Published in final edited form as:
NanoImpact. 2018 January ; 9: 85–101. doi:10.1016/j.impact.2017.11.002.

Integration among databases and data sets to support productive nanotechnology: Challenges and recommendations

Sandra Karcher[a,q], **Egon L. Willighagen**[b], **John Rumble**[c,d], **Friederike Ehrhart**[b], **Chris T. Evelo**[b], **Martin Fritts**[e], **Sharon Gaheen**[e], **Stacey L. Harper**[f], **Mark D. Hoover**[g], **Nina Jeliazkova**[h], **Nastassja Lewinski**[i], **Richard L. Marchese Robinson**[j,k,1,2], **Karmann C. Mills**[l], **Axel P. Mustad**[m], **Dennis G. Thomas**[n], **Georgia Tsiliki**[o,p], **Christine Ogilvie Hendren**[q,*]

[a]Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA

[b]Department of Bioinformatics - BiGCaT, Maastricht University, P.O. Box 616, UNS50, Box 19, NL-6200, MD, Maastricht, The Netherlands

[c]R&R Data Services, 11 Montgomery Avenue, Gaithersburg, MD 20877, USA

[d]CODATA-VAMAS Working Group on Nanomaterials, Paris, France

[e]Clinical Research Directorate/Clinical Monitoring Research Program, Leidos Biomedical Research, Inc., NCI Campus at Frederick, Frederick, MD 21702, USA

[f]Environmental and Molecular Toxicology and School of Chemical, Biological and Environmental Engineering, Oregon State University, Corvallis, OR 97331, USA

[g]National Institute for Occupational Safety and Health, 1095 Willowdale Road, Morgantown, WV 26505-2888, USA

[h]IdeaConsult Ltd., 4 A. Kanchev str., Sofia 1000, Bulgaria

[i]Chemical and Life Science Engineering, Virginia Commonwealth University, Richmond, VA 23284, USA

[j]School of Chemical and Process Engineering, University of Leeds, Leeds LS2 9JT, United Kingdom

[k]School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, James Parsons Building, Byrom Street, Liverpool L3 3AF, United Kingdom

[l]RTI International, 3040 Cornwallis Rd., Research Triangle Park, NC 27709, USA

[m]Nordic Quantum Computing Group AS, Oslo Science Park, P.O. Box 1892, Vika, N-0124 Oslo, Norway

[n]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA

[o]School of Chemical Engineering, National Technical University of Athens, 9 Heroon Polytechneiou Street, Zografou, 15780, Athens, Greece

[*]Corresponding author. christine.hendren@duke.edu (C.O. Hendren). 1 Current. 2 Previous.

[p]Institute for the management of Information Systems, ATHENA Research and Innovation Centre, Artemidos 6 & Epidavrou, Marousi, 15125 Athens, Greece

[q]Center for the Environmental Implications of Nano Technology (CEINT) Duke University, Box 90287, 121 Hudson Hall, Durham, NC 27708-0287, USA

Abstract

Many groups within the broad field of nanoinformatics are already developing data repositories and analytical tools driven by their individual organizational goals. Integrating these data resources across disciplines and with non-nanotechnology resources can support multiple objectives by enabling the reuse of the same information. Integration can also serve as the impetus for novel scientific discoveries by providing the framework to support deeper data analyses. This article discusses current data integration practices in nanoinformatics and in comparable mature fields, and nanotechnology-specific challenges impacting data integration. Based on results from a nanoinformatics-community-wide survey, recommendations for achieving integration of existing operational nanotechnology resources are presented. Nanotechnology-specific data integration challenges, if effectively resolved, can foster the application and validation of nanotechnology within and across disciplines. This paper is one of a series of articles by the Nanomaterial Data Curation Initiative that address data issues such as data curation workflows, data completeness and quality, curator responsibilities, and metadata.

Keywords

Nanomaterials; Nanotechnology; Nanoinformatics; Data integration; Databases; Web services

1. Introduction

Understanding and addressing complexities involved in integrating nanomaterial and non-nanomaterial data resources to enable and advance scientific research is a key focus of nanoinformatics (Thomas et al., 2011a). This article discusses the integration of data resources across nanotechnology, including non-nanotechnology resources. It is one in a series of papers focusing on important aspects of nanoinformatics produced by the Nanomaterials Data Curation Initiative (NDCI), which is part of the National Cancer Institute (NCI) Nanotechnology Working Group (Hendren et al., 2015). Other articles in this series discuss data curation workflows (Powers et al., 2015) and data completeness and quality (Marchese-Robinson et al., 2016).

1.1. Background

The NDCI is currently working to advance nanoinformatics and is exploring the role of data integration as an essential component within the field. The following definition of nanoinformatics (expanded from the Nanoinformatics 2020 Roadmap (de la Iglesia et al., 2011)) has been proposed (Hoover et al., 2015).

"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; confirming that appropriate decisions were made and that desired mission outcomes were achieved; and finally, conveying experience to the broader community, contributing to generalized knowledge, and updating standards and training."

Data integration within nanoinformatics and with outside data resources supports productive nanotechnology, fostering the application and validation of nanotechnology within and across disciplines. Integration of data means combining different data sets such that they are compatible with one another in format and meaning to enable comparison and co-analysis. The nanoinformatics vision is that, beyond achieving individual project goals, the potential exists for broadly-integrated data sets to yield new and unexpected insights from deeper data mining, to generate new hypotheses and knowledge not anticipated by the originating data resources, and to benefit multiple stake-holders. 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.

The overlap of interests among biomedicine, materials science, precision agriculture, and environmental, health, and safety (EHS) research is illustrated in Fig. 1. The figure shows that each field pursues research relating to 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. By looking across all four disciplines, data integration potentially positively affects the entire data life-cycle, from experimental design through data sharing.

Integrating data from different data resources supports multiple goals specific to diverse organizations or projects (Oksel et al., 2015) and is a necessary precursor to deeper data mining to enable interdisciplinary scientific discovery, facilitate regulatory decision making, and provide insight into improving the properties and performance of nanomaterials.

1.2. Importance of data integration to nanotechnology

Nanomaterials (Boholm and Arvidsson, 2016; Rauscher et al., 2012) are becoming ubiquitous in science and technology (Vance et al., 2015; Xia, 2014). Biomedical researchers are making multifunctional nano-materials to diagnose, target, and treat many diseases looking for ways to increase nanomaterial stability and optimize nanomaterial performance while minimizing potential negative effects (Xia, 2014). Other researchers are harnessing similar 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 (Roco et al., 2011).

To design optimal nanomaterials and predict their behaviors, researchers must use data from disparate, non-standardized resources across biomedical, environmental, health and safety, and materials science disciplines. Problems abound. Even when the composition of a nanomaterial is provided, the nomenclature used to describe its components - the base nanomaterial formulation and the material constituents (such as core, coat, shell, and any surface modifiers) - and the relationships among them is not standardized, and in many cases, are incomplete. For example, the surface density of “decorator” molecules on carbon nanotubes is rarely provided, resulting in the need for “guessing” the actual structure when preparing representative structure files for computational modeling (Shao et al., 2013).

A variety of physical-chemical characterization information (such as the size, shape, purity, and surface properties) are included in different resources, but the methods and techniques used to perform the characterization are not always included in sufficient detail or standardized in a way that supports cross-study comparison of reported values (Stefaniak et al., 2013; Marchese-Robinson et al., 2016). Repositories collect and store information in support of their organization’s needs and goals, for which the role of nanomaterials data can differ considerably as indicated using the following examples.

- Biomedical repositories focusing on the reactivity and efficacy of nanomaterials in living systems.
- Environmental repositories with geospatial information of the fateofnanomaterialsintheenvironment.
- Physico-chemical repositories containing physical and chemical properties of nanomaterials.
- Genomic and biological pathways repositories with information on biological structures and reactivity.

These disparate data repositories, when integrated together, can provide greater insights into understanding common endpoints such as nanomaterial toxicity or stability (Izak-Nau et al., 2015). Because of the current lack of standardization and integration of resources, data users must review documentation describing the protocols for storing information in each repository, and sometimes retrieve and review original publications to determine what is and what is not relevant to their research.

1.3. Influence of organizational purpose and goals on data integration

The approaches taken by an organization or project to gather and organize data are governed by the driving scientific questions that need to be answered to further its mission. Some examples of use case scenarios that could benefit from multidisciplinary data integration are shown in Fig. 1. 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 (Marchese-Robinson et al., 2016). Examples are provided below.

- Building an authoritative repository of nanomaterial characterizations.
- Parameterizing models topredictnanomaterial behavior indifferent systems (biomedical, environmental, or other).
- Enabling environmental and health risk assessments
- Improving performance of materials, medicines, or pesticides.

The individual resource goals 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.

Integrating data from different data resources supports multiple goals specific to diverse organizations or projects (Oksel et al., 2015). Using the example provided in Fig. 1, understanding which parameters control stability of a nanomedicine in the human bloodstream could provide insight relevant to predicting nanomaterial dissolution or aggregation in a body of freshwater, transport within a crop field, or efficacy in a material fabrication process.

