International Journal of Nanomedicine

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

NEIMiner: nanomaterial environmental impact data miner

This article was published in the following Dove Press journal:
International Journal of Nanomedicine
17 September 2013
Number of times this article has been viewed

Kaizhi Tang[1]
Xiong Liu[1]
Stacey L Harper[2]
Jeffery A Steevens[3]
Roger Xu[1]

[1]Intelligent Automation, Inc., Rockville, MD, USA; [2]Department of Environmental and Molecular Toxicology, School of Chemical, Biological, and Environmental Engineering, Oregon State University, Corvallis, OR, USA; [3]US Army Engineer Research and Development Center, Vicksburg, MS, USA

Abstract: As more engineered nanomaterials (eNM) are developed for a wide range of applications, it is crucial to minimize any unintended environmental impacts resulting from the application of eNM. To realize this vision, industry and policymakers must base risk management decisions on sound scientific information about the environmental fate of eNM, their availability to receptor organisms (eg, uptake), and any resultant biological effects (eg, toxicity). To address this critical need, we developed a model-driven, data mining system called NEIMiner, to study nanomaterial environmental impact (NEI). NEIMiner consists of four components: NEI modeling framework, data integration, data management and access, and model building. The NEI modeling framework defines the scope of NEI modeling and the strategy of integrating NEI models to form a layered, comprehensive predictability. The data integration layer brings together heterogeneous data sources related to NEI via automatic web services and web scraping technologies. The data management and access layer reuses and extends a popular content management system (CMS), Drupal, and consists of modules that model the complex data structure for NEI-related bibliography and characterization data. The model building layer provides an advanced analysis capability for NEI data. Together, these components provide significant value to the process of aggregating and analyzing large-scale distributed NEI data. A prototype of the NEIMiner system is available at http://neiminer.i-a-i.com/.

Keywords: nanomaterial environmental impact, data integration, data management, content management system, data mining, modeling, model composition

Introduction

As more engineered nanomaterials (eNM) are developed for numerous uses, it is crucial to minimize any unintended nanomaterial environmental impact (NEI) resulting from the application of eNM. To realize this vision, industry and policymakers must base risk management decisions on sound scientific information about the environmental fate of eNM; their availability to receptor organisms, including related concepts such as uptake; and any resultant biological effects, eg, toxicity.[1] This basic knowledge can be effectively conveyed by validated models that describe eNM exposure effects. These models will give decision-makers the tools to grapple with the diverse forms of possible eNM, and to explore the effects of various risk mitigation strategies. However, a tool to assess the environmental risk of eNM faces the following challenges:

- The need for a comprehensive eNM modeling framework to guide analysis. Data mining can assist in the study of NEI. Analysts can use it to build risk assessment models based on experimental data. The design of data mining problems must be based on a high-level modeling framework and the current state of scientific understanding.

Correspondence: Kaizhi Tang
Intelligent Automation, Inc., 15400 Calhoun Drive, Suite 400, Rockville, MD 20855, USA
Tel +1 301 294 5214
Fax +1 301 294 5201
Email ktang@i-a-i.com

Xiong Liu
Intelligent Automation, Inc., 15400 Calhoun Drive, Suite 400, Rockville, MD 20855, USA
Tel +1 301 294 4629
Fax +1 301 294 5201
Email xliu09@gmail.com

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/IJN.S40974

- Lack of organization of NEI data. NEI data are distributed sparsely, and developed with various research objectives under different investigators. To aggregate and normalize this diversified and ever-growing data, the information system must be cost effective, flexible and extensible, and supportive to data sharing and user collaboration.
- Lack of powerful data mining tools to build high-quality models. With the eNM attribute data and environmental impact data collected in a relevant scientific framework, data mining tools can build useful NEI models. However, it is a challenge to build accurate models due to sparse data.
- Overly simplified, disconnected, or otherwise non-comprehensive risk models. Any nanomaterial modeling framework must assemble a group of disparate models, empirical or theoretic, for comprehensive risk assessment. Building this assembly and using it efficiently presents another challenging problem.

To address these challenges, we designed and implemented a comprehensive data mining system, called Nanomaterial Environmental Impact data Miner (NEIMiner), to study the environmental impact of eNM. Figure 1 shows a schematic diagram of NEIMiner, which consists of four components: nanomaterial environmental impact (NEI) modeling framework, data integration, data management and access, and model building. The NEI modeling framework defines the scope of NEI modeling and the strategy of integrating NEI models to form a comprehensive predictability, which is similar to Framework for Risk Analysis of Multi-Media Environmental Systems (FRAMES).[2] The scope of the NEI modeling framework covers nanomaterial physical, chemical and manufacturing properties, exposure and study scenarios, environmental and ecosystem responses, biological responses, and their interactions.[3,4] The NEI modeling framework serves as the guiding force to design, develop, and implement the NEIMiner information system. The three components on the right side in Figure 1 form three important layers for the information system architecture of NEIMiner. The data integration layer brings heterogeneous data sources related to NEI via web services and web scraping technologies. The data management and access layer, which reuses and extends a popular Content Management System (CMS) Drupal, models and enables interactions over the complex data structure for NEI related bibliography and characterization data (eg, nanomaterial physical properties, chemical properties, synthesis methods, etc). The model building layer provides an advanced analysis capability for NEI data. The models built in this layer can be accessed through the data management and access layer.

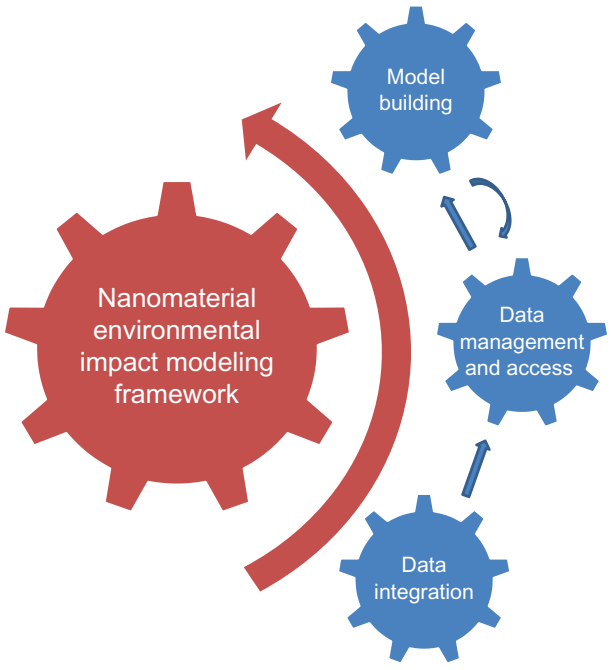

Figure 1 The schematic diagram of nanomaterial environmental impact (NEI)Miner.
Notes: The NEI modeling framework defines the scope of NEI modeling and the strategy of integrating NEI models. The three components on the right side form three layers for the information system architecture of NEIMiner. There is interactive feedback between the layers: the data integration layer provides NEI data, the data management and access layer manages the integrated NEI data, and the model building layer builds models using the integrated NEI data. The models can also be accessed through the data management and access layer.

The key features of NEIMiner include: (1) Model driven data mining system. Instead of using an ad hoc approach to mine the collected data for some useful models, we leveraged the existing NEI modeling framework and collaborated with the domain experts to define the scope and objective for data mining of NEI data. (2) Integrated and collaborative data management. The utilization of Drupal as the development platform provides not only a collaborative data collection and information management system for sparse experimental NEI data, but also an integrated data repository with relational tables for data mining. (3) Optimized data mining process. NEIMiner is supported by ABMiner,[5,6] Intelligent Automation, Inc.'s (IAI) internal data mining tool for model building. ABMiner provides an optimization engine in the meta-learning level to exploit and search data mining models with best performance and efficiency among a wide range of data mining algorithms and their corresponding parameters. (4) Flexible model management. NEIMiner utilizes the model base of ABMiner[5] to expose the NEI models and associated datasets. The generic model base framework provides a solid basis for managing and querying fundamental NEI models.

Also, model composition enables the interoperability of assessment models, whether built by data mining tools or integrated from existing models.

