Semantic integration of US Federal nanomaterials data

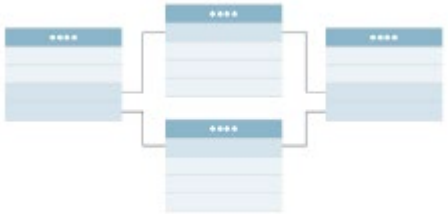
Holly M. Mortensen
Senior Research Biologist
US EPA/Office of Research and Development

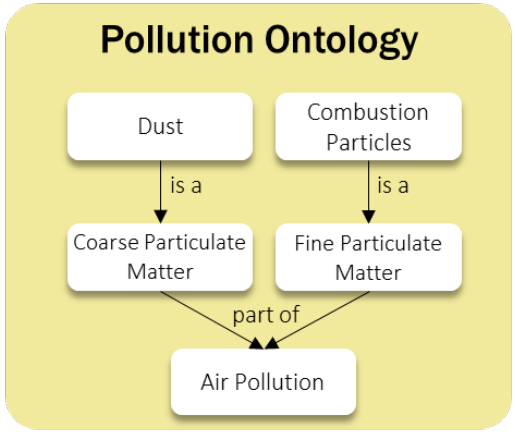
FCSM Research and Policy Conference
Open-Source Software in the Federal Statistical System
Chair: Chris Marcum
University of Maryland, College Park
October 22, 2024, 2pm EST

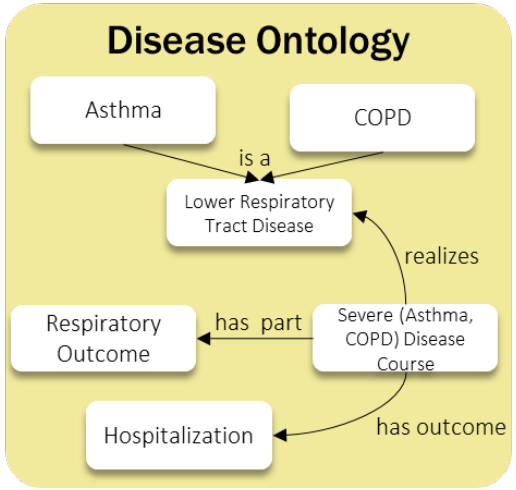
Presentation Outline

- Background*: **Knowledge Organization Systems***
  - *Relational vs. Graph representations-examples*
  - **Environmental Health Data Diversity-EHLC**
- EPA Nanomaterials Knowledgebase- ***NaKnowBase***
  - *Motivations-Nomenclature debacle*
  - Proof of Concept-Semantic/Ontology mapping of *NKB* and the EPA *OntoSearcher*
- ***Consortium Effort**:* NNI NEHI *Database Interoperability Group (DIG)*
  - US Federal Agency NanoEHS Consortium Established
- Progress
  - NNI NanoInformatics Conference Nov. 2023*; Conference Proceeding Pub. 2024*
  - 2024 NNI EHS Research Strategy Update (prev. 2011)
  - *Progress and Future directions*

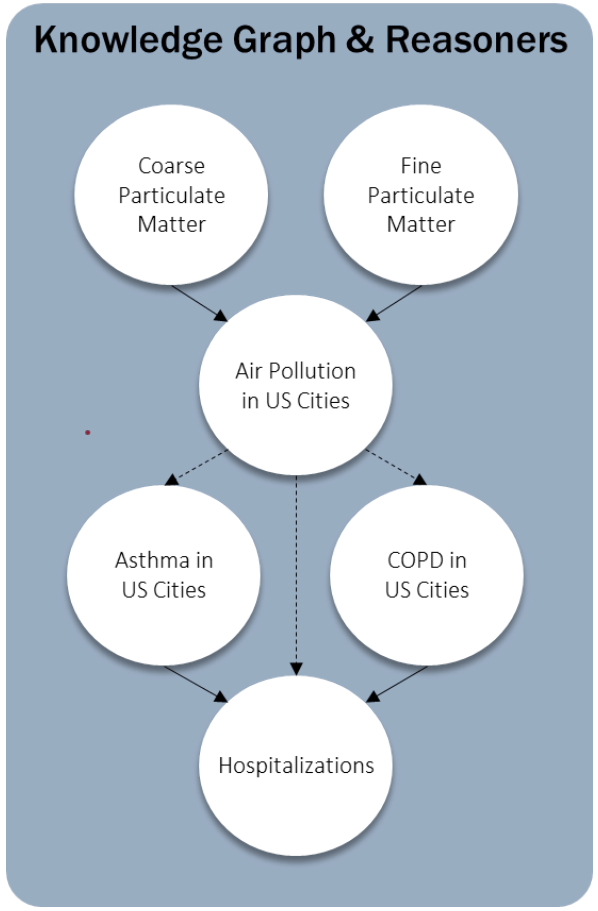
Knowledge Representation







Databases
For Air Quality and Disease Prevalence in US Cities*



→ Defined Relationship ---→ Inferred Relationship

*Captured using common data elements in a data model with minimal information standards captured using controlled vocabularies

Less — Semantic Clarity/Complexity — More

Knowledge Organization | Knowledge Representation

Term List

List of terms with definitions but no defined structure or relationships between terms

VS

Taxonomy

List of terms with definitions organized within hierarchical groups or categories and with relationships between terms defined

VS

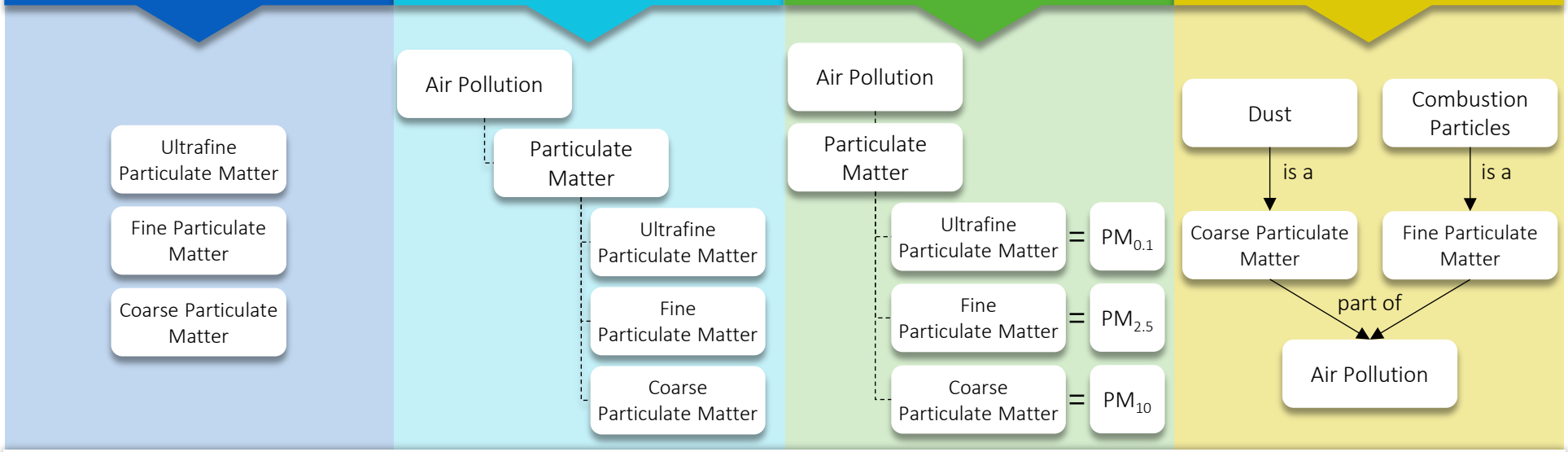
Thesaurus

List of terms with definitions and synonyms defined and with relationships between terms defined

VS

Ontology

Formal representation of a body of knowledge that is human- and machine-readable and with relationships between terms defined



Why Use It?

Consistent	Hierarchical	Associative	Inferential
Collaboration and search are easier when everyone is using a set of terms with an agreed upon definition.	The hierarchical structure helps us integrate data collected at different levels of granularity and search for data using high level categories.	In addition to the hierarchy, the presence of synonyms further improves integration and search over heterogeneous data.	The combination of formal logic and **persistent, unique identifiers** enables inferencing, makes data computable, and reveals novel connections.

