Research Data Discovery and the Scholarly Ecosystem in Canada

A White Paper

Prepared by the Portage Network, Data Discovery Working Group on behalf of the Canadian Association of Research Libraries (CARL)

Eugene Barsky (University of British Columbia)
John Brosz (University of Calgary)
Amber Leahey (Scholars Portal, Ontario Council of University Libraries)

JULY 2016

Portage Network
Canadian Association of Research Libraries
portage@carl-abrc.ca

www.carl-abrc.ca

portage
SERVICES PARTAGÉS POUR LES DONNÉES DE RECHERCHE
SHARED STEWARDSHIP OF RESEARCH DATA
CARL ABRC
CANADIAN ASSOCIATION OF RESEARCH LIBRARIES
ASSOCIATION DES BIBLIOTHÈQUES DE RECHERCHE DU CANADA

Table of Contents

Executive Summary

Research data must be discoverable to be re-used. Data discovery represents the descriptive and technical processing of data and metadata, as well as the tools and infrastructure aimed at improving access and reuse of research data on the web. A Canadian data discovery service would make it easier to find and reuse research data held in institutional and disciplinary repositories. We would like to see a service that provides a coherent, single point of access to authoritative, searchable, browsable, and machine actionable descriptions (metadata) for datasets and implements clear means for accessing them, thus increasing the likelihood of discovery and reuse of research data in Canada.

In this paper, we highlight current opportunities and issues related to developing such a service in Canada. Based on a review of international and national research data repositories and data discovery services, we offer a set of guiding principles, best practices, and recommendations for data discovery:

Common metadata: the descriptive information that accompanies research data should meet minimum standards to enable discovery and support data reuse. This requires a commitment to a core set of metadata components across domains. Metadata tools should accommodate multiple, overlapping metadata namespaces, i.e., descriptive terms assigned, managed, and grouped into collections of classes and attributes. We also recommend building separate, flexible metadata harvesters for indexing specialized repositories, so that domain-specific metadata and granularity can be retained in its original format.

Persistent Identification: the use of global identifiers for researchers and research data. We recommend exploring a national ORCID agreement so that universities and government agencies in Canada can integrate researcher identifiers into institutional and other research management and publishing software. We also recommend registering DOIs corresponding to datasets in participating repositories with DataCite Canada. These DOIs will greatly enhance dataset discoverability via DataCite's metadata partners (e.g. ORCID, VIVO, etc.).

Open Access and Programmatic Interfaces: the use of an application program interface (API) allowing one piece of software to make use of the functionality or data available to another through a set of routines, protocols, and tools. Metadata and data should be programmatically accessible for reuse and development purposes through the provision of APIs among participating repositories and data discovery platforms.

Common licensing: policies and licenses should govern access to data and metadata and, whenever possible, should be minimally restrictive. We recommend the use of Creative Commons licenses for research data as they effectively communicate information about the copyright holders' intentions and clarify usage permissions. Licensing can apply to data and metadata, although we strongly recommend that metadata be provided as openly as possible, with minimal to no restrictions on reuse in order to facilitate discovery.

Collaboration: a joint commitment to shared recognition and cooperation among actors, organizations, data producers, and researchers, sometimes described as "coexistence in the scholarly ecosystem." We emphasize that collaboration will drive improvements for data discovery in Canada. A well-coordinated national project will ensure that all attempts to improve discovery and access to data will be informed and facilitated by stakeholder expectations, participation, and collaboration. Keeping stakeholders engaged and providing clear communication channels are key for the success of a national data discovery service.

This paper is presented with a common goal to make research data as widely discoverable and accessible as possible, thus enhancing opportunities for data reproducibility and reuse. Enhancing data discovery is one approach to facilitating greater interoperability and discovery of scholarly outputs. Building national infrastructure to support research data discovery will greatly enhance opportunities for further integration across the scholarly ecosystem, including support for metadata, global identifiers, and open APIs.

Introduction

Research data must be discoverable to be re-used. Growing pressures and interests to make data more widely available have heightened the need to provide new and improved ways of finding existing research data. While the ability to verify research findings has always been considered a central principle to good scientific practice, there are now increasing calls for openness in research to improve communication, data sharing, and reuse among researchers, most recently by the Canadian Tri-Agencies[1].

[1] Government of Canada (2016). Tri-Agency Statement of Principles on Digital Data Management. Government of Canada, accessed June 15, 2016 from http://www.science.gc.ca/default.asp?lang=En&n=547652FB-1

Data discovery in this paper represents the descriptive and technical processing of data and metadata, and the tools and infrastructure aimed at improving access and reuse of research data on the web. This paper is not meant to be an exhaustive history or review, instead it seeks to summarize a variety of research data repositories and national data discovery services, to inform discussions and offer a set of guiding principles about standards and best practises related to data discovery. It is hoped that this discussion sheds light on current opportunities and issues related to developing such a service in Canada. Repositories and developing data services can leverage this set of principles and recommendations to enhance data discovery.

In Canada, a national data discovery service would help to make research data held in institutional and disciplinary repositories more discoverable by others through aggregating metadata about data collections or datasets. We would like to see a service that will not act as a mega-repository for the datasets themselves. Rather, it would be aimed to increase the likelihood of discovery and reuse of research data in Canada by providing a coherent point of access to authoritative, searchable, browsable, and actionable descriptions (metadata) for datasets and how to access them.

