



Towards FAIR nanosafety data

Nina Jeliaskova¹✉, Margarita D. Apostolova², Cristina Andreoli³, Flavia Barone³, Andrew Barrick⁴, Chiara Battistelli³, Cecilia Bossa³, Alina Botea-Petcu⁵, Amélie Châtel⁴, Isabella De Angelis³, Maria Dusinska⁶, Naouale El Yamani⁶, Daniela Gheorghe⁵, Anna Giusti⁷, Paloma Gómez-Fernández⁸, Roland Grafström^{9,10}, Maciej Gromelski^{11,12}, Nicklas Raun Jacobsen¹³, Vedrin Jeliaskov¹, Keld Alstrup Jensen¹³, Nikolay Kochev¹⁴, Pekka Kohonen^{9,10}, Nicolas Manier¹⁵, Espen Mariussen⁶, Agnieszka Mech¹⁶, José María Navas¹⁷, Vesselina Paskaleva¹⁴, Aurica Precupas⁵, Tomasz Puzyn^{11,12}, Kirsten Rasmussen¹⁶, Peter Ritchie¹⁸, Isabel Rodríguez Llopis⁸, Elise Rundén-Pran⁶, Romica Sandu⁵, Neeraj Shandilya¹⁹, Speranta Tanasescu⁵, Andrea Haase⁷ and Penny Nymark^{9,10}✉

Nanotechnology is a key enabling technology with billions of euros in global investment from public funding, which include large collaborative projects that have investigated environmental and health safety aspects of nanomaterials, but the reuse of accumulated data is clearly lagging behind. Here we summarize challenges and provide recommendations for the efficient reuse of nanosafety data, in line with the recently established FAIR (findable, accessible, interoperable and reusable) guiding principles. We describe the FAIR-aligned Nanosafety Data Interface, with an aggregated findability, accessibility and interoperability across physicochemical, bio-nano interaction, human toxicity, omics, ecotoxicological and exposure data. Overall, we illustrate a much-needed path towards standards for the optimized use of existing data, which avoids duplication of efforts, and provides a multitude of options to promote safe and sustainable nanotechnology.

Science in general is becoming increasingly data intense, which needs to be met by solutions for data management. Building on and applying existing data is key to safety assessments of new and emerging materials to ensure safe and sustainable technologies^{1,2}. The nanosafety community, which addresses the environmental and health safety aspects of nanomaterials (NMs), lacks a structured means to reuse the current wealth of existing safety data. Difficulties exist with regard to the concrete implementation of emerging technical solutions for data management^{3–6}, which are further increased by the complex and multidisciplinary nature of nanosafety data. Efficient and accurate reuse of data involves labour-intensive quality assessments, curation, interpretation and integration of data and metadata^{3,4,7–10}. Core technical challenges that have been reported include the lack of persistent identifiers (IDs) for NMs, lack of standardized user-friendly data retrieval services and reporting formats (the latter leads to a lack of crucial metadata) and data poorly suited to modelling requirements due to data gaps and uncertainty over data quality^{3,6,11}. Application of the recently established FAIR (findable, accessible, interoperable and reusable) guiding principles¹², designed by the informatics community, is expected to substantially improve the rate of data reuse.

The FAIR principles are aspirational, which allows a high level of flexibility with regard to system design specifications. They emphasize the ability of machines to automatically find and use data, and simultaneously support human readability and reuse. It is important to note that FAIR data are not openly available by default, and nor do the FAIR principles provide an automated means for data quality assessment, curation, interpretation or integration. Nevertheless, the so-called FAIRification of data is essential to enable these activities to be performed efficiently, that is, to identify existing data, whether openly available or not, and hence support accurate data reuse^{4,7,13}. FAIRification is, by definition, a type of data curation, albeit with a generalized focus on abiding by the aspirational FAIR principles rather than aiming for a specific reuse scenario. FAIRification focuses on the development of strategies and approaches to enhance and simplify the processes of making data FAIR. FAIR data are at the heart of several international data strategies, which include the recent 'European Union (EU) strategy for data', which prioritizes standardization activities and works towards a more harmonized description of datasets, data objects and IDs to foster data interoperability between sectors through the FAIR principles¹⁴. In addition, FAIR data are considered key components of

¹Ideaconsult Ltd, Sofia, Bulgaria. ²Medical and Biological Research Laboratory, Roumen Tsanev Institute of Molecular Biology, Bulgarian Academy of Sciences, Sofia, Bulgaria. ³Istituto Superiore di Sanità, Rome, Italy. ⁴Mer Molécules Santé, Université Catholique de l'Ouest, Angers, France. ⁵Institute of Physical Chemistry 'Ilie Murgulescu' of the Romanian Academy, Bucharest, Romania. ⁶Department of Environmental Chemistry, Health Effects Laboratory, Norwegian Institute for Air Research, Kjeller, Norway. ⁷Department of Chemical and Product Safety, German Federal Institute for Risk Assessment, Berlin, Germany. ⁸GAIKER Technology Centre, Basque Research and Technology Alliance, Zamudio, Spain. ⁹Department of Toxicology, Misvik Biology, Turku, Finland. ¹⁰Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. ¹¹Group of Environmental Chemometrics, Faculty of Chemistry, University of Gdansk, Gdańsk, Poland. ¹²QSAR Lab Ltd, Gdańsk, Poland. ¹³National Research Centre for the Working Environment, Copenhagen, Denmark. ¹⁴Faculty of Chemistry, Department of Analytical Chemistry and Computer Chemistry, University of Plovdiv, Plovdiv, Bulgaria. ¹⁵Expertise and Assays in Ecotoxicology Unit, French National Institute for Industrial Environment and Risks, Verneuil-en-Halatte, France. ¹⁶Joint Research Centre, European Commission, Ispra, Italy. ¹⁷Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, Madrid, Spain. ¹⁸Institute of Occupational Medicine, Edinburgh, UK. ¹⁹Netherlands Organisation for Applied Scientific Research (TNO), Utrecht, Netherlands. ✉e-mail: jeliaskova.nina@gmail.com; penny.nymark@ki.se

the new EU Industrial Strategy, Chemicals Strategy for Sustainability and Circular Economy Action Plan driven by the EU Green Deal approach, which aims to support substitution and elimination of hazardous substances based on Safe and Sustainable by Design approaches and thus enable safe and sustainable innovation^{1,15–17}.

The eNanoMapper database was established in response to the goals described above to collect data and provision reuse needs of the nanosafety community¹⁸. eNanoMapper is based on a data model originally developed to represent industrial chemicals and implements nanosafety-community-specific solutions aligned with all the FAIR principles, which include community-developed standards, such as a domain-specific ontology, and generated persistent IDs based on Universally Unique Identifiers^{19,20}. In addition, community-accepted user-friendly IDs are also implemented (for example, NM IDs assigned by the European Commission's Joint Research Centre (JRC) Nanomaterials Repository^{21,22}). The data model provides a high potential for interoperability through data serialization into different formats and links to processes and systems for data submission, curation, indexing, federated search, retrieval and analysis workflows^{18,23}. Since 2016, the eNanoMapper data model has been adopted by several European projects, each establishing project-specific eNanoMapper database instances. To provide aggregated findability, accessibility and interoperability across all database instances, the Nanosafety Data Interface was created, and currently represents one of the richest nanosafety data retrieval services that comprises and links to data globally³.

This article captures the know-how from gathering, populating and reusing data in the eNanoMapper database from a variety of EU-funded projects, with a particular focus on the FAIRification of data in the EU Horizon 2020 project NanoReg2 (<http://www.nanoreg2.eu/>). In the project, we adopt a data model with the goal to develop scientifically sound, data-driven methods that enabled the grouping of NMs in the support of safety decisions taken by industry and regulators²⁴. A wide variety of data types were included, based on project-specific criteria (details available in Giusti et al.⁸), which ranged from physicochemical, (eco)toxicological and exposure-related parameters in line with current regulatory requirements for the safety assessment of NMs, to information derived from non-standardized New Approach Methodologies such as omics data and data that describes thermodynamic parameters at the bio–nano interface^{2,8,25}. A FAIR principles-aligned overview of the challenges and lessons learned is provided, along with concrete recommendations for advancing FAIRification and thereby the efficient reuse of nanosafety data, building on recommendations described in the EU–US Nanoinformatics Roadmap 2030⁶. A semi-quantitative evaluation of the data FAIRness in the NanoReg2 database is also provided and compared with a FAIR-compliant omics repository²⁶. The results demonstrate a high level of FAIRness brought to all data types by the eNanoMapper solutions, which include a significantly enhanced FAIRness of nanosafety-relevant omics data. An overview of key terms is provided in Box 1. In addition, the data used to showcase the FAIRification is described in detail in the Supplementary Methods to provide concrete practical examples for those aiming to implement the solutions applied.

Results

The Nanosafety Data Interface. The established Nanosafety Data Interface provides flexible findability of data gathered and/or generated by a wide variety of projects, which include a majority of European nanosafety projects (<https://www.nanosafetycluster.eu/>) and the US cancer Nanotechnology Laboratory portal (caNanoLab; <https://cananolab.nci.nih.gov>) (Supplementary Fig. 1). The Interface was created by compiling, annotating and importing data from multiple nanosafety projects into separate eNanoMapper database instances (described in detail in Supplementary Methods 2.1). The database instances offer a user-friendly web interface and an

Box 1 | Key terms

FAIR

Findable, accessible, interoperable, reusable (<https://www.force11.org/fairprinciples>).

eNanoMapper data model

A set of data elements and relationships that represents chemical substances with a complex composition and associated experimental data, extended to NMs by the EU-funded project eNanoMapper (<http://www.enanomapper.net/>).

eNanoMapper database instance

An installation of AMBIT software (<http://ambit.sf.net>), which implements the eNanoMapper data model. Usually consists of data from individual nanosafety projects and may be protected by role-based access to keep data confidential.

