The Environmental Health Language Collaborative:

Using Harmonized Language to Address Environmental Health Challenges

AGENDA

The Value of Language and Community
Stephanie Holmgren (NIEHS)

What Data Exists for a Given Chemical/Endpoint/Exposure Scenario?
Michelle Angrish (EPA)

Bridging Exposure and Biomarkers of Exposure
Stephen Edwards (RTI International)

Q&A/Discussion

The Environmental Health Language Collaborative

Harmonizing data. Connecting knowledge. Improving health.

The Value of Language and Community

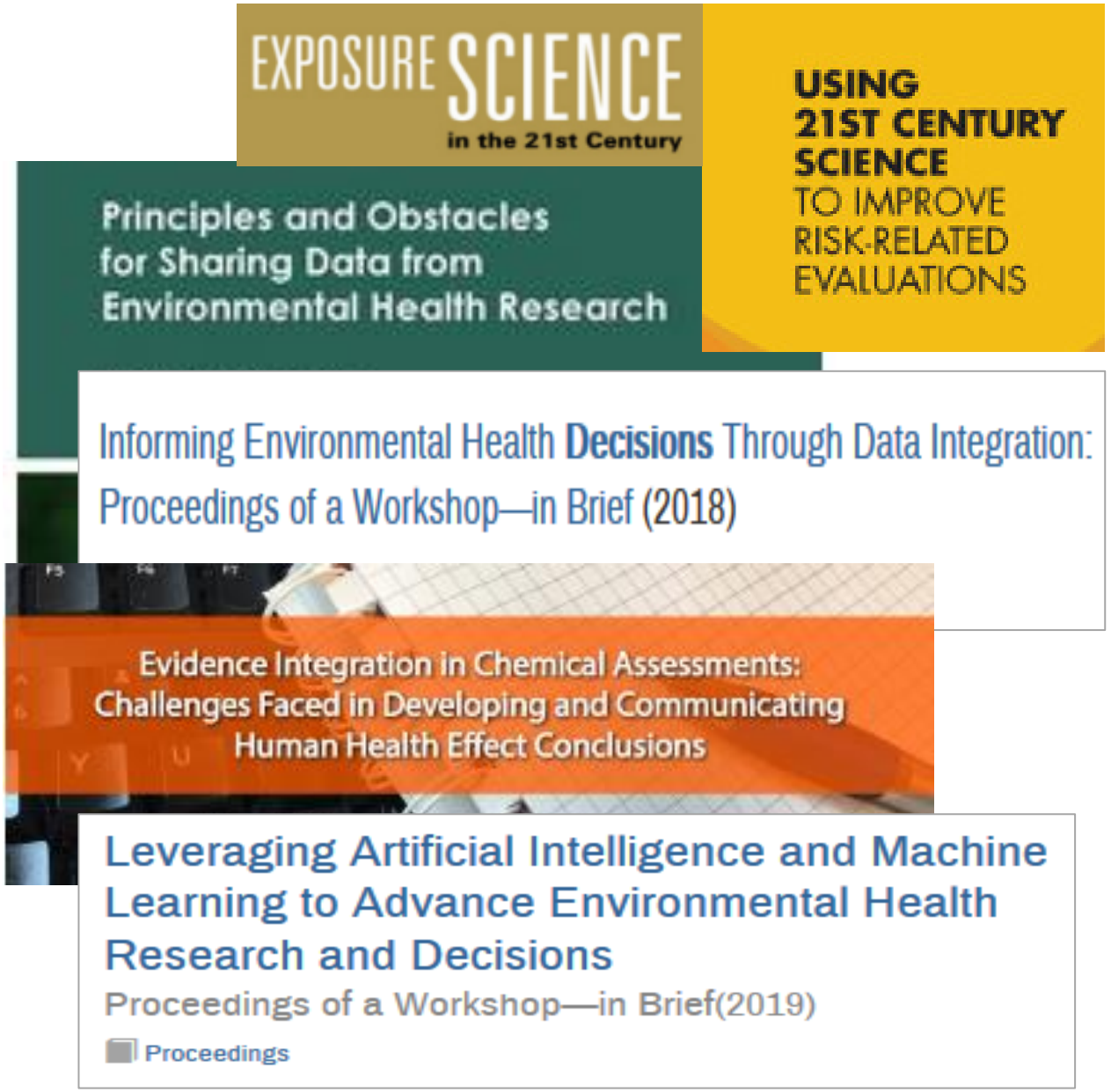
Stephanie Holmgren, NIEHS

Outline

- The Value of Harmonized Language
  - Semantically Speaking
- The Value of Community
  - EHLC – Building a sustainable community
  - EHLC – Developing semantic solutions
- September Workshop

The Value of Language

Collective recognition that the lack of harmonized language for describing environmental health data, findings, and knowledge has been a barrier for research and policy decisions



Contributor – Diverse Data Types

Contributor – Diverse Perspectives

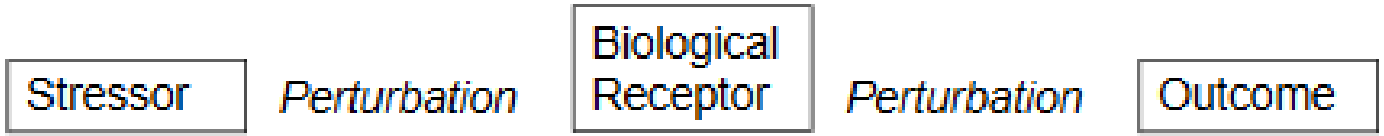
What biological processes are involved in observed changes in endpoints?

What is my biggest exposure risk based on my geographical location or occupation?

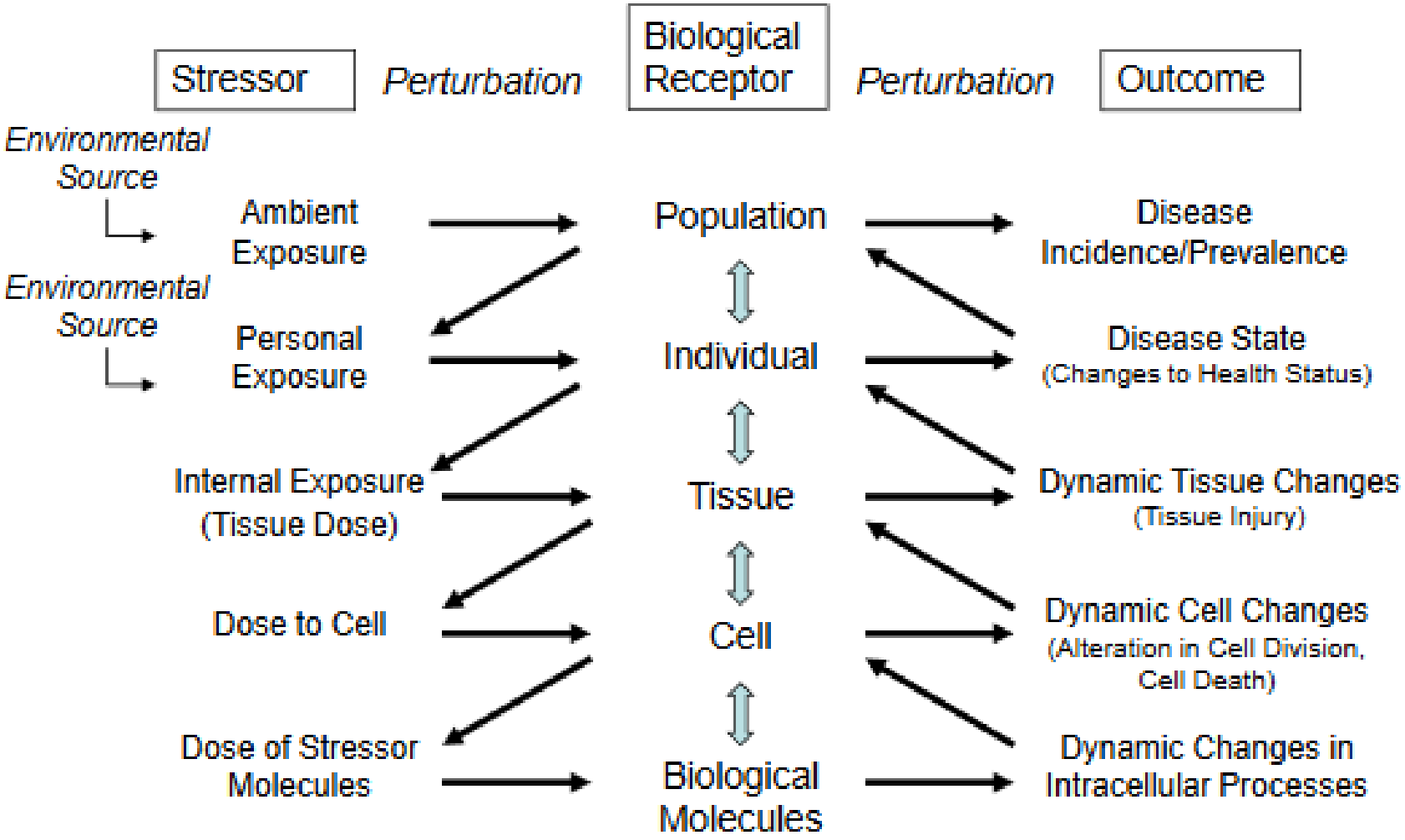
For what components of X industrial emission do we need more information on health outcomes?

What are the health and economic benefits from regulations or policies that reduce exposure to X?

The Complexities of Documenting Exposures



The Complexities of Documenting Exposures



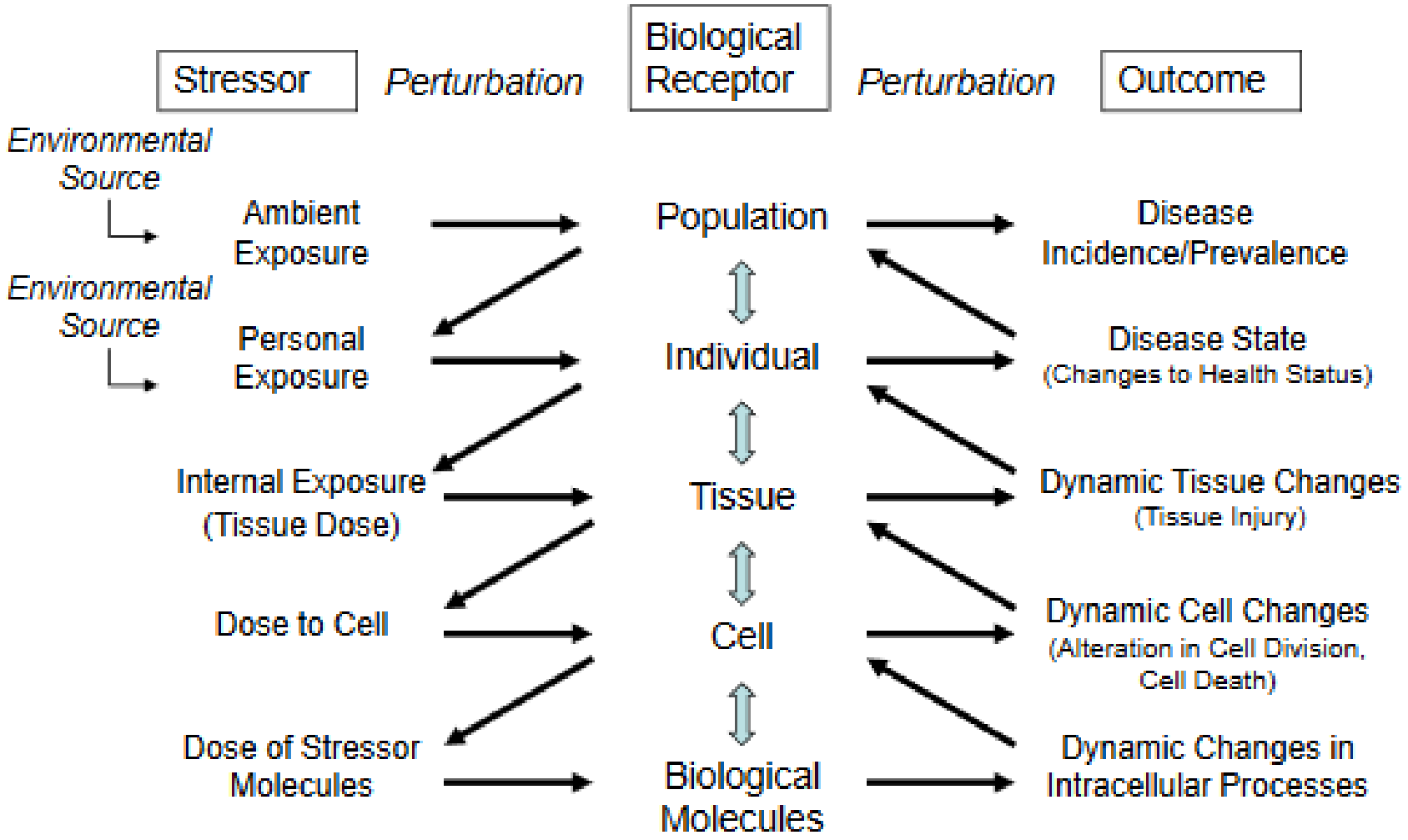
Based on Cohen Hubal (2010), JESEE 20(3): 231-6.