1.4. Purpose and structure of this article

The goal of this article is to capture the current state of data integration in nanoinformatics and provide recommendations for advancing integration within and outside of the nanoinformatics field. As discussed above, the integration of nanomaterial data with other nanomaterial data sets as well as with data from other fields will lead to new and exciting scientific opportunities. This article not only informs the nanoinformatics and nanomaterials testing communities of the challenges involved in data integration, but also identifies concrete actions that will accelerate the integration process. In addition, because much of the material presented herein is based on the results of a survey of most present-day nanomaterials data resources, the authors are confident that many of the current challenges are shared across the emerging communities; further, the characterization of the issues and the recommendations that follow position nano-communities to move forward towards integration, and offer insight to other still-maturing fields characterized by uncertainty.

Following the introduction and a discussion of the importance of integration and the influence of organization purpose and goals on the current state of data integration, the article presents the results of a nanoinformatics community-wide stakeholder survey designed to assess the current practices in integrating data in nanotechnology. Stakeholder demographics are presented (Section 2), followed by stakeholder-identified challenges to integration (Section 3) and stake-holder-identified needed integration functionality (Section 4). Section 5 describes technological and semantic approaches to achieving integration. Section 6 presents stakeholder recommendations for achieving integration, followed by an author-proposed path for moving forward (Section 7). Some closing remarks are provided in Section 8.

2. Stakeholder demographics

To understand the current practices in data integration and to identify challenges and offer recommendations, several organizations that maintain nanomaterial data resources were surveyed. The goal was to identify, define, and provide a stimulus for initiating integration and exchange of data resources across nanomaterial data repositories and with other related, non-nanotechnology data resources. Survey questions included current and recommended functionality and web services enabling data integration as well as perceived challenges associated with integrating primary experimental data sets, or data sets curated from the literature, with other data resources. Appendix A contains a detailed summary of the stakeholder responses to the survey.

Stakeholders who participated in the survey ranged from nanomaterial resources that have extensive experience in integrating data resources to those with limited data integration experience whose focus was primarily on repository development (see Table A-1). The diverse levels of integration capabilities provide insight into the challenges that need to be addressed to integrate across nanomaterial repositories and with other non-nanotechnology resources. A brief summary of the organizations that participated in the survey, which represent some of the most active participants in the field of nanoinformatics, follows.

caNanoLab, a data sharing portal supported by the National Cancer Institute of the U.S. National Institutes of Health, is designed to facilitate information sharing across the international biomedical nano-technology 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.

The Center for Environmental Implications for Nanotechnology (CEINT), located at Duke University (U.S.), focuses on exploring the potential impact of exposure to nanomaterials on ecological and biological systems. The Center is funded by the U.S. National Science Foundation and brings together researchers from several universities, the National Institute of Standards and Technology (NIST), the US Environmental Protection Agency (EPA), the Consumer Product Safety Commission (CPSC), the US Army Corps of Engineers (USACE), 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.

The CSSP/NIPHE, Netherlands (The Center for Safety of Substances and Products), National Institute for Public Health and the Environment (CSSP at NIHE, or RIVM) provides a data resource on eco-toxicity data focusing on nanoparticles in consumer products, used for modeling purposes (QSAR). Also of significance to integrating a broad array of nanomaterial data, RIVM hosts the protocols and other text documents from across the European Union (EU) Seventh Framework Programme 2007–2013 (FP7) NANoREG project; they collaborated with the FP7 eNanoMapper project (see below) that hosts the data as collected using ISA-TAB-Nano inspired (Thomas et al., 2013) MS Excel templates. The templates produced by the NANoREG project are also available (Totaro et al., 2017). (http://publications.jrc.ec.europa.eu/repository/handle/JRC103178).

DECHEMA is a network of experts in chemical engineering and biotechnology and supports several projects applicable to nano-technology such as the DaNa project and the NANORA project (Kühnel et al., 2014). DaNa is a knowledgebase 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.

eNanoMapper was an EU-funded FP7 project comprising eight research and industry institutes, whose aim is to improve data integration and to support safe-by-design development by building up a nanosafety ontology and database and provide web modeling tools for use of these data. Currently the eNanoMapper database at (https://search.data.enanomapper.net) hosts data generated by several EU projects including publicly accessible NANoREG data. Tools for converting Excel templates into ISA-TAB/ISA-Nano JavaScript Object Notation (JSON) and semantic formats are provided. The application programming interface facilitates data usage by the modeling community.

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. Curation is based on a set of minimal information about nanomaterials (Mills et al., 2014). The data of the Registry are also available on the Portal at nanoHUB (https://nanohub.org/) where predictive modelers can find the data in a format that is easy for them to use.

The Nanoparticle Information Library (NIL) (http://nanoparticlelibrary.net/nil.html) is a prototype searchable data resource 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. It is operated by the National Institute for Occupational Safety and Health (NIOSH) in the United States.

3. Stakeholder-identified data integration challenges

In the survey described above, responding nanoinformatics stake-holders identified several technical and operational challenges impacting current data integration efforts, as shown in Fig. 2. These challenges, if not addressed, will continue to hinder the exploitation of the scientific possibilities opened by linking nanomaterials data resources to one another and to data resources in related scientific areas such as biology, medicine, and environmental studies. Each challenge is presented in greater detail below. It should be noted that, in preparing each section that incorporates survey responses, the authors organized and summarized the stakeholders input to consolidate and clarify across the breadth of responses while attempting to maintain the intent and keep as much of the verbiage provided as possible.

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

The primary challenge in achieving data integration in nano-technology is the diversity of ways in which nanomaterial information is represented across data resources 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 also abundant. For example, ZnO and zinc oxide are common ways of referring to the same chemical. Developing appropriate ontologies, including resolution of terminology conflicts, to address the nuances of nanotechnology research is a critical and important key to achieving integration (Thomas et al., 2011a, 2011b; Hastings et al., 2011).

3.2. Lack of unique identifiers for the entities in the domain

Certain aspects of data integration pertaining to semantics, or judgments of meaning, remain challenging regardless of the specific domain, and regardless of the thoroughness of the supporting ontology. For example, interpretation can become ambiguous in 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). Clear guidance for utilizing ontologies and tools must accompany their development to stave off differences introduced by end-user judgment; consider, for example, the difference between the Nano Particle Ontology terms for titanium dioxide (NPO_1485) and titanium oxide nanoparticle (NPO_1486). The NPO terms are intentionally distinct to distinguish between a compound and a nanoparticle and enable semantic integration of data sets. However, without consistent guidance or built-in tools, a user might potentially select either one. The inherent freedom of interpretation in an ontology without associated guidance tools can diminish the confidence in combined data and impact the ability to perform cross-material comparisons.

Other scientific fields have introduced naming conventions for generating unique identifiers based on metadata. The prevalence of informal terminology and the lack of a strong business case to create more formal conventions, however, make this challenging in the nanoscience area. Fields such as genomics are moving forward with generating a Universal Unique Identifier (UUID) for entities not based on metadata. In the context of nanomaterial data resource integration, needed 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 (Stefaniak et al., 2013), 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 (Marchese-Robinson et al., 2016).

3.3. Data are conceptualized in different ways

In conceptualizing data models, both definitions of individual data elements and the relationships between them can be created. As a result, data models created for different purposes or different database owners may have the same data elements but with different relationships between them. There has been a trend away from establishing a fixed relationship, such as a hierarchy, between database elements, a trend that, in some regards, adds to the data mining challenge. Sometimes knowledge of the relationships between attributes is built into the establishment of a hierarchy and that knowledge can be extracted when mining a database to ensure 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 databases, measurements are sometimes stored in replicate so that they can be found in different types of searches. If the seemingly replicated measurements are not handled correctly during analysis, they can lead to bias in statistical computations.

3.4. Information that should be maintained as multiple individual 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 is always 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. 5 mg/L), or included as a range of values in one field (e.g. 7–10 mg/L).

3.5. Lack of publicly available web services for data retrieval

Integration is often hindered by the lack of publicly available web services supporting data retrieval. Without publicly available web services, each user must individually create data retrieval tools every time a data resource is to be integrated. Additionally, even when data services are provided, open frameworks such as the Representational State Transfer framework (REST) (Fielding and Taylor, 2000) are not leveraged to ease development of integration touchpoints.

3.6. Data across organizations has varying levels of quality and completeness

A key challenge for nanoinformatics is finding data that are sufficiently complete and of acceptable quality. At times, data from external resources are not integrated with local systems due to concerns regarding the quality and completeness of those data. For example, a local knowledgebase 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 knowledgebase. When evaluating external data for inclusion in the knowledgebase, 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 because it is often difficult to obtain the necessary metadata to support comparison (a prerequisite for matching and data integration) between material records in different databases. For example, when obtaining information on physico-chemical characterization, it is valuable to have information on the chemical composition of the nanomaterials, such as the presence/absence of coatings, and if the nanomaterial has been transformed. Lack of complete metadata for associated biological tests may affect the clarity, and hence quality, of results (Klimisch et al., 1997) and could preclude an assessment of whether two sets of results were generated under sufficiently similar conditions to allow them to be meaningfully integrated. The lack of proper particle characterization is a key problem (Krug, 2014), 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. Though no system of characterizing quality or completeness has been broadly adopted, relative evaluation approaches have been proposed, including the NNI Data Readiness Level framework, as modeled from technology readiness assessment methods https://www.nano.gov//NKIPortal/DRLs.

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 (Marchese-Robinson et al., 2016).