In this paper, we discuss the design and implementation of NEIMiner. The rest of the paper is organized as follows. We first discuss the design of NEIMiner information system. Then we present the results of system implementation of NEIMiner. Next we discuss the capability of NEIMiner for building and integrating various NEI models. Finally, we conclude the paper and provide possible directions for future research.

Materials and methods

The informatics approach to NEI modeling

We have designed an information system for NEI data analysis with three layers: (1) a data integration layer, (2) a data management and access layer, and (3) a model building layer. Figure 2 shows the system design of these three layers for NEIMiner with the available data sources and tools.

The purpose of the data integration layer is to bring heterogeneous data sources related to NEI to the storage system managed by the data management and access layer. The major effort of data integration is to extract, transform, and load (ETL) data into the common data structure with the aggregated abstraction upon NEI bibliographic and characterization data. We developed different data extraction methods, including web services and web scraping depending on how data can be accessed in different websites. We also developed a flexible management console so that these ETL processes can be automatically executed to update the NEI database.

The data management and access layer consists of a variety of Drupal modules that model the complex data structure for

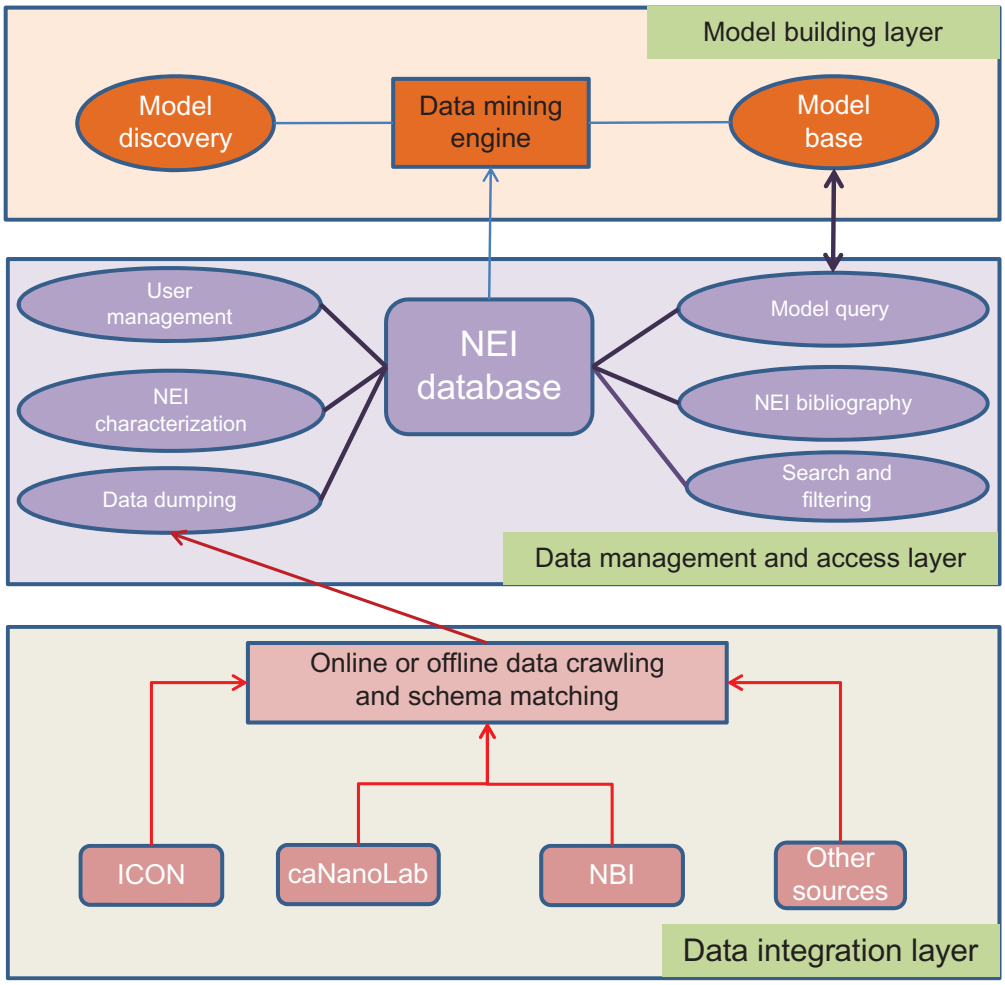

Figure 2 System design for NEIMiner.
Notes: The integration layer integrates heterogeneous data sources related to NEI to the storage system. The data management and access layer provides modules for modeling the complex NEI data and tools for accessing the data. The model building layer provides the analysis capability for NEI data.
Abbreviations: caNanoLab, cancer Nanotechnology Laboratory; ICON, International Council on Nanotechnology; NEI, nanomaterial environmental impact; NBI, Nanomaterial-Biological Interactions Knowledgebase.

NEI related bibliography and characterization data. Drupal is an extensible content management platform with more than 5000 modules available for Internet applications. Drupal has functions to provide web services that can be accessed by any third-party tools. We used existing Drupal modules and also developed our own modules including (1) management of NEI publications and characterizations, (2) web services for automatic data dumping, (3) search and retrieval of NEI contents, (4) user management in multiple levels with different permissions, and (5) NEI risk assessment model querying.

The model building layer provides the analysis capability for NEI data. With the NEI related data, especially characterization data collected, we can map the data to the NEI modeling framework. We integrate the Drupal based database with analytical software tools for model building and management. To build NEI models in the scope of NEI modeling, we apply IAI's internal data mining engine, ABMiner,[5,6] to the NEI data for model building. ABMiner provides an optimization engine which exploits meta-learning to search for the data mining models with best performance and efficiency. This layer also enables the model built in ABMiner to be published and queried in the Drupal system.

Taken together, the data integration layer, the data management and access layer, and the model building layer provide an integrated and comprehensive system for studying environmental impact of nanomaterials. The details of the three layers are discussed in the following sections.

Data integration layer

Data sources

To develop NEIMiner, we surveyed and identified a number of data sources that are related to the environmental impact of eNMs. These data sources are mixed with bibliographic texts and structured experimental data. Among them, we selected three data sources for different considerations to populate the data management and access layer of NEIMiner, namely, the International Council on Nanotechnology (ICON), cancer Nanotechnology Laboratory (caNanoLab), and Nanomaterial-Biological Interactions Knowledgebase (NBI). In the next sections, we describe these three data sources, and discuss why they were selected and how they are used in NEIMiner. We also list some other data sources that can be useful for NEIMiner, but need to be explored further.

The International Council on Nanotechnology (http://icon.rice.edu/) maintains a repository of over 4000 peer reviewed nano-Environmental, Health, and Safety (EHS) publications, which is indexed and searchable by nine categories including nanomaterial type, exposure pathway, exposure or hazard target and method of study. The ICON web service enables browsing of the categorical data and abstracts, as well as summary tools for graphically charting what kinds of papers are published over time.

We chose ICON because its comprehensive, multi-year publication list can provide insights for research on the safety of nanomaterials, from an historical point view. The limited number of nanomaterial characterizations captured in ICON provides a starting point to correlate nanomaterial properties with biological responses.

caNanoLab (https://cananolab.nci.nih.gov/caNanoLab/) is a data sharing portal designed to facilitate information sharing in the biomedical nanotechnology research community to expedite and validate the use of nanotechnology in biomedicine. caNanoLab provides support for the annotation of nanomaterials with characterizations resulting from physicochemical and in vitro assays and the sharing of these characterizations and associated nanotechnology protocols in a secure fashion.

We chose caNanoLab due to the completeness of its data structure. caNanoLab captures protocols, people, publications, and characterizations in its integrated design. caNanoLab not only provides another set of bibliographic and characterization data related to NEI, but also helped us to design the data management and access layer of NEIMiner. A further reason for this choice was the accessibility of caNanoLab via web services, which saves effort in parsing the underlying data compared with other data sources.

NBI (http://oregonstate.edu/nbi) is intended to consolidate and integrate data of nanomaterial effects in experimental animal models and evaluate biological effects from a variety of research platforms (ie, in vivo and in vitro approaches). NBI is intended to provide unbiased informatics approaches to identify the relative importance of characterization parameters on nanomaterial-biological interactions and to determine the capacity of biological assessment platforms to provide us with knowledge on the biological activity and toxic potential of nanomaterials.