Adherence to data discovery principles entails a holistic approach to bringing together disparate sources of data, that by virtue of the data silo effect, are often spread out across different institutional and disciplinary repositories, organizations, and collections of data. Poor data citation practices in published research results (e.g. journal articles, reports) and access to underlying research data for example, have been identified as a significant barrier to scientific verification and replication across a variety of disciplines[2] [3]. Calls for improved data identification, data sharing, and linkages between research data and publications focus on improved management of research data in the scholarly research ecosystem.

Principles of data discovery

Through an evaluation of platforms and standards utilized by data repositories and data discovery services, several important themes emerge. Data discovery is most often related to opening up data, including a commitment to open metadata to enable data discovery and reuse on the web. This is only accomplished through interactions between a variety of actors and is enacted through different

[2] King, G. (1995). Replication, replication. *PS: Political Science & Politics*,*28*(03), 444-452.

[3] Yong, E. (2012). Replication studies: Bad copy. *Nature*,*485*(7398), 298-300. doi:10.1038/485298a

mechanisms (see Figure 1). The following is presented as a set of principles that we believe encompass exemplary data discovery practices.

Set of Principles:

1. Common metadata
2. Persistent identification
3. Open Access
4. Common licensing
5. Collaboration (coexistence in the scholarly ecosystem)

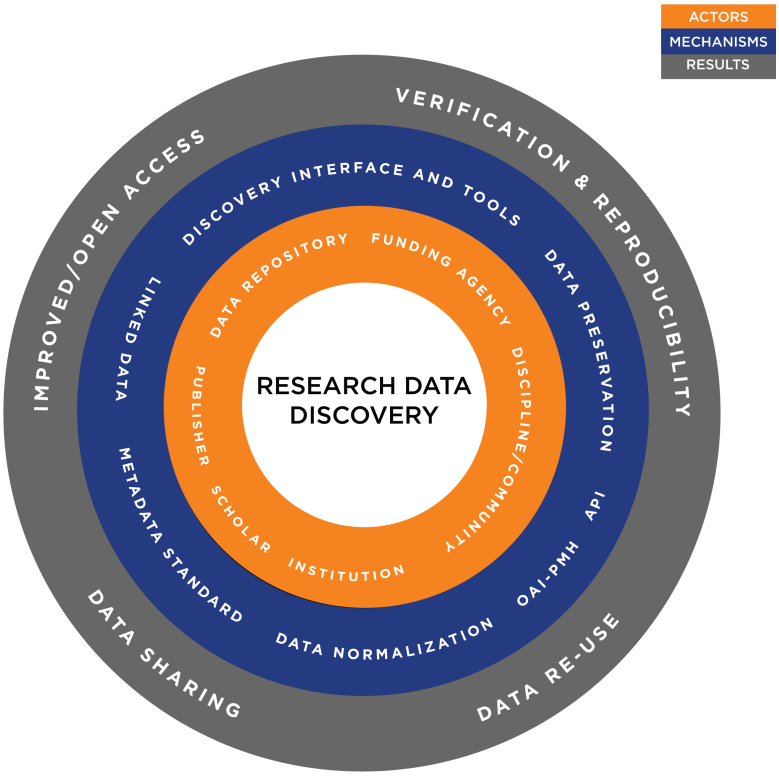

Figure 1 - Research Data and the Scholarly Research Ecosystem

Common metadata refers to the set of descriptive information that accompanies research data. Metadata should meet minimum standards as necessary to enable adequate discovery and to support research data reuse across a variety of disciplines. This does not necessarily require the use of the same standards across domains, but rather, the commitment to a core set of metadata components that enable discovery and reuse.

Persistent identification of data encompasses the use of global identifiers for researchers and research data, to enable data publication in the scholarly research ecosystem (e.g. DOIs, ISNI, ORCID).

Open access of data and metadata borrows from other openness principles (e.g. OECD[4]), such as, access at the lowest possible cost and access that is easy, timely, user-friendly, and preferably web-based. Extending on this principle, open access in discovery systems should be built for both human understanding and machine actionable purposes (e.g. access via APIs).

Common licensing covers the variety of conditions related to sharing, access, and use of metadata and data. Whenever possible, policies should be established to apply licensing to data and metadata that is minimally restrictive. The establishment of national and international data policies around data sharing and reuse will greatly improve efforts to develop shared and common licensing for data across a variety of disciplines. The use of Creative Commons CC0[5] for data and metadata, for example, may be used by anyone who wishes to provide data openly without restrictions.

Collaboration or rather coexistence in the scholarly research ecosystem entails a commitment to shared recognition and cooperation among actors, organizations, data producers, researchers, and so on.

Overall, the purpose of this paper and the presentation of data discovery principles is to bring together key scholarly actors and organizations around a common goal. Foundational to this is making research data as widely discoverable and accessible as possible, enhancing opportunities for research reproducibility and data reuse, and facilitating new knowledge discovery in Canada. Enhancing data discovery is one approach to facilitating greater interoperability and discovery of scholarly outputs, including research data, throughout the entire scholarly ecosystem.

Building a National Data Discovery Service in Canada

Multidisciplinary research requires access to digital data and information from a variety of data sources, communities, and repositories. Demand for data and metadata aggregation requires robust data infrastructure and tools to enable cross-disciplinary research. Examples of aggregating services and federated catalogues include the OCLC OAIster service[6], Europeana Portal[7], the Digital Public Library of

[4] OECD Council Recommendation on Access to Research Data from Public Funding (2006), accessed June 9, 2016 from http://www.oecd.org/sti/sci-tech/38500813.pdf

[5] Creative Commons CC0, accessed June 9, 2016 from https://wiki.creativecommons.org/wiki/CC0_FAQ#What_is_CC0.3F

[6] OCLC OAIster, accessed June 9, 2016 from http://www.oclc.org/oaister.en.html

America (DPLA)[8], and Blacklight[9] - an online public access catalogue framework, to name a few.

