Published in final edited form as:
Comput Sci Discov. ; 6(1): 014008–. doi:10.1088/1749-4699/6/1/014008.

Nanoinformatics workshop report: Current resources, community needs, and the proposal of a collaborative framework for data sharing and information integration

Stacey L Harper[1,2], **James E Hutchison**[3], **Nathan Baker**[4], **Michele Ostraat**[5], **Sally Tinkle**[6], **Jeffrey Steevens**[7], **Mark D Hoover**[8], **Jessica Adamick**[9], **Krishna Rajan**[10], **Sharon Gaheen**[11], **Yoram Cohen**[12], **Andre Nel**[13], **Raul E Cachau**[14], and **Mark Tuominen**[15]

Stacey L Harper: Stacey.Harper@OregonState.edu

[1]Department of Environmental and Molecular Toxicology, Oregon State University, Corvallis, OR, USA.

[2]School of Chemical, Biological and Environmental Engineering, Oregon State University, Corvallis, OR, USA.

[3]Department of Chemistry, University of Oregon, Eugene, OR, USA.

[4]Pacific Northwest National Laboratory, Richland, WA, USA.

[5]RTI International, Research Triangle Park, NC, USA.

[6]Science and Technology Policy Institute, Washington, DC, USA.

[7]US Army Engineer Research and Development Center, Vicksburg, Mississippi, USA.

[8]National Institute for Occupational Safety and Health, Morgantown, WV, USA.

[9]University Libraries, University of Michigan, Amherst, MA, USA.

[10]Department of Materials Science and Engineering, Iowa State University, Ames, IA, USA.

[11]SAIC Frederick, Frederick, MD, USA.

[12]Department of Chemical and Biomolecular Engineering, University of California Los Angeles, CA, USA.

[13]Center for Environmental Implications of Nanotechnology, University of California Los Angeles, CA, USA.

[14]SAIC Frederick, Inc. Advanced Biomedical Computer Center, Information Systems Program, Frederick National Laboratory for Cancer Research, Frederick, MD, USA.

[15]Physics Department, University of Massachusetts, Amherst, MA, USA.

Abstract

The quantity of information on nanomaterial properties and behavior continues to grow rapidly. Without a concerted effort to collect, organize and mine disparate information coming out of current research efforts, the value and effective use of this information will be limited at best. Data will not be translated to knowledge. At worst, erroneous conclusions will be drawn and future research may be misdirected. Nanoinformatics can be a powerful approach to enhance the value of global information in nanoscience and nanotechnology. Much progress has been made through grassroots efforts in nanoinformatics resulting in a multitude of resources and tools for nanoscience researchers. In 2012, the nanoinformatics community believed it was important to critically evaluate and refine currently available nanoinformatics approaches in order to best inform the science and support the future of predictive nanotechnology. The Greener Nano 2012:

Nanoinformatics Tools and Resources Workshop brought together informatics groups with materials scientists active in nanoscience research to evaluate and reflect on the tools and resources that have recently emerged in support of predictive nanotechnology. The workshop goals were to establish a better understanding of current nanoinformatics approaches and to clearly define immediate and projected informatics infrastructure needs of the nanotechnology community. The theme of nanotechnology environmental health and safety (nanoEHS) was used to provide real-world, concrete examples on how informatics can be utilized to advance our knowledge and guide nanoscience. The benefit here is that the same properties that impact the performance of products could also be the properties that inform EHS. From a decision management standpoint, the dual use of such data should be considered a priority. Key outcomes include a proposed collaborative framework for data collection, data sharing and information integration.

1. Workshop description

The Greener Nano 2012 (GN12): Nanoinformatics Tools and Resources Workshop (http://nanoinformatics.org/) was sponsored by the Safer Nanomaterials and Nanomanufacturing Initiative, National Nanomanufacturing Network, University of California Center for Environmental Implications of Nanotechnology and the National Cancer Institute Nanotechnology Working Group with participation from the National Nanotechnology Coordination Office. The workshop brought in over 70 participants nationally and internationally which were representative of academics (US and abroad), government bodies (NNCO, NIESH, NIOSH, State Agencies, military), national laboratories, and industry. An expert panel of materials scientists and engineers active in nanoscience research were invited to participate in the workshop as experts from whom the nanoinformatics community can tap for content knowledge. This article aims to summarize the workshop and present a community-driven vision for the advancement of nanoscience and nanotechnology through a collaborative framework for data collection, data sharing, and information integration.

1.1. Goals

The goals of the workshop were to:

- Establish a better understanding of state-of-the-art approaches to nanoinformatics, and
- Clearly define immediate and projected informatics infrastructure needs for the nanotechnology community.

These goals are consistent with the objectives of the community-based Nanoinformatics 2020 Roadmap (http://eprints.internano.org/607/1/Roadmap_FINAL041311.pdf) [1], including the 2020 Roadmap's working definition that:

- Nanoinformatics is the science and practice of determining which information is relevant to the nanoscale science and engineering community, and then developing and implementing effective mechanisms for collecting, validating, storing, sharing, analyzing, modeling, and applying that information.

1.2 Theme

The theme of *nanotechnology environmental health and safety (nanoEHS)* was used throughout the workshop because of its dual role in advancing knowledge to protect public health and the environment while also providing useful information for nanomaterial design and development.