NanoReg2 database

A set of eNanoMapper database instances and an aggregated search index used for data management within the EU-funded project NanoReg2 (openly available at <https://search.data.enanomapper.net/projects/nanoreg2>).

Omics database instance

A separate eNanoMapper database instance used to gather metadata that links to nanosafety-relevant omics data (openly available at <https://search.data.enanomapper.net/about/omics/> and interoperable with the NanoReg2 database).

Nanosafety Data Interface

Online user interface enabling user-friendly access to the aggregated search index of the (sub)set of eNanoMapper database instances. Usually, the user interface is project specific and protected but can also be publicly available. (Multiple project specific interfaces are at <https://search.data.enanomapper.net/>.)

application programming interface (API). They serve as building blocks to feed the aggregated search and provide interoperability across all or subsets of the instances. A nanosafety community-developed ontology ensured harmonized terminology and served as a synonym list for query expansion²¹. Free text and faceted search applications are available. Multiple data-import formats (IUCLID/OECD HT, W3C-RDF, JSON and MS Excel) and data-export formats (tab-delimited, MS Excel, JSON, RDF and ISA-JSON) are supported (further details in Box 2 and Methods). The import of spreadsheet files was enabled by a configurable parser that maps the spreadsheet data via external configuration files^{18,20,27,28}. It is worth noting that, although the current study focused on FAIRification of data from EU projects and a US-based data portal, the Interface is highly flexible towards international context and harmonization¹⁸.

The NanoReg2 integrated database is available at <https://search.data.enanomapper.net/nanoreg2> and provides authorized and aggregated access to data generated within the project itself and gathered from past projects, which include NANoREG²⁹, ENPRA³⁰, MARINA³¹, NANOGENOTOX³², NanoTEST^{33,34}, NECID^{35,36} and the caNanoLab portal (Fig. 1 and Supplementary Fig. 2). In addition, the database provides access to a publicly available database instance that covers metadata and links to nanosafety-relevant omics data (see the next section and Fig. 1a). The NanoReg2-generated data is publicly released with open access under a CC-BY-NC-SA 4.0 license. (Further details are available in Methods and Supplementary Methods 2.1.)

Box 2 | List of technical abbreviations and terms**Apache Solr**

Open source Java-based platform that supports full-text search, hit highlighting, faceted search, real-time indexing, dynamic clustering, database integration, NoSQL features and rich document handling, <https://lucene.apache.org/solr/>.

REST-API

Representational state transfer-based API.

CC BY-NC-SA 4.0

Creative commons license, Attribution-NonCommercial-ShareAlike 4.0 International, <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

ISA-Tab

Investigation/Study/Assay (ISA) tab-delimited (Tab) format that supports the communication of complex metadata⁵⁹.

JRC ISA-Tab logic

Joint Research Centre-developed openly available Excel-based data logging templates, which apply ISA-Tab (see above) principles⁴³.

IOM data template

Institute of Occupational Medicine-developed data templates widely used within EU projects⁶⁰.

AMBIT

A chemical substances database offering properties prediction and machine learning with REST-API, <http://ambit.sourceforge.net/>.

IUCLID/OECD HT

International Uniform Chemical Information Database, a software for recording, storing, maintaining and exchanging data on intrinsic and hazard properties of chemical substances. IUCLID implements OECD Harmonized Templates, and was co-developed by the European Chemicals Agency and the OECD, <https://iuclid6.echa.europa.eu/>.

RDF

Resource Description Framework is a family of World Wide Web Consortium specifications originally designed as a metadata data model, <https://www.w3.org/RDF>.

W3C

World Wide Web Consortium, international community that develops open standards to ensure the long-term growth of the Web, <https://www.w3.org>.

JSON

JavaScript Object Notation, lightweight data-interchange format, which is easy for humans to read and write.

ISA-JSON

Implementation of the JSON format with an ISA structure, <https://isa-specs.readthedocs.io/en/latest/isajson.html>.

SQL

Structured Query Language, domain-specific language used in programming and designed for managing data held in a relational database management system, or for stream processing in a relational data-stream management system

but not within the nanosafety community due to the poor inclusion of nanosafety-relevant metadata in existing omics repositories, which limits the discovery and reuse of omics data within the field. Expert-based searches and manually created metadata maps were used to semantically implement links to omics data within eNanoMapper (Methods), and an openly available, automatically generated template to gather nanosafety-relevant metadata for omics data was published (<https://search.data.enanmapper.net/projects/nanoreg2/datatemplates/literature/>). The semantic integration allows the nanosafety community to find, access and interoperate relevant omics data with other types of nanosafety data. The raw omics data continue to reside in the original omics repositories, such as Gene Expression Omnibus (GEO)³⁷, ArrayExpress³⁸ and project-specific databases, such as NANOSOLUTIONS (<http://nanosolutionsfp7.com/>)³⁹, and are accessible through active links to the original data repositories. The omics database instance is currently composed of 165 datasets that cover 26 NM types and other materials (for example, asbestos) and provides a basis for interoperable analyses with complementary data in the NanoReg2 database (Fig. 2).

Reusable nanosafety data. The database enables the reuse of nanosafety data for the various purposes outlined in the introduction, and ultimately support in silico modelling, the design of new experiments and safety assessment in line with Safe by Design strategies². Summaries of the FAIR data available in the NanoReg2 database and ready for reuse to diverse levels are presented in Fig. 3.

The FAIRness of the NanoReg2 database was assessed by a semi-quantitative workflow previously applied to omics data repositories²⁶, and yielded a score of 13.5/14 (Table 1). In comparison, the FAIRness of omics data in the GEO repository reached a score of 11/14, as assessed by Berrios et al.²⁶. The FAIRness of omics data in GEO was also explicitly evaluated in relation to reuse for nanosafety and achieved a score of 6/14.

Challenges and lessons learned on the reuse of nanosafety data.

In the current study, the process of FAIRifying nanosafety data led to the identification of 13 challenges, as summarized below in line with the FAIR principles and presented along with lessons learned.

Findability of relevant data was limited by four main challenges: lack of (1) persistent IDs, (2) rich metadata, (3) data IDs and/or ontology included in the metadata and (4) an indexed searchable database. The lack of persistent IDs was previously identified as a bottleneck for discovering nanosafety data and largely attributed to the lack of standardized machine-readable representation of the structural complexity of NMs^{3,6}. Currently, the identification of data derived from the same material relies on a variety of criteria to estimate the sameness of materials, often based on physicochemical characterization and expert-based judgement⁴⁰. The urgent need for persistent IDs was obvious within NanoReg2. The JRC IDs substantially supported the identification of compatible datasets relevant for integration, and enabled the finding of widely diverse data for representative JRC NMs (Supplementary Table 1), which was difficult for NMs without such IDs⁸. The lack of rich metadata led to the need for expert knowledge and enormous curation efforts to identify sources of data, and to gather relevant metadata (which included protocols used to obtain the data) and linked data³ (for example, associated physicochemical characterization data). A detailed review of linked publications (when available) and the manual curation of pertinent metadata were often necessary. However, a general lack of (persistent) IDs and ontology in the metadata also limited the curation efforts, and we quickly realized the value of a harmonized terminology⁴¹ and implementation of the nanosafety community-based ontology (that is, the eNanoMapper ontology)²¹. Predefined structured search interfaces at variable levels of detail provided a user-friendly access to the eNanoMapper data model,

FAIRification of omics data for nanosafety. A case study was performed using omics data to demonstrate FAIRification of a type of data that today is considered highly FAIR on a general level²⁶,