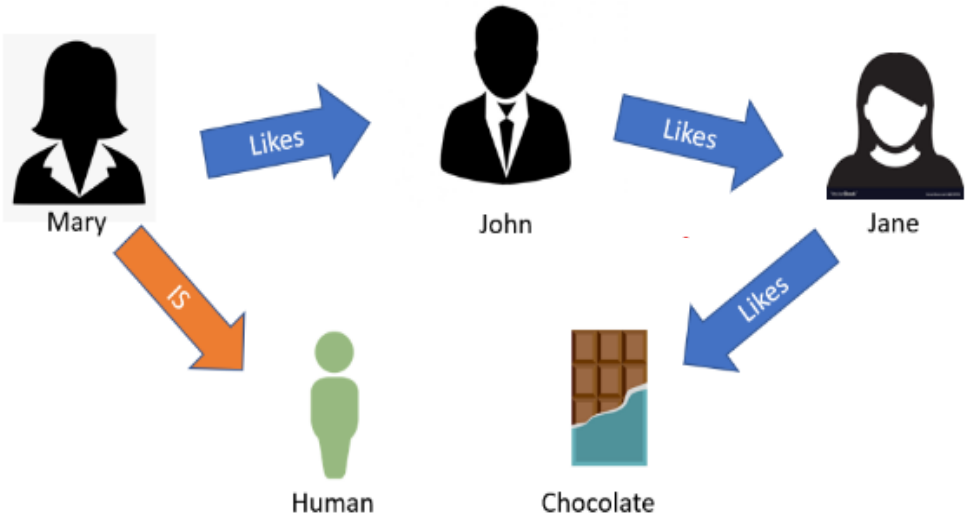
Examples

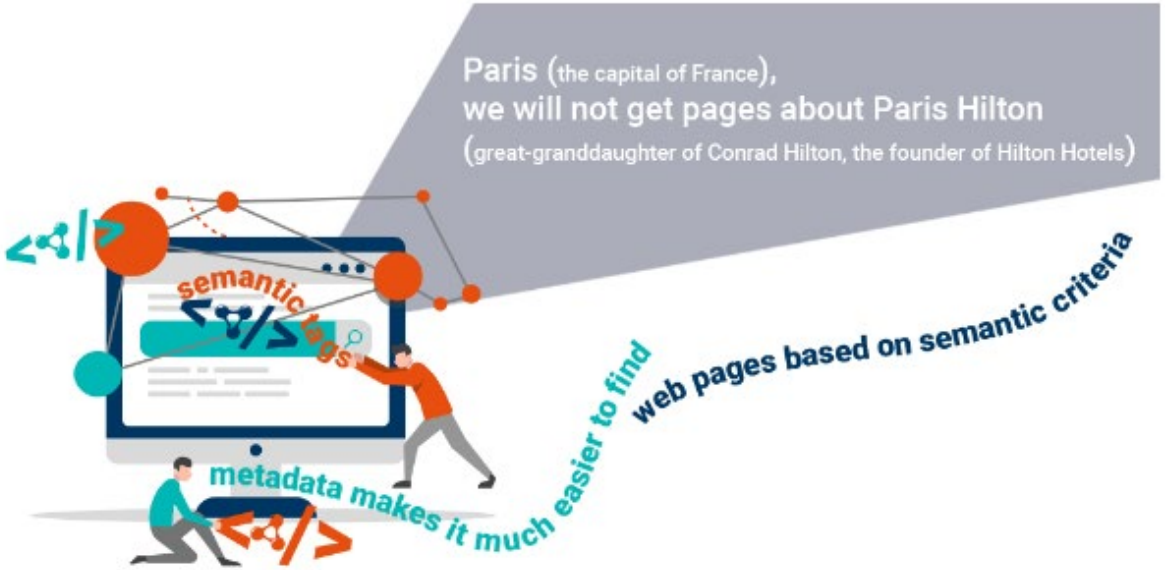
HAWC EHV; LTER Controlled Vocabulary	NCBITaxon	NCIThesaurus; MeSH	OBO Foundry Ontologies; UBERON

Some Terminology…

- Semantic mapping is a way of representing information (concepts or data) as a graph
- Resource Description Framework (RDF) is a directed graph AND a data model for exchanging information on the web
  - triples= subject, predicate, and object
- Is RDF appropriate for nanoEHS data?

  -"***Nomenclature debacle***"- lack of consistent nomenclature across sources exacerbates integration
- ***Concept***-By adding in the metadata component, semantic technologies can address data heterogeneity and interpolation issues



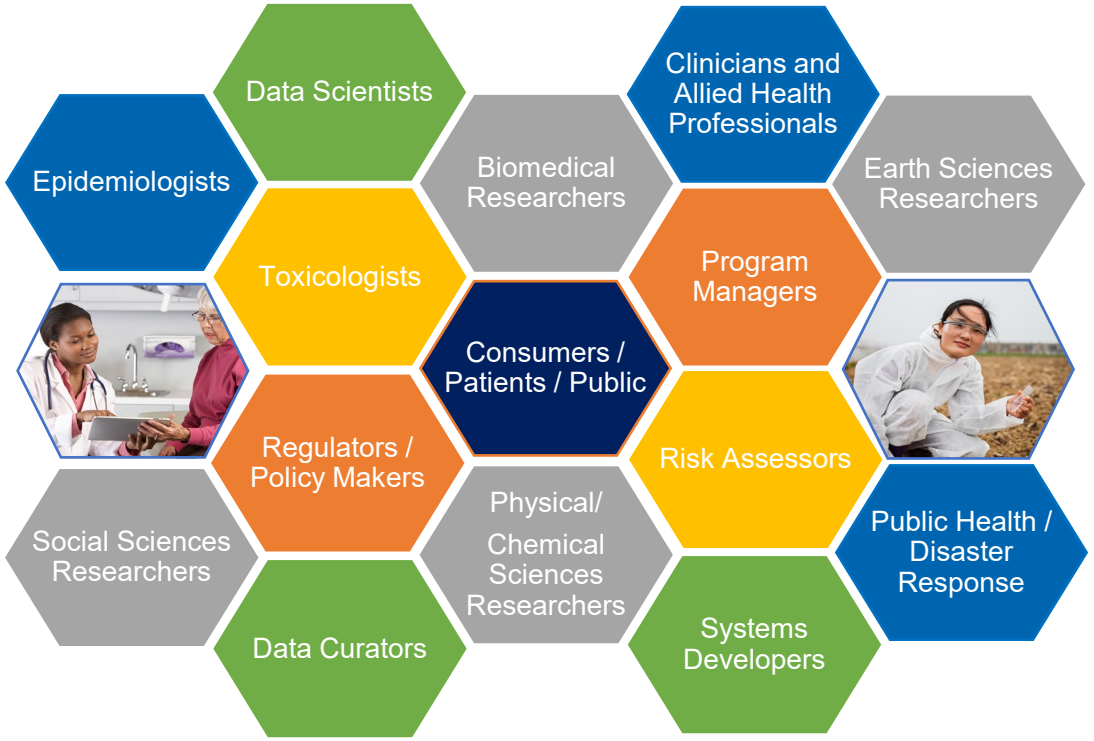


https://medium.com/terminusdb/graph-fundamentals-part-1-rdf-60dcf8d0c459
https://www.ontotext.com/knowledgehub/fundamentals/what-is-the-semantic-web/

Environmental Health Data Diversity: *subfield contributers, language reporting standards, and actors and stakeholders*

The Environmental Health Language Collaborative

Harmonizing Data. Connecting Knowledge. Improving Health.



Provided by Charles Schmitt, NIEHS, and adapted from the NIH Dept of Health and Human Services.

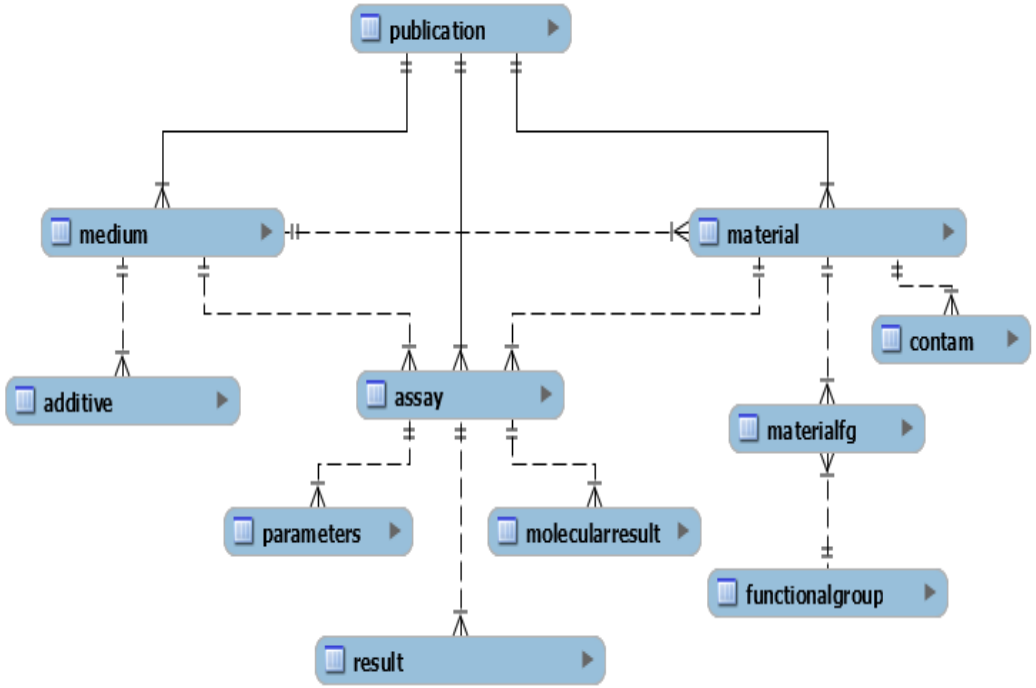
EPA NaKnowBase

A Curated, Relational Database detailing physical-chemical properties of EPA emerging materials research

EPA OntoSearcher

Automated assignment of ontology terms and graph creation

NaKnowBase
nanomaterials knowledge database



Relational SQL database

Curated Data Fields:

- Publication (DOI, author etc.) **>120 peer-reviewed manuscripts (2012-2019)**
- Materials **(>70 unique NM)**
  - Physical/Chemical properties
  - Capping materials
  - Media
  - Contaminants
- Assays **(>160 named assays)**
  - Parameters Measured
- Results

Boyes, W.K., Beach, B., Gayle Chan, G., Thornton, B.M., Harten, P., **Mortensen, H.M**. (*2022*) **An EPA database on the effects of engineered nanomaterials-NaKnowBase.** *Nature Sci Data* **9**, 12. https://doi.org/10.1038/s41597-021-01098-0.

