An interoperable, usable and open framework for linking environmental and health data

Albert Navarro-Gallinad[1,2], Fabrizio Orlandi[1], Maeregu Woldeyes Arisido[2], Luisa Zuccolo[2] and Declan O'Sullivan[1]

[1] ADAPT Centre for Digital Content, School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
[2] Health Data Science Centre, Fondazione Human Technopole, Milan, Italy

Introduction. Linking data for exposure assessment in environmental epidemiology presents a challenge for researchers since environmental and health datasets are generally created for different purpose [1]. This is especially challenging in the context of open science practices towards promoting transparency, quality and efficiency of scientific research, in line with the FAIR principles of Findability, Accessibility, Interoperability and Reusability (FAIR) [2, 3]. During the process of making data FAIR, the most complex step for researchers is to make data interoperable following best practices from World Wide Web Consortium (W3C) standards [4]. The W3C developed the Resource Description Framework (RDF) as the standard graph data model for data interchange publication of information on the Web. Following the W3C standards not only addresses the data interoperability challenge by providing meaning to the data (semantics), but it also provides querying and reasoning capabilities to generate new insights (efficiency) and transparency of the linkage process (explainability). The generation of interoperable data can enhance data linkage tasks in environmental epidemiological studies when integrating datasets from diverse data sources. For data linkage to be effective, the exposure linked data need to be also usable and reusable by researchers without highly specialised expertise in making data interoperable (e.g., RDF W3C standards).

Objectives.

1. To create an interoperable, usable and open framework for linking environmental exposure assessment data to health data.

2. To demonstrate the applicability of our work for a use case that requires meaningful linkage of local area-based environmental data with administrative health data from the Lombardy region (Italy) to assess exposures affecting pregnancy health.

Methods. The Semantic Environmental and Rare Disease data integration Framework (SERDIF) was used to facilitate the data integration step for researchers (Figure 1). The framework is a combination of a methodology, a knowledge graph and a user interface to enable researchers to effectively link environmental and health data following W3C standards for data interchange. The resulting linked data is provided with enough information about the origin of the data and processing steps in a human- (CSV and HTML) and machine-understandable (RDF) format. The data linkage includes selecting a proper and flexible time window and geographical area to filter the environmental observations based on the location and date features of particular health events. This process establishes a new link and constructs an exposure dataset ready to be used for analysis and published following open science practices to be reused by other researchers in different contexts. Furthermore, SERDIF framework was developed to comply with and promote the data governance aspect of the processing of health and environmental data, central to the linkage process. SERDIF is openly available for use by researchers who similarly want to link health and environmental datasets at: https://w3id.org/serdif.

Results. The usability and potential usefulness of SERDIF for researchers has already been evaluated and published by the authors of this abstract [5]. The usability evaluation included three use cases with a total of 30 researchers studying environmental risk factors of two rare diseases, ANCA-associated vasculitis and Kawasaki disease. The positive results indicated that the framework had the potential to be expanded and implemented to any disease of health event. SERDIF was implemented for linking environmental exposures on pregnancy health in the Lombardy region in Italy. Researchers were able to efficiently link administrative health and COVID-19 data with local environmental observations from measurement stations with enough provenance information for its meaningful use. The implementation granted the possibility to deposit environmental datasets into a local folder in CSV, NetCDF and GRIB formats; and link them based on distance from the health event geospatial coordinates or within a Nomenclature of territorial units for statistics (NUTS) for purposes of framing EU regional policies. Furthermore, researchers were able to define a window of exposure to study a period of increased vulnerability for the health of mothers such as prenatal, perinatal and early postnatal life. The resulting linked exposure dataset

included an interoperable distribution (RDF) ready to be deposited in an open data repository towards making the dataset FAIR complying with open science practices.

We are demonstrating the potential of this framework through its application to small area-level linkages of layers of environmental data (from Agenzia Regionale Protezione Ambientale – ARPA – Lombardy) to population-based data on health outcomes in pregnancy and at birth (from administrative health records, such as the Certificato Di Assistenza al Parto -CeDAP- for the Lombardy region).

Conclusions. The open framework provided researchers with interoperable and usable data to facilitate the complex data linkage tasks in environmental exposure studies, as demonstrated in our case study with Lombardy data. Our framework enables substantial advances in environmental epidemiological studies, sustains scalability of European and International projects, and promotes open science-complying best data practices.

