidate other documents and information resources. However, there are also two main issues with MODA and CHADA and to a lesser extent with QMRF: guaranteeing completeness and machine readability. Both of these issues are caused by the fact that

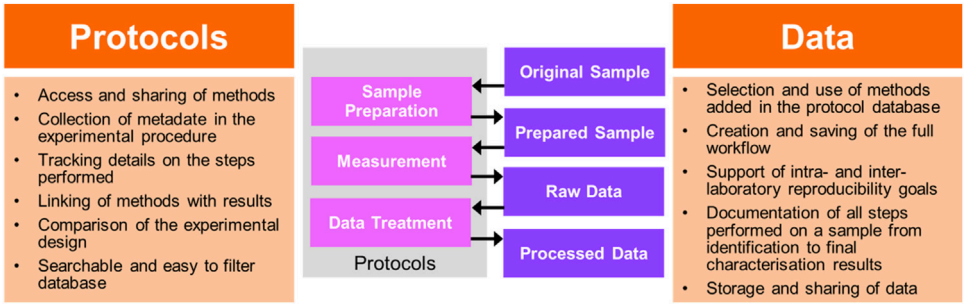


FIGURE 7
Schematic illustration of the interdependencies between the two main sections of a Knowledge Infrastructure (Protocols and Data).

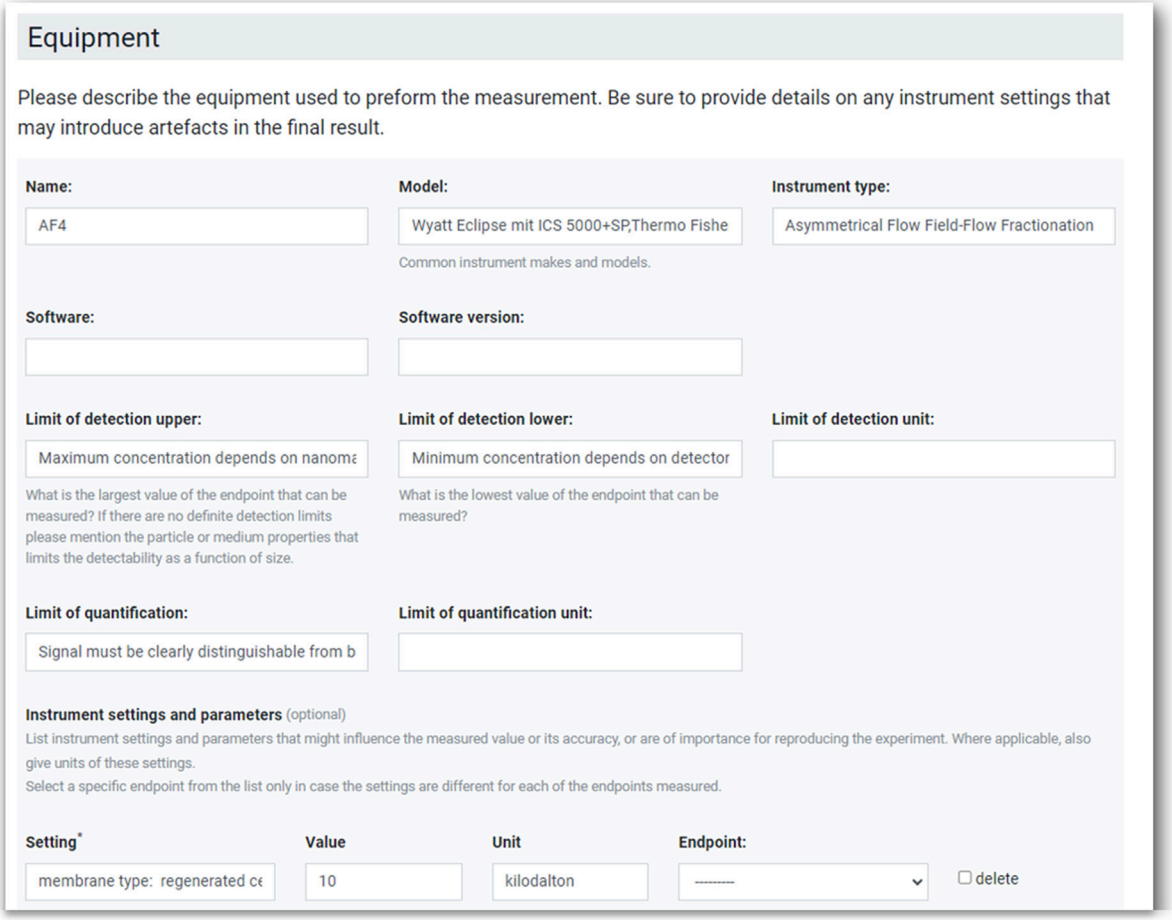


FIGURE 8
Screengrab of an exemplar web interface for metadata, describing experimental equipment and specific instrument settings and parameters.

MODA and CHADA provide a high-level structure and also ask for some specific metadata fields but that, at present, the information is then provided as free text. Therefore, combining the advantages of the structured reporting with better guidance on required (meta)data and its structure, as proposed in Sections 2.1, 2.2, 4) would result in a higher FAIRness.

## 2.4 The environments role in nanomaterials fate

The NanoInformatics Knowledge Commons (NIKC) data template [8] was developed to be used as a guide for curators extracting nanomaterials data from the environmental nano-safety scientific literature. Since the corresponding database has a

strong focus on mesocosm experiments, it was clear from the beginning that the template needed to be able to encode the inseparability of the nanomaterial from its history and current environment. This is realised by introducing the concept of experimental nanomaterial instances, which is also the major difference to other data capture templates. Instances represent significant points of the experiment where the properties of the nanomaterials surroundings have changed which may result in changes to the nanomaterials properties, and can, in this way, cover the complete or a large part of the nanomaterial's life cycle [57]. The spatial or temporal progression can then be visualised in instance maps (Figure 5) showing transformations from one instance to the next (e.g., upon dispersion into a medium, upon contact with an organism, etc.) and identifying the reporting information needed to characterise the material at that stage and in the current medium/environment/biological compartment [8,57].

This clear and easy to visualise representation of the experiment combined with a dynamic and versatile template structure led to the adoption and extension of the concepts by the NanoFASE and the NanoCommons projects. The resulting EU modified version, which is designed for use in primary data capture, makes input of data at the stage of data creation easier, gives more guidance on what metadata should be covered (in the NIKC case, this was determined by what metadata was available in the publications) and allows the automatic semantic annotation when inserted into the NanoCommons Knowledge Base (https://ssl.biomax.de/nanocommons/cgi/login_bioxm_portal.cgi) and is described also in [58].

## 2.5 Meeting requirements from new research foci

All of the above, minimal reporting requirements, (meta)data standards and templates, and input interfaces, have to continuously follow the science and technology to be able to present the state-of-the-art. This needs to be done based on the learnings from previous projects and by adopting and adapting the existing concepts and solutions. Knowledge transfer has to be based on technical solutions allowing extensions and modifications to constantly improve the (meta)data completeness preferable with documentation of the decision process. Such tools could be designed similar to the tools used in ontology development allowing easy browsing, e.g., but also collaborative development of ontology terms and definitions by domain experts, e.g., the Terminology Harmonizer developed in the GRACIOUS project. However, this has to be complemented by a trained team of data stewards and data shepherds as stressed in the first paper of this series [6] guaranteeing that we do not reinvent the wheel every time a new research project starts or a new database is designed. One example, in which this was implemented is the ASINA project. ASINA is part of a cluster of projects focusing on Safe-by-Design approaches for nanomaterials. It uses representative groups of nano-enabled products in the market (coatings and cosmetics) to formulate design hypotheses and to deliver Safe-and-Sustainable-by-design solutions by applying a data-driven approach and methodology [16]. The methodology for designing new nanomaterials/nano-enabled products encompasses the merging of distinct data related to different scopes of the product such as improved functionality (for example, anti-aging capacity for cosmetics or photocatalytic activity for textiles), cost-effectiveness, environmental sustainability, and safety. The data shepherd of ASINA has conceptualised and initiated the design and implementation of data FAIRification processes with multiple stakeholders (i.e., data creators, data analysts, and data re-users) who were previously unaccustomed to the notion of data management and FAIRification (Figure 6A). This resulted in a harmonised set of customised, method-specific templates for exposure (Figure 6B) and antimicrobial functionality (Figure 6C) data and more are under development to cover other use cases [59–61].