8

The EPA NaKnowBase-SQL backend-080121. *U.S. EPA Office of Research and Development (ORD)* https://doi.org/10.23719/1522951.

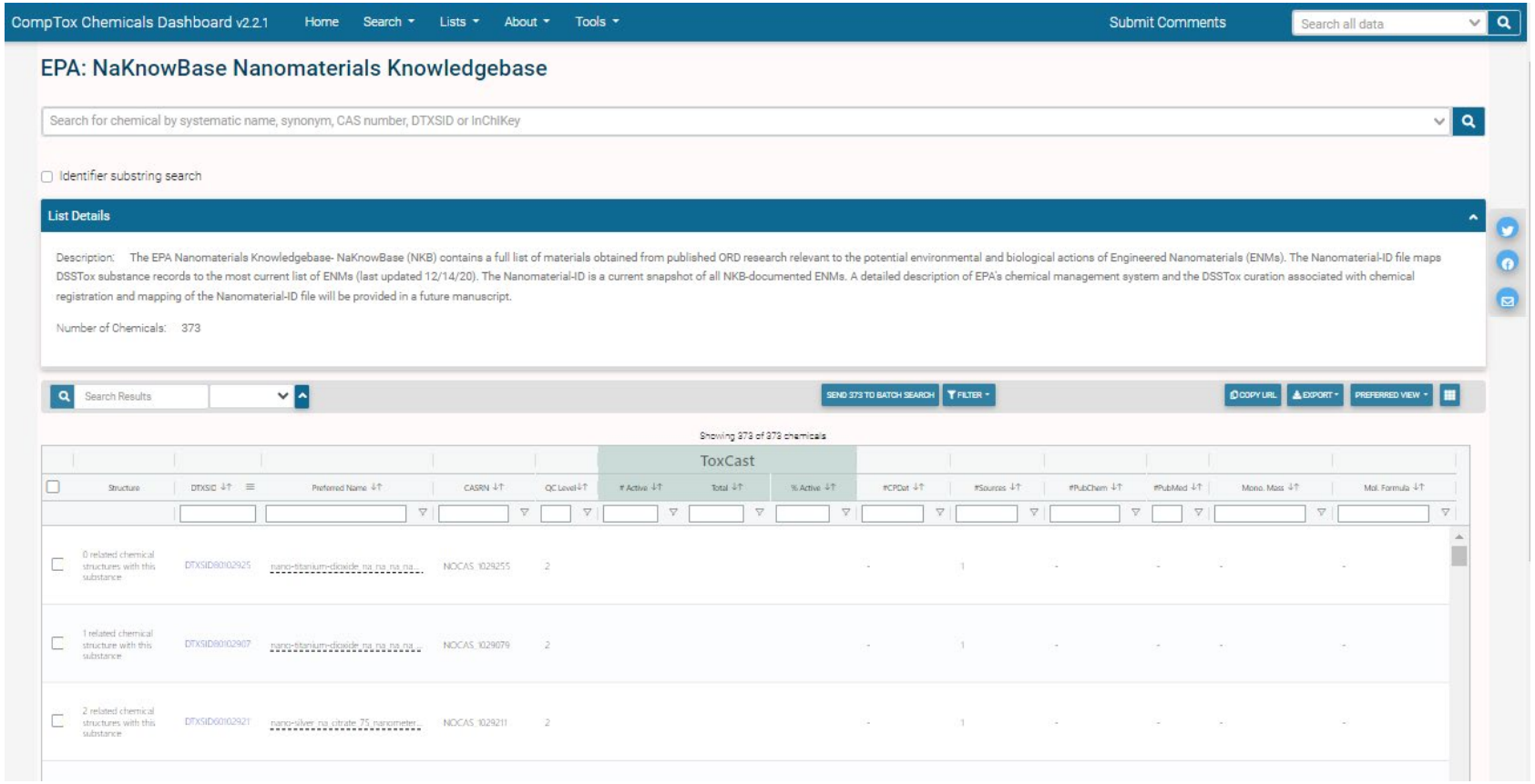
Natural Language Processing (NLP) for Computable and Interoperable Descriptions of EPA nanomaterials

nano-SiO2_NA_NA_NA_nm_ENPRA_A

This nano scaled material is composed of a SiO2 core with no information on the surface coating or capping agent. There is no data on manufacturer reported particle size (diameter). This material was obtained from ENPRA as sample A of the material of the same core, coating/capping, diameter and source information.

Also Discussed in: Watford, et al. (2019) Progress in data interoperability to support computational toxicology and chemical safety evaluation, *TAAP Special Issue in Modernization of Risk Assessment (ed. Holly Mortensen)*

NKB nanomaterials on the EPA CompTox Chemistry Dashboard



The Nanomaterial-ID file maps DSSTox substance records

- **373 ENMs mapped to DSSTox IDs**

EPA's ***DSSTox (Distributed Structure-Searchable Toxicity***) database contains curated chemical substances mapped to chemical identifiers (i.e., chemical synonyms, systematic names, CAS Registry Numbers and others) and, where appropriate, chemical structure representations.

https://comptox.epa.gov/dashboard/chemical-lists/NAKNOWBASE

User Interface Initiated

Naknowbase About How to Cite Tools

1 Material | 2 CoreComposition | 3 = | 4 value | X

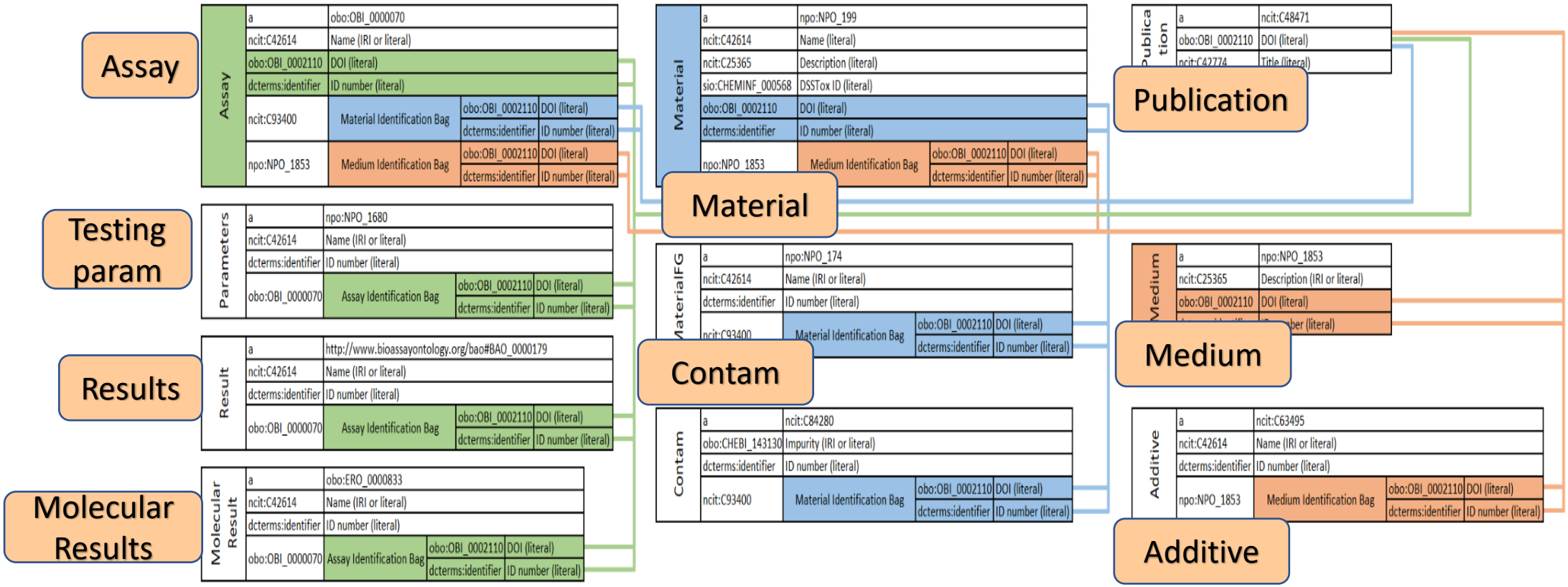
1. List of tables in NKB
2. List of fields in the selected table.
3. Comparison symbols for numeric fields.
4. List of values for non-numeric fields, or a text box for numeric fields.

+ - Submit Restart

Submit a Query...

