Integration among Databases and Data Sets to Support Productive Nanotechnology: Challenges and Recommendations

Egon L. Willighagen*[1], Sharon Gaheen*[2], Sandra Karcher*[3], Christine Ogilvie Hendren[4], Marty Fritts[2], Dennis G. Thomas[5], Stacey Harper[6], Mark D. Hoover[7], Richard L. Marchese Robinson[8], Karmann C. Mills[9], John Rumble[10], Nina Jeliazkova[11], Friederike Ehrhart[1], Georgia Tsiliki[12], Axel P. Mustad[13], Nastassja Lewinski[14] and Chris T. Evelo[1]

Address: [1]Department of Bioinformatics - BiGCaT, Maastricht University, P.O. Box 616, UNS50 Box 19, NL-6200 MD, Maastricht, The Netherlands; [2]Leidos Biomedical Research Inc., Frederick National Laboratory for Cancer Research, Frederick, MD, 21702, USA; [3]Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA; Center for the Environmental Implications of NanoTechnology (CEINT) Duke University, Durham, NC 27708-0287, USA; [4]Center for the Environmental Implications of NanoTechnology (CEINT) Duke University, Box 90287, 121 Hudson Hall, Durham, NC 27708-0287, USA; [5]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, USA; [6]Environmental and Molecular Toxicology and School of Chemical, Biological and Environmental Engineering, Oregon State University, Corvallis, OR 97331, USA; [7]National Institute for Occupational Safety and Health, 1095 Willowdale Road, Morgantown, WV 26505-2888, USA; [8]School of Chemical and Process Engineering, University of Leeds, Leeds LS2 9JT, UK (current); School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, James Parsons Building, Byrom Street, Liverpool, L3 3AF, United Kingdom (previous); [9]RTI International, 3040 Cornwallis Rd., Research Triangle Park, NC 27709, USA; [10]R&R Data Services, 11 Montgomery Avenue, Gaithersburg MD 20877, CODATA-VAMAS Working Group on Nanomaterials, Paris, France; [11]IdeaConsult Ltd. 4 A. Kanchev str. Sofia 1000, Bulgaria; [12]School of Chemical Engineering, National Technical University of Athens, 9 Heroon Polytechneiou Street, Zografou Campus, Athens, 15780, Greece; [13]Nordic Quantum Computing Group AS, Oslo Science Park, P.O. Box 1892 Vika N-0124 Oslo, Norway and [14]Virginia Commonwealth University, Department of Chemical and

Life Science Engineering, 601 West Main Street, P.O. Box 843028, Richmond, Virginia 23284-3028

Email: Egon L. Willighagen - egon.willighagen@maastrichtuniversity.nl; Sharon Gaheen - gaheens@mail.nih.gov; Sandra Karcher - SandraKarcher44@gmail.com

* Corresponding author

Abstract

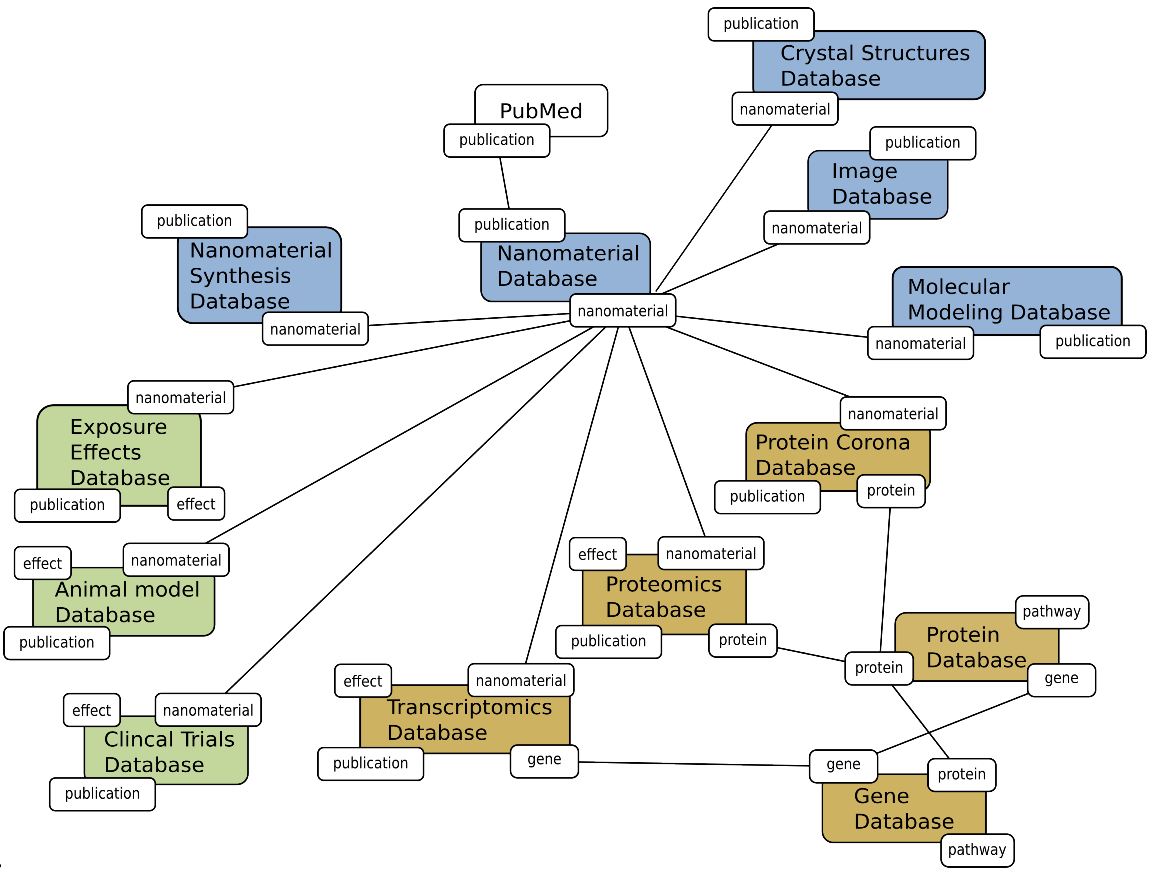
This paper is one of a series of articles by the Nanomaterial Data Curation Initiative. Other articles in this series discuss data curation workflows, data completeness and quality, curator responsibilities, and metadata. Many groups within the broad nanotechnology field are already developing data repositories and tools driven by their individual organizational goals. Integrating these data across disciplines, and with other non-nanotechnology resources, can support multiple objectives by reusing the same information, and can serve as the impetus for novel scientific discoveries through deeper data analyses. This article, framed around the results of a community-based survey of organizations that maintain nanomaterial repositories, discusses current data integration practices in nanoinformatics and in mature fields such as genomics, as well as nanotechnology-specific challenges impacting data integration. Recommendations for achieving integration of existing operational nanotechnology resources, as based on results from the community-wide survey, are presented herein. Nanotechnology-specific data integration challenges, if effectively resolved, can foster the application and validation of nanotechnology within and across disciplines.

Keywords

nanotechnology, nanoinformatics, integration, databases, web services

Introduction

Understanding and addressing complexities involved in integrating nanomaterial and non-nanomaterial data resources to further and enable scientific research is a key focus of nanoinformatics [1]. This paper is one in a series of papers focusing on different aspects of nanoinformatics produced from the Nanomaterials Data Curation Initiative (NDCI), which is part of the National Cancer Institute (NCI) Nanotechnology Working Group [2]. Other articles in this series discuss issues such as data curation workflows [3] and data completeness and quality [4]. The focus of this article is on the integration of databases and data sets across nanotechnology and non-nanotechnology resources. The conceptual integration of resources is shown in Figure 1, with databases shown in large boxes, links shown as lines, and some of the database content shown as corner boxes.



.

Figure 1: Conceptual Integration of Nanotechnology and Non-Nanotechnology Resources.

Figure 1 shows how nanomaterial data repositories can be integrated with other databases to enable interdisciplinary decision making. All repositories containing nanomaterial information are shown with a nanomaterial corner box and are linked with other repositories by nanomaterial, though they often are not specific to nanomaterials. For example, Gene Expression Omnibus (GEO) and ArrayExpress [5] are examples of gene database repositories. Sometimes the genes included in these databases are the focus of studies performed using nanomaterials. The results of those studies may be reported in another database, but the data can be linked using the database content relating to the gene. For example, Figure 1 shows the gene database connecting with the transcriptomics database through information about the gene. It should be noted that the boundaries are not always as clear-cut as indicated in this conceptual diagram, and in reality, there will be many more links than are shown here.

The NDCI is currently working to define nanoinformatics and is exploring the role of data integration as an essential component within the field. The following working definition (expanded from the Nanoinformatics 2020 Roadmap [6]) has been proposed: "Nanoinformatics is the science and practice of determining which information is relevant to meeting the objectives of the nanoscale science and engineering community, and then developing and implementing effective mechanisms for collecting, validating, storing, sharing, analyzing, modeling, and applying the information. Nanoinformatics further involves confirming that appropriate decisions were made and that desired mission outcomes were achieved based on this information. Additional steps in the informatics life cycle including conveying experience to the broader community, contributing to generalized knowledge, and updating standards and training." [7] Successful nanoinformatics endeavors, including data integration, will apply to all of the steps in the process.

The structure of the rest of this paper is as follows. The article begins with an introduction (Section) that discusses why data integration is important and describes common practices for achieving integration. Using the results of a community-wide stakeholder survey, the current practices for integrating data in nanotechnology are presented (Section), followed by stakeholder identified challenges to integration (Section) and a brief description of integration needs (Section). Stakeholder recommendations are reviewed (Section) and the authors' recommendations presented (Section). The article concludes with a few closing remarks by the authors (Section).

Integration of Databases and Data Sets

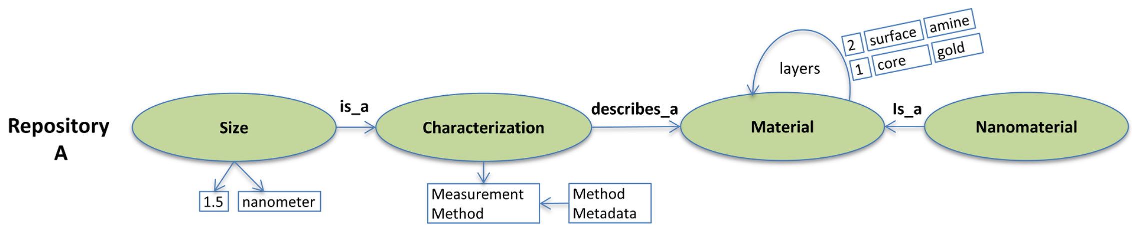
A. Importance and relevance of the integration of databases and data sets to the field of nanoinformatics

Nanomaterials [8,9] are becoming ubiquitous in science and technology [10,11]. Biomedical researchers are making multifunctional nanomaterials that can be used to diagnose, target, and treat many diseases, especially cancer, looking for ways to increase nanomaterial stability and optimize nanomaterial performance while minimizing potential negative effects [11]. Other researchers are harnessing the same useful properties of nanoscale materials for a host of other applications ranging from energy storage to water treatment to improved mechanical strength and flexibility of advanced materials [12].

In order to design an optimal nanomaterial and predict how the nanomaterial will behave, researchers review numerous publications and query disparate nanomaterial repositories across the biomedical, environmental, health and safety, and materials science disciplines. Where the composition of a nanomaterial is provided in a publication and in repositories, the nomenclature used to describe the base nanomaterial formulation, the material constituents (such as core, coat, shell, and any surface modifiers), and the relationships among components are not standardized and mostly incompletely described. For example, the surface density of "decorator" molecules on carbon nanotubes may not be provided, resulting in the need for simplifying assumptions when preparing representative structure files for computational modeling [13].

When storing data in a repository, the selection of the storage format and the design of the database structure is often targeted to meet specific objectives. One method used to organize and store data is as nodes and edges, where data are stored in the nodes and relationships are defined by the edges (the lines connecting the nodes). Another method of organizing data is in tables as columns and rows (fields and records), where insight regarding relationships is built into the structure of the tables.

Some organizations have invested heavily in standardizing the content of databases used by their affiliates. For example, The National Institutes of Health has created an extensive repository of cancer data standards (caDSR) containing appropriate vocabulary and metadata content to "ensure the longevity and agreeability of biomedical research data" (https://wiki.nci.nih.gov/display/caDSR/caDSR+Wiki). Other organizations, such as a group of experimental researchers who are trying to combine their data as part of an integrated study, do not have the funding or expertise to design and plan for multi-repository standardization. Their repository may be an Excel spreadsheet that holds their combined data. These two types of data repositories are represented in Figure 2.



