



Reaping the benefits of Open Data in public health

P Huston¹, VL Edge^{1*}, E Bernier¹

Abstract

Open Data is part of a broad global movement that is not only advancing science and scientific communication but also transforming modern society and how decisions are made. What began with a call for Open Science and the rise of online journals has extended to Open Data, based on the premise that if reports on data are open, then the generated or supporting data should be open as well. There have been a number of advances in Open Data over the last decade, spearheaded largely by governments. A real benefit of Open Data is not simply that single databases can be used more widely; it is that these data can also be leveraged, shared and combined with other data. Open Data facilitates scientific collaboration, enriches research and advances analytical capacity to inform decisions. In the human and environmental health realms, for example, the ability to access and combine diverse data can advance early signal detection, improve analysis and evaluation, inform program and policy development, increase capacity for public participation, enable transparency and improve accountability. However, challenges remain. Enormous resources are needed to make the technological shift to open and interoperable databases accessible with common protocols and terminology. Amongst data generators and users, this shift also involves a cultural change: from regarding databases as restricted intellectual property, to considering data as a common good. There is a need to address legal and ethical considerations in making this shift. Finally, along with efforts to modify infrastructure and address the cultural, legal and ethical issues, it is important to share the information equitably and effectively. While there is great potential of the open, timely, equitable and straightforward sharing of data, fully realizing the myriad of benefits of Open Data will depend on how effectively these challenges are addressed.

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).



Affiliation

¹ Office of the Chief Science Officer, Public Health Agency of Canada, Ottawa ON

*Correspondence:

victoria.edge@canada.ca

Suggested citation: Huston P, Edge VL, Bernier E. Reaping the benefits of Open Data in public health.

Can Commun Dis Rep 2019;45(10):252–6. <https://doi.org/10.14745/ccdr.v45i10a01>

Keywords: Open access, Open Science, Open Data, public health science, big data

Introduction

In June 2013, Canada and the other G8 countries adopted the G8 *Open Data Charter* (1). Open Data is part of a broad global movement that is not only advancing science and scientific communication, but is also transforming modern society and how decisions are made. This global movement is arguably one of the most important advances in evidence-based activities this century. Open Data has been defined as “structured data that is machine-readable, freely shared, used and built on without restrictions.” (2). The two main criteria for Open Data are that they must be freely available online and in a format that allows re-use.

This article provides a brief history of Open Data, explores its potential benefits, challenges and discusses the current state of Open Data in public health in Canada, with a focus on infectious diseases.

A brief history

Openness and sharing of discovery has been at the heart of science since the scientific method was first described by Aristotle (3). However, historically, neither scientific reports nor the data upon which these reports were based have been easily accessible. Scientific research was published in journals where access required paid subscriptions (or was a benefit of paid membership in an association), and databases were considered as the private and intellectual property of those who developed them. Databases were, and often still are, created and stored in different ways, analyzed by different methods, and thus can be deeply siloed.

In the 1970s, Robert Merton, who is considered the founder of the sociology of science, began advancing the idea that research should be freely accessible to all. He asserted that one “Mertonian norm” in the ethos of modern science was that



each researcher must contribute to the “common pot” and give up intellectual property rights to allow knowledge to move forward (4).

The Open Science movement was enabled by the rise of online journals in the 1990s, reflecting the original intent of science in supporting transparency and collaboration in research and scientific communication (5). The Open Science movement was driven by the observation that research was often paid with public funds, and thus taxpayers should not be restricted access to its outcomes by “paywalls”. This led to broad support and demand for open access to scientific publications and the current trend for authors and journals to adopt the Creative Commons license that enables people to freely read and use scientific publications with appropriate attribution (6). We are still in the midst of this transition, with both open access journals and subscription-based journals.

Supporters of the Open Science ethos went a step further by promoting more general access to generated or collected data. Open Data is based on the idea that not only should the results and reports of research be open, but also the underlying data that inform and support them. Nobel Prize winner, Elinor Ostrom, identified that Open Data was a new kind of “public good”. The thinking was that unlike other types of public good, the use of Open Data does not deplete the common stock, but potentially enriches it (7).