We chose NBI due to its combination of physicochemical properties and biological responses, enabling us to build some initial models using these data. The underlying design of NBI includes a wide range of characterization and health impact data.

Other data sources that will potentially be integrated into NEIMiner include:

- The Nanoparticle Information Library (NIL): http://www.nanoparticlelibrary.net/. This database was created by the

National Institute for Occupational Safety and Health. The goal of the NIL is to organize and share information on nanomaterials for the occupational health and safety community.

- Nanomaterial Registry: https://www.nanomaterialregistry.org/. The Nanomaterial Registry is a repository of curated nanomaterial information, gathered by pulling data from a collection of publicly available nanomaterial resources such as NIL and caNanoLab.
- InterNano: http://www.internano.org/. InterNano is an information resource for the nanomanufacturing community.

Extraction methods for data integration

An ETL process involves extracting data from outside sources, transforming it to fit operational needs, and loading it into the database. The procedures of transforming and loading are usually dependent on the structure of the target database and corresponding operations. The procedure of data extraction is dependent on the format and accessibility of available data sources. Since many data sources related to NEI are websites that can be accessed on the Internet via the http protocol, we address the data extraction methods that can obtain data from the Internet. These methods include application programming interface (API) calling via web services and data scraping via parsing web pages.

To increase interoperability and data accessibility, today many websites allow data access via web services. Web services[7] wrap up the underlying applications and libraries as http services so that any other programs can access the underlying applications and libraries by sending messages. For any website identified for NEI, we will check whether web services exist for data access.

caNanoLab is an example that can be accessed via web services. caNanoLab is based on an NCI data sharing project, cancer Biomedical Informatics Grid, by utilizing the APIs of caCORE. The tool caCORE uses Hibernate as the persistence middle for data storage and data access. With the Unified Modeling Language defined related to nanomaterials, caCORE automatically generates a hibernate configuration file in XML, and corresponding Java APIs, web services, and view/edit/delete/search user interfaces. We will leverage the web services of caNanoLab to extract the publication and characterization data curated by the researchers who are willing to share their nanomaterial data in caNanoLab.

For those websites that do not export web service interfaces, we can use web scraping[8] to extract information. Depending on the format and complexity of the website to extract information, we use different sets of techniques. Many available web scraping software tools[9] can be used to customize web-scraping solutions. These tools usually provide some scripting functions that can be used to extract and transform web content. We will use these tools to aid data extraction from a large number of NEI related websites.

Data management and access layer

Management of NEI bibliography and characterization data

To provide flexibility in managing eNM characterization and NEI bibliography data, we use and extend a web Content Management System (CMS), Drupal, to store and manage NEI related data. Specifically, we extend Drupal to manage eNM characterization and NEI bibliography. Drupal is one of the most widely adopted open source CMS solutions, with a vibrant developer and extension community. Drupal provides role-based permission controls and full text search of its content types. Using an existing open source system allows us to build upon proven technologies, while still allowing us the level of customization required to explore new ideas and features. Drupal also provides a level of robustness, as vibrant open sources systems are essentially peer reviewed and peer tested.[10]

Figure 3 shows an extended Entity-Relation diagram that adds NEI entities to the Drupal schema. The basic entities in the Drupal core are content, user, role, and permission type. A user can create a content node via the relation "Author". A content node can be connected with itself for the purpose of versioning. Users can be assigned with a set of roles and each role can have a set of permission types. A content node in Drupal can be easily extended with more data fields for different purposes. For example, the entity "NEI Bibliography", which extends the entity "Content", can represent a bibliographic record with information about contributor, keyword, type, dataset, etc. Similarly, the entity "NEI Characterization", which extends the entity "Content", can represent an NEI characterization record. We utilize the underlying schema in NBI[11] to define the data fields (attributes) in "NEI Characterization". The NBI schema is a multi-dimensional schema used to store data such as nanomaterial properties and interactions. In Drupal, both "NEI Bibliography" and "NEI Characterization" entities are automatically managed by users, roles, and permission types.

Data dumping capability via Drupal web services

Drupal web services enable the programming operations on the contents, comments, users, and so on in the CMS. Drupal web services can be used for dumping a large amount of data related to NEI bibliography and characterizations into the CMS.

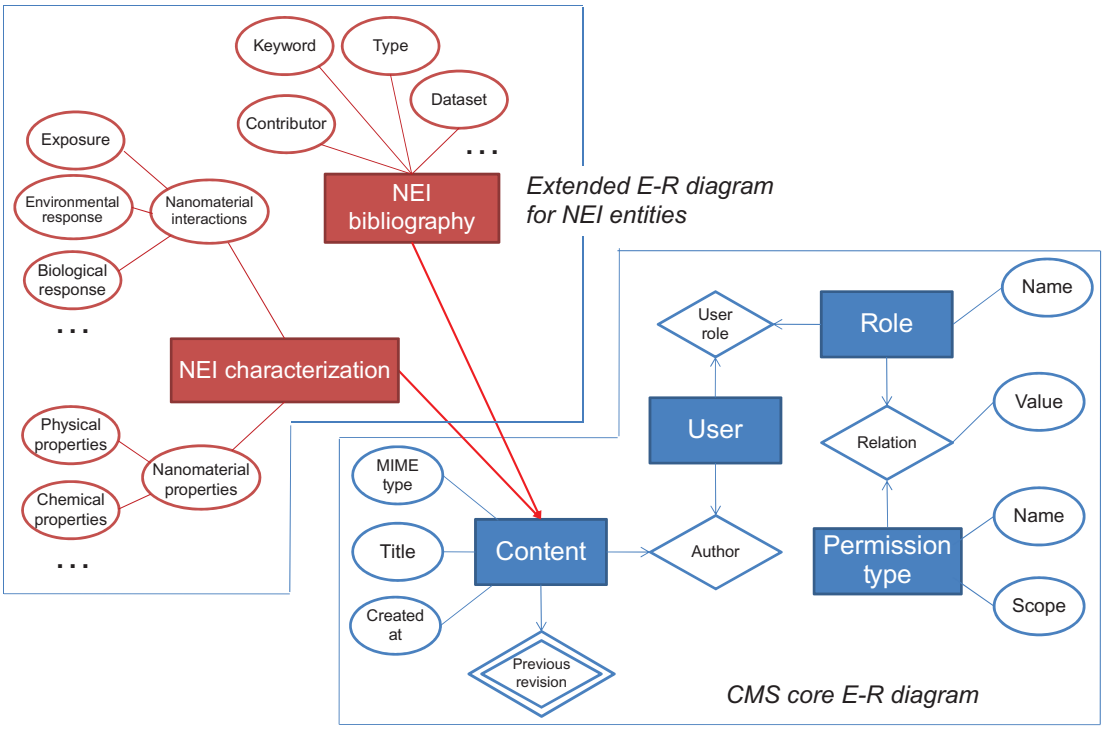

Figure 3 Entity–Relationship (E-R) diagram shows how we extend the existing Drupal CMS E-R diagram to include NEI data.
Notes: The basic entities in the Drupal core are content, user, role, and permission type. The entity "NEI Bibliography" extends the Drupal content to represent a bibliographic record. And the entity "NEI Characterization" extends the Drupal content to represent an NEI characterization record.
Abbreviations: CMS, Content Management System; MIME, Multipurpose Internet Mail Extensions; NEI, nanomaterial environmental impact.

For example, we developed an approach that can automatically read data files or websites, and further save the data as Drupal contents. The data characterization records in Drupal contents can reuse many existing Drupal capability for the purpose of viewing, sharing and visualizing. NEI bibliography data can be dumped into the CMS in the similar manner.

The concept of data dumping capability of NEIMiner is summarized in Figure 4. Given the situation that nano characterizations can be stored in various data sources, we first designed a data schema that aggregates the data fields in these sources. Part of the data fields are shown in the left upper corner of Figure 4. We chose Java as the development language due to its many existing APIs that can be used for data accessing and data parsing. For implementation, we first map a Java class to the corresponding Drupal content type. In our situation, we will map the content type of nano characterization. Second, we parse the data sources using some existing APIs. For example, we can use some CVS reader such as OpenCVS to parse an excel file. Finally, we call Drupal's web service functions to save the data to Drupal as contents.