The Complexities of Documenting Exposures

Medium of Exposure
Place of Exposure
Route of Exposure

Geospatial
Sociodemographic
Behavioral

Timing – Dose
Timing – Life Stage



Based on Cohen Hubal (2010), JESEE 20(3): 231-6.

Challenges = Opportunities

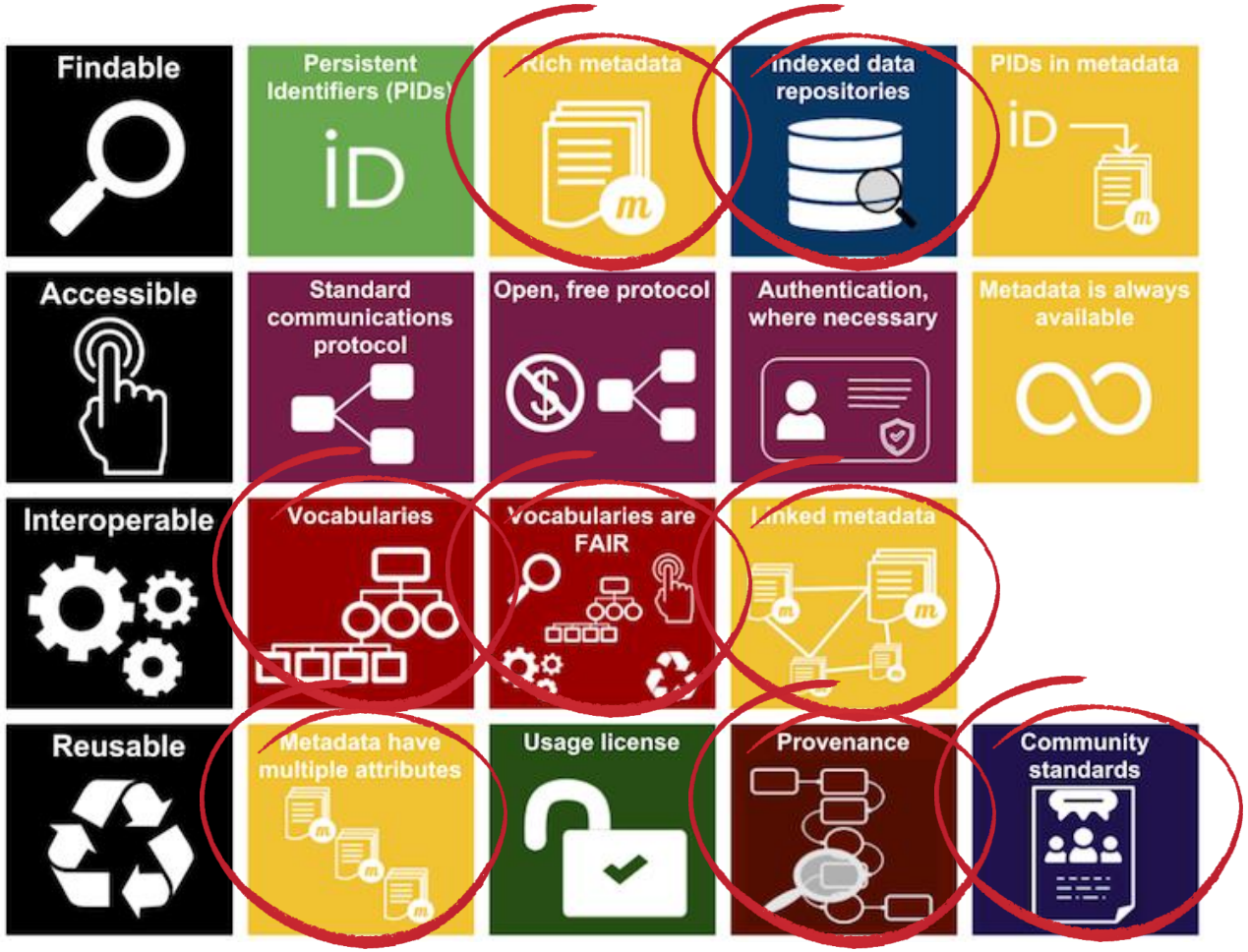
- Researchers in describing and comparing findings
- Data managers in organizing and representing data
- Data wranglers in finding and integrating data for analysis
- Model developers in using reference data collections
- Knowledge graph developers in linking data
- Tool developers in making scientific applications
- Informaticians seeking to automate literature processing and extraction techniques

Final NIH Policy for Data Management and Sharing

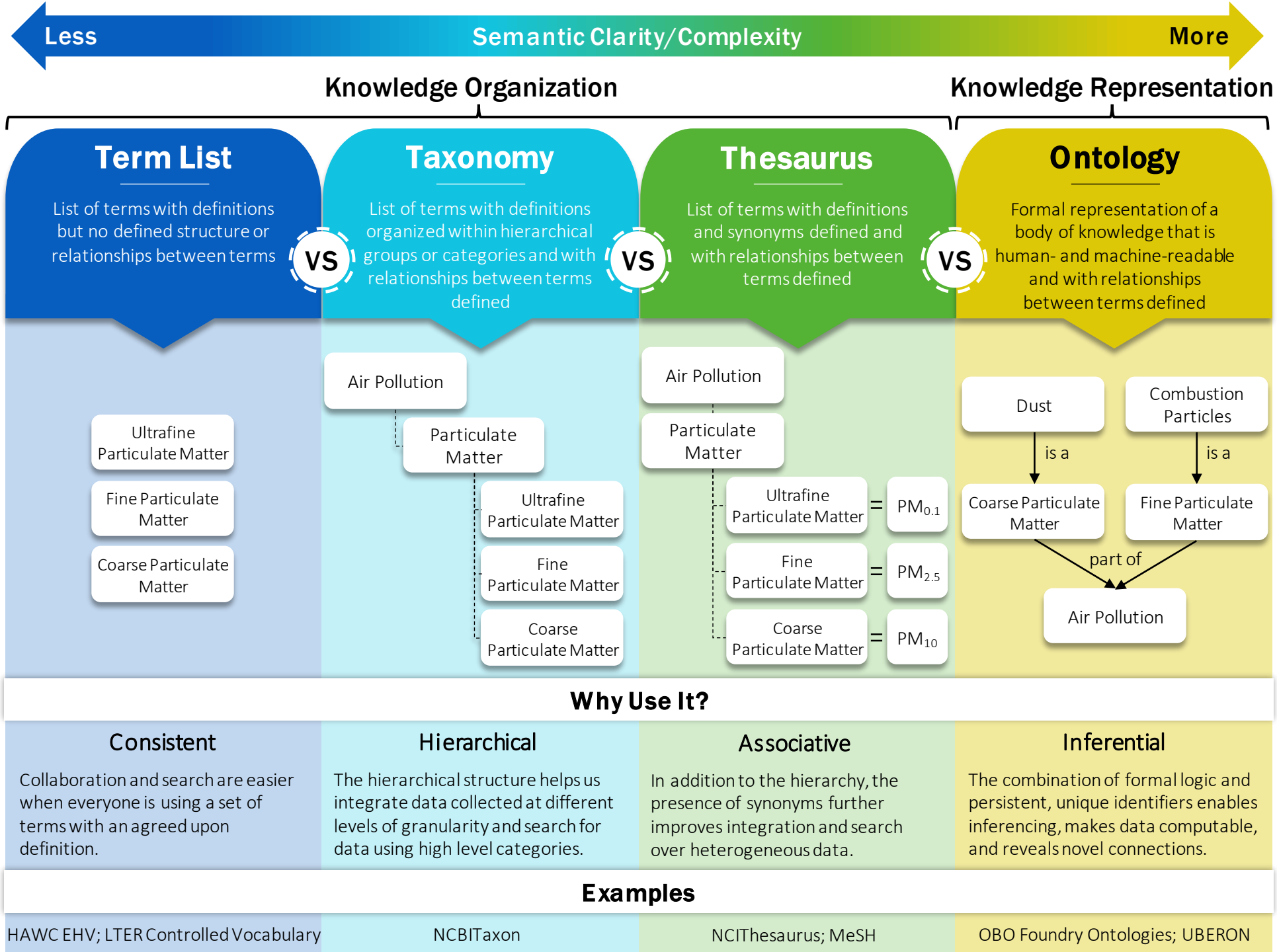
(NOT-OD-21-013)

