Environmental Health Language Collaborative (EHLC): a route to environmental health science data harmonization

Anna Maria Masci
Office of Data Science
National Institute of Environmental Health Science
mascia2@niehs.nih.gov

ICBO 2022

26 September Ann Arbor, Michigan

What is Environmental Health ?

Environmental health is a science that studies the effect of exposure to environmental factors on human health.

Diverse data types

Climate

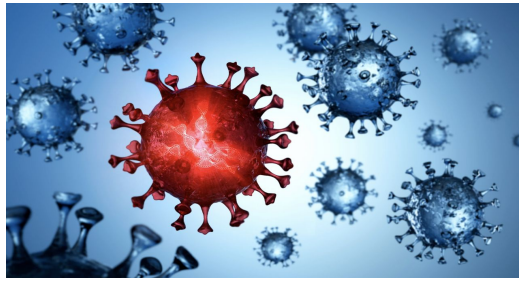
Radiations

Food access and quality

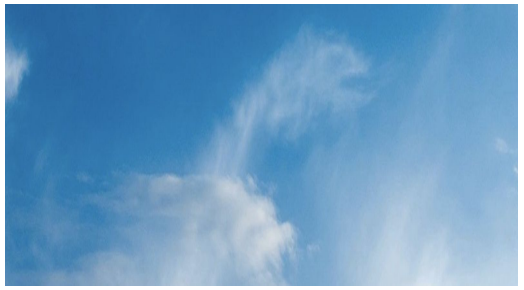
Water quality and access

Financial

Infectious agents



Air quality



Chemicals

Activity

Occupational

The Environmental Health Language Collaborative:

Using Harmonized Language to Address Environmental Health Challenges

EHLC Strategic Goals and Roles

Develop Language-Based Solutions

Foster community-based extension and development of terminologies

Promote and develop methods/tools for applying harmonized language

- **Forum to coordinate**
  - Identify use cases and needs
  - Prioritize activities
  - Strategies and approaches for solutions
- **Platform for collaboration to develop solutions**

Implement Language-Based Solutions

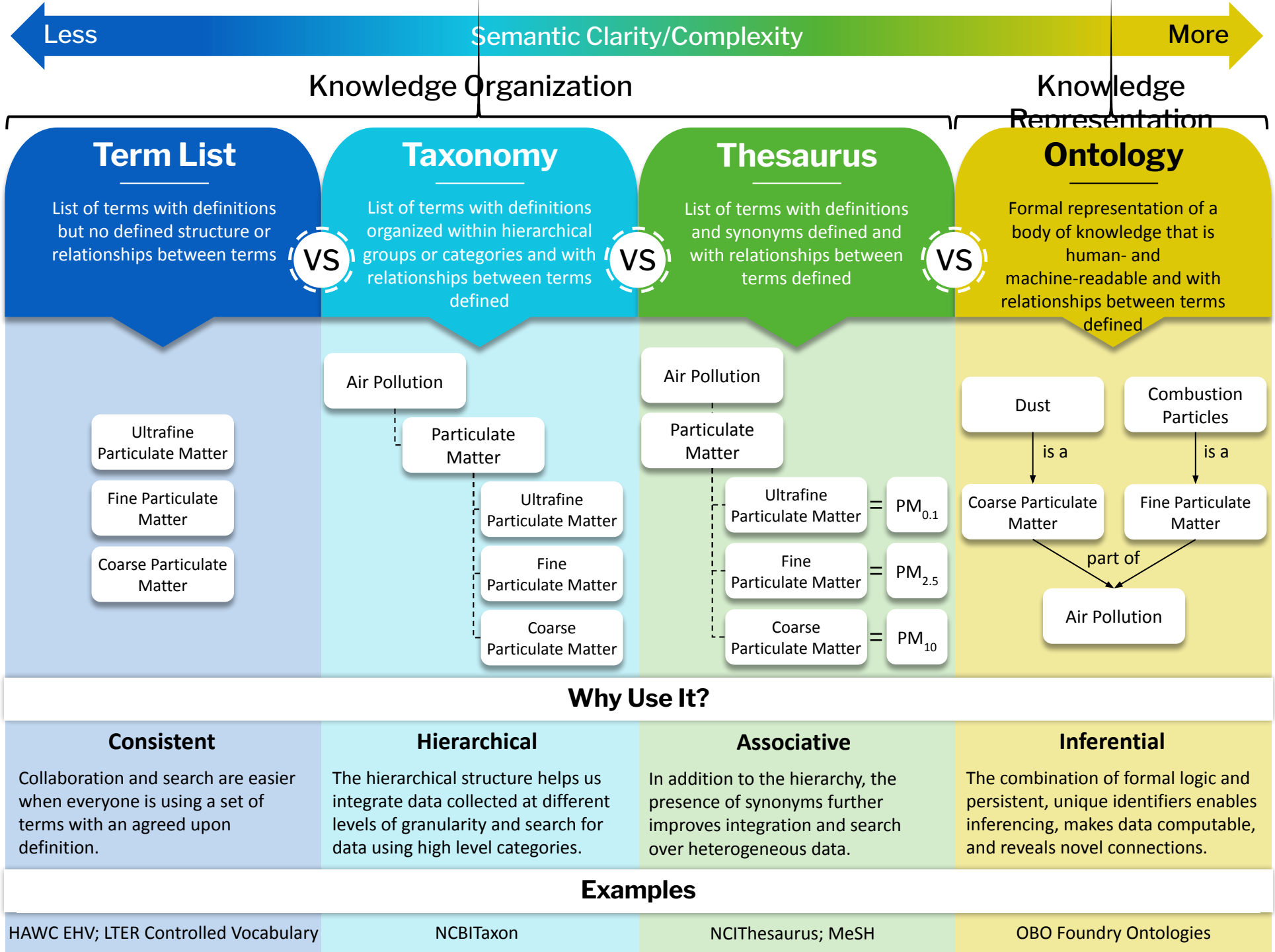
Apply language standards and best practices for accurate environmental health data and knowledge representation