3.7. 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 (which often depends on the research question to be addressed by the resulting integrated data set). 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 the NDCI series (Marchese-Robinson et al., 2016).

3.8. 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 nor cover the data selection process. A commonly accepted minimum documentation standard is needed.

3.9. Protection of 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 rights and investments in generating and updating database content. Often, these restrictions have unclear statements regarding 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 others will take their data and use it in a way that limits or reduces the novelty of their work (Reichman et al., 2011). Some have even suggested that those performing analysis on data they had no role in generating are “research parasites” (Longo and Drazen, 2016). The need to maintain “unique selling points” of a data resource can impede data sharing. One solution to overcome this challenge is to provide a web service with restricted access in support of data retrieval while maintaining a customized interface to maintain unique characteristics of the resource.

3.10. Lack of project funding

Individual projects to build data resources and repositories usually do not have funding allocated to data integration. Furthermore, 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 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.

4. Stakeholder-identified functionality needed to enable data integration

To make progress on data integration, the key challenges noted above must be addressed through individual and collective activities. Stakeholders identified a number of critical functionalities and web services that are needed to enable data integration across nanomaterial repositories. Stakeholders also identified use-case-driven integration needs with non-nanotechnology resources.

4.1. Use of shared controlled vocabularies

To integrate across resources, each resource needs to either adopt shared controlled vocabularies or be able to map to agreed-upon standards; see for example (ASTM Standard E2456, 2006; ISO, 2007). When mapping between controlled vocabularies, it is important to fully document the mappings and develop tools to assist in the mapping and transformation of 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.

4.2. 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 terms (Gruber, 1995). A more detailed discussion of ontologies and how they contain more than just a controlled vocabulary is provided by Thomas and others (Thomas et al., 2011a, 2011b; Hastings et al., 2011). Additionally, search capability should support retrieval of data (e.g. primary nanomaterial characteristics) across each nanomaterial resource and retrieval of detailed information, e.g. study endpoints applicable to the resource, from the same source on. 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, as well as NANoREG data. The site also offers integrated searching over data from several past FP7 projects for H2020 NanoReg2 project partners.

4.3. 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.

4.4. 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 (Thomas et al., 2013; Marchese-Robinson et al., 2015; ASTM Standard E2909, 2013), which allows data to be exported from one data resource 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.

4.5. Tools to analyze and visualize data

Data analysis and visualization tools within and across nanomaterial resources will facilitate cross-material comparisons. Visualizing nano-materials 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.

4.6. 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 in a format supported by modeling tools.

4.7. 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 or quality, or both, can assist in providing this assessment. Approaches could 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 (Marchese-Robinson et al., 2016).

4.8. 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 use of resources, such as ZENODO (https://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. These resources, however, do not support the application of automated data integration techniques (Wilkinson et al., 2016). In addition to annotations on data provenance, data annotations can also be provided to further clarify the quality of the data.

4.9. Web services needed to enable data integration across nanomaterial data repositories

Stakeholders supporting the use of nanotechnology in the biomedicine and the nanosafety communities 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 data resources from environmental and health fields, 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 the metadata identified as part of the ISA-TAB standard.

To support such data integration, ISA-TAB and ISA-TAB-Nano Application Programming Interfaces (APIs) that retrieve entities based on the ISA-TAB and ISA-TAB-Nano JSON schemas (https://github.com/ISA-tools) have been developed. (N.B.: previously ISA-TAB and ISATAB-Nano data sets have been represented using tab-delimited text fields (Sansone et al., 2008; Marchese-Robinson et al., 2015). The nanosafety community has many interests and covers many different scientific domains. Of special interest at this time are web services that meet two essential needs: determination of the similarity between two nanomaterials and locating all data and information associated with one published paper or with a specific 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 1.

4.10. Needs for integrating nanotechnology data repositories with nonnanotechnology resources

Stakeholders also identified a variety of non-nanotechnology resources that must be accessed to support use case driven data integration needs; these are summarized in Table 2.

5. Data integration approaches

A variety of approaches to data integration exist, with new ones continuing to be developed, supported by technologies ranging from manual integration (e.g. via an Excel spreadsheet) to a federated search architecture based on semantic web technologies (http://www.w3.org/standards/semanticweb/) (Cheung et al., 2009; Eyres, 2013). It is beyond the scope of this paper to provide a comprehensive review of these approaches. Instead, approaches that best facilitate the retrieval of integrated data via automated queries (e.g. through data query languages such as SQL or SPARQL (Hartig and Langegger, 2010)) are discussed. Nonetheless, it is important to note that given the preference of many scientists to work in Excel, tools that allow for automated integration of manually prepared Excel data sets into queryable databases are of considerable value (e.g. https://github.com/enanomapper/nmdataparser).

5.1. Data integration technology

The extremes of the spectrum with regard to selecting an architecture that supports data integration through automatic querying are provided below (Doan et al., 2012).

- **Data warehousing** – an approach that loads the content of different data resources into the same physical database. Subsequentlythe "warehouse" database can be queried, which involves querying all loaded data resources concurrently, with results presented to the user.
- **Federated querying** – an approach that sends the same query todifferentdataresourcesattheiroriginallocationsandpresentsthe results to the user in a unified view after they are received.

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 across resources. The integration itself relies on how the different attributes of the mediated schema match the attributes of the resources, 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.

The technology for accessing the data resources can be the same for both approaches. 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 can 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 emergence of new technologies has repeatedly changed technical approaches to data integration. While paradigms based on central data platforms still predominate (Williams et al., 2012; Maglott et al., 2005), the wider data integration community often use a more distributed, more-easily scalable cloud platform (Samwald et al., 2011; Jupp et al., 2014) and other methods, based upon federated search approaches (Cheung et al., 2009; Eyres, 2013). Additionally, between the two extremes of data warehousing and federated query, hybrid architectures exist that combine elements of both pure data warehousing and federated querying.

5.2. Semantic issues in data integration

The choice of integration architecture depends not only on technological approaches, but also on the best approach for addressing semantic issues. In selecting an approach, the following two questions must be considered.

- How can/will entities be matched across data resources?
- How will query results be integrated into coherent answers?

Regardless of the integration approach, all methods require **entity matching** (linking associated information based on database content), or **schema mapping** (virtually altering the schema of one database so that its content can be queried with data from a database with a different schema), or both mapping techniques. Mapping is typically performed using transformation procedures, and there may not exist a simple one-to-one mapping between the final schema and the original data resources. ***This is especially true in scientific and technological disciplines in which very complex concepts (entities) have been modeled differently by different groups at different times, a situation exacerbated by the evolution of new knowledge constantly being developed***.

A key requirement for an integration effort is having a network of schema mapping algorithms, based on individual data and metadata identifiers, that crosslink content from different data resources. In disciplines close to nanotechnology, efforts such as http://identifiers.org/ (Juty et al., 2012) unify how these identifiers are represented, and other resources provide solutions for mapping identifiers from different databases (van Iersel et al., 2010; Chambers et al., 2014; Wohlgemuth et al., 2010). Identifiers, however, typically focus on entities studied, such as chemicals, materials, genes, and proteins. Identifiers for cell lines, assays, and other key entities involved in nanosafety data are less common, though ontologies commonly provide identifiers for them (Hastings et al., 2011; Thomas et al., 2011b; Hastings et al., 2015a).

Developing mapping algorithms has traditionally been done manually; however, active research is producing tools for automatic schema mapping and record linkage by deterministic, probabilistic, and machine learning methods (Christen, 2012). 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. The challenge, of course, is that mapping entities for complex scientific subjects can be very difficult, even for experts. For example, the concept “percentage cumulative mortality,” which may be reported in different ways depending on the experimental time course and time of observation, can require post-processing of retrieved data (Kovrižnych et al., 2013; Truong et al., 2011).

5.3. Data integration in scientific fields closely related to nanotechnology

Solutions to data integration problems in nanotechnology can be informed by practices developed for use in other scientific disciplines. Specifically, semantic issues encountered in other fields are similar to those experienced in the nanoinformatics community, and the types of organizations and individuals interested in data integration are also similar. Integration efforts in three closely related fields are described below.

5.3.1. Integrating small molecule chemical databases—As in nanoinformatics, the major challenge in integrating small molecule chemical data resources is the determination of the equivalency of the composition and structure of two small molecule entities. In the case of small molecule chemicals, the entities are the chemical structures, and the IUPAC International Chemical Identifier, InChI, (http://www.inchi-trust.org) (Heller et al., 2015) can be used as a uniform identifier across databases. When performing integration, the rule for entity matching is, "*if the search results returned include one and the same Standard InChI, then the results are for the same compound*".

Several complexities must be considered when matching based on an InChI. For example, small molecule chemicals may be considered the same, yet still have different InChIs due to their rapidly interconverting structures. While 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 that may readily interconvert in practice - such as differences in protonation state or between open-and-closed ring forms that can equilibrate for sugars in solution. If non-standard InChIs are used, the situation is further complicated; indeed, if the so-called "perception options" are employed, different "standard" InChIs may be generated for the same input structure (Heller et al., 2015). Regardless, integration of small molecule chemical data resources 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 matching (Hersey et al., 2015).

The development of a single, comprehensive identifier for nano-materials, similar to InChI, is an attractive prospect, though the size, complexity, and requisite three-dimensional nature of nanomaterials indicate many challenges for this approach.

