Report on mechanisms to overcome barriers to collect and make available the data

Report Title:	*Report on mechanisms to overcome barriers to collect and make available the data*		
Author(s):	Iseult Lynch		
Responsible Project Partner:	UoB	Contributing Project Partners:	FIOH

Document data:	File name / Release:	EC4SafeNano_D2 4-Barriers to Data sharing-commentsBC.docx		Release No.: 2
	Pages:	29	No. of annexes:	0
	Status:	Final	Dissemination level:	PU
Project title:	**EC4SafeNano**: European Centre for Risk Management and Safe Innovation in Nanomaterials & Nanotechnologies		Grant Agreement No.:	723623
WP title:	The Resources - Mapping the available resources addressing Risk management and Safe innovation		Deliverable No:	D2.4
Date:	Due date:	October 31, 2019	Submission date:	December 18, 2019
Keywords:	Open Data Pilot, Data management, Data sharing, FAIR data			
Reviewed by:	A. Duschl		Review date:	November 30, 2019
	B. Caillard		Review date:	November 30, 2019
Approved by WP leader:	A.-K. Viitanen		Approval date:	December 18, 2019
Approved by Coordinator:	V. Dulio		Approval date:	December 18, 2019

Verneuil-en-Halatte, December 2019

Release History

Release No.	Date	Change
1	December 4, 2019	Initial draft circulated for comments and feedback
2	December 13, 2019	Minor typos and small additions / corrections

Project Contact

INERIS – Institut National de l'Environnement Industriel et des Risques
Parc Technologique ALATA, BP 2
60550 Verneuil-en-Halatte
Tél.: 03 44 55 66 77
Registered in Compiegne, France under B 381 984 921

Coordinator: Dr. Valeria Dulio
Email: valeria.dulio@ineris.fr, Tel: +33 3 44 55 66 47

EC4SafeNano Project

The European Centre for Risk Management and Safe Innovation in Nanomaterials and Nanotechnologies, EC4SafeNano, is a 2016-2019 Coordination and Support Action, funded by the European Commission. EC4SafeNano is coordinated by INERIS, and operated together by major European risk institutes with the support of numerous associated partners, gathering all stakeholders involved in Nanomaterials and Nanotechnologies (regulators, industry, society, research, service providers...).

A central challenge to ensure the sustainable production and use of nanotechnologies is to understand and effectively control the risks along the industrial innovation value chain. Knowledge about nanotechnology processes and nanosafety issues (hazards, fate, risk...) is growing rapidly but the effective use of this knowledge for risk management by market actors is lagging behind.

EC4SafeNano will promote a harmonized vision of expertise in risk assessment and management for the public and private sectors to enable the safe development and commercialization of nanotechnology. The main objective of EC4SafeNano is to design harmonized services in risk assessment and management and a sustainable structure to deliver these services. For that, the project will gather stakeholder needs and expertise resources. It will demonstrate the efficiency of the proposed solution on case studies.

Executive Summary

This report describes the outcomes and best practice recommendations from the EC4SafeNano workshop on overcoming data barriers held as part of the EC4SafeNano data at the NanoSafe Conference in 2018. The workshop involved participants from across the spectrum of nanosafety research, some with data management backgrounds and some more experimental or regulatory in their expertise. The workshop presented some of the state-of-the-art approaches to data sharing and data management being developed in H2020 projects ACEnano and NanoCommons, and included audience participation via the use of mentimeter to gauge participants own concerns regarding barriers to sharing nanosafety data. The key concerns and the free text questions from the workshop participants are discussed in Section 4, and some recommendations and pathways forward are presented in the conclusions. Given EC4SafeNano's proposed role as a distributed centre for excellence in nanosafety, access to high quality data in a timely manner will be central to the centre's ability to add value and provide services at the interfaces between research and policy, research and regulation, and research translation into commercial products. Results and recommendations from the workshop are also feeding into the updated version of the EC4SafeNano Data Management Plan to ensure that the centre can deliver the services required by Users.

Table of Contents

List of Figures

List of Acronyms

Acronym	Definition
DINS	Differences in Nominal Significance
DMP	Data Management Plan
DOI	Digital Object Identifier
EUON	European Union Observatory for Nanomaterials
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable and Re-usable (data)
GDPR	General Data Protection Regulation
GMO	Genetically Modified Organisms
IPR	Intellectual Property Rights
NCI	National Cancer Institute
NDCI	Nanomaterials Data Curation Initiative
NECID	Nano Exposure & Contextual Information Database
SOP	Standard Operating Procedure

1 Introduction

Big data is a phenomenon revolving around the endeavor of accumulating massive amounts of data and using it to understand the objects from which we are gathering data. Combined with the technology of machine learning and artificial intelligence, it is possible to argue that much knowledge is now produced autonomously by the *tools* scientists have made, and not directly by the scientists themselves. As big data has gained prominence, it has spread from the fields of business and computer science into society at large and just about every other science.[2] A core tenet of big data is *access* to data. In the arena of hazard and risk assessment, sharing of data can support improved risk management processes, design of improved materials and much more. For example, Goel et al., discuss the sources, challenges, and discuss the benefits of big data analytics in process safety using four case studies with different applications ranging from incident database analysis, predictive modeling for pump failures, dynamic risk mapping of operating plant, and image analysis. They concluded that the application of big data analytics would provide valuable insights for more informed policy, strategic, and operational risk decision-making leading to a safer and more reliable industry.[3]

The benefits of data sharing have been widely recognized – transparency and cooperation, reproducibility of research, cost-efficiency and preventing redundancies, acceleration of discovery and innovation, and in medicine and public health can even lead to the saving of lives through more efficient and effective public health programs. The added value of Big Data and open data is likely to increase in large, aggregated datasets in which data from different sources (e.g., from different social sectors and industries) are combined.[4] Despite the recognized benefits of data sharing, there are a number of barriers to this, and a number of legal and ethical concerns in terms of the re-use of data.[5] In addition to actual barriers, there are also perceived or anticipated barriers (see Figure 1) which can also prevent researchers or organisations from even trying to share or access shared data.

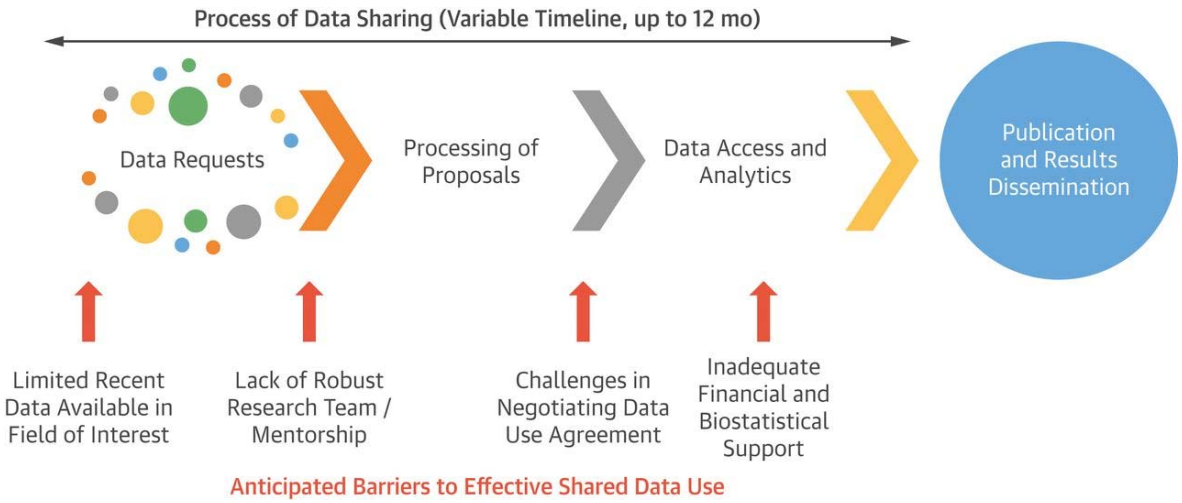

Figure 1: Practicalities and Anticipated Hurdles Related to Shared Data Requests, Access, and Use by Early Career Investigators. Image from McCarthy and Vaduganathan [5].

To help optimize the potential for future sharing and re-use of data, the EC4SafeNano Data Management Plan (DMP) (implemented in WP2, Task 2.3 and described in its final version in Deliverable Report D2.1) helps the consortium partners to consider any problems or challenges that may be encountered with regards to

sharing of their own data and helps them to identify ways to overcome these. Additionally, the DMP outlines the data sharing practices to be implemented by the consortium for the data collected within EC4ASfeNano, much of which is linked to personal data and thus would be subject to (GDPR).