- **Community hub**
  - Identify and promote incentives and support adoption of semantic approaches
  - Offer a resource portal

Advocate Value of Language

Cultivate a vocabulary aware environmental health community

- **Community of Practice**
  - Exchange information, ideas, expertise
  - Foster education and training


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

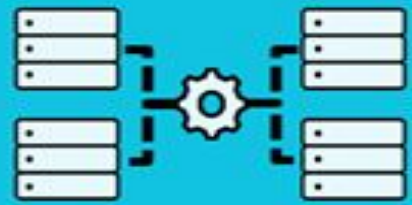
Data Discovery

Using harmonized language to enable data searching and discoverability

Led by Michelle Angrish (US EPA) and Shannon Bell (RTI), this use case is focused on how to improve the findability of exposure data from study reports and databases. This includes facilitating discoverability, screening, and curation of content (whether data or information) related to chemicals, exposures, and endpoints for in vivo and mechanistic studies. The initial approach is to showcase how improving the reporting of key study aspects (leveraging existing minimum information templates) can improve data discovery.

Place-Based Health Research

Developing data tools and strategies to understand how place influences health



Led by Carmen Marsit (Emory University), this use case focuses on data integration and harmonization to better understand how geographic location, time, and structural and societal factors affect environmental exposures. This use case team will develop tools, like a cross-ontology place-based language structure, to improve the rigor and interoperability of place-based environmental health research.

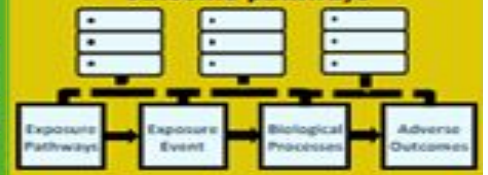
Data Harmonization

Identify and reduce barriers to using harmonized language for dataset annotation

Led by Jeanette Stingone (Columbia University), the purpose of this use case is to facilitate the integration of individual datasets through use of data and metadata standards and annotation of existing datasets. The use case team will anchor their work in epidemiologic studies and pursue the creation of a template-based approach to inform data collection, with the goal of integrating or harmonizing data across studies.