Release Date: **October 29, 2020** | Effective Date: **January 25, 2023**

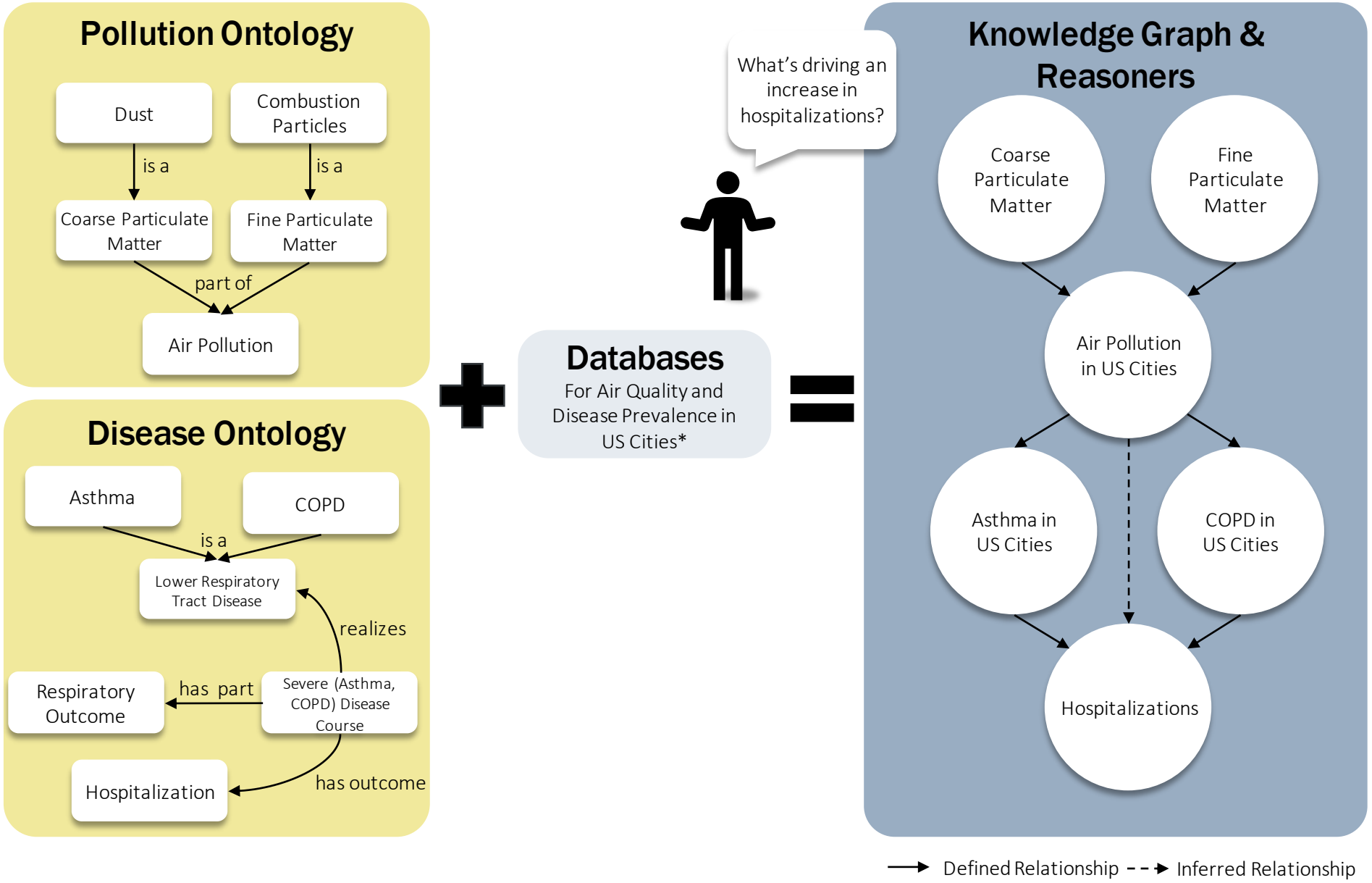
NIH requires researchers to prospectively plan for how scientific data will be preserved and shared through submission of a Data Management and Sharing Plan



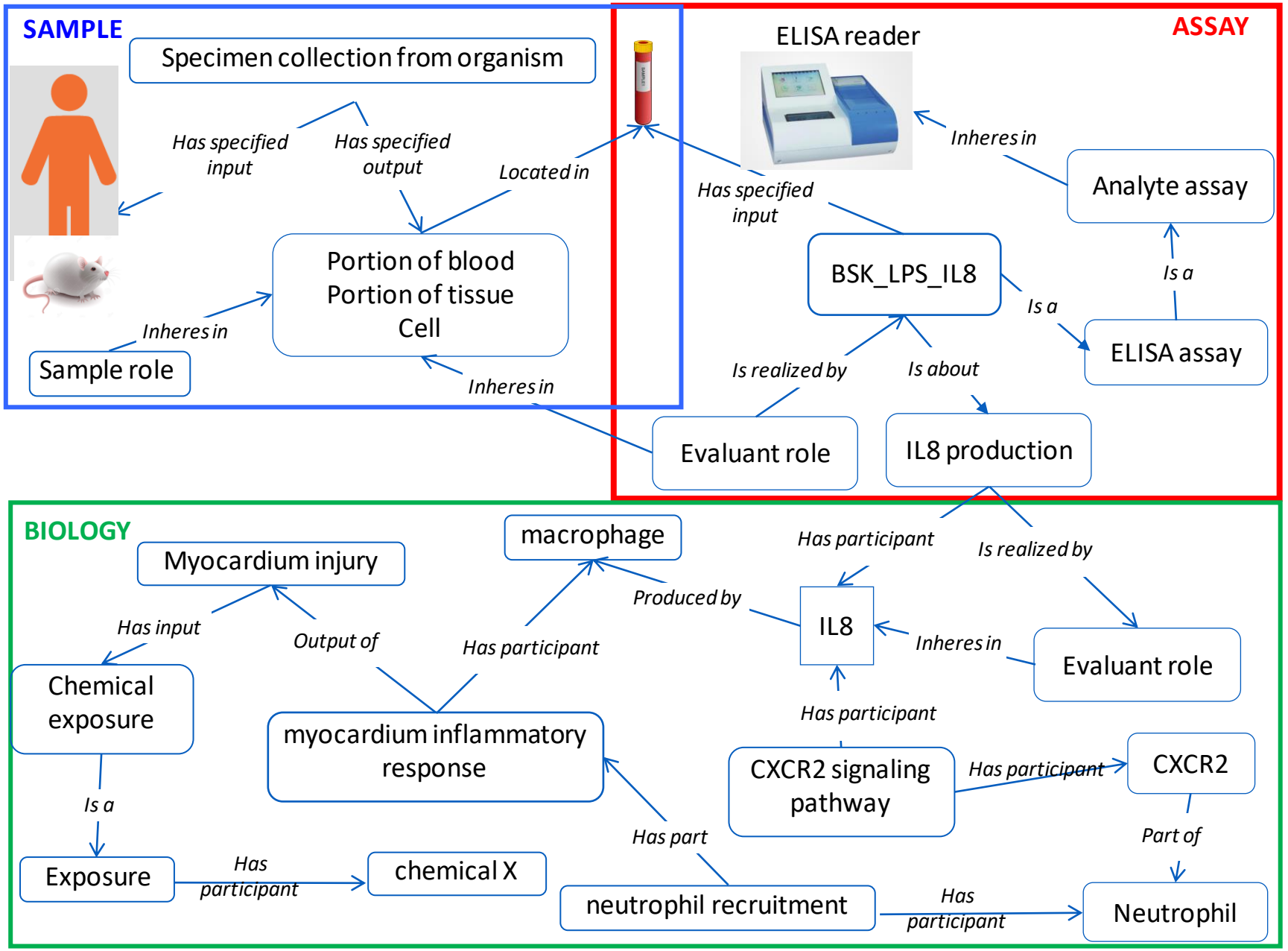
Semantically Speaking


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
Ultrafine Particulate Matter
Fine Particulate Matter
Coarse Particulate Matter
Air Pollution
Particulate Matter
Ultrafine Particulate Matter
Fine Particulate Matter
Coarse Particulate Matter
Air Pollution
Particulate Matter
Ultrafine Particulate Matter = PM0.1
Fine Particulate Matter = PM2.5
Coarse Particulate Matter = PM10
Dust
Combustion Particles
is a
is a
Coarse Particulate Matter
Fine Particulate Matter
part of
Air Pollution
Why Use It?
Consistent
Collaboration and search are easier when everyone is using a set of terms with an agreed upon definition.
Hierarchical
The hierarchical structure helps us integrate data collected at different levels of granularity and search for data using high level categories.
Associative
In addition to the hierarchy, the presence of synonyms further improves integration and search over heterogeneous data.
Inferential
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

Knowledge Representation



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



- Adding annotation makes it easier to find your data and use it in a purposeful way
- Builds connections between in vivo endpoints and in vitro tests aiding NAMs development
- By specifying the relationships between entities and roles they fill, interpreting the outcome of an assay becomes easier for non-domain experts
- Supports artificial intelligence approaches to finding and integrating information and knowledge developing NAMs

The Value of Community

Environmental Health Language Collaborative

Building a Sustainable Community

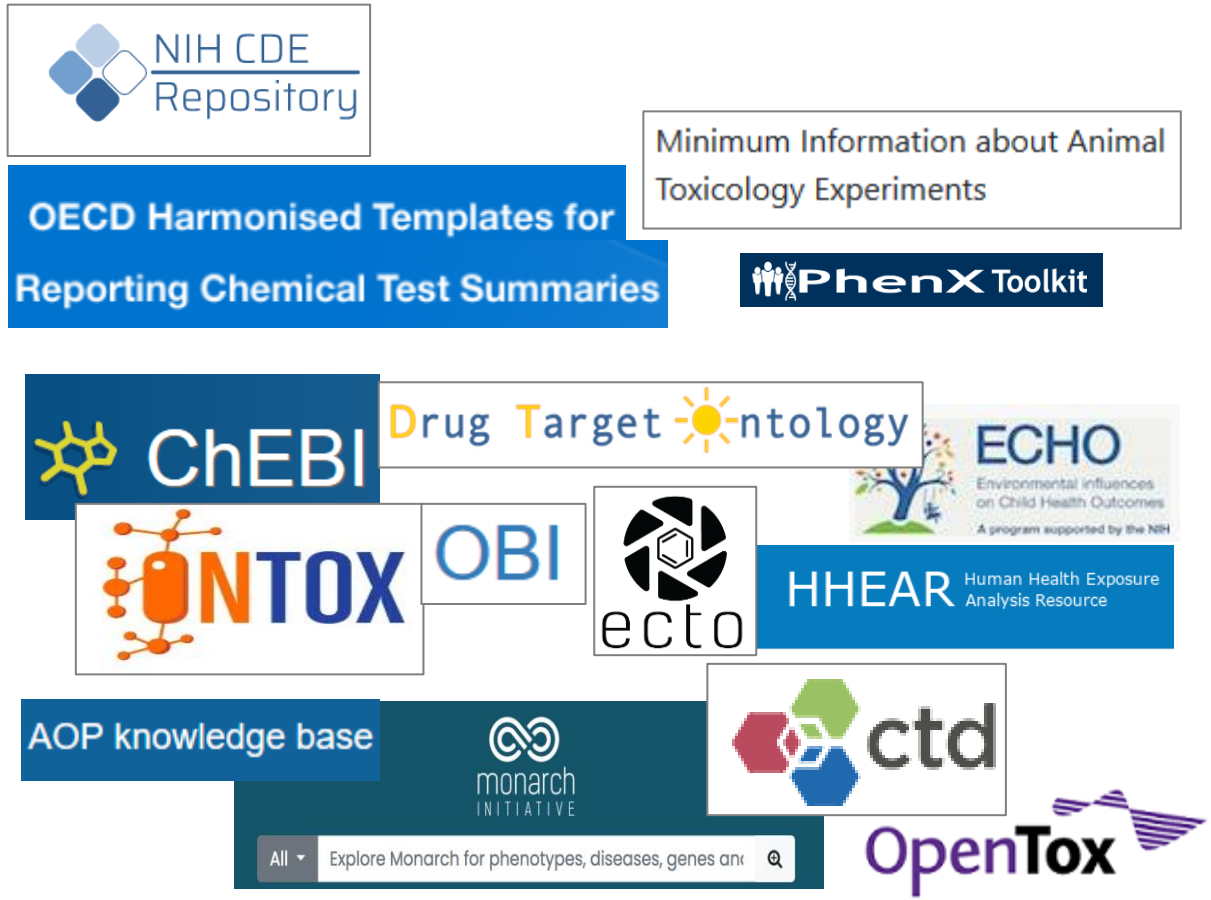
What is Community?

"A community is comprised of an **intentional collective** of people who gather and **"think together"** to address **common interests and goals**. A community commits to empowering its members to **govern** its operations, **guide its development**, and **achieve its purpose**."