As with Open Science activities more broadly, the capacity to produce and share vast amounts of data soon took on a life of its own through enormous advances in technologies and computing. We are now in an age when the sheer volume of data generated daily is staggering (8). Necessarily, the demand for data storage capacity also keeps growing, with the ongoing evolution of new and more sophisticated data generators. Masses of data are increasingly available through digital platforms, wireless sensors, virtual-reality applications and billions of mobile phones (9). The trend towards Open Data is a global phenomenon, supporting opportunities and innovative trends in data analytics that include “big data”, artificial intelligence and machine learning. Increasingly, there is a call for data to be “open by default” and governments are increasingly including Open Data sets on their websites (10,11). The desire, demand and expectation for Open Data are becoming the new normal.

The potential of Open Data for public health

It has long been recognized that population health surveillance is one of the pillars of public health, yet the use and development of new technologies to collect, analyze and share surveillance data has been slow to develop, hindering the effectiveness of

informing public health policy and action (12). Open Data is one effective way to address the need to strengthen public health surveillance.

An early example of a strengthened public health surveillance system through the use of open data is the Behavioral Risk Factor Surveillance System (BRFSS). First developed in 15 states in the United States (US) in 1984, it now includes all US states and territories. Public health officials have used BRFSS for monitoring and responding to public health emergencies in real-time, such as developing the public health response to the effects of Hurricane Katrina in 2005 and monitoring the uptake of the H1N1 vaccine during the influenza pandemic in 2009. Currently, BRFSS data are integrated into the emergency response plan for drought-related threats to public health (13). It has been completely open access since 2014 (14).

Canada also has a number of online databases, including several maintained by the Public Health Agency of Canada (PHAC). The Public Health InfoBase (15), for example, offers easy-to-use tools for accessing and viewing public health data pertaining to chronic diseases, mental health, risk and protective factors and associated determinants of health. By using the search function and selecting criteria through drop down menus, users of the Public Health InfoBase can view data from different data sources in various formats.

In this issue of the *Canada Communicable Disease Report*, Totten et al. describe recent updates to the Canadian Notifiable Disease Surveillance System (CNDSS) and its interactive website (16). Established in 1924, the CNDSS is based on a federal/provincial/territorial collaboration that provides the latest data on key infectious diseases in Canada. Over the years, it has evolved to include an interactive public website that gives anyone the ability to easily create customized figures and tables on multiple diseases and to consider trends by age, sex and year. Currently, this information can be exported into PDF or Excel file formats, but soon it will be possible to download the databases into statistical software programs.

Another example is PulseNet Canada (PNC) run by the PHAC’s National Microbiology Laboratory (NML). This system highlights the successful development of high tech, advanced analytical science, providing real-time molecular surveillance and outbreak detection for foodborne disease, such as *Salmonella* and *Listeria* (17). The NML uses whole genome sequencing (wgs) technology for laboratory-based surveillance. PHAC currently is in the process of releasing all PNC-generated wgs data on outbreak strains originating in Canada to the National Centre for Biotechnology Information’s GenBank (18) online database. These efforts support Open Data and facilitate real-time data sharing with international, provincial and federal partners as well as industry to improve outbreak investigation, give insights into transmission patterns of emerging infections, and strengthen the One Health approach to surveillance.



An increasingly obvious benefit of Open Data is not simply that a single database can be used more widely; it is that these data can be leveraged, shared and combined with other data sets. This creates novel opportunities for scientific collaboration and partnership. For example, surveillance data on sexually-transmitted infections have been paired with data on the number of hits of public health messaging on social media sites to assess the effectiveness of infectious disease outbreak control (19). Open Data from satellites on weather and environmental indicators has been used to help predict increased risk of floods, fires and extreme weather events to trigger and inform mitigation efforts (20).