A survey on barriers to data sharing globally by Springer Nature found that the main challenge to data sharing was identified by the 7,700 respondents as 'Organising data in a presentable and useful way' (46%), with other challenges generally rated: 'Unsure about copyright and licensing' - 37% 'Not knowing which repository to use' - 33% 'Lack of time to deposit data' - 26% 'Costs of sharing data' - 19%. Of those surveyed, 66% had made data available in some form, with 67% publishing data as supplemental material to publications (although this is typically as a pdf file which is difficult to mine from an automatic mining viewpoint), 57% presenting at a conference (which is not re-usable), and 42% providing data on request. 51% of respondents also included datasets in a repository – either an institutional one (26%), a discipline-specific data repository (19%) or a general-purpose data repository such as figshare (6%).[6]

This deliverable report addresses these barriers and their applicability in the field of nanosafety research, and consults with the nanosafety research and stakeholder communities to develop solutions to the identified barriers to data sharing and re-use.

2 Literature findings on generic barriers to data sharing

This section is summarized from van Panhuis, W.G., et al., *A systematic review of barriers to data sharing in public health.* BMC Public Health, 2014. **14**(1): p. 1144.

Few, or none, of the issues related to data sharing are nanomaterials specific, and thus there is enormous scope for learning from, and leveraging of best practice in data management and data sharing from other contiguous areas. Among the most challenging areas for data sharing, where personal data related to health is involved, is that of public health data. A detailed analysis of the barriers to data sharing in public health identified 20 unique real or potential barriers to data sharing which were classified into a taxonomy of six categories: technical, motivational, economic, political, legal, and ethical barriers.[7] These barriers and categories describe a landscape of challenges that is highly dynamic, interconnected, and hierarchical. Although most evidence (68%) was published in peer-reviewed sources, less than a quarter (22%) of all the documents reviewed was based on empirically derived evidence, indicating that a large volume of published expert knowledge has not yet been translated into scientific evidence.[7] Before evaluating the situation in the nanosafety field, and overview of these 6 categories and 20 barriers to data sharing in the field of public health, and if/how the specific barrier applies in nanomaterials safety research, are summarized below.

2.1 Technical barriers

These barriers are, for the most part, well understood as part of resilient challenges in information system capacity and continue to form a major obstacle to the availability and use of (public health) data. Solutions to these barriers have been identified but sustainable implementation and political/financial commitment have been limited.

1. Data not collected. As long as severe limitations persist in (public health) data collection, data sharing will not be considered a priority. Similarly for nanomaterials Environmental Health and Safety (nano-EHS) data, the perception of large data gaps may be an impediment to sharing what data there is. This is also coupled with the general feeling that older data is of less use than newer data due to the lack of accompanying nanomaterials characterisation information. However, this ignores the fact that the more data is available the less individual gaps matter and the fact that there are artificial intelligence methods that work with sparce data and that can impute missing values where needed.[8]

2. Data not preserved or 3, cannot be found. Public health data are often collected for short-term purposes such as outbreak detection with data preservation or archiving not prioritized, and data retrieval systems lacking. This is amplified by relocation of offices, staff turnover, physical damage to paper or electronic files, computer viruses, computer theft, etc.[9] This is a similar situation for much of the academic nano-EHS data, where drivers for data collection are the individual PhD or postdoc project, and the short-term nature of their contracts, lead to similar challenges in terms of long-term data capture. Development of institutional data management policies, and project-specific DMPs are addressing this, but real progress will require educational investment and incentivisation for change. However, these are still focussing on the open access to publications via institutional repositories with less focus as yet on the sharing and re-use of the underpinning datasets.

4. *Language barrier.* Routinely collected public health data are often recorded in local languages, limiting the possibility to integrate and use such data together with other data sets. Similarly, much of the scientific literature in non-English language journals are likely missed and even more inaccessible to machine learning models than currently. Steinberger suggests that's to achieve highly multilingual text mining applications the

most needed resources would be simple, parallel and uniform multilingual dictionaries, corpora and software tools.[10]

5. *Restrictive data format.* Despite major advances in computational resources in public health, a large volume of public health data such as disease surveillance data and administrative data continue to be collected and preserved in hardcopy paper format or in electronic format that may be antiquated or incompatible with modern software systems (e.g. floppy disks, zip disks etc.). This is less of a problem in research generally, although lab notebooks are often still kept in paper copy, and data typed up into excel sheets. The advent of electronic notebooks is addressing this, but change is slow and general adoption of these solutions takes time. For example, the Nano Exposure & Contextual Information Database (NECID) database has published field forms to harmonize the notebooks used for nanomaterials exposure measurements[1], while the EU H2020 project NanoCommons is developing a set of workflows for nano-EHS in electronic notebooks with data and metadata collection templates linked directly to the NanoCommons Knowledgebase and integration of protocols and Standard Operating Procedures (SOPs), as well as calibration data etc. as part of an effort to increase the overall quality of the data and thus confidence in the data.

6. *Technical solutions not available.* Technical software solutions to collect, harmonize (transformation and recoding to enhance inter-operability), integrate (combining harmonized datasets), and share complex and heterogeneous data have been developed in the private or research sector, but have not become widely available to public health agencies.[11] Similarly, nano-EHS researchers are not aware of many of these solutions, which remain in the technical / developer / coder space rather than filtering down to the bench researchers. Additionally, the fact that the existing data is so poorly organized, and that very limited nano-EHS-specific databases exist (e.g. the NECID, eNanoMapper), and/or are maintained only via EU-funded projects limits the ability to develop and apply nano-EHS specific solutions. Some progress has been made via AMBIT templates[2] (via eNanoMapper and NanoREG EU projects) and through international harmonization efforts led by the Nanomaterials Data Curation Initiative (NDCI), which is part of the National Cancer Institute (NCI) Nanotechnology Working Group working group on nanoinformtics,[12] but true automation of the collection and organisation of nano-EHS remains some way off. The aforementioned electronic lab notebooks are an important step in this direction, since data is captured directly from the instruments, linked to the SOPs and annotated to the database schema for easy and automated organization of the data.

7. *Lack of metadata and standards.* Oftentimes, metadata that describe data content, origin, methods, etc. are lacking for public health data and standards for data format, variables, and metadata are insufficiently used, limiting secondary data use and inter-operability. For nano-EHS research, several meta data approaches have been proposed, including the nano-extension to the ISATab format (investigation-study-assay), but again, are insufficiently applied, and evolving as the field matures.[13] A consensus paper on the mammal and gold standard for metadata for nano-EHS is under development (led by the University of Birmingham) as the last in the series of papers on Nanoinformatics planned by the NDCI as part of the NCI Nanotechnology Working group, who have to date addressed topics such as data curation workflows [14] and data completeness and quality.[15]

2.2 Motivational barriers

These include barriers based on personal or institutional motivations and beliefs that limit data sharing. Solutions for this group of barriers lie in building trust or developing transparent legal agreements.

8. *No incentives.* Data sharing requires time and resources that are chronically lacking in public health settings, and which have yet to be built into many funding models and university project management

[1] https://perosh.eu/wp-content/uploads/2017/10/NECID_-fieldform_version-May-2015_final.pdf
[2] http://ambit.sourceforge.net/enanomapper/templates/

processes. Personal and institutional incentives are often required to prioritize data sharing over other duties, particularly if the benefit of data sharing is delayed and uncertain (e.g. possibly more efficient disease control programs) instead of immediately relevant to data providers (e.g. scientific credit or training).

9. *Opportunity cost.* Researchers who have invested time and effort in data collection could anticipate that scientific credit or other opportunities may be lost if data recipients with greater capacity for analysis could gain the majority of credit. This is a particular challenge in low resource settings, and for early career scientists who need papers for their next post. However, increasing recognition of the value of publishing datasets, and the emergence of data-specific publications (e.g. Scientific Data, Data in brief etc.) are helping to address this issue by allow publication of the complete dataset alongside the article describing the findings and analysis of the data.

10. *Possible criticism.* Data providers could be discredited by errors found during secondary use of their data. In the worst case, data sharing could reveal data fabrication or manipulation, although to be fair publication itself exposes this risk also where data are fabricated, and this represents a very tiny majority of the data produced -around 2% of biomedical papers in an analysis were found to have manipulated figures.[16]

11. *Disagreement on data use.* Data providers may disagree with the intended secondary use of their data or may consider their data inappropriate for a certain use.