Safe-and-Sustainable-by-design for advanced materials, as well as other related research fields like nanorisk governance and nanofabrication, are highly multidisciplinary integrating product optimisation, risk and life-cycle assessment as well as ethical and socio-economical aspects needing harmonisation across all these fields as achieved by the ASINA templates. However, innovative research will require to be based on even more and more different fields and, thus, different types of data made available as part of a European or even better global data ecosystem. This is achieved by participating in and harmonising with cross-disciplinary projects on data harmonisation, semantic annotation and FAIRification like OntoCommons and WorldFAIR.

# 3 (Meta)data collection interfaces and integration functionality

The information requirements to define the material, test system, measuring principle and protocol described in the previous section easily results in hundreds of (meta)data fields. For example, Elberskirch et al. identified 300 important parameters for nanosafety research, but also stressed that these would need to be complemented with additional specific parameters in the future [13]. To keep the time demand for inputting all these parameters to a minimum and move the cost-benefit ratio especially for the data providers in the positive direction, the minimum information guidelines and metadata standards need to be complemented with data collection and curation functionality guiding the user but even more importantly giving users the possibility to reuse previous work and automate the data integration as much as possible. This section will describe some of the approaches going in this direction.

## 3.1 Structured (meta)data input via web interfaces

As described in Section 2.4 above, the ACEnano questionnaires implement a system where experiments are described as flexible workflows. The (meta)data is then collected for the steps in these workflows using a deeply interlinked web-based system between protocols and data, as the two components of the knowledge warehouse (Figure 7). The questionnaires are not literally a list of questions, but a number of structured and hierarchical web forms, through which the data provider is guided for filling in the metadata describing the nanomaterial, the individual experimental steps and

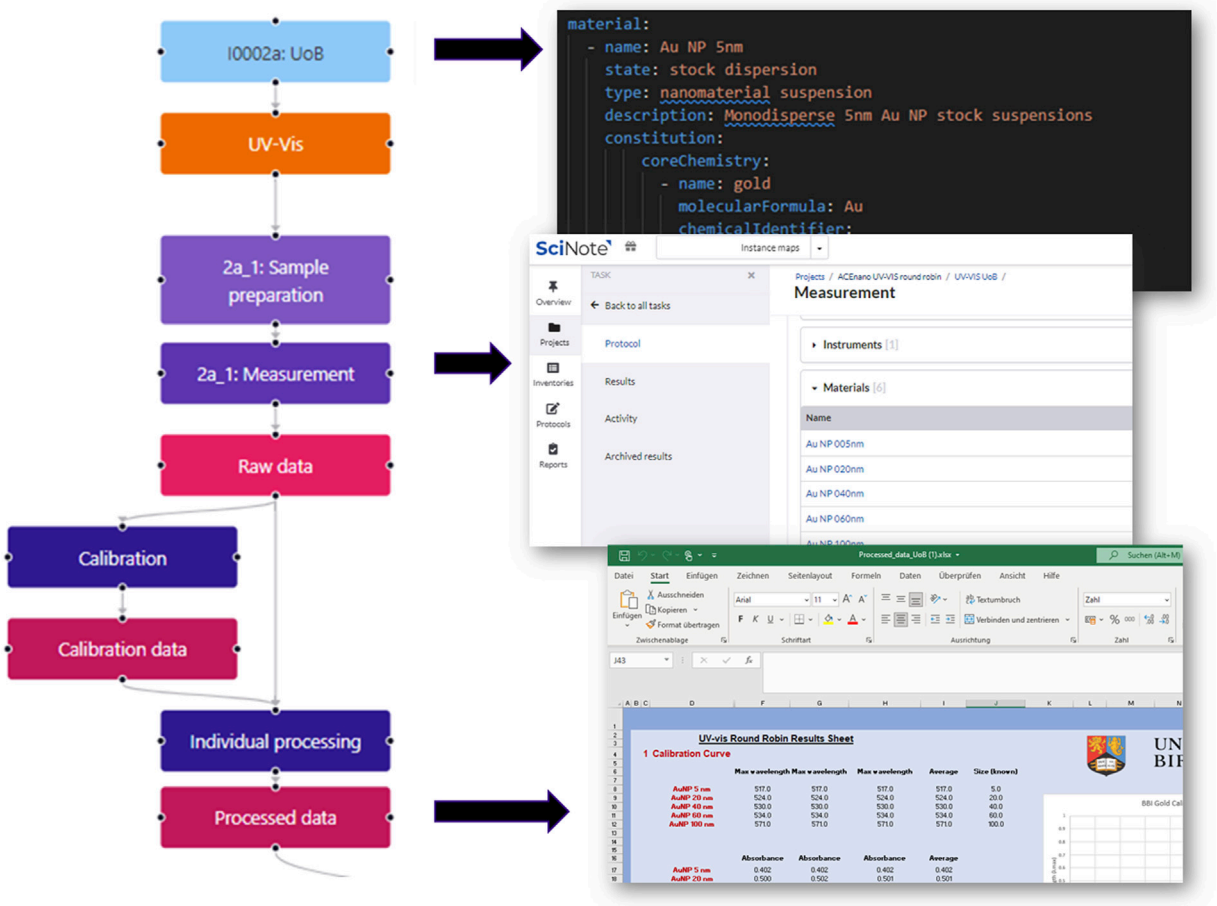


FIGURE 9
Schematic illustration of possible data sources, which can be linked to individual nodes in an instance map, to provide the corresponding data and metadata. At the moment, the (meta)data format is not limited and can be text documents and customised spreadsheets, which are expected to be increasingly replaced by structured and semantically annotated formats produced using data input, management and sharing tools and standards like ELNs and emerging community data standards such as MODA and CHADA. Note: Colours are intrinsic properties of an instance map.

structure of the data, as well as the bibliographical metadata. Previously uploaded similar datasets can be used as guidance by providing the (meta)data structures as well as by being able to reuse parts of the information for scientific equipment, solvents and even complete protocols.

The data upload interfaces are optimised for the corresponding data warehouse and the physicochemical characterisation methods developed and optimised in the project. However, the concept could also be implemented completely independent of a data storage solution collecting the data and then storing it in one of the data transfer formats. Guidance can then come in two forms: i) sections of the web interface can be designed following a well-defined structure, and ii) the interface provides a general structure and the required or suggested (meta)data are loaded based on entries in earlier fields. The first is demonstrated in Figure 3 above, covering the information on the sample. As an example of the second approach, the documentation of lab equipment is presented in Figure 8. It represents a case where the objects described are very diverse and cannot be represented by a single structure. A small number of standard fields like the name, model, software type and version and detection limits are predefined. Specific instrument settings and parameters can then be included but which are required or optional need to be defined outside of the interface.