Repository B

Nanomaterial_Identifier	Core	Shell	Surface	Size
John's gold_amine_coated_1_5	gold	none	amine	1 to 3

Figure 2: Showing the challenges associated with determining if a nanomaterial is the same when described in different repositories.

In the case of repository A, a nanomaterial is composed of materials, in layers. The number of layers is not limited and the description and order of each layer are attributes of the material. The size of the nanomaterial is measured using a specific method, and the associated method metadata is included as attributes of the measurement method. Size is reported as a single value and a unit. In the case of repository B, a nanomaterial is described by a maximum of three layers, a core, as shell, and a surface. No method information is provided in the repository and the size is described as a range. No size unit is given. When data curators of repository B are confronted with the challenge of integrating data from repository A, they face two fundamental challenges: (1) determining the uniqueness and equivalency of the nanomaterials [14] and (2) assessing the value in incorporating data from repository B into repository A.

Physical-chemical characterization information (such as the size, shape, purity, and surface properties) is sometimes included in repositories, but the methods and techniques used to perform the characterization are not always included in sufficient detail or standardized in a way that will allow for cross-study comparison of reported values [15]. Each repository collects and stores information in support of their organization's needs and goals. Some repositories may include the results of experimental studies focused on biomedical research, whereas others may include geospatial information on the fate of nanomaterials in an environmental system. Some repositories focus exclusively on nanomaterials, such caNanoLab, which houses information related to biomedical nanotechnology research. Other repositories, such as the Mouse Genome Informatics (MGI) [16], ArrayExpress, WikiPathways [17], contain information that is not specifically related to nanomaterial research [18]. These disparate data, when integrated together, may provide additional insights into understanding common endpoints such as nanomaterial toxicity or stability [19].

Consider a scenario where a species of mice have been injected with a specific nanomaterial in an *in vivo* laboratory study and the same species were exposed to the same nanomaterial during a mesocosm study. Integrating genomics data with the results of both studies illustrates one of many real world use cases that can benefit from the integration of nanomaterial-specific resources with other relevant resources (e.g. genomics data, clinical trials management systems, chemical repositories) that may not necessarily be specific to nanomaterials. Because of the current lack of standardization and integration of resources, researchers must review documentation describing the protocols for storing information in each repository, and sometimes retrieve and review copious publications to determine what is and what is not relevant to their research. This process is time consuming and redundant. The ability to fully integrate repositories across disciplines would allow rapid association of all relevant experimental results to a specific nanomaterial and could help optimize allocation of resources, for example, integration could enable the prediction of adverse clinical results a priori, allowing resources to go to studies that show a more promising outcome.

As outlined above, multidisciplinary fields are particularly demanding on data integration efforts. In all domains, not just nanotechnology, data integration requires a common language (e.g. ontologies) , as well as standards (formal and de facto) for data exchange, communication channels, and identifiers, among others. New technologies have repeatedly changed technical approaches to data integration. While a paradigm based on central data platforms still predominates [20,21], the wider data integration community often uses a more distributed, more-easily scalable cloud [22,23] and other methods, based upon federated search approaches [24,25].

A key requirement for an integration effort is a shared system that crosslinks among databases. This can be based on database identifiers. In disciplines close to nanotechnology, efforts such as identifiers.org [26] unify how identifiers are represented, and other systems provide solutions for mapping identifiers from different databases [27-29]. Identifiers, however, typically focus on entities studied, such as chemicals, materials, genes, and proteins, but identifiers for cell lines, assays, and other key entities involved in nanosafety data are less common, though ontologies commonly provide identifiers for them [30-32]. Moreover, nanomaterials are not as well-defined as small compound chemicals.

Linking data enables data integration; by integrating data sets, data comparisons are enabled. Linking does not define, of course, which data need to be compared or which data can be connected. Decoupling data integration into two steps, linkage and comparison, allows formalization of a hypothesis into a query. For example, linking two nanomaterial data resources, one containing clinical data and the other embryonic zebrafish toxicity data, by identifying records across both resources as being related to the "same" nanomaterial, allows for a hypothesis (e.g. "toxicity towards embryonic zebrafish is of clinical relevance") [33] to be converted into a query (e.g. "report all nanomaterials where high toxicity with respect to embryonic zebrafish corresponds to a high toxicity in a clinical setting, as a fraction of all nanomaterials with both kinds of data") which compares data retrieved for two endpoints for the same nanomaterial. This approach becomes increasingly powerful if links are made between entities, e.g. nanomaterials, even if they are not identical (the same identifier), but show the same chemical or biological characterization for endpoints of interest, i.e. are functionally equivalent (basically the difference between "the same" and "a close match"). An example of being "the same" would be two databases with data on a nanomaterial from a single paper identified with the same label (see Table 1). An example of "a close match" could be two titanium oxides from the Joint Research Centre Institute for Reference Materials and Measurements (JRC IRMM) with the same vendor identifiers. While having the same identifier, they might not be functionally equivalent, depending upon the extent to which the endpoints of interest were affected by aging, etc [34,35].

Table 1: Levels of Equivalence. The equivalence strengths are intended to indicate how data are intended to be combined, and does not specify why it is that it should be linked like that.

Equivalence Strength	**Semantic Equivalence**	**Description**	**Example**
Strong	Web Ontology Language (OWL) sameAs	Two nanomaterials that share the same properties: all properties for one are valid for the other. Moreover, if one nanomaterial is sameAs with others, the others are equally strong (transitivity).	An example would be the same nanomaterial from a journal article for which information is given in two databases.
Moderate	Simple Knowledge Organization System (SKOS) closeMatch	Two nanomaterials are said to be the same for a certain application. This match is never transitive.	An example could be two nanomaterials from the same production batch, in which the application ignores variation.
Weak	SKOS relatedMatch	Two nanomaterials are merely linked together, with an undefined similarity.	This can link two titanium oxide nanomaterials of different sizes.

A formalization of this approach in terms of Semantic Web technologies has been recently proposed through the introduction of lenses that allow users to turn on and off such equivalents based on which links they deem suited for their research question [36,37]. This approach merges the worlds of ontologies and data, by using Internationalized Resource Identifiers (IRIs), such as that found in the set of Semantic Web technologies [38,39]. The Open PHACTS project has taken this approach and developed an Identifier Mapping Service (IMS) that links databases using IRI-based identifiers [36]. Services such as identifiers.org and the IMS itself provide routes to convert between alphanumeric identifiers (e.g. CHEBI:33128) and IRI-based identifiers (http://purl.obolibrary.org/obo/CHEBI_33128) as defined in the ChEBI ontology [40]. Once these links are operational, allowing comparison of data for a set of similar or identical materials, the cross-comparison can be used for automated data curation. During curation, automated comparisons could be enabled to automatically generate warnings that point the user towards other studies reported in other data sources that contradict those being curated. Assuming the linking and subsequent steps leading to the generation of such a warning are correct, the linking could allow researchers of the earlier study to be automatically notified that new, related data have been added to the database. Data integration for identical (or sufficiently similar) nanomaterials also enables a variety of goals to be achieved that are specific to a particular organization.

B. Influence of organizational purpose and goals on data integration

The approaches taken by an organization or project to gathering and organizing data are governed by the driving scientific questions that need to be answered in order to further its mission. Some examples of use case scenarios that could benefit from multidisciplinary data integration are shown in Figure 3.

Data that are measured, the information derived from those data, and the level of detail targeted for inclusion in a resource are all informed by the purpose for which data are being collected. Such purposes include building an authoritative repository of nanomaterial characterizations, parameterizing models to predict nanomaterial behavior in environmental systems, or improving performance of materials, medicines, or pesticides. The goals of the individual resource also shape the type of data integration of interest, with each project incentivized to link with other data sets to increase the critical mass of data in support of its mission. The vision of the nanoinformatics field is that, beyond achieving individual project goals, the potential exists for broadly-integrated data sets to yield unexpected insights from deeper data mining, generating new hypotheses and knowledge

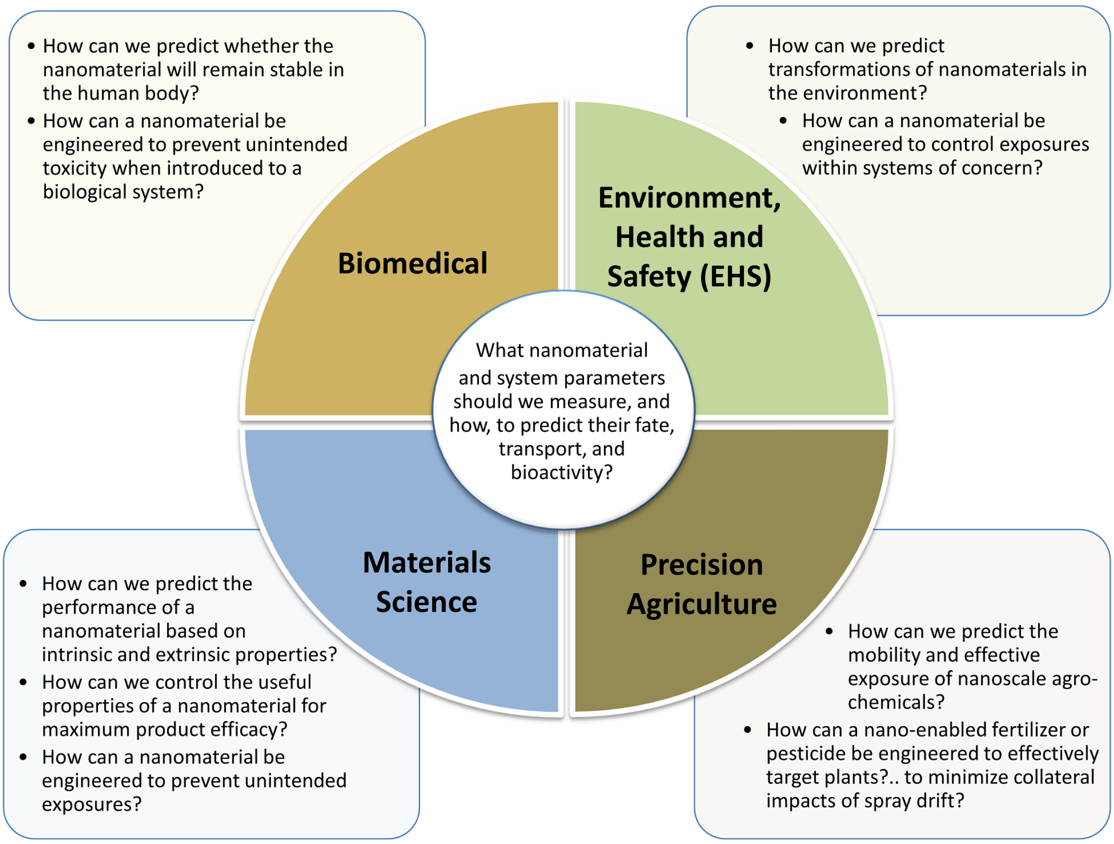


Figure 3: Examples of use cases that can be addressed, and that might mutually benefit through data integration.

not anticipated by the originating data resources and benefiting multiple stakeholders. To realize these secondary benefits of integration, individual projects and disciplines participating in integration efforts must see improvement in their ability to meet their own objectives. Use cases for data integration efforts should therefore be selected such that the different driving forces behind their informatics interests are mutually advanced.

As an example, consider the overlap of interests among biomedicine, materials science, precision agriculture, and environmental, health, and safety (EHS) research as illustrated in Figure 3. Each field pursues research on its discipline-specific questions. Yet at the intersection of these fields is a common kernel of questions and answers that would advance each individual research field as well as open new vistas on a multi-disciplinary basis. Furthermore, by looking across all four disciplines, data integration potentially positively affects the entire data life-cycle, from experimental design through data sharing.

