Gail M. Thornton
School of Library and Information Studies, University of Alberta, Edmonton, Alberta, Canada

Ali Shiri
School of Library and Information Studies, University of Alberta, Edmonton, Alberta, Canada

CANADA’S HEALTH DATA REPOSITORIES: CHALLENGES OF ORGANIZATION, DISCOVERABILITY AND ACCESS (Paper)

Abstract:

Evidence-based medicine relies on health data. Open health data initiatives need to be supported with data repositories that are optimized for searchability and discoverability. Five Canadian open health data repositories were evaluated for metadata-based searching, browsing, and navigational functionalities. In the different national, provincial and institutional open data repositories evaluated, the preliminary findings reveal the variability in these functionalities. This research will contribute to the development of guidelines and best practices for implementing metadata and searching and browsing functionalities for open health data repositories that will ultimately lead to a more interoperable open health data environment.

1. Introduction

Open health data gives health care professionals, biomedical researchers and the general public access to health data that can improve health care and affect health policy (Greenburg & Narang, 2016). Since 2009, the Government of Canada has implemented initiatives promoting open access to government data which includes health data (Government of Canada, 2017). In order to benefit most from these information resources, the open health data must be easy to search and retrieve which depends on the quality of its metadata, i.e. the data about the data (Martin et al., 2017). Identification and documentation of metadata practices and formats for open data repositories will ensure creation of a solid basis upon which subject and semantic interoperability can be implemented. Various interoperability models reported in the literature are supported by identifying useful metadata elements and practices (Nicholson and Shiri, 2003; Hafezi, et al., 2010). The first generation of digital libraries, open archives and content management systems required similar approaches as for open data repositories where the variety of disciplines involved and the vastness of open data necessitate a more systematic and holistic approach to metadata.

One of the key challenges associated with accessing and making effective use of open health data repositories is the identification, ease of access, and discoverability of these emerging digital data environments. As of January 16, 2019, the Directory of Open Access Repositories (OpenDOAR) lists 359 repositories in “Health and Medicine” (OpenDOAR, n.d.-a) which represents a 40% increase in almost 5 years from the 254 repositories listed in April of 2014 (Loan & Sheikh, 2016). Even with existing data directories and registries, discovery and access to subject-specific data repositories such as health and medicine remains a challenge. For instance, Google introduced its Dataset Search service on September 5, 2018 to alleviate some of these issues (Noy, 2018). However, Google's Dataset Search is very limited in its search and retrieval functionalities and makes minimal use of metadata to allow filtering and browsing of the datasets indexed: title, date of publication, data provider, time period covered, description.

Repositories with open access to health data were evaluated where the health data could include administrative, statistical and research data. This paper reports on an evaluation of the discoverability of and information organization and access mechanisms within Canada's open health data repositories. More specifically, the extent to which Canadian open health data repositories provide metadata-based searching, browsing and navigational functionalities to support searchability and discoverability of the data was evaluated. Some of the key questions addressed include: What is the number and nature of metadata elements available in Canada's open health data repositories? What metadata elements are common and different across these repositories? Are there open health data repositories that follow metadata best practices, such as compliance with metadata and interoperability standards, ease of access for a broad range of users, including researchers, practitioners and the general public?

This paper is well aligned with the theme of the CAIS 2019 conference as it addresses a number of areas including the movement of data, such as research data, in contemporary modes of communication, in this case data repositories that focus on health, as well as the exploration and transmission of health data available to improve health and well-being.

2. Methods

2.1 Identifying Repositories

In order to identify open health data repositories in Canada, a number of directories and registries of data repositories were consulted. Initially, OpenDOAR, a global directory of open access repositories was consulted (OpenDOAR, n.d.-b). Using OpenDOAR's filter mechanism, the collection was filtered using "Subjects: Health and Medicine" and "Countries and Regions: Canada". This exercise retrieved only four repositories. Further filtering using "Content Types: Datasets" revealed only one resource: SUMMIT from Simon Fraser University (SFU) in Burnaby, British Columbia (BC). Upon further examination of the SUMMIT repository, no datasets were found under either "Health" or "Medicine". Interestingly, SFU now has a data repository called SFU Radar to complement SUMMIT. SFU Radar contained no datasets or even any entries under "Health Science" or "Kinesiology". Clearly OpenDOAR's directory lacks Canadian-focused data repositories, let alone more specific health data repositories. Next, the Federated Research Data Repository (FRDR) was consulted. The FRDR is a collaboration of the Canadian Association of Research Libraries (CARL) / Portage and Compute Canada (Garnett et al., 2017). Currently, FRDR has 44 collaborating repositories (FRDR, 2017). These repositories include open access data from federal government (e.g. Government of Canada), provincial government (e.g. Alberta, BC), municipal government (e.g. City of Edmonton) and institutions (e.g. University of Alberta Libraries Dataverse, SFU Radar). While these repositories are not specifically Health and Medicine focused ones, some may contain related data (e.g. Government of Canada); however, some may not at this time (e.g. SFU Radar).