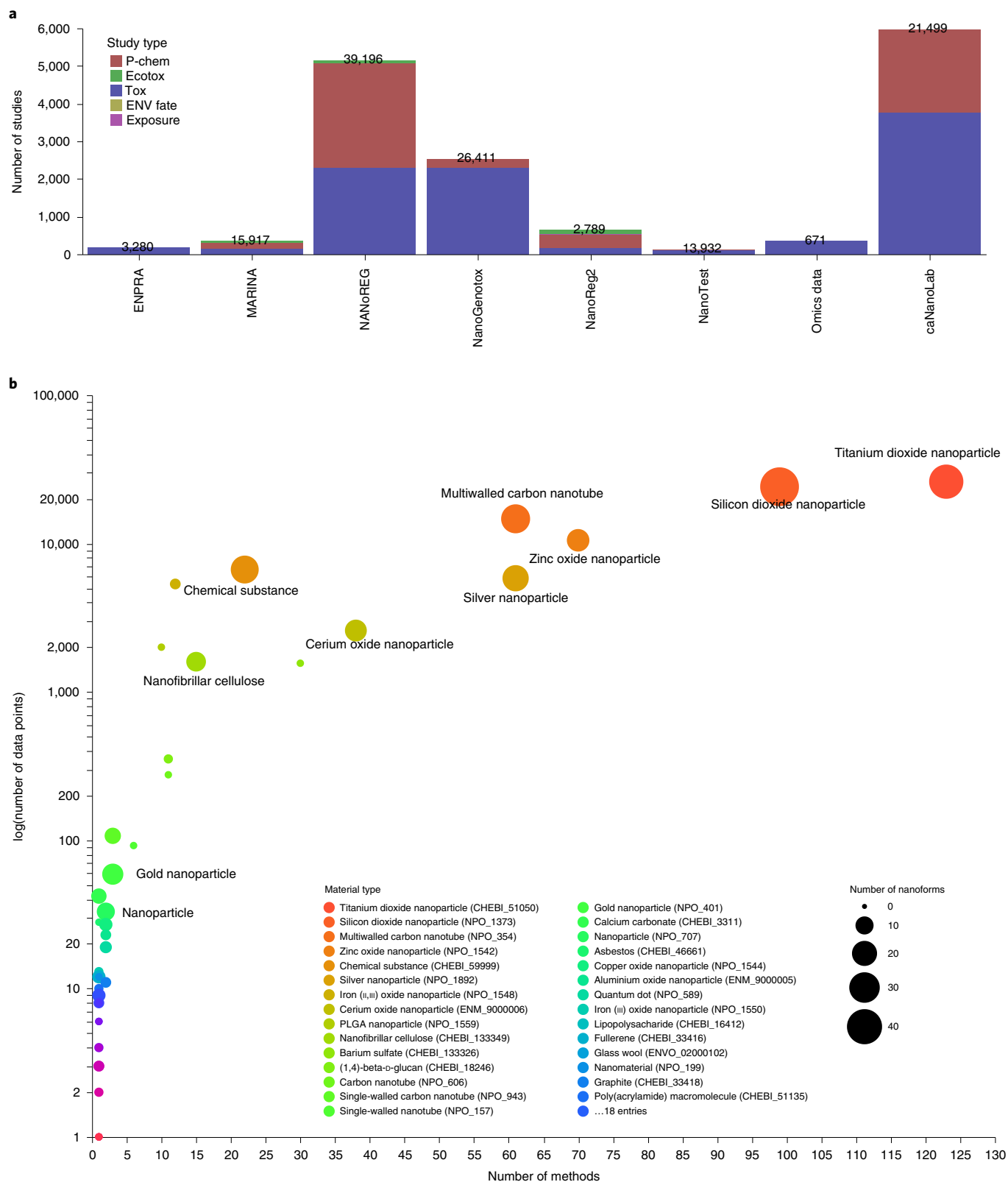


Fig. 1 | Accessible nanosafety data. **a**, Overview of the number of studies that originate from NanoReg2 itself (column annotated 'NanoReg2') and from previous projects, which are now available in the NanoReg2 integrated database. Note that the exposure data from NECID is part of the NanoReg2 column. The number of data points from each project is shown at the top of each column. **b**, Overview of the NanoReg2 database contents in terms of the data available for specific NM categories. Categorization of the NMs is visualized in line with the current annotation of the data (based on the eNanoMapper ontology) when entered into the database. Thus, the annotations are dependent on the current level of curation. For example, the generic term 'chemical substance' covers a wide variety of non-nano substances that are often used as control substances in nanosafety studies, whereas 'nanomaterials' refers to a mixture of NMs in relation to a set of exposure studies included in the database, and 'nanoparticle' covers a set of studies on NMs that are not represented by ontology terms. In addition, apparent overlaps in the annotations may occur for which further curation needs are obvious (NanoReg2 database accessed in December 2020). PLGA, poly(lactic-co-glycolic acid). P-chem, physicochemical; ecotox, ecotoxicological; tox, toxicological; ENV fate, environmental fate.

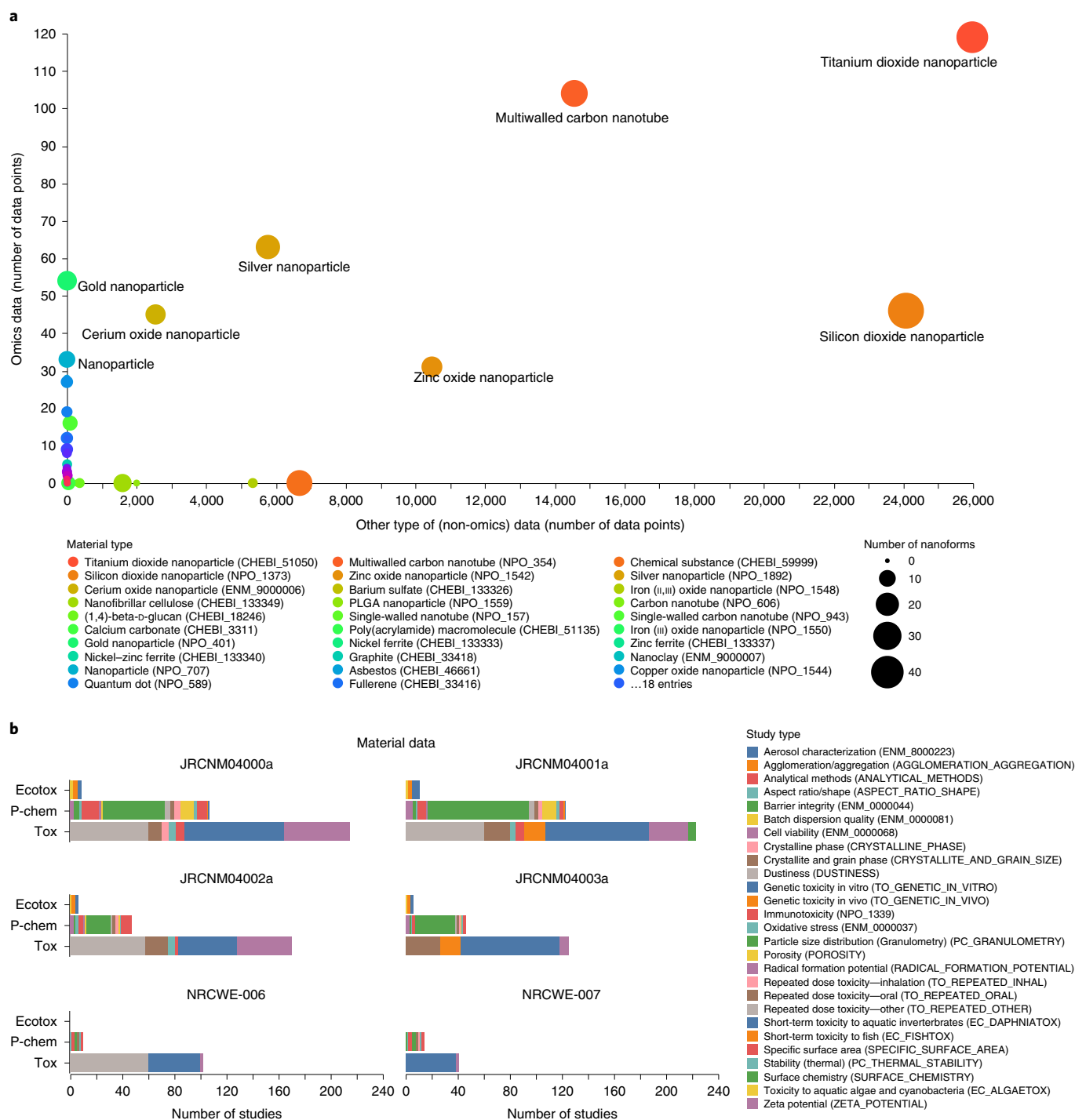


Fig. 2 | Interoperable nanosafety data. **a**, Overview of omics data in relation to, and interoperable with, other types of data available in the NanoReg2 database. (ontology IDs are indicated in brackets for each material type. The terms ‘chemical substance’ and ‘nanoparticle’ refer to non-nano substances included as controls in nanosafety studies and NMs currently not represented by a suitable ontology, respectively.) **b**, Example overview of the available interoperable data for six types of multiwalled carbon nanotubes from the JRC Nanomaterials Repository and the Danish National Research Centre for the Working Environment within the NanoReg2 database. (ontology IDs are indicated in brackets for each study type. The term ‘analytical methods’ refers to methods for the assessment of chemical composition in line with the Organisation for Economic Co-operation and Development (OECD) harmonized template, OHT 87). For further details on the categorization of the methods, please refer to the Supplementary Methods 2.1.

and annotation of the data entities with (multiple) ontology terms substantially supported the efforts, which allowed queries by both free text and ontology terms. Overall, findability issues encountered at the very start of the project were overcome by the improvement and use of the indexed, searchable NanoReg2 database, which saved

time and effort, and reduced uncertainty as to whether all the possible data sources had been explored (Table 1).