The following is a selection of national data discovery aggregating services to enable increased findability and improved access to research data. Currently, there is no central point of access for research data and related resources in Canada, thus it is useful to evaluate approaches adopted elsewhere before such a service is developed here.

Table 1 - National Data Discovery Aggregating Services Comparison[10]

Country/ Region	Provider Name	Service Name	Service Model	Providers / Data Sources	Metadata / Standards	Cost Model / Maintenance
United States	Association of Research Libraries (ARL) & Center for Open Science (COS)	SHARE[11], SHARE Notify	Notification service for research release events (publications, data management plans, presentations, research data), with developer options for database access. SHARE Notification service: Atom-XML feed, Search and Browse Tool, and JSON API.	Over 100 providers (arXiv, CrossRef, PubMed Central, other repositories).	VIVO, ORCID, DataCite, OAI-PMH Dublin Core	Grant funded (Institute of Museum and Library Services (IMLS), & Alfred B. Sloan Foundation), 2016-2017 Curation Associates Program for library professionals as a training and maintenance model

[7] Europeana Portal, accessed June 9, 2016 from http://www.europeana.eu/portal/

[8] DPLA, accessed June 9, 2016 from https://dp.la/

[9] Blacklight Project, accessed June 9, 2016 from http://projectblacklight.org/

[10] The table provides a structured set of information to understand the various initiatives and services being developed across five different national/regional services. It borrows from a variety of sources, mainly online and published, including the Data Service Infrastructure for the Social Sciences and Humanities (DASISH), 2012 *Report, Appendix: Data Archive Description Sheets (DADS)*, pg. 169-179. accessed June 9, 2016 from http://dasish.eu/publications/projectreports/D4.2_-_Report_about_Preservation_Service_Offers.pdf

[11] SHARE, accessed June 9, 2016 from http://www.share-research.org/

Australia	Partnership led by Monash University, along with the Australian National University and the Commonwealth Scientific and Industrial Research Organisation	ANDS[12], Research Data Australia	ANDS has a centralized service - Research Data Australia, an internet-based discovery service that draws data records from more than 90 institutions to aggregate and showcase Australian data nationally and internationally. Research Data Australia covers a broad spectrum of research fields - across sciences, social sciences, arts and humanities.	Draws data records from more than 90 institutions	RIF-CS metadata standard, Catalogue Service for the Web (CSW) harvest, OAI-PMH	Funded by the Australian Government
United Kingdom	JISC, Digital Curation Centre (DCC)	UK Research Data Discovery Service[13]	DCC and UK Data Service pilot project to build a national registry service to aggregate metadata for research data held within UK universities and national, discipline specific data centres.	Nine higher education institutions, seven data centres including UK Data Archive, Archaeology Data Centre, NERC Data Centres	CKAN platform, OAI-PMH, Core Metadata Schema Version 1.0[14] - (mapping from Dublin Core, MODS, DDI-Codebook, DataCite, UK Gemini, EPrints / ReCollect)	Largely funded by the ESRC, the JISC and the University of Essex.

[12] ANDS Research Data, accessed June 9, 2016 from http://www.ands.org.au/

[13] UK Research Data Discovery Service, accessed June 9, 2016 from http://ckan.data.alpha.jisc.ac.uk/dataset

[14] UK Research Data Discovery Core Metadata Schema, accessed June 9, 2016 from https://docs.google.com/document/d/1pdSPfOTDPL8n6MiHDuqRF_zqESIQnb0gKOtVrkIpvBs/edit

Europe	OpenAIRE 2020 Project, Research Data Alliance Europe / WDS Publishing Data Interest Group, & ICSU World Data System	Data Literature Inter-linking (DLI) Service[15]	DLI Service provides an open service for collecting, sharing, and reusing data linkages between published resources and underlying research data. Links and metadata are provided openly for end-users (searching), developers, and content providers (enhance holdings).	Over 20 providers, including CrossRef, PubMed, IEEE, ICPSR, ANDS, PANGAEA, DataCite	DataCite DOIs, OAI-PMH, D-Net infrastructure	OpenAIRE EU funded project
Netherlands	DANS	DANS Search[16], NARCIS, and EASY[17]	DANS provides a variety of services including the National Academic Research and Collaborations Information System (NARCIS), the main national portal for searching for scholarly research outputs from research across the Netherlands.	DANS Search and NARCIS, provides access to scholarly outputs from many institutions. Upload of datasets to EASY by researchers directly; curated by DANS archivists.	Dublin Core; Qualified Dublin Core; DDI attributes with optional additions from FGDC or non-standardised metadata.	KNAW and the Netherlands Organization for Scientific Research (NWO)

The funding models for European national discovery projects and the United States' SHARE service have all been based on national grants. This is not to say that a different model could not be established in Canada, but it is something to think about especially when considering sustainability.

In Canada, there are a variety of institutional and disciplinary data centres and repositories that contain research data for reuse. Table 2 is a list of selected institutional and disciplinary data repositories that this Working Group agrees would

[15] DLI Service, accessed June 9, 2016 from http://dliservice.research-infrastructures.eu/#/

[16] DANS Search service, accessed June 9, 2016 from https://dans.knaw.nl/en/search

[17] DANS. EASY service, accessed June 9, 2016 from https://easy.dans.knaw.nl/ui/home

be in scope for the development of a national data discovery service in Canada. For a full list of data repositories and data centres in Canada, please refer to the National Research Council's Gateway to Research Data Portal[18].