2.2 Selecting Repositories

Canadian health organizations operate at various levels, including municipal, provincial, territorial and national. Comprehensive examination of open health data repositories should be conducted across these various levels to gain a more inclusive and comparative perspective of information organization and accessibility approaches. Initially, five repositories were selected to evaluate, namely two national, two provincial and one institutional. The following are the selected repositories:

- Government of Canada Open Government Portal
- Canadian Institute for Health Information (CIHI)
- Government of Alberta Open Data Portal
- British Columbia (BC) Data Catalogue
- University of Alberta Libraries (UAL) Dataverse

2.3 Evaluation methods

Searching and browsing functionalities of the repositories were evaluated using similar methods used to evaluate medical digital libraries (Ismond & Shiri, 2007). Metadata were evaluated using selected parameters from a previous study on metadata quality in United States open health data repositories (Martin et al., 2017). In particular, the identified and selected Canadian open health data repositories were examined and compared in terms of their searching, browsing and navigation functionalities, the richness of metadata description practices, as well as their metadata-based filtering mechanisms. This examination provides an evidence-based approach to the assessment of the discoverability and access of the repositories.

3. Preliminary findings

The following provides some basic findings about coverage of open health data in various constituencies. Repositories were evaluated on January 16, 2019.

3.1 Filtering health data repositories

Using the Government of Canada Open Government Portal (Government of Canada, n.d.-a), the repository was filtered using “Portal Type: Open Data”, “Subject: Health and Safety” and “Resource Type: Dataset” which retrieved 1,196 records. Of these 1,196 records, 335 records were attributed to the Province of Alberta which was the only provincial jurisdiction represented as indicated by the filter “Jurisdiction: Provincial (335)”. Also, 3 records were attributed to the Canadian Institute for Health Information (CIHI) of the 1,196 records.

CIHI is an independent, not-for-profit organization providing information on the health of Canadians and Canadian health systems (CIHI, n.d.-a). Lucyk et al. (2015) indicate that CIHI is integral to Canada's position as a global leader in administrative health data science. The "Access Data and Reports" page lists the following filters: primary theme, geography, content format, published date (CIHI, n.d.-b). No filter was provided to separate data from reports; however, if only XLSX, XLS, and ZIP "content formats" were considered, then 230 records remained.

Using the Government of Alberta Open Data Portal (Government of Alberta, n.d.-a) and filtering using "Topic: Health and Wellness" resulted in the user being re-directed to the "All Resources" page from the "Open Data" page. This required the extra step of filtering "Information Type: opendata" to get the 358 records (Government of Alberta, n.d.-b). BC Data Catalogue (BC Data Catalogue, n.d.-a) was filtered using "Sectors: Health and Safety", "Dataset types: Datasets" and "Download Permission: Public" to retrieve 66 datasets. (BC Data Catalogue, n.d.-b).

The University of Alberta Libraries (UAL) Dataverse (UAL Dataverse, n.d.-a) was filtered for "Datasets" and "Subject: Medicine, Health and Life Sciences" which retrieved 55 records (UAL Dataverse, n.d.-b).

3.2 Evaluating Repositories

Table 1 shows a comparative table of the various searching, browsing and metadata elements for the selected five open health data repositories. The five data repositories have basic search; however, only UAL Dataverse has advanced search functionality. The browsing options are only the results list for the Government of Canada, Government of Alberta, BC Data Catalogue and UAL Dataverse repositories. CIHI offers 3 "Frequently accessed", 3 "Recently accessed" and 20 "Themes" for browsing their repository.

Table 1: Comparison of searching, browsing, and metadata elements in five repositories.
The text in italics in the Sorting column indicates the default.

Repository	Facets (Filters)	Browsing	Sorting	Metadata on Results List	Metadata on Record
Government of Canada Open Government Portal	Portal Type Collection Type Jurisdiction Organization Keywords Subject Format Resource Type Maintenance and update frequency	Results List	*Relevance* Name ascending Name descending Last modified	Title (link to record) Jurisdiction Description Organization Issued by (Jurisdiction) Resource Formats	Title Description Publisher – Current Organization Name Licence Resources Resource Name Resource Type Format Language Links [Access] (button to download) Additional Information Contact Email Keywords Subject Maintenance and Update Frequency Date Published Date Modified Openness Rating About this Record Record Released Record Modified Record ID Metadata Link to JSON format