5.3.2. Integrating biomolecular databases—The integration of biomolecular databases has faced a number of complex problems, such as mapping things that are related but not identical and mapping things that are similar but not identical. Extending the single identifier approach to more complicated structures (e.g. proteins and genes), requires expanding queries to handle all possible synonyms used within different databases. This is difficult because the nomenclature for biomolecules is still not totally standardized. However, some success has been achieved in establishing a common API for a given type of resource, facilitating integration by alleviating the need for schema matching. Essentially, the API defines a common schema and if all resources of the same kind are compliant with the API, the main burden of semantic mapping is met.

An example of 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. Web services using the REST framework (Fielding and Taylor, 2000) are provided with query requests and responses formatted using the 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 need for such protocols is essential as modeling data integration approaches to capture cytotoxicity effects are now emerging (Kohonen et al., 2017).

The implementation of a central data warehouse that aggregates data from several resources requires ETL processes to assist in aggregating and transforming 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 (https://pubchem.ncbi.nlm.nih.gov/) and ChEMBL (https://www.ebi.ac.uk/chembl/) databases. The Open PHACTS project (https://www.openphacts.org/) provides a common API to a variety of pharmacological data sets. It does not, however, normalize to a single data model, but addresses non-uniformity at the API level (Williams et al., 2012). The European Bioinformatics Institute Resource Description Framework (EBI-RDF) platform uses another approach, maintaining multiple RDF repositories for different resources and allowing federated searching across all of them (Jupp et al., 2014). Entities in the EBI-RDF platform are assigned equivalent identifiers, with identifiers.org service providing mappable URIs, which is essentially implementing the mapping between the distributed resources.

5.3.3. Integrating public life sciences databases—

Life sciences data resources provide further insight into the handling of integration of complex data inherent in living biological systems. In these systems, additional complexity is added by the large number of variables needed to fully characterize a living system.

The Syngenta federated search system (Eyres, 2013) 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/), and provides 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 multidisciplinary researchers, even when the questions driving their research were different. Other similar resources include Ontoforce (http://www.ontoforce.com/) and Euretos (http://www.euretos.com/).

5.4. Data integration lessons for nanoinformatics

Hopefully, the power of integration demonstrated in closely-related fields will stimulate interest in the nanoinformatics community to start significant integration projects. Steps towards integration could begin with developing an understanding of the minimum set of data needed to support integration. The minimum data set will likely vary from resource to resource based on 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 (Brazma et al., 2001)) provides a greater understanding of data requirements. It must be noted that often more information is needed for full data integration than the amount contained in the minimum data requirements. One possible candidate for such a metadata resource is the FAIRsharing platform (https://fairsharing.org/) (Field et al., 2009). Further discussion of data and metadata requirements and their explicit documentation via minimum information checklists is presented in an earlier NDCI article (Marchese-Robinson et al., 2016).

Linking data enables data integration; by integrating data sets, data comparisons are enabled. Linking does not, of course, provide the definitions needed for data to be compared, or even which data can be connected. Decoupling data integration into two steps, linkage and comparison, allows formalization of a hypothesis into a query. For example, consider the linkage of two nanomaterial data resources, one containing clinical data and the other embryonic zebrafish toxicity data. 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") (Harper et al., 2008) 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, 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") (see Table 3).

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 (Batchelor et al., 2014; Brenninkmeijer et al., 2012). This approach merges the worlds of ontologies and data by using Internationalized Resource Identifiers (IRIs), such as those found in the set of Semantic Web technologies (Berners-Lee et al., 2001; Marshall et al., 2012). The Open PHACTS project has taken this approach and developed an Identifier Mapping Service (IMS) that links databases using IRI-based identifiers (Batchelor et al., 2014). 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://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:33128) as defined in the ChEBI ontology (Hastings et al., 2015b).

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 an earlier study to be automatically notified that new related data have been added to the database. These advances in data integration are directly applicable for nanoinformatics and also enable a variety of research goals to be achieved that are specific to a particular organization.

6. Stakeholder recommendations for advancement in nanoinformatics

To assist in providing guidance to the nanotechnology community, stakeholders (in the survey discussed in Section 2) provided recommendations for furthering the integration and exchange of data sets across nanomaterial resources. Recommendations centered on the development of pilot projects supporting data integration and the establishment of a global alliance in nanotechnology for standardizing data formats and web services.

6.1. Obtain commitment to integration from funding bodies and from active project leadership

Stakeholders expressed the opinion that the only way to achieve integration effectively is to take the steps listed below.

1. Be committed to integration.
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.
4. Continue fostering a commitment to maximum possible transparency and community-wide sharing of approaches, intentions, and techniques.

Despite the current competitive funding situation, the nanoinformatics community must work together to maximize what funding resources are available. This good faith collaboration is the necessary key to making enough progress to achieve the momentum needed for long-term success.

6.2. 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 finding resources to devote to collaboration and interoper-ability can be difficult. Based on the U.S. National Nanotechnology Initiative's signature initiative for a knowledge infrastructure (Roco, 2011), there is already a documented need for collaborative resources. Stakeholders clearly believe now is the time to fund pilot collaborative projects focused on data integration. To foster a better understanding of the data life cycle and to be successful in developing meaningful plans for moving forward with existing and new knowledge management resources, these projects must be multidisciplinary and include ontology designers, experimental researchers, and predictive modelers

6.3. Establish a Global Alliance in Nanotechnology (GAIN) to develop integration standards

Similar to the Global Alliance for Genomics and Health (GA4GH) established by the genomics community, the nanotechnology community should form an organization to develop integration standards. A Global Alliance in Nanotechnology (GAIN) would provide the critical mass of interest and commitment necessary to support such development. The first steps of the GAIN would include: developing a common model for representing data and data relationships, creating a standard data dictionary, and establishing web service specifications needed to enable integration.

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

Individual resources should recognize the importance of providing high quality and sufficiently complete data, rather than simply focusing on providing large amounts of data, especially if the data quality is suspect. Assessing data quality, however, 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 (Marchese-Robinson et al., 2016).

6.5. Implement data stewardship

Data stewardship, the management and ownership of data assets in an organization, such that the data are easily available and of necessary quality and consistency, 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 planning (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 (Wilkinson et al., 2016).

7. Authors' recommendations: a path forward for achieving data integration across nanomaterial resources and with nonnanotechnology repositories

Taking into consideration the needs and recommendations of the nanoinformatics stakeholders, a multi-step path forward to achieving meaningful progress in integrating nanomaterial data resources is proposed. The phases identified in Fig. 3 provide a roadmap towards integration. Each phase is discussed in greater detail below.

7.1. Phase 0: establishment of an organization dedicated to achieving data integration in the nanomaterial domain

Based on the authors' experience with multi-partner projects in Europe and the United States, the authors recommend the nanoinformatics community establish a multi-stakeholder, multi-disciplinary, international group focused on nanotechnology data integration. As described above, this envisioned group, referred to here as the Global Alliance in Nanotechnology (GAIN) would provide the visibility and energy needed to start the process towards meaningful data integration in nanoinformatics. The GAIN could be an independent group or part of an existing working group such as the U.S. Nano WG https://wiki.nci.nih.gov/display/icr/nanotechnology+working+group or the NSC (Savolainen et al., 2013) focused on achieving data integration goals. The advantage of having a group such as the GAIN is that the synergy and diversity of data resources necessary to test integration approaches would be present. For example, the breadth of nanomaterials entities (see Phase 1 below) would be rich enough so models developed would encompass relevant domains.

In the stakeholder survey, all stakeholders agreed to participate in a Global Alliance pending availability of funding and time. Active groups include the stakeholders previously identified plus the NanoSafety Cluster (NSC, http://nanosafetycluster.eu/) Databases 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 (https://us-eu.org/communities-of-research/overview/), and the US Nano WG, the CODATA/VAMAS Working Group developing the Uniform Description System for Nanomaterials (http://www.codata.org/nanomaterials) (Rumble et al., 2014). Alliances among 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.

7.2. 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 the GAIN would be the development of a common model that identifies nanomaterial entities and their relationships. It is recommended that the common model provide a flexible structure that can more readily be changed as the model evolves. The design of the common model would prioritize identifying components that cross multiple fields, such as nanomaterial composition and physico-chemical characterizations (Stefaniak et al., 2013). 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. Instead, 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 towards detailed integration projects and tools.

7.3. 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 basic queries to retrieve nanomaterial data sorted by nanomaterial characteristics and other properties. Web services can be further expanded to accommodate use-case-dependent data exchange with non-nanotechnology resources. In support of data exchanges with these resources, established interfaces could be published and organizations could collaborate with resource providers to develop a common interface that facilitates re-use.

7.4. Phase 3: implementation of web services through pilot projects

Once an initial web service is designed, pilot projects should be started quickly to validate the common model and design specifications for web services. To reach the ultimate goal of integrated querying across nanomaterial resources will require an evolutionary approach. Pilot projects provide the mechanism for testing and making adjustments.

7.5. Phase 4: publication and demonstration to the broader science community

Once pilot projects have been successfully completed, the GAIN would publish information about the system and demonstrate the system functionality to the wider science community.