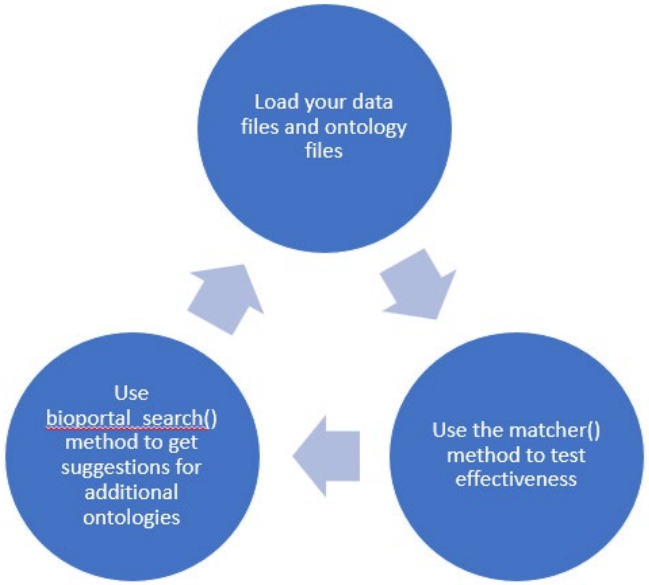
Internet Subdomain Request Initiated---HOLD:
naknowbase.epa.gov

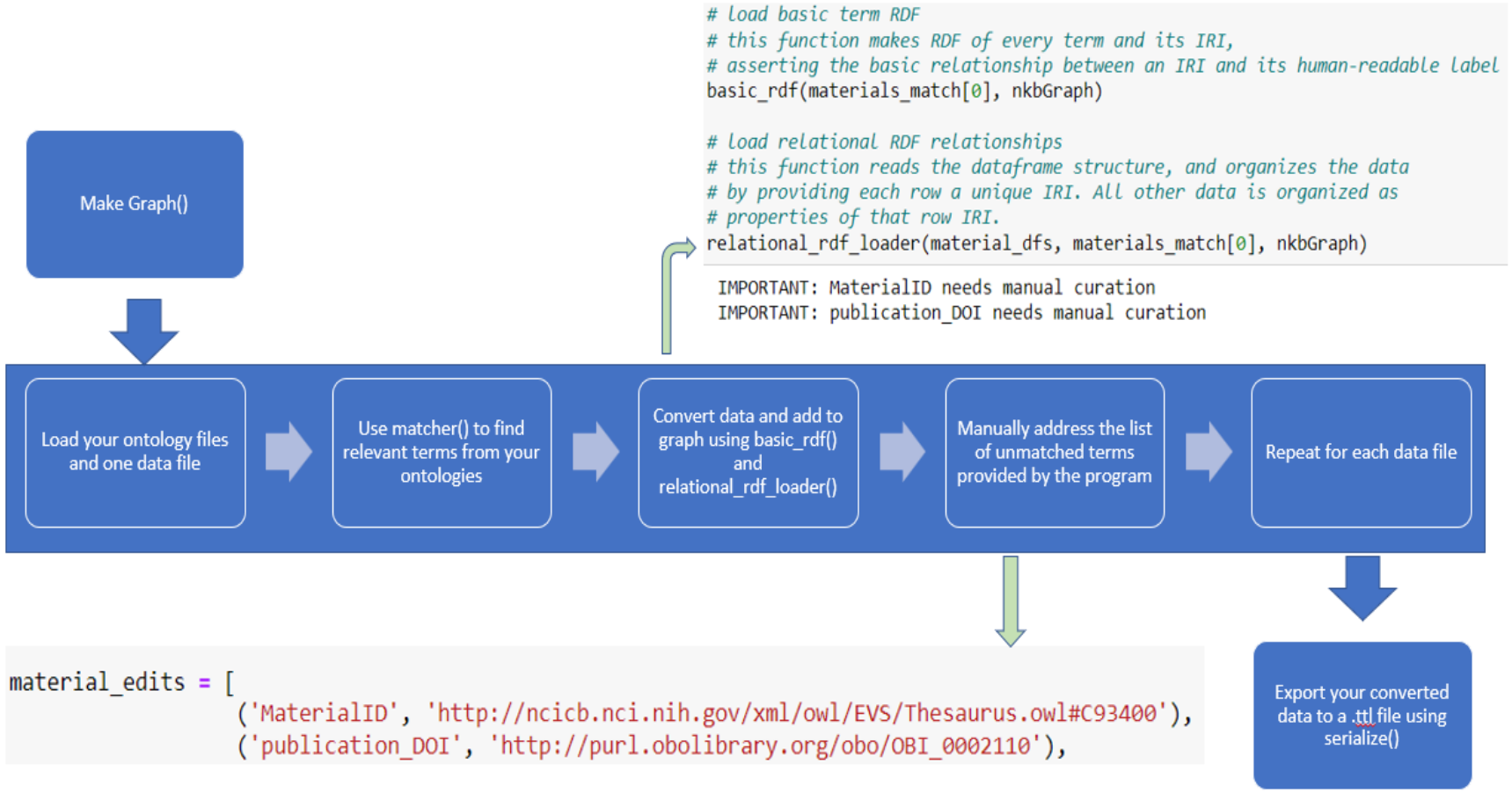
Semantic mapping of NKB using the OntoSearcher tool



Mortensen HM, Beach B, Slaughter W *et al.* Translating nanoEHS data using EPA NaKnowBase and the resource description framework [version 1; peer review: 2 approved]. *F1000Research* 2024, **13**:169 (https://doi.org/10.12688/f1000research.141056.1)

EPA OntoSearcher for Automated Term Mapping and Software-guided graph creation





- Automated mapping with match score to Assists in manual curation
- Convert data into RDF
- .csv input to .ttl output

Mortensen HM, Beach B, Slaughter W *et al.* Translating nanoEHS data using EPA NaKnowBase and the resource description framework [version 1; peer review: 2 approved]. *F1000Research* 2024, **13**:169 (https://doi.org/10.12688/f1000research.141056.1)

Transversing separate datasets: *Interoperability using Federated Queries*

Federated SPARQL Query – AOP-DB pathways by AOP-DB gene and NKB material

```
#determine which NKB material/AOPDB gene combinations are associated with the most pathways
aopdb_fed = """
SELECT distinct ?geneID (COUNT(?pathwayname) as ?p) ?DOI ?material ?DTXSID
WHERE {
    SERVICE <https://aopdb.rdf.bigcat-bioinformatics.org/sparql> {
        ?chemgeneassoc a <http://semanticscience.org/resource/SIO_001257>.
        ?chemgeneassoc <http://semanticscience.org/resource/CHEMINF_000446> ?CAS.
        BIND(URI(CONCAT("http://identifiers.org/cas:", STRAFTER(str(?CAS), "https://identifiers.org/cas:"))) as ?CAS2).
        ?chemgeneassoc <http://edamontology.org/data_1027> ?geneID.
        ?pathway <http://edamontology.org/data_1027> ?geneID.
        ?pathway a <http://purl.obolibrary.org/obo/PW_0000001>.
        ?pathway dc:title ?pathwayname.
    }
    ?material sio:CHEMINF_000446 ?CAS2.
    ?material a npo:NPO_199.
    ?material npo:NPO_1808 ?core.
    ?material obo:OBI_0002110 ?DOI.
    ?material sio:CHEMINF_000568 ?DTXSID.
}
GROUP BY ?geneID
ORDER BY DESC(?p)
LIMIT 5
"""
qres = g.query(aopdb_fed)
for row in qres:
    print(f"{row.DTXSID} of {row.DOI} | gene {row.geneID} |\n# pathways: {row.p}")
```

```
http://identifiers.org/comptox/DTXSID501028989 of 10.1186/s12951-014-0047-3 | gene https://identifiers.org/ncbigene/196 |
# pathways: 560
http://identifiers.org/comptox/DTXSID501028989 of 10.1186/s12951-014-0047-3 | gene https://identifiers.org/ncbigene/208 |
# pathways: 283
http://identifiers.org/comptox/DTXSID501028989 of 10.1186/s12951-014-0047-3 | gene https://identifiers.org/ncbigene/154 |
# pathways: 196
http://identifiers.org/comptox/DTXSID301028969 of 10.1016/j.watres.2012.12.041 | gene https://identifiers.org/ncbigene/180359 |
# pathways: 56
http://identifiers.org/comptox/DTXSID501028989 of 10.1186/s12951-014-0047-3 | gene https://identifiers.org/ncbigene/948 |
# pathways: 53
```

Federated Query calling for nanomaterials and corresponding gene targets that hit the greatest number of pathways

Mortensen HM, Beach B, Slaughter W *et al.* Translating nanoEHS data using EPA NaKnowBase and the resource description framework [version 1; peer review: 2 approved]. *F1000Research* 2024, **13**:169 (https://doi.org/10.12688/f1000research.141056.1)