Search and filtering of NEI contents

NEIMiner needs to have powerful searching and filtering tools to help users to find their interested NEI entities to view, flag, comment, and rate. Drupal has various modules that can help searching and filtering if the content is built on nodes. In NEIMiner, we reused and extended the following modules:

- Faceted search, also called faceted navigation or faceted browsing, is a technique for accessing a collection of information represented using a multidimensional classification scheme, allowing users to explore by filtering available information. A faceted classification system allows the assignment of multiple classifications to an object, enabling the classifications to be ordered in multiple ways, rather than in a single, pre-determined, taxonomic order.
- The module view can be used to develop customized filtering capability. The Views module provides a flexible method for Drupal site designers to control how lists and tables of content are presented. We can give parameters to Views and the filtering interface will be automatically generated. With the Views module, we can easily develop multi-dimension filtering user interfaces for both NEI bibliographic and characterization data.

User management of NEI related participants

To enable user interactions in NEIMiner, we reused and extended Drupal modules to manage users. Basically, users

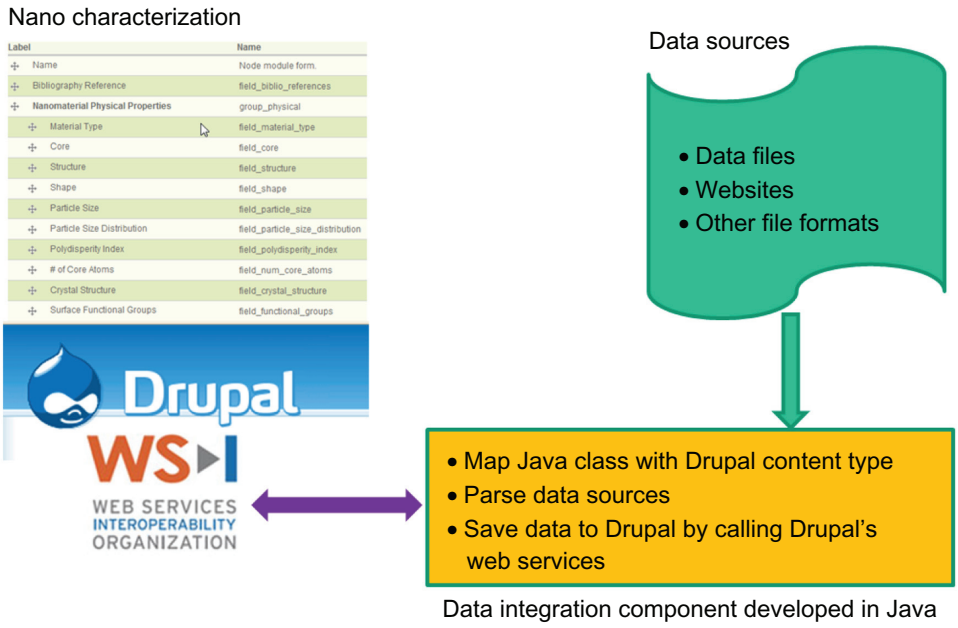

Figure 4 Conceptual diagram of data dumping capability of NEIMiner.
Notes: Various data sources related to NEI characterization are parsed by the data integration component. The results are saved to a data schema that aggregates the data fields in the data sources.
Abbreviation: NEI, nanomaterial environmental impact.

are related to roles and roles are related to permissions. In Drupal, special care was taken to ensure the permission management was consistent across all sections of the system (nodes, menus, menu items). Also each security module uses a cascading security scheme to allow global security and the option to override or define exceptions to the security model for individual items. NEIMiner is developed for all categories of NEI professionals. Different professionals will play different roles and have different permissions. We can gradually incorporate more roles into the system. Currently, we designed the following roles:

- Policy makers: this role represents government decision makers whose main task is to view the evaluation results to make appropriate decisions.
- Nano experts: represents researchers and scientists either from academics or industry in the area of nanotechnology whose main task is to rate the samples, publications, and experiments.
- Risk analysts: represents environmental and health analysts whose main task is to create risk analysis models.

NEI model query

In the Drupal system of NEIMiner, we enable the management and query service of many NEI risk models trained from NEI related datasets (see section "model building layer" for the discussion of model building). We reuse ABMiner's "model base" tool to store, manage, and provide remote access to data mining models for NEI. The model base makes data mining models easy to use in domain oriented applications. It enables a well-trained data mining model, together with its underlying data set and performance metrics, to be wrapped up and accessible across various platforms through web services. With the power of model base, the well-trained NEI models can be accessed anywhere in the website.

Model building layer

Model building

The NEI prediction model building module automates the processing of deriving, sharing, and visualizing high-quality prediction models from experimental data. In NEIMiner, we used the core feature of ABMiner, agent based meta-optimization model searching, to build useful NEI models. Specifically, we used a two-stage approach to use ABMiner. In the first stage, we identified a set of algorithms that can be applied to mine NEI datasets. Representative algorithms include nearest neighbor algorithms, tree algorithms, and support vector machines. Second, we reused the meta-optimization strategy[6] to select algorithms and tune parameters.

We developed a meta-optimizer to handle the problem of the selection of data mining models with multiple algorithms and parameters. The meta-optimizer aims to find an optimal data mining model which has the best

performance (eg, highest prediction accuracy) for a given dataset. The data mining meta-optimization problem consists of parameter optimization and algorithm optimization. Parameter optimization refers to finding the parameter settings that will result in optimal performance for a given data mining algorithm. Most data mining models require the setting of input parameters. For example, the parameters in a support vector machine include penalty parameters and kernel function parameters. These parameters usually have significant influences on the performance of the algorithm. Algorithm optimization refers to selecting the algorithm with the best performance from a list of applicable algorithms, each of which is considered for parameter optimization. Algorithm optimization automates the process of selecting one or several optimal algorithms/models. With the parameter optimization embedded, the process of algorithm optimization iterates within a list of feasible algorithms and finds one or multiple algorithms with the highest modeling performance.

Model base

ABMiner provides a model base component that enables a well-trained data mining model together with its underlying data set and performance metrics to be wrapped up and accessible across various platforms. Previously, we developed the model base based on the technology of web services so that the data mining models can be queried and presented everywhere, including desktop, Internet, and mobile devices.[5] The functional components of the model base are divided into two separate groups: the business tier module and the web tier module. The business tier module hosts the core functions, including the storage and management of the knowledge models and the associated datasets, and the computations incurred by dataset visualizations and model queries per users' requests. The computation results are consumed by the web tier module via the web services hosted in the business tier module. The web tier module generates the web pages based on the computation results from the business tier module and returns the pages to the users' web browsers. In addition, users can use ABMiner to deploy the trained knowledge models and the associated datasets via the web services in the business tier module.

Capturing the prediction relationship from different perspectives, we may build many NEI models in different categories. NEIMiner website has a powerful management capability for web contents using Drupal technology. With the NEI models represented as web contents, NEIMiner will have the capability to manage these NEI models through taxonomy, search, and collaborating. It is very helpful to automatically import NEI models produced by ABMiner into NEIMiner Drupal site so that these models can be managed by Drupal.

Results

Implementation of the data integration layer

We have successfully developed software components that can call the web service of caNanoLab grid service. All the caNanoLab data structures are exposed to web services. Through a CaNanoLabServiceClient object, we are able to connect to a caNanoLab web service and operate the data in the storage. Our software can read all the data in the caNanoLab including nano samples with their corresponding characteristics, all the publications related to these samples, and protocols to collect those samples.

We have developed software tools to scrape bibliography data from ICON. ICON is a large corpus of publications related to the safety of nanomaterials with over 4000 records. We scraped all their records and made them available in NEIMiner. The procedures to scrape ICON data can be described as:

- Step 1: analyzing the data fields to get a complete class design for ICON publication.
- Step 2: developing the control mechanism to iterate page by page for publication records.
- Step 3: parsing each page of publication record to populate the objects of the designed class for ICON publication.
- Step 4: dumping the objects scraped to the NEIMiner Drupal site.

In addition, we implemented software tools to integrate the NEI characterization data extracted from the NBI knowledgebase. The data contain the nanomaterial properties, experimental design parameters, dosage, and impact measurements. For example, there is an experimental dataset on the toxic effects experienced by embryonic zebrafish due to exposure to nanomaterials. Several nanomaterials were studied, such as metal nanoparticles, dendrimer, metal oxide, and polymeric materials.