Disclosures. The authors have declared no conflicts of interest.

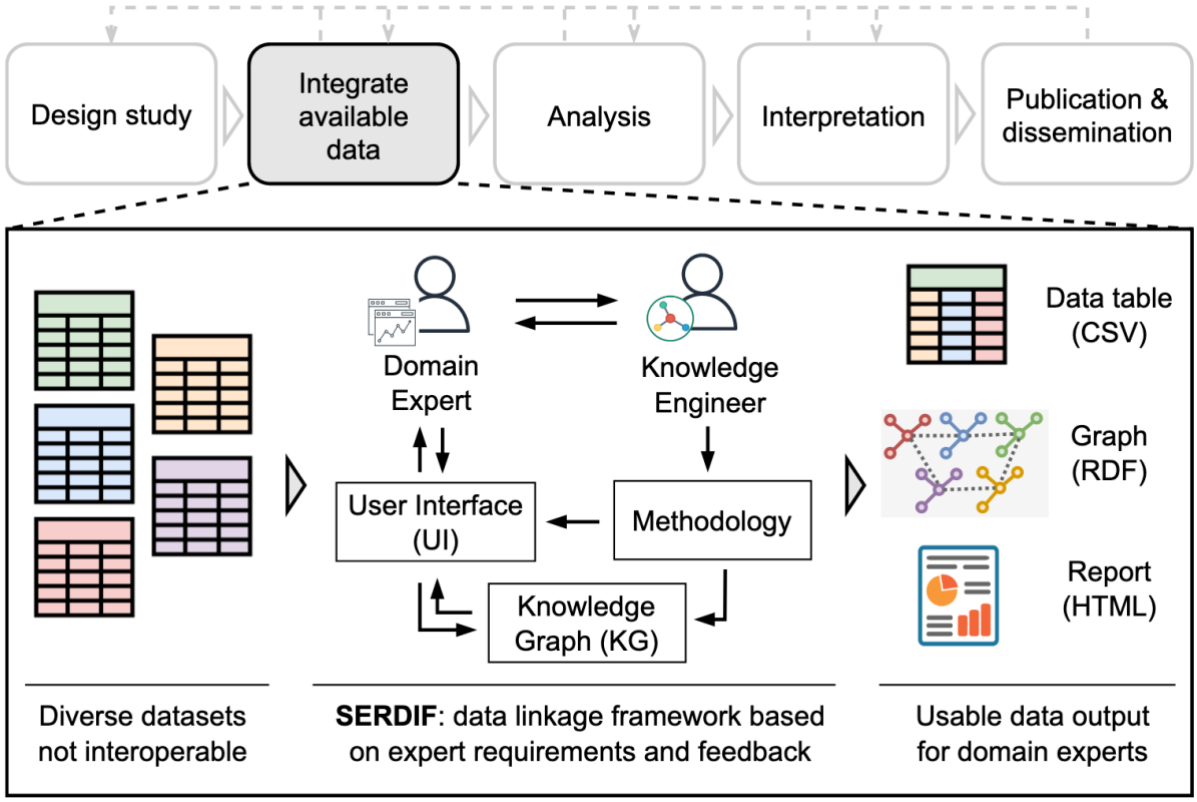

Figure 1. Location of the data linkage framework within an exposure assessment workflow.

Bibliography

[1] Standing Committee on Emerging Science for Environmental Health Decisions, Board on Life Sciences, Board on Environmental Studies and Toxicology, Division on Earth and Life Studies, and National Academies of Sciences, Engineering, and Medicine. Informing Environmental Health Decisions Through Data Integration: Proceedings of a Workshop in Brief. National Academies Press 2018. https://doi.org/10.17226/25139
[2] Wilkinson M. D. et al., The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 2016; 3(1):160018. https://doi.org/10.1038/sdata.2016.18
[3] Jacobsen A. et al., FAIR principles: Interpretations and implementation considerations. Data Intelligence, 2020; 2(1):10–29. https://doi.org/10.1162/dintr00024
[4] W3C World Wide Web Consortium (W3C), Semantic Web Standards 2023. https://www.w3.org/standards/semanticweb/
[5] Navarro-Gallinad A. et al., Evaluating the usability of a semantic environmental health data framework: Approach and study. Semantic Web 2023; 14(5); 787-810. https://doi.org/10.3233/SW-223212
Topic. General Epidemiology