# Data Standards to Support Integrated Source to Outcome Modeling

David Hines[1], Jaleh Abedini[1*], Shannon Bell[1], Carol M. Hamilton[1], Nomi Harris[3], Sierra Moxon[3], Chris Mungall[3], James Rineer[2]
[1]RTI International, Research Triangle Park, NC | [*]Currently at US EPA, ORD, Research Triangle Park, NC | [3]Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Abstract Number: 4028
Poster number: J522

## Objectives

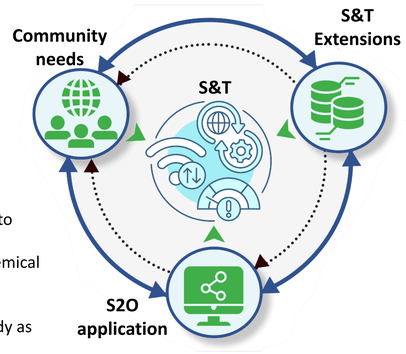
**Coordinate standards across EHS study areas**

**Expand standards & terminologies**

**Improve data interoperability**

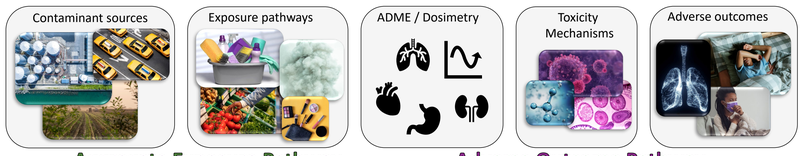
## Goals

- Environmental Health Science (EHS) encompasses numerous subdomains along the source-to-outcome (S2O) continuum, each generating unique and nuanced data.
- Standards and terminologies (S&T) across subdomains (e.g. chemical release, exposure, outcome) are not well coordinated, limiting data **interoperability**.
- We aim to improve the precision of semantic descriptions along the S2O continuum to facilitate communication among humans and machines and strengthen predictive capabilities.
- This work bolsters data **interoperability,** a component of the FAIR (Findable, Accessible, **Interoperable**, Reusable) principles, by:
  1. Engaging the expert and stakeholder communities to define standard and terminology (S&T).
  2. Expanding the Biolink Model to better describe chemical fate, exposure events, and biomarkers within environmental contexts.
  3. Establishing a functional workflow using a case study as a test system.



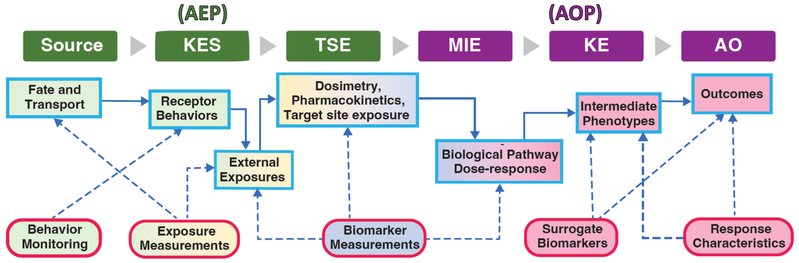

## Approach

### Source-to-Outcome Continuum




**Aggregate Exposure Pathway (AEP)**

**Adverse Outcome Pathway (AOP)**



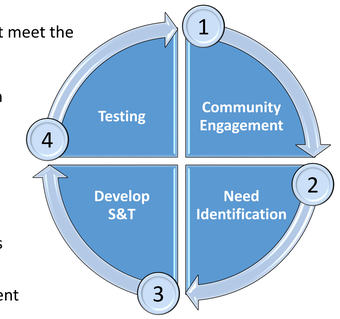

Source to Outcome continuum and subdomains colored by phases of the work. Boxes in light green reflect subdomains covered in Source-Exposure phase; yellow reflect subdomains covered in the Exposure-Dosimetry phase; blue reflect subdomains covered in the Biomarkers-Pathways phase; and pink reflect subdomains covered in the Phenotypes-Outcomes phase. Boxes with multiple colors indicate they are covered by adjacent phases. KES= key exposure state, TES= target site exposure, MIE= molecular initiating event, KE= key event, AO= adverse outcome.