Implementation of the data management and access layer

NEI bibliography module

We have implemented a module to model the bibliography related to NEI. The bibliography module models various types of publications, including books, book chapters, journal articles, conference papers, and so forth. Figure 5 shows the interface for creating publication entries. For each publication type, we

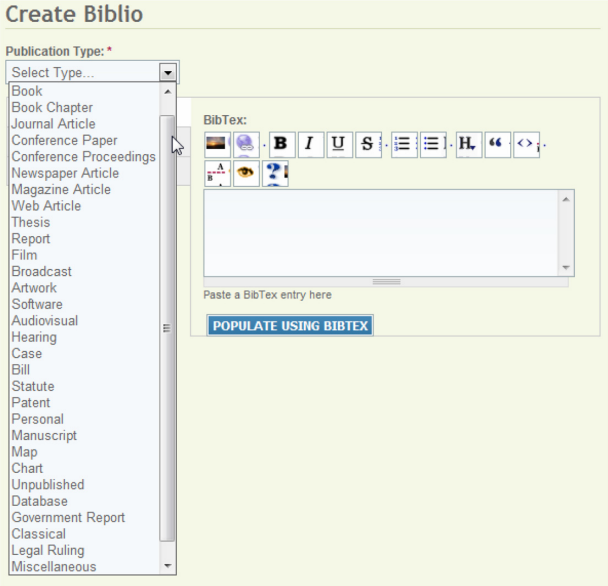

Figure 5 The user interface of creating various types of publications.
Notes: For each publication type, we model the corresponding data fields. For example, the journal article includes fields such as title, year of publication, journal name, and publisher. With this interface, users can create bibliography data related to nanomaterial environmental impact analysis.

modeled different data fields. For example, the fields to fill in for a journal article include the title, year of publication, journal name, volume, start page, issue, pagination, date published, type of article, ISBN number, ISSN, accession number, keywords, URL, DOI, short title, edition, number of volumes, publisher, and place published. With this module, we can easily input the publications related to NEI analysis.

NEI characterization module

To capture the full range nano characterization for environmental impact, we developed a user interface for compiling NEI data (see Figure 6). With the interface, users can easily fill the data fields in the following categories:

- Nanomaterial physical properties
- Nanomaterial chemical properties
- Exposure and study scenario
- Environmental/ecosystem response
- Biological response

The selection of the eNM parameters and fields is based on the data schema defined in NBI, which has been used to store NEI characterization data. The NBI data schema has a hierarchical view of nanomaterial properties and nanomaterial-biological interactions. The material type can be carbon, metal, dendrimer, semiconductor, and other types (eg, biocompatible or biodegradable materials).[12] The data fields that describe the physical properties include particle size (eg, size distribution, polydisperity), shape (eg, sphere, cubic, asymmetric triangle, rod, whiskers), structure (eg, core, shell, hollow), core composition (eg, number of core atoms, crystal structure), surface composition (eg, functional groups, density of ligands, surface area), manufacturing process (eg, chemical synthesis, reduction, arc discharge, ablation), and purity. The data fields that describe the chemical properties include surface chemistry (eg, reactivity, redox potential, surface charge), solubility (eg, hydrophobicity, liposolubility), and stability. The data fields that describe the exposure/study scenario include media/matrix (eg, solution/solvent, zeta potential, agglomeration state) and exposure (eg, duration, continuity, length, age/life stage, route, dose). The data fields that describe the environmental/ecosystem response include fate and transport (eg, adsorption, transformation), bioavailability/uptake, and biomagnification. The data fields that describe the biological response include genomic response (eg, up-regulated or down-regulated genes), cell death (eg, apoptosis, necrosis), and whole organism response.

The tool we have developed has the flexibility to add and change the data fields, giving us the capability to handle the complex characteristics for nano risk assessment.

Faceted search of the NEI bibliography

To enable the easy access of the NEI bibliography, we implemented a faceted search on the nanomaterial publications. A user can type a keyword in the search box or click any category in the page to view the NEI contents in that category. Figure 7 shows the user interface of the keyword "particle" with multiple navigation areas and statistics. The user can easily refine the searching results or start a new search.

Multidimensional view of the NEI characterization data

Multidimensional data visualization is always important for data analysis. In the present study, we integrated Java Applets for data visualization. For the purpose of demonstration, we can display and enable drag-and-drop for a 3D chart and histogram chart for the dataset, collected from the NBI knowledgebase.[11] Figure 8 provides an example scatter 3D chart to show the relationship between concentration, size, and biological impact when nanomaterials are applied to animals. Figure 9 shows an example histogram chart for biological impact.

Implementation of the model building layer

Model building

To explore the process of building risk assessment models for nanomaterials, we utilized the dataset from NBI to build

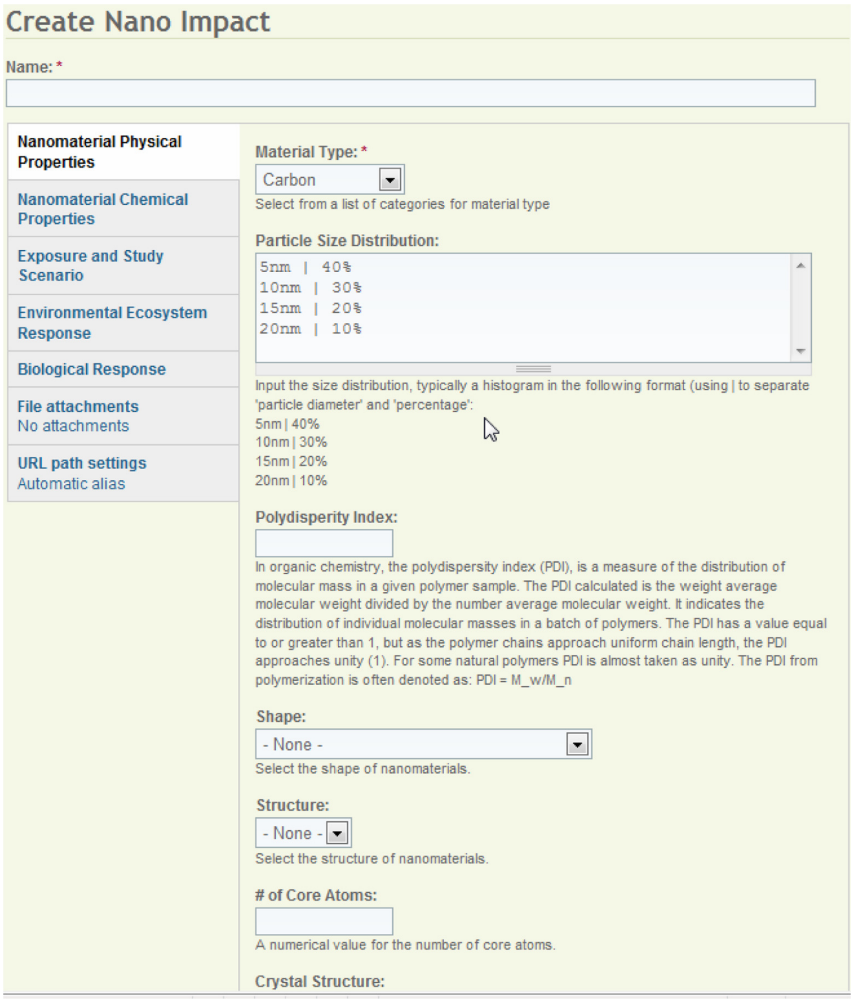

Figure 6 The user interface for filling nanomaterial environmental impact data.
Note: The interface allows users to fill the data fields in different categories such as nanomaterial physical properties, chemical properties, exposure and study scenario, ecosystem response and biological response.

prediction models. The NBI dataset captured the data fields on nanomaterial characterization (eg, purity, size, shape, charge, composition, functionalization, agglomeration state), synthesis methods, and nanomaterial-biological interactions (beneficial, benign or deleterious) defined at multiple levels of biological organizations (molecular, cellular, organismal). IAI's data mining tool, ABMiner, provides hundreds of learning algorithms including classification, clustering, numerical prediction, and many more. We can compare the performance of different algorithms for predicting the adverse effects of nanomaterials.