2.3 Economic barriers

These barriers concern the potential and real cost of data sharing and solutions depend on the recognition of data value and on sustainable financing mechanisms.

12. *Possible economic damage*. Data sharing in public health (and likely also nano-EHS) is challenged by the economic damage that this may cause to data providers. Public sharing of disease outbreak data, for example, can result in economic damage due to reduced tourism and trade.[17] The global SARS outbreak led to estimated economic losses of 50 billion USD between 1998 and 2004 and Foot & Mouth Disease in the UK resulted in losses of 30 billion USD between 1998 and 2003.[17] The possibility of such significant economic implications due to (over) reactive market forces could cause great reluctance among health agencies to rapidly release disease data. There are similar concerns in nanosafety – including fear of negative public opinion and a repeat of the situation with genetically modified organisms (GMO) where the European public rejected all GMO foods outright, potential loss of business etc., although in fact regulatory uncertainty is greatest barrier to industry.[18] Sharing of data is thus encouraged under REACH, but with clear costs as well as benefits to sharing data, e.g., in industrial consortia to register chemicals.[3]

13. *Lack of resources.* The process of data sharing requires human and technical resources for data preparation, annotation, communication with recipients, computer equipment, internet connectivity, etc. These resources cost money and such IT support is not yet common in universities.

2.4 Political barriers

These are fundamental structural barriers embedded in the public health governance system that are grounded in a political or socio-cultural context. Solutions for these barriers are not clear-cut and will require global and national processes to build consensus and political will.

14. *Lack of trust.* Trust between a data provider and user greatly enables data sharing. In the absence of trust, providers could anticipate potential misinterpretation, misuse or intentional abuse of the data.

[3] https://echa.europa.eu/documents/10162/13585/article_industry_consortia_en.pdf/238c930b-abf0-4308-b715-313abf3237ac

15. *Restrictive policies.* Agencies may have developed official policy guidelines that restrict data sharing, resulting from various possible underlying factors such as a general sense of distrust, negative prior experiences, or other factors. In nano-EHS (and EHS generally) there is an underlying issue of commercial sensitivity in terms of the data associated with regulatory dossiers, and the consortia that generate the dossiers and register specific nanomaterials invest considerable time and resources into the generation of the data which has very high commercial value. Thus, sharing of the data is not straightforward and the data itself has monetary value.

16. *Lack of guidelines.* Frequently, official guidelines on data sharing simply do not exist, are unclear or inconsistent. The balance between making data accessible, safeguarding privacy, and protecting intellectual, time and financial investments by public health staff is often not well regulated or standardized, resulting in protective policies on sharing of public health data in general.[19]

2.5 Legal barriers

These barriers are legal instruments used to restrict data sharing, resulting from the underlying willingness (or not) to share data. Solutions to this group of barriers include legal instruments to facilitate data sharing and are highly dependent on solutions to underlying political barriers.

17. *Ownership and copyright.* Agencies that collect public health data are often responsible for the protection of individual and community privacy and may feel that a guardianship or ownership role is bestowed on them by the public. This could result in a default of restricting access to most data.[11] Copyright can be used to restrict rather than expand access to data. In practice, it is often not well documented or known who owns public health data, resulting in inconsistent ad-hoc guidelines.[20]

18. *Protection of privacy.* A clear distinction between data containing personal identifiers and fully anonymous data may not always be possible, leading to restrictive policies on all types of data due to privacy concerns. Aggregated data without personal identifiers may not be sufficiently detailed for certain applications. Existing tools and standards for the de-identification of personal identifiers such as statistical data masking [21] may not be known or available in many contexts.[22]

2.6 Ethical barriers

These are normative barriers involving conflicts between moral principles and values. Solutions for these barriers will involve a global dialogue among all stakeholders on the ethical principles that should govern data sharing.

19. *Lack of proportionality.* The issue of proportionality, the careful deliberation in assessing the risks and benefits that derive from the amount and type of data requested compared to the potential impact of its secondary use, has been identified as a guiding ethical principle for public health data sharing.[4] Public health agencies may disagree with data requestors about the proportional risks and benefits of the secondary use of data and its impact on public health.[23]

20. *Lack of reciprocity.* Data sharing practices have not always been fair, and data producers have often felt exploited in transactions where they receive little credit or benefit from their work, while data users that can rapidly analyze data and publish results benefit from academic credit and career advancement as has happened in the past.[24]

[4] Global health data access principles. http://www.gatesfoundation.org/global-health/Documents/data-access-principles.pdf,

3 Organisation of the Barriers to data sharing workshop

The workshop was organized as a satellite event to the 2018 NanoSAFE conference in Grenoble, under the overall title of the EC4SAfeNano day on 5th November 2018[5], and was one of three sequential workshops held on the day (the others being *Establishment and operation of EC4SafeNano Focus Network* and *Blueprint for Nanosafety Platform Development & Sustainability)*. Further details of the organization process, overview of attendees etc. is captured in the overall report from the event (EC4SafeNano conference). Below an overview of the agenda is presented to provide focus for the subsequent discussion of the data barriers identified and the key areas of concern for participants.

Overcoming barriers to making data FAIR – integrating data management into data generation workflows-a joint workshop with H2020 NanoCommons

(Organised by Iseult Lynch, I.Lynch@bham.ac.uk, UoB with support from Anna-Kaisa Viitanen, FIOH and Geethu Balachandran, EUvRI)

Aim:

The workshop aims to gain feedback on utility and user acceptability of proposed solutions to knowledge managementand FAIR data and based on the user / stakeholder feedback to develop recommendations regarding solutions to maximise data sharing and data accessibility for the entire community and all stakeholders. Among the barriers to be considered, and for which best-practice solutions will be developed, are:

Primary publication of data before the data are made available in a database;

Labelling of the data (ontology);

Confidentiality of the data;

Security of the data storage and access (trust that the data are protected);

Format of the data for long-term accessibility and sharing etc.

Agenda:

15:30 –16:15	**Opening and introduction to the data life cycle and the NanoCommons solution –online notebooks and their applicability to experimental scenarios of increasingly complexity up to mesocosms**	**Iseult Lynch & Anastasios Papadiamantis (UoB)**
16:15 –16:35	ACEnano approach integrating and streamlining data analysis and output formats for nanomaterials characterisation	Thomas Exner (Douglas Connect)
16:35 –16:55	Adapting data management tools and platforms to industry stakeholders – stand-alone versus cloud applications	Antreas Afantitis (NovaMechanics)
16:55 –17:25	Discussion on the solutions presented and other key barriers –suggestions to improve / adapt etc. Collection of other examples of low-cost / Open Source tools for data	Facilitated by Iseult Lynch (UoB)

[5] https://www.nanosafetycluster.eu/wp-content/uploads/Events/EC4SafeNano%20Day_5%20November%202018_Agenda.pdf?_t=1538174280

	management / data sharing from the stakeholder community.	
17:25 –17:30	Wrap-up and key recommendations	Iseult Lynch (UoB)

Expected Outcome:

A set of recommendations for EC4SafeNano data providers on how to capture, process and share their data to maximise its FAIRness, i.e., its Findability, Accessibility, Interoperability and Re-usability, which is essential to enable EC4SafeNano partners to provide their proposed services. Potential case studies where solutions are still missing will also be identified, which could be taken up within NanoCommons, e.g. with EC4SafeNano as the "User" of the NanoCommons research infrastructure expertise. Feedback to NanoCommons and ACEnano on their tools and services will also be provided.

The workshop posed a series of questions to the audience to stimulate the discussion, and also invited the audience to pose questions to us for discussion and subsequent follow-up in this deliverable report. Details of the questions and the answers developed within EC4SafeNano to address them are presented in Section 4. The tool mentimeter was used to prepare the questions in advance of the session, and participants were able to vote or answer the questions using their smart phones and a link to the sessions-specific set of questions.

As the last session of a long, intensive day, we were very pleased to have such a good turn-out for the workshop – over 30 participants engaged in the session, with most (>25) participating in polling also, which is an excellent response rate. Some of the questions posed by the audience were challenging and required some thought and effort to devise response to that address the issue or show where progress is being made. Section 4 of this deliverable report covers the responses from the EC4SafeNano Barriers to Data sharing working.