- Groups of overlapping EHS subdomains will be examined in phases; initial work has begun with the exposure and dosimetry subdomains
- For each subdomain, data and metadata S&T are evaluated for gaps and extended to:
  1. ensure that data collected are interoperable across other subdomains
  2. promote the capture of metadata that are relevant for S2O analyses
  3. facilitate machine readability and broader applications
  4. support integrated quantitative modeling of exposure, effect, and outcome.
- Developed S&T will be tested in a functional use case, which will expand across subdomains throughout the work

| Subdomain grouping map |
|---|
| External exposure, dosimetry |
| Dose-response, pathway perturbations |
| Phenotypes, adverse outcomes |
| Exposure pathways, external exposure |

## Work Cycle

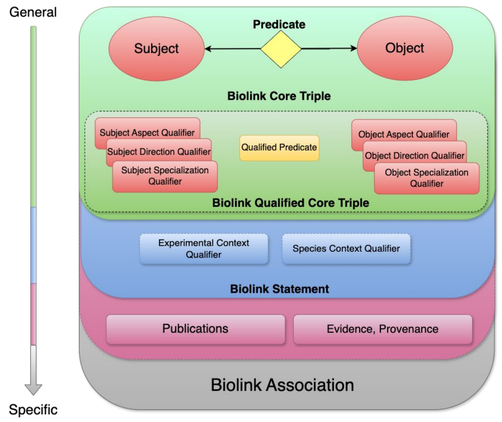
- At each phase, we deploy a work cycle to guide the development of S&T that meet the needs of community users




1) We engage community members through workgroups (WGs) focused on identifying gaps and needs in the standards for focal subdomains
   - WGs are provided with a state-of-the-science summary
   - The first WG will convene in Spring 2025. If interested in learning more contact DataStandards@rt.org.
2) Identify community needs from WG feedback
3) Develop S&T frameworks within the **Biolink Model** for focal subdomains
4) Test the interoperability of developed S&T in a quantitative use case
   - Use data gaps identified in the use case to inform further development

## Semantic development

- The **Biolink Model** is an open-source standard for harmonization across biomedical knowledge graphs that includes terms and relationships useful for computationally representing mechanistic pathways.




**Biolink Core Triple:** includes knowledge statement grounded with biomedical ontology types and terms

***"hexachlorobenzene affects CDKN2a"***

**Biolink Qualified Core Triple:** includes core subject, predicate, and objects, plus their qualifiers

***"hexachlorobenzene causes increased methylation of a mutant form of CDKN2a"***

**Biolink Statement:** includes fully qualified Biolink core triple and qualifiers that refine the entire knowledge statement.

***"hexachlorobenzene causes increased methylation in mutant form of the CDKN2a promoter in HeLa cells"***

**Biolink Association:** includes fully qualified Biolink statement and evidence, provenance, and confidence assertions about the statement.

***"hexachlorobenzene causes increased methylation in mutant form of the CDKN2a promoter in HeLa cells from Taylor, S, et al. (2001)"***

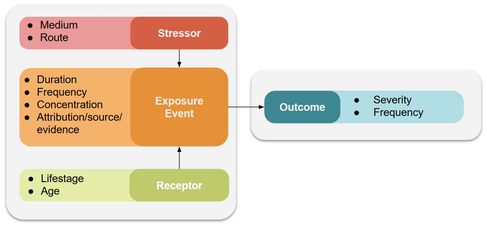
- **Biolink Model** provides a standardized schema and semantic model for developing tools to integrate biological knowledge across diverse resources.
- Organized as a class and attribute hierarchy with mappings to external vocabularies (e.g., OBO, GO, UMLS).
- Built using LinkML, a flexible schema modeling language that supports JSON, YAML, RDF, and database implementations.

| Biolink Model Strengths |
|---|
| • Facilitates interoperability among biomedical databases |
| • Provides a structured framework for representing biological entities and their relationships |
| • Supports multiple serialization formats, enhancing data exchange |
| • Allows for the validation, harmonization, and distribution knowledge with robust provenance. |

## Example expansion:

**Goal:**

Use ontological modeling from Environmental Conditions, Treatments, and Exposures Ontology (ECTO) to describe Exposure Events in terms of their duration, frequency, type, medium, route, and outcomes in Biolink Model.




Semantic Model of an Exposure Event. This figure shows the basic structure of an exposure event according to the Exposure Ontology (ExO; Mattingly et al. 2012), which is the exposure model used in ECTO. Figure adapted from Chan et al. (2023).

**Tasks:**

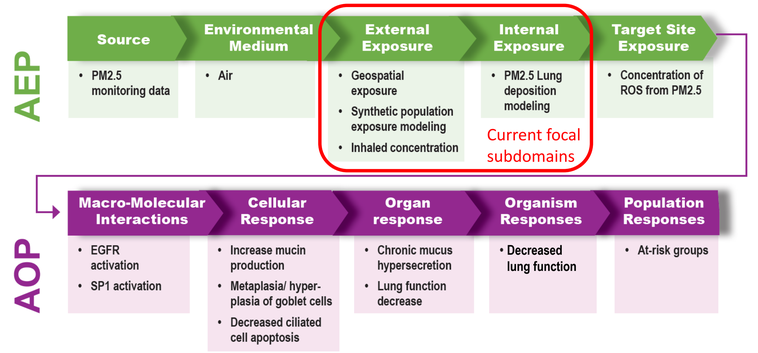
1. Map the current `biolink:ExposureEvent` class and descendent classes to ECTO
2. Capture necessary population and demographic information necessary to characterize exposure metadata
3. Bridge outcomes, phenotypes, and results in a set of harmonized Biolink categories and predicates useful in Biolink Exposure Event Associations