Such use cases can guide initial pilot projects for nano-specific data integration, recognizing the direct near-term value to participating projects, as well as demonstrating the benefits data sharing brings to measurements outside the domain in which they were made. For example, a biomedical nanomaterial data repository integrated with other nanomaterial data resources relevant to biomedicine (e.g. toxicity) and non-nanomaterial data resources (e.g. gene expression and biomedical images) would open interesting pathways to finding effective safe disease treatments. Integrating data from different data resources for equivalent nanomaterials supports multiple goals specific to diverse organizations or projects [14]. Using the example provided in Figure 3, understanding which parameters control stability of a nanomedicine in the human bloodstream could provide insight when predicting nanomaterial dissolution or aggregation in a body of freshwater, transport within a crop field, or efficacy in a material fabrication process. Other examples of potential mutually beneficial integration projects include the following:

- Calculation of a therapeutic index by integrating data from toxicology and clinical studies (http://bioportal.bioontology.org/ontologies/NCIT?p=classes&conceptid=http%3A%2F%2Fncicb.nci.nih.gov%2Fxml%2Fowl%2FEVS%2FThesaurus.owl%23C18223)
- Development of computational models for predicting nanomaterial effects, using consistent physico-chemical measurements in well-characterized media, based on integrating data from physico-chemical and biological characterizations studies, where common nanomaterials might be established based upon product names and/or batch identifiers [41-43]
- Predictions of nanomaterial transformations and effective exposures made through integrating fate and transport data across a variety of systems of interest (bloodstream, aerosolized irrigation stream, polymeric matrix, water column, sediments) with implications for establishing treatment efficacy, product performance, and collateral toxicity [44].

The power of these and other use cases will hopefully pique sufficient curiosity to start significant integration projects. Steps toward integration might then begin with developing an understanding of the respective minimal data standards. Understanding these resource-specific data requirements similarly requires an appreciation for the driving purpose of the resource. It has been suggested that a method for capturing minimum data requirements by resource (e.g. MIAME for microarray data [45]) supports the resource categorization and provides a greater understanding of data requirements. One possible candidate for such a metadata resource is the BioSharing platform (https://www.biosharing.org/) [46]. Further discussion of data and metadata requirements and their explicit documentation via minimum information checklists is presented in an earlier article in the NDCI series [4]

C. Established methods for the integration of databases and data sets

A variety of different approaches have been developed to integrate data, supported by a variety of different kinds of technologies, ranging from manual integration within an Excel spreadsheet (e.g. based on "VLOOKUP" matching of identifiers) to a federated search architecture based on semantic web technologies (https://www.w3.org/standards/semanticweb/) [24,25]. The focus of this paper is upon approaches that best facilitate the retrieval of integrated data via automated queries (e.g. the data query languages SQL or SPARQL [47]); hence, these latter approaches will frame the following discussion. Nonetheless, it is important to note that, given the preference of many scientists for data collection in Excel, tools that allow for automated integration of manually prepared Excel data sets into queryable databases are of considerable value (https://github.com/enanomapper/nmdataparser) [48].

The extremes of the spectrum with regard to selecting an architecture that will support data integration through automatic querying are data warehousing and federated query [49].

- The data warehousing approach involves loading the content of different data resources into the same physical database. Subsequently the "warehouse" database can be queried, which involves querying all loaded data resources concurrently, with results presented to the user.
- Federated querying is implemented by sending queries to the different data resources at their original locations and presenting the results to the user in one unified view as soon as they are received.

The technology for accessing the data resources may be the same for both approaches, e.g. the data warehouse approach may use extract-transform-load (ETL) procedures, connecting to external data resources via web services and loading the results into the warehouse, while federated querying may use wrappers for accessing several distinct databases residing on the same machine and combine results only when presenting them to the user. Hence, a web service is a method for accessing the data, but its use does not imply anything about the data integration paradigm after data retrieval.

The data warehouse paradigm accomplishes the integration by transforming all the data resources into a physical schema (i.e. tables and relationships for relational databases, or XML schema, etc.). The federated query approach relies on a "mediated schema", i.e., a virtual schema, embedded in the application, which does not store any data, but presents to the user a unified view of the domain and allows queries to be specified. The integration itself relies on how the different attributes of the mediated schema match the attributes of the sources, and if the grouping of the attributes corresponds to similar groupings of attributes in the data resources. This is known as "semantic mapping" and is the hardest task within the integration. Regardless of the integration approach, all methods require entity matching (linking associated information based on database content) or mapping (virtually altering the schema of one database so that its content can be queried with data from a database with a different schema). Mapping is typically performed using transformation procedures. There may not exist a simple one-to-one mapping between the final schema and the original data resources. For example, suppose percentage cumulative mortality data were required from the Nanomaterial-Biological Interactions (NBI) Knowledgebase data resource (http://nbi.oregonstate.edu/), in order to include those data with embryonic zebrafish toxicity data curated from the literature [50] in a common data warehouse. Since the NBI knowledge base [51] provides mortality data in terms of the raw numbers of dead/live organisms at 24 hours post-fertilization, and the additional number of zebrafish that were observed to be dead at 120 hours post-fertilization, determining the total number of zebrafish observed to be dead at 120 hours post-fertilization would require mathematical processing before being returned to the user in a schema requiring a field "percentage cumulative mortality" to be populated.

Developing mapping algorithms has traditionally be done manually, however, active research is producing tools for automatic schema mapping and record linkage by deterministic, probabilistic and machine learning methods [52]. In the case of unstructured data resources, e.g. text, the workflow first performs data extraction and entity recognition and then proceeds with the mapping.

Between the two extremes of data warehousing and federated query, many hybrid architectures exist combining elements of both pure data warehousing and federated querying. The choice of integration architecture depends on:

1. How the entities can/will be matched across databases.

2. How the query results will be integrated.

Federated searching can be illustrated with an application to query several online chemical databases for small molecule chemical compound properties via an Application Programming Interface (API) and presenting integrated results on a single web page. Here the entities are the chemical structures, and the IUPAC International Chemical Identifier (InChI, http://www.inchi-trust.org/) [53] can be used as a uniform identifier across databases. The matching rule is "if the search results returned include one and the same Standard InChI, then the results are for the same compound". A data warehouse implementation would use the API to retrieve the results, store them into a database, and then allow the user to query the database. A federated approach would use the API to retrieve the results and present them in a unified format to the user. Although this example may seem straightforward, there are number of complexities that must be considered when matching based on an InChI. For example, small molecule chemicals which may be considered the same, yet correspond to rapidly interconverting structures, may still fail to match based upon InChIs. Whilst InChIs are designed to be invariant to different ways of representing chemicals based on small molecular structures, including taking into account tautomeric forms which are expected to rapidly equilibrate, they cannot account for all differences in chemical structure which may readily interconvert in practice - such as differences in protonation state (e.g. salicylic acid will exist in dynamic equilibrium with its deprotonated form under physiological conditions) or between open-and-closed ring forms, which can equilibrate for sugars in solution. If non-standard InChIs are used, the situation is further complicated [53]. In spite of the challenges discussed here, integration of small molecule chemical databases based on matching their Standard InChIs is currently viewed as best practice and may be combined with other software tools to enforce further standardization of chemical structures that may facilitate desired matches [54].

Extending this approach to more complicated structures, e.g. proteins and genes, would require expanding the queries to handle all possible synonyms used by different databases.

Establishing a common API for a given type of resource facilitates integration because it alleviates the need of schema matching. Essentially, the API defines a common schema and if all resources of the same kind are compliant with the API, the main hurdle of semantic mapping is lifted. An example implementation of this approach in the genomics field is the Global Alliance for Genomics and Health (GA4GH) Data Working Group (http://ga4gh.org/#/), which is establishing common web services in support of genomic data integration and exchange. Example web services using the Representational State Transfer (REST) framework [55] are provided with query requests and responses formatted using the JavaScript Object Notation (JSON). The common web services allow the genomics community to exchange reads, variants, and reference information, provided all data resources follow the API specification.

The implementation of a central data warehouse or repository that aggregates data from several resources requires extract, transform, and load (ETL) processes to assist in aggregating and transforming the data based on matching rules. Data are typically transformed into a common data model (e.g. relational database or a triple store); examples of this approach are PubChem and ChEMBL databases. The Open PHACTS project provides a common API to a variety of pharmacological data sets. However, it does not normalize to a single data model, but addresses the non-uniformity at the API level [20]. The European Bioinformatics Institute Resource Description Framework (EBI-RDF) platform uses a different approach, maintaining multiple RDF repositories for different resources and allowing federated searching across all of them [23]. It is mandatory for all of the entities in the EBI-RDF platform to be assigned equivalent identifiers via identifiers.org service, which is essentially implementing the mapping between the distributed resources.

The Syngenta federated search system [25] is an example of addressing the challenge of integrating internal company data with public life science databases. The system has moved from data warehousing (even if that offers faster reporting) towards federated search technologies. The architecture includes several internal relational database repositories, translated into RDF dynamically via D2RQ (http://d2rq.org/) [56], and providing adapters in order to combine all internal and external data resources into a distributed SPARQL endpoint. The implementation of this federated architecture for data integration was found to offer clear benefits to Syngenta's chemists and biochemists.

Current practice for data integration in the nanotechnology field: perspectives of key stakeholders

To understand the current practices in data integration and to identify challenges and offer recommendations, several organizations that maintain nanomaterial repositories were asked to respond to a questionnaire on data integration. The goal was to assist in defining and initiating integration and exchange of data resources across nanomaterial data repositories and with other non-nanotechnology data resources. Questions included current and recommended functionality and web services enabling data integration and exchange as well as perceived challenges associated with integrating primary experimental data sets, or data sets curated from the literature, with existing nanomaterial and non-nanomaterial data repositories. The following sections provide details on the organizations who participated in the survey along with summarized results of their feedback. Information on each nanomaterial resource is provided in Table 2.

Table 2: Nanomaterial Resources Responding to Data Integration Questionnaire

Nanotechnology Resource	Resource Description	Integration Capabilities
caNanoLab - caNanoLab Data Portal (https://cananolab.nci.nih.gov/) caNanoLab Wiki (https://wiki.nci.nih.gov/display/caNanoLab/caNanoLab+Wiki+Home+Page) caNanoLab Data Dictionary Resources: caNanoLab Glossary (https://wiki.nci.nih.gov/display/caNanoLab/caNanoLab+Glossary) NCI Thesaurus (https://ncit.nci.nih.gov/ncitbrowser/pages/home.jsf?version=15.05d) NCI caDSR (https://cdebrowser.nci.nih.gov/CDEBrowser/) Design Document with Domain Model caNanoLab Code Repository (https://github.com/NCIP/cananolab)	caNanoLab is a data sharing portal designed to facilitate information sharing across the international biomedical nanotechnology research community to expedite and validate the use of nanotechnology in biomedicine. caNanoLab provides support for the annotation of nanomaterials with characterizations resulting from physico-chemical, in vitro and in vivo assays and the sharing of these characterizations and associated nanotechnology protocols in a secure fashion.	Provides REST-based web services supporting general sample search and retrieval of sample composition and characterizations by sample ID. Supports retrieval of samples associated with a publication. Integrates with ScienceDirect publications through an Elsevier bi-directional link and uses the PubMed and PubChem interfaces.

CEINT CEINT Wiki (http://www.ceint.duke.edu/)	CEINT is a center-wide effort focused on exploring the potential impact of exposure to nanomaterials on ecological and biological systems. The center is funded by the National Science Foundation and the US Environmental Protection Agency, and brings together researchers from several universities, NIST, the EPA, as well as other key domestic and international partners. CEINT supports fundamental research regarding the behavior of nanomaterials in laboratory studies and also in complex ecosystems. One of the goals of the center is to develop a web-based risk assessment tool that can be used to elucidate the potential risk associated with the release of nanomaterials into the environment.	Integration within the CEINT NIKC resource is achieved by custom API development for each collaborative project with targeted data sets.
Center for Safety of Substances and Products, National Institute for Public Health and the Environment (CSSP/NIHE) Netherlands Center Information (http://www.rivm.nl/en/About_RIVM/Organisation/Centres/Centre_for_Safety_of_Substances_and_Products) Software Model for Estimated Exposure from Consumer Products (http://www.rivm.nl/en/Topics/C/ConsExpo) Nanotool for Spray Products (http://www.rivm.nl/en/Topics/C/ConsExpo/Nano_tool)	The CSSP NIHE provides a database on ecotoxicity data focusing on nanoparticles in consumer products. The database provides a repository for modeling purposes (QSAR).	Does not provide any web services. In case of gathering/uploading toxicity data, the OCHEM database is commonly used. The database also allows for modeling and selection of descriptors.