Table 2 - List of Relevant Canadian Institutional Data Repositories / & Data Centres[19]

Provider / Institution	Repository name	Disciplines covered	Repository model	Metadata / Standards	Size[20]	Data Discovery
University of British Columbia Library	Abacus / Dataverse Repository	Multi-disciplinary	Curated / controlled access to deposit / multi-institutional	DDI-Codebook (mapping to Dublin Core, DataCite, ISO 19115)	1,875 studies, 30,555 files	Open access, OAI-PMH, Public API, Common metadata (required fields), Variable-level discovery, Handles, DOI
Ontario Council of University Libraries (OCUL)	<odesi>	Social Science / Multi-disciplinary	Curated / controlled access to deposit	DDI-Codebook (mapping to Dublin Core, MARC)	3,535 datasets	Some restrictions to data, open access to metadata OAI-PMH, Public API, Common metadata (best practice), Variable-level discovery
Ontario Council of University Libraries (OCUL)	Scholars Portal Dataverse Repository	Multi-disciplinary	Self-deposit / multi-institutional	DDI-Codebook (mapping to Dublin Core, DataCite, ISO 19115) Other disciplinary standards (astronomy, biomedical, earth sciences, journal)	473 studies, 6,342 files	Open access, OAI-PMH Public API Common metadata (required fields), Variable-level discovery, Handles, DOI

[18] NRC Gateway to Research Data, accessed June 9, 2016 from https://dr-dn.cisti-icist.nrc-cnrc.gc.ca/eng/home/collection/Gateway%20to%20Research%20Data/

[19] This is not an exhaustive list. These were chosen as relevant repositories to consult with further in the development of a collections policy for a national data discovery service. This table borrows from the Research Data Canada (RDC), Standards and Interoperability Committee (SINC), *Research Data Repositories: Review of current features, gap analysis, and recommendations for minimum requirements*. 2015. Accessed June 9, 2016 from http://www.rdc-drc.ca/download/review-of-research-data-repositories-2015/?wpdmdl=669

[20] Items in repository, if known, at the time of writing (June 2016).

University of Alberta Libraries	University of Alberta Dataverse Repository	Multi-disciplinary	Curated / controlled access to deposit	DDI-Codebook (mapping to Dublin Core, DataCite, ISO 19115)	247 studies, 2,407 files	Open access, OAI-PMH Public API Common metadata (required fields), Variable-level discovery, Handles, DOI
Statistics Canada, Data Liberation Initiative (DLI)	DLI	Multi-disciplinary	Curated / controlled access to deposit	DDI-Codebook	Unknown	Some restrictions to data, open access to metadata, OAI-PMH, Public API Common metadata (best practices) Variable-level discovery
Canadian Research Data Centre Network (CRDCN)	Research Data Centre Master File Metadata Repository (Multi-institutional network to access RDCs data)	Social Science/ Multi-disciplinary	Curated / controlled government access only	Statistics Canada Integrated Metadata Database (IMDB), DDI-Codebook DDI-Lifecycle	114 surveys	Restricted access to data, open access to metadata OAI-PMH Public API Common metadata (best practices) Variable-level discovery
Polar Data Catalogue (PDC)	Canadian Cryospheric Information Network / University of Waterloo	Earth Sciences/ Multi-disciplinary	Curated / controlled access to deposit	PDC Metadata, FGDC, ISO 19115	2,442 datasets	OAI-PMH DOI
Nordicana D	Centre for Northern Studies, Université Laval	Multi-disciplinary	Curated / controlled access to deposit	Unknown	Unknown	Most principals N/A; DOI for data and publications
Simon Fraser University	RADAR[21]	Multi-disciplinary	Curated / controlled access to deposit	DDI, Dublin Core	269 items	Open access Open formats

[21] RADAR, accessed June 9, 2016 from http://researchdata.sfu.ca/

Oceans Network Canada	Oceans 2.0 Data Search[22]	Multi-disciplinary	Curated / controlled access to deposit	Unknown	Unknown	Quality Assurance of Real Time Oceanographic Data (QARTOD)
Canadian Astronomy Data Centre[23]	CADC	Astronomy / Physics	Curated / controlled access to deposit	Unknown	115 Instruments providing unknown amount of data	TAP[24], direct download
University of Calgary	Xenbase	Biology / Life Sciences	Curated / controlled access to deposit	Unknown	Unknown	Open Access

The research data repositories highlighted above represent large collections of research data that operate mainly to support research in a particular domain or subject area, that serve a particular region or community (in the case of institutions or consortia), or that are embedded in government or other jurisdictions, preventing cross-disciplinary or external deposit or collections development. It is important to note, that while repositories are often the curators of research data and provide the means to share and access data, there is a need to have a shared research data discovery service in Canada that encompasses the vast variety of data, from across a range of disciplines, to support central data discovery and access.

Issues and considerations

There are several considerations to make when attempting to bring together metadata from across repositories and data centres in Canada. This includes having a reason to bring together the data, the scope, and the audience. Not all Canadian data are found in Canadian repositories, for example, data collected by Canadian researchers, or about Canada, can be found in a multitude of places, including international repositories such as Dryad[25], FigShare[26], and PANGAEA[27]. Assuming that the focus is just on Canadian repositories and data centres would be an oversight as there are additional considerations and issues to take into account.