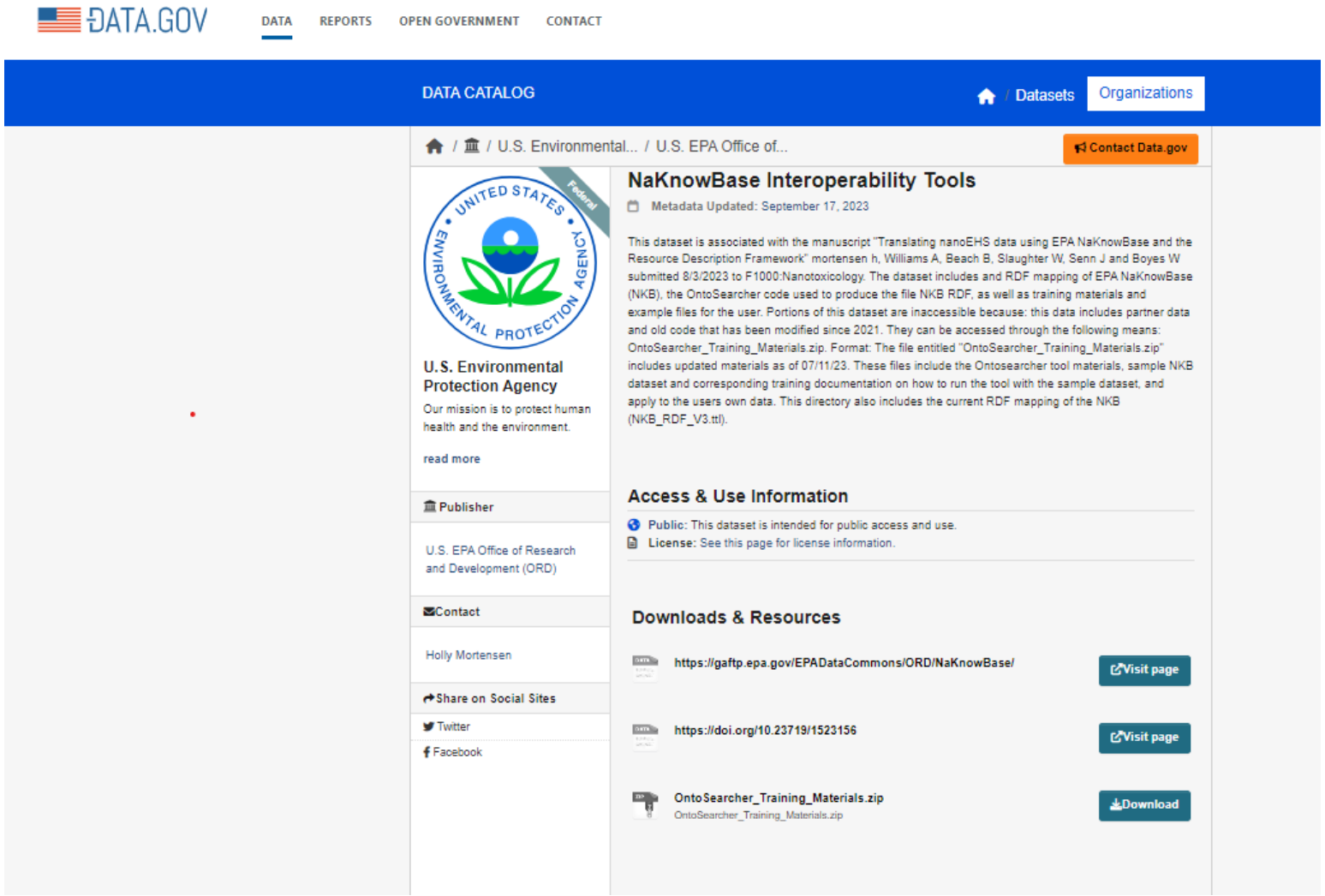
EPA Interoperability tools and Relational DB products

NaKnowBase nanomaterials knowledge database

Onto Searcher

DATA.GOV: NKB Data Catalog

https://catalog.data.gov/dataset/naknowbase-interoperability-tools

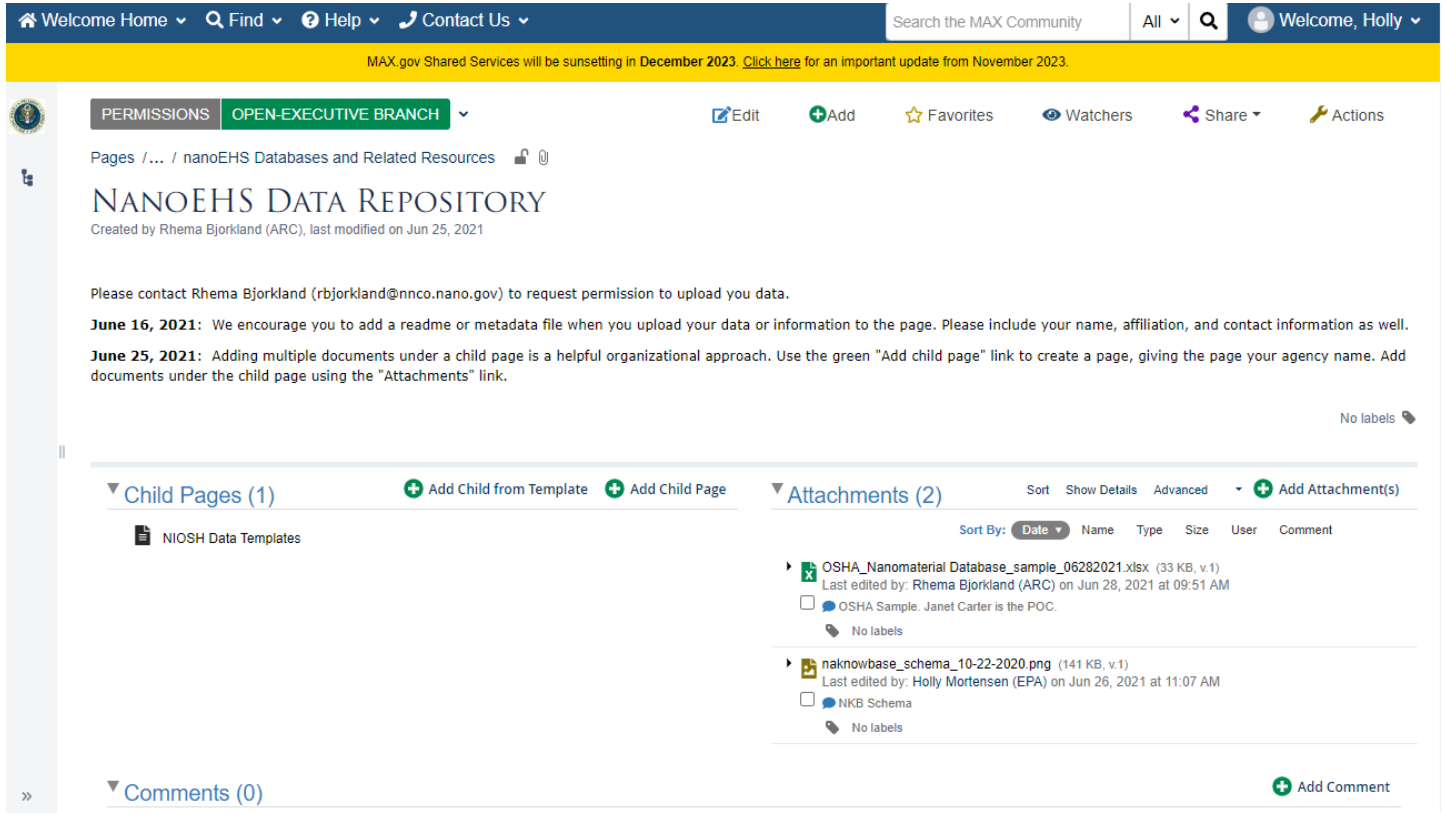


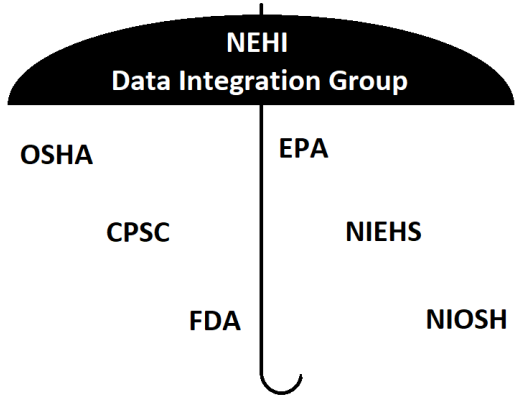
US Federal Agency NanoEHS Consortium

Informatics plan described in
2030 US EU Roadmap and 2024 NNI EHS Research Strategy

Introductory Results and Project plan

US Federal Agency NanoEHS Consortium Established



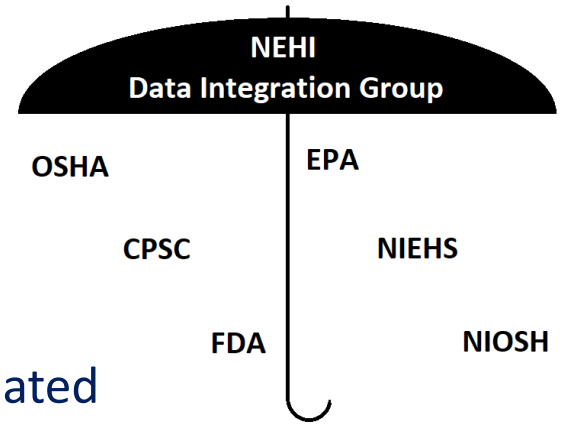


- **2022 NEHI-DIG Consortium Formed**
- Initial Data repository established **(MAX.gov)**
- **Participating Federal Agencies: EPA, NIOSH, OSHA, CPSC, FDA, NIH**

https://www.nano.gov/NNINanoinformaticsConference

Described in the upcoming **NNI EHS Strategy update 2023: Informatics and Modeling.*
Federal Register/Vol. 89, No. 115/Thursday, June 13, 2024/

Motivations for forming the DIG Consortium



- Shared vision of Federal partners – team science, and data sharing
  - ***breaking down data silos!***
- ***Mechanistic Interaction*** of engineered nanomaterials (ENMs) is not yet fully elucidated
- Lack of information on ***toxic relevance*** (*e.g.* which disease outcomes are relevant for ENMs?)
- *Computationally structured, semantic annotation* can improve our ability to understand this biology
  - Promoting FAIR (Finable, Accessible, Interoperable, and Reuseable) data management and sharing principles for ENM would simplify data integration with other knowledge systems
  - EPA NKB as proof of concept- *We can do this!*
  - Stay current with EU progress in this area (??)