DECHEMA http://www.dechema.de/en/ Nano-safety Wiki (http://www.nanora.eu/nano-safety)	DECHEMA is a network of experts in chemical engineering and biotechnology. DECHEMA supports several projects applicable to nanotechnology such as the DaNa project and the NANORA project [57]. DaNa is a Knowledge base of applied nanomaterials on health and environment. The NANORA project provides web facilities supporting the Nano Region Alliance, an alliance that facilitates market entrance for nanotechnology subject matter experts.	The DaNa project has been providing the web service for the NANORA project to implement the Danavis Database on the NANORA website based on JSON as data exchange format.
eNanoMapper Ontology http://bioportal.bioontology.org/ontologies/ENM Database https://apps.ideaconsult.net/enanomapper/ Search https://search.data.enanomapper.net Modeling http://enanomapper.net/modeling	eNanoMapper is a European FP7 project of eight research and industry institutes. The aim is to improve data integration and to support safe-by-design development by building up a nanosafety ontology, a database and provide tools for use of this data (e.g. modeling approaches).	There is a REST-based API and nanomaterials have URIs allowing a linked data approach. External databases can be indexed by uploading, for example, nanomaterial characterization or via search integration.

Nanomaterial Registry Websites: http://www.nanomaterialregistry.org Partner Portal at nanoHUB	The Nanomaterial Registry is a publicly-available database of nanomaterial characterization and biological/environmental interaction data. Data in the Registry are curated from niche databases, literature, catalogs, and reports by trained scientists. Data are curated based on a set of minimal information about nanomaterials. The data of the Registry are also available on the Portal at nanoHUB, where predictive modelers can find the data in a format that is easy for them to use.	Integration with the Registry is achieved on a case by case basis. Future development will include a JSON interface for analysis tools and data submission templates.
Nanoparticle Information Library http://nanoparticlelibrary.net/	The NIL is a prototype searchable database of nanoparticle properties and associated health and safety information designed to help occupational health professionals, industrial users, worker groups, and researchers organize and share information on nanomaterials, including their health and safety-associated properties.	Integration with the NIL is achieved on a case by case basis.

A. Stakeholder demographics

Stakeholders who participated in the survey ranged from nanomaterial resources that have extensive experience in integrating databases and data sets to those with limited data integration experience whose focus was primarily on repository development (Table 2). The diverse levels of integration capabilities provide insight into the challenges that need to be addressed in order to integrate across nanomaterial repositories and with other non-nanotechnology resources.

B. Stakeholder experience in nanomaterial data integration

The surveyed nanomaterial data resources exhibited a variety of experience in data integration including integrating primary data sets and web services supporting data integration. Stakeholders were asked for information on existing resource functionality supporting data integration including data standards, controlled vocabulary, and common identifiers. They were also asked to identify available web services supporting cross-nanomaterial resource exchange and current efforts supporting integration with non-nanotechnology resources.

Uploading / Downloading Data Sets

When using a data warehousing architecture, the ability to upload and download data sets is an initial step towards integration as support for this feature requires the identification of data formats and representation of common data elements. Federated approaches may not require the actual movement of the data, but also requires identification of data formats and common data elements. Stakeholders responded to questions relating to integration of primary data sets, including services available in-house or services that are publicly available (Table 3). These stakeholder experiences provide insights into the level of readiness the nanotechnology community has achieved with regards to integrating databases and data sets.

Web Services Supporting Data Exchange

The missions of the stakeholder groups are highly diverse, with web services being of high priority for some and not for others. The data exchange capabilities of each resource, as provided by each stakeholder, are summarized in Table 3, and capabilities relating specifically to web services are described in the following section.

Table 3: Summary of Stakeholder Responses to Upload, Download, and Mapping Questions: Does the nanomaterial data resource provide the following?

	Does the nanomaterial resource provide the following?				
Nanomaterial data resource	**Uploading, downloading, or mapping**	**Definitions of the database fields**	**Controlled vocabularies, taxonomies and/or ontologies**	**Nanomaterial identifier uniqueness**	**Integration with any non-nanotechnology resources**
caNanoLab	web-based forms for uploading and downloading nanomaterial composition, characterizations, publications and protocols	extensive documentation is available[a]	uses NPO and the NCI Thesaurus (http://ncit.nci.nih.gov/ncitbrowser/pages/home.jsf?version=15.05d)	Uses a pattern containing source information and a numeric identifier resulting in a unique identifier. The pattern for the sample name is: abbreviation(s) of institution names, name of the first author (without middle name), custom abbreviation of journal title, year of publication, and sample sequential number, e.g. SNL_UNM-CAshleyACSNano2012-01.	caNanoLab integrates loosely with six non-nano resources[b].

CEINT	mapping from NBI data set	not yet	uses ontologies such as MO, NPO, UO, and ChEBI	nanomaterial associated to data source and assigned a unique identifier	not currently
CSSP/NIPHE, Netherlands	commonly use the OCHEM database for uploading toxicity data	provides a list a fields available for storing toxicity data	uses field headings as a means of controlling vocabulary	identifier assigned based on particle core composition	no
DECHEMA	no	relational model documented in Kimmig et al. [58] and Atli et al. [59]	uses the scientific wording for materials and nanomaterials, toxicology, biology[c]	not a central issue of the DECHEMA work	no
eNanoMapper	extends the OpenTox platform which has the means to download and upload data	overview of the data model documented in Hastings et al. [40]	uses the eNanoMapper ontology (composed of NPO, ChEBI, BFO, IAO, CHEMINF and others)	uses an IUC substance UUID[d]	not currently
Nanomaterial Registry	export for physico-chemical characterization	Nanomaterial Registry glossary (https://www.nanomaterialregistry.org/resources/Glossary.aspx)	uses a controlled vocabulary[e]	uses unique numeric IDs[f]	not currently
Nanoparticle Information Library	Accomplished on a case-by-case basis	Provided as drop-down lists of available fields	Uses the NPO as well as user-specified terms	Unique NIL entry numbers are assigned	The NIL integrates directly with data resources on hazardous materials [60][g].

[a]The caNanoLab Design document (https://github.com/NCIP/cananolab/tree/master/docs/design) includes the object model which represents class names and attributes associated with the data model. All class names and attributes are maintained in the NCI caDSR (https://cdebrowser.nci.nih.gov/CDEBrowser/). Concepts are defined in the NCI Thesaurus (http://ncit.nci.nih.gov/ncitbrowser/pages/home.jsf?version=15.05d). caNanoLab also provides a user-friendly glossary (https://wiki.nci.nih.gov/display/caNanoLab/caNanoLab+Glossary).

[b]caNanoLab integrates with PubMed and ScienceDirect for access to publications, Elsevier for linking caNanoLab data to publications, PubChem for chemical information, The Collaboratory for Structural Nanobiology - CSN (http://uqbar.ncifcrf.gov/Advanced_Structure_Analysis/HOME.html) for displaying 3D models of specific nanomaterials, and Nanotechnology Characterization Laboratory (NCL, http://ncl.cancer.gov/working_assay-cascade.asp) assay cascade and JoVE (http://www.jove.com/) for nanotechnology protocols.

[c]DECHEMA has a very diverse target group ranging from interested laymen, stakeholders to other scientists; wording is adjusted in order to tell a comprehensive story without confusing the laymen on the one and hand and not losing the scientific correctness.

[d]eNanoMapper is based on semantic web technologies including dereferenceable Internationalized Resource Identifiers (IRIs) and the Resource Description Framework (RDF). The substance UUID does not reflect the uniqueness of the material structure, but is an identifier of the material in the database. The substances (materials) are described with their composition (e.g. core, shell, and functionalization) and are linked to the chemical structures of their components. These can be used to decide if the nanomaterials are the same or similar.

[e]The NPO has been mapped to the Nanomaterial Registry and it was determined that a little over 80 terms used by the Registry are not yet part of the breadth of the NPO.

[f]It is the intent of the Nanomaterial Registry not to judge equivalence between any two nanomaterials from different data resources, as the characterization results can be wildly different based on sample medium and characterization protocol.

[g]The NIL integrates with the NIOSH Pocket Guide to Chemical Hazards (NPG, http://www.cdc.gov/niosh/npg/) and with the Registry of Toxic Effects of Chemical Substances (RTECS, http://www.cdc.gov/niosh/rtecs). The current hosting, administration, and maintenance of the NIL web resource outside of the CDC/NIOSH website is being conducted by Oregon State University in conjunction with its program to characterize nanomaterials.

caNanoLab Web Services

caNanoLab implements an internal and external API leveraging REST (see Table 4). The internal API retrieves web forms in JSON format, while the external API retrieves web forms in HTML format. caNanoLab exposes web services that retrieve publicly available information. All other web services are used internally and are not exposed. caNanoLab does not publish documentation on web services other than The caNanoLab Design document which documents the system architecture and object model. Internal web services are based on method calls on object model attributes. Other NCI projects supporting genomics use Apiary for documenting web services. caNanoLab uses the PubMed API to retrieve publications and interfaces with PubChem to retrieve information on chemicals associated with nanomaterial composing elements.

Table 4: Web Services provided by caNanoLab (https://cananolab.nci.nih.gov/caNanoLab/#/)

Search Type	Possible Search Criteria	Notes and Links
protocol	protocol name	https://cananolab.nci.nih.gov/caNanoLab/\# /searchProtocol
sample	specific sample, composition, and/or characterization	https://cananolab.nci.nih.gov/caNanoLab/\# /advancedSampleSearch. Returns sample information by sample ID.
publication	sample name. nanomaterial characteristics	https://cananolab.nci.nih.gov/caNanoLab/\# /searchPublication Retrieves publication information and associated samples by PubMed ID or DOI.

CEINT Web Services

CEINT does not currently provide web services for data set sharing; however, CEINT does provide a web-enabled service for use by CEINT members that allows them to connect with other researchers who identify as working on the same research questions, with the same materials, and with the same methods. This service facilitates Center-wide data integration through direct upstream collaboration, even in the absence of prescribed data templates that would support more automated integration. CEINT uses web services provided by others, included the Nanomaterial Registry, the Integrated Taxonomic Information System, Ontobee, caNanoLab, USDA Geospatial Data Gateway, and the Project on Emerging Nanotechnologies.

CSSP/NIPHE, Netherlands Web Services

The Center for Safety of Substances and Products, National Institute for Public Health and the Environment, Netherlands does not offer web services; however, the OCHEM database is publicly available.

DECHEMA Web Services

DECHEMA does not provide any web services per se for the DaNa project. In the case of the NANORA project, a web service was specifically created, together with an interface to implement the DaNaVis database on the NANORA website using JSON as the data exchange format. The backend web services and customized interface for the NANORA website are not publicly available but the frontend user interface is freely accessible. There is no publicly available documentation for the web service for the NANORA project. DECHEMA uses a content-management system for the DaNa website (Joomla + several plug-ins, bootstrap framework). The DaNa website is accessible for everyone without any usage restrictions. The DaNaVis database and tools use a Django-framework (Python as the programming language), REST API- and JSON-based data interchange between client and application server, client-side JavaScript widget. More details on the database and tool design have been published [58,59]. DECHEMA does not use any web services provided by other organizations

eNanoMapper Web Services

eNanoMapper provides web services based on the OpenTox API. eNanoMapper inherits and, where needed, extends the machine readable API. The supported return formats include JSON , JSON-LD and RDF/XML, CSV, XLSX. Methods exist for a number of entity types, including substances, which is how eNanoMapper models a nanomaterial. The API is REST-like. eNanoMapper separates the API design from the server implementation; AMBIT is one of the reference implementations of eNanoMapper services [61], and on the server-side uses Apache's Tomcat. The API implements user authentication and authorization. This means that an eNanoMapper instance (it is a platform rather than a single system), allows for both public data and confidential data that can be shared with only a selected group of researchers. The example http://data.enanomapper.org/ instance currently hosts several public data sets, available under an Open Data license or waiver.

The eNanoMapper server currently does not use other web services, besides being able to retrieve chemical structures from public databases (e.g. PubChem). However, this may change when eNanoMapper moves towards a more distributed platform later in the project.

The full details of the eNanoMapper API, including a description of the computational services implementation (which uses and integrates a variety of technologies and also reads and writes from/to data services) are published [62]. Interactive API documentation is available online (http://enanomapper.github.io/API/). A webinar using the API to visualize data in web pages is available on YouTube (https://www.youtube.com/watch?v=quy7G2mZ0gk), and a complete list of models that can be used for prediction can be found in the Swagger documentation (http://app.jaqpot.org:8080/jaqpot/swagger/).

Nanomaterial Registry

The Nanomaterial Registry does not currently have data exchange web services other than the export tools described in Table 3. However, a JSON interface is in development for the connection with data analysis tools. The Registry website does provide a web service search tool that allows for keyword and specific measurement values to be searched, as well as allowing the user to browse nanomaterials by a variety of characteristics. Nanomaterial Registry data are also batch exported to a portal at nanoHUB, where users can interact with and download the data in different ways.