## 3.2 Organising metadata into a complete study

Electronic lab notebooks (ELNs) are another important tool for achieving on-the-fly data management, which addresses the needs of the data providers and increases harmonisation and digitalisation. ELNs offer, at least in parts, experimentalists functionalities such as i) creating a protocol library which they can implement in their experimental workflows, documenting any deviations and then sharing with others in the group, project or even publicly, ii) pre-designed and pre-annotated data templates that already integrate aspects of data annotation and FAIRness and iii) a means of converting all this information into reports and other digital objects. Working together in teams and being able to comment on the work of others can identify even small differences in the protocols, which might lead to different results, providing quality control and can lead to the development of SOPs

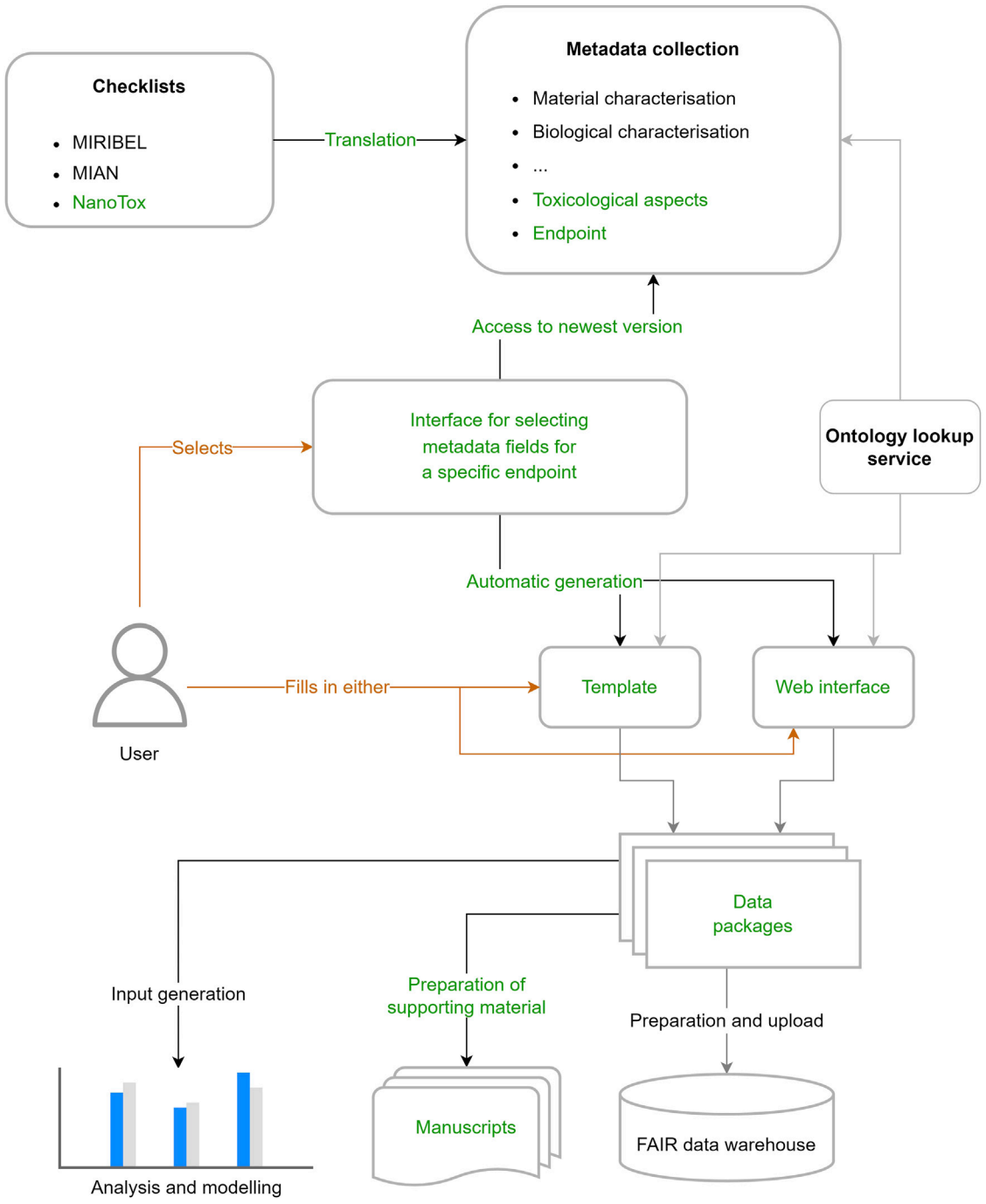


FIGURE 10
Schematic illustration of infrastructure (relations) for continuous development and improvement of minimum reporting checklists and (meta)data schemas and for providing those to users for guiding (meta)data reporting.

and standard testing guidelines. Many commercial and open-source ELNs exist and reviewing them all is beyond the scope of this publication. At the time of writing, SciNote, eLabFTW, Labfolder, Benchling, Sapio, and Chemotion are often used stand-alone ELN software solutions and others are available bundled with laboratory information management systems. As always, different solutions provide different features based on the needs of their users and have strengths and weaknesses when applied to a specific area. We highlight this with two examples. For example, the SciNote ELN has proven beneficial for reporting nanomaterial characterisation and bio-nano interaction experiments (e.g., [62]) but is limited with respect to managing chemical registries and reactions. The latter is the focus domain of the Chemotion ELN for example. This shows that there is no one-solution-fits-all and especially in interdisciplinary projects, the combination of multiple ELNs and, in more general, integration of these tools in the broader data

management and FAIRification ecosystem will be needed. The latter includes guidance on (meta)data input according to minimum reporting guidelines and (meta)data standards as well as using standard (meta)data exchange formats, linking to sophisticated data processing and analysis workflows, and providing the information in for of (meta)data, which can be provided as part of the dataset to public data warehouses.

Pulling all this together can be facilitated by the instance map concept described in Section 2.4 above. The high-level structure of these maps can organise even very complex studies, in which multiple researchers or even scientific groups collaborate. To showcase this, a simple approach was integrated in the newly developed instance map tool. It not only makes the creation of maps simple but also allows linking of the instances, materials, environment, properties and the newly introduced protocols and data, combinedly called nodes in the following, to (meta)data. Each node has a set of standard properties like licence and contributors but most importantly a reference to a data file. At the moment, no limitations are set regarding the data file type. Thus, text files and spreadsheets can be used in the same way as references to entries in ELNs like SciNote and/or Chemotion, or even in protocol repositories (e.g., Protocols.io) (Figure 9). Figure 9 demonstrates the ability of the instance maps to reproduce the protocol and data workflow of the ACEnano knowledge base, using data from a EV-vis round robin study [63]. Instance Maps can be extended to introduce additional steps, e.g., for describing the internal and external calibration needed for UV-Vis experiments, as well as external calibration, quality control and maintenance of biological systems. They can be even extended to represent the full structure of the CHADA templates presented in Figure 4 above. Since each node has its own attached file, the most appropriate format can be chosen, e.g., ELN pages of protocols, spreadsheets for tabular data, or even files for serialisation of complex structures like JSON and YAML.

## 3.3 (Meta)data integration

The goal of data management is not storing the data somewhere but to make it available for use and reuse. From the data user perspective, establishing a semantic and FAIRification framework makes more sense for large settings like projects and public warehouses, and is considered out of reach for individuals and most individual institutions. In reality, this is not the case and understanding the semantic mapping requirements for each use case is helpful for selecting data collection and curation tools in the most efficient and appropriate way. When (meta)data is collected according to a defined data model (following the guidance from Section 2) and preferable in an annotated format, mechanisms to integrate heterogeneous information both across data types and across data sources can be implemented as demonstrated in the NanoCommons Knowledge Base [55]. These map the data schema of the dataset to the semantic model implemented in the database and the data is restructured according to this mapping to fit the needs of the specific data users the database is designed for. Currently, the NanoCommons Knowledge Base provides manually developed mappings for different data types and sources. For data managed directly within the data warehouse of the KnowledgeBase, this results in physicochemical, toxicological, omics, and computational data from different data sources being mapped to the same nanomaterial. As of now the "same" nanomaterials means a nanomaterial identified by the same ID while Instance specific information such as ageing procedures are not necessarily mapped and therefore associated data sits in parallel Instances connected to the same nanomaterial. Additionally, the NanoCommons KnowledgeBase provides a data catalogue to search and browse data hosted by external data management systems in the same interface as the data hosted by NanoCommons.