4 Output from the Barriers to data sharing workshop

The first question posed to workshop participants during the discussion was to tell us (type in) the first word that popped into their head when they thought about data barriers, and there words then formed a Word Cloud, shown in Figure 2. The size of the word represents the relative frequency with which it appeared in the text entered by the workshop participant. Clearly the most significant barrier / concern identified by the workshop participants related to the data quality, and how to ensure quality of the data in databases. This also appeared as alternative terms such as bad data, consistency, purpose (of the data) etc. Confidentiality (confidential, privacy) were also significant concerns, as were intellectual property rights (IP / IPR).

Figure 2: Word cloud based on inputs from 23 participants in the barriers to data sharing workshop. The Word Cloud was based on first impressions relating to barriers to data sharing.

Participants were next asked to indicate on a scale of 0-5 how strongly they agreed or disagreed with a set of statements regarding barriers to data sharing, including the need for training on data sharing as part of university curricula, and what might incentivize data sharing in both academia and industry. The results of this are shown in Figure 3 below.

Participants were then asked to rank the suggested barriers in order of importance from most to least significant, with the results shown in Figure 4. Not surprisingly, issues related to intellectual property rights related to data deposited in databases was the most significant barrier identified. Finally, participants were

asked to suggest questions for detailed discussion and analysis, with responses to these questions presented in the last part of this report. Six questions were submitted, as presented below, although not all were specific to nanomaterials / nanosafety.

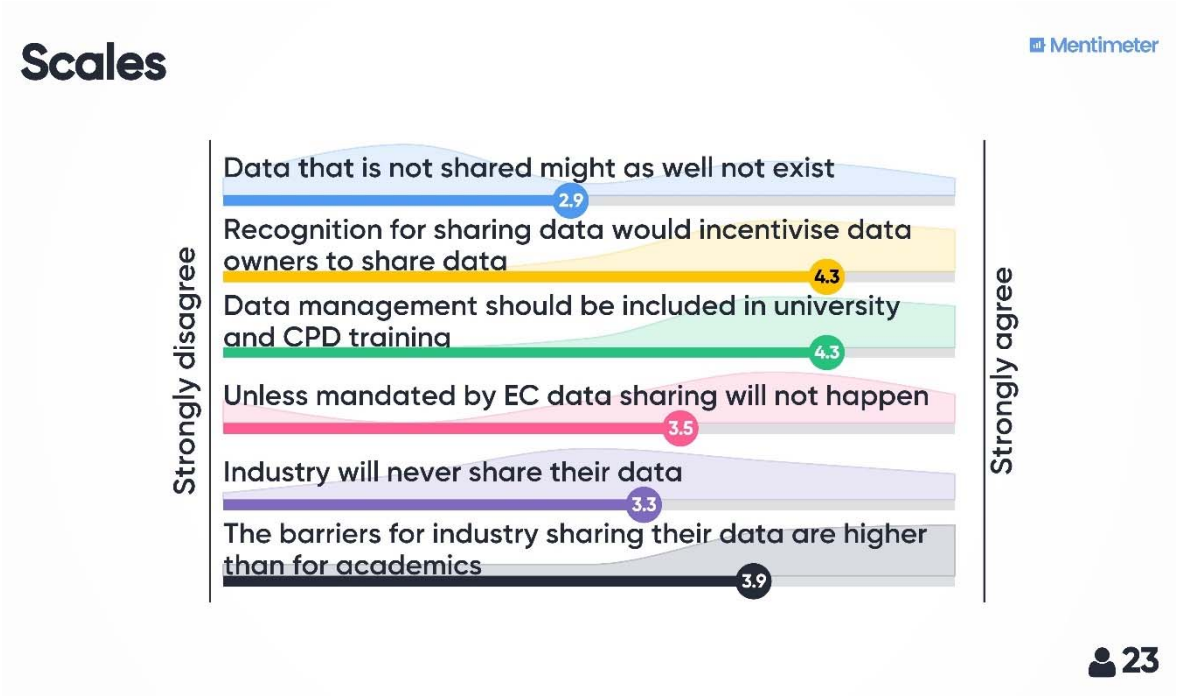

Figure 3: Responses of the participants of the workshop on barriers to data sharing on how strongly they agree or disagree on a scale of 0-5 with the statements shown below.

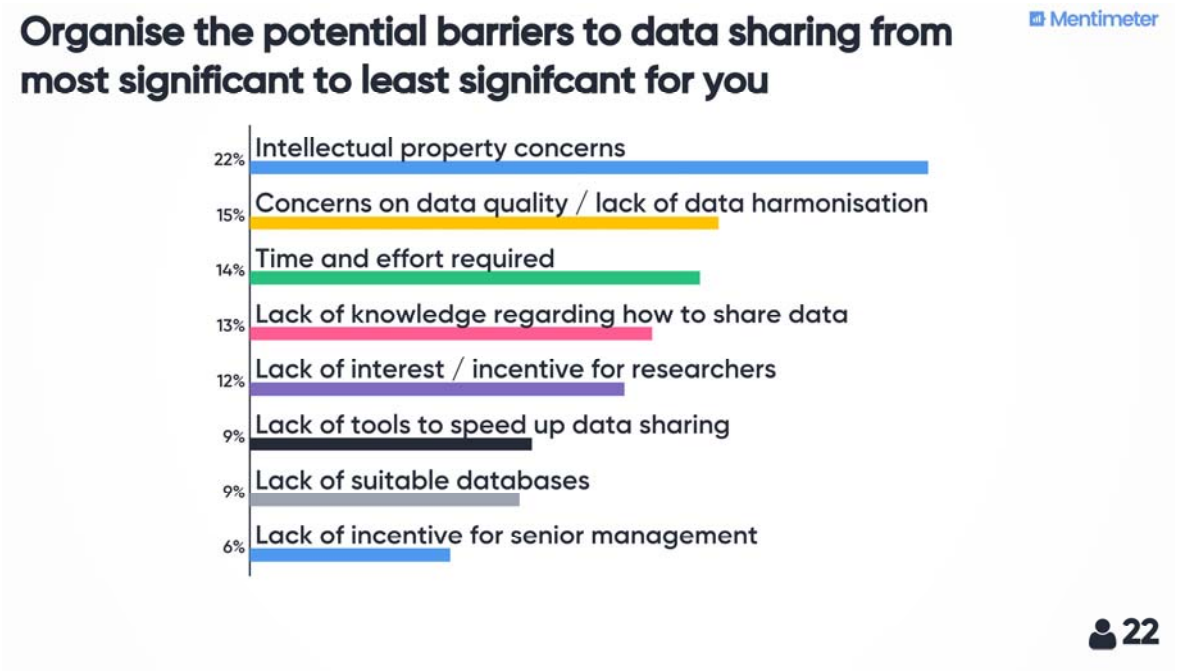

Figure 4: Ranking of the potential barriers to data sharing by the participants of the workshop. Not surprisingly IPR issues were the most significant, as also reflected in the free text questions submitted by the audience and addressed in the next section.

5 Responses to the Questions from the audience on their biggest concerns regarding data sharing

5.1 Q1 - How to deal with trash in - trash out?

Database owners should have a plan in place to either automatically clean up the database or periodically do a manual scrub to avoid reporting on inaccurate data. A key part of this is understanding the types of errors in datasets and data management and how to address them.

This section is adapted from Brown et al., *Issues with data and analyses: Errors, underlying themes, and potential solutions*. PNAS, 2018 115:2563-2570.

A variety of themes or taxa of errors have been proposed.[25] For examples, errors in datasets can be related to measurement, study design, replication, statistical analysis, analytical choices, citation bias, publication bias, interpretation, and the misuse or neglect of simple mathematics.[26] Others have categorized errors by the stage of the research process – for example, Bouter et al. classified "research misbehaviors" in four domains: reporting, collaboration, data collection, and study design.[27] Many items within various themes or taxa overlap: One person's research misbehavior may be classified as another's statistical error.[25]

5.1.1 Errors producing "bad data."

We define bad data as those acquired through erroneous or sufficiently low-quality collection methods, study designs, or sampling techniques, such that their use to address a particular scientific question is scientifically unjustifiable. In one example, self-reported energy intake has been used to estimate actual energy intake. This method involves asking people to recall their dietary intake in one or more ways, and then deriving an estimate of metabolizable energy intake from these reports. The method, compared with objective measurements of actual energy intake, turns out to be invalid,[28] not just "limited" or "imperfect." The measurement errors are sufficiently large and nonrandom that they have led to consistent and statistically significant correlations in the opposite direction from the true correlation for some relationships. Moreover, the relations between the errors and other factors are sufficiently numerous and complex that they defy simple corrections. Concerns about this method were raised decades ago, and yet its use is continued.