8. Closing remarks

The challenges identified by the nanoinformatics community must be recognized and overcome before integration across nanomaterial and other non-nanotechnology resources in a practical and usable manner can be accomplished. The technical and operational challenges summarized in Fig. 2 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. Some of the tools needed to take advantage of high quality nanotechnology data exist, but full exploitation of true data sharing and integration to develop new scientific knowledge lies in the future. This paper has analyzed 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 its web site: https://nciphub.org/groups/nanotechnologydatacurationinterestgroup/wiki/MainPage.

Acknowledgements and funding sources

The 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 and funding of its Open Access publication status.

The first three authors are listed in order of contributions. The remaining authors, excluding the last, are listed alphabetically.

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.

The authors would like to gratefully acknowledge several organizations who participated in the data integration survey including the:

Center for the Environmental Implications of Nano Technology (CEINT) which is funded by the National Science Foundation (NSF) and the Environmental Protection Agency (EPA) under NSF Cooperative Agreement DBI-126625and EF-0830093;

Nanomaterial Registry, which is funded by the National Institutes of Health (NIH) under contract 84HHSN268201000022C;

caNanoLab has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government;

Center for Safety of Substances, and Products, National Institute of Public Health and the Environment, Netherlands;

DECHEMA; and

eNanoMapper, funded by the European Union Seventh Frame work 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-NMP-2012-SMALL-6) under grant agreement no. 309837 (NanoPUZZLES project).

JR acknowledges support from the FutureNanoNeeds project funded by the European Union Seventh Framework Programme (FP7-NMP-2013-LARGE-7) under grant agreement no. 604602.

NJ acknowledges support from the NanoReg2 project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement 646221.

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

Appendix A.

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. Many of the answers were summarized in Sections 3 and 4, leading to the recommendations found in Sections 6 and 7. In this Appendix, additional details about the stakeholder responses are provided.

Appendix A. A.1.: Stakeholder experience in nanomaterial data integration

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. 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 nonnanotechnology resources. Reponses to questions relating to experience in data integration, including integrating primary data sets and web services supporting data integration are provided in Table A-1.

Appendix A. A.2.: 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. Stakeholders were asked for information on existing resource functionality supporting data integration including data standards, controlled vocabulary, and common identifiers. Federated approaches may not require the actual movement of the data, but require 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 A-2). These stakeholder experiences provide insights into the level of readiness the nanotechnology community has achieved with regards to integrating databases and data sets.

Appendix A. A.3.: 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 below along with capabilities relating specifically to web services.

Appendix A. A.3.1.: caNanoLab Web Services

caNanoLab implements an internal and external API leveraging REST (see Table A-3). 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.

Appendix A. A.3.2.: CEINT web services

CEINT is developing beta web services for collaborator data set curation and sharing; this functionality is under development and intended for active collaborators to use during research, not for the broader public. 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 up-stream collaboration, even in the absence of prescribed data templates that would support more automated integration. CEINT uses web services provided by others, including eNanoMapper, the Nanomaterial

Registry, the Integrated Taxonomic Information System, Ontobee, caNanoLab, USDA Geospatial Data Gateway, and the Project on Emerging Nanotechnologies.

Appendix A. A.3.3.: CSSP/NIPHE, Netherlands (The Center for Safety of Substances and Products), National Institute for Public Health and the Environment, Web Services

CSSP/NIPHE, Netherlands does not offer web services; however, the OCHEM database (https://www.ochem.eu/) is publicly available.

Appendix A. A.3.4.: DECHEMA web services

DECHEMA does not provide any web services per se for the DaNa project. In the case of the NANORA project (http://www.nanora.eu/), 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 (http://www.nanoobjects.info/en/) 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 (Atli et al., 2011; Kimmig et al., 2014). DECHEMA does not use any web services provided by other organizations.

Appendix A. A.3.5.: 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 OpenTox services (Jeliazkova and Jeliazkova, 2011) and more recently eNanoMapper database (Jeliazkova et al., 2015), 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 https://data.enanomapper.net/ instance currently hosts several public data sets, available under an Open Data license or waiver. Several more instances are currently available, hosting data from past EU FP7 funded project, and integrated view of these is provided at https://search.data.enanomapper.net/. The eNanoMapper server currently does not use other web services, besides being able to retrieve chemical structures from public databases (e.g. PubChem).

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 (Jeliazkova et al., 2015)(Chomenidis et al., 2017).

Appendix A. A.3.6.: Nanomaterial registry

The Nanomaterial Registry does not currently have data exchange web services other than the export tools described in Table A-1. 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.

Appendix A. A.3.7.: 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. Web links to the primary data and to the principle investigators who have provided data to the NIL are included.

References

ASTM Standard E2456, 2006. E2456. 2006. Stand. Terminol. Relat. Nanotechnology. ASTM Int, West Conshohocken, PA. 10.1520/E2456-06R12.

ASTM Standard E2909, 2013. E2909. 2013. Stand. Guide Investig. Tab-Delimited Format Ofr Nanotechnologies ISA-TAB-Nano Stand. File Format Submiss. Exch. Data Nanomater. Charact. ASTM Int, West Conshohocken, PA. 10.1520/E2909-13.

Atli A, Nau K, Schmidt A, 2011. Navigation Along Database Relationships-An Adaptive Framework for Presenting Database Contents as Object Graphs. WEBIST, pp. 372–379.

Batchelor C, Brenninkmeijer CY, Chichester C, Davies M, Digles D, Dunlop I, Evelo CT, Gaulton A, Goble C, Gray AJ, 2014. Scientific lenses to support multiple views over linked chemistry data. In: International Semantic Web Conference. Springer, pp. 98–113.

Berners-Lee T, Hendler J, Lassila O, 2001. The semantic web. Sci. Am 284, 28–37.

Boholm M, Arvidsson R, 2016. A definition framework for the terms nanomaterial and nanoparticle. NanoEthics 10, 25–40. 10.1007/s11569-015-0249-7.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Vingron M, 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat. Genet 29, 365–371. 10.1038/ng1201-365. [PubMed: 11726920]

Brenninkmeijer C, Evelo C, Goble C, Gray AJ, Groth P, Pettifer S, Stevens R, Williams AJ, Willighagen EL, 2012. Scientific lenses over linked data: an approach to support task specific views of the data. A vision In: Proceedings of 2nd International Workshop on Linked Science, 10.1007/978-3-319-11964-9_7.

Chambers J, Davies M, Gaulton A, Papadatos G, Hersey A, Overington JP, 2014. UniChem: extension of InChI-based compound mapping to salt, connectivity and stereochemistry layers. J. Cheminformatics 6, 43. 10.1186/s13321-014-0043-5.

Cheung K-H, Frost HR, Marshall MS, Prud'hommeaux E, Samwald M, Zhao J, Paschke A, 2009. A journey to Semantic Web query federation in the life sciences. BMC Bioinforma 10, S10. 10.1186/1471-2105-10-S10-S10.

Chomenidis C, Drakakis G, Tsiliki G, Anagnostopoulou E, Valsamis A, Doganis P, Sopasakis P, Sarimveis H, 2017. Jaqpot Quattro: a novel computational web platform for modeling and analysis in nanoinformatics. J. Chem. Inf. Model 57, 2161–2172. 10.1021/acs.jcim.7b00223. [PubMed: 28812890]

Christen P, 2012. Data Matching: Concepts and Techniques for Record Linkage, EntityResolution, and Duplicate Detection. Springer Science & Business Media

Doan A, Halevy A, Ives Z, 2012. Principles of Data Integration Elsevier.

Eyres TP, 2013. Extracting more value from data silos: using the semantic web to link chemistry and biology for innovation. EMBnet J 19, 36–39. 10.14806/ej.19.B.725.

Field D, Sansone S-A, Collis A, Booth T, Dukes P, Gregurick SK, Kennedy K, Kolar P, Kolker E, Maxon M, .Wilbanks J, 2009. Omics data sharing. Science 326, 234–236. 10.1126/science.1180598. [PubMed: 19815759]

Fielding RT, Taylor RN, 2000. Architectural Styles and the Design of Network-based Software Architectures University of California (Irvine Doctoral dissertation).

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Zhang J, 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5, R80. 10.1186/gb-2004-5-10-r80. [PubMed: 15461798]

Gruber TR, 1995. Toward principles for the design of ontologies used for knowledge sharing? Int. J. Hum. Comput. Stud 43, 907–928. 10.1006/ijhc.1995.1081.

Harper SL, Dahl JA, Maddux BL, Tanguay RL, Hutchison JE, 2008. Proactively designing nanomaterials to enhance performance and minimise hazard. Int. J. Nanotechnol 5, 124–142. 10.1504/IJNT.2008.016552.

Hartig O, Langegger A, 2010. A database perspective on consuming linked data on the web. Datenbank-Spektrum 10, 57–66.

Hastings J, Chepelev L, Willighagen E, Adams N, Steinbeck C, Dumontier M, 2011. The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web. PLoS One 6, e25513. 10.1371/journal.pone.0025513. [PubMed: 21991315]

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

Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C, 2015b. ChEBI in 2016: improved services and an expanding collection of metabolites. Nucleic Acids Res 44, D1214–D1219. 10.1093/nar/gkv1031. [PubMed: 26467479]

Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D, 2015. InChI, the IUPAC international chemical identifier. J. Cheminformatics 7, 23 (doi: 10.1186/s13321-015-0068-4).

