Data: The Common Thread & Tie That Binds Exposure Science

Vasu Kilaru – National Exposure Research Laboratory, US EPA

Zachary A. Collier – US Army Engineer Research & Development Center

Data Collection and Management

- Charter:

There is a compelling need for the development of Policies and Standard Operating Procedures for the collection and management of exposure and exposure-related data. In the near term, exposure science needs to develop strategies to expand exposure information rapidly to improve understanding of where, when, and how exposures occur and their health significance. Data generated and collected would be used to evaluate and improve models of exposure for use in generating hypotheses and developing policies. ***This data collection and management group is tasked to review the current status of policies and procedures associated with exposure data, and to identify critical needs and opportunities to partner across the government to establish a common set of standard operating procedures to facilitate the sharing and utilization of existing and future exposure data.***

Question: What is Exposure Science Data?

Answer: A lot of things

- *Broadly speaking it **"could"** include many things*
  - SHEDS / HEM
  - Measurement studies...HEDS (Human Exposure Database System)
  - Media specific model outputs (CMAQ, water quality models...etc)
  - Exposure Factors Handbook
  - NHANES....CDC survey of health and nutrition
  - Non-targeted analysis? Pharmacokinetics data?
  - Civic Science – data collected from non-experts and non-government?

- *Progress in ES21 will require a thoughtful framework that encompasses all these various data sources and types in a manner that provides greater openness and utility....a real challenge*
  - Encompasses everything from data to parameterize models, to measurement data on metabolites in blood and urine, to product use categories, to surveys.....all of which form a piece of the exposure puzzle

Issue with the existing paradigm

- ***Data Access***:   What data are currently available and where?
- ***Data Openness***:   How to incentivize documentation & sharing?
- ***Data Quality***:   How to ensure we have data quality we need?
- ***Data Sensitivity***:   Ownership, "de-identification", protection?
- ***Big Data & 7 Vees***:
  - Velocity….the speed or frequency of data generation
  - Volume…..the size and quantity of new data
  - Veracity….the quality of data
  - Variety…structured, unstructured, various formats….a lot of work
  - Variability….referring to ontological issues in meaning and context
  - Visualization….and analytics used to discern patterns in complexity
  - Value….or the creation of knowledge

Subgroup Recommendations: Short Term

- *Making Data Available and Discoverable:* Enabling data to more easily be published and found
  - Establish an ES21 web presence with links to relevant exposure science information and ongoing research efforts....but where will it be located and who will manage?
  - Institutionalize the public provision and documentation of data as part of the grant process
  - Establish an Exposure Science Data Community of Practice with members across the Federal research community.
  - Longer term efforts include education (e.g., webinars, workshops), newsletters, and other dissemination efforts to keep the research community engaged in the latest developments in the field.

Subgroup Recommendations: Medium Term

- *Data / Metadata Standards*: critical to the understanding the usability of data and data quality
  - Ontology issues are thorny and complicated. What to do?
  - Standards divided by exposure routes (inhalation, dermal, and ingestion), and further subdivided by chemical classes and/or endpoints.
  - Harmonizing these standards into a unified reporting standard.
  - Rating system? Allow user feedback (e.g., "thumbs up" or "thumbs down") on exposure data sets?
  - Incentivize open data & adherence to data standards?
  - Example: The Semantic Web…..an extension of the world wide web through standards. The standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF). The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries

Subgroup Recommendations: Long Term

- *Data Science to create knowledge*: Frameworks to integrate data and create useful knowledge from the variety of exposure science data
  - Traditional data mining and data analytics applied to exposure science
  - Cognitive Computing, the general term that includes the various forms of automated machine learning and pattern recognition
  - Need to better link exposure science data with risk and decision models to enhance decision support

Planning Matrix

	Short Term (next two to three years)	Mid Term (next three to six years)	Long Term (beyond six years)
Most		• Advanced big data analytics and algorithm development • Ontological reconciliations • Institutionalize data provision and documentation in grants	• Unified data standard • Data Science and Cognitive Computing for Exposure Science
Medium		• Dermal data standard • Metadata standard – including user feedback options for data quality and ease of use	• Ingestion data standard
Least	• ES21 web with links to exposure science data & ongoing research • Initiate Community of Practice	• Inhalation data standard	• Outreach efforts including education, newsletters, and other dissemination efforts

Level of Difficulty (vertical axis label)

Accomplishments & Path Forward

- Accomplishments
  - Shared list of ongoing data collection and management efforts
  - Identified research gaps and identified potential solutions
  - Developed a glossary of data management terms

- Path Forward
  - Participate in cross-subgroup case study
  - Identify areas for continuing collaboration

Talk Amongst Yourselves
I'm Verklempt

If Love makes the world go around
Standards stop it from spinning out of control