Nanoparticle Information Library

The Nanoparticle Information Library website is publicly accessible to everyone with the request that any use of the data be attributed to the primary source associated with the data entry. Online search capabilities within the NIL are based on attributes of nanomaterial structure, elemental composition, method of synthesis, and nanomaterial size-related features including primary particle diameter, agglomerate diameter, and specific surface area. Weblinks to the primary data and to the principle investigators who have provided data to the NIL are included.

Stakeholder identified data integration challenges

Stakeholders identified several technical and operational challenges impacting current data integration efforts, as shown in Figure 4. These challenges, if not addressed, will continue to plague the nanotechnology informatics community and greatly hinder scientific discoveries. Each are discussed in greater detail below.

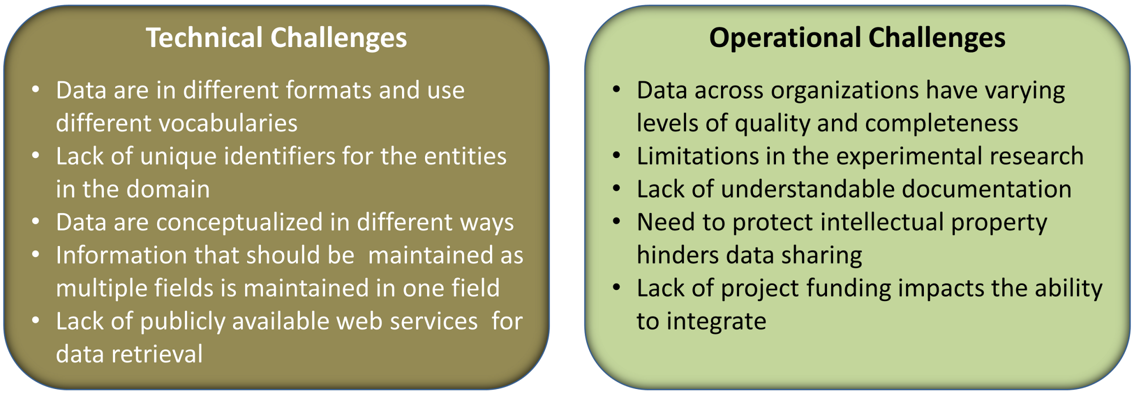


Figure 4: Technical and operational challenges impacting data integration.

Data are in different formats and use different (or no) common vocabularies or ontologies

The primary challenge in achieving data integration, selection, and aggregation in the nanotechnology domain is the diversity of ways in which nanomaterial information is represented across data repositories and the lack of standardization to a common model that represents nanomaterial entities, their attributes, and relationships. These issues include multiple meanings for the same word (or abbreviation) and different words (or abbreviations) having the same meaning. For example, cytotoxicity can have different specific meanings when different bioassays were used to measure it. Similarly, examples of synonyms with the same meaning are abundant too, and include for example, ZnO, zinc oxide, and nano-zinc oxide. Developing appropriate ontologies, including resolution of terminology conflicts, to address the nuances of nanotechnology research are an important key to achieving integration.

Lack of unique identifiers for the entities in the domain

Certain difficult aspects of data integration remain challenging regardless of the specific domain, including deciding when entities (e.g. nanomaterials, cells, samples, people, etc.) in different data contexts should be mapped as "the same" or "different" e.g. if their names have narrower or broader meanings. An example would be the difference of titanium dioxide (NPO_1485) and titanium oxide nanoparticle (NPO_1486) made by the NanoParticle Ontology. This difficulty impacts the ability to perform cross material comparisons. Other fields have introduced naming conventions for generating unique identifiers based on metadata; however the different metadata used across studies made this challenging. As such, other fields such as genomics are moving forward with generating a Universal Unique Identifier (UUID) for entities not based on metadata associated with the UUID to support queries in support of entity comparisons. In the context of nanomaterial data resource integration, metadata might include the results of physico-chemical characterization required to establish whether the nanomaterials are "the same" or "sufficiently similar" to be matched during data integration. However, the question of which physico-chemical properties need to match [15], not to mention complexities associated with different measurement techniques and experimental protocols, make uniquely identifying and matching nanomaterials a significant scientific challenge. Further discussion of metadata (including batch identifiers) that could support unique identification and matching of nanomaterial database records is provided in an earlier article in the NDCI series [4].

"Data" are conceptualized in different ways

There has been a trend away from establishing a fixed hierarchy between database elements; a trend that, in some regards, adds to the data mining challenge. Sometimes expert knowledge is built into the establishment of hierarchical relationships, and that knowledge can be extracted when mining a database to assure that data are appropriately aggregated when performing statistical analyses. Often times, databases are designed to support searching, but not specifically to support mining. In these types of database repositories, measurements are sometimes duplicated so that they can be pulled in many kinds of searches. Duplicate measurements, if not handled correctly during analysis, can lead to bias in statistical computations.

Information that should be maintained as multiple fields is maintained in one field

Integration of data can be hampered by differences in data granularity. A common issue is that information in one repository may be stored in one "field", but be split into multiple "fields" in another repository. Additionally, in some repositories, numerical data are stored without a separate unit "field". For example, some repositories use a field name such as "Concentration" and expect the user to know that the result should always be in a specific unit, such as "mg/l". In other cases, a measured result is combined with a unit and stored together in the same field (e.g. 7 mg/l), or include a range of values in one field (e.g. 7-10 mg/l).

Lack of publicly available web services for data retrieval

Integration is often hindered by the lack of publicly available web services supporting data retrieval. Additionally, even when data services are provided, open frameworks such as REST are not leveraged to ease development of integration touchpoints [55].

Data across organizations has varying levels of quality and completeness

Finding data that are sufficiently complete and of acceptable quality is a key challenge for nanoinformatics. At times data from external repositories are not integrated with local systems due to concerns regarding the quality and completeness of those data. For example, a local knowledge base can implement a screening procedure that carefully selects high quality data from the scientific literature; data from publications not meeting the specific quality criteria are deemed unsuitable and are not curated into the knowledge base. When evaluating external data for inclusion in the knowledge base, if they do not come with an indicator or ranking of the reliability of those data, and if the ranking is not in line with the screening procedure used by the curators, it is difficult to determine if and how those data should be incorporated.

Lack of data completeness also poses a challenge to data integration as it is often difficult to obtain the necessary information to support comparison (a pre-requisite for matching and data integration) between material records in different databases. For example, when obtaining information on physico-chemical characterization, it is important to have information on the chemical composition of the particles, such as the presence/absence of coatings, and if the particle has been transformed. In addition, lack of complete metadata for associated biological tests may be considered to affect the clarity, hence quality, of results [63] and preclude an assessment of whether two sets of results were generated under sufficiently similar conditions to allow them to be meaningfully integrated in support of analysis, for example, the relationships between material characteristics and biological effects. It is also critical to have information on the media properties that might affect the result of (toxicity) testing as well as standardized methods of collecting the information. A lack of proper particle characterization is a key problem [64], and the consequence is that often a database contains more blank fields (no information) than actual data. This lack of high quality and complete data sets discourages integration.

A thorough discussion of the challenges associated with assessing the completeness and quality of nanomaterial data was presented in an earlier paper in the NDCI series [4].

Limitations in the experimental research

There are limitations in the experimental research process, such as biological variance, uniform characterization, and technological and methodological constraints. One major challenge related to data quality and completeness is defining the minimum data requirements for integration. The continuing evolution of knowledge of the important independent variables that must be controlled to make a measurement or assay accurate and reproducible can change these data requirements. As is customary in science, it takes time for new scientific insights to reach every lab, and as with any novel field, nanotechnology is evolving and maturing. This maturing process is evident in the nanosafety field as well as in bioinformatics; the first generation of results may not be optimal, but they must be used as a basis for improvement or the field will not progress. Another major challenge in nanoinformatics is that researchers are continuing to refine measurement techniques, which could change the comparability of measure results over time. These kinds of issues are related to the concepts of data quality and completeness, which were discussed - along with recommendations for progress - in an earlier article in the NDCI series [4].

Lack of usable documentation

The available documentation for external resources often just introduces the resource and provides instructions for its use, but does not convey adequate information to understand the conceptualization behind the database design. A commonly accepted minimum documentation standard would be helpful.

Need to protect intellectual property hinders data sharing

Although data sharing encourages the public to use and exploit knowledge contained in a database, restrictions may be in place to protect intellectual property and investments in generating and updating database content. Often, these restrictions have unclear statements about ownership, copyright, and licensing. Researchers are sometimes reluctant to share data until they are completely done analyzing and reporting their results out of fear that someone will take their data and use it in a way that limits or reduces the novelty of their work [65]. Some have even suggested that those performing analysis on data they had no role in generating are "research parasites" [66]. The need to maintain "unique selling points" of a database can impede data sharing. One solution to overcome this challenge is to provide a web service with restricted accesses in support of data retrieval while maintaining a customized interface to maintain the unique characteristics of the resource.

Lack of project funding

Individual projects to build data resources and repositories usually do not have funding allocated to data integration. Further, it is not clear which people in the management and funding chain are the correct contacts for expanding a project scope to include integration. This is also a primary constraint for driving standardization towards a common model. The funding issues extend beyond the necessity to win monetary support that is shared by all research endeavors because these projects can often be seen as investments in infrastructure or tools and are thus perceived to fall outside the purview of basic science funding. Data projects, however, are actually significant exploratory investigations into scientific questions and not just IT projects. Data resources are a major future source of scientific knowledge, and integration across numerous sources expands research opportunities.

Stakeholder nanomaterial data integration needs

To address key challenges, stakeholders identified the functionality and web services needed to enable data integration across nanomaterial repositories. Stakeholders also identified use case driven integration needs with non-nanotechnology resources.

Functionality Needed to Enable Data Integration across Nanomaterial Repositories

Use of shared controlled vocabularies

To integrate across resources, each resource needs either to adopt shared controlled vocabularies or to be able to map to agreed-upon standards. When mapping between controlled vocabularies, it is important to fully document the mappings and develop tools to assist in the mapping and transformation of the data. Although tool development to automate mapping of terms and schemas requires significant work, time is saved in the long run as standards evolve. Adoption of a common language is important, as well as using open standards for data exchange.

Data search and retrieval by ontological terms

Most nanomaterial resources support basic search and retrieval by nanomaterial, characterization, protocol, and publication. To facilitate search and retrieval across resources, it is necessary for resources to support searching by ontological term. Additionally, search capability should support retrieval of data (e.g. primary particle characteristics) across each nanomaterial resource and retrieval of detailed information from the same source on study endpoints applicable to the resource. For example, in the case of toxicity data, it is necessary to support retrieval of particle fate characteristics during testing as well as information on the test medium. eNanoMapper's search system allows searching using ontologies, taking into account synonyms. The demonstration server at https://search.data.enanomapper.net/ allows simultaneous searching over data collected by eNanoMapper and by caNanoLab.

User friendly web-based data submission forms

Nanomaterial resources should provide user friendly tools supporting the submission of data on nanomaterials, characterizations, protocols, and publications via web-based forms. These forms should constrain data entry by requiring use of a controlled vocabulary.

Data import and export tools

Resources should provide support for the validation, import, and export of data in standard data file formats such as ISA-TAB-Nano [67,68], which would allow data to be exported from one database directly into another. It is understood that the development of such tools would require a significant amount of work for resources not currently supporting standards like ISA-TAB-Nano.

Tools to analyze and visualize data

Data analysis and visualization tools within and across nanomaterial resources will facilitate cross material comparisons. Visualizing nanomaterials in 3D and displaying scatter plots and distribution plots across data would assist in optimizing nanomaterial design. Analytic tools need to support the work of many disciplines, including chemistry, biology, toxicology, medicine, and physics.

Data modeling tools

Data modeling tools assist in predicting nanomaterial behavior in different biological and environment systems. The integration of nanomaterial resources with data modeling tools requires that each resource provide access to sufficiently high quality and complete data sets [4].

Facilities for rating data sets for data quality and completeness

Prior to integrating with an existing nanomaterial resource, it is important to understand the data quality and completeness of the resource. Facilities that rate data for completeness and/or quality can assist in providing this assessment. This may include rating against minimum information as well as feedback from users who try to reproduce those data. However, assessing data completeness and quality is decidedly non-trivial. A thorough examination of this issue is presented in another article in the NDCI series [4].

Data Annotations

It is important that data are clearly annotated with statements such as possible provenance, including ownership and licensing or rights waiving where applicable. Understandably, data can be proprietary, and if so should be clearly marked as proprietary. The growing use of resources, such as ZENODO (http://zenodo.org/) and FigShare (https://figshare.com/), which allow users to assign a specific license to their research data, is arguably indicative of a growing awareness of the importance of clarity regarding rights to data usage within the scientific community - although these resources do not support the application of automated data integration techniques [69]. In addition to annotations on data provenance, data annotations can also be provided to further clarify the quality of the data.