Some of the many potential benefits of Open Data in public health are summarized in the textbox below.

Textbox: Summary of the potential benefits of Open Data in public health

- Increases opportunities for scientific collaboration and partnerships
- Enriches research and analytical capacity
- Improves early detection of health and environmental threats
- Improves option analysis and monitoring real-time response
- Informs interventions and policy decisions
- Improves evaluation capacity and performance indicators
- Increases capacity for public participation
- Enables transparency and improves accountability

Challenges of Open Data

While the possibilities of Open Data are vast and promising, there are numerous challenges that need to be addressed to truly reap the benefits. They can be grouped into three key areas: making the technological shift; making the social and cultural shift that includes not only social norms, but also legal and ethical issues; and avoiding the pitfalls.

Making the technological shift

Open Data requires significant resources to set up databases for public use and combinability. Appropriate technological infrastructure is necessary, including software programs, high capacity computers and cloud-based solutions to store and analyze large amounts of data. Open Data also requires clear standards to ensure transparency regarding the source, how the data are generated, its combinability with other data and its limitations. Finally, there is a need for training to develop different types of expertise in systems and analytics. Some databases, such as the CNDSS, can easily generate quite simple

graphs and trends. However, with the use of more complex databases, the combining of databases or the use of large amounts of data, analytics has become more sophisticated and this requires development of analytical capacity.

Making the social and cultural shift

Although the call for Open Data began as a popular movement, there is still hesitancy in making some databases freely available. Not everyone wants to, or is able to, share their data. Developing excellent databases take a lot of time, work, resources and skill. If people share their hard-earned databases, will they get appropriate recognition? There has to be some motivation to spend time developing databases without the worry that its use will only enable others to get credit for the analysis and publication of those data. There is also the legitimate concern that open data could be used inappropriately, if the purpose for which the data was collected and the limitations of the data are not well-understood.

The hesitancy to share data is also often linked to legal and ethical issues. Who owns these data? Is there legislative support for data sharing? Especially with healthcare and public health databases, there are concerns about safeguarding privacy and confidentiality. There is a recognition that the call for openness and transparency needs to be tempered by the need to respect privacy and confidentiality. Generally, there are careful protocols for ensuring non-identifiability, but what if this is not done adequately, or the efforts to ensure confidentiality can be circumvented? This hesitancy highlights the need for clear standards and policies.

There is a concern about equity. Without the infrastructure capacity or expertise to access and make use of the data, is it really open to all? This also introduces a number of questions. What type and scope of data are being gathered? Whose interests are being prioritized? These and other aspects regarding equity will be explored during this year's International Open Access Week where the theme is "Open for Whom? Equity in Open Knowledge" (21). Equity is being addressed by international initiatives, such as the Open Government Partnerships that help to support scientists and other governments in less resource-rich environments (22).

Avoiding the pitfalls

There are two obvious pitfalls with the Open Data movement that need to be managed. The first is the need for common language, definitions, principles and tools—a common understanding of data management and best practices for data sharing agreements. This common approach is particularly important in situations where multiple disciplines are involved, where there are often different assumptions, different methodologies and practices, and when the same or similar terms can have different meanings.



Secondly, with so much focus on infrastructure, management and analytical capacity, there is a need to ensure that efforts are made to communicate the results of data-driven research effectively. With data creation growing at unprecedented rates, we are gathering more data than we can digest and deliver in an understandable way. For the analysis of open data to have optimal uptake, there is a need to advance ways of presenting data that will ensure that it is both succinct and understandable. With more and more data available, data are often combined from different disciplines, which means greater creativity in summarizing data—not only with tables and figures, but also visual abstracts, infographics, dashboards and more.

Discussion

Open Data represents a fundamental and massive shift in how we conduct research, make decisions, develop policy and evaluate our interventions. There is increasing pressure and expectation by the public for researchers and governments to show and share the data and information that public funds have generated. The potential benefits of making data open and accessible are very exciting; however, the challenges in making this happen are substantial and should not to be underestimated.