Hendren CO, Powers CM, Hoover MD, Harper SL, 2015. The nanomaterial data curation initiative: a collaborative approach to assessing, evaluating, and advancing the state of the field. Beilstein J. Nanotechnol 6, 1752–1762. 10.3762/bjnano.6.179. [PubMed: 26425427]

Hersey A, Chambers J, Bellis L, Patricia Bento A, Gaulton A, Overington JP, 2015. Chemical databases: curation or integration by user-defined equivalence? Drug Discov. Today Technol 14, 17–24. 10.1016/j.ddtec.2015.01.005. [PubMed: 26194583]

Hoover MD, Myers DS, Cash LJ, Guilmette RA, Kreyling WG, Oberdörster G, Smith R, Cassata JR, Boecker BB, Grissom MP, 2015. Application of an informatics-based decision-making framework and process to the assessment of radiation safety in nanotechnology. Health Phys 108, 179–194. 10.1097/HP.0000000000000250. [PubMed: 25551501]

van Iersel MP, Pico AR, Kelder T, Gao J, Ho I, Hanspers K, Conklin BR, Evelo CT, 2010. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics 11, 5. 10.1186/1471-2105-11-5. [PubMed: 20047655] de la Iglesia D, Harper S, Hoover MD, Klaessig F, Lippell P, Maddux B, Morse J, Nel A, Rajan K, Reznik-Zellen R, 2011. Nanoinformatics 2020 Roadmap 10.4053/rp001-110413.

ISO T, 2007. 80004–1: Nanotechnologies-Vocabulary-Part 1: Core Terms. 2007. Int.Stand. Organ, Geneva Switz.

Izak-Nau E, Huk A, Reidy B, Uggerud H, Vadset M, Eiden S, Voetz M, Himly M, Duschl A, Dusinska M, Lynch I, 2015. Impact of storage conditions and storage time on silver nanoparticles' physicochemical properties and implications for their biological effects. RSC Adv 5, 84172–84185. 10.1039/C5RA10187E.

Jeliazkova N, Jeliazkova V, 2011. AMBIT RESTful web services: an implementation of the OpenTox application programming interface. J. Cheminformatics 3, 18. 10.1186/1758-2946-3-18.

Jeliazkova N, Chomenidis C, Doganis P, Fadeel B, Grafström R, Hardy B, Hastings J, Hegi M, Jeliazkov V, Kochev N, Willighagen E, 2015. The eNanoMapper database for nanomaterial safety information. Beilstein J. Nanotechnol 6, 1609. 10.3762/bjnano.6.165. [PubMed: 26425413]

Jupp S, Malone J, Bolleman J, Brandizi M, Davies M, Garcia L, Gaulton A,Gehant S, Laibe C, Redaschi N, Jankinson A, 2014. The EBI RDF platform: linked open data for the life sciences. Bioinformatics 30, 1338–1339. 10.1093/bioinformatics/btt765. [PubMed: 24413672]

Juty N, Le Novère N, Laibe C, 2012. Identifiers. org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res 40, D580–D586. 10.1093/nar/gkr1097. [PubMed: 22140103]

Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, Bryant SH, 2015. PubChem substance and compound databases. Nucleic Acids Res 44, D1202–D1213. 10.1093/nar/gkv951. [PubMed: 26400175]

Kimmig D, Marquardt C, Nau K, Schmidt A, Dickerhof M, 2014. Considerations about the implementation of a public knowledge base regarding nanotechnology. Comput. Sci. Discov 7, 014001. 10.1088/1749-4699/7/1/014001.

Klimisch H-J, Andreae M, Tillmann U, 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul. Toxicol. Pharmacol 25, 1–5. 10.1006/rtph.1996.1076.

Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Wishart D, 2011. DrugBank 3.0: a comprehensive resource for “Omics” research on drugs. Nucleic Acids Res 39, D1035–D1041. 10.1093/nargkq1126. [PubMed: 21059682]

Kohonen P, Parkkinen JA, Willighagen EL, Ceder R, Wennerberg K, Kaski S, Grafström RC, 2017. A transcriptomics data-driven gene space accurately predicts liver cytopathology and drug-induced liver injury. Nat. Commun 8. 10.1038/ncomms15932.

Kovrižnych JA, Sotníková R, Zeljenková D, Rollerová E, Szabová E, Wimmerová S, 2013. Acute toxicity of 31 different nanoparticles to zebrafish (Danio rerio) tested in adulthood and in early life stages–comparative study. Interdiscip. Toxicol 6, 67–73. 10.2478/intox-2013-0012. [PubMed: 24179431]

Krug HF, 2014. Nanosafety research—are we on the right track? Angew. Chem. Int. Ed53, 12304–12319. 10.1002/anie.201403367.

Kühnel D, Marquardt C, Nau K, Krug HF, Mathes B, Steinbach C, 2014. Environmental impacts of nanomaterials: providing comprehensive information on exposure, transport and ecotoxicity-the project DaNa2. 0. Environ. Sci. Eur 26, 21. 10.1186/s12302-014-0021-6.

Longo DL, Drazen JM, 2016. More on data sharing. N. Engl. J. Med 374, 1896. 10.1056/NEJMc1602586.

Maglott D, Ostell J, Pruitt KD, Tatusova T, 2005. Entrez gene: gene-centered information at NCBI. Nucleic Acids Res 33, D54–D58. 10.1093/nar/gki031. [PubMed: 15608257]

Marchese-Robinson RL, Cronin MT, Richarz A-N, Rallo R, 2015. An ISA-TAB-Nano based data collection framework to support data-driven modelling of nanotoxicology. Beilstein J. Nanotechnol 6, 1978. 10.3762/bjnano.6.202. [PubMed: 26665069]

Marchese-Robinson RL, Lynch I, Peijnenburg W, Rumble J, Klaessig F, Marquardt C, Rauscher H, Puzyn T, Purian R, Åberg C, Harper S, 2016. How should the completeness and quality of curated nanomaterial data be evaluated? Nano 8, 9919–9943. 10.1039/C5NR08944A.

Marshall MS, Boyce R, Deus HF, Zhao J, Willighagen EL, Samwald M, Pichler E, Hajagos J, Prud'hommeaux E, Stephens S, 2012. Emerging practices for mapping and linking life sciences data using RDF—a case series. Web Semant. Sci. Serv. Agents World Wide Web 14, 2–13. 10.1016/j.websem.2012.02.003.

Miller AL, Hoover MD, Mitchell DM, Stapleton BP, 2007. The Nanoparticle Information Library (NIL): a prototype for linking and sharing emerging data. J. Occup. Environ. Hyg 4, D131–D134. 10.1080/15459620701683947. [PubMed: 17924276]

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

Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey M-A, Chute CG, Musen M, 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 37, W170–W173. 10.1093/nar/gkp440. [PubMed: 19483092]

Oksel C, Ma CY, Wang XZ, 2015. Current situation on the availability of nanostructure–biological activity data. SAR QSAR Environ. Res 26, 79–94. 10.1080/1062936X.2014.993702. [PubMed: 25608859]

Powers CM, Mills KA, Morris SA, Klaessig F, Gaheen S, Lewinski N, Hendren CO, 2015. Nanocuration workflows: establishing best practices for identifying, inputting, and sharing data to inform decisions on nanomaterials. Beilstein J. Nanotechnol 6, 1860. 10.3762/bjnano.6.189. [PubMed: 26425437]

Rauscher H, Sokull-Klüttgen B, Stamm H, 2012. The European Commission's recommendation on the definition of nanomaterial makes an impact. Nanotoxicology 7, 1195–1197. 10.3109/17435390.2012.724724. [PubMed: 22920756]

Reichman OJ, Jones MB, Schildhauer MP, 2011. Challenges and opportunities of open data in ecology. Science 331, 703–705. 10.1126/science.1197962. [PubMed: 21311007]

Roco MC, 2011. The long view of nanotechnology development: the national nano-technology initiative at 10 years. In: Roco MC, Hersam MC, Mirkin CA (Eds.), Nanotechnology Research Directions for Societal Needs in 2020: Retrospective and Outlook Springer, Netherlands, Dordrecht, pp. 1–28. 10.1007/978-94-007-1168-6_1.

Roco MC, Mirkin CA, Hersam MC, 2011. Nanotechnology Research Directions for Societal Needs in 2020 Springer, Netherlands.

Rumble J, Freiman S, Teague C, 2014. The description of nanomaterials: a multi-disciplinary uniform description system. 2014 IEEE Int. Conf. Bioinforma. Biomed. BIBM 34–39. 10.1109/BIBM.2014.6999372.

Samwald M, Jentzsch A, Bouton C, Kallesøe CS, Willighagen E, Hajagos J,Marshall MS, Prud'hommeaux E, Hassanzadeh O, Pichler E, 2011. Linked open drug data for pharmaceutical research and development. J. Cheminformatics 3, 19. 10.1186/1758-2946-3-19.

Sansone S-A, Rocca-Serra P, Brandizi M, Brazma A, Field D, Fostel J, Garrow AG, Gilbert J, Goodsaid F, Hardy N, Wiemann S, Members of the RSBI Working Group, 2008. The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?" OMICS. J. Integr. Biol 12, 143–149. 10.1089/omi.2008.0019.