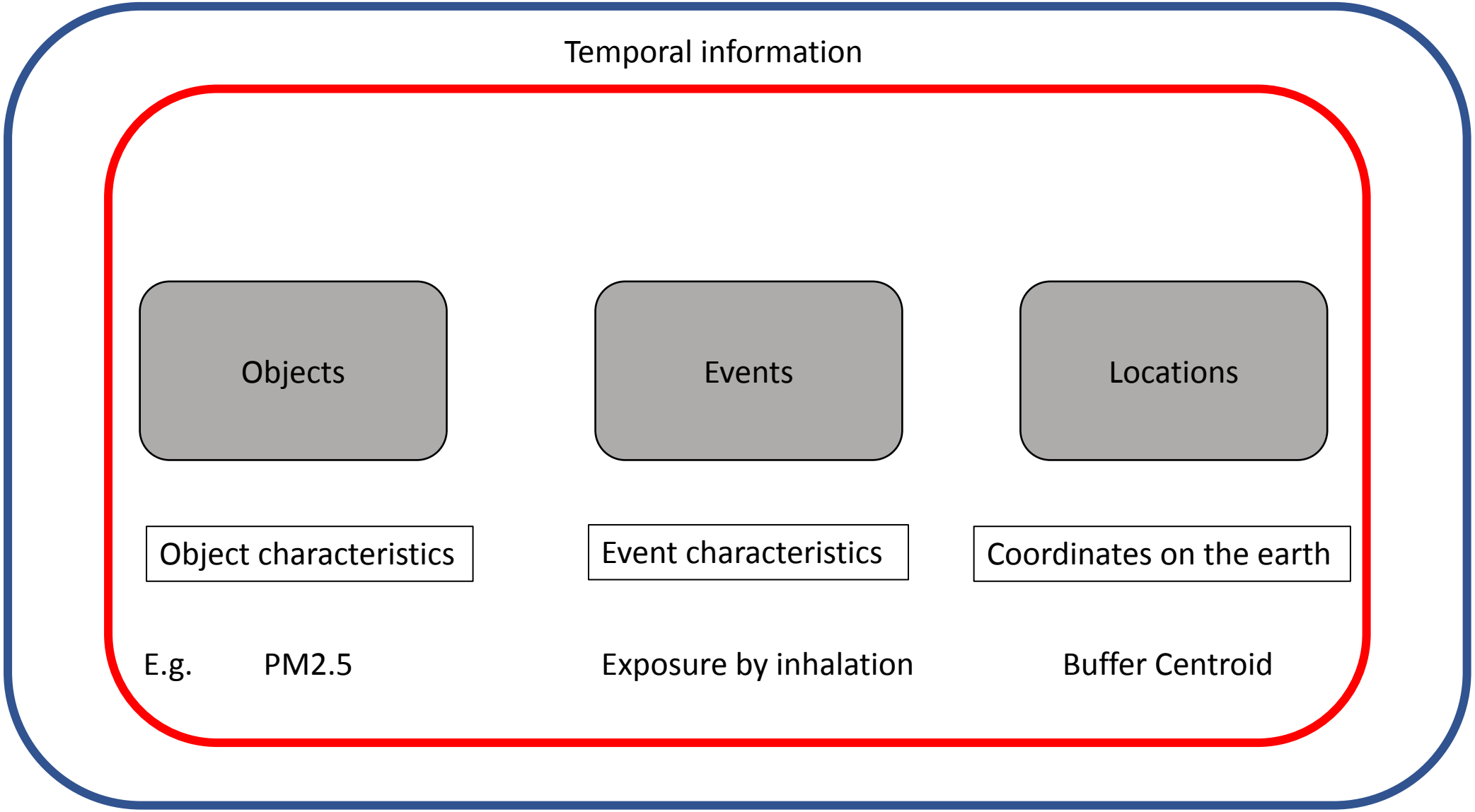
Biomarkers and Biological Processes

Exploring how harmonized datasets lead to better understanding of adverse outcome pathways



Led by Chirag Patel (Harvard University) and Stephen Edwards (RTI), this use case explores semantics for capturing the biomarkers and biological processes perturbed by an exposure event. This will be divided into two parts: creating a semantic description of an exposure event and linking that event to an adverse outcome. This use case was inspired by the Adverse Outcomes Pathway (AOP), which describes an exposure event, and the subsequent "key events," which detail what happens from the exposure to the adverse outcome.