1.3 Pre-workshop webinars

Two pre-workshop webinars were held to establish a shared understanding of foundational concepts that would maximize participant engagement in, and results from a one-day workshop. The first webinar provided 15 minute overviews of current nanoinformatics tools with detailed information on our current state-of-the-science. Webinars on the following resources were recorded and are available at http://nanoinformatics.org/2012/webinar:

- *caNanoLab* - Sharon Gaheen, SAIC Frederick;
- *GoodNanoGuide* - Kristen Kulinowski, Science and Technology Policy Institute, Rice University;
- *InterNano* - Mark Tuominen, University of Massachusetts Amherst;
- *ISA-TAB-Nano* - Sharon Gaheen, SAIC Frederick; Nathan Baker, Pacific Northwest National Laboratory;
- *NanoHUB* - Gerhard Klimeck, Purdue University;
- *Nanomaterial-Biological Interactions Knowledgebase* - Stacey Harper, Oregon Nanoscience and Microtechnologies Institute, Oregon State University;
- *Nanomaterial Registry* - Michele Ostraat, RTI International;
- *Nanoparticle Information Library* - Mark Hoover, National Institute for Occupational Safety and Health;
- *NEIMiner* - Kaizhi Tang, Intelligent Automation, Inc.; and
- *MendNano* - Yoram Cohen, University of California Los Angeles.

The second webinar provided live interactions with the developers of those nanoinformatics tools and resources prior to the workshop, allowed for preliminary discussions on the workshop topics, included an overview of the workshop events. These discussions enabled participants to be well prepared for group discussions during the formal workshop.

1.4 Opening and keynote addresses

The GN12: Nanoinformatics Tools and Resources Workshop opening remarks and charge to the group were presented by Stacey Harper, Oregon State University, after which Sally Tinkle, Deputy Director of the National Nanotechnology Coordination Office, presented an overview of the Nanotechnology Knowledge Infrastructure Signature Initiative and the Material Genome Initiative. Dr Tinkle's presentation and follow-on workshop discussions centered on identifying synergies and opportunities for these diverse scientific communities to work together on topics of mutual interest and benefit. A luncheon keynote address on "The Materials Genome Initiative Interface" was presented by Krishna Rajan of Iowa State University to provide a highly relevant example of the opportunities to apply the emerging principles of nanoinformatics in a practical and increasingly important setting.

1.5 Breakout sessions

The body of the workshop consisted of a series of discussion-oriented breakout sessions focused on:

i. data lifecycle to support a sustainable cyber-toolbox,

ii. use of nanoinformatics for predictive modeling, and

iii. the integration of nanoinformatics efforts.

2. Data lifecycle to support a sustainable cyber-toolbox for nanoinformatics

2.1 The data lifecycle

The data lifecycle, in the context of this workshop, refers to the way in which the community deals with data collected from a wide variety of studies and experiments that could ultimately be useful for developing conceptual or mathematical models that are beneficial for the nanoscience community. The discussion groups in this session examined participants' experiences and expectations about an *optimal* data lifecycle in comparison with current approaches for dealing with data. The discussions revealed existing infrastructure gaps and participants identified present barriers to sustainable nanoinformatics.

Each discussion group in this session began with a generic 'data pipeline' (generated during the second webinar) that represented the community's current use of data (Figure 1A). Figure 1B illustrates an improved strategy that was developed by group consensus to allow for the reuse of data and enhance the community capacity to establish structure-property relationships and predictive models for nanomaterials.

As shown in Figure 1A, the generic 'data pipeline' at present, involves raw data generated by independent researchers or research groups being processed, analyzed, and published by those groups. Few research programs currently have the resources, infrastructure, expertise or expectation to provide open access to the raw and processed data outside of the research team. As a result, published findings of processed data must be extracted from the primary literature to make use of the data for computational analysis and future predictive modeling. In fact, informatics approaches are being applied to automate this process. A major limitation of curating data from publications is the loss of data and data integrity. For example, a published investigation may highlight certain data in the study that fit the purpose of the article and may not necessarily report additional useful data that were collected and that could be of high value to other interested groups. In addition, there is a clear reluctance in the research community to publish data that does not demonstrate the sought after effects (e.g., identifying toxic nanomaterial). Yet, negative effect data are as important, if not more important, with respect to regulatory and risk management decision making. Studies on nanomaterial hazards could be minimized if all data were made more broadly available and used to build conceptual and mathematical models to improve our understanding. Likewise, the integration of such data would accelerate the translation of research findings to management practices; allowing for a rational prioritization of nanomaterials for testing and application purposes.

Through discussions and debate, each group worked through improvements to the current data lifecycle. Many workshop participants concluded that there was not just a single optimal lifecycle for data because different classes of nanomaterials and different potential applications pose different data requirements. For instance, nanomaterial production methods that result in batch-to-batch variation will require measurement validation methods. Data requirements for physicochemical information or biological assay results may be quite different depending on the specific application and intended use.

In the revised framework shown in Figure 1B, raw data are still processed, analyzed and published by individual researchers or groups; however, the data are also submitted to a federated, interoperable system of data repositories for broader access and analysis for modeling. Such a strategy will allow for inter-laboratory comparisons and definition of the error and variability within and between studies. This information and data validation is necessary for optimizing and informing subsequent studies. Re-informed data objectives are a reflection on the data in the context of why the data were needed. For example, researchers

would focus on the hypothesis that drove the research, while industry would place greater importance on whether the material properties were useful for their product needs. Although this step may be field specific, much can be gained by aligning the common data elements that would be useful for field-specific purposes as well as serve the dual role of being useful for predictive modeling and establishing structure-property relationships [2].

2.2 Barriers to rapid progress in nanoinformatics

Additional discussions involved identifying barriers to rapid progress in nanoinformatics including inadequate material description to establish cause and effect, the lack of available raw datasets which are not typically published in the literature, as well as insufficient protocols, reference materials, standards for data acquisition, processing and reporting.

Additional limitations include batch-to-batch variability and potential transformation of the nanomaterials in exposure media which constrain comparisons among datasets. As stated above, datasets may not be adequate or complete when comparing different classes of materials or analyzing different application scenarios.