---

[22] Oceans 2.0 Search, accessed June 9, 2016 from http://dmas.uvic.ca/DataSearch

[23] CADC, accessed June 9, 2016 from http://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/

[24] TAP, accessed June 9, 2016 from http://www.ivoa.net/documents/TAP/20100327/REC-TAP-1.0.html

[25] Dryad, accessed June 9, 2016 from http://datadryad.org/

[26] FigShare, accessed June 9, 2016 from https://figshare.com/

[27] Pangaea, accessed June 9, 2016 from https://www.pangaea.de/

Further engagement with partners managing institutional and disciplinary repositories in Canada is required to represent the full breadth of use cases and issues for such a national service. It should be noted, that in all the national data discovery service cases outlined in Table 1, consultations with the community and repository stakeholders were of primary importance.

The following is a list of issues, in no apparent order, which will require further community consultation:

- Data duplication;
- Dataset granularity - especially for large volumes of data (e.g. genomics data);
- Platforms and tools;
- Willingness to participate, share, and contribute metadata[28];
- Metadata standards;
- Aggregation level and methods;
- Maintenance and sustainability model.

The multidisciplinary scope of a national discovery service presents unique challenges for infrastructure, since dataset granularity can be difficult to manage for all user groups, disciplines, and stakeholders. It should be noted that it may not be feasible to find a 'one fits all approach' for all data in Canada. Nevertheless, we can begin to discuss these issues and develop solutions for disciplinary requirements as they arise.

Platforms and tools are also important to discuss as it directly impacts the ability to adhere to data discovery principles. For example if a data discovery service uses platforms and tools that do not support open access, it will be difficult to meet our principles and standards. This Working Group is not evaluating platforms or repository tools, although such assessments could be in the scope of the advisory bodies tasked with assisting the development of such services.

Recommendations

We recommend to begin a project to scope what a national data discovery service would entail, and ideally through the formation of defined advisory groups for data repositories and data centres with whom to consult. These groups, or a larger combined advisory group, will be consulted throughout the project phases, including scoping collections, metadata and tools, stakeholder and repository engagement, technical development, and piloting. Rather than reinventing the wheel, the advisory

[28] This working group did not actively engage with any of the repositories or institutions described in the above table (beyond our own affiliations), and therefore consultation is required to get a sense of the willingness to participate and contribute to a national service.

groups will assess the national data discovery services in existence (UK, ANDS, DANS, etc.) and try to build on these efforts to establish some shared understanding.

Recommended representation on Advisory Groups:

- Users (researchers, faculty, librarians, etc.)
- Repositories / Data Centres (a selected group to begin piloting a service)
- Institutions (academic institutions, from across Canada's regions).

Enhanced Data Discovery & Visualization Systems

A state-of-the-art discovery system does not simply provide a list of resources with a text search box and keywords. The latest search technologies also provide support to seekers, helping them formulate and refine queries.

The following is a set of features identified as particularly useful for data discovery:

- Advanced search options (such as date range, variable-level / data element searching, geographic coverage, etc.)
- Visualization of standard metadata (display of fields, values, links, etc., for understanding);
- Faceting (by type / format, collection, repository, geographic coverage, topic, date, etc.);
- High level overviews (study-level), with the ability to drill-down to data elements;
- Linked representations that bridge data and related resources, in order to contextualize data and related research; and
- Open APIs that allow the reuse of metadata and search results for the development of applications.

Table 3 (below) contains a sample of systems and techniques that provide examples of enhanced data discovery. Most of these systems do not specifically address the discovery of research data, however, their treatment of bibliographic metadata for more traditional scholarly research materials is directly applicable to collections of datasets.

Table 3 - Enhanced Discovery Systems

Project	Description	Data[29]	Beta[30]	API	Link
Bohemian Bookshelf	Serendipitous discovery through linked, modular visualizations	N	Y	N	http://www.alicethudt.de/BohemianBookshelf
Collection Diver	Search interface that highlights the search process; providing strong visual feedback on filters and facets.	N	Y	N	http://hci.uni-konstanz.de/downloads/CollectionDiver.mp4
PivotPaths	Faceted exploration of bibliographic citation networks	N	Y	N	http://mariandoerk.de/pivotpaths
PivotSlice	Visual construction of dynamic data queries featuring integration of online data sources, live search, and graphic interaction histories.	N	Y	N	http://vialab.science.uoit.ca/portfolio/pivotslice
VisGets	Coordinated visual interactive elements for creating search queries that provide both graphical summaries and query formulation.	N	Y	N	http://innovis.cpsc.ucalgary.ca/Research/VisGets
Visual-overview for government data	Use of visual dashboards for previewing datasets, enabling people to quickly evaluate datasets for quality and applicability to their purpose.	Y	Y	N[31]	http://dl.acm.org/citation.cfm?id=2757407 https://github.com/niclabs/visual-overview

[29] Indicates whether the project has been specifically designed for datasets discovery.

[30] Indicates whether the project is a research project or prototype (Y) as opposed to a production level system (N).

[31] Demonstration source code is available on GitHub under an Apache 2.0 license.

Ariadne's Thread	Provides network graphs of bibliographic entities in order to encourage exploration of context. Prototypes interoperate w/ ArticleFirst, WorldCat, and Astrophysics.	N	Y	N	http://www.oclc.org/research/themes/data-science/ariadne.html
Scopus	Provides options to analyze search results of abstract & citation database allowing interaction through charts for selections based on year, source, author, affiliation, country, document type, and subject area.	N	N	Y	https://www.elsevier.com/solutions/scopus
UBC Open Collections	Open Collections brings together locally created and managed content from the University of British Columbia Library's four open access repositories (DSpace, CONTENTdm, Dataverse and AtoM).	Y	N	Y	https://open.library.ubc.ca

In addition to tools designed for the discovery of datasets, several data repositories either provide their own or make use of external tools to integrate data analysis and visualization functionality, allowing users to explore a dataset in situ with analysis occurring on the repository server. This improves discovery and access to research data by allowing the researcher to explore data elements on the web without specialized software for reading the data. In addition to these kinds of visualization systems, data download offer users the flexibility to explore and assess the applicability of data for reuse.