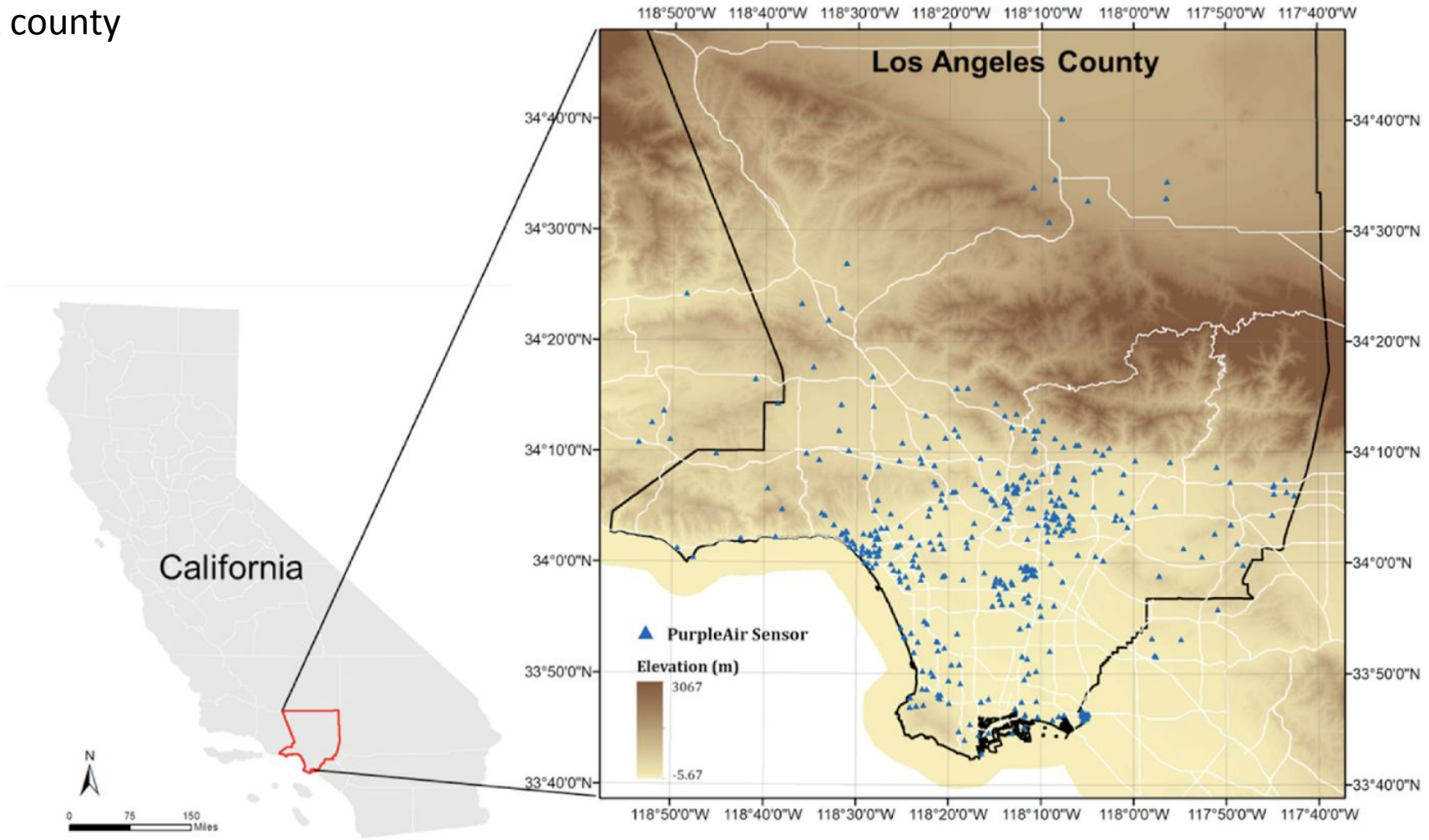
GEOSPATIAL DATA

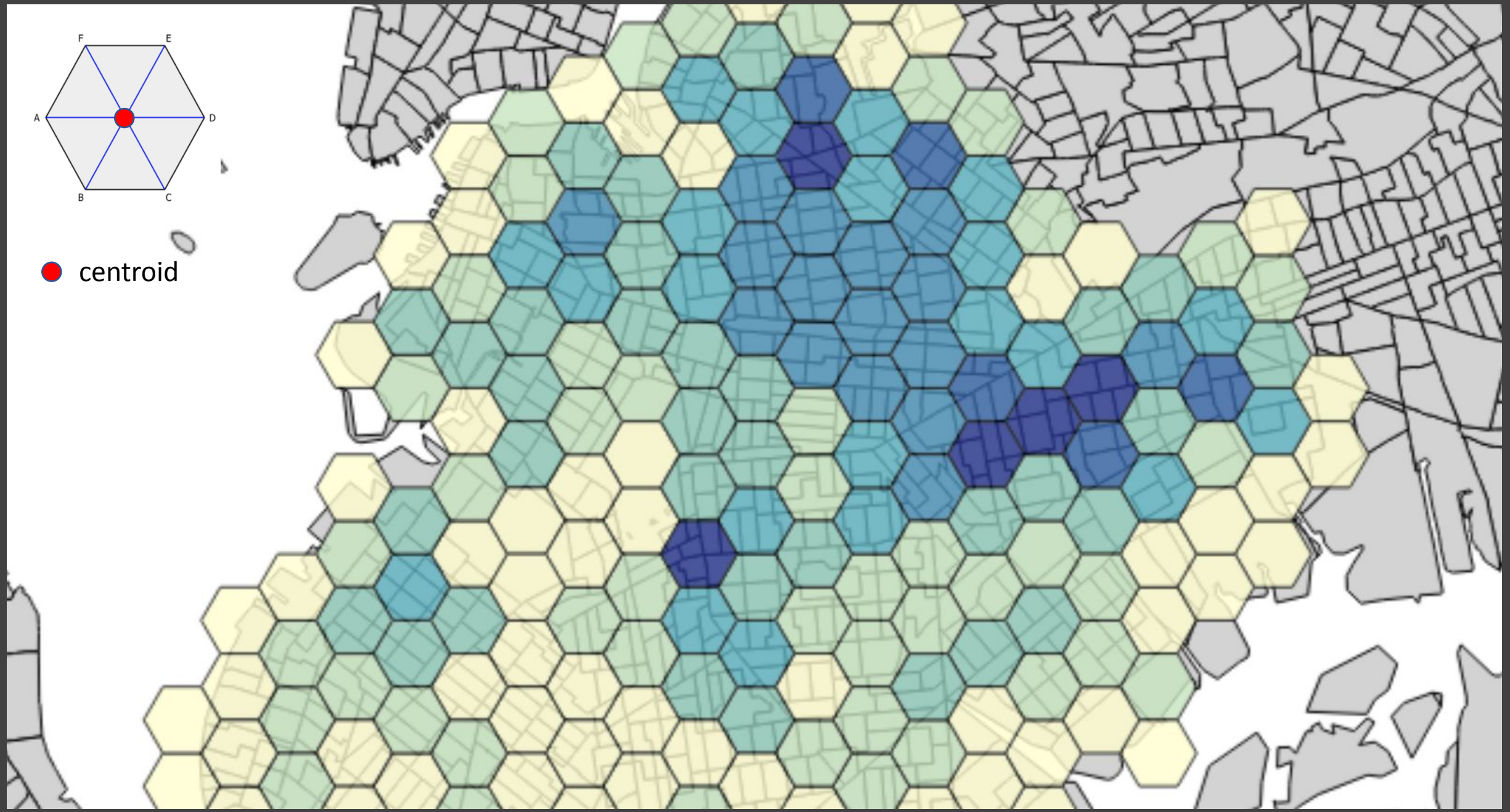


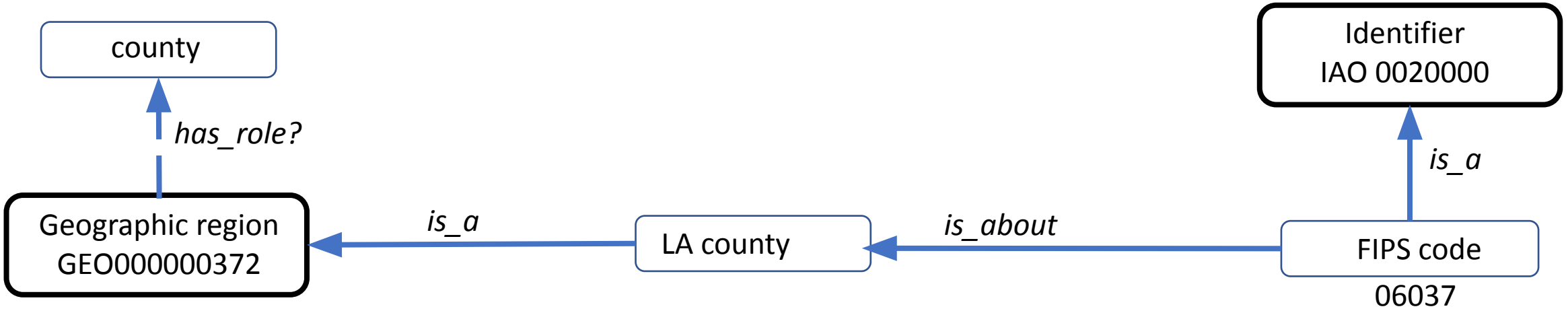
PMID: 334766
65