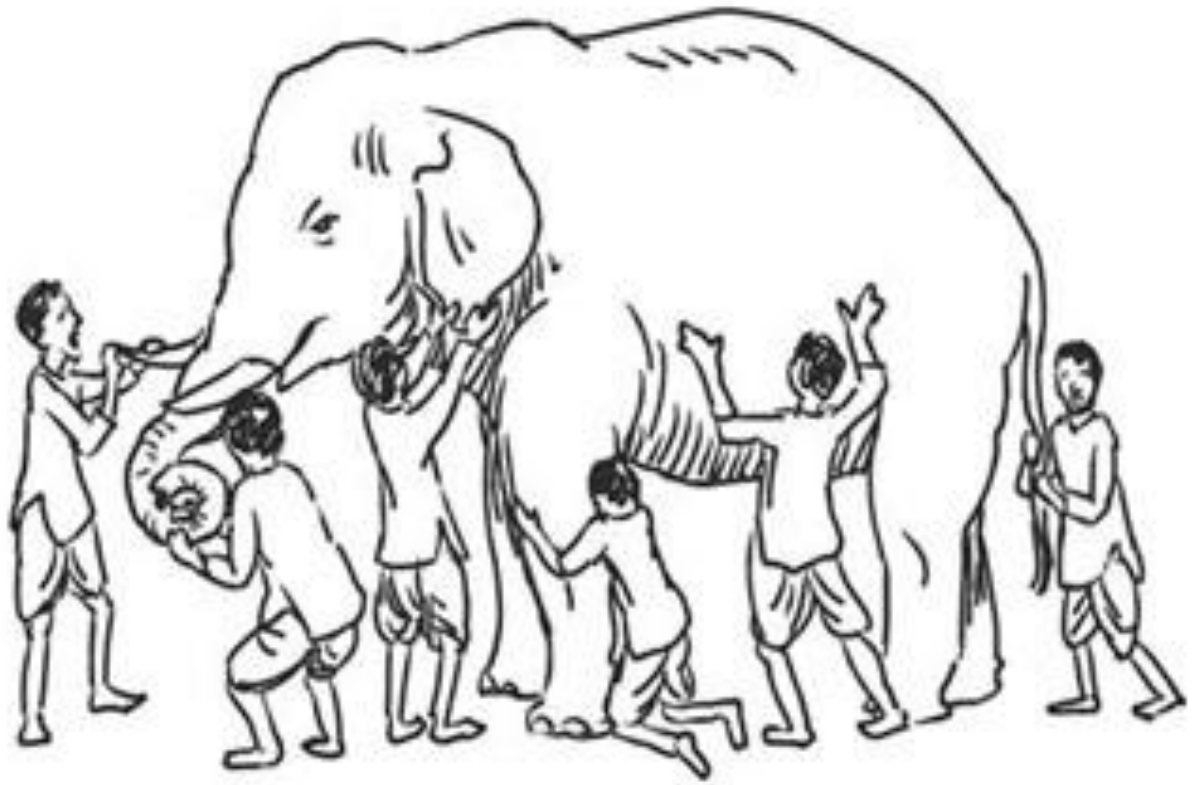
Sources: Educopia Field Guide and Pyrko et al (2017)

Building on the shoulders …

- Common Data Elements
- Minimal Information Standards
- Standard Vocabularies
- Ontologies
- Knowledge bases



Why start this effort?



Forum to:

- engage diverse perspectives
- raise awareness of efforts
- identify opportunities
- seek synergies
- represent EHS-needs
- pinpoint gaps

Environmental Health Language Collaborative

Vision

What do we aspire to achieve?

The vision of the Environmental Health Language Collaborative is to **leverage community-driven environmental health language standards to catalyze knowledge-driven discovery and improve public health.**

Mission

What is our fundamental purpose?

The mission of the Environmental Health Language Collaborative is to **advance integrative environmental health sciences research by developing and promoting adoption of a harmonized language**.

Goals

Develop Language-Based Solutions

Foster community-based extension and development of knowledge organization systems (KOS)

Promote and develop methods/tools for applying harmonized language in research

Implement Language-Based Solutions

Apply language standards and best practices for accurate environmental health data and knowledge representation

Advocate Value of Language

Cultivate a vocabulary-aware environmental health community

Roles

Forum to coordinate

- identifying use cases and needs
- prioritizing activities
- strategies and approaches for solutions

Platform for collaboration to develop semantic solutions to address identified needs

Community hub to

- identify and promote incentives and support adoption and use of semantic approaches
- identify and apply metrics to gauge success
- offer a resource clearinghouse

Community of practice to

- exchange information, ideas, expertise
- foster education and training

How will the Collaborative work?

Research Data Alliance

rd-alliance.org

Mission: to build the social and technical bridges to enable open sharing and re-use of data to accelerate data-driven innovation.

Goals:

- exchange knowledge and share discoveries
- discuss barriers and potential solutions
- explore and define policies, and
- harmonize standards to enhance/facilitate global data sharing, interoperability, and re-use.

Research Data Alliance

rd-alliance.org

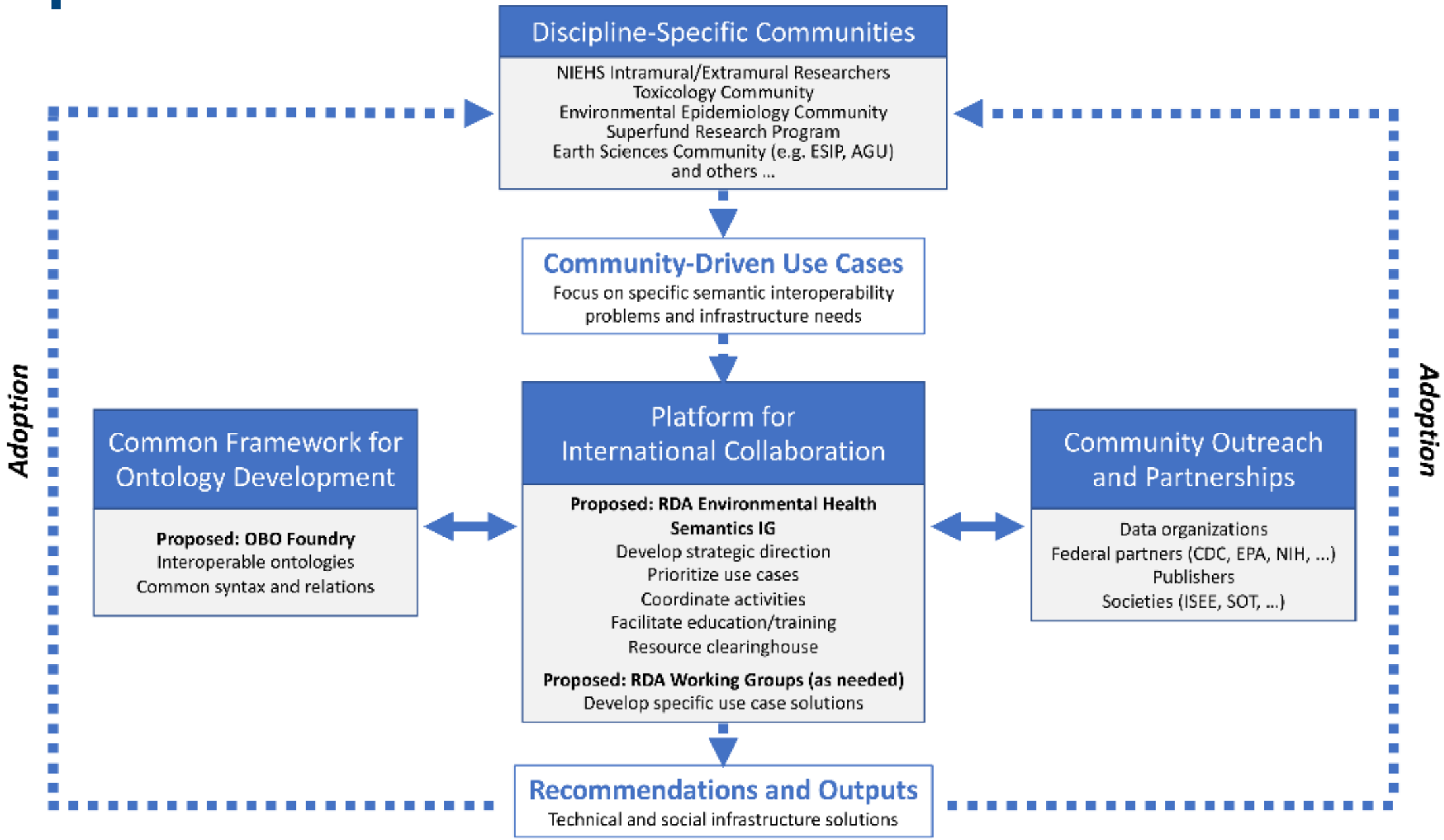
Membership: volunteer, community-driven, international initiative - individuals (11,445 members from 145 countries) and organizational and affiliate members (61)

Plenaries: meet every 6 months (April and November)

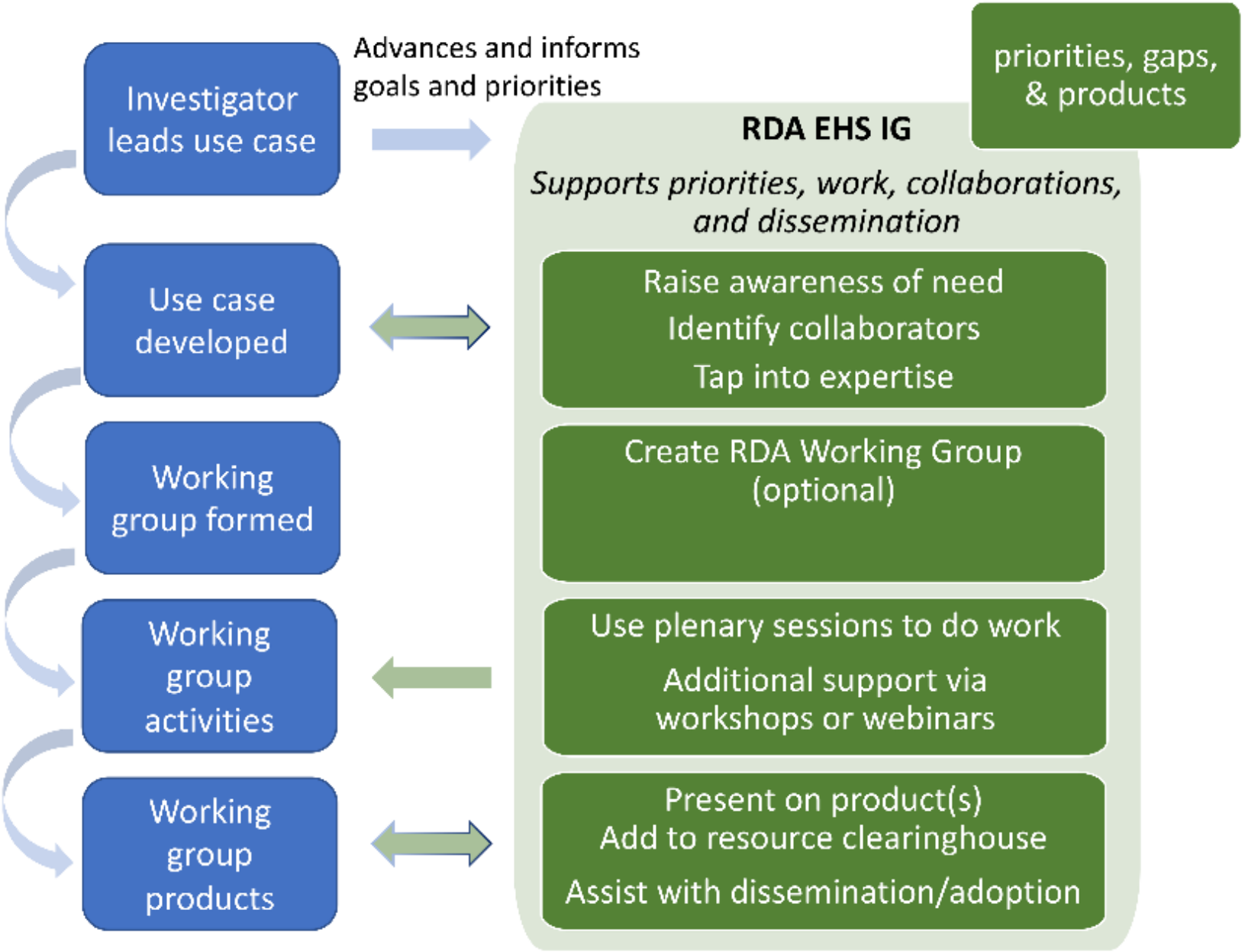
RDA 18th (virtual) Plenary Meeting
3-18 November, 2021

