Standardized, linkable, analysis-ready environmental data for understanding and preventing disease and building resilience to climate change

A white paper submitted to the New Digital Research Infrastructure Organization (NDRIO)

Jeffrey R. Brook[1], Dany Doiron and Eleanor Setton

[1] Designated contact person: Jeff.brook@canue.ca

Current Issues

Better understanding of the links between environment and both non-communicable (NCD) and communicable diseases, factors that alter susceptibility, and the effectiveness of possible interventions is a critical path towards reducing the burden due to disease on individuals, families, the health care system and the economy. Identifying new risks resulting from climate change are also critical to inform adaptation measures and to build long-term resilience. A deeper, more-holistic characterization of the physical and socioeconomic environments that individuals experience ('environmental exposure') is essential to achieving these benefits as is the timely and secure linkage to individual-level electronic health records (EHRs).

The Canadian Urban Environmental Health Research Consortium (CANUE) was initiated to help address these challenges and represents a working example for how environmental exposure data can be centralized, enhanced, linked with commensurate health data and widely shared. CANUE takes in environmental and area-level 'social' (demographic/socioeconomic) data from a wide variety of sources, some in ready-to-use condition and others independently derived from massive digital datasets such are global remote sensing observations, with the majority requiring substantial formatting and processing prior to being used as inputs to derive environmental models and metrics of exposure and for linkage to individual-level health data.

This white paper focuses on these environmental exposure data and their linkage to confidential health data in the context of Digital Research Infrastructure (DRI) challenges and needs. It draws from experience gained in establishing CANUE and is intended to complement other white papers in this series which have articulated the DRI perspectives relevant to health data. This includes those prepared by Wong et al., Rosella et al. and Greiver et al., who tackle different aspects of electronic health records (EHRs) from the under-utilized resource of primary care data to the needs for linkage and interoperability of diverse sources of health/medical information to facilitate more-rapid and efficient analyses through advanced artificial intelligence (AI) approaches.

CANUE (https://canue.ca/) relies on project staff as well as academic collaborators and government data producers across Canada to develop standardized, analysis-ready environmental and socio-economic metrics relevant to health outcomes from heart and lung disease, to obesity and mental health. We currently use the following DRI tools, services and resources regularly to fulfill our mandate:

Data handling and processing

- Storage (project, nearline and archive) - local personal computers, personal Google Drive space, purchased portable hard drives, Compute Canada, and Google cloud storage.
- File transfer/sharing via GLOBUS for large datasets, Filezilla, SFTP, and DropBox.
- Remote connections to shared computer workspaces via WinSCP, Remote Desktop, institutional virtual private networks (VPNs).

- High performance local PCs for data preparation and analysis, relying on institutions or individual team members for data backup processes.
- High performance computing infrastructure from Compute Canada (GPUs, parallel processing of large jobs).
- Data processing and analysis tools - R, MATLAB, python, java, html, Jupyter notebook, Docker, ArcGis, QGis, Google Earth Engine, computer vision/ML (PSPnet, YOLO), SQL, MSAccess, Excel, etc.
- IT/Network support from local institutions and Compute Canada staff.
- Domain hosting by for-fee commercial hosts, and via ssh connections to databases hosted on Compute Canada virtual machines.

Data Access and Linkage

- Secure data transfer mechanisms.
- Health data holders secure research environments (i.e., PopData BC, Statistics Canada secure data linkage environment (SDLE), Research Data Centres, CanPATH) and as described by Wong et al., Rosella et al. and Greiver et al. in this White Papers series.

Data management

- Web-enable PHP metadata and spatial data browser to manage and provide data access to validated users.
- Standard digital metadata.
- Unique identifiers for datasets/research publications, made available at no charge via current DRI in Canada.

A key challenge for CANUE, and similar research organizations/projects, is the lack of project IT staff to manage the complex computing infrastructures required to conduct our daily activities. We use a constellation of local PCs and cloud-based virtual machines, operating on Windows and Linux, and the learning curve every time we need to add a new program or reach storage or file transfer limits and need to find new solutions is costly in terms of time and effort. While the virtual (cloud) computing infrastructure we use is supplied by Compute Canada, it is technically up to CANUE staff to install operating systems, set up file permissions/access, install necessary programs and any related resources (i.e. python libraries), perform regular updates for all operating systems and programs and troubleshoot when these produce conflicts, and set up back up procedures for virtual machines. We have only been able to use this extremely valuable DRI by leveraging Compute Canada staff for help in almost every aspect of managing the virtual computing and storage environment.