Other common examples of bad data include confounding batch effects with study variables of interest [29] and cell-line misidentification or contamination.[30] For confounding or contamination, the data are bad from failed design and are often unrecoverable.

Bad data represent one of the most egregious of themes of errors because there is typically no correct way to analyse bad data, and often no scientifically justifiable conclusions can be reached about the original questions of interest. It also can be one of the more difficult errors to classify, because it may depend on information like the context in which the data are being used and whether they are fit for a particular purpose.

5.1.2 Errors of data management

Errors of data management tend to be more idiosyncratic than systematic. Errors we have seen (and sometimes made) are the result not of repeating others' errors, but of constructing bespoke methods of handling, storing, or otherwise managing data. In one case, a group accidentally used reverse-coded

variables, making their conclusions the opposite of what the data supported.[31] In another case, authors received an incomplete dataset because entire categories of data were missed; when corrected, the qualitative conclusions did not change, but the quantitative conclusions changed by a factor of >7.[32] Such idiosyncratic data management errors can occur in any project, and, like statistical analysis errors, might be corrected by reanalysis of the data. In some cases, idiosyncratic errors may be able to be prevented by adhering to checklists (as proposed in ref. [33]).

Errors in long-term data storage and sharing can render findings nonconformable because data are not available to be reanalysed. Many metaanalysts, including us, have attempted to obtain additional information about a study, but have been unable to because the authors gave no response, could not find data, or were unsure how they calculated their original results. Brown *et al* note that they themselves have struggled on occasion to find their own raw data from older studies and welcome advances in data management, data repositories, and data transparency.[25]

5.1.3 Errors of statistical analysis

Errors of statistical analysis involve methods that do not reliably lend support to the conclusions. These can occur if the underlying assumptions of the analyses are not met, the wrong values are used in calculations, statistical code is mis-specified, incorrect statistical methods are chosen, or a statistical test result is misinterpreted, regardless of the quality of the underlying data. First, misanalysis of cluster-randomized trials [34] may inappropriately and implicitly assume independence of observations. Worse still, when there is only one cluster per group, clusters are completely confounded with treatment, resulting in zero degrees of freedom to test for group effects. Second, effect sizes for metaanalyses may inappropriately handle multiple treatment groups (e.g., assuming independence despite sharing a control group) or fail to use the correct variance component in calculations. In turn, the metaanalytic estimates from these effect-size calculations may be incorrect, and have sometimes required correction.[35] Third, it is inappropriate to compare the nominal significance of two independent statistical tests as a means of drawing a conclusion about differential effects.[36] This "differences in nominal significance" [DINS] error is sometimes committed in studies with more than one group, in which final measurements are compared with baseline separately for each group; if one is significant and one is not, an author may erroneously conclude that the two groups are different.

5.1.4 Considerations for "gatekeeper" functions.

Gatekeeper functions create circumstances in which people have no choice but to "do the right thing." Such solutions have already been implemented in several domains, such as requirements for registration of trials. Requirements for depositing of raw data and publication of statistical code have been implemented by some journals. Some funders and contracts require the posting of results, such as for studies registered in ClinicalTrials.gov. One thing these functions have in common is increasing the amount of information reported—"increased transparency." After all, it is difficult to identify errors if insufficient information is provided to be able to evaluate the science.

These gatekeeper functions are important for forcing some actions. However, without the intrinsic buy-in from cultural shifts or extrinsic incentives, researchers may only comply within the letter of the requirements, rather than the intended spirit of rigor. Tightening requirements too far may risk the creation of a system that will fail to be flexible enough to accommodate a variety of scientific areas. Gatekeeping functions may increase this burden, and have been criticized as the bureaucratization of science.[37] Burdens can be alleviated by additional resources, such as new job roles tailored to requirements within institutions, much like an interdisciplinary approach alleviates the need for a single scientist to be a polymath.

5.2 Q2 - Recycling vs possible environmental impact for each type of ENM and article/products - how to deal with this in the near future?

Not related to data barriers, so not a topic for this report, but something to be considered by others within the EC4SafeNano project as part of the overall menu of tools and services offered to address stakeholder needs.

5.3 Q3 - Can FAIR data management tools presented here be used for other biological data? so students in other research fields can be taught generically in FAIR data management.

Making FAIR data a reality requires a major change in the practice of many research communities, institutions and funders. Some disciplines have made great progress already in the sharing and reuse of research data; important lessons can be learnt from these examples. Data storage, preservation, and dissemination can be tackled at a generic, cross-disciplinary, disciplinary level or at a more granular, sub-disciplinary level. Successful implementation of the FAIR principles generally requires significant resources at the disciplinary level to develop the data-sharing framework (i.e. principles and practices, community-agreed data formats, metadata standards, tools, data infrastructures, etc.). Disciplinary interoperability frameworks are essential to the realisation of FAIR. Such frameworks have been developed in certain disciplines and often rely on a shared research culture and shared research and data infrastructures. Nevertheless, as fields shift their boundaries and the scientific grand challenges of the 21st century require collaboration across traditional disciplines (e.g. involving the social sciences in medical, scientific or engineering research), attention needs to be paid to the extremely challenging task of developing FAIR data frameworks across disciplines and for interdisciplinary research. Care should be taken to articulate interoperability frameworks in ways that adopt common standards and enable brokering across disciplines to break down silos. Coordination on the development of standards and infrastructure as the FAIR ecosystem is implemented via the EOSC, and in similar initiatives globally, will be critical.

Training for FAIR:

The High Level Expert Group report on the EOSC mentions an urgent need to educate and equip up to 500,000 data stewards in Europe. Hence, GO FAIR will strongly focus on training. The GO TRAIN pillar will coordinate the training to use and provide FAIR data and services. The guiding principles of the GO TRAIN pillar were drafted at a GO FAIR-related training workshop on 3 February 2017 in Paris, as follows.

The main GO TRAIN guiding principles are:

1. The GO TRAIN pillar will be most effective if it will focus on the skills needed by the following five types of professionals:
    - research data specialist's based at institutions,
    - institutional research data advisor's,
    - people that are responsible for long-term data stewardship,
    - all researchers (training includes respect for professional data stewards),
    - organisational managers (capacity building and best practice in skills management).
2. The GO TRAIN approach
    - addresses the needs of researchers and other professionals in an increasingly non-traditional academic sector,
    - performs an advocacy function in addition to training,
    - adopts a coordinating role,

- provides some form of certification of materials and activities,
- promotes and encourages the following training approaches: blended learning, train-the-trainers, peripatetic career pathways (i.e., trainers travelling from place to place), and new & innovative mechanisms for skills acquisition.

3. The workshop report lists several organisations that could form an effective GO FAIR Training Implementation Network, including their current data training portfolios. A matrix of these organisations and the required training activities should be developed.
4. The primary objective of the GO TRAIN pillar is to direct and accelerate the widespread implementation of data training in the identified areas. The optimal institutional arrangement to exert this function requires further discussion. The five priority activities are
   - refine and focus training requirements,
   - signpost and accredit to assist discovery and adoption of materials; create or use existing training material directories,
   - training advice, career & capacity building,
   - coordinated and scaling delivery of training; create and maintain trainers network,
   - encourage and perform train-the-trainer activities.

Examples of available training:

https://marketplace.eosc-portal.eu/services?providers=80

The ELIXIR Training Platform was established to develop a training community that spans all ELIXIR member states (see the list of Training Coordinators). It aims to strengthen national training programmes, grow bioinformatics training capacity and competence across Europe, and empower researchers to use ELIXIR's services and tools. The current programme builds on the work established during the previous work programme (2014-2018).

https://elixir-europe.org/platforms/training.

5.4 Q4 - Numerous purposes of use in a unique data base structure. Is it possible?

The databases / knowledgebases presented as part of the workshop by ACEnano and NanoCommons were researcher-focused databases but also have stated goals of providing solutions for industry and regulators, as part of the impact expected by the EU. However, since different users will have different quality needs from the data, and the quality scores and degrees of completeness needed vary for different purposes, the simplest solution is to have a set of scoring (and FAIRness) criteria that will allow users to determine easily if a specific dataset is suitable for their purposes.

NanoCommons is building on the various scoring systems available including Klimish, GUIDEnano etc.