https://www.rd-alliance.org/plenaries/rda-18th-plenary-meeting-virtual

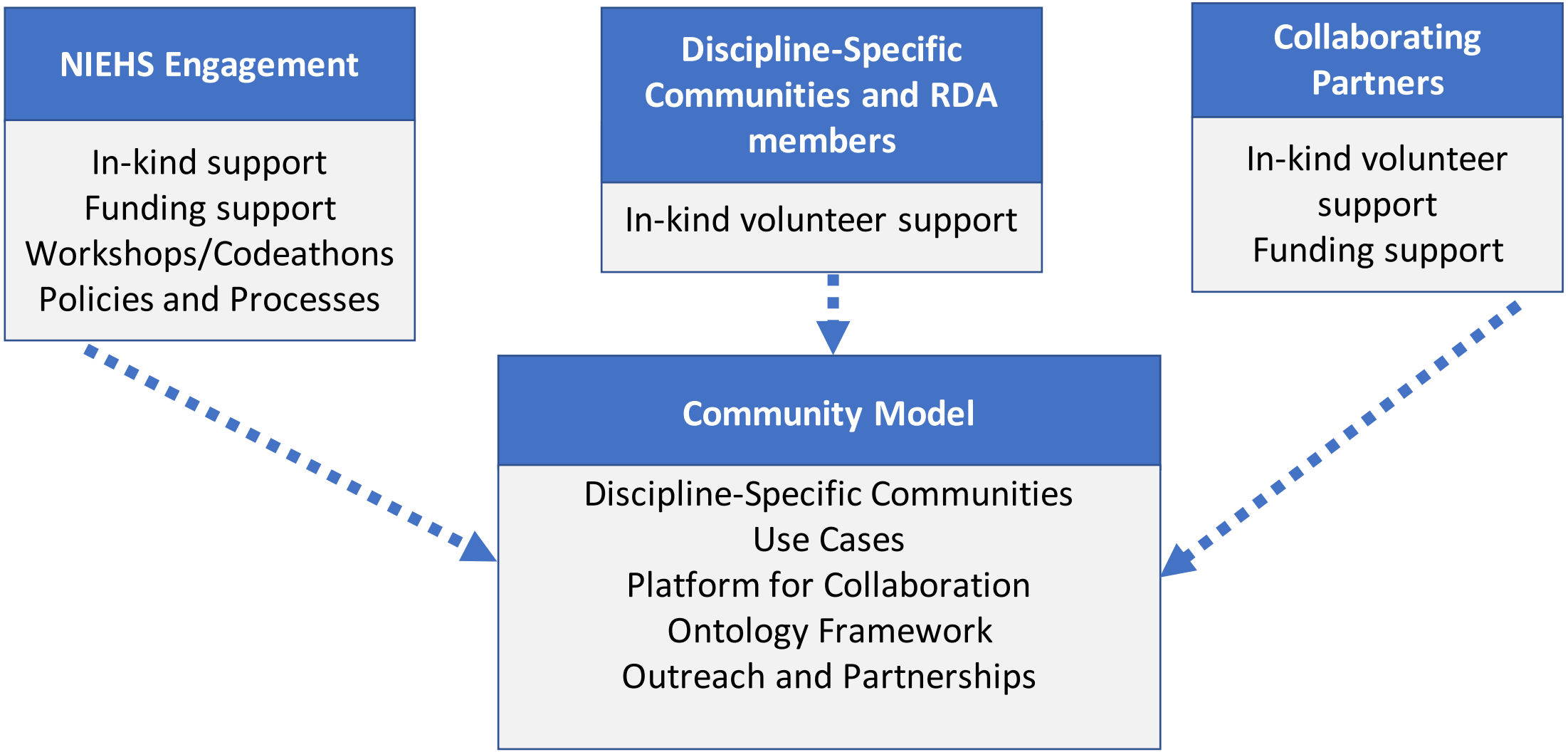
Proposed Model



Model in Practice



Sustaining the Community Model



Environmental Health Language Collaborative
Developing Semantic Solutions

Importance of Use Cases for the EHLC

Use Case Development

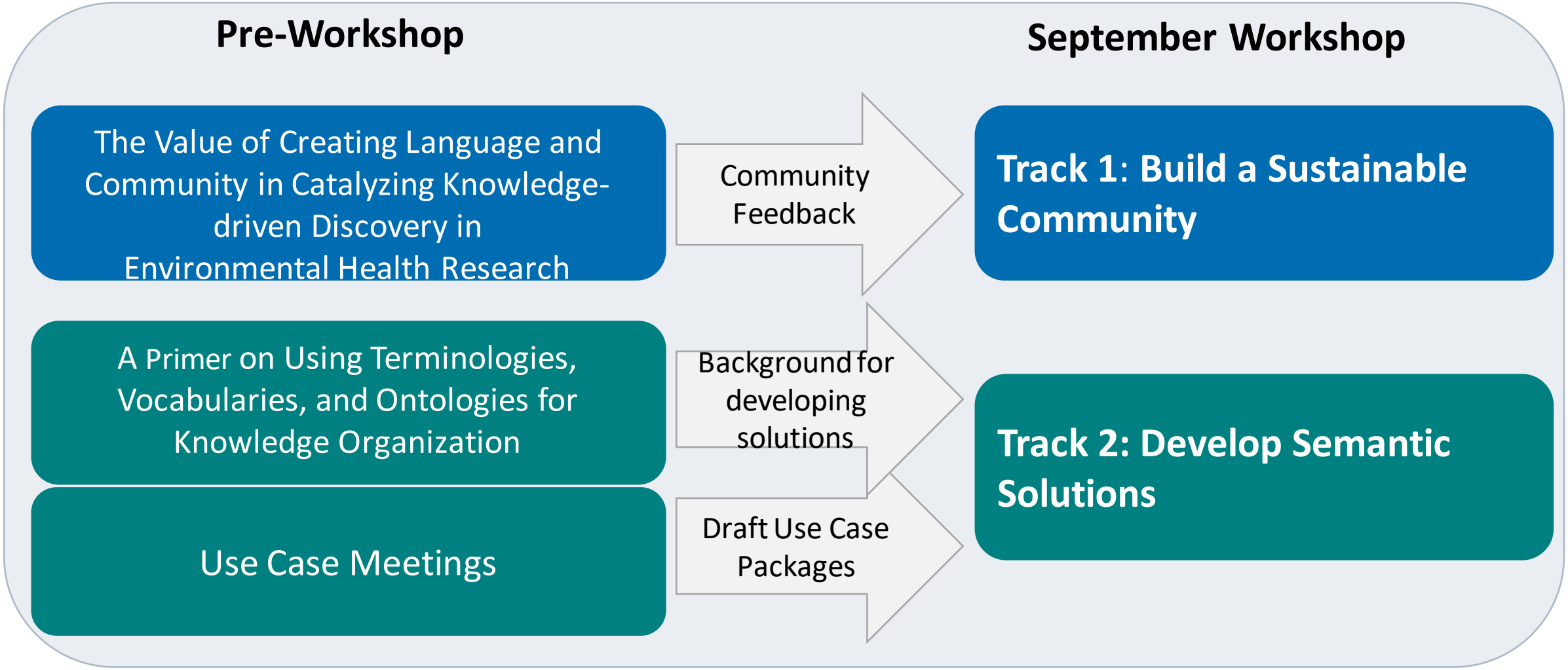
* https://www.niehs.nih.gov/research/programs/ehlc/use-cases/index.cfm

Current Use Cases

- What data exists for a given chemical/endpoint/exposure scenario? (**Michelle Angrish, EPA**)
- What are the biological processes and biomarkers associated with exposure and how do they relate to the potential for an adverse outcome associated with a given exposure
  (**Steve Edwards, RTI** and **Chirag Patel, Harvard**)
- Data and tools needed to harmonize place-based health research (**Carmen Marsit, Emory**)
- How do we combine individual-level data from multiple independent studies to understand how exposures X+Y impact health outcome Z?
  (**Jeanette Stingone, Columbia**)

September 9-10 Workshop

Catalyzing Knowledge-Driven Discovery in Environmental Health Sciences through a Harmonized Language



Workshop Goals and Outputs

Track 1: **Build a Sustainable Community**
Begin formation of a collaborative and cross-disciplinary community that will identify, develop, and champion the extension and use of language approaches within and across environmental health research.

- Achieve community agreement on the purpose and scope of the Collaborative as well as plan for how the Collaborative will work and define its success

Track 2: Develop Semantic Solutions
Define use cases in environmental health sciences research and begin identifying semantic needs, gaps, and next steps for implementing solutions.

- Make progress on the initial use cases and develop post-workshop action plans
- Begin compiling list of other use cases and needs

Sustaining the effort

Catalyzing Knowledge-Driven Discovery in Environmental Health Sciences Through a Harmonized Language

September 9 — 10, 2021
Virtual Workshop

Become Involved

- Attend the Pre-worksh
  Workshop Events

March 4, 2020

Journal article | Open Access

Computable Exposures Workshop Report

Thessen, Anne E; Grondin, Cynthia J; Kulkarni, Resham D; Brander, Susanne; Truong, Lisa; Vasilevsky, Nicole A; Callahan, Tiffany J; Chan, Lauren E; Westra, Brian; Willis, Mary; Rothenberg, Sarah E; Jarabek, Annie M; Burgoon, Lyle; Korrick, Susan A; Haendel, Melissa A

Informing Environmental Health Decisions Through Data Integration

Proceedings of a Workshop—in Brief (2018)

NIH National Institute of Environmental Health Sciences
Your Environment. Your Health.

Workshop for the Development of a Framework for an Environmental Health Science Language

September 15-16, 2014
North Carolina State University

Collaborative Next Steps

Build a Sustainable Community

- Refine vision, mission, goals, and roles
- Agree on community model governance
- CDISC Presentation

Build Semantic Solutions

- Use Case Working Groups
- Identify low-activation ideas - quick implementation, high impact
- Identify ontologies relevant to EHS

Become Involved

- Email Stephanie (Holmgren@niehs.nih.gov)
  - Volunteer to participate on a use case or topic working group
  - Submit ideas for use cases/semantic needs
- Join the EHLC email listserv - https://tinyurl.com/nfxp8ycf
- Learn more about the Collaborative at https://www.niehs.nih.gov/research/programs/ehlc/index.cfm
- **Spread the word!**

Acknowledgement

NIEHS

Shannon Bell

Gwen Collman

Chris Duncan

Jennifer Fostel

Richard Kwok

Ruth Lunn

Anna Maria Masci

Alison Motsinger-Reif

Charles Schmitt

Vickie Walker

Contract Support

Canden Byrd (ICF)

Ryan Cronk (ICF)

Courtney Lemeris (ICF)

Kim Osborn (ICF)

Jessica Wignall (ICF)

Rebecca Boyles (RTI)

Anne Thessen (CU Anschutz)

Community Advisors

Michelle Angrish (EPA)

Stephen Edwards (RTI) International

Carmen Marsit (Emory)

Rachel Morello-Frosch (UC Berkeley)

Chirag Patel (Harvard)

Jeanette Stingone (Columbia)

Robyn Tanguay (OSU)

Thank you!