At the same time, data storage and processing needs are rapidly and continually increasing. The growth in relevant digital data sources and the need to accelerate their processing and use to understand the role of environment in maximizing health throughout the lifecourse[1] is outstripping resources. New data streams such as high resolution satellite and street-level imagery and video combined with machine learning techniques are providing, for the first time, data on local environmental conditions for much of the urbanized world.[2] For example, daily global satellite imagery is now available at 0.5 to 3 meter spatial resolutions.[3, 4] Street-level imagery is also becoming ubiquitous, via proprietary sources such as Google Street View and openly via crowdsourcing efforts like Open Street Cam. Many larger cities operate video cameras and these continuous streams of data are typically not archived but are overwritten on a regular basis. Using these images, computer programs can be trained to identify urban features, which can be turned into geospatial data and used to estimate urban exposures appropriate for environmental health research. Machine learning techniques and algorithms applied to satellite, street view, and video images already have been used to estimate greenness,[5] walkability,[6] urban heat island intensity,[7] and to even predict spatial distribution of social and environmental health inequities.[8] Approaches to quantifying environmental exposure at the individual level are also expanding due to an explosion in digital platforms such as internet access and GPS-enabled smartphones. Aggregated cell-phone data is commercially available.[9] representing highly-resolved information, both temporally and spatially, on where people go throughout each day. These data can then be integrated with spatial models of air pollution, noise, greenspaces, walkability and many other environmental data to produce continuous characterizations of individual 'exposomes'.

Future DRI State

The model established by CANUE, which involves active partnerships with health data holders, is increasingly being recognized as a necessary path forward to drive research and development.[10–13] As outlined in the other white papers referenced above, the scope of the health data that can and should be served by Canada's future DRI is also primed for a dramatic increase in size and scope and to gain the most knowledge from these data, environmental exposure information is essential. Mining these health datasets for new insights into the causes of NCDs, and subsequent preventative measures under current and future climates, and to track trends in NCDs will be seriously lacking, however, if they do not have the ability to explore, as holistically as possible, the environmental factors at play.

Experience gained through CANUE has identified the key features required of a DRI system; specifically, a more efficient use of environmental exposure data for health research over the next decade requires progress in four key areas: metadata and data access portals, linkage with health databases, harmonization of exposure measures and models over large areas, and leveraging 'big data' streams for exposure characterization and evaluation of temporal changes.

In order to optimize the utility of existing and emerging environmental data for health research, structures and mechanisms should be put in place to ensure that data are findable, accessible, interoperable, and reusable (FAIR).[10] The functionalities of a future DRI portal and supporting software can be best described within the FAIR framework: how the data are 1) found, 2) accessed, 3) made interoperable, and 4) made reusable.

Privacy concerns and related jurisdictional differences will continue to pose challenges to DRI supporting environmental health research and is clearly a major concern in managing individual-level health DRI. Given that this issue is best addressed by holders of health data, the key challenge for environmental exposure researchers will be to instill the willingness of the health data generators, who have numerous and important motives (e.g., personalized medicine, ingestion of more and more data, such as genetics and primary care records, optimization of treatment and of health care costs and, implementation of AI tools) to advocate for ongoing integration and advancement of environmental exposure data (i.e., within their DRI) and subsequent research collaborations related to environmental health. Related to this will be the challenge of sustaining a quick pace for this data integration. Both health and environment data evolve rapidly and research utilizing these data is strongest and has the greatest potential for impact if it is up-to-date as possible in terms of the technology used to generate the data and how current the patient follow-up data are.

In the best of all scenarios, a future DRI supporting environmental health research would

- Remove data storage barriers to enable the acquisition and processing of big geographic data (i.e., national, daily high resolution satellite imagery, street-cam video files and cell-phone GPS data).
- Provide seamless one-stop storage/backup/transfer on a cloud platform.

- Enable secure connections allowing integrated analysis of confidential and non-confidential databases, using individual geographic location information (e.g., postal code, lat/long, etc.) as a common link.
- Use a central metadata repository and data access system.