Table 4 - Repositories & Data Visualization Systems

Tool	Repository	Description	Link
CKAN	CKAN	Several web browser-based tools for viewing tables, charting, and mapping.	http://ckan.org/
R	Dataverse	Integrated w/ Dataverse to provide advanced statistical analysis functionality.	https://www.r-project.org/
Two Ravens	Dataverse	Aimed to provide advanced statistical analysis functionality for quantitative data; makes use of interactive charts for filtering and display results.	http://datascience.iq.harvard.edu/about-tworavens
Chemistry Solution Pack	Islandora	Adds chemistry-specific functionality such a 3D viewing of molecules and checkmol analysis (analyzes molecular structure files for presence of functional groups and structural elements).	https://github.com/discoverygarden/islandora_solution_pack_chemistry
Islandora Data Solution Pack	Islandora	In-browser viewing of tabular data (spreadsheets).	https://github.com/axfelix/islandora_solution_pack_data
Geoserver	Nesstar	Provides geospatial mapping functionality.	http://geoserver.org/
Nesstar	Nesstar	Create charts, and sub-tables. Statistical functionality includes cross-tabulations, correlations, regressions and application of variable weights.	http://www.nesstar.com/

Issues and considerations

In the context of data discovery, the volume of material in such systems can pose challenges, for example, lists and text searches can be unwieldy. Systems that are inherently multidisciplinary require different techniques to assist users in dealing with differences in vocabularies, word usages and representations of a dataset. Consequently the process of visualization for discovery systems is not the simple creation of a chart or plot with an overview of the datasets; rather these are interactive, visual elements that can be used to access individual resources, filter unwanted items, navigate collections, and assist seekers in querying the collection.

In terms of enhanced data discovery and visualization systems, there are two important areas:

- Breaking down data silos and encouraging the linkage and reuse of data and related collections, particularly in interdisciplinary research;
- Facilitating the linkage of data to other research outputs, making data citation possible (through persistent identifiers) and referencing easier, thereby incorporating data in research achievements and impact assessment overall.

Moreover, for such a service to be genuinely useful to the Canadian education and research community, it is crucial that the user research community has a central role in setting out requirements and providing feedback on all aspects of the service development.

Data Discovery Principles & Further Recommendations

Common metadata

Agreed minimum standards of RDM-related metadata are necessary to enable adequate discovery and to support research administration and management throughout the research data lifecycle. The UK National Data Discovery project[32] has recently compared various metadata schemas from institutional research data repositories and looked common elements among them. Other data discovery systems take different approaches to common or mandatory metadata requirements and to the aggregation of various metadata standards from different metadata producers and suppliers.

[32] JISC Blog, accessed June 9, 2016 from https://rdds.jiscinvolve.org/wp/2016/03/18/how-much-metadata-is-enough/

As shown in Table 1, ANDS uses the RIF-CS schema which has many mandatory elements. This is because ANDS is addressing a national solution from a creation-through-preservation lifecycle of research data. This includes discovery, value, access and reuse standards that require administrative and disciplinary metadata to be included in one record.

On the other hand, DataCite only has five mandatory metadata fields. DataCite[33] is an important standard that particularly addresses discovery and linking on the web. It should be noted that individual disciplines often have their own metadata standards and ontologies; it remains to be seen whether local technical solutions can accommodate these across a broad range of disciplines without making deposit workflows overly complex.

A note about granularity:
In addition, many data repositories in Canada support granular variable-level metadata descriptions, such as the Data Documentation Initiative (DDI) standard. OCUL's <odesi>, UBC's Abacus, and Statistics Canada's Data Liberation Initiative (DLI) repository provide DDI encoding of datasets to enable variable-level description and reuse. Therefore it is important to note that granularity should also be considered and left flexible to support these descriptions that enable rich data discovery.

Nevertheless, does not mean that variable-level descriptive metadata is superfluous. Granularity should be incorporated in context, and not to the detriment of usability. This is also the case for other disciplines where differences in what qualifies as a dataset can become overwhelming, especially with large volumes of data (e.g. genomics data, crystals, etc.).

A note about multilingual systems and data:
In the Canadian context, there is a need to provide users with content and a search interface in both official languages, English and French. There is also the case for other languages. To facilitate powerful discovery, the systems underlying the display and visualization of metadata and data should handle multiple languages.

Platforms for displaying metadata for example should provide users with the same or a closely similar experience in either language. This requires careful consideration and review of how the system is configured, rules for defining what gets indexed and is searchable, and the overall experience related to searching, user interface interactions and dynamic displays.

[33] DataCite Metadata Standard, accessed June 9, 2016 from https://www.datacite.org/

There are several data repository internationalization projects currently underway in Canada to provide multilingual tools and software for research data management, including partnerships between the University of Alberta, Université de Montréal, and Scholars Portal (OCUL).

Across all of these projects, the aim is to help researchers find and use relevant datasets independently of original language of the research source. This includes tackling issues related to metadata standards and language qualification at the field-level, controlled vocabularies, dataset character encoding, data delimiters, and more. CERIF[34] of euroCRIS[35], an international relational data model for research administrative information, may provide insight into integrating research datasets from French and bilingual institutions and repositories.