Please feel free to reach out to me with questions or further discussion.

Stephanie Holmgren
NIEHS, Office of Data Science
Holmgren@niehs.nih.gov

The Environmental Health Language Collaborative

Harmonizing data. Connecting knowledge. Improving health.

What Data Exists for a Given Chemical/Endpoint/Exposure Scenario?

Dr. Michelle Angrish

The views and opinions expressed here do not reflect official US Environmental Protection Agency policy.

Challenge/purpose

- Understanding the health effects of environmental exposure requires finding and integrating relevant information
- Finding that information can be a challenge because one must
  1. know **where** to look and **how** to find it,
  2. have the resources to **collect, screen, and curate** the information, and
  3. assimilate that information so that it is **accessible and usable.**
- Such a workflow is further complicated because **study reports** are the typical form of information

Purpose is to develop solutions toward **identifying, connecting, and making use of environmental health science resources**

Final desired output

We will aim to develop tools and strategies to facilitate **interoperability of existing databases**.

Workshop goal

We will aim to identify and define concepts and features that are common across representative environmental health datasets that are needed to achieve resource interoperability.

Progress

Key points raised, gaps, and challenges

Defined use case question

What are the needs

Key points raised

1. Defining the **end goal** of data acquisition/solutions and needs will be **fit for purpose**
2. Understanding the players/roles and **their different** needs when designing tools/resources
3. Solutions will likely be a **blend of 20th century approaches** (standards, structures) with **modern techniques** (AI/NLP)
4. **Curation is critical** and still resource intensive

Gaps

1. Data producers and consumers **lack information** on sources, tools, and “best practices” limiting adoption
2. Lack of **structures to** require/encourage use of **standardized terminology** (e.g., requirements by publishers or funding agencies)
3. **Availability of data** in public space along with well-curated training data to support method development

Challenges

1. Sorting out the **subject domain-specific differences**
2. Encouraging use of **unique/specific identifiers** and appropriate metadata
3. Ensuring the **context needed** to use data is provided/identified in search
4. How do we **stop feeding the unstructured data problem**?

Next steps

1. Assembly of **resources, core trainings** that are available to stakeholders to support finding and creating structured data
2. Development of “standards” or tools to **support creation and sharing of structured data**

If interested in participating, email

Stephanie Holmgren, holmgre1@niehs.nih.gov and

Michelle Angrish, angrish.michelle@epa.gov

Thank you!

The Environmental Health Language Collaborative

Harmonizing data. Connecting knowledge. Improving health.

Bridging Exposure and Biomarkers of Exposure

Stephen Edwards, RTI

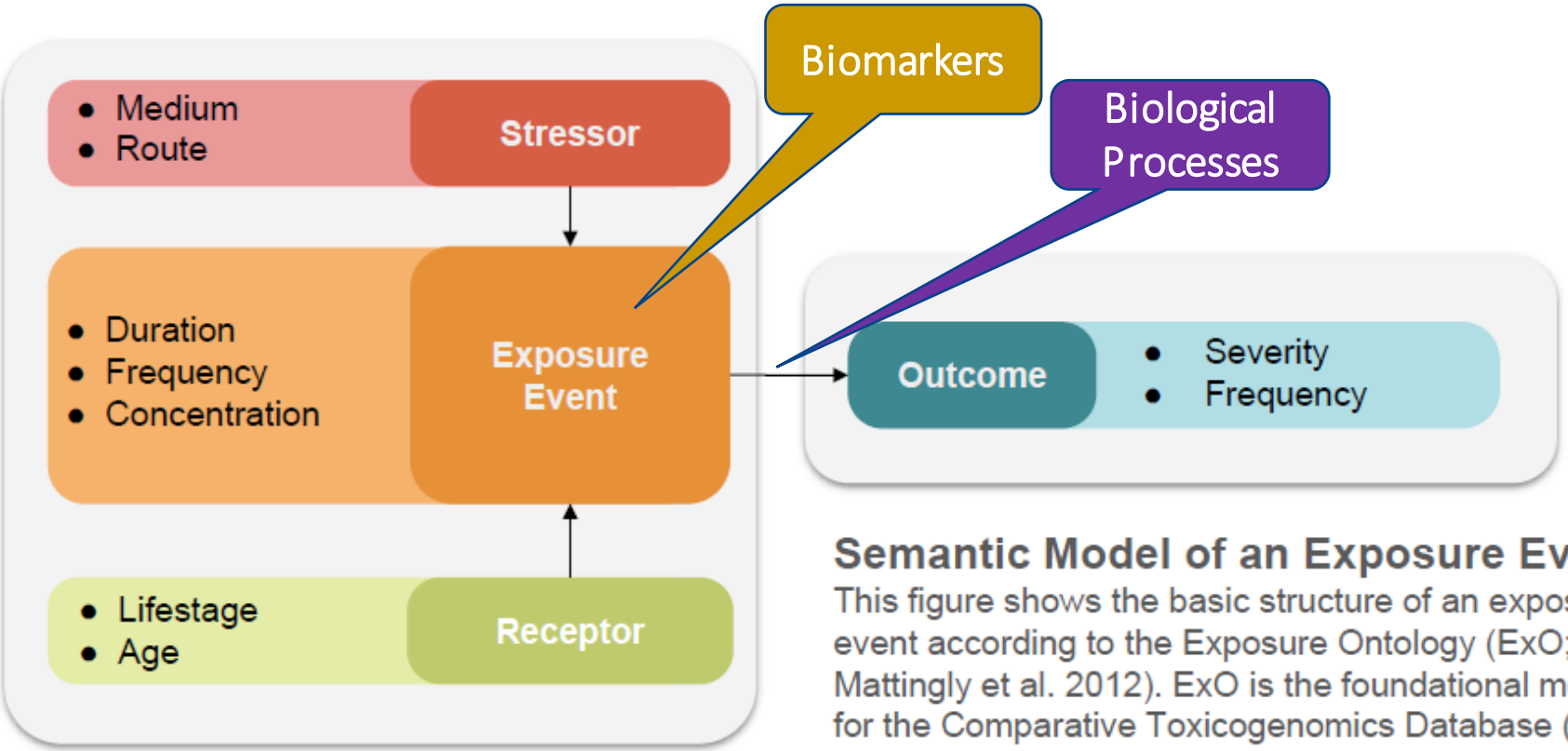
“What are the biological processes and biomarkers associated with exposure and how do they relate to the potential for an adverse outcome associated with a given exposure?”

Chirag Patel, Harvard
Stephen Edwards, RTI

Why are we exploring this use case?

- This use case is intended to build upon the other use cases and consider a more complex question
  - Will run in parallel with the other use cases but with a **longer timeline**
  - Will **utilize interim results** from the other use cases and **provide feedback** on their general utility
  - Will provide a 'Big Hairy Audacious Goal' for the initiative
    - Collins and Porras "Built to Last: Successful Habits of Visionary Companies" (1994)

Why are we exploring this use case?



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). ExO is the foundational model for the Comparative Toxicogenomics Database (CTD; Mattingly et al. 2006).

From preworkshop presentation by Anne Thessen

See Thessen et al. *Environmental Health Perspectives* 128:125002 (2020) https://doi.org/10.1289/EHP7215

Benefit of developing solutions around this use case

This use case provides a **longer-term horizon** to both **guide and expand upon the other use cases**

1. Provide additional context for the short-term use cases
2. Identify additional short-term use cases
3. Build upon results from the short-term use cases immediately

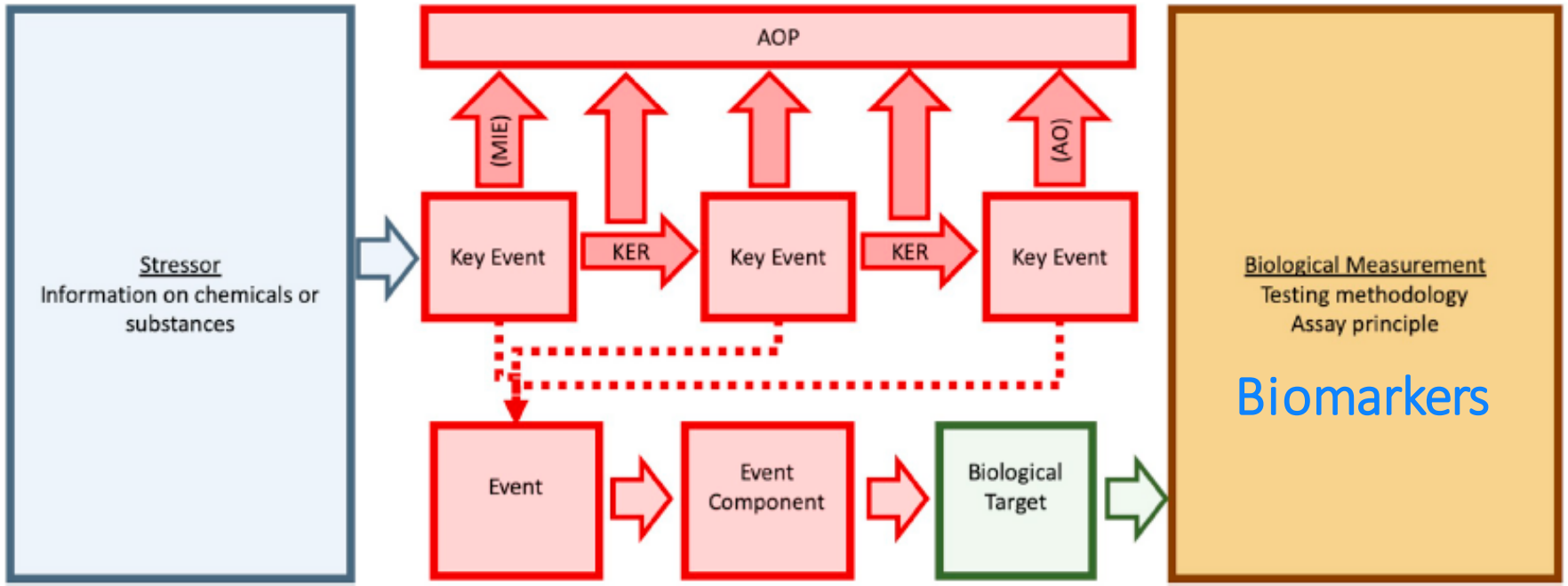
Workshop goal

Connect measured biomarkers to exposure-response relationships with a

- **semantic description of exposure events**
- that **incorporates** the associated **biomarkers** and **biological processes**
- to support the **integration** of existing data resources