Consensus during the workshop was reached regarding the need for appropriate materials definitions and reference materials that could feed into the informatics process and the need for standards to calibrate methods and to normalize datasets. Likewise, standardized vocabulary and terminology would allow for more consistency and less variability between datasets.

Another barrier to an optimal data lifecycle is assurances that proprietary data will be protected from disclosure if that is a concern. Efforts, such as those by National Institute of Environmental Health Sciences Centers for Nanotechnology Health Implications Research (NCNHIR) to fund the collection and dissemination of data, can catalyze the nanoscience community by providing much needed datasets and an example of coordinated, integrated research.

2.3 Emerging needs for a nanotechnology data ‘incubator’

One of the concepts which emerged from the GN12 discussions on data lifecycle is the need for a nanotechnology data ‘*incubator”* (Figure 2). Approaches to making progress towards an optimal data lifecycle include leveraging current islands of data that do exist to establish weighted models of high utility to the nanoscience community. The nanoscience community currently lacks high quality datasets that comprise testing of a sufficiently large number of well-characterized nanomaterials to support predictive modeling. While not all data are suitable for cause-effect analysis or modeling, there remains an immediate need to utilize available data to inform decision making.

As shown in Figure 2, datasets that are incomplete, lack validation, or are of low quality are considered to be in the incubator to ensure that the quality of data is not deluded for predictive informatics [3]. For instance, published works from 10 years ago may lack characterization of the nanomaterials altogether or may have been characterized or analyzed using inappropriate methodologies. These data could still be informative in the **zone of exploration** but would remain restricted to the incubator because they are not immediately useful beyond the scope of research for which they were collected and thus, there would not be a draw to build on such datasets. Data currently available in the virtual incubator could be used to discover interesting trends and generate hypotheses that guide development of new approaches for both the characterization of nanomaterials and the modeling of their behavior as illustrated in the example arrows in Figure 2. Sharing the datasets available in existing data repositories could speed the translation of data into knowledge (Figure 2A). The

development of effective data mining tools and integration methodologies will likewise accelerate this translation.

The incubator is a federation concept and not a formal data repository as it draws on data from a variety of resources to offer solutions to specific stakeholder sectors. In this way, data that comes out of the incubator is use-inspired and can be consolidated to allow for weight-of-the-evidence evaluations that inform end users. In this **zone of use-case driven research**, exploratory models (e.g., heat maps, clustering approaches, radar plots, self-organizing maps) can be applied to gain an understanding of data relationships.

The spokes emanating from the incubator can be tailored to a specific use scenario and material application. Although synergistic transformation of data into knowledge will occur when the spokes for multiple applications or multiple materials overlap, this is not currently the norm. For instance, information on potential drug candidates for cancer therapeutics could be combined with EHS considerations to develop a series of drugs that has been optimized for performance and safety (Figure 2 Arrow B). Research areas that have defined overlap in data requirements can leverage the same pools of data. For example, aquatic impacts that are analyzed for nanomaterials of interest for clean water applications could be used to inform the development of nanomaterials for environmental remediation (Figure 2 Arrow C).

Based on use-case driven research, commonalities and differences required as model input parameters can be determined and predictive models can be developed as data moves to the **zone of validation**. Figure 2 Arrow D illustrates the integration of data based on performance, EHS and cost in establishing new solar technologies. As data converge around specific materials or applications, knowledge gained can feed back into the zone of exploration.

3. Use of nanoinformatics for predictive modeling

Breakout discussions in the session on the use of nanoinformatics for predictive modeling were organized to focus on expected outcomes from predictive toxicology and mechanistic models in order to understand the level of nanomaterial representation/description that is required to parameterize such models. Real world examples were discussed to illustrate the value of nanoinformatics for informing nanoscience and nanotechnology design.

Each discussion group had a different theme to ensure coverage of a broad range of topics under predictive modeling, including:

- **i.** nanomanufacturing supported by informatics,
- **ii.** predicting nanomaterial biodistribution,
- **iii.** nanomaterial structure-property relationships, and
- **iv.** nanomaterial environmental fate modeling.

In each group, there was recognition of the importance of both first principles and phenomenological data for predictive modeling as well as for developing data repositories that would allow for the integration of negative data into a conceptual understanding of nanomaterial behavior. Considerations of positive and negative controls in nanomaterial characterization present a significant challenge to the predictive modeling community. The complexities associated with nanomaterials transport and potential transformations in different environmental compartments and across their life cycle requires the development of specialized models for biodistribution, and for fate and transport (e.g., pharmacokinetic/pharmacodynamics models, dosimetry models, multi-media fate and transport models, and

molecular simulations). Finally, research samples used to generate data have to be adequately defined and described in order to enable the development of accurate and useful correlations and predictive models regardless of the informatics infrastructure.

3.1. Nanomanufacturing supported by informatics

The use case for nanomanufacturing supported by informatics focused on predicting the properties of quantum dots from synthesis control conditions and the resulting impact on application performance and EHS. The benefit of dual-use of data was highlighted, in this case, through the use of informatics to support predictions of product properties as well as the prediction of their environmental and human health effects. Quantum dots were selected for this case study because some modeling and simulations for structure-property currently exist. Quantum dots have specific customizable electron energy levels and optical properties due to "quantum confinement" which makes them useful for applications such as solar cells, solid-state lighting, imaging dyes, and more.

Alternative nanoparticles illustrating this same use case could include nanoparticles of SiO_2, ZnO, TiO_2, etc. In nanomanufacturing, modeling of process-property relationships is informative for scaling up, reproducibility and sustainable design (Figure 3) and leverage both first principles and phenomenological models, again, benefitting product performance and informing environmental, health and safety risk assessments. From a decision management standpoint, the dual use of such data should be considered a priority.