**Applicability in use case:**

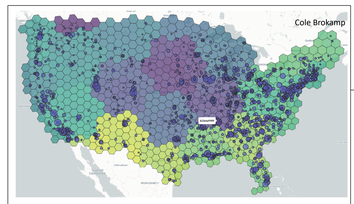
- Biolink ExposureEvent classes are modeled as edge attributes.
- Expanding Biolink to include the ECTO modeling approach that characterizes ExposureEvents as classes with duration, frequency, concentration, attribution as well as lifestage/age, medium/route will enable representation of PM2.5 exposure data

## Use Case

- We are constructing a quantitative **test system** centered around a use case for $PM_{2.5}$ and decreased lung function to evaluate functionality and Identify gaps in data standards



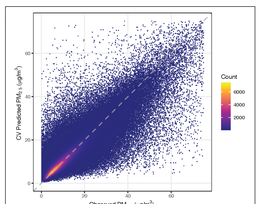

## Synthetic test population




- Example H3 grid for the U.S. used in the Degauss geomarker application to provide PM2.5 daily concentration predictions
- Blue circles represent the locations of EPA monitoring locations used in training the model
- Results are aggregated here to H3 level 3, while predictions are performed at H3 level 8 (see Wake County example below)

**Spatiotemporal model construction**

**PM2.5 Interpolation**




- A 2-D histogram of all cross validated model predictions and PM2.5 observations. Each hexagon bin is colored according to the number of points contained within it.

Cole Brokamp – Degauss and Related
https://doi.org/10.1016/j.envadv.2021.100155
https://github.com/geomarker-io/st_pm_hex?tab=readme-ov-file

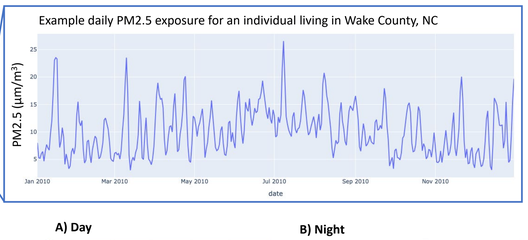
345,643 Households in Wake County, NC

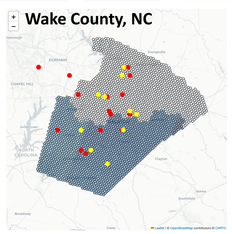
879,193 Persons

164 Schools

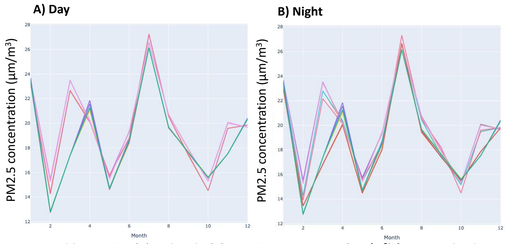
61,399 Workplace locations

EPA AQS data: 252,219,390 PM2.5 Records




Day Locations (Red) School and Workplace Night Locations (Yellow) for 10 Households




Monthly max Day (A) and Night (B) PM2.5 concentration (μm/m³) for 10 randomly selected synthetic households (23 individuals) in 2010

## Next Steps:

- Work with **subject matter experts** (SMEs) within subdomains to develop S&T that meet the current needs of the field.
- Continue expanding Biolink Model to capture exposure data
- Get involved! To learn more:

**Contact us at : DataStandards@rti.org**
**GitHub site : https://s2o-datastandards.github.io/**

**References:**

Brokamp, C. *A high resolution spatiotemporal fine particulate matter exposure assessment model for the contiguous United States.* (2022) Environmental Advances, 7.
Chan LE, Thessen AE, Duncan WD, Matentzoglu N, Schmitt C, Grondin CJ, Vasilevsky N, McMurry JA, Robinson PN, Mungall CJ, Haendel MA. *The Environmental Conditions, Treatments, and Exposures Ontology (ECTO): connecting toxicology and exposure to human health and beyond.* (2023) J Biomed Semantics, 14(1):3.
Mattingly, CJ, McKone TE, Callahan MA, Blake JA, Cohen Hubal EA. *Providing the missing link: the exposure science ontology ExO*. (2012): 3046-3053.



This work is funded by the National Institute of Environmental Health Sciences under award number 1R24ES036127

For more Information, contact David Hines at dhines@rti.org