Proposed approach to achieve workshop goal

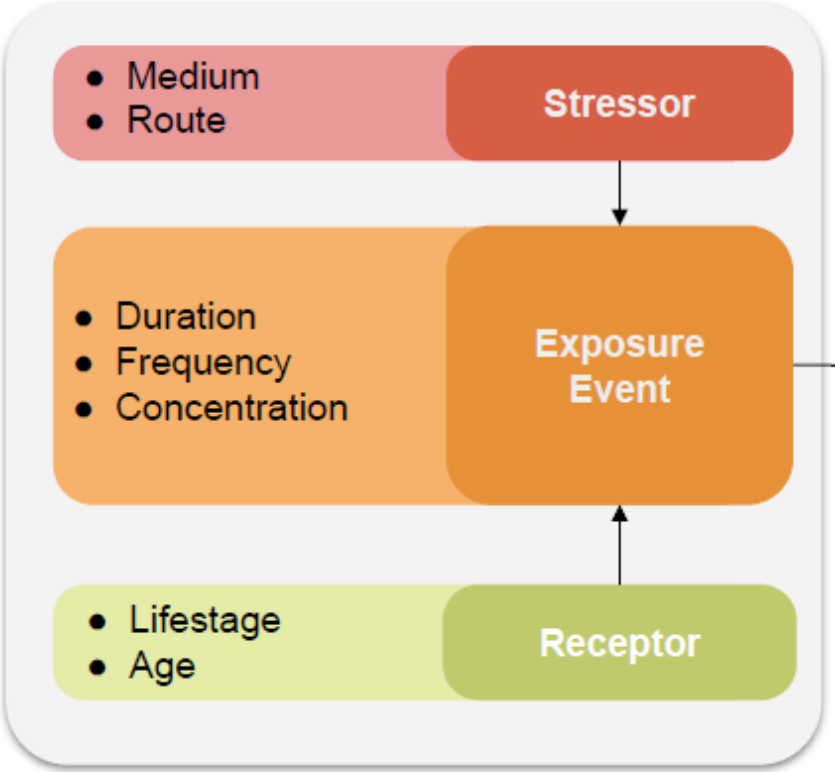
Extend the **semantic description of the exposure event** to explicitly include measurements as previously done for adverse outcome pathways



From Watford, et al. *Toxicology and Applied Pharmacology* 380:114707 (2019) https://doi.org/10.1016/j.taap.2019.114707

Proposed approach to achieve workshop goal

Extend the **semantic description of the exposure event** to explicitly include measurements as previously done for adverse outcome pathways



Biomarkers of Exposure

- **Parent**
- **Metabolites**
- **Chemical signatures**
- **Response signatures***

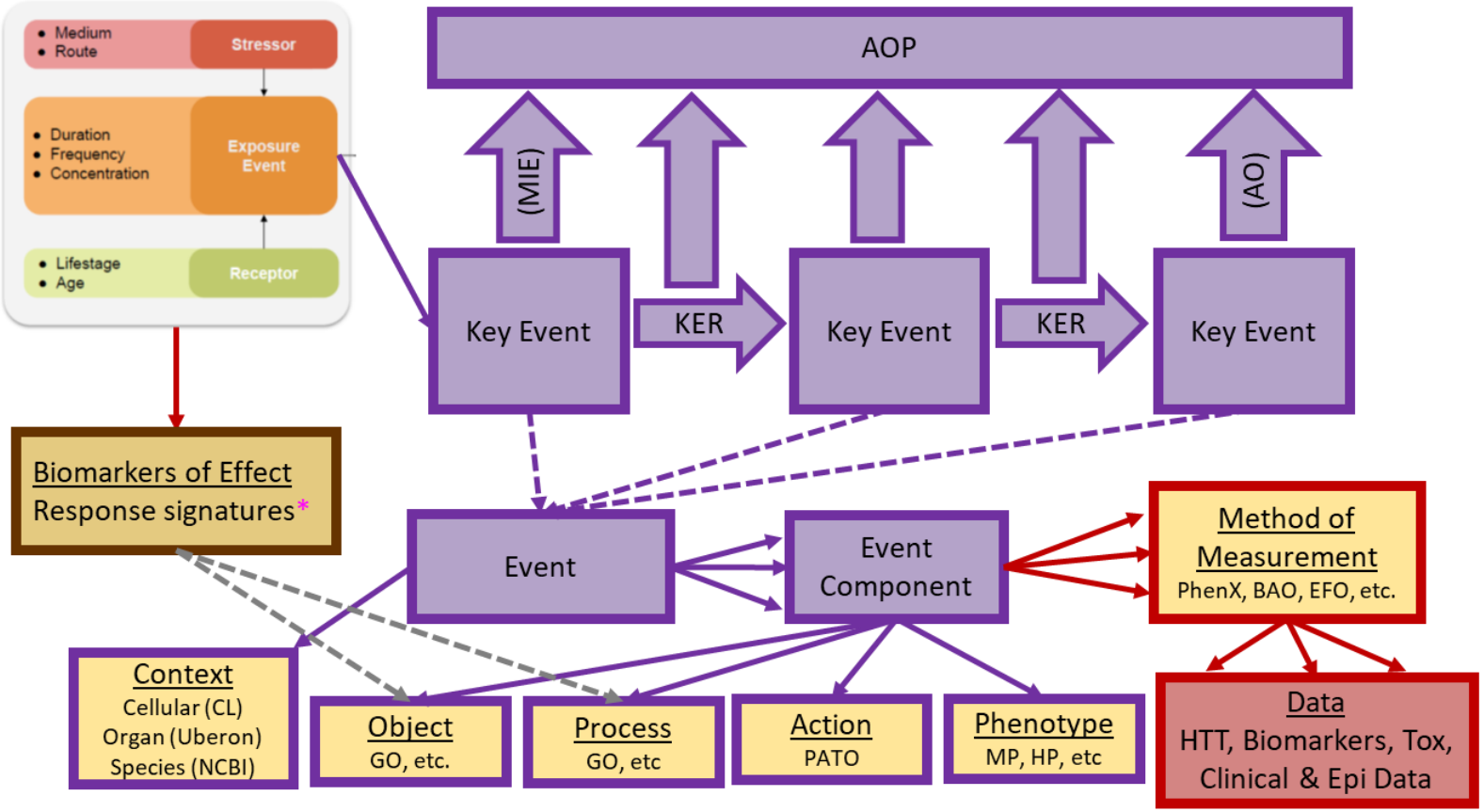
Considerations

- **Biological matrix**
- **Timing of exposure and measurement**
- **Pharmacokinetics**
- **Understanding of the biomarker**
- **Covariates impacting measurement (e.g. hydration)**
- **…**

* **Also biomarkers of effect**

Proposed approach to achieve workshop goal

Semantically link the exposure event to adverse outcomes by connecting the perturbed biological processes with toxicity mechanisms



Proposed approach to achieve workshop goal

Split into two breakout groups to consider both perspectives

1. Semantic description of the exposure event
   a) What information is needed to **interpret biomarker measurements**?
   b) How do we ensure that **measurements** can be **connected** back to databases containing information about **exposure potential**?
2. Semantically link the exposure event to adverse outcomes
   a) How to define the biological processes in terms that **connect to mechanisms** of disease such as AOPs?
   b) Can we **harmonize** different representations of mechanisms such as **AOPs, Causal Activity Models, and Monarch Phenotypes->Genotypes**

Example sub use cases

Breakout Group 1

1. What biomarkers are directly indicative of exposure to a given chemical? Biomarkers can include direct measurement of the chemical or its metabolites and can be identified associatively or experimentally through epidemiological or experimental approaches, respectively.

2. What are the exposures that are associated with the observed biomarkers in an epidemiological study? One may observationally or experimentally find biomarkers associated with health and disease – what are potential exposures that may also induce changes in the biomarkers?

Breakout Group 2

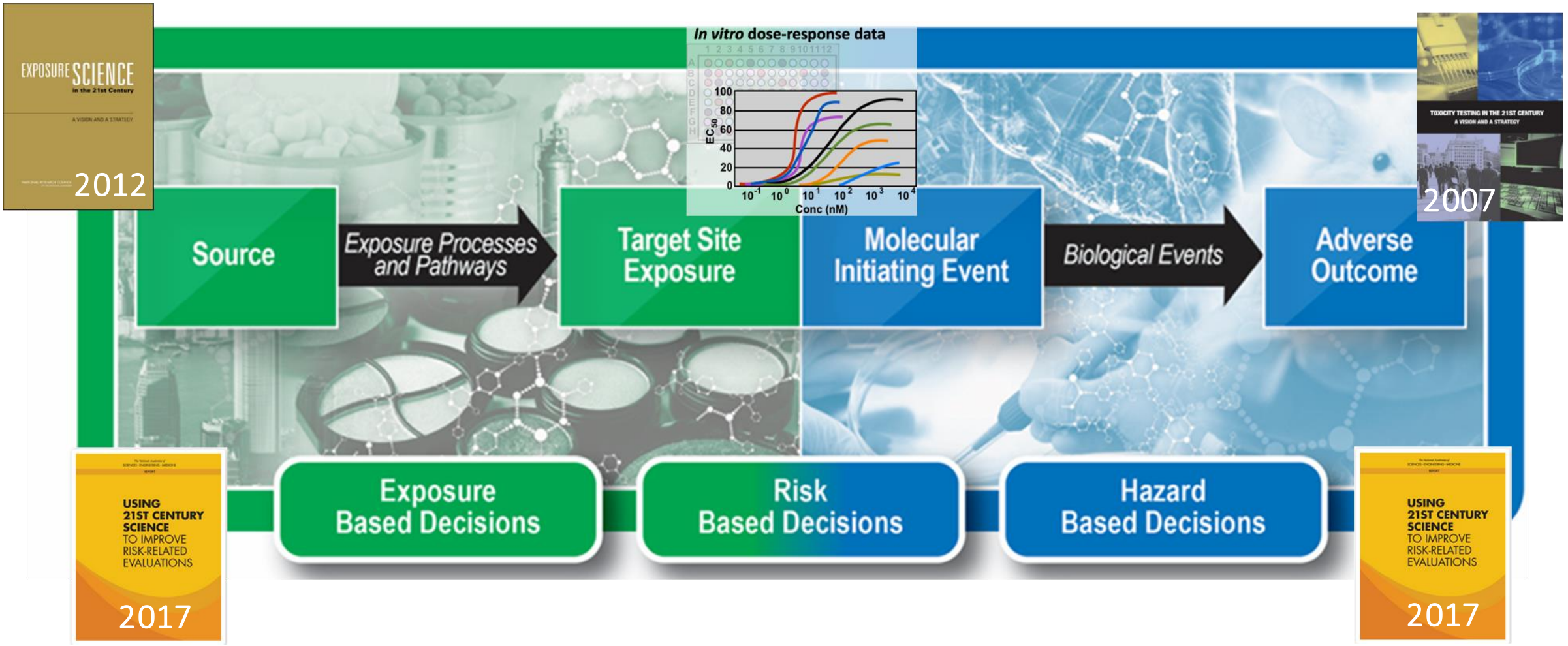
1. Map signatures of 'omic changes to chemical exposure: Query for organ-specific signatures of 'omic biomarkers, across the metabolome or the transcriptome, that are indirectly or directly associated with exposure.

2. What biological processes are linked to biomarkers that are indicative of the exposure? If an exposure is causal for a change in state, their biomarkers must also be directly or indirectly associated with biological processes. Given biomarkers that are indicative of exposure to a chemical or class of mechanistically related chemicals, query for all biological processes that are associated with changes in the biomarker(s).