The process for proactively attaining data that could be leveraged for both performance and safety inputs begins with identifying common overlaps, such as average size and size distribution. Evaluation of data needs across diverse stakeholders can help identify which input properties are mutually useful and which ones are less relevant for the intended use. Models can then be applied to optimize manufacturing based on properties, cost, safety and sustainability. Data should be collected according to protocols that fit into database informatics system so that it is useful for subsequent modeling and data mining [4]. Systems may need to implement strategies for anonymous sharing of data. For instance, toxicological information and the associated variables could be made available while other properties/ variables could be hidden as confidential business information for proprietary protection.

3.2. Predicting nanomaterial biodistribution

A biomedical use case was discussed as an example of predictive modeling in nanotechnology. The modeling focused on predicting the biodistribution of a nanomaterial based on specific nanomaterial physicochemical properties. The primary goal of a predictive biodistibution model is to improve the effectiveness of the desired biological outcome (e.g. the appropriate cell was targeted, therapeutic efficacy was achieved) while minimizing undesirable outcomes (e.g. toxic side effects). The model, if successfully designed, could translate changes in nanomaterial structure to the desired biological outcome; however, the model would have to consider conditional behaviors (i.e. correlative vs. causative).

It was argued that traditional pharmacokinetic models lack the complexity required to describe a nanomaterial sufficiently since they typically capture the entrance and exit of chemicals into compartments but lack the ability to capture material movement, or account for material degradation. A model predicting the biodistribution of a nanomaterial would require characterization data on the nanomaterial within the plasma and an understanding of the biodistribution, cellular uptake, transport and metabolism of the nanomaterial. Additionally, the model would have to capture the polyvalent nature of a nanomaterial and take into consideration nanoparticle transformations such as core degradation and protein corona formation. Fluid dynamic models may need to be considered as an extension to

current biomedical modeling in order to model nanomaterial trajectories. Fluid dynamic models may also assist in modeling the mechanical, or physical, toxicity inherent to nanomaterials that are not traditionally considered for small molecules.

It was also recognized that there is a need to effectively model the physiological conditions and bioavailability in different species to gain a weight-of-the-evidence perspective. Additional model considerations relative to the biomedical use case include the need to improve the prediction of *in vivo* outcomes based on cell culture or other *in vitro* data. This may be an area where public-private partnerships are immediately needed and additional communities of interest could be leveraged as these concerns are not unique to nanomaterials.

3.3. Modeling nanomaterial structure-property relationships

Broadly- and narrowly-focused use cases were evaluated to understand the requirements for parameterization of predictive modeling in nanotechnology [5]. When addressing the extent of nanomaterial representation required to develop predictive models, it is desirable to work backwards from the scientific endpoint. A hypothesis or a mechanistic model would first be established which would define the needed model parameters, and therefore the datasets to be collected. It is evident that nanomaterial composition, physicochemical properties, *in vitro* characterization (e.g. blood impacts, immunotoxicological assessments) and *in vivo* characterization (e.g. pharmacokinetic, toxicological assessments) need to be effectively captured. However, the essential data and the means of collecting them may strongly depend upon the type of nanomaterial in question and its intended use. The stability and durability of the nanomaterial also needs to be assessed, as does the inherent and derived nanomaterial surface chemistry.

It is necessary to realize in developing models that the potential impact on ecological receptors will be species dependent and route of exposure dependent. Given that the bulk of data currently available is not well-suited to predictive modeling for whole animal impacts, the development of proper structure-activity relationships should be considered in parallel with the use case driven research in order to develop the most meaningful conceptual framework. The ability to make predictions will be sorely limited if datasets are collected in only a single use case fashion. If data are collected in this way, the material type and use scenario are very narrow. In order for models to be broadly predictive, they need to be trained on broader training sets with clearly defined applicability domain [5]. The generation of meaningful structure level language (defining the S in SARs) also requires a reevaluation of nanomaterial descriptors to more fully capture a nanomaterials structure and complexity (Figure 4), especially in applications to biomedical safety and health problems. 3D models with overlays of electronic energy states distributed across the surface of a nanomaterial may more fully represent the features that drive their interactions with other particles, with molecules or with living systems. Descriptors should be linked to vocabularies (see section 4.1) so that data exchange can occur without the loss of data integrity and in order to accommodate batch-to-batch variability.

3.4. Nanomaterial environmental fate modeling

A use case focused on the predicting the environmental fate of nanomaterials was discussed as an exemplar of predictive modeling for nanotechnology [6].

Additional example success stories were identified from the breakout sessions on predictive modeling including: biological description of functionalized buckyballs, interlaboratory comparisons and impact assessments [5,7] and web-based platforms for nanomaterial data normalization and analysis [8]. Although it is anticipated that such exemplars will continue

to grow in number, many unique challenges for nanomaterials still exist, including protocol reproducibility, testable predictions, *in vitro→in vivo→in silico* translation, heterogeneity of materials and data, data uncertainty, model integration, positive and negative control availability, lack of systematic datasets, issues with data comparability, data access, data sharing and data interoperability.

4. Nanoinformatics integration

All workshop participants engaged in a final group session that focused on the integration of nanoinformatics tools and technologies. The goals of this session were to identify standards that could be used to catalyze nanoscience and support nanoinformatics approaches and modeling, to establish a plan for data sharing and informatics integration, and to identify gaps or barriers to nanoinformatics integration.

Discussions centered on the interface between major initiatives that could offer mutual benefits and a coordinated path forward. The Nanotechnology Knowledge Infrastructure Initiative and the Material Genome Initiative, in particular, have multiple connections with the grassroots efforts of the nanoinformatics community. These include the need for data sharing, a standard describing the minimal information required for utility in predictive models, definition of quantitative structure-property relationships, and the integration of multi-scale complex models.