Recommendations
Organizations implementing one common metadata standard provide a litany of complications, for example, UBC's Open Collections, UK National Research Data Discovery Service, and others. We recommend exploring a set of metadata tools and a platform that can accommodate multiple, overlapping metadata namespaces, with the expectation that a generic but already existing schema, such as Dublin Core[36], MODS[37] or METS[38], would be described for all objects, using crosswalks if necessary. We also recommend building flexible metadata harvesters for indexing specialized repositories as needed, so that domain-specific metadata and metadata granularity can be retained in its original format for discovery.

Careful consideration will also need to be given to issues related to multilingual content and data, and to a consistent user experience independent of language.

Persistent identification

A global unique identifier, such as a DOI or ORCID, provides a unique and stable mechanism to identify objects and people on the web. This means it will not change if the item or object is moved or renamed.

DataCite, an international organization comprised of and supported by a variety of scholarly actors including government, journal publishers, data producers,

[34] CERIF, accessed June 9, 2016 from http://www.eurocris.org/cerif-cornerstone-creation-research-information-infrastructures

[35] eutoCRIS association, accessed June 9, 2016 from http://www.eurocris.org/what-eurocris

[36] Dublin Core Metadata Standard, accessed June 9, 2016 from http://dublincore.org/

[37] MODS Metadata Standard, accessed June 9, 2016 from http://www.loc.gov/standards/mods/

[38] METS Metadata Standard, accessed June 9, 2016 from http://www.loc.gov/standards/mets/

institutions, and libraries, promotes research data identification using the Digital Object Identifier (DOI) standard, which supports both data and publication identification on the web. A number of data repositories, libraries and data centres in Canada are aiming to provide DOIs for datasets. The hope is that this will eventually lead to shared standards for the identification of datasets on the web. Identification is only one piece of the challenge to improved data discovery, however.

The DOI standard (ISO 26324:2012[39]) is the foundation of the DataCite linking service, which allows location and tracking of both cited and citing references in the scholarly record. The DOI system provides a framework for persistent identification, managing intellectual content and most importantly managing metadata. DOIs are widely used in scholarly publishing to cite journal articles and research data.

ORCID[40] is a non-profit initiative, similar to DataCite, that provides a registry of unique researcher identifiers providing a transparent method of linking research activities and published scholarly outputs. ORCID, similar to authority files that have existed in libraries for many years, has the ability to span disciplines, research sectors, and national boundaries to solve researcher name ambiguity and thus assure that each author and researcher derives full credit for his or her work. ORCID IDs can also streamline the process of identifying a researcher's publications when tracking their citations, or calculating their h-index, or in creating a CV for funding agencies. It is also being explored as a means of linking disparate sources of data and scholarly outputs when no technical linkage has existed previously, greatly enabling system interoperability.

Issues and considerations

There are a number of considerations around best practices for minting DOIs and ORCID identifiers that will need to be addressed, specifically around authority control and data duplication. A dataset with multiple DOIs issued, for example, is not considered a best practice for identifying data. Best practices and some level of authority is crucial when issuing DOIs to datasets. We suggest identifying the primary version of the dataset and assigning a DOI to this version only. Where there is an unavoidable need to publish a dataset in different locations each with a separate DOI, we recommend that metadata for each appearance of the dataset should indicate the association[41].

---

[39] DOI ISO Standard, accessed June 9, 2016 from http://www.iso.org/iso/catalogue_detail?csnumber=43506

[40] ORCID, accessed June 9, 2016 from http://orcid.org/

[41] CISTI. "Republished and duplicate datasets" Accessed June 9, 2016 from https://cisti-icist.nrc-cnrc.gc.ca/obj/cisti-icist/doc/datacite/datasets.pdf

We also note that there is a great deal of data duplication currently among data repositories in Canada. This should be actively assessed and, if possible, agreements be arranged between parties to avoid passing duplicated metadata onto a national data discovery service.

Recommendations

We recommend exploring a national ORCID agreement[42] that universities and government agencies in Canada can use to integrate researcher identifiers into institutional and other research management and publishing software. With every researcher having an ORCID-style identifier, all data deposit and publication routes could be easier and clearer.

We also recommend registering DOIs with Datacite Canada[43] for the datasets existing in the participating repositories in Table 2 above, especially if the repository or data centre itself is without a means to do this. Assigning DOIs to research datasets will also greatly enhance their discoverability via Datacite metadata partners, e.g. DataOne, ORCID, VIVO and more. However, it should be noted that this should be done in collaboration with participants and not to the detriment of data identification best practices.

Lastly, it is recommended that a collaborative, community approach be undertaken to come up with a national sustainable technical solution for issuing DOIs and ORCIDs in Canada.

Open Access and Programmatic Interfaces

An application program interface (API) allows one piece of software to make use of the functionality or data available to another through a set of routines, protocols, and tools. A good API provides a set interface with which to work, making it easier to develop a software program and freeing future parties from having to understand and work with the entire system. In a discovery system, an API might provide mechanisms for registering datasets, searching the collection's metadata, or retrieving the information (such as identifiers, associated files, metadata, etc.) for a particular dataset. Most importantly it allows third parties to build tools upon existing systems and allows different systems to interoperate through the exchange of information in a structured, consistent manner.