Savolainen K, Backman U, Brouwer D, Fadeel B, Fernandes T, Kuhlbusch T, Landsiedel R, Lynch I, Pylkkänen L, 2013. Nanosafety in Europe 2015–2025: Towards Safe and Sustainable Nanomaterials and Nanotechnology Innovations

Shao C-Y, Chen S-Z, Su B-H, Tseng YJ, Esposito EX, Hopfinger AJ, 2013. Dependence of QSAR models on the selection of trial descriptor sets: a demonstration using nanotoxicity endpoints of decorated nanotubes. J. Chem. Inf. Model 53, 142–158. 10.1021/ci3005308. [PubMed: 23252880]

Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu W-L, Wright LW, 2007. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J. Biomed. Inform 40, 30–43. 10.1016/j.jbi.2006.02.013. [PubMed: 16697710]

Stefaniak AB, Hackley VA, Roebben G, Ehara K, Hankin S, Postek MT, Lynch I, Fu W-E, Linsinger TP, Thünemann AF, 2013. Nanoscale reference materials for environmental, health and safety measurements: needs, gaps and opportunities. Nanotoxicology 7, 1325–1337. 10.3109/17435390.2012.739664. [PubMed: 23061887]

Thomas DG, Klaessig F, Harper SL, Fritts M, Hoover MD, Gaheen S, Stokes TH,Reznik-Zellen R, Freund ET, Klemm JD, Paik D, Baker NA, 2011a. Informatics and standards for nanomedicine technology. Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol 3, 511–532. 10.1002/wnan.152. [PubMed: 21721140]

Thomas DG, Pappu RV, Baker NA, 2011b. NanoParticle Ontology for cancer nanotechnology research. J. Biomed. Inform 44, 59–74. 10.1016/j.jbi.2010.03.001. [PubMed: 20211274]

Thomas DG, Gaheen S, Harper SL, Fritts M, Klaessig F, Hahn-Dantona E, Paik D, Pan S, Stafford GA, Freund ET, Klemm JD, Baker NA, 2013. ISA-TAB-Nano: a specification for sharing nanomaterial research data in spreadsheet-based format. BMC Biotechnol 13 (1). 10.1186/1472-6750-13-2.

Totaro S, Crutzen H, Sintes Riego J, 2017. Data Logging Templates for theEnvironmental, Health and Safety Assessment of Nanomaterials 10.2787/505397.

Truong L, Harper SL, Tanguay RL, 2011. Evaluation of embryotoxicity using the zebrafish model. Drug Saf. Eval. Methods Protoc 271–279. 10.1007/978-1-60761-849-216.

Vance ME, Kuiken T, Vejerano EP, McGinnis SP, Hochella MF Jr, Rejeski D, Hull MS, 2015. Nanotechnology in the real world: redeveloping the nanomaterial consumer products inventory. Beilstein J. Nanotechnol 6, 1769. 10.3762/bjnano.6.181. [PubMed: 26425429]

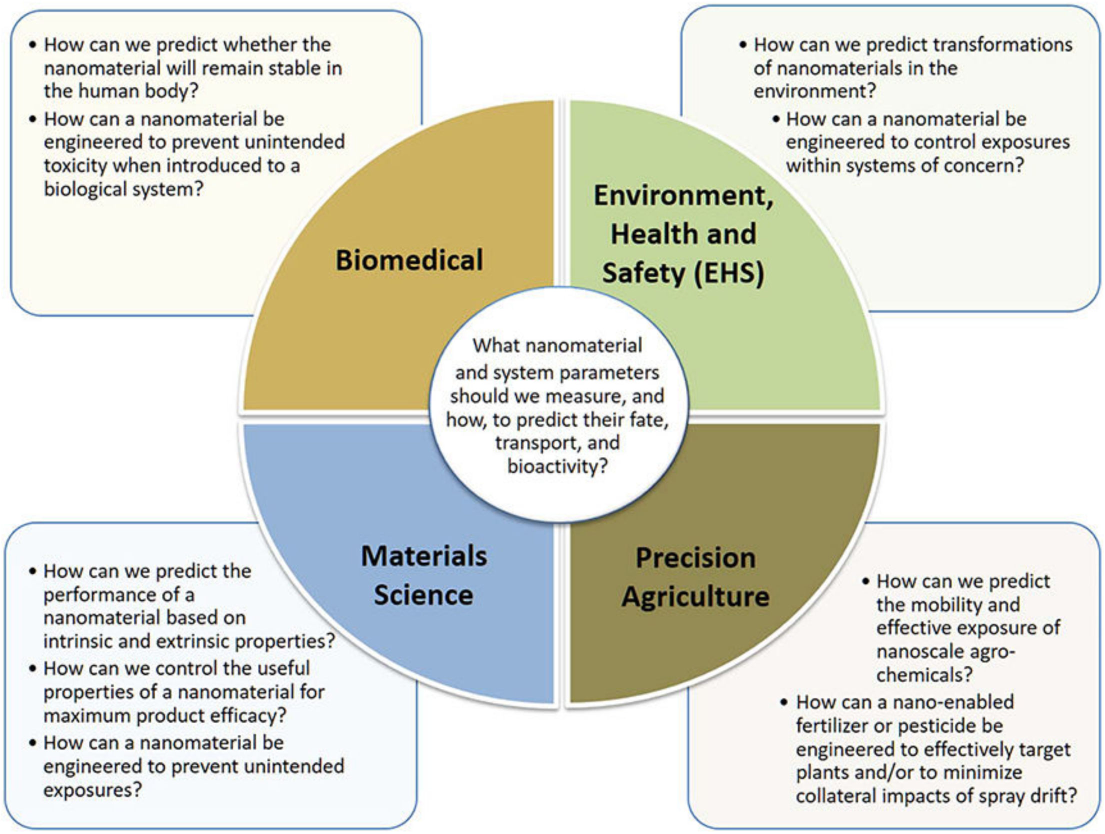
Wang Y, Suzek T, Zhang J, Wang J, He S, Cheng T, Shoemaker BA, Gindulyte A, Bryant SH, 2014. PubChem bioassay: 2014 update. Nucleic Acids Res 42, D1075–D1082. 10.1093/nar/gkt978. [PubMed: 24198245]

Wilkinson MD, Dumontier M, Aalbersberg Ij.J., Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Mons B, 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3. 10.1038/sdata.2016.18.

Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, Evelo CT, Blomberg N, Ecker G, Goble C, 2012. Open PHACTS: semantic interoper-ability for drug discovery. Drug Discov. Today 17, 1188–1198. [PubMed: 22683805]

Wohlgemuth G, Haldiya PK, Willighagen E, Kind T, Fiehn O, 2010. The Chemical Translation Service—a web-based tool to improve standardization of metabolomic reports. Bioinformatics 26, 2647–2648. 10.1093/bioinformatics/btq476. [PubMed: 20829444]

Xia Y, 2014. Are we entering the nano era? Angew. Chem. Int. Ed 53, 12268–12271. 10.1002/anie.201406740.



What nanomaterial and system parameters should we measure, and how, to predict their fate, transport, and bioactivity?

- How can we predict whether the nanomaterial will remain stable in the human body?
- How can a nanomaterial be engineered to prevent unintended toxicity when introduced to a biological system?

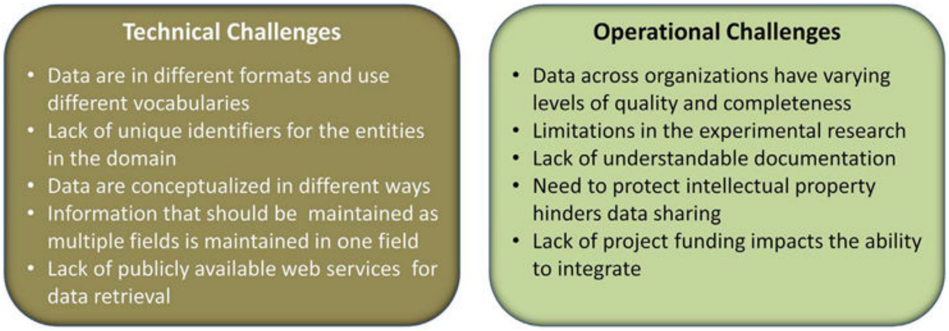
- How can we predict transformations of nanomaterials in the environment?
- How can a nanomaterial be engineered to control exposures within systems of concern?

- How can we predict the performance of a nanomaterial based on intrinsic and extrinsic properties?
- How can we control the useful properties of a nanomaterial for maximum product efficacy?
- How can a nanomaterial be engineered to prevent unintended exposures?

- How can we predict the mobility and effective exposure of nanoscale agro-chemicals?
- How can a nano-enabled fertilizer or pesticide be engineered to effectively target plants and/or to minimize collateral impacts of spray drift?

Fig. 1.

Examples of use cases that can be addressed and might mutually benefit from data integration.



Technical Challenges

- Data are in different formats and use different vocabularies
- Lack of unique identifiers for the entities in the domain
- Data are conceptualized in different ways
- Information that should be maintained as multiple fields is maintained in one field
- Lack of publicly available web services for data retrieval

Operational Challenges

- Data across organizations have varying levels of quality and completeness
- Limitations in the experimental research
- Lack of understandable documentation
- Need to protect intellectual property hinders data sharing
- Lack of project funding impacts the ability to integrate

Fig. 2.