Environmental Research 195 (2021) 110653

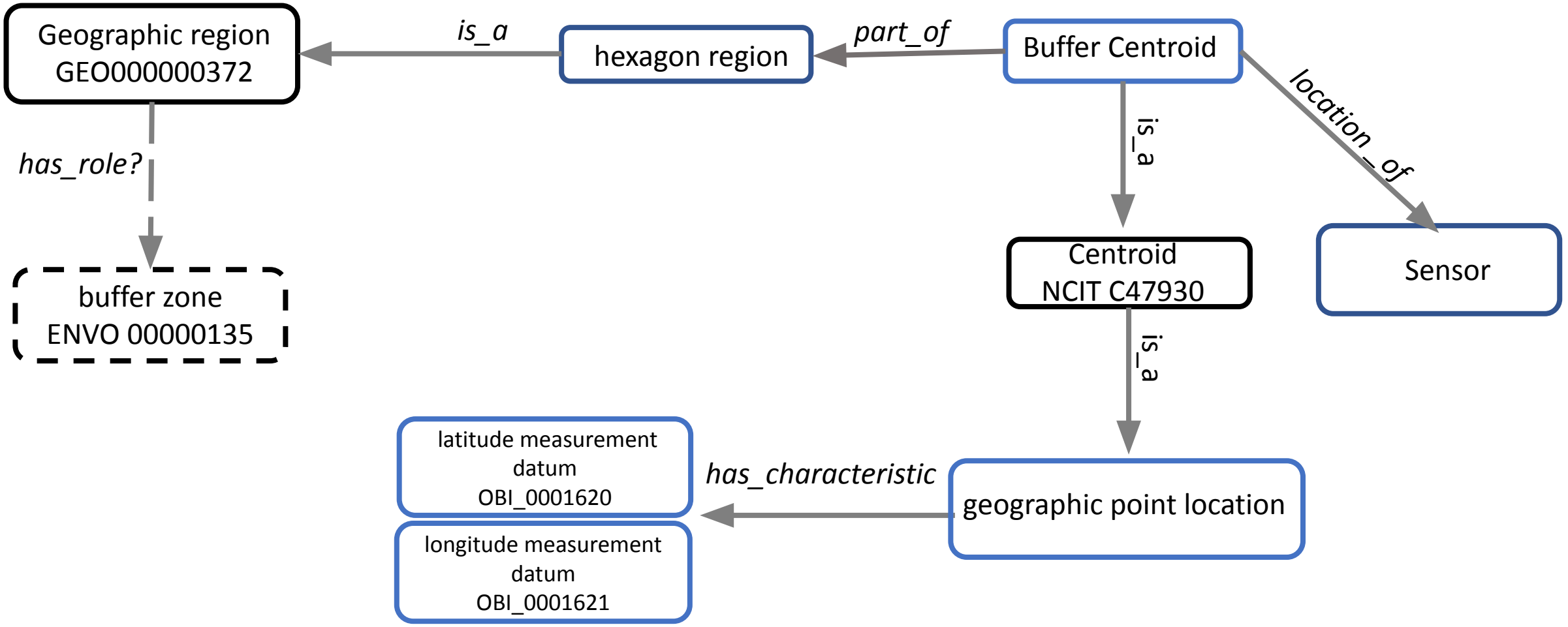
Study area : LA county




F
E
A
D
B
C
centroid



FIPS: Federal Information Processing System



Buffer Zone : A geographical region that serves the purpose of keeping two or more other areas (often, but not necessarily, countries) distant from one another, for whatever reason