FAIRness score:

There are some ongoing efforts to score data in terms of its Openness and FAIRness, as well as suggestions that the requirements for what is FAIR in different research communities should be agreed bottom-up within the research community. One such approach is the 5-star deployment scheme presented in Table 1 below, which assesses the Findable aspect of FAIR, and another is the Go-FAIR initiative, which is a network of networks working on FAIR developing Transformation tools to make data FAIR. The FAIR Badge System is proposed as a proxy for data quality assessment, and aims to operationalise the original FAIR principles to ensure no interactions among the four dimensions in order to ease scoring. The system then considers Reusability as the resultant of the other three, i.e. the average FAIRness score is then (F+A+I)/3=R . Efforts are also underway to develop an automated tool for assessing FAIRness, the fairdata webtool, which is currently available as a prototype.

Table 1: The 5-star deployment scheme from the Open which assesses how "Findable" data is

★	Available on the web (whatever format) *but with an open licence, to be Open Data*
★★	Available as machine-readable structured data (e.g. excel instead of image scan of a table)
★★★	As (2) plus non-proprietary format (e.g. CSV instead of Microsoft Excel)
★★★★	All the above, plus: Use open standards from W3C (RDF and SPARQL) to identify things, so that the data can be referenced by the users
★★★★★	All the above, plus: Link your data to other available data to provide context

Data quality scores

Klimisch et al. (1997) [IL5] developed a scoring system to assess the reliability of data from toxicological and ecotoxicological studies. This rating system has been extended to physico-chemical studies and is now accepted by many regulatory authorities and organisations. Indeed, ECHA's IUCLID 6 software includes rationale for assigning Klimisch scores for toxicology testing, based on utilisation of standard assays, and thus much of the data being utilised in NanoCommons will not have a high **Klimisch score**. However, this approach will be integrated as the infrastructure is intended to have utility for industry and regulators as well as the research community, and indeed the modelling community developing QSARs for use in regulatory risk assessment require high quality data. To automate the process of assigning Klimisch scores, NanoCommons will integrate the ToxRTool" (Toxicological data Reliability Assessment Tool), an Excel-based tool developed by ECVAM to provide comprehensive criteria/guidance for evaluation of the inherent quality of toxicological data.

NanoCommons is also building on work done previously in the GUIDEnano project (coordinated by NanoCommons partner Leitat), shown schematically in Figure 5. Note, however, that within the GUIDEnano project the approach was tested and refined based only on 137 peer-reviewed articles, and as such some developmental work is required to adopt it to experimental data sets and to align it to the NanoCommoasens Knowledge Base, ontology and semantic structures.

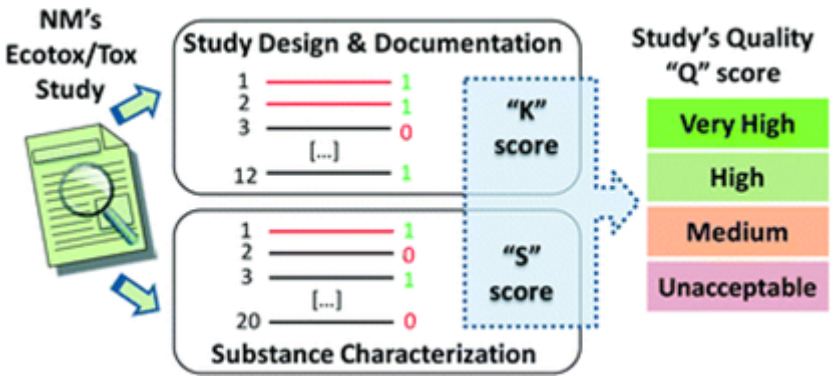

Figure 5: The scoring system related to test design and reporting considerations developed following the principles of the Klimisch score (K score). In addition, a scoring system based on the physicochemical properties that have been characterised and reported for the nanomaterial, including properties characterised in the exposure medium was developed within the FP7 project GUIDEnano (S score). These two scores (K and S) are combined to obtain an overall quality score (Q score) that can be used to select studies, to weight different studies, and/or to introduce uncertainty factors in the risk assessment process. This approach is also being implemented for NanoCommons datasets and databases. From Fernández-Cruz et al, 2018[1].

5.5 Q5 - There is need for data but lack of harmonized data. Can we develop procedures to assess the quality of what is now available?

The answer to this question aligns with that of Q3.

EU H2020 funded projects such as NanoREG and GRACIOUS[6] have been developing standardized reporting templates for nanosafety data, aligned where possible with the OECD guidelines.

5.6 Q6 - Concern over domination of final regulation submission by an enclosed nano platform council over the members?

EC4SafeNano is not itself building a nano-governance council, as there have been three EU projects funded to address this topic (NanoRIGO, Gov4Nano and RiskGone). Instead, it's goal is to provide a network of member states organizations who to design harmonized services in risk assessment and management and a sustainable structure to deliver these services. For that, the project will gather stakeholder needs and expertise resources. It will demonstrate the efficiency of the proposed solution on case studies and provide data into the risk governance council to facilitate their decision-making processes. Similarly, any regulatory framework would come only from the regulatory agencies (ECHA, EFSA, EMA etc.).

Users have the choice of whether / how to database their data and as such can choose to deposit their data into a repository whose structure and approach / ethos suits their individual needs. That said, mandated solutions, such as the use of IUCLID6 for data related to regulatory dossiers are, by necessity, highly structured. Tools and / or data repositories aligned to the structures of IUCLID6 are emerging as a means to facilitate data entry / data transfer. IUCLID is the essential tool for any organisation or individual that needs to record, store, submit, and exchange data on chemical substances in the format of the OECD Harmonised Templates. IUCLID6 itself is also continuously evolving: Version 6.4 includes an extended and more complete web interface, as well as an update to the format.

As it's not clear how the question relates to the topic of barriers to data sharing, it is not discussed any further here.

5.7 Q7 - Can privacy and IP issues be relieved if data are directly treated in e.g. machine learning tools?

Re-use of data

The meaning of "Informed Consent" and its implication for reusing human subject open data from and for biomedical research has been explored in the context of volunteer patient data and its re-use.[38] Patients and volunteers donate their data in the context of research designs that are vetted and approved by ethics committees. When research data are released in open access – especially observational data - these can be reused to explore a number of new hypotheses. As we know from previous studies, biomedical data can be reused in many unpredictable ways – new research communities are formed around pre-existing data, and the free availability of research data increases innovation, knowledge integration, and reproducibility. At the same time, openness of data could also expose donors to surveillance and discriminatory research practices that not only have ethical implications, but also were never agreed upon by the donors at the moment of data collection.

[6] https://op.europa.eu/en/publication-detail/-/publication/1d240ad5-e32f-11e9-9c4e-01aa75ed71a1/language-en

Taxonomy of the re-use of data

Custers and Vrabec proposed a taxonomy for data reuse as follows.[4] From a data owners (controller's) perspective, the major types of data reuse are as follows:

† Data recycling—using data several times for the same purpose,

† Data repurposing—using data for different purposes than for which they were initially collected,

† Data recontextualization—using data in another context than in which they were initially collected.

The second and third meanings of data reuse are expected to be of most added value in the European data economy. This distinction between several types of data reuse from a data owners' perspective may be useful for interpretation of the privacy principles, particularly the use limitation principle. Current legislation does not distinguish between data repurposing and data recontextualization. It may be argued that the use limitation principle is, in a way, not specific enough.

Even though the use limitation principle restricts the use of personal data for purposes other than those specified, 81 Article 7(3), and thus addresses function creep, Custers and Vrabec suggest that there is a major difference (particularly from a data subject's rights perspective) between reuse in the same context or in a different context (for instance, when a data controller sells the personal data to other, and when personal data are transferred to other countries). Consequently, it may be argued that asking consent for data reuse should be more much more explicit in cases of data recontextualization than in cases of data repurposing. Asking consent for data repurposing, in turn, should be more explicit than consent for data recycling, for which consent may be assumed in most cases.

From a data subject's perspective, the major types of data reuse are as follows:

† Data sharing or data disclosure—data subjects have the ability to (directly) allow use and reuse of their personal data,

† Data portability—data subjects have the ability to use and reuse their personal data across devices and services,

† The right to be forgotten—data subjects have the ability to block data use and reuse.

Strictly speaking, the third item (the right to be forgotten) is not a type of data reuse, but a right to block data reuse. As such, data sharing/data disclosure and data portability are all ways to reuse data and may promote Big Data, whereas the right to be forgotten is an instrument that has exactly the opposite effect, i.e., it intends to limit the use of Big Data.