---

[42] ORCID Membership, accessed June 9, 2016 from https://orcidpilot.jiscinvolve.org/wp/2015/02/03/next-steps-for-orcid-adoption-orcid-consortium-membership-for-the-uk/

[43] DataCite Canada, accessed June 9, 2016 from http://www.nrc-cnrc.gc.ca/eng/publications/library_services/datacite/index.html

Many repositories and platforms, including the UBC Open Collections project[44], use APIs to allow users and developers to run powerful queries, perform advanced analysis, and build custom views, apps, and widgets with full access to the Open Collections' metadata and transcripts. In the UBC case, a request is a URL sent to the web server over HTTP with the expectation of getting resource items back in a machine and human-readable form. The URL supplies the web server with everything it needs to create and return a correct response. This is known as a RESTful approach to API design.

While custom-made, system specific APIs are often necessary to expose functionality, there are a variety of standardized APIs that are relevant to discovery systems. One particularly relevant standard is the Open Archives Initiative (OAI) protocol for the exchange and harvesting of metadata, referred to as OAI-PMH[45]. This standard provides consistent, structured, and interoperable formats for metadata exchange and consequently is used for many existing aggregating services for harvesting data (Table 1) and repositories for exposing data (Table 2).

Issues & Considerations

- Not all research data reside in open repositories, and often data centres do not have APIs or support protocols such as OAI-PMH;
- Standardized APIs provide common interfaces across a variety of systems, but require time to become standards and may not expose all the latest features and discovery elements;
- Harvesting metadata doesn't address issues or concerns about metadata quality, completeness, or a common metadata across repository systems;
- Harvesting should be scheduled according to a timely update and refresh cycle, and ideally this should be automated.

Recommendations

We recommend that participating repositories and any national research data discovery platform provide an application programming interface (API) by which metadata and data can be programmatically accessed for a variety of reuse and development purposes. In addition to providing system-specific API elements, API standards such as OAI-PMH should be utilized where applicable.

[44] UBC Open Collections API, accessed June 9, 2016 from https://open.library.ubc.ca/research

[45] OAI-PMH, accessed June 9, 2016 from http://www.openarchives.org/OAI/openarchivesprotocol.html

Licensing

We believe that nobody yet has solved all the complexities of making data openly available and reusable. In fact, nobody has yet even figured out what all the questions are. There is a gap between data that is available and data that is reusable, in the absence of good and robust metadata, including administrative metadata about licensing, data are often not easily reproducible or reusable. The Creative Commons licenses model[46] is well established, with a range of standard, easily expressed and understood, legally enforceable licenses. Creative Commons (CC) licenses do two things: they allow creators to share their work easily and they allow everyone to find work that is free to use without permission.

The Dataverse repository, used by UBC Abacus and others in Canada, employs CC0 from Creative Commons as its default license for open research datasets. However, it also allows researchers to choose a variety of other Creative Commons licences from a drop down menu when datasets are uploaded. Encouraging open licences for datasets will greatly enable easy and timely reuse.

Recommendations

We recommend the use of Creative Commons licenses for research data because they effectively communicate information about the copyright holders' intentions, clarifying which data may be used and which require permission. CC licenses help authors and creators manage their copyrights and share their creative work without losing control over it. Furthermore, Creative Commons licenses provide a contact for permission when appropriate. Licensing can apply to data and metadata, although we strongly recommend that metadata be provided as openly as possible, with minimal to no restrictions on reuse to facilitate discovery.

Next steps

A national discovery platform in Canada will enable researchers to access metadata describing datasets across a variety of disciplines. This has great potential to expand and accelerate the generation of knowledge on a national and international basis. Participating institutions and research data centres would likely see a marked increase in traffic to their datasets because of increased exposure. Research administrators may benefit from improved statistics about the re-use and impact of research data generated at their institution, providing another driver for the recognition of research datasets as primary outputs of research and as an institutional asset in their own right.

[46] Creative Commons Licenses, accessed June 9, 2016 from https://creativecommons.org/licenses/

We emphasize that collaboration will be the driving force to improving data discovery in Canada. A well coordinated national project will ensure that all attempts to improve discovery and access to data will be informed and facilitated by stakeholder expectations, participation, and collaboration. This will ensure that data discovery infrastructure in Canada moves forward with collaborative and informed decision-making. Keeping stakeholders engaged and providing clear communication channels are key for the success of a national data discovery service.

We recommend expanding the Portage Discovery Expert Group[47] to proactively invite participation from other stakeholders and to serve as a national data discovery expert group. This expanded group will in turn create working groups (users, repositories / data centres, and research institutions) tasked with the following:

- Select repositories / data centres for inclusion in a pilot service;
- Create a collection development policy, to decide what data should be included and excluded;
- Evaluate and decide on metadata elements for discovery (e.g. spatial, temporal, descriptive, technical) and work with experts to build a metadata model;
- Engage and communicate with Canadian higher education institutions and national and domain specific data centres about requirements gathering and research use cases for a pilot service;
- As best as possible, through further consultation, address issues around data duplication, common metadata, licensing, and adoption of appropriate standards with the data community in Canada (CARL Portage, RDC, DLI, Compute Canada, Environment Canada, and others).

The changing practice of research increasingly requires the data and other sources that constitute the evidence underpinning findings to be made available for verification and reuse. We repeat - in order to be re-used, research data must be discoverable. We are excited to see this work commencing in Canada and are looking forward to further collaborations.

Acknowledgements:

We would like to thank Diane Sauvé (Université de Montréal), Chuck Humphrey (Portage), and Martha Whitehead (Queen's University) for peer reviewing our paper and providing very valuable feedback on its content, directions and style.

[47] Portage Data Discovery Expert Group, accessed on June 10, 2016 from https://portagenetwork.ca/about/network-of-expertise/expert-groups-membership/