Using an experimental dataset on the toxicity of nanomaterials to embryonic zebrafish, we conducted case studies on modeling the specific toxic endpoints such as mortality, delayed development, and morphological malformations. The dataset contains test results on different nanomaterials including metal nanoparticles, dendrimer, metal oxide, and polymeric materials. The algorithms are numerical prediction algorithms, which can mathematically quantify the relationship between the input variables (eg, nanomaterial properties, dosage, exposure route, and timing) and the output label (eg, a numeric measure of toxicity). The results show that we can achieve high prediction accuracy for certain biological effects, such as 24 hours post fertilization (hpf) mortality, 120 hpf mortality, and 120 hpf heart malformations. Figure 10 shows the result of predicting the 24 hpf mortality using different algorithms. The prediction accuracy is measured by the correlation between the actual and predicted toxicity values. We can see that the algorithm of IBK offers the best performance. IBK is a K-nearest neighbor predictor that assigns an input to the output label most common among its K nearest neighbors.

Currently, a wide range of nanomaterials are being tested in a broad array of in vivo animal systems and in vitro assays

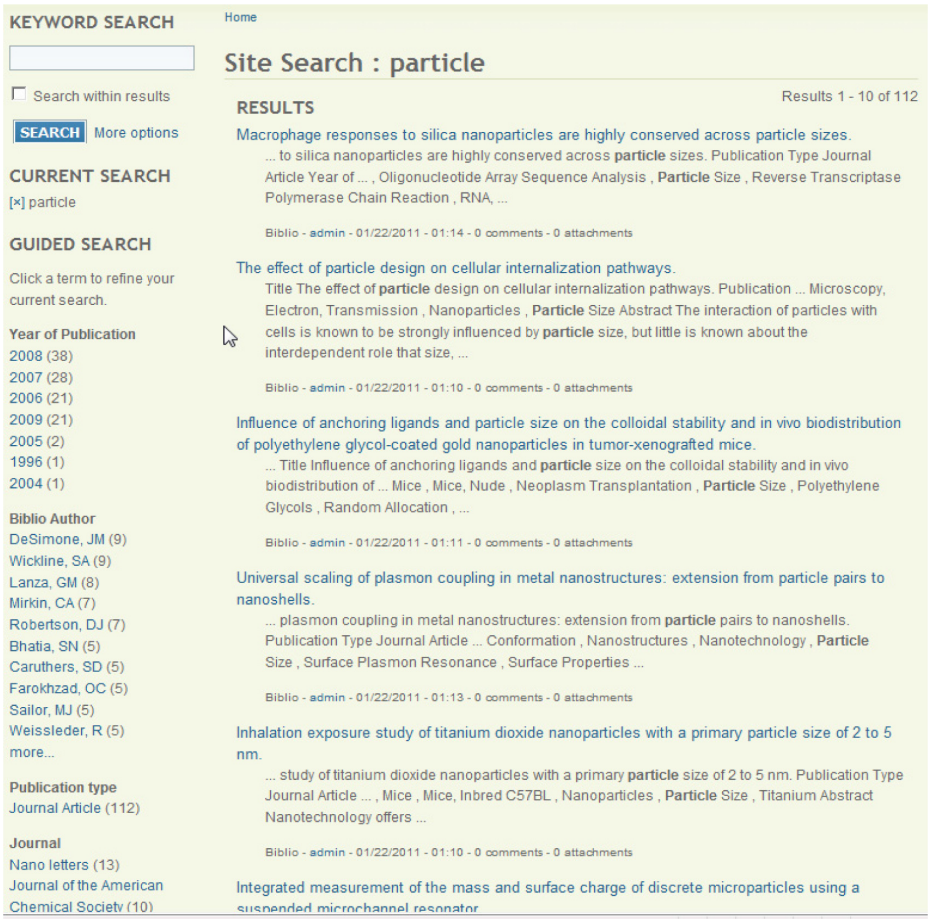

Figure 7 The user interface of the keyword "particle" with multiple navigation areas and statistics.
Note: Users can type any keyword(s) to retrieve articles related to nanomaterial environmental impact.

by different organizations. Knowledge of nanomaterial-biological interactions requires the inclusion and consideration of the entire body of data produced from global efforts in this area. Compilation of this data will allow for the determination of more robust nanomaterial structure-activity relationships. In the future, we will expand our modeling to other data sources to improve the reliability of models.

Model query

With the comparison results in Figure 10, we selected a tree based regression model M5P to predict the environmental impact from nano characterizations. Figure 11 shows the structure of M5P for the NEI prediction model and we can see the overall branches of the model. Figure 12 shows the user interface of the model query. By adjusting the values for each variable, we can see the change of the impact value.

Discussion

NEIMiner provides the comprehensive capability to query NEI bibliographic and characterization data. Furthermore, NEIMiner provides the capability to build and query various NEI prediction models. Using the data from the Nanomaterial-Biological Interactions knowledgebase (NBI), the present study demonstrates NEIMiner's capability to build models for predicting the exposure risk of nanomaterials. NEIMiner also supports the development of other types of prediction models, such as those defined in the FRAMES framework.[1] FRAMES is a 3MRA (multi-media, multi-pathway, and multi-receptor risk analysis) modeling system that connects various elements in four different layers: source, transport, food chain, and exposure/risk. In the source layer, we identify chemical models, air quality models, and human activity models. In the transport layer, we identify consensus transport model and consensus fate model. In the food chain layer, we can develop the models of bioavailability, biomagnifications, and exposure assessment. In the exposure and risk layer, we are interested in the fundamental models such as exposure-related dose estimating models, stochastic human exposure and dose simulation models, and also hazardous air pollutant exposure models.

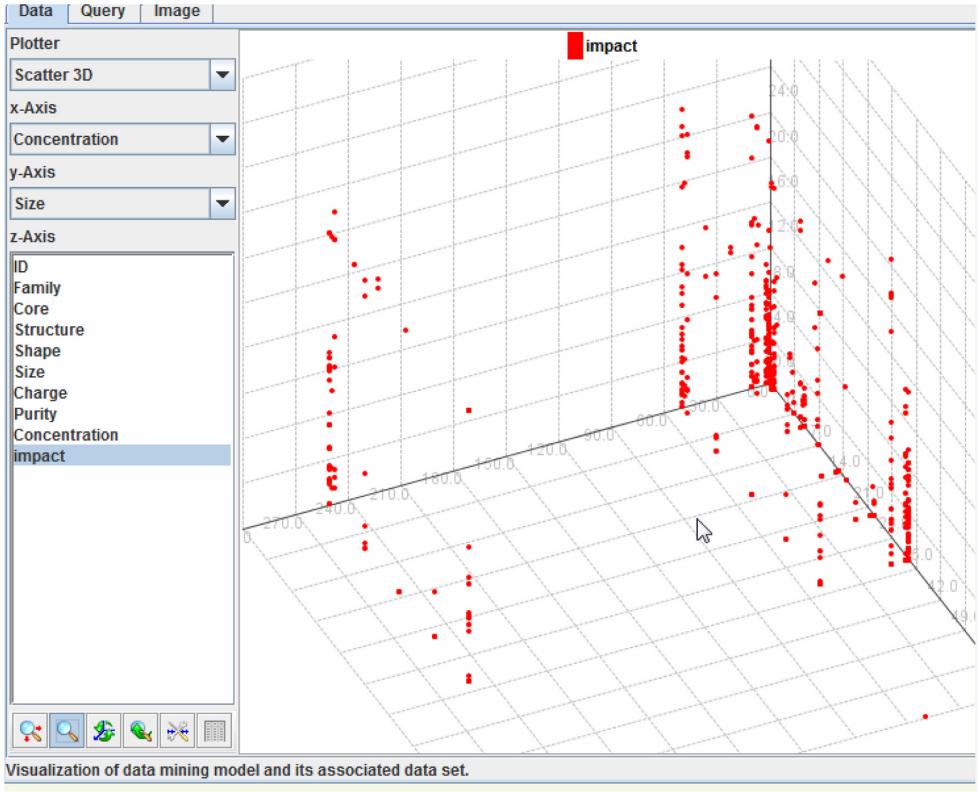

Figure 8 An example scatter 3D chart that shows the relationship between concentration, size, and biological impact when nanomaterials are applied to zebrafish embryos.
Note: Users can define their own x-axis, y-axis, and z-axis to view the 3D plot.