So where are we in terms of addressing the challenges and reaping the benefits of Open Data in public health? With respect to the technological shift, there has been a lot of progress, but appropriate technology and infrastructure is still being developed at all levels of government. Some areas of public health science, such as bioinformatics, are well ahead in current activities and in future planning for technologies and infrastructures. Other areas are less well developed. In addition, a socio-cultural shift is still underway and there remain those who are still hesitant to share their data.

Addressing concerns around legal and contractual obligations will require careful and considered legislative change in some domains. For example, a recent federal plan to advance Open Data identified the need to update the *Statistics Act* (23). For public health, specifically, work is underway to balance Open Data with regulatory limitations, and address privacy and confidentiality concerns. In avoiding the numerous potential pitfalls, developing a common language and applying best practices in data sharing, we are in the early days. In Canada, the *Multilateral Information Sharing Agreement* (MLISA) is likely to be a landmark document that identifies best practices for the sharing of public health surveillance information amongst the federal, provincial and territorial governments; however, the details of this agreement are still being advanced (24). For example, MLISA includes appropriate attribution, which is a hallmark of the Creative Commons license, but which has not been a widespread feature of Open Data. The MLISA also includes safeguards to promote and ensure appropriate use

of data. These features have gone a long way to address the concerns of those who created the databases that their work will be acknowledged and used appropriately.

In terms of effectively communicating the results, a lot of progress has been made since the early days when data sets were simply placed on the web with little explanation. Although there has been a perennial need to make scientific communications accessible, this need becomes even more acute with the data revolution that is currently underway. We need to find more ways to summarize data and make the key messages evermore succinct and memorable.

With Open Data still very much in development in public health, what are the next steps? When considering the increased use of Open Data demands balanced against limited resources, there is a need to better understand the type, degree of uptake and use of Open Data. Good, reliable and freely accessible public health data could be useful to students and researchers (undergraduate to postdoctoral), federal, provincial and territorial governments, non-profit organizations, healthcare and public health professionals, as well as journalists. The idea of the “public good” derived from Open Data is attractive in principle, but is it actually being used and to what extent? It would also be interesting to assess if more access to health data increases engagement in personal and public health. Further to this, innovative projects that ask for public support and involvement in open data generation or analysis, through activities such as crowd-sourcing (25) or hack-a-thons (26), could extend the reach and resources of public health.

Conclusion

Technologies and science will continue to contribute to the explosive generation of data. The possibilities that these data create have captured the scientific imagination. The global trend to embrace Open Science and Open Data reflects the inherent desire by many to work collaboratively to address complex issues, recognizing the benefit of multiple perspectives, the leveraging of resources, the advancement of research methodologies and the benefits of timely, robust data to inform decisions made in many domains. Public health has started to reap the many anticipated benefits that openness and transparency of data present; and work continues to address the significant challenges involved in making a successful transition towards this “new normal”. Stay tuned.

Authors' statement

All authors conceptually developed the manuscript together. PH wrote the first draft, VLE and EB contributed to subsequent drafts and all signed off on the final version. P Huston was the Editor-in-Chief of the *Canada Communicable Disease Report* at the time this was written, but recused herself from all



editorial decisions on the manuscript. M Deilgat, the incoming Editor-in-Chief took the editorial decisions on the manuscript.

Conflict of interest

None.