4.1 Standards to catalyze nanoscience and support nanoinformatics approaches and modeling

Utilizing the genomics “big data” experience, there is a pressing need to standardize the minimal information that satisfactorily describes a nanomaterial in order to interpret and compare results across different nanomaterial formulations. For gene expression microarray data, adherence to the Minimum Information About a Microarray Experiment (MIAME) is required for works to be published. There are a number of extensions to the MIAME requirement for specific fields anddomains of interest. These agreed upon community standards arose to reduce ambiguity in datasets as well as to improve the usability of the data from data-intense microarray experiments.

Because instrumentation plays a crucial role in data files that are produced in a microarray experiment, there was a need to have access to original instrument scanning data files. This approach has the added benefit of allowing other researchers to go back to processed data outputs from an instrument for evaluation of error and variability as well as to be available for other researchers to go back and mine the datasets by asking different questions of the same data. The development of data standards is timely as International standards bodies and major research programs such as the Nanomaterial Registry are working to define a materials specification, including methods and protocols, that provides this information [9]. At a least, scientists should be encouraged to collect data with the minimal metadata standards in mind; at best, this data would be deposited in a data repository prior to publication and supporting documentation would include the entire investigation dataset as part of a complete publication process [10].

Current nanoinformatics standards and approaches should be expanded upon and developed further. These areas provide opportunities for communities to leverage results from group activities instead of reinventing a process or product. Examples of this include the ISA-TAB-Nano and the Nanoparticle Ontology activities [11, 12, 13]. A solutions oriented approach is most useful for drawing communities of interest together. Guidelines and incentives could be used to promote standards utilization [13]. Examples would include the

requirement for a satisfactory set of nanomaterial information for publications and the potential use of standard naming conventions (e.g. IUPAC) for nanomaterials.

4.2. Plan for data sharing and informatics integration

In addition to defining the minimal information required to describe a nanomaterial, there is an immediate need for an expert system to manage the information and knowledge linked to data. Standardized tools that allow for processing of raw data should be made broadly available to collect this information; e.g. remove outliers or develop cluster maps. When these tools and platforms become publically available, they can be cited in the same manner as a peer-reviewed publication. Data management tools allow for the analysis of raw data in addition to the modeling tools to analyze matrix information which enables researchers to contextualize how their nanomaterials function relative to other nanomaterials.

4.3. Gaps or barriers to nanoinformatics integration

A major challenge to nanoinformatics integration is that data are not provided in a consistent format that supports data integration across databases. Data repositories can offer a wealth of information on nanomaterials; however, most are currently limited to the research groups or individuals who collect the data. Data sharing is not yet a broadly accepted practice in the scientific community due to the ‘publish or perish’ mentality that still impacts university researchers and ‘intellectual property’ issues that limit accessibility to industry data. At present, curation of data into data repositories is done manually primarily by extraction from the literature. A significant challenge to data curation, in addition to the time and personnel investments, may be the lack of familiarity with what data to enter.

Pathways should be established that improve the accessibility and interest of scientists to enter the data directly into a database from the research bench. Developing user friendly informatics services to create a repository or spreadsheet that simplifies data contributions is warranted. Linking data with analysis tools and predictive models could also provide added benefits to the researchers and may make them less reluctant to contribute. An additional idea to incentivize researchers to contribute data is to provide a digital object identifier for contributors so that credit is given when needed [2]. Another way to overcome such barriers is to have fully-funded curation efforts; such as those taken on and supported by the federal government to establish the Protein Data Banks. Journal data requirements could also be improved with community input and become logical avenues to move such curation forward.

5. Key workshop outcomes

Informatics capabilities that need a significant amount of coordination and are beyond the scope of an individual researcher or even groups of researchers include data accessibility, data sharing, database interoperability and data fusion. Functionally integrating research to produce sufficiently large and robust datasets will require targeted cross-cutting grants. Grants that aim to integrate across diverse fields of research will allow for the assessment of material properties from multiple, use-inspired perspectives. Systematic studies on a series of nanomaterials would provide the data necessary for parameterizing predictive models if the studies are designed with informatics in mind. In addition, informatics can serve an important role in identifying data gaps and developing key correlations across multiple scales.

A strategy of moving forward could focus on a case study of a material that industry regards as important; for example, a nanomaterial that the Environmental Protection Agency needs to regulate (or identify if it should be regulated). The nanomaterial of focus needs to be evaluated in detail employing synthesis methods that can afford modification of specific

physicochemical features; thus, information about the structural or chemical properties and their relationship to biological or environmental impacts can be elucidated. The materials selected for these studies could lead to the development of a reference material library that is broadly available. The development of reference materials is a priority consideration and requires the input of significant resources. Requirements for an ideal reference materials include: industry views the material as being important, EPA needs to make a regulatory decision regarding its use, infrastructure tools are available (preferred but not essential), and feasibility of modifying the material to vary its physicochemical properties in a systematic manner.

In order to achieve the above objectives, it is important to involve diverse stakeholders as early as possible. Thus, a community driven approach must have significant and broad representation.

Acknowledgments

The GN12: Nanoinformatics Tools and Resources Workshop was sponsored by the Safer Nanomaterials and Nanomanufacturing Initiative, National Nanomanufacturing Network, UC Center for Environmental Implications of Nanotechnology and the NCI Nanotechnology Working Group in coordination with the Nanotechnology Knowledge Infrastructure Signature Initiative. Greener Nano 2012 was made possible through funding provided by the National Science Foundation through grant numer NSF CMMI 1025020 and the Air Force Research Laboratory grant number FA8650-05-1-5041 provided to the Oregon Nanoscience and Microtechnologies Institute / Safer Nanomaterials and Nanomanufacturing Initiative.