# 4 Putting it all together

In contrast to the previous parts of this paper, in this section we do not describe existing solutions, but instead present ideas which are not yet realised or are only in early prototypes and demonstrators, to lay out the vision for how things could develop further. Firstly, we describe a way that the community should work together towards consensus (meta)data standards; secondly, we present further work that is needed to make the nanosafety data more machine actionable and, thirdly, an approach to building a distributed FAIR data ecosystem around these machine-actionable standards is outlined.

## 4.1 Continuous development and harmonisation of (meta)data schemas

As described above, many projects and organisations have developed and published minimal reporting checklists, (meta)data standards and standard reporting templates. However, especially in the cases where this is done for the internal data management of projects like NANoREG, NanoFASE and ASINA, these are only available to the project consortia and are only released at the time the data becomes publicly available (e.g., in scientific publications). Reuse is then only possible after the data is released, which itself is only done after the corresponding papers have been published, leading to a large time gap between development and a possible general adoption. In addition, the guidance is often only provided in form of descriptions in papers and as predefined templates, which do not provide information on what the development was based on and how it integrated earlier approaches. To circumvent this, we propose to use a similar approach as applied for the development of ontologies, which use clear track records of changes as well as tools like the GRACIOUS terminology harmonizer for community- and consensus-driven creation of terms and term definitions. As shown in Figure 10, community repositories could be built to provide the possibility of FAIR sharing of checklists and schemas. These would offer the functionality of fairsharing.org, in which these standards still need to be listed additionally and from which inspiration should be extracted with respect to how to run such repositories of standards and how other communities have structured their standards. However community-specific repositories would go beyond such general listings since they could document the usage of the schema within the community as a measure of community consensus and, even more importantly, offer support for selection of schemas using knowledge for our specific community, for comparing schemas, showing the history of modifications and the reasons for these (e.g.,

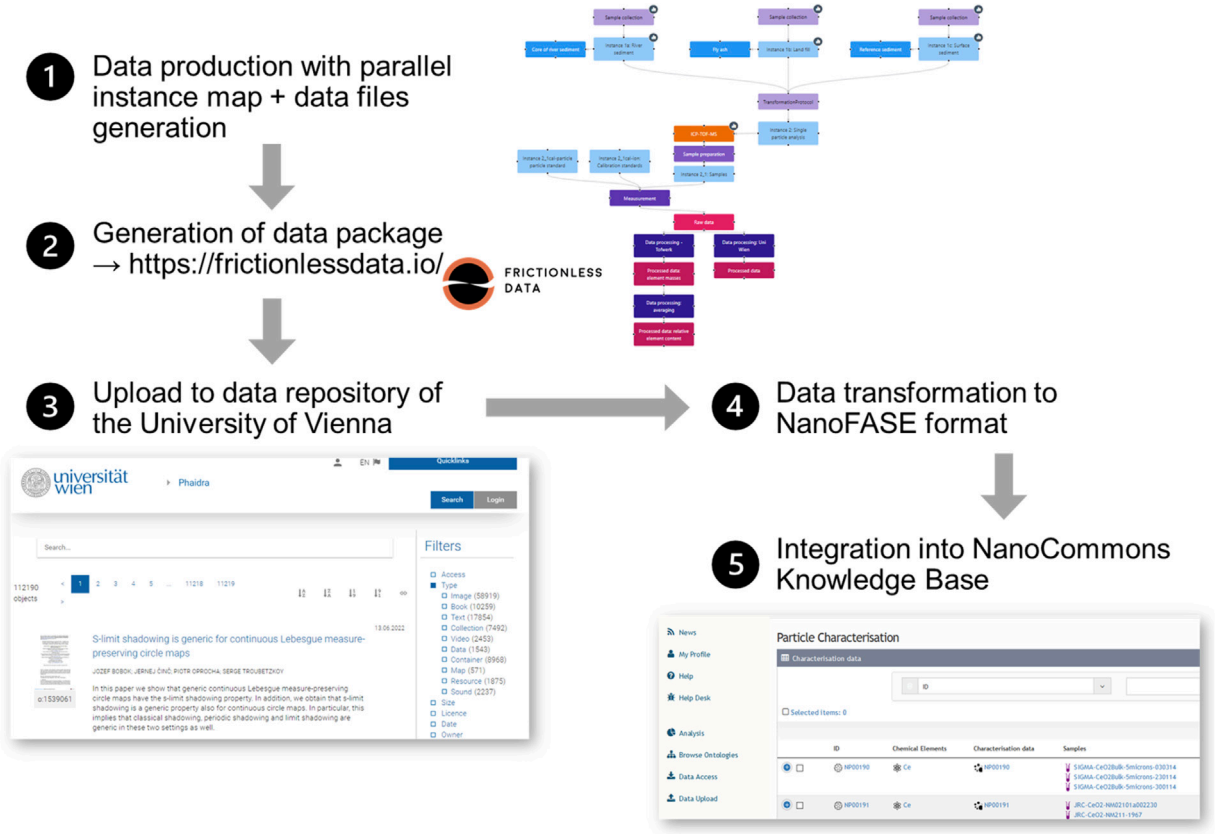


FIGURE 11
Schematic illustration of the steps of the semi-automatic workflow starting with internal data collection using an instance map for the overall structure, upload into the institutional repository and indexing in the NanoCommons KnowledgeBase for improved findability and accessibility.

when templates from ended projects are reused and modified by new projects), and even highlight areas where flexibility is acceptable to adapt the schema to a new assay/endpoint. As a FAIR digital object, schemas would then have full provenance trails, metadata on their development history (which previous schemas have been considered) including versioning, and are semantically annotated to enable comparison of different schemas. Users can then search and browse the repositories and when they have found an appropriate schema, use it to generate (meta)data curation templates or an interactive web form for data input. Using input functionality, harmonised and interoperable (meta)data will be directly produced, which can then be packed into data packages and used in downstream analysis workflows, as an information source for paper writing and to store the data in public data warehouses. There are quite some steps still necessary to achieve such an infrastructure but we hope that this paper shows that we have all the components ready to go.

## 4.2 Machine-actionable data

Another aspect which needs improvement when looking at protocols and SOPs in ELNs but also reporting templates like MODA and CHADA, is machine readability since they currently often use free text fields. Work is ongoing to use MODA and semantic annotation using the Elementary Multiperspective Material Ontology (EMMO) [64] as a basis for metadata schemas of material modelling and a corresponding computer-actionable template iMODA [65]. In the same way, CHADA combined with the Characterisation Methodology Domain Ontology (CHAMEO) [66] can increase the data interoperability based on standardisation of terminology for metadata and classification with taxonomies and ontologies [67] as has already been introduced with iCHADA as a unique means of interoperable metadata structure for robust data management, data traceability and robust data quality. Both, iMODA and ICHADA, are meant to become topics of CEN standardisation workshops (organised by the NanoMeCommons project) with their specifications published as CEN Workshop Agreements (CWAs). However, it has to be seen how EMMO and CHAMEO will interplay with other ontologies like OSMO [64] for simulations, models and optimisation and especially the eNanoMapper ontology [5] currently used as the main ontology for nanosafety and developed by the EU NanoSafety Cluster.