How to Bridge the Gap

In advancing to a future EDI ecosystem most-supportive of environmental health research the vast amounts of data in both of these realms need to flow into a common platform. For researchers to use the data both the standard highly secure facilities for in-person work and common de-identified data products will need to be retained. Traditionally, these latter data products have been created on an as needed basis. CANUE has worked with partners to produce some standard environmental data that is held with the health data custodians. However, to best enable future progress, current, national level efforts that bring together provincial health data (e.g., Canadian Institute for Health Information (CIHI), the Health Data Research Network (HDRN) Canada), as well as future platforms as described in the white papers referenced above, need to be leveraged to an even greater extent than at present. While it is feasible to *virtually* integrate data across provinces and even countries, and issues of how to best harmonize such data are being addressed, the expansion to include harmonized environmental data would be best served by having the same organization or consortium lead the work (i.e., as opposed to one for health and one for environmental exposures). The actual scope of the data that can support future research is well-portrayed in the figure in the white paper by Greiver et al. Once an organizational lead (virtual or otherwise) for compilation and management of these integrated data is created, a key role would be to maintain one 'intake' window for data developers, thereby bringing considerable national efficiency. This organization or agency would be best positioned to frequently enable and motivate data providers, including resources to insure data are not lost, to share data widely and derive innovative new data metrics. Through this effort Canadian researchers would also be better able to focus on identifying critical gaps in data and methods, new opportunities from national and international trends, including international partnerships, and new needs for standardizing the archiving and exchange processes.

To bridge the gaps revealed through CANUE's experience to-date, key infrastructure (software and system) capabilities include:

- Automated creation of digital metadata providing structured variable dictionaries and assigning unique and persistent identifiers for each dataset on upload to the portal.
- Interfaces for conducting complex spatial and metadata searches to facilitate data exploration, interpretation, and identification of variables/datasets of interest.
- Use of controlled vocabularies and compatible metadata standards; thereby supporting multi-platform data browsing and eventual environmental data integration (e.g., exposure datasets across Canada and internationally).
- Automated systems enabling researchers to request and subsequently download data once authorized, via a web-based portal interface, including a dynamic mapping application to allow for visualization of the data to refine access requests and uploading of spatial boundaries for extracting data for specific areas of interest. Health data holders (e.g., prospective cohorts and administrative databases) will have on-demand access via this process that, once approved, they can extract for research.
- Software to process new datasets not yet included in the data portal via scheduling functions that automatically and regularly pulls source environmental data available on open platforms.
- Automated processes for indexing of spatial datasets to commonly used linkage fields such as postal codes and small area census boundaries.

- Systems and procedures to facilitate routine/automated linkage of environmental exposure files with health databases (EHRs as well as health data held by observational cohorts).
- Infrastructure that substantially enhances the reusability of the data by providing a digital data request tracking system. This tracking system should capture information on incoming data requests, document the datasets downloaded, generate dynamic metadata/citations that documents specific queries used to produce the required data extracts. The system should also prompt users to return and update the system with related outputs (e.g., scientific reports and publications) over time.

Digital infrastructure is clearly critical to supporting research needs; however, expertise to support the infrastructure cannot be ignored. While computer scientists, data scientists and data managers are all essential, the resulting data are best improved and shared when those working on the infrastructure also have content expertise across disciplines. Researchers accessing the data can bring some of this, but all are better-served when those working day-to-day building the infrastructure have a deep understanding of the substantive content and the research questions it/they are supporting. Developing and motivating this expertise can thus not be over-looked.

References

1. Jia P. Spatial lifecourse epidemiology. Lancet Planet Health. 2019;3:e57–9.

2. Weichenthal S, Hatzopoulou M, Brauer M. A picture tells a thousand…exposures: Opportunities and challenges of deep learning image analyses in exposure science and environmental epidemiology. Environ Int. 2019;122:3–10.

3. Maxar. https://www.maxar.com/. Accessed 14 Jul 2020.

4. Apte JS, Messier KP, Gani S, Brauer M, Kirchstetter TW, Lunden MM, et al. High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data. Environ Sci Technol. 2017;51:6999–7008.

5. Li X, Zhang C, Li W, Ricard R, Meng Q, Zhang W. Assessing street-level urban greenery using Google Street View and a modified green view index. Urban For Urban Green. 2015;14:675–85.

6. Yin L, Wang Z. Measuring visual enclosure for street walkability: Using machine learning algorithms and Google Street View imagery. Appl Geogr. 2016;76:147–53.

7. Chakraborty T, Lee X. A simplified urban-extent algorithm to characterize surface urban heat islands on a global scale and examine vegetation control on their spatiotemporal variability. Int J Appl Earth Obs Geoinformation. 2019;74:269–80.

8. Suel E, Polak JW, Bennett JE, Ezzati M. Measuring social, environmental and health inequalities using deep learning and street imagery. Sci Rep. 2019;9. doi:10.1038/s41598-019-42036-w.

9. Why StreetLight: Our Data. StreetLight Data. https://www.streetlightdata.com/our-data/. Accessed 14 Dec 2020.

10. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3. doi:10.1038/sdata.2016.18.