Accessibility was limited by two challenges: (5) difficulties in retrieving data and (6) the lack of awareness of data from previous projects. Even when existing data were stored in project-specific

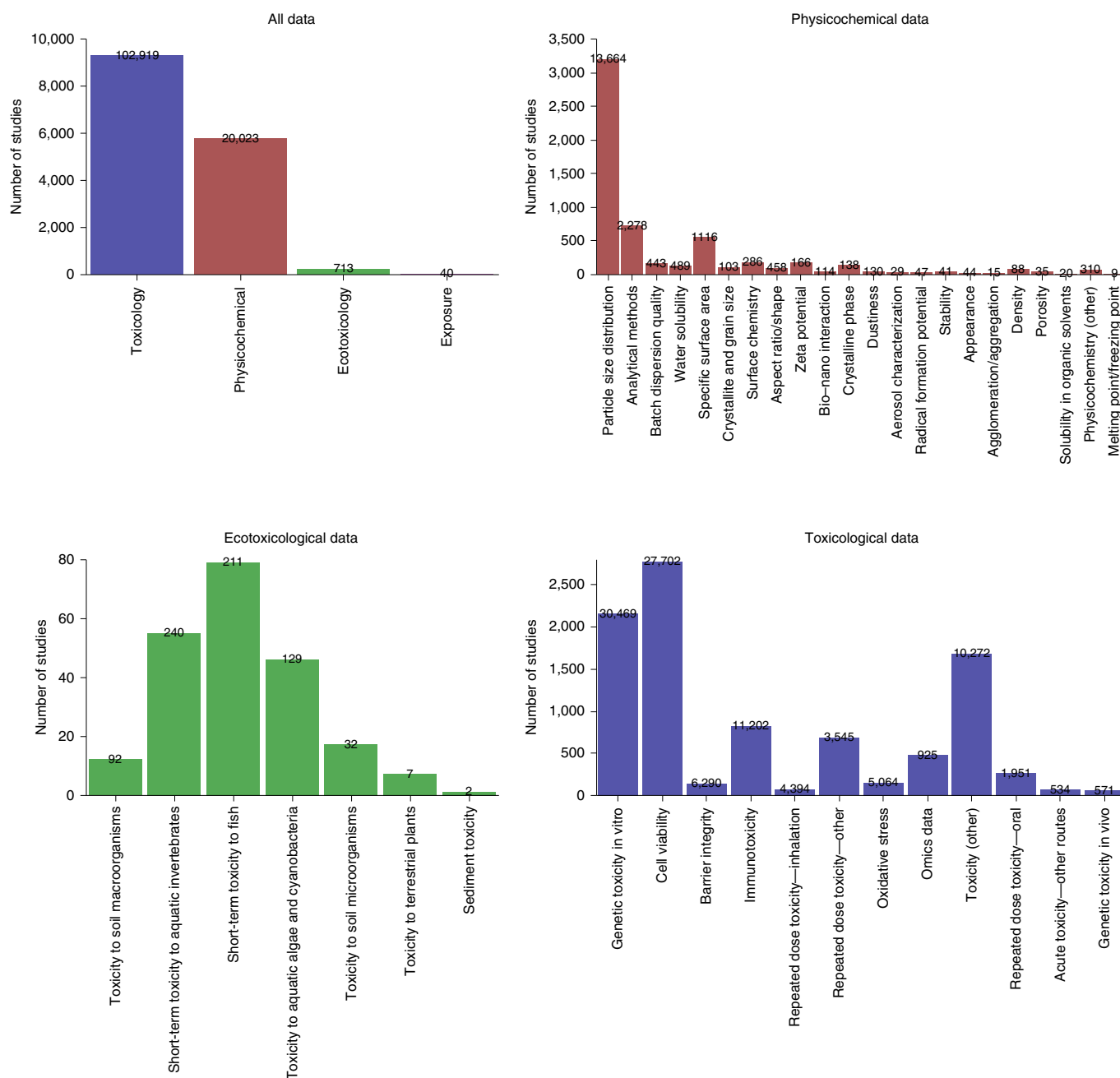


Fig. 3 | Reusable nanosafety data. Examples of data summaries from the NanoReg2 database, which provide an overview of the data and study types currently included in the database and available for reuse in various approaches, which include quality assessment, further fit-for-purpose curation, interpretation and integration for the purpose of various risk-assessment practices. The number of data points for each type of data is indicated above the columns in the first panel. Further interactive summaries are available through the Nanosafety Data Interface (<https://search.data.enanomapper.net/projects/nanoreg2/dashboard/>, database accessed December 2020).

repositories and announced to be openly available, advice from experts was required to establish whom to contact and where to access the data. Consequently, the lack of accessible metadata led to unawareness of the whole dataset. The high value of making project (meta)data easily accessible through the Nanosafety Data Interface was obvious in NanoReg2 (Table 1). The database instances allowed for both published and unpublished data to be accessible to different levels, from the level of raw data to metadata only, depending on the user rights. Subsequently, unpublished data could also be considered for reuse, as openly accessible metadata raise the awareness of existing compatible data and allow researchers to communicate and collaborate.

Interoperability was challenged by three shortages: (7) harmonized terminology, (8) harmonized reporting formats/criteria and (9) links and/or tools for external information retrieval. Similar to findability, the lack of interoperable data representation and harmonized terminology seriously limited interoperability, which is dependent on machine-readable semantic annotations, such as ontologies. Even data from the same discipline may be annotated according to individual researchers' preferences and needs, and the initial purpose of the data often defines the level of annotation⁴². The community-developed ontology²¹ allows for the identification of compatible data types. Difficulties in using the available reporting formats⁴³ and lack of harmonized reporting criteria also

Table 1 | FAIRness of nanosafety data

FAIR principles	NanoReg2 database ^a	Omics repository (GEO) ^b	Omics repository (GEO) for nanosafety	Comment
Findable				
F1. (Meta)data are assigned a globally unique and persistent ID	●	●	●	All datasets are assigned unique and persistent IDs in both the NanoReg2 database (Universally Unique Identifiers) and in GEO (GSExxx)
F2. Data are described with rich metadata (defined by R1 below)	●	●	○	Metadata specific to NMs are added through the eNanoMapper semantic model and ontology by means of linking and the semantic integration of diverse data types; domain-specific metadata are missing in GEO
F3. Metadata clearly and explicitly include the IDs of the data they describe	●	●	○	Metadata available in the NanoReg2 database always include data IDs, such as the omics data ID (GSExxx), which is sometimes missing in publications; GEO does not provide links to the Nanosafety Data Interface
F4. (Meta)data are registered or indexed in a searchable resource	●	●	○	The NanoReg2 database allows for indexed and searchable metadata, based on a domain-specific vocabulary (that is, the eNanoMapper ontology); GEO does not register nor index domain-specific vocabularies
Accessible				
A1. (Meta)data are retrievable by their ID using a standardized communications protocol	●	●	●	The NanoReg2 database and GEO use standardized communication protocols to access data
A1.1 The protocol is open, free and universally implementable	●	●	●	The NanoReg2 database and GEO use open source communication protocols to access data
A1.2 The protocol allows for an authentication and authorization procedure where necessary	●	●	●	The NanoReg2 database and GEO allow for authentication and authorization procedures where necessary
A2. Metadata are accessible, even when the data are no longer available	●	●	●	The NanoReg2 database and GEO retain metadata even if data files are removed
Interoperable				
I1. (Meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation	●	○	○	The NanoReg2 database implements a domain-specific nanosafety ontology based on the knowledge representation language OWL; GEO does not implement formal semantics ^b
I2. (Meta)data use vocabularies that follow FAIR principles	●	○	○	The NanoReg2 database makes use of an open and community-developed semantic vocabulary (that is, the eNanoMapper ontology); GEO does not implement domain-specific vocabularies ^b
I3. (Meta)data include qualified references to other (meta)data	○	○	○	The NanoReg2 database includes references and links to other types of nanosafety (meta)data, with specified relations through ontology properties; GEO does not offer semantic links to other systems/data ^b
Reusable				
R1. Meta(data) are richly described with a plurality of accurate and relevant attributes	●	●	○	The NanoReg2 database allows for a rich description of nanosafety data; GEO allows for a rich description but does not provide the means to include links to the Nanosafety Data Interface
R1.1. (Meta)data are released with a clear and accessible data-usage license	●	●	○	The NanoReg2 database will include clear licensing for nanosafety data (for example, in line with the NANoREG data ^c); GEO is governmental and licensing is prohibited ^b
R1.2. (Meta)data are associated with detailed provenance	●	●	○	The NanoReg2 database allows for the inclusion of detailed provenance (for example, DOI, year, data owner and links to SOPs); GEO includes detailed provenance linked to the omics data itself, but lacks the nanosafety-specific details
R1.3. (Meta)data meet domain-relevant community standards	●	●	○	The NanoReg2 database meets nanosafety domain-relevant community standards; GEO meets omics community standards, but lacks nanosafety standards
Overall FAIRness score	13.5	11	6	
Comparison of the level of FAIRness of the NanoReg2 database, the omics repository GEO and GEO for use in nanosafety analyses. The table is inspired by Berrios et al. ²⁶ . ●, pass; ○, partial; ○, fail. ^a A visual guide to FAIR Principles (with eNanoMapper, https://search.data.enanmapper.net/fair/) ^b Berrios et al. ²⁶ . ^c NANoREG data shared under CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/).				

hampered interoperability. Criteria for inclusion of specific meta-data through the use of commonly agreed templates greatly enhanced the interoperability of diverse, but compatible, data types (Table 1). Finally, the lack of links and tools for external information retrieval was noted. Importantly, we emphasize that it is neither feasible nor desirable to store all the nanosafety data in one database, as there will always be multiple, fit-for-purpose repositories, such as those for omics or exposure data^{37,38}. Thus, a nanosafety-focused database should be interoperable with other databases to allow for the identification and retrieval of related datasets. Such flexibility towards interoperability with external data is one of the greatest challenges for all data management efforts, exemplified by the challenge in NanoReg2 to reuse omics data for modelling and grouping purposes⁸. Interoperability of omics data with other nanosafety data was seriously challenged by the lack of nanosafety-specific metadata (see above and Supplementary Methods 2.1.4).