## 4.3 Distributed data storage

The NanoCommons Knowledge Base showed how data, which is available from different sources, can still be brought together into a one-stop shop for searching, browsing and finally accessing data without actual copying of data, and, in this way, how to break down the

boundaries between the data silos. To make this approach more general and separated from a specific solutions would have multiple advantages: i) all benefits from early and on-the-fly data management and sharing would be directly available, ii) other demands like requests to make data available on institutional repositories could be fulfilled at the same time, and iii) multiple views on the same data could be provided by different tools like a material-centric data structure like in the NanoCommons Knowledge Base and an endpoint-centric structure as often needed as input for nanoinformatics applications as is the goal of the nanoPharos database. To bring these conceptual ideas into reality, a pilot was run as part of the NanoCommons project (Figure 11). The first two steps of this pilot covered i) data preparation, ii) curation following the experimental workflow using a highly adaptable data schema to provide the needed flexibility and iii) generation of a data package using the frictionless data specification (https://frictionlessdata.io/introduction/) able to store different data resources with standardised metadata into one sharable file. This was done manually using a data schema optimised for the specific kind of data, characterisation of environmental samples without defined nanomaterial content, guided by (meta)data standards but not (yet) following one exactly. The following steps were then completely automated with upload of the data to the institutional data repository of the University of Vienna (step 3), transformation it into the NanoFASE format using an automated but custom-made workflow (step 4) and indexing it in the NanoCommons KnowledgeBase (steps 5). The custom-made steps were needed because of the custom data model and the (yet) missing semantic annotation requiring a manual mapping between the custom and NanoCommons KB data models similar to the one described in Section 3.3. Continuous development and improvement of the harmonised metadata schemas and ontologies (Section 4.1) and semantic annotation of the data (Section 4.2) will render these manual interventions unnecessary at one point of time and open multiple options to transform the data into a format best for different stakeholders and applications. As already mentioned above, different data sharing solutions could then provide data from different datasets in an enriched form according to the needs of their user base with NanoCommons KnowledgeBase more for risk assessors and regulators (material-centric view) and nanoPharos for model developers (endpoint-centric view). Separating in this way data input and curation completely from data access and (re-)use allows the optimisation of the support tools for each of the tasks individually as long as the metadata provided is complete enough to satisfy the needs of all applications.

# 5 Conclusion

In this paper we have attempted to demonstrate a roadmap *via* which the current "silos" of data management templates, ELNs, reporting formats, and repositories for protocols and data, can be bridged and integrated into a functional FAIRification ecosystem that addresses the needs of both data generators and data (re)users. The two hypotheses generated at the outset to explain the slow uptake of existing solutions into data management best practice by data generators and data users provided a useful framework against which to assess the various tools and solutions developed in a suite of EU-funded projects over the last 15 years. As with all good hypotheses, we could test these, and decide whether to accept or reject them, based on the available evidence. The good news is that both hypotheses can be accepted, and the roadmap presented herein provides the next steps to implement these necessary changes and interlinks to make the silos interoperable and enhance the machine-actionability of nanosafety data.

**Hypothesis 1** - Existing solutions need to be re-designed to be both generic and customisable in order to address the broadest set of data provider's needs (the data provider perspective)

- Templates implementing reporting standards and using standard data transfer file formats, like those from ASINA [59,60], are good ways to guide data providers to report the required metadata in a harmonised structure.
- As (meta)data can be recorded in many different formats, even if the information covered by the (meta)data is the same, we should start thinking more about the (meta)data model represented by the templates than the templates themselves. In this way, we could move towards more machine-actionable solutions like electronic lab notebooks for reporting protocols as long as they structure the metadata according to a defined model and the information can then be extracted to build parts of the metadata.
- The documentation of the (meta)data models in a machine-readable/FAIR way would allow comparison of the different templates and evaluation of the progression of the field, e.g., from the first versions from NANoREG to the newest installations from the ASINA project. These (meta)data models could be used by new projects to create their customised templates even if using a different data management system, which would avoid the constant reinventing of templates which is leading to incompatibility and data silos. Approaches to annotate such (meta)data schemas, at different levels of detail, can be reused such as the DDI-Codebook and DDI-Lifecycle specifications or from the OntoCommons project and are envisioned to be implemented in the MACRAME project. This will become part of the BODA specification that is being established as a data format for assays describing interactions of (nano)materials with biological systems and, in this way, complements the system of data documentation formats started with MODA and CHADA.
- New research and regulatory foci like advanced materials and Safe-and-Sustainable-by-Design increase the need for interdisciplinary data sharing (including data from the social sciences and humanities). Thus, approaches for (meta)data documentation, at least on a high level, need to be aligned and harmonised across all these disciplines, which is facilitated by the work of the WorldFAIR project on developing a Cross-Discipline Interoperability Framework [68] based on the DDI specifications mentioned above and reusing, at least in parts (meta)data standards (e.g., for provenance data), and existing ontologies.

**Hypothesis 2** - Existing solutions can be made interoperable through recording of rich metadata and a deeper understanding of the concept of data re-use (the data user perspective)

- Making data from different data solutions findable in a single user interface as provided, besides more general, non-nano-specific solutions like Google data search and Zenodo, by the

NanoCommons Knowledge Base (see also [58] in this issue) and the NanoSafety data Interface is a step towards provision of easy access to a larger amount of FAIR data.

- This interoperability of data repositories has to be massively extended to cover more resources in order to avoid pushing data into suboptimal data management solutions as a result of basing the decision on where to put the data on what is indexed in this meta-search catalogues rather than determining the optimal solution for this specific data. This listing of available FAIR Enabling Resources could be achieved by data documentation as already described for Hypothesis 1, i.e., clear specification of the (meta)data model used in a data warehouse or even in a specific dataset. In this way, also other external requirements like the push to use institutional data repositories or data solutions endorsed by the funding agencies can be fulfilled even if these solutions are not nanosafety specific as demonstrated in the NanoCommons Transnational Access project with the data centre PHAIDRA of the University of Vienna presented above.
- Enhanced interoperability does not, however, solve the issue that most data is only available from data solutions organised for a very specific use case. Data transfer formats and data models also provide the answer here. If the data files as provided by the user are seen as the primary data source, multiple data solutions can access them and restructure the data according to the needs of their users, e.g., one solutions could offer the material-centric view for risk assessors and another the endpoint-centric view for creating training sets for nanoinformatics applications by automatically restructuring the original data accordingly. This will only work with clearly defined and, in the optimal scenario, semantically annotated (meta)data models. Reusing what is already available and then collaboratively work within the nanosafety community and also across communities on filling in the gaps could lead to quick benefits since even a partial harmonisation and semantic annotation will be enough to make some applications like the cross-resource searching, browsing and accessing possible and full interoperability and computer-actionability can be achieved over time. But this is only possible if we (the community of nanosafety researchers) can agree on what are the most important interoperability goals to cover first and what (meta)data would need to be annotated to enable these.

## 5.1 Beyond nanosafety

Many of the issues discussed above are caused by the intrinsic multidisciplinarity of nanosafety research. This results in many different data types (nanomaterials functionality, physicochemical characterization, human and environmental toxicity, exposure, life-cycle assessment, circularity) as well as methods (*in chemico*, *in vivo*, *in vitro*, *ex vivo*, *in silico*) and throughputs each with their specific requirements for data documentation and management. Other disciplines are facing similar issues and interdisciplinary research across natural science, humanities and social science will make sharing and combining data of very different origin inevitable. To facilitate transfer of knowledge, we provide recommendations in support of intra- and cross-discipline interoperability:

1. The needs of data providers and data users are very distinct and data management solutions should take this into account. Data input should be structured according to the experimental workflow and fit directly into standard lab procedures. Data templates, reporting guidelines and checklists are helpful to identify the required (meta) data but are often too complex to be used as the primary data curation format. ELNs or similar approaches used in the specific discipline help to structure the data input especially if complemented with advanced tools to visualize the experimental workflows and allow linking to other data files and resources, as provided by the instance map concept which enables on-the-fly data curation and documentation. Being structured, the (meta)data can be easily re-organized, often using automated procedures, to conform to a selected template but even more importantly, to the needs of specific data users. These data users are also not a homogenous group and there is no one-size-fits-all solution to provide the data in an optimal way for all potential users/their disciplines.
2. Data standards, templates and database/data warehouse solutions might look very different because they implement a specific way of representing the data tailored for one of the groups described in point 1 (data providers using different methodologies and different data users such as risk assessors and modellers in the nanosafety case). However, the underlying (meta)data models are often, at least in part, quite overlapping. Documenting (meta)data models (the Data Documentation Initiative (DDI), CODATA and ONTOCOMMONS provide general approaches on how to do this) will allow mapping of common features of different datasets onto each other and in this way enable an automated enrichment of datasets with information from other sources. Reusing common substructures (e.g., for documenting data provenance, material and sample origins, experimental steps) will simplify this mapping and, at the same time, allow integration of method-specific aspects to give the user the flexibility to completely describe their experiments. This is especially important for researchers working on the development of new methods.
3. Separation of data input and use will also help research fields to become less dependent on specific solutions. We had to present the recommended concepts and approaches with specific templates or database implementations here since they can only be judged on existing examples. However, relying on one solution, even if it seems to be optimal and provide all required features at the time the decision is made, will limit the sustainability and reusability in the future, for example, if a specific database goes out of business or is no longer being maintained. Generalization of the concepts described in this publication and similar concepts defined by other interdisciplinary communities, combined with respecting the FAIR, TRUST and CARE and, whenever possible, openness principles, will allow data providers to use many different data storage solutions from general options like Zenodo, figshare or EOSC provided services (e.g., B2SHARE), domain-specific options (e.g., NanoCommons KnowledgeBase for nanosafety) or institutional repositories (as demonstrated above with PHAIDRA from the University of Vienna). Data managers and providers of data sharing solutions could then focus on the needs of their "customers" by indexing the data and restructuring it to support specific applications, e.g., by being able to easily find all data for a specific endpoint as training input for machine-learning and

Artificial Intelligence. However, this is only possible if 1) the data model is well described, as part of the metadata associated with the dataset, 2) if the metadata follows standards as much as possible, and 3) is supported by the creation of meta-services, which index all relevant data sources so that the customized data services can find them. For the latter, Google Dataset Search could, in principle, be used but we believe that domain-specific solutions would be more beneficial since they can profit from community-endorsed standards like specialized unique identifiers (NInChI, European Registry of Materials) and preselection of the most relevant sources.

4. What we describe in the first 3 points requires an ecosystem with many components including, but not limited to, repositories of documented data standards/models, templates and checklists to guide (meta)data collections, data input tools (like electronic lab notebooks, workflow visualization (instance maps), and data transfer standards), data storage solutions, data indexing solutions listing relevant data resources, and customized data retrieval solutions for the different data users as well as general and domain-specific unique identifiers, ontologies and other FAIR enabling and supporting resources. This will not come into existence overnight, but the modular concept allows a smooth transition, continuous expansion and improvement (by replacing one early solution by a new improved one) and adaptation to future needs. Efficient implementation needs to be driven by real user demands and solution-providers willingness to address specific existing pain points. Only then, will widespread adoption of the approaches occur. While the first users of the ecosystem will come from inside the domain and thus solutions need to satisfy their needs, over time there should be an increasing number of off-domain stakeholders who profit from the information and knowledge produced, even if this is on a different level of data aggregation and detail. Therefore, while we often advocate for domain-specific solutions, these should be developed with the Cross-Domain Interoperability Framework and similar cross-disciplinary recommendations in mind. This must be complemented by continuous communication and training activities, to support the transition from "talking about FAIR" to implementing FAIR in daily research practice.

## Data availability statement

The original contributions presented in the study are included in the article, further details of the described data management approaches are available in the NanoCommons User Guidance Handbook, at: https://nanocommons.github.io/user-handbook/data-management/nanosafety-data-concepts/. Further inquiries can be directed to the corresponding authors.

## Author contributions

Conceptualization: TE and IL; methodology: TE, AP, JA, CC, GC, AC, PD, LF, IF, HS, and AA; formal analysis: TE, AP, PD, GM, DM, and IL; resources: TE, AP, GM, JA, CC, GC, AC, PD, LF, IF, HS, and AA; writing—original draft: TE and IL; writing—review and Editing: TE, AP, GM, JA, NB, CC, GC, AC, PD, LF, SF, IF, FK, VL, DM, JR, HS, BS-M, SV, MW, AA, and IL (all co-authors); visualization: TE, LF, GC, IF, SF, and IL; project administration: TE and IL; funding acquisition: IL, AA, TE, and SF. All authors contributed to the article and approved the submitted version.

## Funding

Funding from the European Commission via the Horizon 2020 research infrastructure project NanoCommons (Grant Agreement No. 731032), the Horizon 2020 research and innovation projects NanoSolveIT (Grant Agreement No. 814572), ACEnano (Grant Agreement No. 720952), NanoMECommons (Agreement No. 952869), and ASINA (Grant Agreement No. 862444). Additional support was provided by the Horizon Europe projects WorldFAIR (Grant Agreement No. 101058393) and the Innovate UK support for UoB participation in WorldFAIR (Grant No. 1831977), MACRAMÉ (Grant Agreement No. 101092686) and the Innovate UK support for UoB participation in MACRAMÉ (Grant No. 10066165), and nanoPASS (Grant Agreement No. 101092741). Support from IUPAC (Project No. 2022-001-2-800), CODATA TaskGroup on Extension of InChI for Nanomaterials and VAMAS TWA 34 new project 17 (proposal for NInChI) are also acknowledged.

## Acknowledgments

The authors acknowledge Barry Hardy (Edelweiss Connect) and Eugenia Valsami-Jones (University of Birmingham) from the ACENano project for useful discussions, as well as the following NanoCommons partners and Transnational Access users for discussions that supported the refinement of the concepts presented herein: Andrea Haase (BfR), Claus Svendsen, Lee Walker and Amaia Green Etxabe (UK CEH), Georgios Konstantopoulos, Nikos Nikoloudakis (NTUA), Konstantinos Kotsis (UCD), Martin Himly (PLUS), Beatriz Alfaro and Andreas Falk (BNN), Jan Schüürman and Frank von der Kammer (Uni Vienna).

## Conflict of interest

Author TE was employed by Seven Past Nine GmbH. Authors AP and AA were employed by NovaMechanics Ltd. Authors NB and BS–M were employed by TEMAS Solutions GmbH. Author SF was employed by AcumenIST SRL. Author IF was employed by Transgero Limited. Author DM was employed by Labvantage—Biomax GmbH. Author JR was employed by R&R Data Services.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

# References


1. Wilkinson M, Dumontier M, Aalbersberg I, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* (2016) 3:160018. doi:10.1038/sdata.2016.18

2. Schultes E. The fair Hourglass: A framework for fair implementation. *FAIR Connect* (2023) 1:13–7. doi:10.3233/FC-221514

3. Valsami-Jones E, Lynch I, Charitidis CA. Nanomaterial ontologies for nanosafety: A rose by any other name. *J Nanomed Res* (2016) 3(5):00070. doi:10.15406/jnmr.2016.03.00070

4. Thomas DG, Pappu RV, Baker NA. NanoParticle Ontology for cancer nanotechnology research. *J Biomed Inform* (2011) 44:59–74. doi:10.1016/j.jbi.2010.03.001

5. Hastings J, Jeliazkova N, Owen G, Tsiliki G, Munteanu CR, Steinbeck C, et al. eNanoMapper: harnessing ontologies to enable data integration for nanomaterial risk assessment. *J Biomed Semant* (2015) 6:10. doi:10.1186/s13326-015-0005-5

6. Papadiamantis AG, Klaessig FC, Exner TE, Hofer S, Hofstaetter N, Himly M, et al. Metadata stewardship in nanosafety research: Community-driven organisation of metadata schemas to support FAIR nanoscience data. *Nanomaterials* (2020) 10(10):2033. doi:10.3390/nano10102033

7. EC. *Commission staff working document: Supporting and connecting policymaking in the Member States with scientific rese*. Brussels (2023). Avaliable at: https://knowledge4policy.ec.europa.eu/sites/default/files/SWD_2022_346_final.PDF (Accessed May 28, 2023).