Technical and operational challenges impacting data integration.

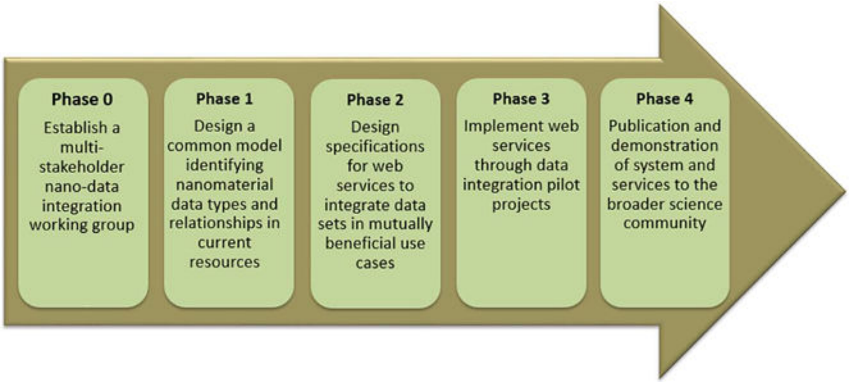

Phase 0
Establish a multi-stakeholder nano-data integration working group
Phase 1
Design a common model identifying nanomaterial data types and relationships in current resources
Phase 2
Design specifications for web services to integrate data sets in mutually beneficial use cases
Phase 3
Implement web services through data integration pilot projects
Phase 4
Publication and demonstration of system and services to the broader science community

Fig. 3.

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

Table 1

Common web services envisaged by nanoinformatics stakeholders, as reported in the survey, as being needed to support integration of nanomaterial data in the biomedical nano-technology and nanosafety domains.

Web service method	Description
Creation of an identifier	Creates a Universally Unique Identifier (UUID) for any entity such as a material, characterization, protocol, or publication
Characterization retrieval	Retrieves characterizations for a material by material type and characterization type (e.g. size) and returns characterization data in JSON and XML format
Get data by DOI	Returns (pointers to) entries in the database with information about or from a specific publication
Get data by PubMed ID	Returns (pointers to) entries in the database with information about or from a specific publication
Get identifier	Retrieves a UUID for any entity such as a material, characterization, protocol, or publication
Get ISA-TAB-Nano file	Retrieves ISA-TAB-Nano files associated with a publication (DOI, PubMed)
Get investigation Get material	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. 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
Get protocol	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
Get publication	Retrieves publications associated with a material, characterization, and/or protocol, and returns a DOI, PubMed ID, and/or URL to the publication
Get study	Retrieves a study associated with a specific assay type and/or nanomaterial type and re-turns a study in JSON or XML format; the JSON and XML format would be based on metadata from ISA-TAB-Nano
Search by chemistry	Retrieves nanomaterials based on chemical structure or chemical similarity. Supports a function such as: “Find the most similar structure in database X”

Table 2

Non-nanotechnology resources needed to support use case driving data integration.

Non-nanotechnology resource	Description or example
Life Sciences and Chemistry Databases	Life science and chemistry databases in general, containing information about human biology (both experimental data, as well as knowledgebases) and chemistry (functionality, chemical structure, etc.) (Kim et al., 2015; Wang et al., 2014); needed to inform the design of new nanomaterials to avoid potential negative influences on human health
Image archives	The National Biomedical Imaging Archive (NBIA) (https://imaging.nci.nih.gov/ncia/login.jsf); the Cancer Image Archive (TCIA) (http://www.cancerimagingarchive.net/), or other image archives 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	The Molecular Imaging and Contrast Agent Database (MICAD) (https://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	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 to link nanomaterial data to nanomaterial publications; an example of this is the caNanoLab interface with Science Direct publications through Elsevier
Clinical Trials Management Systems (CTMS)	OpenClinica (https://www.openclinica.com/) to access clinical data associated with the use of nanomaterials in human clinical trials
Genomic Data/ Biomarker Repositories	Repositories such as the NCI Genomic Data Commons (https://gdc.cancer.gov/) to maintain molecular data for transfection and targeting characterization involving nanomaterials and NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) to achieve high-throughput functional genomics data
Chemical and Agent Repositories	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 (Knox et al., 2011) 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 bio-distribution and toxicity; integrating with 3D modeling tools such as The Collaboratory for Structural Nanobiology https://ncifrederick.cancer.gov/dsitp/abcc/abcc-groups/simulation-and-modeling/collaboratory-for-structural-nanobiology/to facilitate the display on nanomaterial structures in 3D leveraging a Protein Data Bank (PDB) file, offering prediction options for adverse effects of nanomaterials (Chomenidis et al., 2017)
Analysis and visualization tools	Includes various tools such as R (https://www.r-project.org/, an environment for statistical computing), and Bioconductor (Gentleman et al., 2004), Data-Driven Documents https://d3js.org/ 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 such as the NCI Thesaurus http://evs.nci.nih.gov/ftp1/NCI_Thesaurus (Sioutos et al., 2007), BioPortal http://bioportal.bioontology.org/ (Noy et al., 2009), and Ontobee http://www.ontobee.org/. This will allow databases to link to term references and accession numbers. (N.B.: as discussed above, ontology annotations support data integration. Hence integration of two resources with ontology terms supports wider integration.

Table 3

Levels of equivalence. The equivalence strengths are meant to indicate how data are intended to be combined and do not specify why they should be linked in that manner.

Equivalence strength	Semantic equivalence	Description	Example
Strong	Web Ontology Language (OWL) “same as”	Two nanomaterials that share the same properties: all properties for one are valid for the other; moreover, if one nanomaterial is the “same as” other nanomaterials, the others are equally strong (transitivity).	A nanomaterial reported in a journal article for which information is compiled in two databases
Moderate	Simple Knowledge Organization System (SKOS) “close match”	Two nanomaterials are said to be the “same” only for a certain specified application; this match is never transitive	Two nanomaterials from the same production batch, in which the application ignores intra-batch variation
Weak	SKOS “related match”	Two nanomaterials are merely linked together, with an undefined similarity	Two nanomaterials from the Joint Research Centre - Health, Consumers & Reference Materials Directorate - 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. (Izak-Nau et al., 2015)

Table A-1

Integration capabilities of responding nanoinformatics resources.

Nanotechnology resource	Integration capabilities
caNanoLab https://cananolab.nci.nih.gov/	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 Science Direct publications through an Elsevier bi-directional link and uses the PubMed and PubChem interfaces.
CEINT NIKC (NanoInformatics Knowledge Commons) http://www.ceint.duke.edu/	Integration within the CEINT NIKC resource is achieved by cross-training lead curators within key collaborator teams in the consistent manual curation process utilizing shared templates and consistent valid values. The reviewed combined data set is then ported to the NIKC via custom API for these targeted data sets.
Center for Safety of Substances and Products, National Institute for Public Health and the Environment (CSSP/NIHE) http://www.rivm.nl/en/About_RIVM/Organisation/Centres/	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://nanopartikel.info/en/projects/current-projects/dana-2-0	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 Database https://data.enanomapper.net/ Search integration of several databases: https://search.data,enanomapper.net	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	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/	Integration with the NIL is achieved on a case-by case-basis.

Table A-2

Summary of stakeholder responses to upload, download, and mapping questions: Does the nanomaterial data 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 https://ncit.nci.nih.gov/ncitbrowser/	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 (with- out middle name), custom abbreviation of journal title, year of publication, and sample sequential number, e.g. SNLJJNM-CAshleyACSNano2012–01.	caNanoLab integrates loosely with six non-nano resources[b].
CEINT	Mapping directly from NBI data set; curating literature data; directly integrating data from customized templates built with active collaborators	Under development	Uses ontologies such as MO, NPO, UO, ChEBI, and eNanoMapper; also compatible with ISA-TAB-Nano.	Nanomaterial associated to data source and assigned a unique identifier	Includes some non-nanomaterial data
CSSP/NIPHE, Netherlands	Commonly uses 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 Kühnel et al. (2014)	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 Jeliazkova et al. (2015)	Uses the eNanoMapper ontology (composed of NPO, ChEBI, BFO, IAO, CHEMINF and others)	Uses a substance UUID[d]	Not currently
Nanomaterial Registry	Export for physico-chemical characterization	Nanomaterial Registry glossary https://nanohub.org/groups/nanomaterialregistry	Uses a controlled Vocabulary[e]	Uses unique numeric IDs[f]	Not currently
Nanoparticle Information-on 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

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
					materials (Miller et al., 2007)[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 (https://ncifrederick.cancer.gov/dsitp/abcc/abcc-groups/simulation-and-modeling/collaboratory-for-structural-nanobiology/) for displaying 3D models of specific nanomaterials, and Nanotechnology Characterization Laboratory (NCL, http://ncl.cancer.gov/) assay cascade and JoVE (https://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 and not losing the scientific correctness.

d eNanoMapper is based on semantic web technologies including referenceable 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 approximately 8–10 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 (https://www.cdc.gov/niosh/npg/default.html) 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.

Table A-3

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

Search type	Possible search criteria
Protocol	Protocol name
Sample	Specific sample, composition, and/or characterization
Publication	Sample name. Nanomaterial characteristics