Overall, the reuse of data was limited by four main challenges: (10) poorly described (meta)data that limited data gap analysis, (11) variation in the levels of data processing (raw, preprocessed or final result-based format), (12) lack of tools and/or workflows for interpretation and (13) lack of clear and accessible data-usage licenses. As the reuse of data depends on all the other components of the FAIR principles, poorly described (meta)data naturally also relate to the reuse challenges. Owing to the lack of harmonized reporting formats, which leads to a varying inclusion of metadata, and the lack of links to standard operating procedures (SOPs), time-consuming, manual quality assessment and curation efforts were needed to generate data gap maps. Tracing a few crucial metadata attributes significantly supported the efforts, and included links to SOPs, links to data generated within the same experiment and harmonized assay-specific metrics (for example, standardized dose metrics). Various levels of data processing limited the data interpretation and integration and required manual expert-based curation. In general, raw or preprocessed data was preferred over final result-based data. However, a lack of approaches to interpret and integrate data limited the reuse of raw data. Finally, clear and accessible data-usage licenses were often lacking and the benefits of Creative Commons licenses, for example, CC BY-NC-SA 4.0 for the NANoREG data⁴⁴, was recognized (Table 1).

Discussion

Since the publication of the FAIR principles in 2016¹², awareness and implementation efforts have flourished, which demonstrates the immense need for harmonized data management strategies across every field of research (<https://www.go-fair.org/go-fair-initiative/>)^{5,45,46}. In the EU, the FAIR principles have been recognized at policy level and form the basis of the European Open Science Cloud initiative, which has become one of the pillars of the EU Data Strategy within the European Green Deal⁴⁷. The large amount of data and knowledge generated by the nanosafety community over the past decades has great potential, together with the applied FAIR principles, to serve as a basis for the efficient governance and regulation of NMs and other new advanced materials. FAIR principles are key to fulfilling data requirements in relation to the safety assessment of NMs by regulatory bodies that operate in different nanotechnology sectors. Here, we provide an overview of the challenges and lessons learned during the past decade, which leads to the presentation of a set of sound recommendations to advance the reusability of nanosafety data (Table 2).

The first set of recommendations is directly coupled to the previously identified challenges, that is, lack of persistent IDs and (use of) harmonized reporting formats and data retrieval systems, and all relate to findability and accessibility. The implementation of persistent IDs for all the materials used, as well as for the metadata, which includes protocols, methods and model systems, is needed (relating to principles F1 and F2). The newly formed European

Registry of Materials (ERM) is a step towards providing materials with IDs and is currently implemented in ongoing EU-funded projects⁴⁸. Furthermore, a proper description of the data generated, that is, semantic characterization of raw (source) data, metadata and linked data is particularly central for findability. Agreed ontologies enable 'true semantic' findability and permit a user to find all the data tagged with a specific concept and/or query²⁶. The eNanoMapper ontology responds to the need for such terminology and supports the automated generation of 'rich metadata' (relating to F2 and F3). However, currently, data submitters have a relatively low motivation to provide 'rich metadata', which has been attributed to shared data having reduced citation rates, partly due to the poor availability of persistent IDs²⁶. The FAIR principles support the enhanced citation of shared data through better data management, the implementation of persistent IDs and a broader awareness of the value of reusing data, supported by funding agencies, governments and publishers⁶. The recommendations that relate to these challenges include the active implementation of persistent IDs for materials, which increases awareness and incentives for data generators to provide (rich) metadata, and provide tools for an efficient semantic characterization of the (meta)data. Guidance on reporting criteria to reach high-quality assay documentation has been published^{12,49} and can be implemented through federated system functions and metadata schemas available to all according to the FAIR principles^{18,26}. Finally, the lack of indexed searchable data-retrieval systems was addressed directly by the eNanoMapper database semantic model, which provided tools to search, access and retrieve data without the need for prior knowledge of projects (relating to F4 and A1). The Nanosafety Data Interface also offers standard authorization and authentication procedures to protect unpublished data, but still provides access to relevant stakeholders (relating to A1.2).

The next set of recommendations align with the identified poor suitability of available data for modelling purposes due to data gaps and variable data formats, and couple to interoperability. In this respect, the omics and bioinformatics community has served as a good example for nanosafety^{4,13}. The high-dimensional nature of omics data encouraged the community to quickly respond to the need for standardized file formats (both for raw and meta-data) and repositories²⁶. Publishers also established standardized reporting formats and database upload requirements for manuscript submissions, which gives data generators the opportunity and incentives (and responsibility) to provide data in standard community-compatible formats²⁶. Thus, recently omics repositories were shown to be FAIR on a general level^{26,50}, complying in particular with the F, A and R principles. However, challenges for interoperability within specific communities, such as the nanosafety community, remain due to the lack of domain-specific metadata, ontology and community-defined reporting standards²⁶. These 'unFAIR' nanosafety omics data prevent the community from finding and integrating relevant omics data without tedious, expert-based search efforts in the otherwise FAIR omics repositories. This unique situation enabled us to showcase how the FAIRness of data can be significantly enhanced through the implementation of metadata specifications and ontology. Thus, using significant resources and domain expertise, metadata linked to omics data relevant to the nanosafety community was gathered and curated into the eNanoMapper data model. The data were shown to reach an enhanced level of FAIRness with a score of 13.5/14 compared to 11/14 for omics data in general, and 6/14 for nanosafety-relevant omics data directly retrieved from GEO (Table 1). In particular, an increased level of interoperability was provided by proper field-specific semantic characterization using the community-developed ontology (I1 and I2)²¹.

The last set of recommendations relates to interoperability and, in particular, reusability and builds on previous recommendations given in relation to NM grouping strategies⁸. Two of the main challenges related to NM grouping were repeatedly associated with

Table 2 | Challenges, lessons learned and future recommendations for FAIRification of nanosafety data

Challenges	Associated FAIR principle	Lessons learned	Recommendations
1. Lack of persistent IDs	F1	Persistent IDs, such as the JRC IDs, support the identification of various types of data for the same NM from different sources	The implementation of persistent IDs at the start of projects for all the materials and/or substances used and their batches is recommended, for example, in line with the current efforts to establish an ERM ⁴⁸ ; persistent IDs should also be implemented for metadata (for example, protocols, methods, model systems and so on) ²⁸
2. Lack of rich metadata limits discovery	F2	Enormous curation efforts are needed to gather and include metadata and linked data	Incentives for data generators to include rich metadata are needed; for example, increased citation rates when data are reused through the use of persistent IDs, which efficiently and consistently link to the original publication of the data
3. Highly variable vocabularies to describe metadata requires expert knowledge	F3	Efficient use and implementation of ontologies (for example, the eNanoMapper ontology) greatly supports findability and decreases the need for expert knowledge	Efficient semantic characterization using the available ontologies for both raw and linked (meta)data should be implemented, for example, by harmonized data-entry templates and electronic lab notebooks
4. Lack of indexed searchable database	F4	Gathering data and linking to other data and/or knowledge sources in the Nanosafety Data Interface strongly supports findability	Harmonized gathering and linking of (meta) data should be encouraged, for example, through spreading the awareness of the value of reusing data to individual researchers and research consortia, as well as to funding agencies, governments and publishers
5. Data from previous projects are not easily retrievable, and requires expert knowledge, for example, through contact with individual researchers	A1	(Meta)data from previous projects are made easily accessible through the Nanosafety Data Interface	The research community, and especially project consortia, should recognize the value of federated system functions and make metadata schemas available to all according to the FAIR principles
6. Unpublished (meta)data are often completely restricted and can go unnoticed and unused due to lack of awareness	A1.2	The Nanosafety Data Interface allows for the authentication and authorization to access unpublished data; inclusion of (openly available) metadata for the unpublished data enhances awareness of the data and supports future reuse	
7. Non-harmonized vocabularies limit interoperability	I1, I2	Implementation of the eNanoMapper ontology greatly enhances interoperability by enabling the identification of compatible data types	Efficient semantic characterization using the available ontologies for both raw and linked (meta)data should always be implemented, for example, by using electronic lab notebooks
8. Difficulties in using available reporting formats and the lack of harmonized reporting criteria hamper interoperability	I3	Harmonized reporting of metadata raises the interoperability of a dataset and strongly supports the integration of diverse data types	Further work on the implementation of user-friendly harmonized reporting formats and criteria for diverse types of data is needed
9. Interoperability is limited by a lack of links to and tools for retrieving external information	I3	Links to information, knowledge and/or data from other databases benefits interoperability, for example, omics databases	Efforts to advance harmonized reporting formats and criteria should consider the inclusion of reference and links to external information of use, which include, for example, the Adverse Outcome Pathway Knowledgebase ⁵⁶ , life science ⁵⁷ and chemical toxicity databases ⁵⁸
10. Poorly described (meta)data limit the identification of data gaps	R1	Generating data gap maps is manual and time consuming; rich metadata support the assessment of data gaps	The development of automatically generated data gap maps, which include the assessment of completeness (and quality), supported by richly described (meta)data, supports the efficient new generation of data to fill gaps
11. Various levels of data processing limit interpretation	R1	Interpretation of data in the Nanosafety Data Interface requires tedious manual curation work	Harmonized reporting formats should include flexibility towards the inclusion of various levels of processing; in general, raw data are always preferred (although this requires tools and/or guidance for interpretation—see recommendation 12)
12. The lack of analysis tools and workflows that interpret, convert and/or compact data into suitable domain-relevant formats limits reuse	R1, R1.3	Links and interoperability with tools and workflow recommendations for data analysis supports reuse	Harmonized reporting formats should include reference and/or links to available analysis tools and workflows for the interpretation of data
13. The lack of a clear and accessible data-usage license limits reuse	R1.1	A clear data-usage license linked to the NANoREG data has enabled efficient reuse for a wide variety of purposes	All data should be linked to clear data-usage licenses under, for example, the Creative Commons framework