This work was also funded in part by the NCI-NIH (Contract No. HHSN261200800001E). The contents of this publication do not necessarily reflect the views or policies of these funding agencies, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. The authors would like to acknowledge that the summary presented herein was synthesized by the authors with intellectual contributions from all workshop participants

References

1. de la Iglesia, D.; Harper, S.; Hoover, MD.; Klaessig, F.; Lippell, P.; Maddux, B.; Morse, J.; Nel, A.; Rajan, K.; Reznik-Zellen, R.; Tuominen, MT. Nanoinformatics 2020 Roadmap. 2011. Available at: http://eprints.internano.org/607/.
2. Baker NA, Klemm JD, Harper SL, Gaheen S, Heiskanen M, Rocca-Serra P, Sansone SA. Standardizing data. Nature Nanotechnology. 2013; 8:73–74.
3. Kim W, Choi B-J, Hong E-K, Kim S-K, Lee D. A taxonomy of dirty data. Data Mining and Knowledge Discovery. 2003; 7:81–99.
4. Thomas DG, Gaheen S, Harper SL, Fritts M, Klaessig F, Hahn-Dantona E, Paik DS, Pan S, Stafford GA, Freund ET, Klemm JD, Baker NA. ISA-TAB-Nano: A specification for sharing nanomaterial research data in spreadsheet-based format. BMC Biotechnology. 2013; 13:2–25. [PubMed: 23311978]
5. Liu R, Rallo R, Weissleder R, Tassa C, Shaw S, Cohen Y. Nano-SAR development for bioactive nanoparticles with considerations of decision boundaries. Small. 2013
6. Liu HH, Cohen Y. Multimedia environmental distribution of nanomaterials. Nanotechnology 2012: BioSensors, Instruments, Medical, Environment and Energy. 2012; 3:304–306.
7. Shatkin, JA. Nanotechnology: Health and Environmental Risks. CRC Press; 2008.
8. Liu R, Hassan T, Rallo R, Cohen Y. HDAT: Web-based high throughput screening data analysis tools. Computational Science and Discovery. 2013 *In Press*.
9. Ostraat ML, Mills KC, Guzan KA, Murry D. The Nanomaterial Registry: Facilitating the Sharing and Analysis of Data in the Diverse Nanomaterial Community. Int J Nanomed. 2013 in press.
10. Zimmerman AS. New knowledge from old data: The role of standards in the sharing and reuse of ecological data. Science, Technology & Human Values. 2008; 33:631–652.
11. Thomas DG, Klaessig F, Harper SL, Fritts M, Hoover MD, Gaheen S, Stokes TH, Reznik-Zellen R, Freund ET, Klemm JD, Paik DS, Baker NA. Informatics and Standards for Nanomedicine

Technology, Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology. 2011; 3(5): 511–532.

12. Thomas DG, Pappu RV, Baker NA. NanoParticle Ontology for cancer nanotechnology research. Journal of Biomedical Informatics. 2011; 44:59–74. [PubMed: 20211274]
13. Leonelli, S. Packaging small facts for re-use: Databases in model organism biology. How well do facts travel?. In: Howlett, P.; Morgan, M., editors. The dissemination of reliable knowledge. Cambridge: University Press; 2010. p. 325-348.