Web Services Needed to Enable Data Integration across Nanomaterial Repositories

Stakeholders supporting the use of nanotechnology in the biomedicine and the nanosafety community indicated that the Biomedical Community needs common web services supporting the exchange of nanomaterials, characterizations, protocols, and publications in support of cross material comparison. By integrating with other nanomaterial repositories supporting biomedicine and with other repositories from environmental and health, the biomedical community hopes to better predict the bio-distribution and toxicity of nanomaterials in model organisms, including humans. Additionally, the biomedical community would like to obtain detailed information on the investigation, studies, and assays based on metadata identified in the ISA-TAB standard. To support data integration, ISA-TAB and ISA-TAB-Nano Application Programming Interfaces (APIs) are under development that retrieve entities based on the ISA-TAB and ISA-TAB-Nano JavaScript Object Notation (JSON) schemas (https://github.com/ISA-tools). The Nanosafety Community has many interests and covers many different scientific domains. But of special interest, at this moment, for linking databases, are web services oriented at two central entities in publishing: most similar nanomaterials, and anything about the same paper or experimental protocol. Common web services envisaged by these stakeholders as being needed to support integration of nanomaterial data in the biomedical nanotechnology and nanosafety domains are presented in Table 5.

Needs for Integrating Nanotechnology Repositories with Non-nanotechnology Resources

Stakeholders identified a variety of non-nanotechnology resources that must be accessed to support use case driven data integration needs; these are summarized in Table 6.

Table 6: Non-Nanotechnology Resources needed to support use case driving data integration.

Non-nanotechnology Resource	**Description**
Life Sciences and Chemistry Databases	Life science and chemistry databases in general, containing information about human biology (both experimental data, as well as knowledge bases) and chemistry (functionality, chemical structure, etc.) [70,71]. Needed to inform the design new nanomaterials to avoid potential negative influences on human health.

Image Archives	Such as the National Biomedical Imaging Archive (NBIA) (https://ncia.nci.nih.gov/ncia/login.jsf), The Cancer Image Archive (TCIA) (http://www.cancerimagingarchive.net/), or other image archive to display MRIs or other image modalities of subjects in which nanomaterials are used for diagnostic and/or therapeutic purposes. A "public domain" image archive illustrating images used in articles, e.g. SEM-pictures would assist in visualizing particle characterizations (see http://www.enanomapper.net/library/image-descriptor-tutorial).
Image Contrast Agent Repository	For example, the Molecular Imaging and Contrast Agent Database (MICAD) (http://www.ncbi.nlm.nih.gov/books/NBK5330/) to obtain information on image contrast agents to compare with nanomaterials used in diagnostic imaging.
Model Organisms Repository	Such as the Mouse Genome Informatics (MGI) (http://www.informatics.jax.org/) resource to access information on animal models used in in vivo characterizations involving nanomaterials.
Publication Sources	PubMed LinkOut or publication vendors such as Elsevier (http://www.elsevier.com/books-and-journals/content-innovation/data-base-linking) to link nanomaterial data to nanomaterial publications. An example of this is the caNanoLab interface with ScienceDirect publications through Elsevier.
Clinical Trials Management Systems (CTMS)	Such as OpenClinica to access clinical data associated with the use of nanomaterials in human clinical trials.
Genomic Data / Biomarker Repositories	Such as the NCI Genomic Data Commons to maintain molecular data for transfection and targeting characterization involving nanomaterials.
Chemical and Agent Repositories	Such as PubChem, ChemSpider, ChEBI, and vendor repositories like Sigma Aldrich to obtain information on chemicals used in nanomaterial compositions. Integrate with small molecule repositories like DrugBank [72] to compare a small molecule (e.g. magnevist) with a nanomaterial formulation that associates with the small molecule (e.g. dendrimer magnevist complex).

Modeling Tools	Modeling and simulation tools as well as 3D structural modeling tools. Integrating with modeling and simulation tools will assist in modeling the effects of nanomaterial size, shape, and other properties on biodistribution and toxicity. Integrating with 3D modeling tools such as The Collaboratory for Structural Nanobiology - CSN (http://uqbar.ncifcrf.gov/Advanced_Structure_Analysis/HOME.html) facilitates the display on nanomaterial structures in 3D leveraging a Protein Data Bank (PDB) file.
Analysis and Visualization Tools	Includes various tools such as R (https://www.r-project.org/) [73], an environment for statistical computing, and Bioconductor [74], D3.js, and other tools to analyze and visualize nanomaterial data in support of nanomaterial comparisons.
Ontology / Taxonomy Resources	To obtain an up-to-date database of ontologies in a table type format so that one can easily review them. This includes resources like the NCI Thesaurus (https://ncit.nci.nih.gov/ncitbrowser/) [75], BioPortal (http://bioportal.bioontology.org/) [76], and Ontobee (http://www.ontobee.org/). This will allow databases to link to term references and accession numbers.

Stakeholder recommendations for the nanotechnology community in furthering integration

To assist in providing guidance to the nanotechnology community, stakeholders provided recommendations for furthering the integration and exchange of data sets across nanomaterial resources. Guidance centered around the development of pilot projects supporting data integration and the establishment of a global alliance in nanotechnology for standardizing data formats and web services.

Obtain commitment to integration

Stakeholders expressed that the only way to achieve integration effectively is to:

1. Be committed to achieving integration,

Table 5: Common web services envisaged by these stakeholders as being needed to support integration of nanomaterial data in the biomedical nanotechnology and nanosafety domains.

Web Service Method	Description
createIdentifier	Creates a Universally Unique Identifier (UUID) for any entity such as a material, characterization, protocol, or publication
getCharacterization	Retrieves characterizations for a material by material type and characterization type (e.g. size) and returns characterization data in JSON and XML format.
getDataByDOI	Returns (pointers to) entries in the database with information about or from a specific publication.
getDataByPubMedID	Returns (pointers to) entries in the database with information about or from a specific publication.
getIdentifier	Retrieves a UUID for any entity such as a material, characterization, protocol, or publication
getIsaTabNano	Retrieves ISA-TAB-Nano files associated with a publication (DOI, PubMed)
getInvestigation	Retrieves an investigation associated with a specific disease and/or nanomaterial type and returns an investigation in JSON or XML format. The JSON and XML format would be based on metadata from ISA-TAB-Nano.
getMaterial	Retrieves materials by material type (e.g. dendrimer) or property (e.g. size) and returns a material in JSON or XML format. The JSON and XML format would represent the minimal information about a material.
getProtocol	Retrieves protocols by protocol type (e.g. in vitro) and returns a protocol document and list of materials characterized with the protocol if requested. The protocol document can be returned in a format that uses a common workflow language (e.g. CWL) and/or as a document file.
getPublication	Retrieves publications associated with a material, characterization, and/or protocol, and returns a DOI, PubMed ID, and/or URL to the publication.
getStudy	Retrieves a study associated with a specific assay type and/or nanomaterial type and returns a study in JSON or XML format. The JSON and XML format would be based on metadata from ISA-TAB-Nano.
searchByChemistry	Retrieves nanomaterials based on chemical structure or chemical similarity. Supports a function such as: "Find the most similar structure in database X".

2. Have the funding in place to complete the effort,

3. Get the right people (i.e. hands-on developers and nanomaterial experimental experts) together to work through details of conceptual design and controlled vocabulary, and

4. Continue fostering a commitment to maximum possible transparency and community-wide sharing of approaches, intentions, and techniques, despite the concurrent need of individual teams to remain competitive for what will certainly represent limited funding opportunities.

This good faith collaboration is the necessary key to making enough progress to achieve the momentum needed for success.

Initiate pilot integration projects

Initiating pilot projects in data source integration efforts is critical. As it stands, individual data resources are funded for individual purposes and collaboration and interoperability can be difficult. Based on the U.S. NNI's signature initiative for a knowledge infrastructure [77], there is already a documented need for collaborative resources. Now is clearly the time for funding pilot collaborative projects focused on data integration. These should include databases, repositories, ontology designers, experimental researchers, and predictive modelers for a better understanding of the data life cycle and for development of meaningful plans to go forward with existing and new knowledge management resources.

Establish GAIN - a Global Alliance in Nanotechnology - to develop integration standards

Similar to the genomics community that established a Global Alliance in Genomics and Health (GA4GH), the nanotechnology community should form an organization to develop integration standards. A Global Alliance in Nanotechnology (GAIN) would provide a critical mass of interest to develop:

1. A common model for representing data and their relationships,
2. A standard data dictionary, and
3. Web service specifications enabling integration.

In the stakeholder survey, all stakeholders agreed to participate in a Global Alliance pending availability of funding and time. The eNanoMapper project already actively participates in various collaborations, including the NanoSafety Cluster (NSC) Database Working Group (along with participation in other NSC working groups), the US-EU Communities of Research working group on Databases and Computational Modeling for NanoEHS, the US NanoWG, the CODATA/VAMAS Working Group developing the Uniform Description System for Nanomaterials [78] and applied for associate partnership with the CEN/CENELEC node in Europe of the International Standards Organisation (ISO). Alliances with these organizations can be strengthened to avoid unnecessary duplication of effort across the broader community with the primary objective of supporting and enabling concrete open source projects around ontologies, nanoinformatics tools, and data integration.

Focus on providing high quality and complete data sets in data repositories to encourage integration

Individual repositories should recognize the importance of providing high quality and sufficiently complete data, rather than simply focusing on providing large amounts of data. However, assessing data quality is a complex issue as is the related topic of data completeness. It should also be recognized that the requirements for data to be considered complete and the degree of quality required may be contingent upon the intended purpose of the data. The extent to which different data resources may have legitimately different definitions of data completeness, based upon their different objectives, underscores the importance of nanoinformatics data resource developers collectively recognizing the value of data integration and the need to ensure the necessary data and metadata required to support integration are documented. A thorough examination of these challenges and a set of recommendations to promote and extend best practice is presented in another article in the NDCI series [4].

Implement data stewardship

Data stewardship should be central to any nanomaterial project. Good stewardship requires that all researchers involved in the project actively participate throughout the process, from beginning to conclusion. This effort involves experimental design, data management plans (including planning for data sharing and adoption of scientific methods in handling data), data citation, and more. Stewardship implies setting aside resources for these tasks. Some will be monetary resources, e.g. for cloud storage, data hosting, possibly commercial support in making data available in community formats, but other actions should be a core part of the daily research of all the people involved in the project. Postponing planning for data (handling, retrievability, and storage) inevitably jeopardizes good stewardship and increases costs substantially [69].

Recommendations: A Path forward for achieving data integration across nanomaterial resources and with non-nanotechnology repositories

Taking into consideration needs of the stakeholders (Section), a multi-step path forward to achieve meaningful progress in integrating nanomaterial data resources is proposed. The four phases identified in Figure 5 provide a roadmap for achieving data integration. Each phase is discussed in greater detail below.

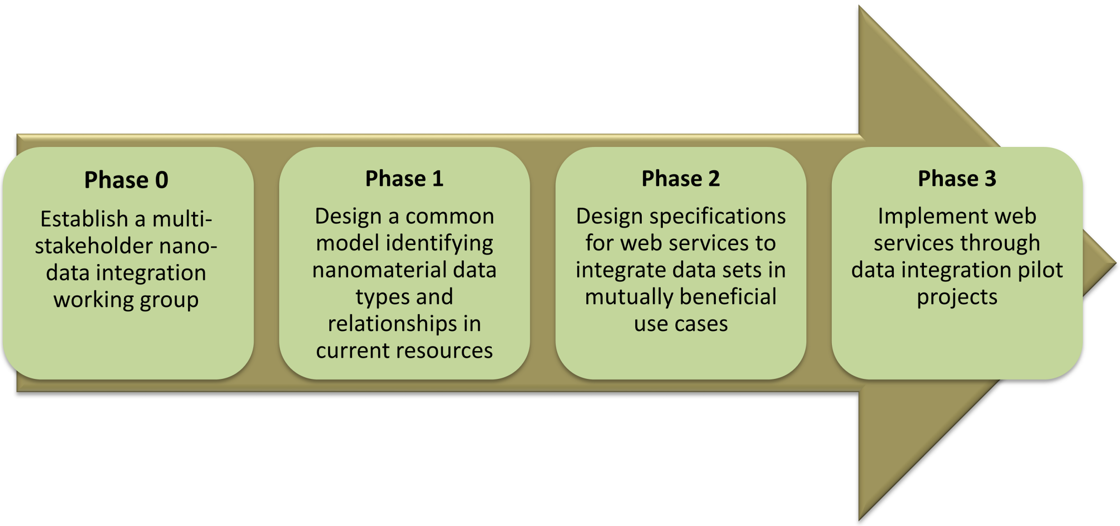


Figure 5: Roadmap of recommendations for achieving data integration across nanomaterial and non-nanomaterial repositories.