8. Amos JD, Tian Y, Zhang Z, Lowry GV, Wiesner MR, Hendren CO. The NanoInformatics Knowledge Commons: Capturing spatial and temporal nanomaterial transformations in diverse systems. *NanoImpact* (2021) 23:100331. doi:10.1016/j.impact.2021.100331

9. Karcher S, Willighagen EL, Rumble J, Ehrhart F, Evelo CT, Fritts M, et al. Integration among databases and data sets to support productive nanotechnology: Challenges and recommendations. *NanoImpact* (2018) 9:85–101. doi:10.1016/j.impact.2017.11.002

10. Fadeel B, Farcal L, Hardy B, Vázquez-Campos S, Hristozov D, Marcomini A, et al. Advanced tools for the safety assessment of nanomaterials. *Nat Nanotech* (2018) 13:537–43. doi:10.1038/s41565-018-0185-0

11. Afantitis A, Melagraki G, Isigonis P, Tsoumanis A, Varsou DD, Valsami-Jones E, et al. NanoSolveIT Project: Driving nanoinformatics research to develop innovative and integrated tools for *in silico* nanosafety assessment. *Comput Struct Biotechnol J* (2020) 18:583–602. doi:10.1016/j.csbj.2020.02.023

12. Kochev N, Jeliazkova N, Paskaleva V, Tancheva G, Iliev L, Ritchie P, et al. Your spreadsheets can Be fair: A tool and FAIRification workflow for the ENanoMapper database. *Nanomaterials* (2020) 10(10):1908. doi:10.3390/nano10101908

13. Elberskirch L, Binder K, Riefler N, Sofranko A, Liebing J, Minella CB, et al. Digital research data: From analysis of existing standards to a scientific foundation for a modular metadata schema in nanosafety. *Part Fibre Toxicol* (2022) 19:1. doi:10.1186/s12989-021-00442-x

14. Furxhi I. Health and environmental safety of nanomaterials: O data, where art thou? *NanoImpact* (2022) 25:100378. doi:10.1016/j.impact.2021.100378

15. Baer DR, Munusamy P, Thrall BD. Provenance information as a tool for addressing engineered nanoparticle reproducibility challenges. *Biointerphases* (2016) 11:04B401. doi:10.1116/1.4964867

16. Furxhi I, Perucca M, Blosi M, de Ipiña JL, Oliveira J, Murphy F, et al. ASINA project: Towards a methodological data-driven sustainable and safe-by-design approach for the development of nanomaterials. *Front Bioeng Biotechnol* (2021) 9:805096. doi:10.3389/fbioe.2021.805096

17. Dumit VI, Ammar A, Bakker MI, Bañares MA, Bossa C, Costa A, et al. From principles to reality. FAIR implementation in the nanosafety community. *Nano Today* (2023) 51:101923. doi:10.1016/j.nantod.2023.101923

18. van de Sandt S, Dallmeier-Tiessen S, Lavasa A, Petras V. The definition of reuse. *Data Sci J* (2019) 18:22. doi:10.5334/dsj-2019-022

19. Schöch C. Wiederholende forschung in den digitalen geisteswissenschaften (2017). Avaliable at: https://christofs.github.io/wiederholende-forschung-dhd/ (Accessed August 4, 2023).

20. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* (2016) 533:452–4. doi:10.1038/533452a

21. Totaro S, Crutzen H, Riego Sintes J. *Data logging templates for the environmental, health and safety assessment of nanomaterials EUR 28137 EN*. Luxembourg (Luxembourg): Publications Office of the European Union (2017). p. 103178. doi:10.2787/505397

22. OECD. Test guidelines for chemicals (2023). Avaliable at: https://www.oecd.org/env/ehs/testing/oecdguidelinesforthetestingofchemicals.htm (Accessed May 28, 2023).

23. Gottardo S, Ceccone G, Freiberger H, Gibson P, Kellermeier M, Ruggiero E, et al. *GRACIOUS data logging templates for the environmental, health and safety assessment of nanomaterials*. Luxembourg: EUR 29848 EN, Publications Office of the European Union (2019). doi:10.2760/142959

24. Basei G, Rauscher H, Jeliazkova N, Hristozov D. A methodology for the automatic evaluation of data quality and completeness of nanomaterials for risk assessment purposes. *Nanotoxicology* (2022) 16(2):195–216. doi:10.1080/17435390.2022.2065222

25. Jeliazkova N, Longhin E, El Yamani N, Rundén-Pran E, Moschini E, Serchi T, et al. *Template wizard: Co-creation of data collection templates for safety assessment of (nano) materials*. Nature Protocols (2023).

26. Oberdörster G, Maynard A, Donaldson K, Castranova V, Fitzpatrick J, Ausman K, et al. Principles for characterizing the potential human health effects from exposure to nanomaterials: Elements of a screening strategy. *Part Fibre Toxicol* (2005) 2:8. doi:10.1186/1743-8977-2-8

27. Cong Y, Pang C, Dai L, Banta GT, Selck H, Forbes VE. Importance of characterizing nanoparticles before conducting toxicity tests. *Integrated Environ Assess Manage* (2011) 7(3):502–3. doi:10.1002/ieam.204

28. Andraos C, Yu IJ, Gulumian M. Interference: A much-neglected aspect in high-throughput screening of nanoparticles. *Int J Toxicol* (2020) 39(5):397–421. doi:10.1177/1091581820938335

29. Chetwynd AJ, Wheeler KE, Lynch I. Best practice in reporting corona studies: Minimum information about Nanomaterial Biocorona Experiments (MINBE). *Nano Today* (2019) 28:100758. doi:10.1016/j.nantod.2019.06.004

30. Lowry GV, Gregory KB, Apte SC, Lead JR. Transformations of nanomaterials in the environment. *Environ Sci Technol* (2012) 46(13):6893–9. doi:10.1021/es300839e

31. Markiewicz M, Kumirska J, Lynch I, Matzke M, Köser J, Bemowsky S, et al. Changing environments and biomolecule coronas: Consequences and challenges for the design of environmentally acceptable engineered nanoparticles. *Green Chem* (2018) 20:4133–68. doi:10.1039/C8GC01171K

32. Svendsen C, Walker LA, Matzke M, Lahive E, Harrison S, Crossley A, et al. Key principles and operational practices for improved nanotechnology environmental exposure assessment. *Nat Nanotechnol* (2020) 15:731–42. doi:10.1038/s41565-020-0742-1

33. Marchese Robinson RL, Lynch I, Peijnenburg W, Rumble J, Klaessig F, Marquardt C, et al. How should the completeness and quality of curated nanomaterial data be evaluated? *Nanoscale* (2016) 8:9919–43. doi:10.1039/C5NR08944A

34. ECHA. Appendix on nanoforms to the guidance on registration and substance identification (2022). Avaliable at: http://echa.europa.eu/contact (Accessed May 28, 2023).

35. EC. Communication from the commission: The European green deal, COM(2019) 640 final (2019). Avaliable at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52019DC0640&from=EN (Accessed May 28, 2023).

36. Taylor CF, Field D, Sansone S-A, Aerts J, Apweiler R, Ashburner M, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: The MIBBI project. *Nat Biotechnol* (2008) 26:889–96. doi:10.1038/nbt.1411

37. Wilson W. *Workshop on ensuring appropriate material characterization in nano-toxicity studies*. Washington DC: held at the Woodrow Wilson International Center for Scholars (2008). Avaliable at: https://characterizationmatters.wordpress.com/parameters/(Accessed May 28, 2023).