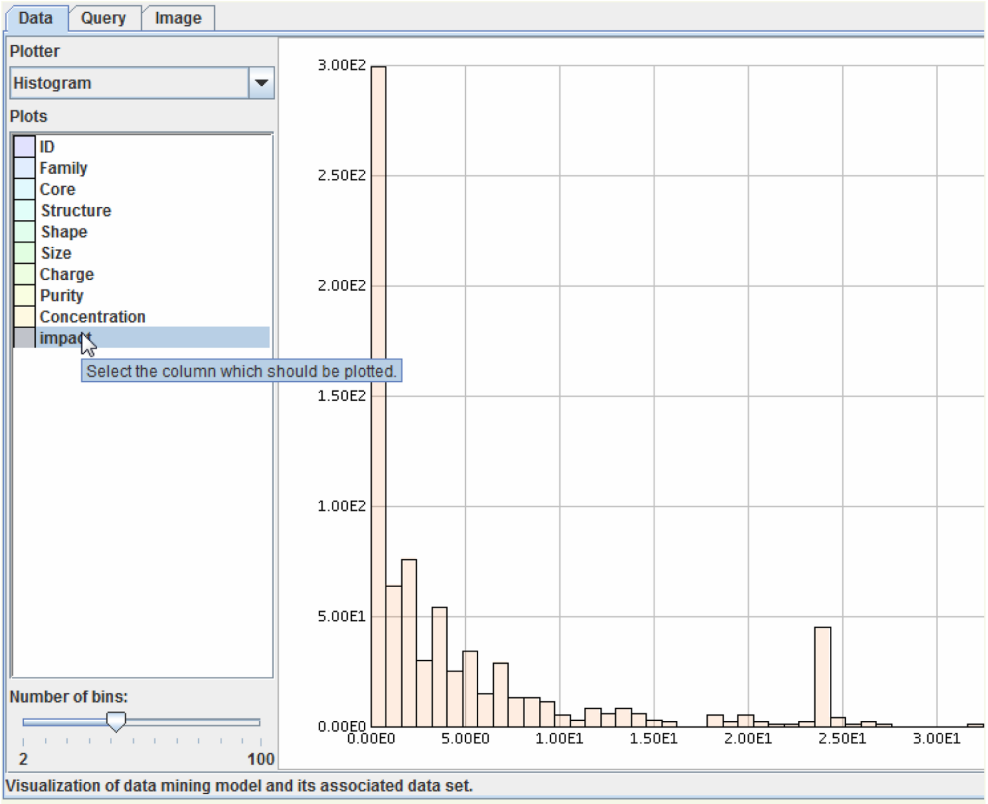

Figure 9 An example histogram chart for biological impact when nanomaterials are applied to zebrafish embryos.
Note: Users can select their own data field to plot the histogram.

Learner Name	Selected	Execution time (seconds)	Correlation
weka.classifiers.lazy.IBk	☑	0.899	0.841
weka.classifiers.trees.M5P	☑	9.058	0.816
weka.classifiers.rules.M5Rules	☑	18.237	0.807
weka.classifiers.meta.Bagging	☑	3.049	0.79
weka.classifiers.functions.GaussianProcesses	☑	32.937	0.755
weka.classifiers.lazy.KStar	☑	6.069	0.707
weka.classifiers.trees.REPTree	☑	0.358	0.701
weka.classifiers.meta.RandomSubSpace	☑	1.831	0.696
weka.classifiers.meta.AdditiveRegression	☑	0.565	0.611
weka.classifiers.functions.LinearRegression	☑	3.505	0.598
weka.classifiers.lazy.LWL	☑	3.75	0.507
weka.classifiers.functions.SVMreg	☑	21.235	0.506
weka.classifiers.functions.SMOreg	☑	25.448	0.497
weka.classifiers.trees.DecisionStump	☑	0.102	0.432
weka.classifiers.rules.ConjunctiveRule	☑	0.155	0.408
weka.classifiers.meta.RegressionByDiscretization	☑	0.535	0.244
weka.classifiers.functions.RBFNetwork	☑	0.452	0.07

Figure 10 Comparison of numerical prediction algorithms for nanomaterial environmental impact modeling.
Notes: The list shows the performance of different algorithms for predicting the overall impact of nanomaterials on the embryonic zebrafish. The overall impact is an aggregation of individual toxic endpoints such mortality, delayed development, and morphological malformations. The prediction accuracy is measured by the correlation between the actual and predicted impact values. The abbreviations in the list are defined by the Weka software package. Refer to: http://weka.sourceforge.net/doc.dev/overview-summary.html. Weka is open source software issued under the GNU General Public License. Refer to: http://www.cs.waikato.ac.nz/ml/weka/.
Abbreviations: IBK, instance-based classifier; M5P, M5 model trees; M5Rules, generating rules from M5 model trees; KStar, k instance-based classifier; REPTree, reduced error pruning tree; LWL, locally-weighted learning; SVMreg, support vector machine for regression; SMOreg, sequential minimal optimization for regression; RBFNetwork, radial basis function network.

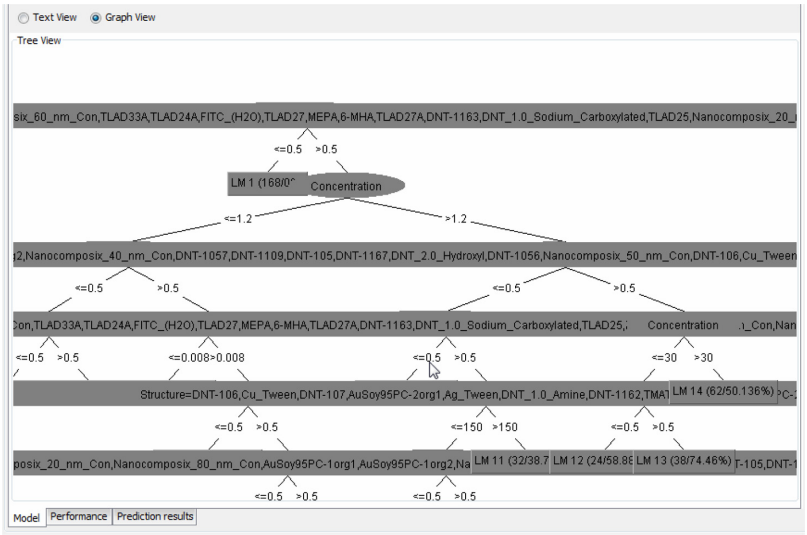

Figure 11 The structure of M5P model for nanomaterial environmental impact prediction.
Notes: M5P is a tree algorithm that can predict nanomaterial impact based on the input attributes of nanomaterial properties and exposure scenarios. Users can view the overall branches of the tree model.

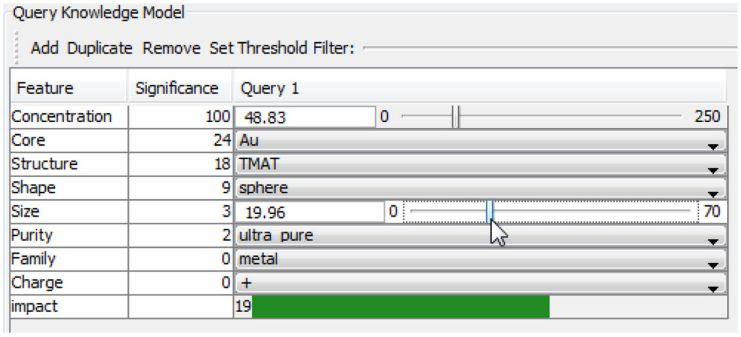

Figure 12 The query interface for a trained nanomaterial environmental impact prediction model.
Note: The interface allows users to dynamically predict the impact value for different combinations of attribute values.