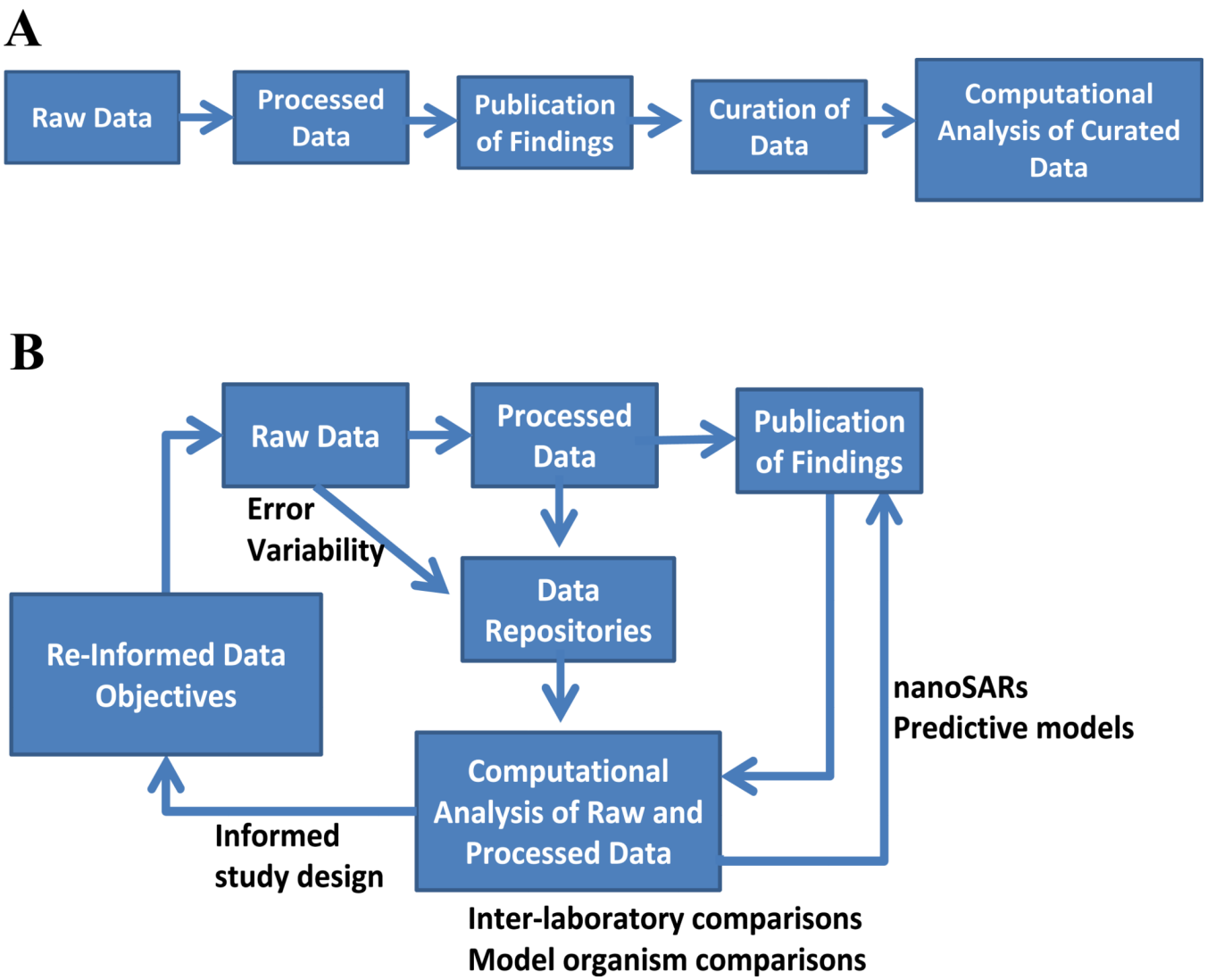

Figure 1.
A) Diagram of the current data lifecycle used as a starting point for GN12 discussions. B) Diagram of an optimized data lifecycle created through GN12 discussions to improve the value of data.

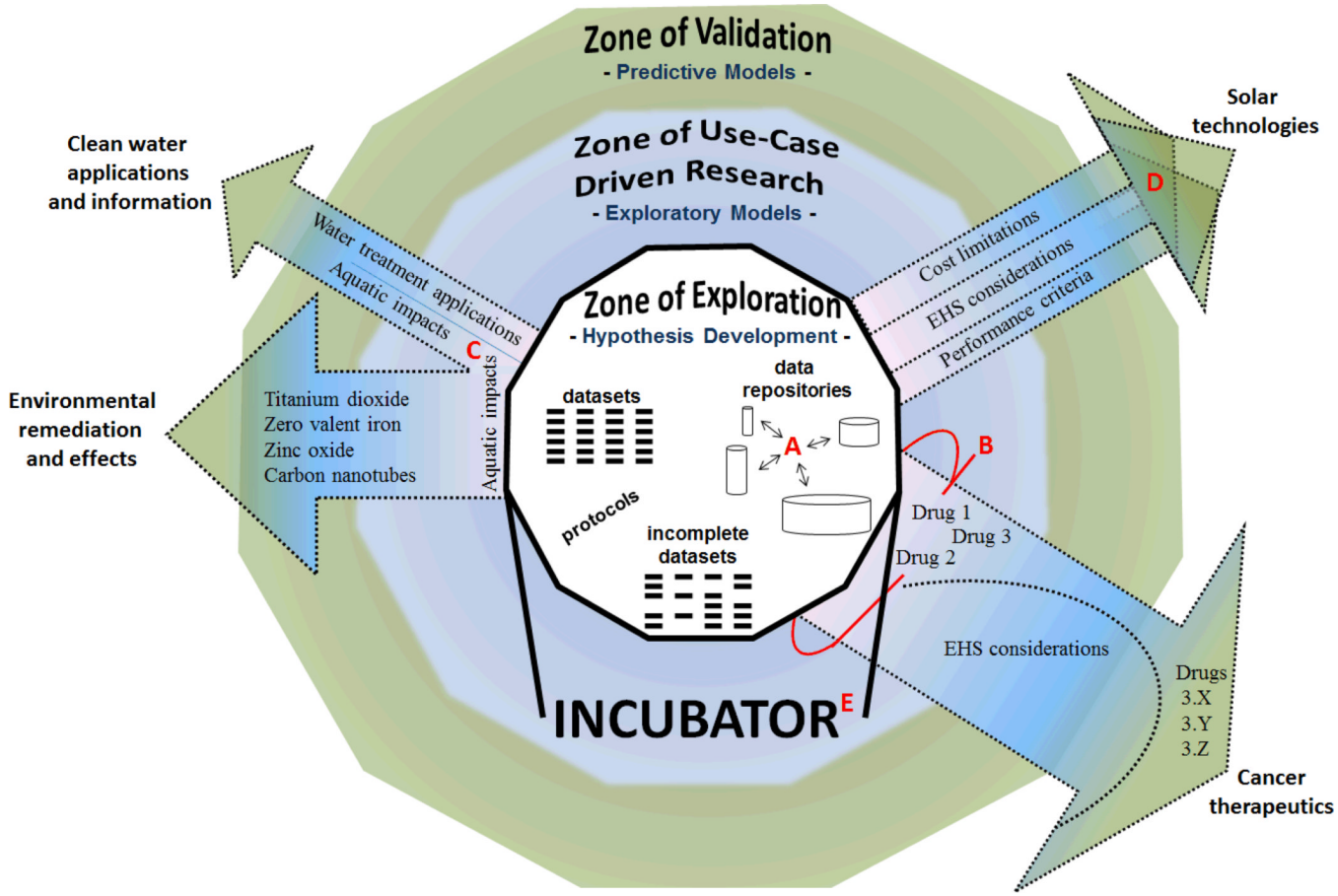

Figure 2.
Diagram of the incubator concept that supports greater knowledge generation as information proceeds from the zone of exploration to use-case driven research and into validation. See text for explicit description.