Phase 0: Establishment of an organization dedicated to achieving data integration in the nanomaterial domain

The time has come to establish a multi-stakeholder, multi-disciplinary, international group focused on nanotechnology data integration. As described above the Global Alliance in Nanotechnology (GAIN) would provide the visibility and energy to start the process towards meaningful data integration in nanoinformatics. GAIN could be an independent group or part of an existing working group such as the Nano WG (https://wiki.nci.nih.gov/display/ICR/Nanotechnology+Working+

Group) and the NanoSafety Cluster (http://nanosafetycluster.eu/) [79] focused on achieving data integration goals. Initial goals include development of a common model to describe the nanomaterial domain with associated web services supporting data exchange across specific nanomaterial sources.

Phase 1: Design of a common model that identifies nanomaterial entities and their relationships within existing resources

One of the first tasks for an organization such as GAIN would be development of a common model that identifies nanomaterial entities and their relationships. It is recommended that the common model be a graph model that depicts nanomaterial entities and nodes and associated relationships as edges (Figure 6). A graph model can provide a flexible structure that can more readily changed as the model evolves. The design of the common model can prioritize identifying the nodes and edges that cross multiple fields such as nanomaterial composition and physico-chemical characterizations [15]. Concepts from ISA-TAB-Nano and other ontologies and description systems can be leveraged to represent entities associated with investigations, studies, assays, and materials. It is important to note that this common model is not envisaged as a single, authoritative, federated cyberinfrastructure to facilitate integration in an automated manner. Rather, this model is intended to provide a centralized community-wide understanding of the nanoinformatics space, capturing an overview of the data types implicated, and providing insight into where it makes sense to dedicate resources toward detailed integration projects and tools.

Phase 2: Design specifications for web services that implement the common model

Once the common model is established, specifications for common web services can be developed, including defining service endpoints based on entities in the common model. Web service specification should be prioritized to focus on a basic query to retrieve nanomaterials by nanomaterial characteristics and other properties. Web services can be further expanded to accommodate use-

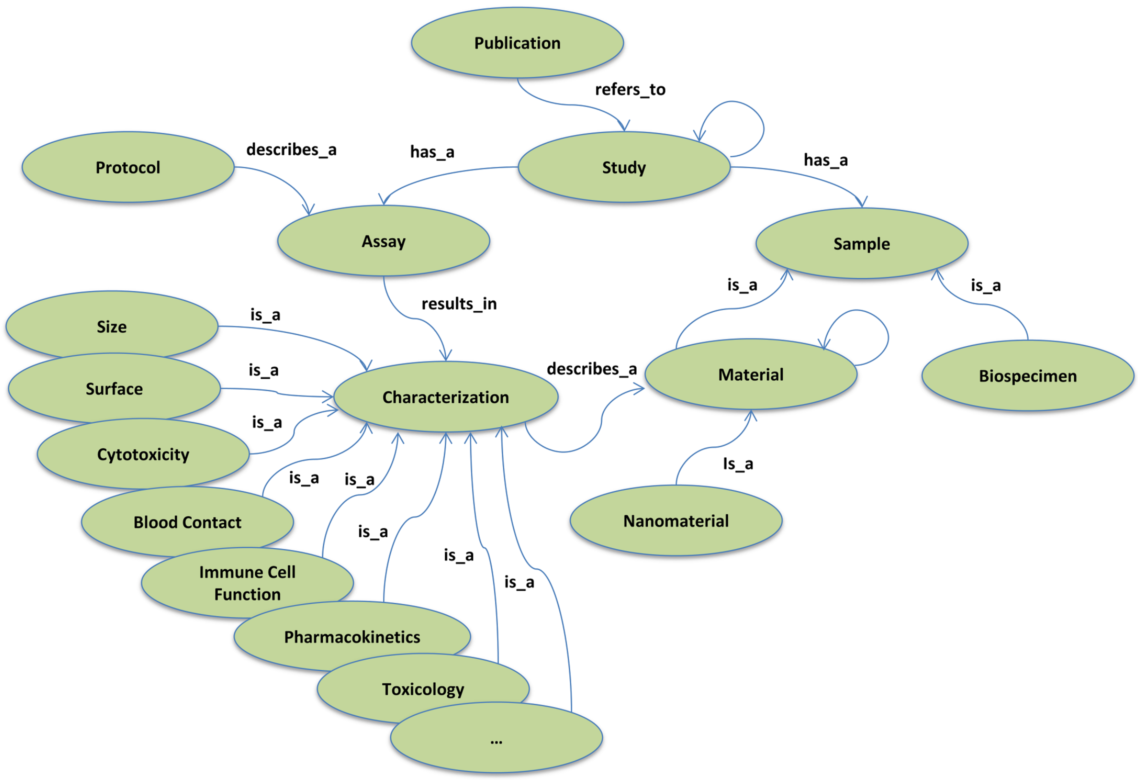


Figure 6: Example graph model depicting nodes (e.g. nanomaterial) and edges (describes_a).

case-dependent data exchange with non-nanotechnology sources. In support of data exchanges with non-nanotechnology sources, established interfaces could be published and organizations could collaborate with resource providers on developing a common interface to facilitate re-use.

Phase 3: Implementation of web services through pilot projects

Once an initial web service is designed, pilot projects should be started as soon as possible to implement the web service with an ultimate goal of querying across nanomaterial resources. Pilot projects should focus on developing re-usable software that can be extended in support of other pilot efforts. Software should be made available as open source and published as a GitHub (https://github.com/) repository. Lessons learned from pilot efforts should result in improvements to the common model and web services design specifications.

Closing remarks

The various challenges recognized by members of the nanoinformatics community are hampering efforts to integrate across nanomaterial and other non-nanotechnology resources in a meaningful way. The technical and operational challenges summarized in Figure 4 are significant barriers to scientific progress in designing new and higher impact nanomaterials and in understanding how nanomaterials interact with biological, environmental, and other systems. The tools to take advantage of high quality nanotechnology data exist but cannot be exploited unless true data sharing and integration is possible. This paper analyzes these challenges and outlines a path forward to real progress.

The authors encourage readers to share feedback or join the National Cancer Informatics Program (NCIP) Nanotechnology Working Group (https://nciphub.org/groups/nanowg/overview) and learn more about the Nanomaterial Data Curation Initiative, in particular, by visiting https://nciphub.org/groups/nanotechnologydatacurationinterestgroup/wiki/MainPage.

Acknowledgements

Authors would like to gratefully acknowledge several organizations who participated in the data integration survey including the: Center for the Environmental Implications of NanoTechnology (CEINT) which is funded by the National Science Foundation (NSF) and the Environmental Protection Agency (EPA) under NSF Cooperative Agreement DBI-1266252 and EF-0830093; Nanomaterial Registry, which is funded by the National Institutes of Health (NIH) under contract HHSN268201000022C; caNanoLab which is funded in whole or in part with Federal funds from the National Cancer Institute (NCI), NIH, under Contract No. HHSN261200800001E; Center for Safety of Substances, and Products, National Institute of Public Health and the Environment, Netherlands; DECHEMA; and eNanoMapper, funded by the European Union's Seventh Framework Programme for research, technological development and demonstration (FP7-NMP-2013-SMALL-7) under grant agreement no. 604134. RLMR is grateful for funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement #309837 (NanoPUZZLES project).

Authors would also like to acknowledge the Nano WG leaders for their time and expertise in providing the necessary tools that supported collaboration on this article. Authors would like to give special thanks to Dr. Mervi Heiskanen, the Nano WG lead from the NCI Center for Biomedical Informatics and Information Technical (CBIIT) and Dr. Stephanie Morris, from the NCI Office of Cancer Nanotechnology Research (OCNR) for their leadership and support for Nano WG initiatives including the development of this article.

The views, opinions, and content in this article are those of the authors and do not necessarily represent the views, opinions, or policies of their respective employers or organizations. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. government.

References

1. Thomas, D. G.; Klaessig, F.; Harper, S. L.; Fritts, M.; Hoover, M. D.; Gaheen, S.; Stokes, T. H.; Reznik-Zellen, R.; Freund, E. T.; Klemm, J. D.; Paik, D. S.; Baker, N. A. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology* **2011**, *3* (5), 511–532.

2. Hendren, C. O.; Powers, C. M.; Hoover, M. D.; Harper, S. L. *Beilstein Journal of Nanotechnology* **2015**, *6*, 1752–1762.

3. Powers, C. M.; Mills, K. A.; Morris, S. A.; Klaessig, F.; Gaheen, S.; Lewinski, N.; Ogilvie Hendren, C. *Beilstein Journal of Nanotechnology* **2015**, *6*, 1860–1871.

4. Marchese Robinson, R.; Lynch, I.; Peijnenburg, W.; Rumble, J.; Klaessig, F.; Hendren, C.; Marquardt, C.; Rauscher, H.; Puzyn, T.; Purian, R.; ÃĚberg, C.; Karcher, S.; Vriens, H.; Hoet, P.; Hoover, M.; Harper, S. *Nanoscale* **2016**, *8*, 9919–9943.

5. Rustici, G.; Kolesnikov, N.; Brandizi, M.; Burdett, T.; Dylag, M.; Emam, I.; Farne, A.; Hastings, E.; Ison, J.; Keays, M.; Kurbatova, N.; Malone, J.; Mani, R.; Mupo, A.; Pereira, R. P.; Pilicheva, E.; Rung, J.; Sharma, A.; Tang, Y. A.; Ternent, T.; Tikhonov, A.; Welter, D.;

Williams, E.; Brazma, A.; Parkinson, H.; Sarkans, U. *Nucleic Acids Research* **2013**, *41* (D1), D987–D990.

6. de la Iglesia, D.; Harper, S.; Hoover, M. D.; Klaessig, F.; Lippell, P.; Maddux, B.; Morse, J.; Nel, A.; Rajan, K.; Reznik-Zellen, R. et al. *Nanoinformatics 2020 Roadmap*; National Nanomanufacturing Network, 2011.

7. Hoover, M. D.; Myers, D. S.; Cash, L. J.; Guilmette, R. A.; Kreyling, W. G.; Oberdörster, G.; Smith, R.; Cassata, J. R.; Boecker, B. B.; Grissom, M. P. *Health physics* **2015**, *108* (2), 179–194.

8. Boholm, M.; Arvidsson, R. *NanoEthics* **2016**, *10* (1), 25–40.

9. Rauscher, H.; Sokull-Klüttgen, B.; Stamm, H. *Nanotoxicology* **2012**, *7* (7), 1195–1197.

10. Vance, M. E.; Kuiken, T.; Vejerano, E. P.; McGinnis, S. P.; Hochella Jr, M. F.; Rejeski, D.; Hull, M. S. *Beilstein journal of nanotechnology* **2015**, *6* (1), 1769–1780.

11. Xia, Y. *Angewandte Chemie International Edition* **2014**, *53* (46), 12268–12271.

12. Roco, M. C.; Mirkin, C. A.; Hersam, M. C. *Nanotechnology research directions for societal needs in 2020 retrospective and outlook*; World Technology Evaluation Center ; Springer, 2011.

13. Shao, C.-Y.; Chen, S.-Z.; Su, B.-H.; Tseng, Y. J.; Esposito, E. X.; Hopfinger, A. J. *J. Chem. Inf. Model.* **2013**, *53* (1), 142–158.

14. Rumble, J.; Freiman, S. *Data Science Journal* **2012**, *11* (0), ASMD1–ASMD6.

15. Stefaniak, A. B.; Hackley, V. A.; Roebben, G.; Ehara, K.; Hankin, S.; Postek, M. T.; Lynch, I.; Fu, W.-E.; Linsinger, T. P. J.; Thünemann, A. F. *Nanotoxicology* **2013**, *7* (8), 1325–1337.

16. Bult, C. J.; Eppig, J. T.; Kadin, J. A.; Richardson, J. E.; Blake, J. A.; Group, M. G. D. et al. *Nucleic acids research* **2008**, *36* (suppl 1), D724–D728.