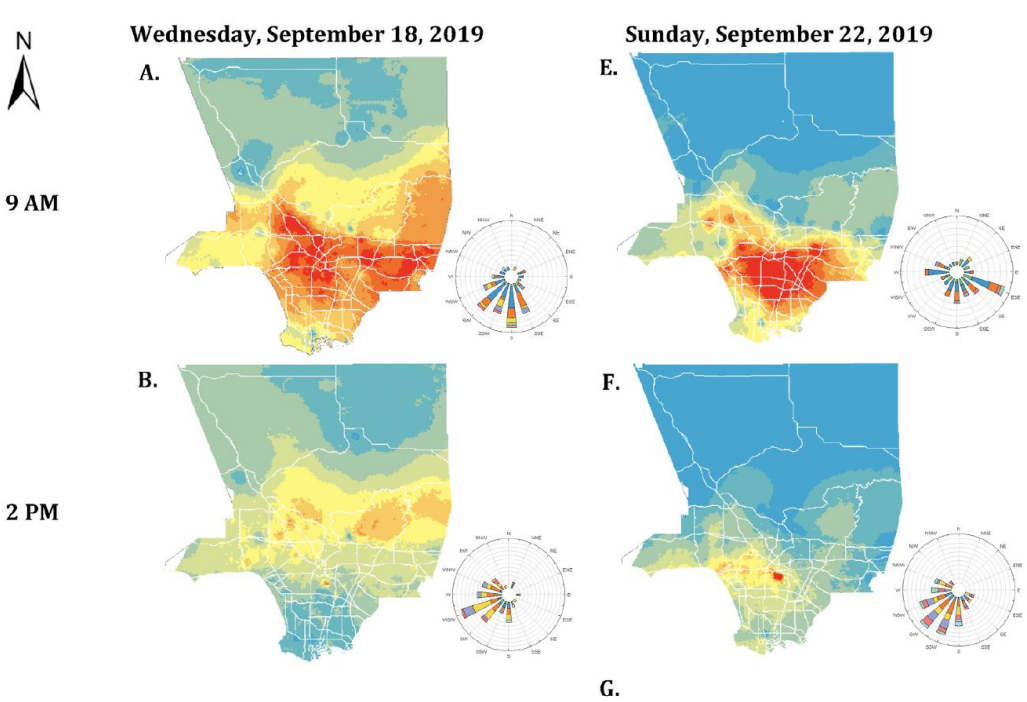
VARIABLE

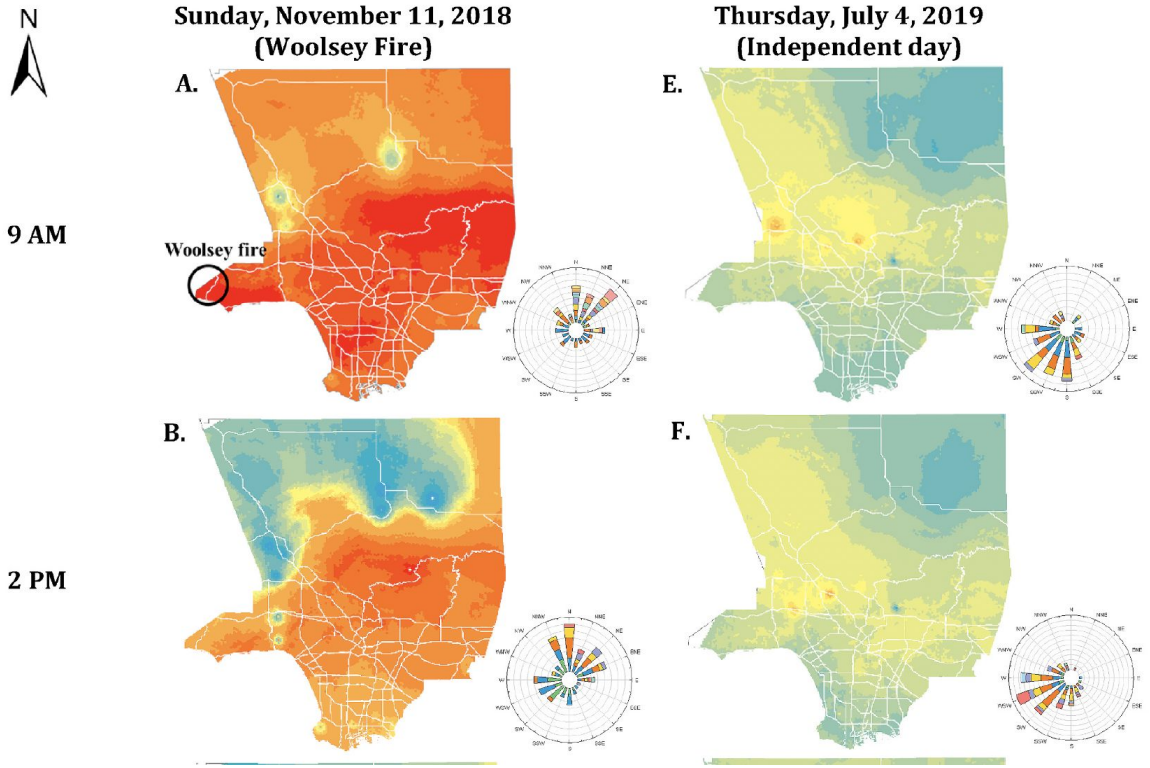
- Temperature
- Humidity
- Wind speed
- Wind direction
- Precipitation
- Traffic (Type of Road)
- Temporal

U.S. Census Bureau's Census Feature Class Codes (CFCC) provide information on the classification of a feature.

CFCC Description
A00 Road, major and minor categories unknown
A01 Road, unseparated
A02 Road, unseparated, in tunnel
A03 Road, unseparated, underpassing
A04 Road, unseparated, with rail line in center
A05 Road, separated
A06 Road, separated, in tunnel
A07 Road, separated, underpassing
A08 Road, separated, with rail line in center
A10 Primary road with limited access or interstate highway, major category
A11 Primary road with limited access or interstate highway, unseparated
A12 Primary road with limited access or interstate highway, unseparated, in tunnel
A13 Primary road with limited access or interstate highway, unseparated, underpassing
A14 Primary road with limited access or interstate highway, unseparated, with rail line in center
A15 Primary road with limited access or interstate highway, separated
A16 Primary road with limited access or interstate highway, separated, in tunnel
A17 Primary road with limited access or interstate highway, separated, underpassing
A18 Primary road with limited access or interstate highway, separated, with rail line in center
A20 Primary road without limited access, U.S. and state highway, major category