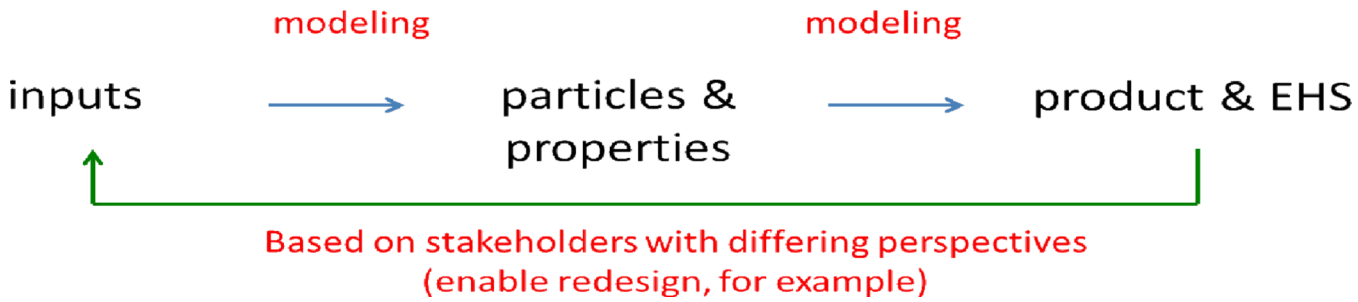

Figure 3.
Flow diagram of model inputs into nanomaterial design that leverage dual purpose data for product performance and EHS.

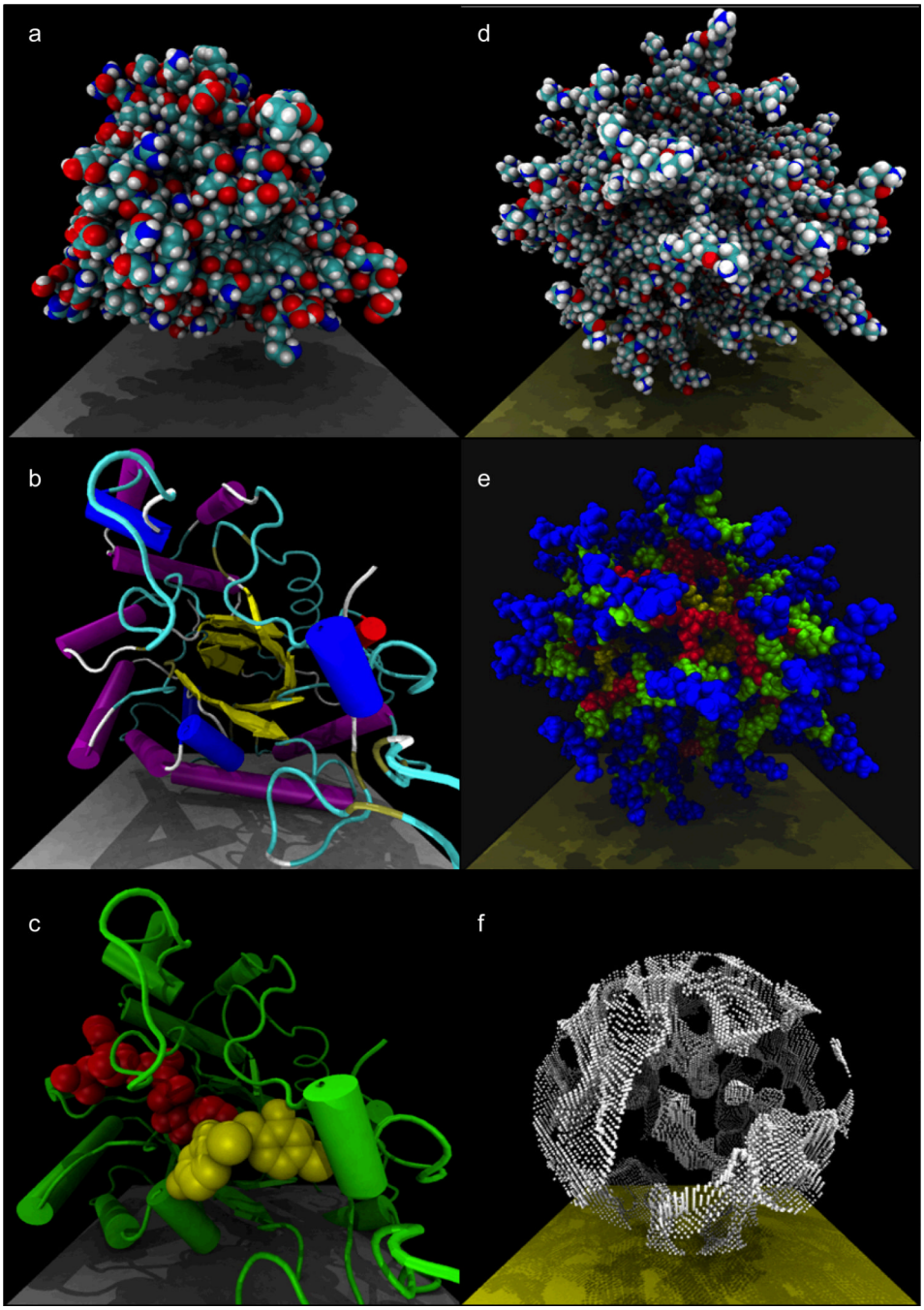

Figure 4.
Comparative structure characterization of a biomolecule (Aldore Reductase,-AR- left column, a to c) and a nanoparticle (PAMAM G4, right column, d to f). The structural elements of the protein are immediately recognizable. b: Tim barrel, in yellow; secondary structure elements. AR activity pockets can be easily identified as well. c: NAPD –in red-binding pocket; Ligand –in yellow-binding pocket. PAMAM annotation does not convey a similar level of biologically relevant information. e: PAMAM G4 color coded by generation; f: PAMAM voids.