NIOSH: EPA processing with OntoSearcher

EPA OntoSearcher: CSV to RDF Conversion

NIOSH Dataset

This document uses EPA's OntoSearcher application to convert multiple CSVs, derived from an Excel workbook of nanomaterial research data provided by NIOSH, into **Resource Description Framework (RDF)**. RDF is a data format which uses unique web addresses, called **Internationalized Resource Identifiers (IRIs)**, to identify pieces of unique information. Associating data with these unique identifiers and publishing that data in RDF format allows for any data regarding the same entity (that shares an IRI) to be interoperable.

EPA OntoSearcher is a prototype application developed at the **Dr. Holly Mortensen** lab at **EPA ORD CPHEA** to expedite the conversion of relational data into RDF. The application provides functions for importing CSV data, importing ontology and RDF data, search algorithm functions to compile a dictionary of IRI's for csv terms, and functions that build RDF from csv data and term-IRI associations.

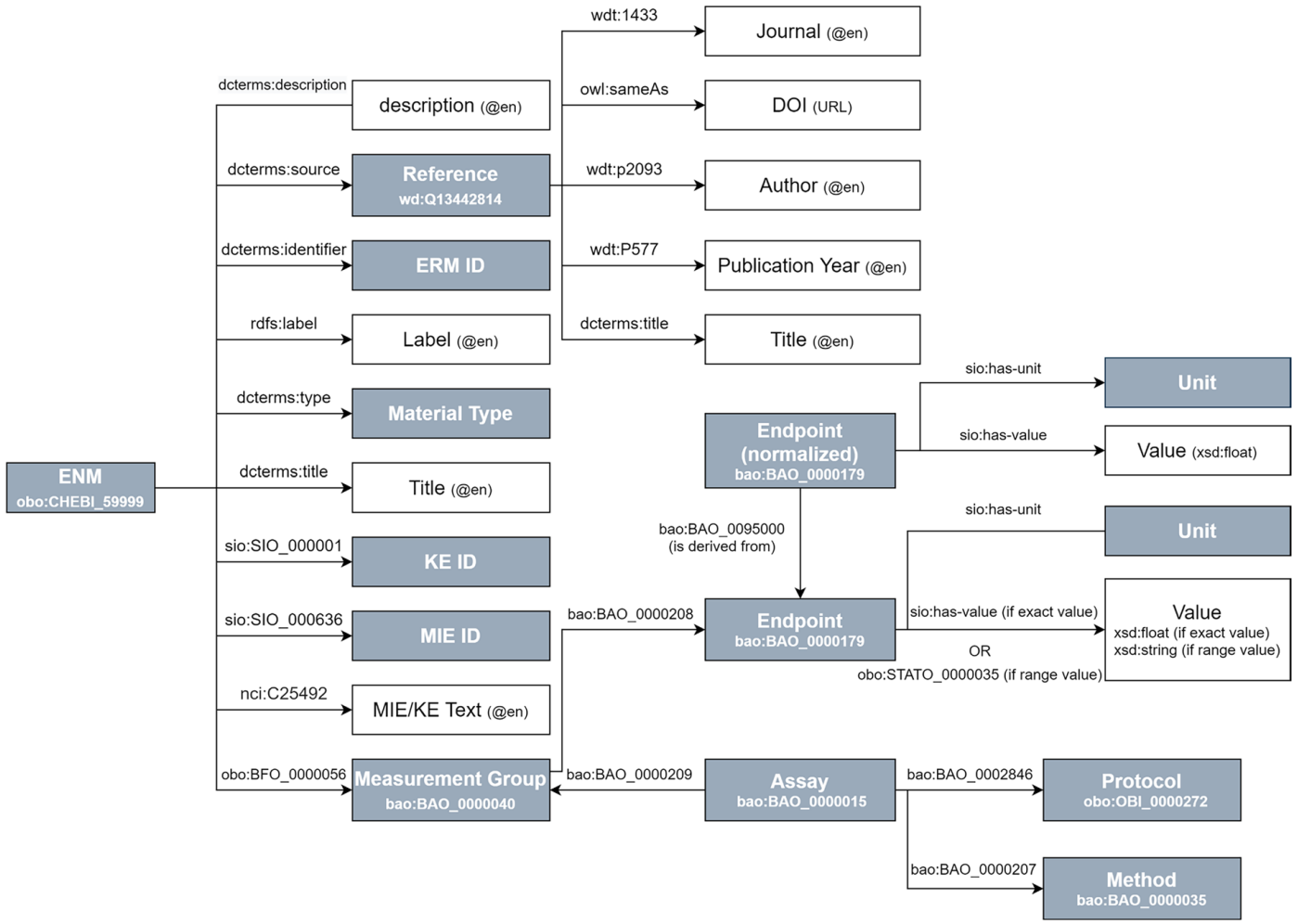
This document will showcase all of this functionality, as well as how to query RDF data using SPARQL, the RDF query language.

Why are we here?

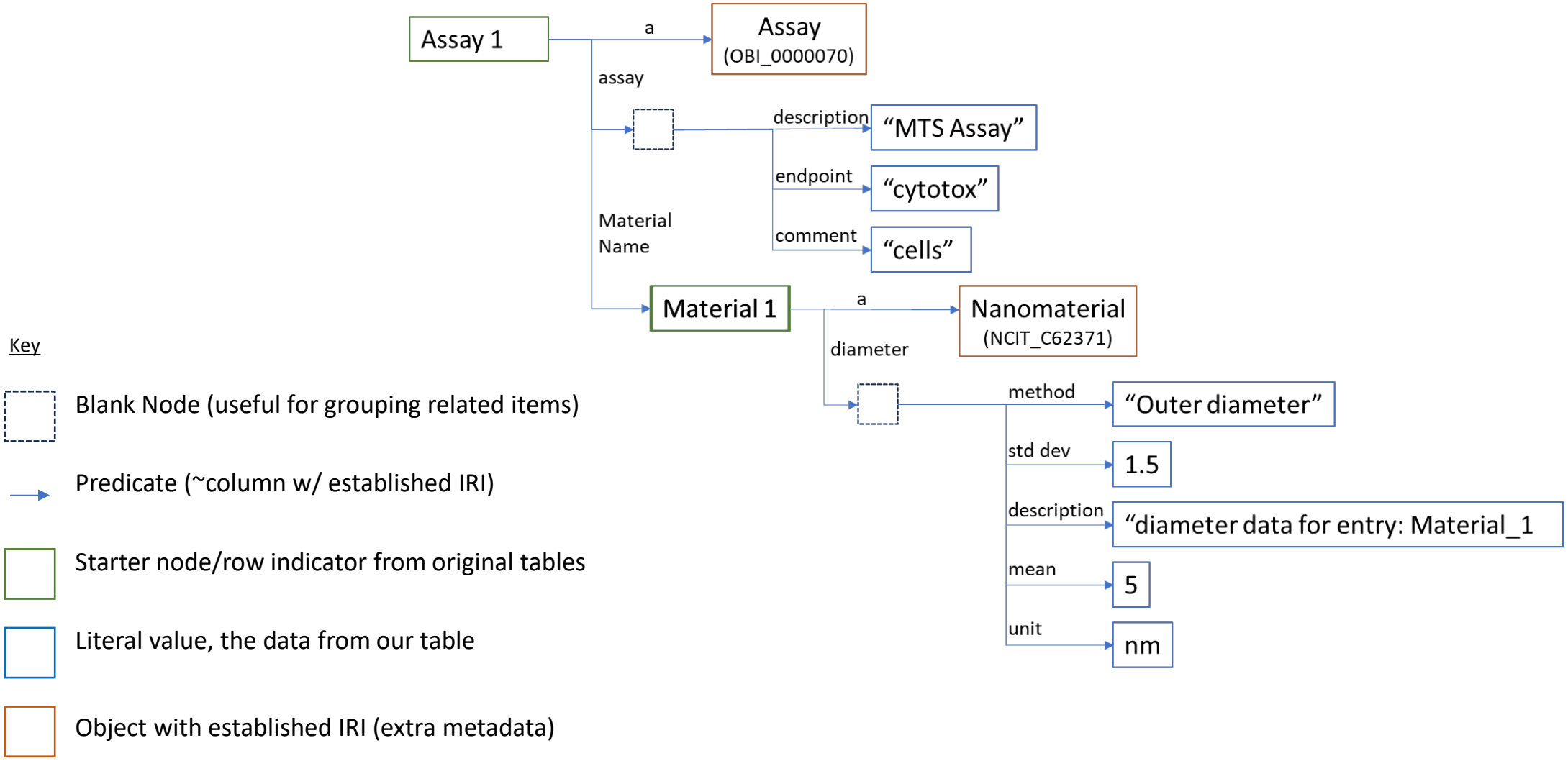
to answer *three questions*

- How should we **format our relational data** to make interoperability easier?
- What is **RDF/OWL** and **what utlitity does it have** for my needs?
- **How can I convert my data** into RDF/OWL (without breaking a sweat)?

```
# import EPA OntoSearcher modules, and other packages
from onto import ontolister, ontocontext
from csv_importer import load_data
from find import matcher
from onto_api import bioportal_search, unpack_superclass
from onto_api import bioportal_sample, dict_samp, bio_summary
from rdf_print import table_from_file, term_editor, term_lookup
from rdf_print import basic_rdf, relational_rdf_loader
from rdf_print import primenode, node_one, node_two, multi_editor
```



NIOSH: EPA manual interrogation



NNI nanoinformatics conference 2023

https://doi.org/10.1016/j.comtox.2024.100316

NNI National Nanotechnology Initiative

ABOUT NNCO | Search...