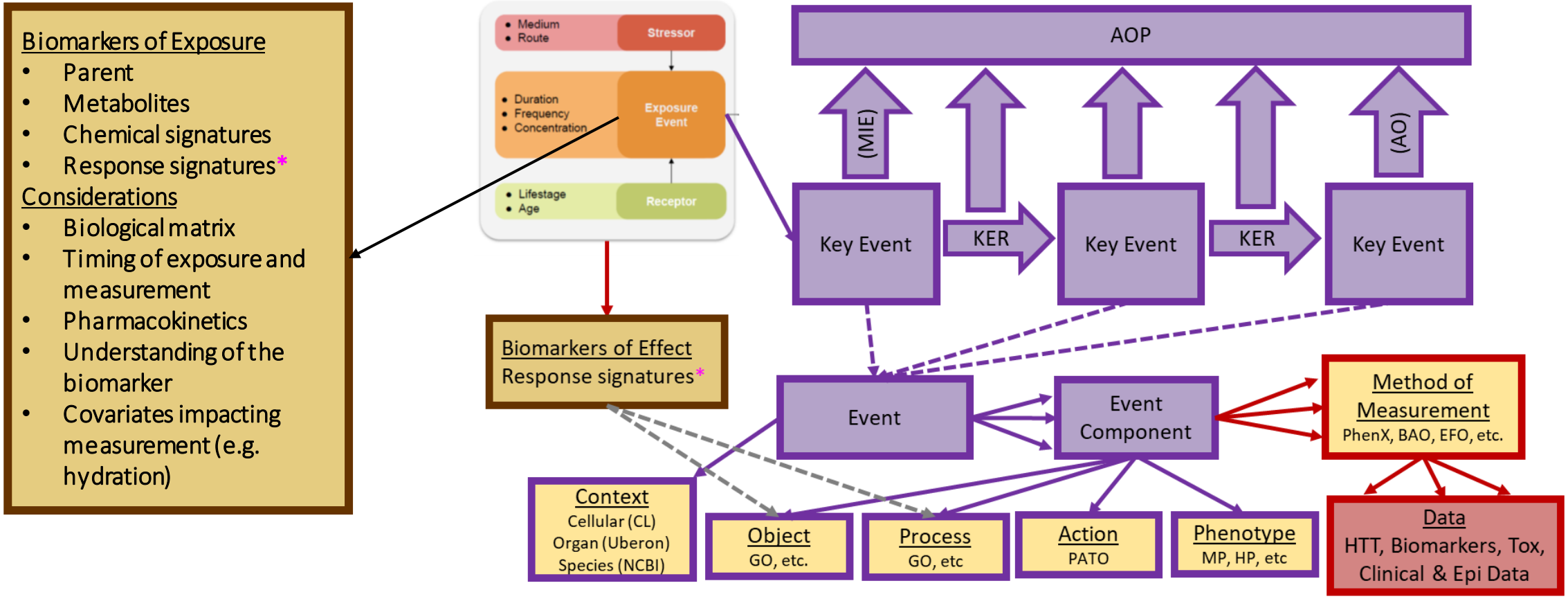
Key points raised

1. Is an “exposure event” the same as an AOP initiating event? No.
    a) Not all exposures result in adversity.
    b) One exposure can have multiple outcomes.
    c) One outcome can result from many exposures.
2. ‘Omics measurements hold great promise for connecting exposure events and the biological impacts of those events.
    a) Can fill gaps in our knowledge where targeted biomarkers are not yet available
    b) Our EHS language must be precise enough to guarantee these types of data are correctly interpreted
3. The ability to combine and query data across model organisms is very important.
    a) The AOP framework accommodates this extremely well

Exposure event vs. molecular initiating event

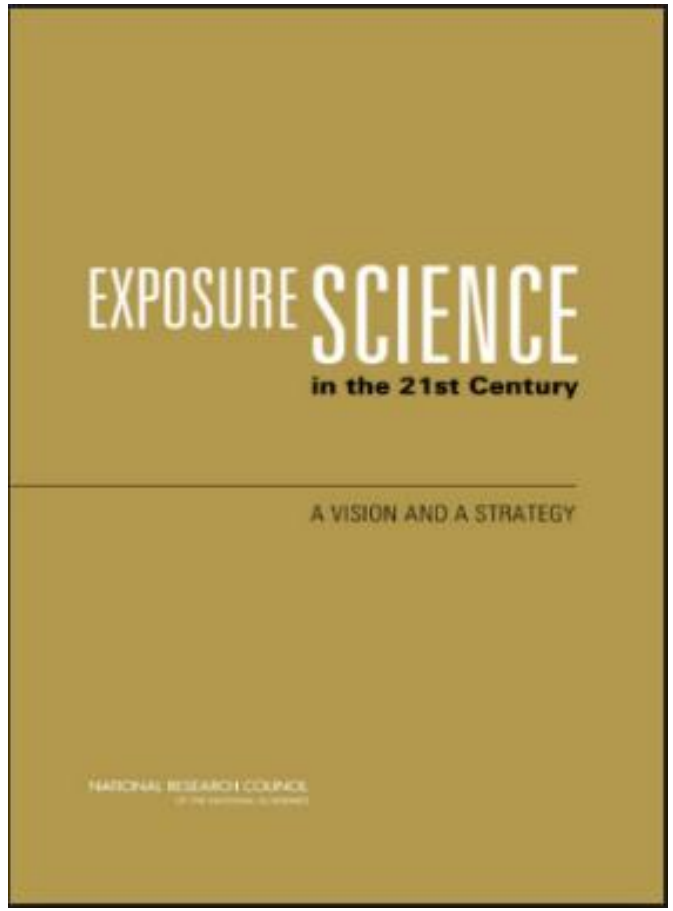


Example of precise language – measurement vs. event

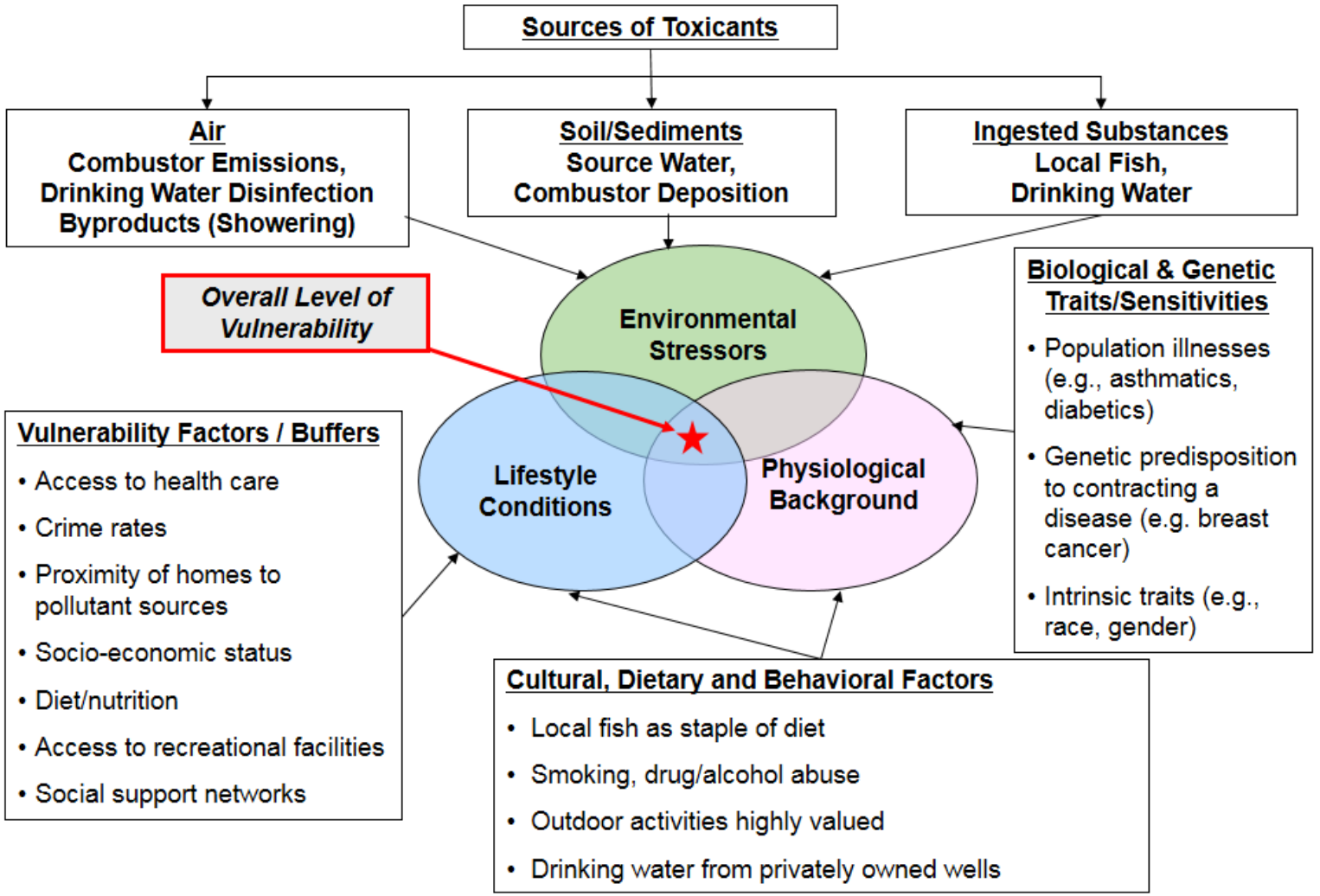


Key points raised - Biomarkers

1. Susceptibility vs. exposure: Toxic agent, metabolites, and secondary markers with markers for susceptibility throughout.
2. Should include exposure pathways and other contextual information.
3. Separate the marker from what the marker can represent.
4. Need to build from existing resources such as NAS Exposure Science and EPA Cumulative Risk Framework



EPA Cumulative Risk Framework (2003)



Graphic courtesy of Annie Jarabek

U.S. EPA. 2003. Framework for Cumulative Risk Assessment. U.S. EPA/ORD/Risk Assessment Forum. Washington, DC. EPA/600/P-02/001F.

Key points raised – Use Cases

1. **PM:** synergies with separate use case focused on exposure routes
2. **Carbon monoxide** - **cardiovascular** and susceptibility markers
3. **Smoking** and **chronic outcomes**: 'omic markers may be indicative of both exposure and biological response.
4. **Phthalates** and **asthma**: may be possible with NIEHS sponsored data, including HHEAR/ECHO

5. **Should include**: AOPs, omic data, large cohorts (e.g. HHEAR, ECHO), and model system databases (e.g., epigenome roadmap)

Gaps

1. Methods to link/annotate actual data (from labs) with ontologies
2. Identify what types of numerical or statistical models are needed
3. We are still identifying biomarkers, even for known exposures
4. Examples that include all desired information (exposure routes, biomarkers from known exposures, biomarkers of different types…)
5. How do we disseminate complex ‘omics information?
6. How do we know when we are successful? (e.g., the dimensionality of ‘omics and biomarker data may be large and how informative they might be in addressing a use case may not be known)

Challenges

1. Capturing and integrating the information from the people who are experiencing the adverse outcome and connecting that data to research measurements.
2. Integrating different types of data in order to tease apart association and causation. When we find a method that works, how do we repeat and generalize?
3. Modeling complex biomarkers that might be cell or tissue specific (and dependent on the route of exposure)
4. Dimensionality of the problem
5. Noise in our measurements
6. Define how detailed these models need to be to be useful
7. Data is at different scales, modalities, organisms, and tissues
8. AND SCOPE!!

Data and Knowledge Resources (examples)

Knowledge resources

- Monarch Initiative

Cohort data

- HHEAR, ECHO

Model system and/or experimental resources

- NIEHS Target II

Next steps

1. Identify participants who want to work on a specific sub use case (1-4)
   - Reach out to community (HHEAR, ECHO, Superfund) for translational component
2. Articulate the scientific applications
   - e.g., Smoking, Phthalates, PM
   - Create context for the translational work
3. Articulate data sources beyond those mentioned
4. Coordinate with the other use cases
5. Determine a timeline for this use case

If interested in participating, email

Stephanie Holmgren
holmgre1@niehs.nih.gov,

Chirag Patel
Chirag_Patel@hms.harvard.edu,

AND

Steve Edwards
swedwards@rti.org

Thank you!


OPEN
FOR
DISCUSSION
?