References

1. Government of Canada. Canada's Action Plan on Open Government 2014-16. <https://open.canada.ca/en/content/canadas-action-plan-open-government-2014-16>
2. Government of Canada. Open Data 101. <https://open.canada.ca/en/open-data-principles>
3. Barnes J, editor. Complete Works of Aristotle. 6th printing (with corrections). Princeton, NJ: Princeton University Press; 1995.
4. Merton RK. The Sociology of Science: Theoretical and Empirical Investigations. Chicago, IL: University of Chicago Press; 1973.
5. Gezelter D. The OpenScience Project. What, exactly, is Open Science? July 28, 2009. <http://openscience.org/what-exactly-is-open-science>
6. Commons C. About the Licenses. <https://creativecommons.org/licenses/>
7. Ostrom E. Governing the Commons. The Evolution of Institutions for Collective Action. New York, NY: Cambridge University Press; 1990. DOI
8. IBM. 2.5 quintillion bytes of data created every day. How does CPG & Retail manage it? Industry Insights. <https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/>
9. McKinsey Global & Company. The Age of Analytics: Competing in a data-driven world. McKinsey Global Institute: 2016. <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world>
10. European Commission. Open Data. Digital Single Market. <https://ec.europa.eu/digital-single-market/en/open-data>
11. Centers for Disease Control and Prevention. Data Sets. <https://open.cdc.gov/data.html>
12. Canada's Fundamental Science Review. Investing in Canada's Future. Strengthening the Foundations of Canadian Research. April 10, 2017. <http://www.sciencereview.ca/eic/site/059.nsf/eng/home>
13. Centers for Disease Control and Prevention. Office of Surveillance, Epidemiology and Laboratory Services. The Behavioral Risk Factor Surveillance System: History (fact sheet) (Accessed 09-06-2019). <https://www.cdc.gov/brfss/factsheets/pdf/brfss-history.pdf>
14. Centers for Disease Control and Prevention. 2017 BRFSS Survey Data and Documentation. https://www.cdc.gov/brfss/annual_data/annual_2017.html
15. Government of Canada. Public health Infobase. <https://health-infobase.canada.ca/>
16. Totten S, Medaglia A, McDermott S. Updates to the Canadian Notifiable Disease Surveillance System and its interactive website. *Can Commun Dis Rep* 2019;45(10):257-61. DOI
17. Public Health Agency of Canada. PulseNet Canada. PHAC; modified May 20, 2019. <https://www.canada.ca/en/public-health/programs/pulsenet-canada.html>
18. National Center for Biotechnology Information. GenBank Overview. <https://www.ncbi.nlm.nih.gov/genbank/>
19. Sachdeva H, Benusic M, Ota S, Stuart R, Maclachlan J, Dubey V, Andonov A. Community outbreak of hepatitis A disproportionately affecting men who have sex with men in Toronto, Canada, January 2017–November 2018. *Can Commun Dis Rep* 2019;45(10):262-8. DOI
20. Kotchi SO, Bouchard C, Ludwig A, Rees EE, Brazeau S. Using Earth observation images to inform risk assessment and mapping of climate change-related infectious diseases. *Can Commun Dis Rep* 2019 May;45(5):133-42. DOI PubMed
21. Scholarly Publishing and Academic Resources Coalition. Theme of 2019 International Open Access Week. <https://sparcopen.org/news/2019/theme-of-2019-international-open-access-week-to-be-open-for-whom-equity-in-open-knowledge/>
22. Open Government Partnership. Launching the Community of Practice for Water and Open Gov. <https://www.opengovpartnership.org/>
23. Privy Council Office. Report to the Clerk of the Privy Council: A Data Strategy Roadmap for the Federal Public Service. Government of Canada. <https://www.canada.ca/en/privy-council/corporate/clerk/publications/data-strategy.html>
24. Pan-Canadian Public Health Network. Partners in Public Health. Multilateral Information Sharing Agreement (MLISA). 2014. <http://www.phn-rsp.ca/pubs/mlisa-eng.pdf>
25. Brabham DC, Ribisl KM, Kirchner TR, Bernhardt JM. *Am J Preventative Med*. 2014; 24(2):179-87. <https://www.sciencedirect.com/science/article/abs/pii/S0749379713005898>
26. Consortium for Affordable Medical Technologies. Massachusetts General Hospital Global Health. Opioid Challenge Summit and Hack-a-thon. Sept 9-11, 2016. Boston MA. <http://www.globalhealthmgh.org/camtech/wp-content/uploads/2016/10/OpioidEpidemicHack-Report-FINAL.pdf>