About Nanotechnology | National Nanotechnology Initiative | Events & Initiatives | News & Impact | Reports & Resources

NNI Nanoinformatics Conference

November 15, 2023
In-person, Washington, D.C.

Computational Toxicology 30 (2024) 100316

Contents lists available at ScienceDirect

Computational Toxicology

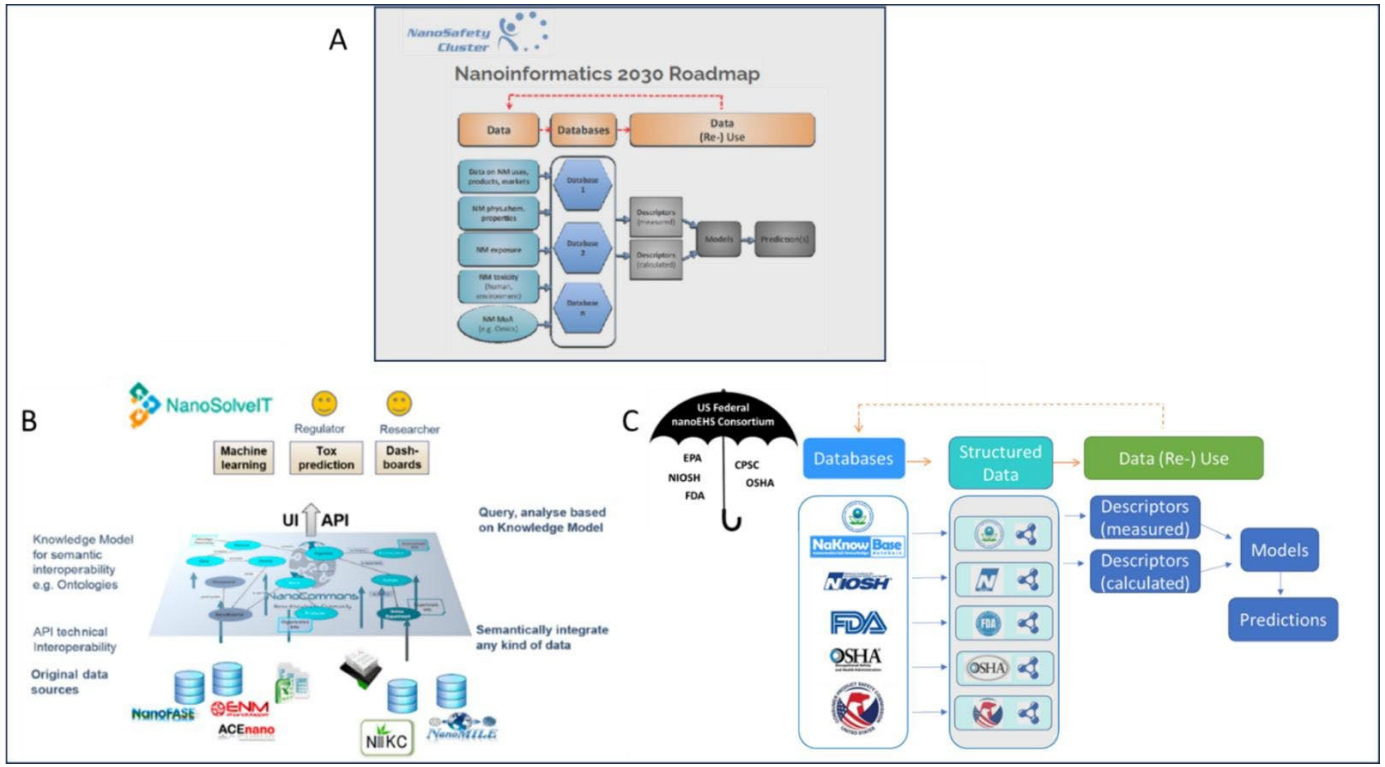
journal homepage: www.sciencedirect.com/journal/computational-toxicology

Short Communication

NNI nanoinformatics conference 2023: Movement toward a common infrastructure for federal nanoEHS data computational toxicology: Short communication

Holly M. Mortensen [a,*,1], Jaleesia D. Amos [b,2], Thomas E. Exner [c,3], Kenneth Flores [d,4], Stacey Harper [e,5], Annie M. Jarabek [f,6], Fred Klaessig [g,7], Vladimir Lobaskin [h,8], Iseult Lynch [i,9], Christopher S. Marcum [j,10], Marvin Martens [k,11], Branden Brough [l], Quinn Spadola [l,12], Rhema Bjorkland [m,13]

[a] Public Health and Integrated Toxicology Division, Center for Public Health and Environmental Assessment, Office of Research and Development, US Environmental Protection Agency, Research Triangle Park, NC, United States
[b] Civil & Environmental Engineering, Duke University, Durham, NC 27708, United States
[c] Seven Post Nine GmbH, Rebacher 68, 79650 Schopfheim, Germany
[d] NSF MPS Ascend Fellow, Arizona State University, Tempe AZ 85287, United States
[e] Department of Environmental and Molecular Toxicology and the School of Chemical, Biological and Environmental Engineering, Oregon State University, Corvallis, OR 97331, United States
[f] Health and Environmental Effects Assessment Division, Center for Public Health and Environmental Assessment, Office of Research and Development, US Environmental Protection Agency, Research Triangle Park, NC 27711, United States
[g] Pennsylvania Bio Nano Systems, PA, United States
[h] School of Physics, University College Dublin, Dublin 4, Ireland
[i] University of Birmingham, Birmingham, UK
[j] Office of the Chief Statistician of the United States, Office of Management and Budget, Washington, DC 20503, United States
[k] Department of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, Maastricht, the Netherlands
[l] National Nanotechnology Coordination Office, Alexandria, VA, United States
[m] Contractor to the National Nanotechnology Coordination Office, Alexandria, VA, United States



Mortensen, H.M., Amos, J.D., Exner, T.E., Flores, K., Harper, S., Jarabek, A.M., Klaessig, F., Lobaskin, V., Lynch, I., Marcum, C.S., Martens, M., Brough, B., Spadola, Q., Bjorkland, R. (2024) NNI nanoinformatics conference 2023: Movement toward a common infrastructure for federal nanoEHS data computational toxicology: Short communication, Computational Toxicology, Volume 30,2024,100316.

EU US Roadmap Nanoinformatics 2030

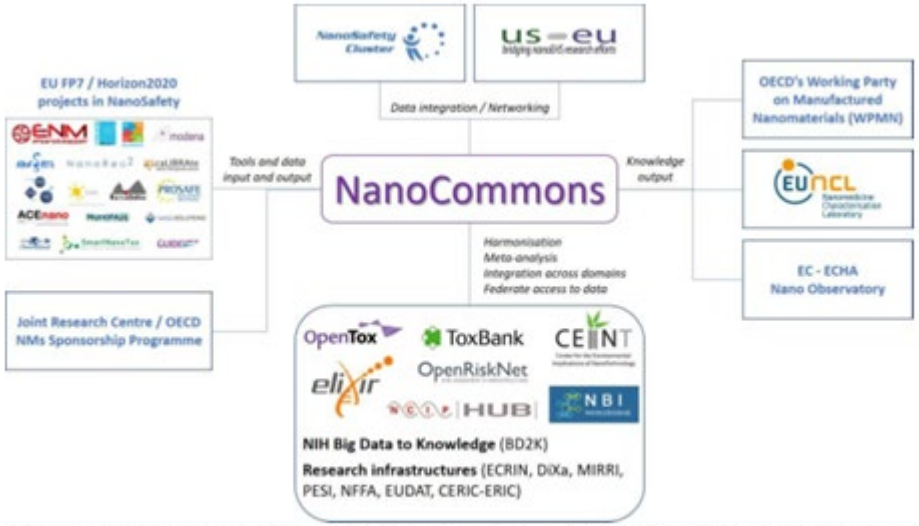
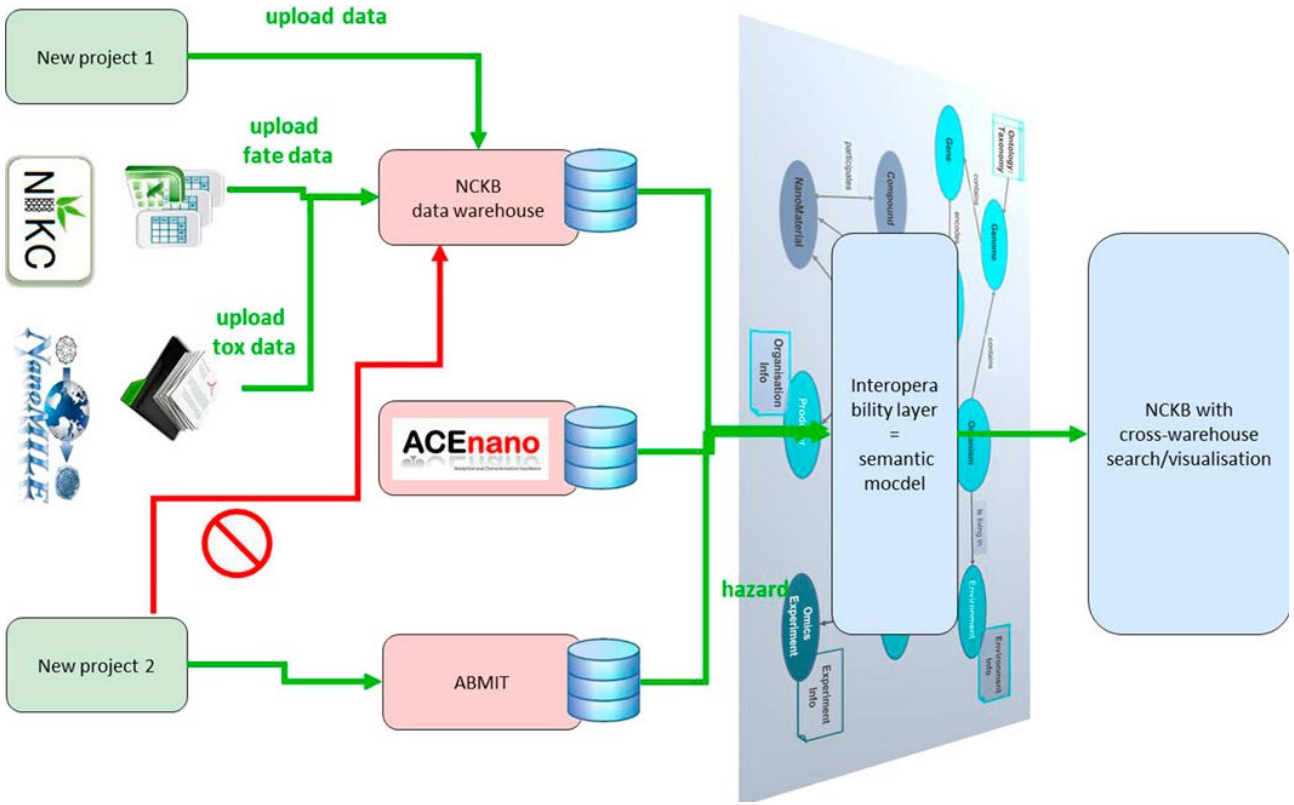