the limited availability and serious concerns regarding quality of data. A recommended solution included future publicly accessible databases that host both raw data and metadata that support data interpretation and reanalysis, directly in line with the EU–US Nanoinformatics Roadmap 2030⁶, European Open Science Cloud Initiative¹⁴ and EU Industrial Strategy¹⁵. Here, we recommend the eNanoMapper data model as a best-practice example, being robust enough to maintain widely diverse forms of physicochemical and toxicological data, and flexible towards the inclusion of variable levels of links to external data, information, knowledge and analysis tools (which relate to I3, R1 and R1.3). These recommendations aim to enhance data availability and are expected to directly support the issues of poor data quality, which often relate to the lack of metadata reporting⁴⁹. In addition, we recommend and provide solutions that include both raw and processed data (R1). Raw data allows for re-interpretation, which may become relevant depending on the reuse scenario. For example, toxicity data are typically based on dose–response assessments and the final results are often reported as concentration-based effects, such as IC₅₀ (50% inhibitory concentration). However, new approaches to score and integrate complete dose–response results are in development^{8,51} and allow for comparability across widely diverse data types⁵². Finally, the reusability of data is greatly enhanced by data-usage licenses, such as Creative Commons licenses. Legal uncertainty exists for reusing non-licensed data, even when announced as openly available by academic researchers¹². This is clearly demonstrated by the NANoREG data, which have successfully been reused for the development of the Nanosafety Data Interface (described here) to design harmonized data-reporting templates⁴³, develop a harmonized domain-specific terminology⁴¹, establish SOPs for NMs⁴⁴ and develop computational models applicable to the safety assessment of NMs⁵³.

In conclusion, the experience compiled here serves as an example and sets the scene for the ongoing continued efforts to gather and generate inherently FAIR nanosafety data to support the efficient governance and regulation of NMs (<https://www.gov4nano.eu/>, <https://riskgone.eu/2020/>, <https://www.nanoinformatix.eu/>, <https://www.h2020gracious.eu/> and <https://nanosolveit.eu/>, <https://www.patrols-h2020.eu/>)¹. The recommendations provided are valid and aligned with the successful reuse of nanosafety data for a variety of purposes, as described above. To support the continued efforts, a FAIR implementation network for nano- and advanced materials was recently established (the AdvancedNanoIN, <https://www.go-fair.org/implementation-networks/overview/advancednano/>), which directly benefits from the recommendations provided here. Ultimately, FAIR nanosafety data support the sustainable reuse of publicly funded data for, for example, in silico modelling to inform potential further information needs in line with integrated approaches to testing and assessment^{2,6}. In addition, FAIR nanosafety data effectively contribute to digitalization and sustainable industrial processes through enabling an efficient information exchange⁵⁴. For this reason, the solutions offered by the Nanosafety Data Interface were recently adopted by the EUON (EU Observatory for Nanomaterials) (<https://euon.echa.europa.eu/enanomapper>) to support the increased use of safety information in product innovation and design¹, and facilitate the implementation of Safe by Design concepts spurred by the European Union^{15,16}. This article provides solutions and a concrete basis for practical FAIR data-driven safety decisions regarding nano- and advanced materials⁵⁵—decisions that contribute to a safe and sustainable world.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of

data and code availability are available at <https://doi.org/10.1038/s41565-021-00911-6>.

Received: 18 June 2020; Accepted: 28 March 2021;

Published online: 20 May 2021

References

- Soeteman-Hernandez, L. G. et al. Safe innovation approach: towards an agile system for dealing with innovations. *Mater. Today Commun.* **20**, 100548 (2019).
- Nymark, P. et al. Toward rigorous materials production: new approach methodologies have extensive potential to improve current safety assessment practices. *Small* **16**, 1904749 (2020).
- Karcher, S. et al. Integration among databases and data sets to support productive nanotechnology: challenges and recommendations. *NanoImpact* **9**, 85–101 (2018).
- Powers, C. M. et al. Nanocuration workflows: establishing best practices for identifying, inputting, and sharing data to inform decisions on nanomaterials. *Beilstein J. Nanotechnol.* **6**, 1860–1871 (2015).
- Mahony, C., Currie, R., Daston, G., Kleinstreuer, N. & van de Water, B. Highlight report: ‘Big data in the 3Rs: outlook and recommendations’, a roundtable summary. *Arch. Toxicol.* **92**, 1015–1020 (2018).
- Haase, A. & Klaessig, F. *EU–US Roadmap Nanoinformatics 2030* (EU Nanosafety Cluster, 2017); <https://doi.org/10.5281/zenodo.1486012>
- Marchese Robinson, R. L. et al. How should the completeness and quality of curated nanomaterial data be evaluated? *Nanoscale* **8**, 9919–9943 (2016).
- Giusti, A. et al. Nanomaterial grouping: existing approaches and future recommendations. *NanoImpact* **16**, 100182 (2019).
- Haase, A. & Lynch, I. Quality in nanosafety—towards reliable nanomaterial safety assessment. *NanoImpact* **11**, 67–68 (2018).
- Comandella, D., Gottardo, S., Rio-Echevarria, I. M. & Rauscher, H. Quality of physicochemical data on nanomaterials: an assessment of data completeness and variability. *Nanoscale* **12**, 4695–4708 (2020).
- Tropsha, A., Mills, K. C. & Hickey, A. J. Reproducibility, sharing and progress in nanomaterial databases. *Nat. Nanotechnol.* **12**, 1111–1114 (2017).
- Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
- Hendren, C. O., Powers, C. M., Hoover, M. D. & Harper, S. L. The Nanomaterial Data Curation Initiative: a collaborative approach to assessing, evaluating, and advancing the state of the field. *Beilstein J. Nanotechnol.* **6**, 1752–1762 (2015).
- European Open Science Cloud (EOSC) Strategic Implementation Plan* (European Commission, 2019); <https://op.europa.eu/en/publication-detail/-/publication/78ae5276-ae8e-11e9-9d01-01aa75ed71a1/language-en>
- A New Industrial Strategy for Europe* (European Commission, 2020); https://ec.europa.eu/info/sites/info/files/communication-eu-industrial-strategy-march-2020_en.pdf
- A New Circular Economy Action Plan for a Cleaner and More Competitive Europe* (European Commission, 2020); https://ec.europa.eu/environment/circular-economy/pdf/new_circular_economy_action_plan.pdf
- Chemicals Strategy for Sustainability Towards a Toxic-Free Environment* (European Commission, 2021); <https://ec.europa.eu/environment/pdf/chemicals/2020/10/Strategy.pdf>
- Jeliazkova, N. et al. The eNanoMapper database for nanomaterial safety information. *Beilstein J. Nanotechnol.* **6**, 1609–1634 (2015).
- Jeliazkova, N. et al. Linking LRI AMBIT chemoinformatic system with the IUCLID substance database to support read-across of substance endpoint data and category formation. *Toxicol. Lett.* **258**, S114–S115 (2016).
- Kochev, N., Jeliazkova, N. & Tsakovska, I. in *Big Data in Predictive Toxicology* (eds Neagu, D. & Richarz, A.-N.) 69–107 (The Royal Society of Chemistry, 2020).
- Hastings, J. et al. eNanoMapper: harnessing ontologies to enable data integration for nanomaterial risk assessment. *J. Biomed. Semant.* **6**, 10–10 (2015).
- Totaro, S. et al. The JRC Nanomaterials Repository: a unique facility providing representative test materials for nanoEHS research. *Regul. Toxicol. Pharm.* **81**, 334–340 (2016).
- Chomenidis, C. et al. Jaqpot Quattro: a novel computational web platform for modeling and analysis in nanoinformatics. *J. Chem. Inf. Model* **57**, 2161–2172 (2017).
- Mech, A. et al. Insights into possibilities for grouping and read-across for nanomaterials in EU chemicals legislation. *Nanotoxicology* **13**, 119–141 (2019).
- Precupas, A. et al. Thermodynamic parameters at bio–nano Interface and nanomaterial toxicity: a case study on BSA interaction with ZnO, SiO₂, and TiO₂. *Chem. Res. Toxicol.* **33**, 2054–2071 (2020).
- Berrios, D. C., Beheshti, A. & Costes, S. V. FAIRness and usability for open-access omics data systems. *AMIA Annu Symp. Proc.* **2018**, 232–241 (2018).