The differentiation of data reuse from a data owner's perspective is of more obvious use (both from a practical and technological perspective and for the further interpretation of privacy principles) than the differentiation from a data subject's perspective. Custers and Vrabec suggest that this is a direct result of the fact that most privacy principles address data controllers regarding what they may or may not do with personal data. This assumes that restricting data processing for data owners guarantees better privacy and personal data protection of data subjects.[4] This assumption seems to be controversial in the era of Big Data, in which many Big Data firms have sound privacy policies (from a legal perspective), but data subjects still have many privacy concerns (from a social perspective).

Intellectual property (IP) issues for databases and data re-use

Extracted from: *Big Data & Issues & Opportunities: Intellectual Property Rights* by Debussche & César

Debussche and César state that while the general rules governing the protection of database are established at international level, EU law provides for a specific protection of databases which goes beyond other international legal instruments,[39] via the Database Directive[7] with the objective of harmonising the

[7] Directive 96/9/EC of the European Parliament and of the Council on the legal protection of databases [1996] OJ L 077/20 (Database Directive)

protection of databases in all Member States. Databases, within the broad meaning of the Database Directive, are protected in the EU by:

(i) copyright, where such copyright protection echoes the one recognised in the international treaties; and

(ii) a *sui generis* right.

While copyright protects the (original) structure of the database, the *sui generis* right aims to cover the investment made in its creation. These two rights are independent, and can be applied separately. They will however apply cumulatively if the conditions for both regimes are simultaneously met.

The term of the *sui generis* protection is much shorter than that of the copyright protection. It is limited to 15 years as from the first of January of the year following the date of completion of the database. However, such protection may in practice be much longer. According to the Database Directive, any substantial change to the contents of the database, that could be considered to be a new investment, will cause the term of protection to run anew.[8] In practice, should such protection be applied in a big data context, this could result in providing an indefinite protection, given that databases are usually dynamic, hence, leading in all likelihood to "substantial changes to the contents of the database".

Copyright protection of databases

Copyright protection is granted to databases which, as such, by reason of the selection or arrangement of their contents, constitute the "author's own intellectual creation".[9] A database structure may be protected under copyright even if the elements contained therein are in the public domain or are otherwise not protected by copyright.

However, the originality criterion might be more difficult to fulfil in case of automatically created electronic databases that contain data selected by software, without the actual involvement of an author. In such situations it seems more likely to award copyright protection to the underlying software (algorithm written in a way allowing for selection of specific data/types of data), than to the database itself.

This is particularly relevant in a big data context. Indeed, the development of technology has enabled data analytics of unstructured data. Accordingly, while protection of datasets is particularly relevant, the protection of the database structure has become less relevant and more difficult when confronted to new types of databases, unforeseen by the (over twenty-year-old) Database Directive.

Sui generis protection of databases

The second type of protection introduced by the Database Directive is the protection awarded on the basis of a *sui generis* right[10], rewarding the substantial investment of the database maker in creating the database. It was developed in order to prevent free-riding on somebody else's investment in creating the database and exists in parallel to the copyright protection on the structure of the database.

In order for a database to be protected by the *sui generis* right, an investment must be made in the creation of the database. The jurisprudence of the Court of Justice of European Union has clarified that an investment in the creation of the data as such does not suffice to merit protection under the *sui generis* right.[11] Such reasoning would entail that the *sui generis* right does not apply to machine-generated databases, as it could be argued that the data included in such databases are 'created' instead of 'obtained'. This could have a

[8] Article 10(3) of the Database Directive stipulates indeed that "*any substantial change, evaluated qualitatively or quantitatively, to the contents of a database, including any substantial change resulting from the accumulation of successive additions, deletions or alterations, which would result in the database being considered to be a substantial new investment, evaluated qualitatively or quantitatively, shall qualify the database resulting from that investment for its own term of protection*".

[9] No other criteria shall be applied to determine the eligibility of databases for that protection (Database Directive, art 3(1))

[10] The term "*sui generis* right" is a generic one and means "the right of its own kind".

[11] Case C-46/02 *Fixtures Marketing Ltd v. Oy Veikkaus AB* [2004] ECLI:EU:C:2004:694; Case C-338/02 *Fixtures Marketing Ltd v. Svenska Spel AB* [2004] ECLI:EU:C:2004:696; Case C-444/02 *Fixtures Marketing Ltd v. Organismos Prognostikon Agonon Podosfairou* [2004] ECLI:EU:C:2004:697; Case C-203/02 *British Horseracing Board Ltd and others v William Hill Organization Ltd* [2004] ECLI:EU:C:2004:695, para 42

broader effect on the data economy, which relies on digitisation processes such as Internet of Things devices, big data, and artificial intelligence; as it becomes increasingly difficult to distinguish between the generation and the obtainment of data in the context of such processes.[12]

That being said, there is no automatic exclusion from *sui generis* protection when the database's creation is linked to the exercise of a principal activity in which the person creating the database is also the one creating the materials that are processed in the database. It is however always the responsibility of that person to demonstrate a substantial investment (qualitative and/or quantitative) in the obtaining, verification or presentation of the content, independent from the resources used to create the content.[13]

Debussche and César foresee that it will become increasingly difficult to satisfy the *sui generis* right protection requirements in a data economy context, given that the processes of obtaining, verifying and/or presenting the data will happen more and more automatically, as they will be normally conducted using an algorithm. In many cases, it might be true that the investment in creating the raw material exceeds the investment made in segmenting and aligning that pre-existing raw material. In those cases, it might be more difficult to rely on the *sui generis* protection.

In Debussche and César's view it is regrettable that the Database Directive, which was drafted in the 90s, does not accommodate for the technical evolution and thus everything that is possible with data and databases today. For instance, it is unclear how techniques of enrichment, partitioning, harmonisation, homogenisation, etc. of data would fit within the criteria of obtaining, verification or presentation of the database contents. Moreover, the criterion of 'verification' may become less and less pertinent, especially in a big data context which allows analytics of unstructured data.

In summary then, it is possible to protect both IP and personal data in the context of data re-use via databases. However, The developments in the area of Big Data call for new technological models (regarding standardization and adequate IT infrastructure), new economic models (regarding corporate secrecy and IP rights), new social models (regarding public support), and new legal models (regarding personal data protection) in which the reuse of data is encouraged rather than hindered. Notably, these issues are not specific to nanomaterials or nanomaterials safety data, and as such are being addressed in the wider databasing communities. Official providers of data services such as databases are aware of the regulatory landscape and are contributing to the development of appropriate solutions.

[12] Staff Working Document - Evaluation of Directive 96/9/EC on the legal protection of databases {SWD(2018) 147 final}; http://edz.bib.uni-mannheim.de/edz/pdf/swd/2018/swd-2018-0146-en.pdf

[13] Case C-203/02 Horseracing Board Ltd and others v William Hill Organization Ltd [2004] ECLI:EU:C:2004:695, para 35

6 Summary and next steps

This report describes the outcomes from the EC4SafeNano workshop on data barriers. The workshop presented emerging data management tools and solutions under development in H2020 projects running in parallel to EC4SafeNano, including the e-infrastructure project NanoCommons, and the nanomaterials characterization project ACEnano, as well as the Nanoinformatics project NanoSolveIT. Some thought-provoking questions came up in the discussion, which have been addressed in this report, including aspects of data re-use and the legalities around this, as well as considerations of who owns the data deposited into databases.

This work will feed into the ongoing activities in NanoCommons and the EU Open Science Cloud, as well as supporting activities of the three risk governance projects RiskGONE, NanoRIGO and Gov4Nano. A summary paper on some of the key issues related to barriers to data sharing for nanosafety and nano-EHS is being prepared based on the outcomes from the workshop and the literature performed in order to address the questions raised by the audience.

22% of workshop participants ranked Intellectual property concerns as their number 1 barrier to data sharing, followed by 15% citing concerns on data quality / lack of data harmonization, and 14% ranked the time and effort required to share data as their most significant barrier. Lack of incentives for researchers (12%) and senior management (6%) were not especially significant barrier, and a lack of tools to speed up data sharing and suitable databases were identified by 9% of participants each. These findings are consistent with barriers identified in other contiguous areas such as public health data.

References

1. Fernández-Cruz, M. L., Hernández-Moreno, D., Catalán, J., Cross, R.K., Stockmann-Juvala, H., Cabellos, J., Lopes, V.R., Matzke, M., Ferraz, N., Izquierdo, J.J., Navas, J.M., Park, M., Svendsen, C., Janer, G., Quality evaluation of human and environmental toxicity studies performed with nanomaterials – the GUIDEnano approach. *Environ. Sci.: Nano* 2018, *5* 381-397.