Figure 16: Schematic illustration of the positioning of NanoCommons and how it will provide an integrating platform for the nanosafety knowledge community in Europe and internationally.

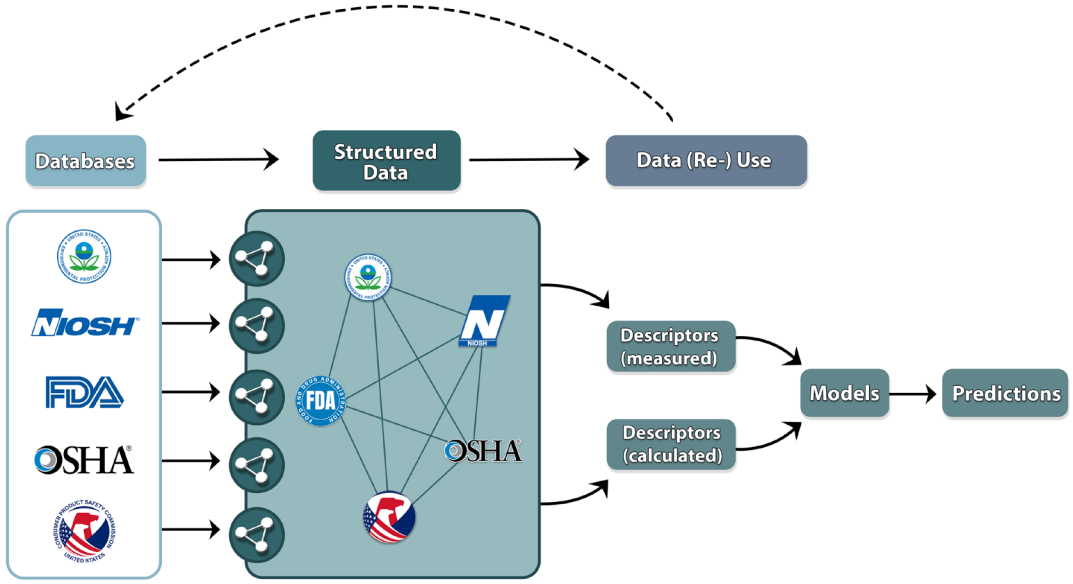


Maier, et al. Front. Phys., 12 November 2023
https://doi.org/10.3389/fphy.2023.1271842

2024 NNI EHS Research Strategy

NATIONAL NANOTECHNOLOGY INITIATIVE
ENVIRONMENTAL, HEALTH, AND SAFETY
RESEARCH STRATEGY: 2024 UPDATE

DRAFT FOR PUBLIC COMMENT, 06/10/24



- Identifies Nanoinformatics as a cross-cutting theme
- Needs
  - Expand and Strengthen the Collaborative Informatics Infrastructure
  - Boost informatics and data infrastructure for robust risk assessment and decision-making
  - Ensure alignment with FAIR and TRUST principles:

Requires the development of standardized protocols for data reporting and sharing, ensuring that data generated from nanotechnology research is easily findable and accessible to a broad range of stakeholders

2024 Project proposed

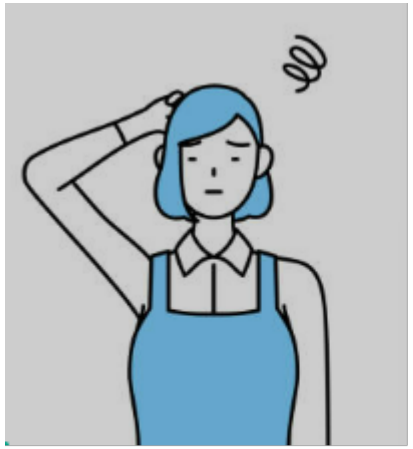
ATTACHMENT I- Project Topic

Federal Government Data Integration and Usage Platform for Emerging and Nanomaterial Environmental and Health Safety

1. A signed ***MOU or other agreement vehicle*** establishing data sharing, collection, usage.
2. A ***web tool*** to support partner data sharing and accessibility, as well as semantic interoperability using the *proof-of-concept* tools.
3. A ***data model*** for contributing federal partner nanoEHS data.
4. A prototype data usage dashboard or other presentation and tools.
5. ***Communication Plan for*** implementation and deployment across federal agencies.
6. Training data (if applicable), and any other data created under this award.

Future work and Questions/Concerns

- ***Staffing and project support***-NNI interns assigned to expedite processing (Prakash Pranav -UC San Diego)
  - (*Post-OntoSearcher*) Manual Curation – ***Agency Data owners are the lag***
- Are one-to-one reuse of existing ontologies the way to go?
  - Different ontologies use different structures- ***Combining terms reused from other ontologies with new domain-specific terms proves consistently problematic***
  - Large number of terms map into the ontology! But…the new ontology has a ***specific application domain***
    - Precisely the domain-specific terms that are most relevant, but not available somewhere else.
- **What is needed:** ***Agency coordination and needed expertise***



Planned virtual all-day event

EHLC Use case focusing on AOP-biomedical entity mapping (Early 2025) –contact **mortensen.holly@epa.gov**

LINKS to EPA projects

- EPA NaKnowBase and related tools
  https://catalog.data.gov/dataset/naknowbase-interoperability-tools

- AOP-DB web user interface https://aopdb.epa.gov/

  (permanently decommissioned as of 2024)

- The AOP-DB v.2 stressor linkages are provided through the CompTox Chemicals Dashboard-Number of Chemicals: 349:
  https://comptox.epa.gov/dashboard/chemical_lists/AOPSTRESSORS

- AOP-DB SPARQL Endpoint:

  https://github.com/BiGCAT-UM/AOP-DB-RDF

  https://aopwiki.org/info_pages/8

Acknowledgements

EPA
Will Boyes
Antony Williams
EPA-NPD
Annette Guiseppi-Elie
Kathy Dionisio
EPA Students/ Contractors
Bradley Beach
Weston Slaughter
Jonathan Senn
Bradley Sutliff

US Federal Partners
Treye Thomas **CPSC**
Joanna Matheson **CPSC**
Janet Carter **OSHA**
Jay Vietas **CDC/NIOSH**
Kuempel, Eileen **CDC/NIOSH**
Nathan Drew **CDC/NIOSH**
Anil Patri **FDA**

Duke-CIENT/InFRAMES
Mark Weisner
Jaleesia Amos

International Partners
Thomas Exner
Egon Willighagen
Marvin Martens
Andrea Haase
Penny Nymark

NNI-NEHI
Branden Brough
Quinn Spadola
Rhema Bjorkland
Geoff Holdridge

EPA National Program Support: **Chemical Safety and Sustainability**, RA3: Emerging Materials and Technologies, CSS 403.1: Evaluate environmental impacts of emerging materials on humans and ecological species, Product 403.1.4: **Improved NaKnowBase Data Integration to Meet Program and Partner Needs**
AND RA8: Informatics, Synthesis, and Integration: CSS.408.2 - Knowledge delivery and interoperability in support of chemical safety decisions, Product 408.2.22 **Development of infrastructure support for EMT: EPA NaKnowBase**