27. Jeliaskova, N. eNanoMapper—parsers for different NM data formats *GitHub* <https://github.com/enanomapper/nmdataparser>
28. Kochev, N. et al. Your spreadsheets can be FAIR: a tool and FAIRification workflow for the eNanoMapper Database. *Nanomaterials* **10**, 1908 (2020).
29. Gottardo, S. et al. NANoREG Framework for the Safety Assessment of Nanomaterials (Joint Research Centre, 2017); <https://doi.org/10.2760/245972>
30. Kermandadeh, A. et al. A multilaboratory toxicological assessment of a panel of 10 engineered nanomaterials to human health—ENPRA Project—the highlights, limitations, and current and future challenges. *J. Toxicol. Environ. Health B* **19**, 1–28 (2016).
31. Bos, P. M. J. et al. The MARINA risk assessment strategy: a flexible strategy for efficient information collection and risk assessment of nanomaterials. *Int. J. Environ. Res. Public Health* **12**, 15007–15021 (2015).
32. Nessler, F. NANoGENOTOX European joint action: what could we learn from all these data?. *Toxicol. Lett.* **229**, S35 (2014).
33. Juillerat-Jeanneret, L. et al. Biological impact assessment of nanomaterial used in nanomedicine. Introduction to the NanoTEST project. *Nanotoxicology* **9**, 5–12 (2015).
34. Dusinska, M. et al. Towards an alternative testing strategy for nanomaterials used in nanomedicine: lessons from NanoTEST. *Nanotoxicology* **9**, 118–132 (2015).
35. *Nano Exposure & Contextual Information Database (NECID)* (PEROSCH, accessed 1 March 2020); <https://perosh.eu/research-projects/perosh-projects/ncid/>
36. Pelzer, J. Structure and functionality of the Nano Exposure and Contextual Information Database (NECID). *Gefährst. Reinhalt. Luft* **73**, 302–304 (2013).
37. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
38. Kolesnikov, N. et al. ArrayExpress update—simplifying data submissions. *Nucleic Acids Res.* **43**, D1113–D1116 (2015).
39. NANOSOLUTIONS Data Repository (NANOSOLUTIONS, accessed 1 March 2020); <http://nanosolutionsfp7.com/>
40. Fernández-Cruz, M. L. et al. Quality evaluation of human and environmental toxicity studies performed with nanomaterials—the GUIDEnano approach. *Environ. Sci. Nano* **5**, 381–397 (2018).
41. Gottardo, S., Quiros Pseudo, L., Totaro, S., Riego Sintes, J. & Crutzen, H. NANoREG Harmonised Terminology for Environmental Health and Safety Assessment of Nanomaterials (European Commission, 2017); <https://doi.org/10.2788/71213>
42. Krebs, A. et al. Template for the description of cell-based toxicological test methods to allow evaluation and regulatory use of the data. *ATLA* **36**, 682–699 (2019).
43. Totaro, S., Crutzen, H. & Riego Sintes, J. *Data Logging Templates for the Environmental, Health and Safety Assessment of Nanomaterials* (Joint Research Centre, 2017); <https://publications.jrc.ec.europa.eu/repository/handle/JRC103178>
44. NANoREG Results Repository (RIVM, 2017); <https://www.rivm.nl/en/about-rivm/mission-and-strategy/international-affairs/international-projects/nanoreg>
45. Wilkinson, M. D. et al. Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Sci. Data* **6**, 174 (2019).
46. *Criteria for FAIR Research Data* (Swedish Research Council, 2019); <https://staff.ki.se/the-fair-principles>
47. Collins, S. et al. *Turning FAIR into Reality. Final Report and Action Plan from the European Commission Expert Group on FAIR Data* (European Commission, 2018); <https://doi.org/10.2777/1524>
48. Willighagen E., Jeliaskova N. NanoCommons—nanomaterial identifiers, basis for European Registry of Nanomaterials (ERM) *GitHub* <https://github.com/NanoCommons/identifiers/blob/master/registry>
49. Nymark, P. et al. *caLIBRAte D5.3—Document on Quality Criteria for Data* (EU Nanosafety Cluster, 2017); <https://doi.org/10.5281/zenodo.3859951>
50. Ammar, A. et al. A semi-automated workflow for FAIR maturity indicators in the life sciences. *Nanomaterials* **10**, 2068 (2020).
51. Nymark, P. et al. Grouping of representative nanomaterials is efficiently executed by combining high-throughput-generated biological data with physicochemical data. *Toxicol. Lett.* **314**, abstr. OP02-02 (2019).
52. Marvel, S. W. et al. ToxPi Graphical User Interface 2.0: dynamic exploration, visualization, and sharing of integrated data models. *BMC Bioinf.* **19**, 80 (2018).
53. Lamon, L. et al. Grouping of nanomaterials to read-across hazard endpoints: from data collection to assessment of the grouping hypothesis by application of chemoinformatic techniques. *Part. Fibre Toxicol.* **15**, 37 (2018).
54. Antikainen, M., Uusitalo, T. & Kivikytö-Reponen, P. Digitalisation as an enabler of circular economy. *Procedia CIRP* **73**, 45–49 (2018).
55. Falzetti, M., Keiper, W., Igartua, A. & Alliance for Materials (A4M) Consortium. *Opinion Paper on Governance and Strategic Programming of Materials Research and Innovation in Horizon Europe* (EUMAT, 2019); https://www.eumat.eu/media/uploads/descargas/2019_02_a4m_position_paper_v44.pdf
56. Carusi, A. et al. Harvesting the promise of AOPs: an assessment and recommendations. *Sci. Total Environ.* **628–629**, 1542–1556 (2018).
57. Martens, M. et al. WikiPathways: connecting communities. *Nucleic Acids Res.* **49**, D613–D621 (2021).
58. Davis, A. P. et al. Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Res.* **49**, D1138–D1143 (2021).
59. Sansone, S.-A. et al. Toward interoperable bioscience data. *Nat. Genet.* **44**, 121–126 (2012).
60. Jeliaskova, N., Haase, A., Ritchie, P., Shahzad, R. & Nymark, P. *NanoReg2 D1.8—Report on the Defined ISA-TAB Nano Templates* (European Commission, 2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2021

Methods

The Nanosafety Data Interface and NanoReg2 database instance. The Nanosafety Data Interface is enabled by the Apache Solr backend and brings together multiple installations, referred to as eNanoMapper database instances⁵⁵ (Box 1), which are populated with data from specific projects. The databases offer a large number of FAIR-aligned solutions, such as user-friendly web interfaces and REST-APIs, and use the same software as previously described (<https://www.enanomapper.net/>)¹⁸. In the following, the solutions applied to FAIRify data for the purpose of reaching the NanoReg2 objectives are described, and include reference to detailed technical publications where relevant. An overview of the technical abbreviations and terms is provided in Box 2.

During the early stages of the NanoReg2 project, a Data Solution Team was established (led by co-author A.H., and including N.J., P.N. and P.R.) to coordinate and solve the project data needs and the Team took the decision to apply the eNanoMapper data model and solutions⁶⁰. Previously established database instances that contained data generated in NANOREG, NanoTEST, MARINA, ENPRA and NANOGENOTOX were installed (with permission from the project coordinators if required) and updated several times within the project to support the data needs⁸. Consequently, in support of data findability, a NanoReg2-specific integrated database instance was installed and further populated with data generated within the project itself, exposure data that originated from NECID³⁵ and metadata for omics data that resided in omics repositories (see below for a further description and Supplementary Methods 2.1). Although the NANOREG data and the omics metadata are publicly accessible (CC BY-NC-SA 4.0 license), data from other projects, which include the NanoReg2 integrated view, are currently protected and access is granted on a case-by-case basis.

The eNanoMapper data model represents NMs as chemical substances (that is, one or more chemical structures with identified role(s), for example, core or coating) and links experimental data to the substance(s) rather than to the individual components, in line with the REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) definition of a substance. The approach is consistent with the fundamental concepts of 'test' and 'measurement' defined elsewhere (for example, ISA-Tab⁵⁹) and aligns with several of the FAIR principles, including F4 and I3 (Table 1). As the eNanoMapper data model is based on an open source software, AMBIT (<http://ambit.sourceforge.net/>), all the implementation details (including the database structure) can be examined. Although the overall structure was not changed in NanoReg2 (as compared to the original publication of the data model¹⁸), multiple minor improvements related to the annotation and arrangement of experimental entries were implemented and are reflected in the version control of the source code and subsequent releases of AMBIT software (<https://sourceforge.net/projects/ambit/>), which demonstrates the overall flexibility of the model⁶¹. It is worth noting that a user guidance for retrieving data from the NanoReg2 database was prepared and is publicly available⁶². The guide assists the user in finding the most relevant assay in relation to a set of selected risk-assessment tools that were evaluated within the project.

Furthermore, in support of the solutions for FAIRification provided by the eNanoMapper data model, a similar set-up that applies an aggregated search view of eNanoMapper database instances is currently hosted for the EUON (see Discussion). Owing to the requirements of the EUON, the set-up involves separate database instances (that is, not those used for the Nanosafety Data Interface) and the user interface is visually adapted to fit the overall appearance of the European Chemicals Agency pages, while using the same open source JavaScript library (<https://github.com/ideaconsult/jToxKit>) for the visualization of chemical substances and NM safety data.

Overview of data gathered and generated in NanoReg2. It is beyond the scope of the current article to describe the gathering and generation of data in detail. Thus, here we provide a brief overview of the data used to populate the NanoReg2 database for the purpose of demonstrating the FAIRification of nanosafety data and to identify challenges. In particular, it is important to note that the NanoReg2 database does not provide fully curated data, but rather a starting point and large cohesive datasets (in terms of NMs tested and SOPs applied) to support further fit-for-purpose curation efforts. Curation is a continuous process and several such efforts are ongoing (<https://www.gov4nano.eu/>, <https://www.nanoinformatix.eu/> and <https://nanosolveit.eu/>) and align with the overall concept of curation workflows and their iterative nature¹³. Further details on the gathered and generated data, as well as on the FAIRification workflow employed, can be found in the Supplementary Methods 2.2.