					DCAT (JSON-LD) DCAT (XML)
Canadian Institute for Health Information (CIHI)	Primary theme Geography Content format Published date	Frequently accessed Recently accessed Themes	*Relevance* Date	Title (link to download) Date Description Tags (including) Primary theme Geography Content format	
Government of Alberta Open Data Portal	Information Type Topics Publisher Formats Audience Publication Type Date Added to Catalogue	Results List	Date last updated ascending *Date last updated descending* Date added to portal ascending Date added to portal descending Publication date ascending Publication date descending Title ascending Title descending Last Modified Relevance	Title (link to record) Information Type Formats Views Last Modified Description Tags	Title *Summary Tab* Description Tags Resources Resource name (link to download) [More Information] or [Preview] [Download] (button to download) downloads *Detailed Information Tab* Title and Dataset Information Alternative Title Date Modified Update Frequency Publisher/Creator Information Creator Publisher Subject Information Topic Start Date End Date Spatial Coverage Resource Dates Date Created Date Added to Catalogue Date Issued Date Modified Audience information Language Identifiers Usage/Licence Usage Considerations License Contact Contact Name Contact Email *Related Tab* list/link related records
British Columbia (BC) Data Catalogue	License Sectors Dataset Types Formats Organizations Download permission	Results List	Relevance Popular Name Ascending Name Descending *Published Date* Last Modified	Title (link to record) Dataset Types Sectors Formats Description Record Published	Title Dataset Types Sectors Views Published by Licensed under Description Tags Activity Stream Data and Resources Filename (file size) [Explore > Preview or Download] Additional Information Data Quality Lineage Statement More Information Contact Information Name Email Organization Suborganization Access & Security Who can view this dataset?

					Who can download this dataset? Metadata Information Record Published Record Last Modified Resource Status
University of Alberta (UAL) Libraries Dataverse	Metadata Source Publication Date Author Name Subject Keyword Term Deposit Date	Results List	Name (A-Z) Name (Z-A) *Newest* Oldest	Title (link to record) Metadata Source Publication Date Citation Description	Title Version Citation [Cite Dataset] Description Subject Keyword Related Publication *Files Tab* Search bar Number of Files Filename file format, file size, date, downloads [Download] (button to download) *Metadata Tab* [Export Metadata] Dublin Core DDI JSON Schema.org JSON-LD Citation Metadata Dataset Persistent ID Title Alternative title Other ID Author Contact Description Subject Keyword Related Publication Producer Production Date Production Place Grant Information Time Period Covered Date of Collection Kind of data Software Geospatial Metadata Social Sciences and Humanities Metadata Life Sciences Metadata *Terms Tab* Terms of Use Restricted Files + Terms of Access Guestbook *Versions Tab* Dataset Summary Contributors Published

The number of facets for filtering varies greatly with four facets for CIHI and nine for Government of Canada (Table 1). Two repositories used the term *subject* for filtering (Government of Canada and UAL Dataverse). The two provincial government repositories do not use the term subject but rather “Topics” (Alberta) or “Sectors” (BC). Four repositories filter using *format*, three filter using *type* and three filter using *publisher* (Table 1).

All of the repositories except UAL Dataverse included relevance ranking as a sorting option for the results list. While Government of Canada and CIHI had relevance sorting as the default, the remaining repositories employed a descending date-based default sorting (Table 1).

All repositories included the *title* and *description* in the metadata on the results list (Table 1). For all of the repositories except CIHI, the title in the results list was hyperlinked to the record. CIHI titles in the results list were hyperlinked to the file download. Interestingly, *title* and *description* are consistent across all five repositories in the results list. An interoperable interface could be provided to search across these five repositories.

Three repositories explicitly refer to metadata on the record (Table 1): Government of Canada, BC Data Catalogue, UAL Dataverse. Under “Metadata Information”, BC Data Catalogue refers to published and modified dates for the record and status of the resource. With greater effort to address metadata on the record, Government of Canada (under “Metadata”) and UAL Dataverse (under “Export Metadata”) provide links to export metadata in different standards (Table 1): three for Government of Canada and four for UAL Dataverse. This suggests that some priority was placed on clarifying, within the record itself, the use of metadata standards and support for interoperability. Given the role of metadata for searchability, findability and discoverability of open data, the implementation of explicit and easy to find mechanisms to access metadata in data repositories is particularly emphasized.

4. Conclusion

This study examined and compared five Canadian open health data repositories across institutional, provincial and national levels in terms of information access and metadata practices. These preliminary findings will improve the understanding among researchers, librarians and data managers of the application of metadata in open health data repositories as well as the challenges associated with finding and discovering open health data. This research will contribute to the development of guidelines and best practices for developing and implementing metadata for open health data repositories in order to pave the way for an interoperable open health data environment. Further research and development in this area will be guided by considering a framework accounting for preservation infrastructures, unique identifiers, interoperability architecture and the definition of a set of data-specific metadata.

