NanoImpact 9 (2018) 85–101

Contents lists available at ScienceDirect

NanoImpact

journal homepage: www.elsevier.com/locate/nanoimpact

Research paper

# 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
[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

ARTICLE INFO

*Keywords:*
Nanomaterials
Nanotechnology
Nanoinformatics
Data integration
Databases
Web services

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.

## 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

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

https://doi.org/10.1016/j.impact.2017.11.002
Received 25 August 2017; Received in revised form 12 October 2017; Accepted 10 November 2017
Available online 11 November 2017

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 stakeholders. To realize these secondary benefits of integration, individual projects and disciplines participating in integration efforts must see improvement in their ability to meet their own objectives.

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 nanomaterials 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

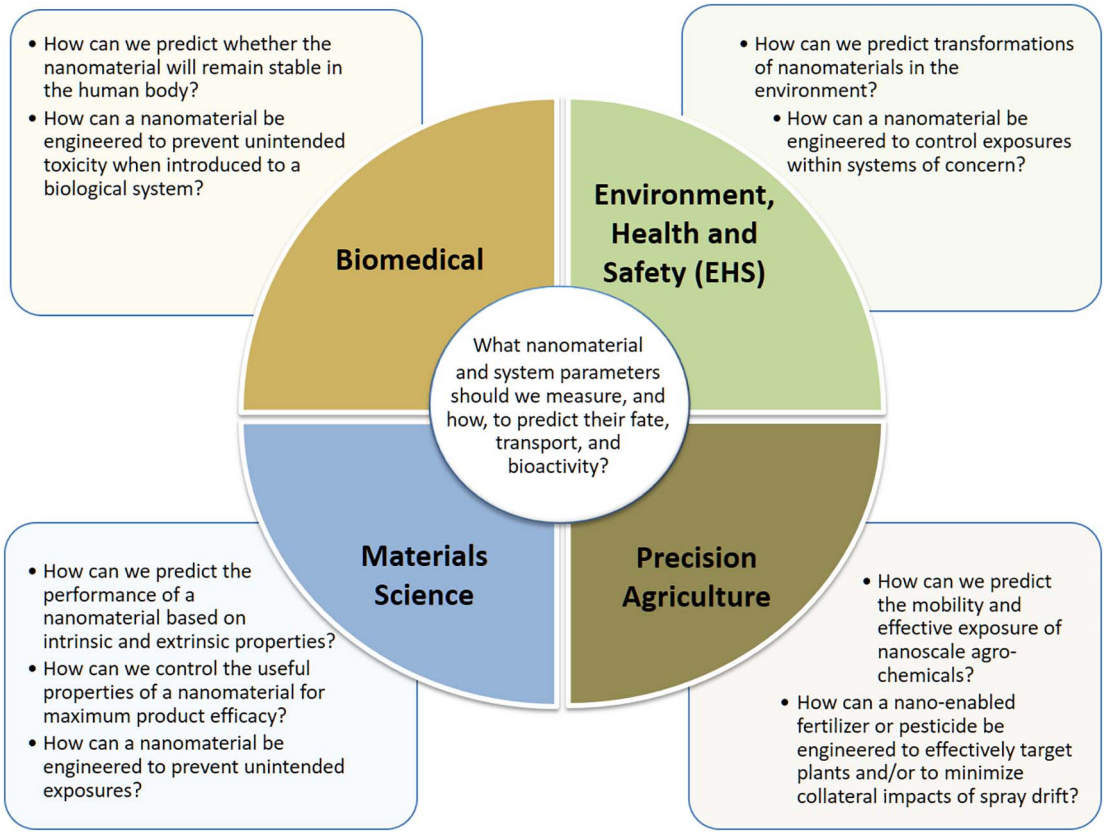


**Fig. 1.** Examples of use cases that can be addressed and might mutually benefit from data integration.