38. Mills KC, Murry D, Guzan KA, Ostraat ML. Nanomaterial registry: Database that captures the minimal information about nanomaterial physico-chemical characteristics. *J Nanopart Res* (2014) 16(2):2219. doi:10.1007/s11051-013-2219-8

39. Rumble J, Freiman S, Clayton T. Towards a Uniform description system for materials on the Nanoscale. *Chem Int* (2015) 37(4):3–7. doi:10.1515/ci-2015-0402

40. CODATA-VAMAS. *Working group on the description of nanomaterials, rumble J. Uniform description system for materials on the Nanoscale* (2016). Avaliable at: https://zenodo.org/record/56720 (Accessed August 4, 2023).

41. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* (1988) 28(1):31–6. doi:10.1021/ci00057a005

42. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. InChI - the worldwide chemical structure identifier standard. *J Cheminform* (2013) 5:7. doi:10.1186/1758-2946-5-7

43. Lynch I, Afantitis A, Exner T, Himly M, Lobaskin V, Doganis P, et al. Can an InChI for nano address the need for a simplified representation of complex nanomaterials across experimental and nanoinformatics studies? *Nanomaterials* (2020) 10:2493. doi:10.3390/nano10122493

44. Murray RA, Escobar A, Bastús NG, Andreozzi P, Puntes V, Moya SE. Fluorescently labelled nanomaterials in nanosafety research: Practical advice to avoid artefacts and trace unbound dye. *NanoImpact* (2018) 9:102–13. doi:10.1016/j.impact.2017.11.001

45. Langevin D, Lozano O, Salvati A, Kestens V, Monopoli M, Raspaud E, et al. Inter-laboratory comparison of nanoparticle size measurements using dynamic light scattering and differential centrifugal sedimentation. *NanoImpact* (2018) 10:97–107. doi:10.1016/j.impact.2017.12.004

46. Wang AY-T, Murdock RJ, Kauwe SK, Oliynyk AO, Gurlo A, Brgoch J, et al. Machine learning for materials scientists: An introductory guide toward best practices. *Chem Mater* (2020) 32(12):4954–65. doi:10.1021/acs.chemmater.0c01907

47. OECD. OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationships models (2004). Avaliable at: https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf (Accessed May 28, 2023).

48. ECHA. *European Chemicals Agency, How to use and report (Q)SARs. Practical guide 5*. European Chemicals Agency (2016).

49. ECHA. Read-across assessment framework (RAAF) (2017). Avaliable at: https://echa.europa.eu/documents/10162/17221/raaf_en.pdf/614e5d61-891d-4154-8a47-87efebd1851a (Accessed May 28, 2023).

50. Triebe J, Worth A, Janusch Roi A, Coe A. *JRC QSAR model database: EURL-ECVAM DataBase service on ALternative methods to animal experimentation: To promote the development and uptake of alternative and advanced methods in toxicology and biomedical sciences: User support &tutorial, EUR 28713 EN*. Luxembourg: Publications Office of the European Union (2017). doi:10.2760/905519

51. OECD. *OECD guideline no. 497: Defined approaches on skin sensitisation*. OECD (2021).

52. CEN. CEN workshop agreement on materials modelling - terminology, classification and metadata (2018). Avaliable at: https://www.cencenelec.eu/media/CEN-CENELEC/CWAs/RI/cwa17284_2018.pdf (Accessed May 28, 2023).

53. Combemale B, Kienzle JA, Mussbacher G, Ali H, Amyot D, Bagherzadeh M, et al. A hitchhiker's guide to model-driven engineering for data-centric systems. *IEEE Softw* (2020) 38:71–84. doi:10.1109/MS.2020.2995125

54. CEN. CEN workshop agreement on materials characterisation - terminology, metadata and classification (2021). Avaliable at: https://www.cencenelec.eu/media/CEN-CENELEC/CWAs/ICT/cwa17815.pdf (Accessed on 28 May, 2023).

55. Sosnowska A, Jagiello K, Peijnenburg W, Grafström R, Dalmaar C, Jensen KA, et al. How the EMMC MODA can be used for physics-based and data-based models for risk assessment? *Toxicol Lett* (2021) 350:S82. doi:10.1016/S0378-4274(21)00438-0

56. Romanos N, Kalogerini M, Koumoulos EP, Morozinis AK, Sebastiani M, Charitidis C. Innovative Data Management in advanced characterization: Implications for materials design. *Mater Today Commun* (2019) 20:100541. doi:10.1016/j.mtcomm.2019.100541

57. Amos JD, Zhang Z, Tian Y, Lowry GV, Wiesner MR, Hendren CO. *Knowledge and instance mapping: Architecture for premeditated interoperability of disparate data for materials* (2023). Submitted.

58. Maier D, Exner TE, Papadiamantis AG, Ammar A, Tsoumanis A, Doganis P, et al. All members of the NanoCommons project. Knowledge for safer materials – "the NanoCommons" knowledge base. *Front Phys* (2023).

59. Furxhi I, Arvanitis A, Murphy F, CostaBlosi AM. Data shepherding in nanotechnology. The Initiation. *Nanomaterials* (2021) 11(6):1520. doi:10.3390/nano11061520

60. Furxhi I, Koivisto AJ, Murphy F, Trabucco S, Del Secco B, Arvanitis A. Data shepherding in nanotechnology. The exposure field campaign template. *Nanomaterials* (2021) 11(7):1818. doi:10.3390/nano11071818

61. Furxhi I, Varesano A, Salman H, Mirzaei M, Battistello V, Tomasoni IT, et al. Data shepherding in nanotechnology: An antimicrobial functionality data capture template. *Coatings* (2021) 11(12):1486. doi:10.3390/coatings11121486

62. Martinez DST, Da Silva GH, de Medeiros AMZ, Khan LU, Papadiamantis AG, Lynch I. Effect of the albumin corona on the toxicity of combined graphene oxide and cadmium to *Daphnia magna* and integration of the datasets into the NanoCommons knowledge base. *Nanomaterials* (2020) 10:1936. doi:10.3390/nano10101936

63. Quevedo AC, Guggenheim E, Briffa SM, Adams J, Lofts S, Kwak M, et al. UV-vis spectroscopic characterization of nanomaterials in aqueous media. *J Vis Exp* (2021) 176:e61764. doi:10.3791/61764

64. Horsch MT, Niethammer C, Boccardo G, Carbone P, Chiacchiera S, Chiricotto M. Semantic interoperability and characterization of data provenance in computational molecular engineering. *J Chem Eng Data* (2020) 65:1313–1329. doi:10.1021/acs.jced.9b00739

65. Charitidis C, Sebastiani M, Goldbeck G. Fostering research and innovation in materials manufacturing for Industry 5.0: The key role of domain intertwining between materials characterization, modelling and data science. *Mater Des* (2022) 223:111229. doi:10.1016/j.matdes.2022.111229

66. Del Nostro P, Goldbeck G, Toti DC. Chameo: An ontology for the harmonisation of materials characterisation methodologies. *Appl Ontology* (2022) 17(301):401–21. doi:10.3233/AO-220271

67. Al-Zubaidi R-Smith N, Gramse G, Kienberger F, Moerman D, Douheret O. CHADA characterisation data and description of a characterisation experiment for impedance spectroscopy for organic photovoltaic samples (1.0). *Zenodo* (2020). doi:10.5281/zenodo.4304043

68. Arofan G, Hodson S. Cross-domain interoperability framework (CDIF) working documents (a WorldFAIR deliverable). *February* (2023) 18. doi:10.5281/zenodo.7652742