Reference List:

BC Data Catalogue. (n.d.-a). Datasets. Retrieved from https://catalogue.data.gov.bc.ca/dataset

BC Data Catalogue. (n.d.-b). Datasets. Retrieved from https://catalogue.data.gov.bc.ca/dataset?type=Dataset&sector=Health+and+Safety&download_audience=Public

CIHI. (n.d.-a). About CIHI. Retrieved from https://www.cihi.ca/en/about-cihi

CIHI. (n.d.-b). Access Data and Reports. Retrieved from https://www.cihi.ca/en/access-data-and-reports

FRDR. (2017). Canadian Research Data Repositories. Retrieved from https://www.frdr.ca/discover/html/repository-list.html?lang=en

Garnett, A., Leahey, A., Savard, D., Towell, B., & Wilson, L. (2017). Open metadata for research data discovery in Canada. *Journal of Library Metadata*, 17(3-4), 201-217. doi:10.1080/19386389.2018.1443698

Government of Alberta. (n.d.-a). Open Data. Retrieved from https://open.alberta.ca/opendata

Government of Alberta. (n.d.-b). All Resources. Retrieved from https://open.alberta.ca/dataset?dataset_type=opendata&topic=Health+and+Wellness

Government of Canada. (2017, December 19). Open Data 101. Retrieved from https://open.canada.ca/en/open-data-principles#toc95

Government of Canada. (n.d.-a). Open Government Portal. Retrieved from https://open.canada.ca/data/en/dataset?q=

Government of Canada. (n.d.-b). Open Government Portal. Retrieved from https://open.canada.ca/data/en/dataset?portal_type=dataset&q=&_subject_limit=0&_organization_limit=0&subject=health_and_safety&res_type=dataset&_res_type_limit=0

Greenberg, C. J. & Narang, S. (2016). World Health Organization member states and open health data: An observational study. *Epidemiology Biostatistics and Public Health*, 13(3). doi:10.2427/11950

Alipour-Hafezi, M., Horri, A., Shiri, A. & Ghaebi, A. (2010). Interoperability models in digital libraries: An overview. *The Electronic Library*, 28(3), 438-452.

Ismond, K. P. & Shiri, A. (2007). The medical digital library landscape. *Online Information Review*, 31(6), 744-758. doi:10.1108/14684520710841748

Loan, F. A., & Sheikh, S. (2016). Analytical study of open access health and medical repositories. *Electronic Library*, 34(3), 419-434. doi:10.1108/EL-01-2015-0012

Lucyk, K., Mingshan Lu, Sajobi, T., & Quan, H. (2015). Administrative health data in Canada: lessons from history. *BMC Medical Informatics & Decision Making*, 15(1), 1-6. doi:10.1186/s12911-015-0196-9

Martin, E.G., Law, J., Ran, W., Birkhead, G.S., & Helbig, N. (2017). Evaluating the quality and usability of open data for public health research: A systematic review of data offerings on 3 open data platforms. *Journal of Public Health Management and Practice*, 23(4), e13. doi:10.1097/PHH.0000000000000388

Nicholson, D. & Shiri, A. (2003). Interoperability in subject searching and browsing. *OCLC Systems & Services*, 19(2), 58-61.

Noy, N. (2018, September 5). Making it easier to discover datasets. Retrieved from https://www.blog.google/products/search/making-it-easier-discover-datasets/

OpenDOAR. (n.d.-a). Subjects matches any of "Health and Medicine". Retrieved from http://v2.sherpa.ac.uk/cgi/search/repository/advanced?screen=Search&repository_name_merge=ALL&repository_name=&repository_org_name_merge=ALL&repository_org_name=&content_types_merge=ANY&content_subjects=10&content_subjects_merge=ANY&org_country_browse_merge=ANY&satisfyall=ALL&order=preferred_name&_action_search=Search

OpenDOAR. (n.d.-b). Directory of Open Access Repositories. Retrieved from http://v2.sherpa.ac.uk/opendoar/

University of Alberta Libraries (UAL) Dataverse. (n.d.-a). UAL Dataverse. Retrieved from https://dataverse.library.ualberta.ca/

University of Alberta Libraries (UAL) Dataverse. (n.d.-b). UAL Dataverse. Retrieved from https://dataverse.library.ualberta.ca/dataverse/ualib?q=&fq0=subject_ss%3A%22Medicine%2C+Health+and+Life+Sciences%22&types=datasets&sort=dateSort&order=desc