17. Kutmon, M.; Riutta, A.; Nunes, N.; Hanspers, K.; Willighagen, E. L.; Bohler, A.; Mélius, J.; Waagmeester, A.; Sinha, S. R.; Miller, R.; Coort, S. L.; Cirillo, E.; Smeets, B.; Evelo, C. T.; Pico, A. R. *Nucleic Acids Research* **2016**, *44* (D1), D488–D494.

18. Jeliazkova, N.; Doganis, P.; Fadeel, B.; Grafstrom, R.; Hastings, J.; Jeliazkov, V.; Kohonen, P.; Munteanu, C. R.; Sarimveis, H.; Smeets, B.; Tsiliki, G.; Vorgrimmler, D.; Willighagen, E. The first eNanoMapper prototype: A substance database to support safe-by-design. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*; IEEE, 2014; pp 1–9.

19. Izak-Nau, E.; Huk, A.; Reidy, B.; Uggerud, H.; Vadset, M.; Eiden, S.; Voetz, M.; Himly, M.; Duschl, A.; Dusinska, M.; Lynch, I. *RSC Adv.* **2015**, *5* (102), 84172–84185.

20. Williams, A. J.; Harland, L.; Groth, P.; Pettifer, S.; Chichester, C.; Willighagen, E. L.; Evelo, C. T.; Blomberg, N.; Ecker, G.; Goble, C.; Mons, B. *Drug Discovery Today* **2012**, *17* (21-22), 1188–1198.

21. Maglott, D.; Ostell, J.; Pruitt, K. D.; Tatusova, T. *Nucleic acids research* **2005**, *33* (suppl 1), D54–D58.

22. Samwald, M.; Jentzsch, A.; Bouton, C.; Kallesoe, C.; Willighagen, E.; Hajagos, J.; Marshall, M.; Prud'hommeaux, E.; Hassanzadeh, O.; Pichler, E.; Stephens, S. *Journal of Cheminformatics* **2011**, *3* (1), 19+.

23. Jupp, S.; Malone, J.; Bolleman, J.; Brandizi, M.; Davies, M.; Garcia, L.; Gaulton, A.; Gehant, S.; Laibe, C.; Redaschi, N.; Wimalaratne, S. M.; Martin, M.; Le Novère, N.; Parkinson, H.; Birney, E.; Jenkinson, A. M. *Bioinformatics* **2014**, *30* (9), 1338–1339.

24. Cheung, K.-H.; Frost, H. R.; Marshall, M. S.; Prud'hommeaux, E.; Samwald, M.; Zhao, J.; Paschke, A. *BMC Bioinformatics* **2009**, *10* (Suppl 10), S10+.

25. Eyres, T. P. *EMBnet.journal* **2013**, *19*, 36–39.

26. Juty, N.; Le Novère, N.; Laibe, C. *Nucleic Acids Research* **2012**, *40* (D1), D580–D586.

27. van Iersel, M. P.; Pico, A. R.; Kelder, T.; Gao, J.; Ho, I.; Hanspers, K.; Conklin, B. R.; Evelo, C. T. *BMC Bioinformatics* **2010**, *11* (1), 5+.

28. Chambers, J.; Davies, M.; Gaulton, A.; Papadatos, G.; Hersey, A.; Overington, J. *Journal of Cheminformatics* **2014**, *6* (1), 43+.

29. Wohlgemuth, G.; Haldiya, P. K.; Willighagen, E.; Kind, T.; Fiehn, O. *Bioinformatics* **2010**, *26* (20), 2647–2648.

30. Hastings, J.; Chepelev, L.; Willighagen, E.; Adams, N.; Steinbeck, C.; Dumontier, M. *PLoS ONE* **2011**, *6* (10), e25513+.

31. Thomas, D. G.; Pappu, R. V.; Baker, N. A. *Journal of Biomedical Informatics* **2011**, *44* (1), 59–74.

32. Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. *Nucleic Acids Research* **2016**, *44* (D1), D1214–D1219.

33. Harper, S. L.; Dahl, J. A.; Maddux, B. L. S.; Tanguay, R. L.; Hutchison, J. E. *International Journal of Nanotechnology* **2008**, *5* (1), 124+.

34.

35. Izak-Nau, E.; Huk, A.; Reidy, B.; Uggerud, H.; Vadset, M.; Eiden, S.; Voetz, M.; Himly, M.; Duschl, A.; Dusinska, M. et al. *RSC Advances* **2015**, *5* (102), 84172–84185.

36. Batchelor, C.; Brenninkmeijer, C.; Chichester, C.; Davies, M.; Digles, D.; Dunlop, I.; Evelo, C.; Gaulton, A.; Goble, C.; Gray, A.; Groth, P.; Harland, L.; Karapetyan, K.; Loizou, A.; Overington, J.; Pettifer, S.; Steele, J.; Stevens, R.; Tkachenko, V.; Waagmeester, A.; Williams, A.; Willighagen, E. Scientific Lenses to Support Multiple Views over

Linked Chemistry Data. In *The Semantic Web - ISWC 2014*; Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., Goble, C., Eds.; Springer International Publishing, 2014; Vol. 8796, pp 98–113.

37. Brenninkmeijer, C.; Evelo, C.; Goble, C.; Gray, A. J. G.; Groth, P.; Pettifer, S.; Stevens, R.; William, A. J.; Willighagen, E. L. Scientific Lenses over Linked Data: An Approach to Support Task Specific Views of the Data. A Vision. In *Linked Science 2012 - Tackling Big Data*; CEUR-WS.org, 2012; Chapter 5.

38. Berners-Lee, T.; Hendler, J.; Lassila, O. *Scientific American* **2001**, *284* (5), 34–43.

39. Marshall, M. S.; Boyce, R.; Deus, H. F.; Zhao, J.; Willighagen, E. L.; Samwald, M.; Pichler, E.; Hajagos, J.; Prud'hommeaux, E.; Stephens, S. *Web Semantics: Science, Services and Agents on the World Wide Web* **2012**, *14*, 2–13.

40. Hastings, J.; Jeliazkova, N.; Owen, G.; Tsiliki, G.; Munteanu, C. R.; Steinbeck, C.; Willighagen, E. *Journal of Biomedical Semantics* **2015**, *6* (1), 10+.

41. Oksel, C.; Ma, C. Y.; Wang, X. Z. *SAR and QSAR in Environmental Research* **2015**, *26* (2), 79–94.

42. Crist, R. M.; Grossman, J. H.; Patri, A. K.; Stern, S. T.; Dobrovolskaia, M. A.; Adiseshaiah, P. P.; Clogston, J. D.; McNeil, S. E. *Integrative Biology* **2013**, *5* (1), 66–73.

43.

44. Hendren, C. O.; Lowry, G. V.; Unrine, J. M.; Wiesner, M. R. *Science of The Total Environment* **2015**, *536*, 1029–1037.

45. Brazma, A.; Hingamp, P.; Quackenbush, J.; Sherlock, G.; Spellman, P.; Stoeckert, C.; Aach, J.; Ansorge, W.; Ball, C. A.; Causton, H. C. et al. *Nature Genetics* **2001**, *29* (4), 365–371.

46. Field, D.; Sansone, S.-A. A.; Collis, A.; Booth, T.; Dukes, P.; Gregurick, S. K.; Kennedy, K.; Kolar, P.; Kolker, E.; Maxon, M.; Millard, S.; Mugabushaka, A.-M. M.; Perrin, N.; Remacle, J. E.; Remington, K.; Rocca-Serra, P.; Taylor, C. F.; Thorley, M.; Tiwari, B.; Wilbanks, J. *Science (New York, N.Y.)* **2009**, *326* (5950), 234–236.

47. Hartig, O.; Langegger, A. A Database Perspective on Consuming Linked Data on the Web. In *Datenbank-Spektrum*; Springer-Verlag, 2010; Vol. 10, pp 57–66.

48.

49. Doan, A.; Halevy, A.; Ives, Z. G. *Principles of data integration*; Morgan Kaufmann, 2012.

50. Kovrižnych, J. A.; Sotníková, R.; Zeljenková, D.; Rollerová, E.; Szabová, E.; Wimmerová, S. *Interdisciplinary Toxicology* **2013**, *6* (2), 67–73.

51. Truong, L.; Harper, S. L.; Tanguay, R. L. Evaluation of Embryotoxicity Using the Zebrafish Model. In *Drug Safety Evaluation*; Gautier, J.-C., Ed.; Humana Press, 2011; Vol. 691, pp 271–279.

52. Christen, P. *Data Matching*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2012.

53. Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. *Journal of Cheminformatics* **2015**, *7* (1), 23+.

54. Hersey, A.; Chambers, J.; Bellis, L.; Bento, A. P.; Gaulton, A.; Overington, J. P. *Drug Discovery Today: Technologies* **2015**, *14*, 17–24.

55. Fielding, R. T. Architectural styles and the design of network-based software architectures. Ph. D. Thesis, University of California, Irvine, 2000.

56.

57. Kühnel, D.; Marquardt, C.; Nau, K.; Krug, H. F.; Mathes, B.; Steinbach, C. *Environmental Sciences Europe* **2014**, *26* (1), 21.

58. Kimmig, D.; Marquardt, C.; Nau, K.; Schmidt, A.; Dickerhof, M. *Computational Science & Discovery* **2014**, *7* (1), 014001.

59. Atli, A.; Nau, K.; Schmidt, A. Navigation along Database Relationships-An Adaptive Framework for Presenting Database Contents as Object Graphs. In *WEBIST*; Institute for Systems and Technologies of Information, Control and Communication, 2011; pp 372–379.

60. Miller, A. L.; Hoover, M. D.; Mitchell, D. M.; Stapleton, B. P. *Journal of occupational and environmental hygiene* **2007**, *4* (12), D131–D134.

61. Jeliazkova, N.; Jeliazkov, V. *J. Cheminformatics* **2011**, *3*, 18.

62. Jeliazkova, N.; Chomenidis, C.; Doganis, P.; Fadeel, B.; Grafström, R.; Hardy, B.; Hastings, J.; Hegi, M.; Jeliazkov, V.; Kochev, N. et al. *Beilstein journal of nanotechnology* **2015**, *6* (1), 1609–1634.

63. Klimisch, H.-J.; Andreae, M.; Tillmann, U. *Regulatory toxicology and pharmacology* **1997**, *25* (1), 1–5.

64. Krug, H. F. *Angewandte Chemie International Edition* **2014**, *53* (46), 12304–12319.

65. Reichman, O.; Jones, M. B.; Schildhauer, M. P. *Science* **2011**, *331* (6018), 703–705.

66. Longo, D. L.; Drazen, J. M. *New England Journal of Medicine* **2016**, *374* (3), 276–277.

67. 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. et al. *BMC biotechnology* **2013**, *13* (1), 1.

68. Marchese Robinson, R. L.; Cronin, M. T. D.; Richarz, A.-N.; Rallo, R. *Beilstein Journal of Nanotechnology* **2015**, *6*, 1978–1999. doi:10.3762/bjnano.6.202.

69. Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M. A.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.;

Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; âÂĂÂŹt Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. *Scientific Data* **2016**, *3*, 160018+.

70. Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. *Nucleic Acids Research* **2016**, *44* (D1), D1202–D1213. doi:10.1093/nar/gkv951.

71. Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H. *Nucleic Acids Research* **2014**, *42* (D1), D1075–D1082. doi:10.1093/nar/gkt978.

72. Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V. et al. *Nucleic acids research* **2011**, *39* (suppl 1), D1035–D1041.

73. *R: A Language and Environment for Statistical Computing*; 2016.

74. Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J. et al. *Genome biology* **2004**, *5* (10), R80.

75. Sioutos, N.; de Coronado, S.; Haber, M. W.; Hartel, F. W.; Shaiu, W.-L.; Wright, L. W. *Journal of biomedical informatics* **2007**, *40* (1), 30–43.

76. Noy, N. F.; Shah, N. H.; Whetzel, P. L.; Dai, B.; Dorf, M.; Griffith, N.; Jonquet, C.; Rubin, D. L.; Storey, M.-A.; Chute, C. G. et al. *Nucleic acids research* **2009**, gkp440.

77. Roco, M. C. *Journal of Nanoparticle Research* **2011**, *13* (2), 427–445.

78. Rumble, J.; Freiman, S.; Teague, C. *Chemistry International* **2015**, *37* (4), 3–7.

79. Savolainen, K.; Backman, U.; Brouwer, D.; Fadeel, B.; Fernandes, T.; Kuhlbusch, T.; Landsiedel, R.; Lynch, I.; Pylkkänen, L. *Nanosafety in Europe 2015-2025: Towards Safe and Sustainable Nanomaterials and Nanotechnology Innovations*; Finnish Institute of Occupational Health: Helsinki, 2013.