An Infrastructure for Reproducible Exposomic Research

Ramkiran Gouripeddi[1,2], Phillip Warner[2], Randy Madsen[2], Peter Mo[2], Nicole Burnett[1,3] Jingran Wen[1], Albert Lund[1], Ryan Butcher[2], Mollie Cummins[1,2,4], Julio Facelli[1,2], Katherine Sward[1,2,4]

[1]Department of Biomedical Informatics, [2]Center for Clinical and Translational Science, [3]Department of Chemical Engineering, [4]College of Nursing; University of Utah, Salt Lake City, Utah, USA

Introduction

- Understanding effects of the modern environment on human health requires a complete picture of environmental exposures, behaviors, and socio-economic factors.
- *Exposome*: encompasses life-course of environmental exposures & lifestyle beginning prenatally; complements the genome by providing a comprehensive description of exposure history[1].
- Exposomic research requires integrating diverse data types to support different research use-cases.
- Data gaps and sparseness are common with exposure monitoring and challenge generation of sufficiently complete exposomes.
- Systematically using available data with an understanding of their limitations could enable research reproducibility.

Conclusion

- A generalizable and metadata-driven platform for integrating multi-scale and multi-omics data provides a robust pipeline for reproducible research data.
- Informs end-user not only of the specifics about the data but also its limitations.

References

1. C. P. Wild, “The exposome: from concept to utility,” Int. J. Epidemiol., vol. 41, no. 1, pp. 24–32, Feb. 2012.
2. An Informatics Architecture for an Exposome, R. Gouripeddi, Session II06 – Secondary Use of Data for Research (Interactive Learning), AMIA 2016 Joint Summits on Translational Science, March 22nd, 2016, San Francisco. https://www.amia.org/sites/default/files/2016-joint-summits-program-book.pdf

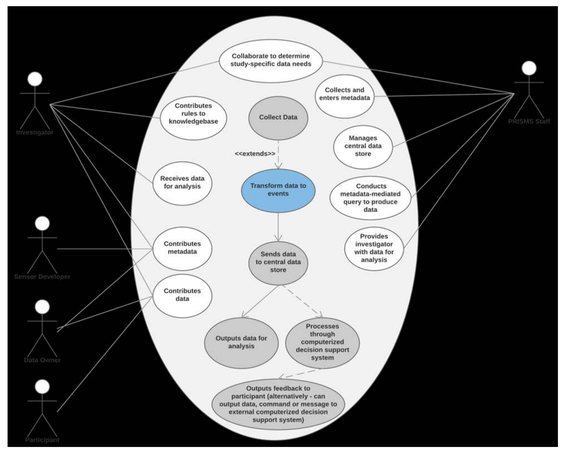
Genome ↔ Phenome ↔ Exposome

Exposome: Totality of human environmental exposures from conception onwards, complementing the genome[4].

$$Exposure = f\left(\begin{array}{c} Quantity\ of\ Air\ Pollutant, \\ Duration, \\ Frequency, \\ Person\ and\ Biological\ Charateristics \end{array}\right)$$

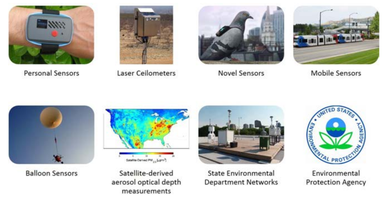
⟷

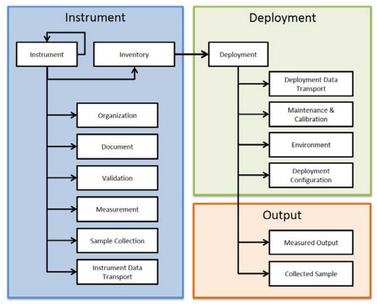
- Selection of Relevant Sensor Data Sources
- Modeling for a High Spatio-temporal Grid
- Characterizing Uncertainty
- Data Integration to Support Ease of Use



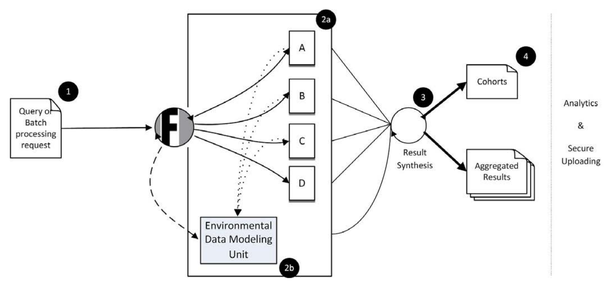
Use-case Archetypes

Air Quality Sensors

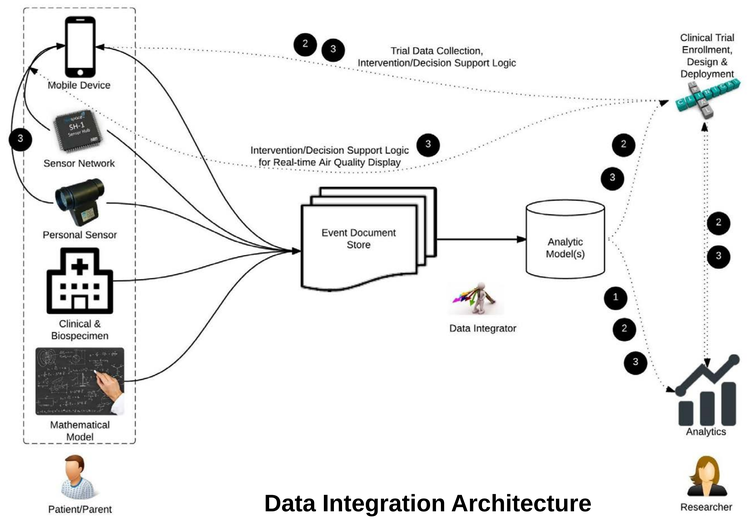




Sensor Common Data Model



Data Integration Workflow



Data Integration Architecture

Informatics Infrastructure

We are developing a scalable computation infrastructure in order to systematically generate air quality exposomes for the NIH Pediatric Research using Integrated Sensor Monitoring Systems (PRISMS) program.

- **Use cases**: Research use-cases clarify requirements and work flows.
- **Data Models**: Conceptual data models integrate diverse sensor data as related to individuals and populations. Standards support integration across centers.
- **Metadata Management**: Graph implementation of OpenFurther’s Metadata Repository[2] for authoring and storage of metadata to support proper use of heterogeneous data.
- **Integration Platform**: A metadata-driven big data infrastructure based on the OpenFurther[2] (OF) platform.
- **Integrated Data Store**: OF generates an event-document store (EDS) to support different use-cases. EDS captures spatio-temporal variations of events (e.g. air pollutant concentrations, asthma symptoms), and locations of the individuals and populations.
- **Mathematical Modeling**: Fills gaps in measurements and characterizes uncertainties in the data.

Acknowledgements

PRISMS is supported by NIH/NIBIB U54EB021973. OpenFurther is supported by NCATS UL1TR001067, NCRR/NCATS UL1RR025764, 3UL1RR025764-02S2, AHRQ R01 HS019862, DHHS 1D1BRH20425, U54EB021973, UU Research Foundation.

Contact: Ram Gouripeddi
ram.gouripeddi@utah.edu

The University of Utah Biomedical Informatics