2. Skaug Sætra, H., Science as a Vocation in the Era of Big Data: the Philosophy of Science behind Big Data and humanity's Continued Part in Science. *Integr Psychol Behav Sci.* 2018, *52*, 508–522.

3. Goel, P., Datta, A., Mannan, M.S. In *Application of Big Data analytics in process safety and risk management*, Conference: 2017 IEEE International Conference on Big Data (Big Data), 2017.

4. Custers, B., Vrabec, H.U., Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection. *International Data Privacy Law* 2016, https://ssrn.com/abstract=3046774

5. McCarthy, C. P., Vaduganathan, M., Navigating Data Sharing in Cardiology From a Trainee's Perspective. *Journal of the American College of Cardiology* 2018 *71*.

6. Nature, S. *Whitepaper: Practical challenges for researchers in data sharing Recent white paper baed on survey of >7,700 researchers worldwide;* https://researchdata.springernature.com/users/69154-springer-nature/posts/31633-whitepaper-practical-challenges-for-researchers-in-data-sharing; 2018.

7. van Panhuis, W. G.; Paul, P.; Emerson, C.; Grefenstette, J.; Wilder, R.; Herbst, A. J.; Heymann, D.; Burke, D. S., A systematic review of barriers to data sharing in public health. *BMC Public Health* 2014, *14* (1), 1144.

8. Walport, M., Brest, P., Sharing research data to improve public health. *The Lancet* 2011, *377*, 537-539.

9. Baldwin W, D. J. *Poverty, gender, and youth: Demographic data for development in sub-Saharan Africa.*; 2009.

10. Steinberger, R. *Challenges and methods for multilingual text mining*

11. Abo-Farha, S. A.; Abdel-Aal, A. Y.; Ashour, I. A.; Garamon, S. E., Removal of some heavy metal cations by synthetic resin purolite C100. *Journal of Hazardous Materials* 2009, *169* (1), 190-194.

12. Karcher, S.; Willighagen, E. L.; Rumble, J.; Ehrhart, F.; Evelo, C. T.; Fritts, M.; Gaheen, S.; Harper, S. L.; Hoover, M. D.; Jeliazkova, N.; Lewinski, N.; Marchese Robinson, R. L.; Mills, K. C.; Mustad, A. P.; Thomas, D. G.; Tsiliki, G.; Hendren, C. O., Integration among databases and data sets to support productive nanotechnology: Challenges and recommendations. *NanoImpact* 2018, *9*, 85-101.

13. Thomas, D. G.; Gaheen, S.; Harper, S. L.; Fritts, M.; Klaessig, F.; Hahn-Dantona, E.; Paik, D.; Pan, S.; Stafford, G. A.; Freund, E. T.; Klemm, J. D.; Baker, N. A., ISA-TAB-Nano: A Specification for Sharing Nanomaterial Research Data in Spreadsheet-based Format. *BMC Biotechnology* 2013, *13* (1), 2.

14. Powers, C. M. M., K A.; Morris, S.A.; Klaessig, F.; Gaheen, S.; Lewinski, N.; Ogilvie Hendren, C., Nanocuration workflows: Establishing best practices for identifying, inputting, and sharing data to inform decisions on nanomaterials. *Beilstein J Nanotechnol.* 2015, *6*, 1860-1871.

15. Marchese Robinson, R. L.; Lynch, I.; Peijnenburg, W.; Rumble, J.; Klaessig, F.; Marquardt, C.; Rauscher, H.; Puzyn, T.; Purian, R.; Åberg, C.; Karcher, S.; Vriens, H.; Hoet, P.; Hoover, M. D.; Hendren, C. O.; Harper, S. L., How should the completeness and quality of curated nanomaterial data be evaluated? *Nanoscale* 2016, *8* (19), 9919-9943.

16. Bik, E. M.; Casadevall, A.; Fang, F. C., The Prevalence of Inappropriate Image Duplication in Biomedical Research Publications. 2016, *7* (3), e00809-16.

17. Marsh Inc:. *The Economic and Social Impact of Emerging Infectious Diseases*. Marsh Inc.: New York, NY, 2008.

18. Hoerr, R. A., Regulatory uncertainty and the associated business risk for emerging technologies. *Journal of Nanoparticle Research* **2011,** *13* (4), 1513-1520.

19. Strobl, J.; Cave, E.; Walley, T., Data protection legislation: interpretation and barriers to research. **2000,** *321* (7265), 890-892.

20. Stansfield, S., Who owns the information? Who has the power? *Bull World Health Organ* **2008,** *86* (3), 170-171.

21. Anderson, M., Seltzer, W., Federal Statistical Confidentiality and Business Data: Twentieth Century Challenges and Continuing Issues. *Journal of Privacy and Confidentiality* **2009,** *1*.

22. El Emam, K.; Mercer, J.; Moreau, K.; Grava-Gubins, I.; Buckeridge, D.; Jonker, E., Physician privacy concerns when disclosing patient data for public health purposes during a pandemic influenza outbreak. *BMC Public Health* **2011,** *11* (1), 454.

23. Willison, D., Ondrusek, N., Dawson, A., Emerson, C., Ferris, L., Saginur, R., Sampson, H., Upshur, R. *A Framework for the Ethical Conduct of Public Health Initiatives*; Public Health Ontario: Toronto, Canada, 2012.

24. Butler, D.; Cyranoski, D., Flu papers spark row over credit for data. *Nature* **2013,** *497* (7447), 14-15.

25. Brown, A. W. K., K.A.; Allison, D.B., Issues with data and analyses: Errors, underlying themes, and potential solutions. **2018,** *115* (11), 2563-2570.

26. Allison, D. B. B.-R., J.; Burlingame, B.; Brown, A. W.; le Coutre, J.; Dickson, S. L.; van Eden, W.; Garssen, J.; Hontecillas, R.; ; Khoo, C. S. H. K., D.; Kussmann, M.; Magistretti, P.J.; Mehta, T.; Meule, A.; Rychlik, M.; Vögele, C., Goals in Nutrition Science 2015–2020. **2015,** *2* (26).

27. Bouter, L. M.; Tijdink, J.; Axelsen, N.; Martinson, B. C.; ter Riet, G., Ranking major and minor research misbehaviors: results from a survey among participants of four World Conferences on Research Integrity. *Research Integrity and Peer Review* **2016,** *1* (1), 17.

28. Dhurandhar, N. V.; Schoeller, D.; Brown, A. W.; Heymsfield, S. B.; Thomas, D.; Sørensen, T. I. A.; Speakman, J. R.; Jeansonne, M.; Allison, D. B.; the Energy Balance Measurement Working, G., Energy balance measurement: when something is not better than nothing. *International Journal of Obesity* **2015,** *39* (7), 1109-1113.

29. Pulit, S. L. L., M.; Menelaou, A.; De Bakker, P.I.W. , Association Claims in the Sequencing Era. *Genes* **2014,** *5*, 196-213.

30. Hughes, P.; Marshall, D.; Reid, Y.; Parkes, H.; Gelber, C., The costs of using unauthenticated, over-passaged cell lines: how much more data do we need? **2007,** *43* (5), 575-586.

31. Verhulst, B.; Eaves, L. J.; Hatemi, P. K., Correlation not Causation: The Relationship between Personality Traits and Political Ideologies. **2012,** *56* (1), 34-51.

32. Bent, S.; Tiedt, T. N.; Odden, M. C.; Shlipak, M. G., The Relative Safety of Ephedra Compared with Other Herbal Products. *Annals of Internal Medicine* **2003,** *138* (6), 468-471.

33. Rosen, D., The Checklist Manifesto: How to Get Things Right. *JAMA* **2010,** *303* (7), 670-673.

34. Allison, D. B.; Brown, A. W.; George, B. J.; Kaiser, K. A., Reproducibility: A tragedy of errors. *Nature* **2016,** *530* (7588), 27-29.

35. Zalewski, B. M.; Chmielewska, A.; Szajewska, H., The effect of glucomannan on body weight in overweight or obese children and adults: A systematic review of randomized controlled trials. *Nutrition* **2015,** *31* (3), 437-442.e2.

36. Gelman, A.; Stern, H., The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician* **2006,** *60* (4), 328-331.

37. Editors, T., The Registration of Observational Studies—When Metaphors Go Bad. **2010,** *21* (5), 607-609.

38. Pasquetto, I. V., Beyond privacy: the emerging ethics of data reuse. . In *Workshop presented at the Cochrane Colloquium 2018, Edinburgh.*, 2018.

39. Debussche, J., César, J., Big Data & Issues & Opportunities: Intellectual Property Rights 2019.