Briefly, data from six previous EU projects and from various public and project-specific omics repositories were gathered^{30–32,34,35,62}. The gathered data covered a wide variety of physicochemical, (eco)toxicological, human-toxicity-focused omics and exposure data for a wide variety of NMs. Next, data were generated specifically within the NanoReg2 project to fill identified data gaps for a set of 19 NMs, and covered physicochemical, bio-nano interaction^{26,64}, human toxicity and ecotoxicological data.

Data upload into the NanoReg2 database. Although the eNanoMapper database supports multiple structured import formats (for example, IUCLID, RDF and JSON (Box 2)), most nanosafety data from past and ongoing projects is collected

using customized spreadsheet templates, which currently encompass over 1,400 Excel files. The import of Excel files was enabled by a configurable parser that mapped the spreadsheet data via external configuration files and was continuously updated with bug fixes and new functionality (available at <https://github.com/enanomapper/nmdataparser>). In line with the FAIR principles, for example, F2 and F4, the data were subsequently indexed with the help of the mapping procedure, which assigned one or more ontology annotations to the NMs and to different measurement entries (for example, endpoints and methods), which can then be accessed through the aggregated search API. The indexing process is described in detail in Kochev et al.²⁸.

Briefly, Excel (.xls or .xlsx) spreadsheets that roughly follow either JRC ISA-Tab logic templates⁴³ or IOM dose-response templates⁶ (Box 2) were imported into the respective eNanoMapper database instances by creating JSON configuration files for each file or a set of similar files and using the configurable parser²⁸. Although the Excel/JSON importing capability is integrated with the eNanoMapper database application and can be used through a web-form upload, here the parser was integrated into a command line client to automate the import of several hundreds of files. Additionally, unit tests were developed to automatically import the files into a test database instance and to automatically verify sets of different criteria. New functionalities of the parser were also implemented to handle the specifics of the diverse files. Although the configuration JSON files provide a mechanism to map the Excel files structure into the eNanoMapper data model, the process of mapping data to the correct fields and understanding of the eNanoMapper data model requires expert-based manual curation (further details in Supplementary Methods 2.1). In particular, a major challenge was the diverse structures among the Excel files, which often only superficially followed the original design of the JRC or IOM templates. Data generators often customize the structure or the content according to specific needs, which requires the creation of separate configurations suitable for each batch of files (often between 1–5, and in a few cases ~10).

The overall workflow steps were as follows: the configuration files were created manually (and, more recently, semi-automatically), stored in a version control system and included in the automatic import scripts unit tests. The result of importing into a test database was verified for consistency and, if necessary, new automatic checkpoints were introduced. This process was iterative, as there was frequently a need to contact the data providers (or experts within NanoReg2) to obtain specific knowledge of the metadata (file content, assays and so on; Supplementary Methods 2.1). To maintain dialogue with data providers and/or experts, a collaboration system was installed, and access was granted to project partners willing to help. The collaboration system allows us to create tasks, which typically include questions about specific files and the follow-up answers or discussion. As the files and configurations are hosted in a version control maintained by the same collaboration system, it is easy to cross-reference updates of the configurations (or files) to specific knowledge exchange.

Disparate terminology in the files was handled through specific 'dictionary' mapping files for objects, which are automatically mapped when importing, and subsequently annotated with the eNanoMapper ontology (for example, NM names and types, endpoint categories, methods and cell type). Where possible, the experiments were linked to SOPs. For a small number of experiments, a quality flag was set to indicate potential problems with the data (based on feedback). Although the entire import workflow is automated once the configuration files and dictionaries are in place, the creation of these involves manual work and expertise. This process is now being automated through an online Template Wizard that applies community harmonized data entry templates and predefined configuration files (<https://search.data.enanomapper.net/projects/enanomapper/datatemplates/>)²⁸.

In addition to separate data files, the content in an SQL database obtained from the previous project NANOREG was converted into an eNanoMapper database SQL with an SQL script, originally developed during the eNanoMapper project⁶⁵. The process involved curation of the data to improve FAIRness, which included conversion to a common data model to address interoperability and reusability, harmonized IDs of, for example, NMs, cells, methods and so on, and add missing (meta)data and ontology annotation.

Semi-quantitative assessment of the FAIRness of the NanoReg2 database.

Several efforts are ongoing to establish quantifiable FAIR metrics to assess the FAIR maturity^{45,50}, but have not been adopted to evaluate specifically nanosafety data. Additionally, the FAIR principles are currently aspirational within the nanofield and the assessment of FAIRness also remains, at most, semi-quantitative. Thus, inspired by previous work to assess the FAIRness of omics databases²⁶, a semi-quantitative assessment was performed on the NanoReg2 database. Briefly, a scale of pass, partial pass and fail was used to rate the compliance of the data in the database with the 14 FAIR principles. Pass was used when no evidence of failure of a test by the metric was seen. Partial pass was used if the metric had multiple steps or components, and the database passed some, but not all, or if there was evidence that some data inputs for the metric yielded a pass, but others did not. Fail was used only if no evidence was found that the database was compliant with any part of the metric, for any inputs we tried. The ratings were then combined into a consensus rating, based on discussion among the co-authors. Finally, the overall FAIRness score used in the original publication²⁶ was applied, that is, the sum of each pass (1 point), partial pass (0.5 points) and fail (0 points).

In addition, the original assessment of the omics repository GEO²⁶ is included here to provide a comparison of the NanoReg2 database FAIRness with one of the most FAIR databases within life sciences (Table 1). We also assessed how FAIR the GEO database is to the nanosafety field specifically. This was done in line with the above-described approach to provide a comparable semi-quantitative FAIR score. A detailed description of the assessment of each of the 14 FAIR principles was compiled and included in the results table (Table 1).

Data availability

The datasets described here are available through the Nanosafety Data Interface and the NanoReg2 database (<https://search.data.enanmapper.net/projects/nanoreg2>). Data generated within NANOREG, NanoReg2 and the omics metadata are publicly available under CC BY-NC-SA 4.0 license. The NanoReg2 generated data is also available as SQL (DOI: 10.5281/zenodo.4713745, accessed 23 April 2021). Data that originate from the projects NanoTEST, ENPRA, MARINA and NANOGENOTOX are currently restricted from public use.

Code availability

The eNanoMapper data model is implemented in the open-source chemical substance management software AMBIT (<http://ambit.sf.net>). Machine readability for data retrieval and analysis is facilitated via an open source JavaScript client library (<https://github.com/ideaconsult/jToxKit>) and a Python client library (<https://github.com/ideaconsult/pynanomapper>).

References

- Jeliazkova, N. & Jeliazkov, V. AMBIT RESTful web services: an implementation of the OpenTox application programming interface. *J. Cheminformatics* **3**, 18 (2011).
- Shandilya, N. et al. NanoReg2 D3.2—Database/Structural Model and Report Describing the Relationships between Functionality, Physicochemical Properties and Hazard, and Allowing for Integration in the Safe Innovation Approach (2018); <https://doi.org/10.5281/zenodo.3854938>
- NANOREG D6.05 Database sql (RIVM, accessed 23 November 2019); <https://www.rivm.nl/en/documenten/nanoreg-d605-database-sql>
- Tanasescu, S. et al. in *Nanomaterials—Functional Properties and Applications* (eds Zaharescu, M. et al.) 85–97 (Publishing House of the Romanian Academy, 2020).
- Jeliazkova, N. et al. *eNanoMapper D3.4—ISA-Tab Templates for Common Bioselected Set of Assays* (European Commission, 2014); <https://doi.org/10.5281/zenodo.375814>

Acknowledgements

The work leading to this article has received funding from the European Union's Horizon 2020 Research and Innovation programme, Grant Agreements no. 646221 (NanoReg2, 2015–2019), no. 814401 (Gov4Nano, 2019–2022) and no. 814425 (RiskGONE 2019–2023). In addition, the European Union's 7th Framework Programme projects NANOREG (2013–2017, Grant Agreement no. 310584), NanoTEST (2008–2012, no. 201335) and ENPRA (2009–2012, no. 228789), the European Union's Health Programme Joint Action project NANOGENOTOX (2010–2013, no. 2009 21 01) and the US NIH NCI caNanoLab portal are acknowledged for providing data.

Author contributions

N.J. and P.N. conceptualized the study, interpreted the results and drafted the manuscript. A.H. designed and coordinated the work and acquisition of the data. M.D.A., C.A., F.B., A.B., C. Battistelli, C. Bossa, A.B.-P., A.C., I.D.A., M.D., N.E.Y., A.G., P.G.-F., D.G., R.G., M.G., N.R.J., V.J., K.A.J., N.K., P.K., N.M., E.M., A.M., J.M.N., V.P., A.P., T.P., K.R., P.R., I.R.L., E.R.-P., R.S., N.S. and S.T. contributed to the acquisition and analysis of data, as well as the formulation of the methods and results. All the authors approved the final version of the manuscript and agree to being accountable for their own contributions.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41565-021-00911-6>.

Correspondence and requests for materials should be addressed to N.J. or P.N.

Peer review information *Nature Nanotechnology* thanks Wojciech Chrzanowski and Iseult Lynch for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.