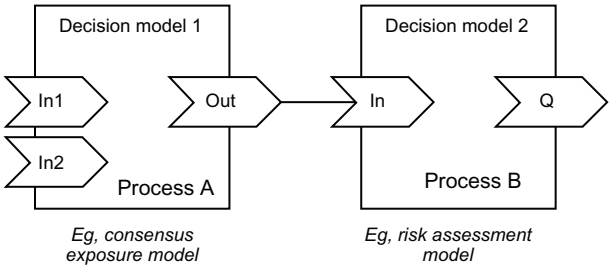

Figure 13 A conceptual dataflow diagram.
Notes: The output variable of process A can be bound to the input variable of process B. For example, the dosage (output) predicted by a consensus exposure model can be used as an input to the risk assessment model.

With the fundamental models derived from model building, we can apply model composition techniques to form more comprehensive models. We use a directed graph – the process graph – based on object-oriented technology to model these models and their connections. The process graph metaphor allows us to represent each model as a node, quickly assembling a comprehensive NEI risk assessment scenario visually. In our dataflow design, each node will represent a prediction model at different levels, which can be any pattern or model specified above. The user will be able to select the nodes from the model base, and link the nodes together for decision making purpose. For example, in Figure 13, process A may represent a decision model (eg, consensus exposure model), and its output variable can be bound to an input variable of another decision model in process B (eg, risk assessment model). With a library of models, the user should be able to conduct basic decision making through model composition. As an example of model composition, fundamental models such as human exposure and dose simulation models can be aggregated to form a consensus exposure dose model and further aggregated to form risk assessment models.

Conclusion

We have developed an integrated information system for NEI utilizing various technologies, including data integration, CMS, data mining, and data visualization, based on a comprehensive NEI modeling framework. Utilizing our internal data mining tool and a powerful open source Internet development platform, we have developed a prototype software tool for NEI data management and modeling. The following are our major contributions.

First, we have designed the NEI modeling framework to organize the eNM characterization data and to define the potential data mining problems. The modeling scope provides a holistic view of data and models related to NEI. Such efforts will have a significant value for the aggregation and analysis of NEI data, especially as NEI data is sparsely distributed and ever-growing.

Second, we have successfully extracted data in an automatic way from multiple NEI data sources and aggregated them in a central information system. These data sources include caNanoLab, ICON, and NBI; each holds unique data and relevance for eNM safety. The capability of integrating multiple data sources will foster interesting and meaningful innovation through interdisciplinary research.

Third, we have designed and implemented an online tool for NEIMiner based on the cost-effective CMS Drupal. Even at this early stage, our web tool demonstrates many useful features for NEI data integration, modeling, and data visualization. We will benefit from modules in many areas of Internet applications and from the solid and extensive modular system, which will allow us to develop more useful features related to content search, user management, user interaction and data visualization.

Fourth, we have successfully reused ABMiner to import and analyze a highly relevant dataset from NBI to build an interesting NEI prediction model. All the features in ABMiner, including data extraction, feature selection, meta-optimization, model selection, model deployment, and model query, have proved to be very useful for mining NEI data. The ABMiner model base provides a useful bridge between the offline and online tools for NEI modeling.

In the future, we will continue to improve the NEIMiner information system in these ways.

First, we will acquire more data to refine and expand the impact prediction models. Currently, we utilized data on the toxic potential of nanomaterials as indicated by the embryonic zebrafish assay. Although this model represents an in vivo system and is perceived as a very sensitive model, we will seek to enhance our data considerations to include additional animal models and impact studies related to humans.

Second, we will strengthen the development of structure-activity relationships for nanomaterials (nano-SAR)[13] on both data collection and model development. In a manner similar to the experimental approaches for hazard assessment, nano-SARs attempt to correlate toxicity end points to the underlying structures of nanomaterials.[14,15] The nano-SARs require sufficiently large experimental databases of reasonable diversity (eg, with respect to the heterogeneity of nanoparticles and biological receptors) and suitable nanoparticle descriptors.[16] Due to the complexity of chemical and morphological structures of nanoparticles, it would be most beneficial to develop nano-SARs for individual nanoparticle classes along with appropriate validation of the

applicability domain of such models and selection of suitable NP descriptors.[16]

Third, we will adapt FRAMES from traditional environmental risk assessment to nanomaterial risk assessment. We will reorganize the elements of FRAMES relevant to modeling nanomaterial environmental impact, which connects various elements in four different layers: source, transport, food chain, and exposure/risk. In this framework, we will be able to incorporate models, whether simulation or analytical, to the system.

Finally, we will adopt ideas from the recent NanoInformatics 2020 Roadmap[1] as we innovate within the NEIMiner roadmap. The Nanoinformatics 2020 Roadmap was designed to articulate comprehensive industrial, academic, and government needs for a successful nanoinformatics enterprise. One goal for the nanoinformatics community is to provide the architecture and vision that allows for data discovery, standardization, verification, and data sharing. The major rationale for this goal is to support the development of prediction models. These models, once subject to validation and verification, can provide information that complements experimental data and potentially motivates subsequent research.

Acknowledgments

The authors wish to thank the US Army Corps of Engineers for their support under research contract No W912HZ-11-P-0009 and W912HZ-12-C-0004; National Institute of Environmental Health Sciences grants ES017552-01A2, ES016896-01 and P30 ES03850; and AFRL FA8650-05-1-5041. An earlier version of this paper was presented at the Institute of Electrical and Electronics Engineers (IEEE) Workshop on Nanoinformatics for Biomedicine, in conjunction with the 2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2012).

Disclosure

The authors report no conflicts of interest in this work.

References

1. National Nanomanufacturing Network, Nanoinformatics 2020 Roadmap, http://www.internano.org/content/view/510/251/. Accessed November 25, 2012.
2. Framework for Risk Analysis of Multi-Media Environmental Systems (FRAMES), http://www.epa.gov/extrmurl/research/3mra.html. Accessed November 25, 2012.
3. Liu X, Tang K, Harper S, Harper B, Steevens J, Xu R. Predictive modeling of nanomaterial biological effects. IEEE Workshop on Nanoinformatics for Biomedicine, in Conjunction with IEEE International Conference on Bioinformatics and Biomedicine (BIBM); October 4–7, 2012; Philadelphia, PA, USA.
4. Tang K, Liu X, Harper S, Steevens J, Xu R. NEIMiner: a model driven data mining system for studying environmental impact of nanomaterials. IEEE Workshop on Nanoinformatics for Biomedicine, in Conjunction with IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2012; Philadelphia, PA, USA.
5. Tang K, Liu X, Tang Y, et al; ABMiner: A scalable data mining framework to support human performance analysis. International Conference on Applied Human Factors and Ergonomics; July 17–20, Miami, FL, USA.
6. Liu X, Tang K, Buhrman JR, Cheng H. An agent-based framework for collaborative data mining optimization. Proceedings of the IEEE International Symposium on Collaborative Technologies and Systems (CTS). May 17–21, 2010; Chicago, Illinois, USA.
7. Web Service. Available from: http://en.wikipedia.org/wiki/Web_service. Accessed November 25, 2012.
8. Web Scraping. Available from: http://en.wikipedia.org/wiki/Web_scraping. Accessed November 25, 2012.
9. ForNova. How to compare and choose scraping tools? Available from: http://www.fornova.net/blog/?p=18. Accessed November 25, 2012.
10. Drupal Peer Review. Available from: http://groups.drupal.org/peer-review. Accessed March 6, 2013.
11. Nanomaterial-Biological Interactions Knowledgebase. Available from: http://oregonstate.edu/nbi. Accessed November 25, 2012.
12. Kumari A, Yadav SK, Yadav SC. Biodegradable polymeric nanoparticles based drug delivery systems. *Colloids Surf B Biointerfaces*. 2010; 75(1):1–18.
13. Liu R, Rallo R, George S, et al. Classification nanoSAR development for cytotoxicity of metal oxide nanoparticles. *Small*. 2011;7(8): 1118–1126.
14. Selassie CD. History of quantitative structure-activity relationships. In: Abraham DJ, editors. *Burger's Medicinal Chemistry and Drug Discovery*. New York: Wiley; 2003:1–48.
15. Cattaneo AG, Gornati R, Sabbioni E, et al. Nanotechnology and human health: risks and benefits. *Journal of Applied Toxicology*. 2010;30(8):730–744.
16. Puzyn T, Leszczynska D, Leszczynski J. Toward the development of "nano-QSARs": advances and challenges. *Small*. 2009;5(22): 2494–2509.