A21 Primary road without limited access, U.S. and state highways, unseparated
A22 Primary road without limited access, U.S. and state highways, unseparated, in tunnel
A23 Primary road without limited access, U.S. and state highways, unseparated, underpassing
A24 Primary road without limited access, U.S. and state highways, unseparated, with rail line in center
A25 Primary road without limited access, U.S. and state highways, separated
A26 Primary road without limited access, U.S. and state highways, separated, in tunnel

Hourly mean of PM2.5 concentration in LA county





PMID: 33476665

Acknowledgements

Stephanie Holmgren. Office of Data Science NIEHS, Durham, NC, USA

Charles Schmitt Office of Data Science NIEHS, Durham, NC, USA

Rima Habre University of Southern California, Los Angeles, CA, USA

Anne E Thessen University of Colorado Anschutz Medical Campus, Center for Health AI, Aurora CO, USA

Rebecca Boyles Center for Data Modernization Solutions, RTI International. Durham, NC. USA

Carmen Marsit Gangarose Department of Environmental Health, Emory University Rollins School of Public Health, Atlanta, GA, USA

Have Ideas? Want to get involved?

EHLC website at https://www.niehs.nih.gov/research/programs/ehlc/index.cfm

EHLC email distribution list for information about events, opportunities to contribute, and products.

Contact Anna Maria Masci (NIEHS) at **mascia2@niehs.nih.gov** Stephanie Holmgren (NIEHS) at holmgre1@niehs.nih.gov


THANK
YOU!