Research Data Management Program – Competitive Funding Call 1

Statement of Work

Name of project:

Federated Geospatial Data Discovery for Canada

Institution / Organization (Lead Contractor):

University of British Columbia Library
1961 East Mall
Vancouver, BC Canada V6T 1Z1

Table of Contents:

Project Summary

In this proposal, we are working towards three (3) priority areas for CANARIE RDM funding consideration: Enriching (Meta)data and Discovery, Data Deposit and Curation, and Federated Repositories / Interoperability.

Introduction

To make data accessible, one of the key components is discoverability. Traditionally, research data repositories have relied on text-based searching, familiar to users everywhere as the default method of searching in Google. However, more and more users are searching for data that is geospatial in nature, with an explicit (or sometimes implicit) geographic component. This type of data arrives in many, sometimes seemingly unrelated forms. For example, the migration paths of humpback whales, the distribution of maple-syrup yields, infrared satellite imagery, distribution of artifacts in an archaeological site or the flow routes of water due to sea level rise. However, what all of these seemingly disparate forms of data share is **location**.

Current data repositories most commonly in use in Canada lack a map-based interface for which users can search, and more importantly, see, whether data has a location based component which is essential for their research. Such an interface would be beneficial to all research disciplines, from anthropology to zoology.

The goal of our project is to create an extensible software method to find and display this location-aware data in a search interface which is both map and text based, combining research data with the functionality of a product such as Google Maps.

Structure

The project will have several components working in concert:

- Software which will query the Dataverse repository to determine if geospatial information is present within the digital object (eg, a study, or a data deposit)
- The software will query and harvest any geospatial data in the primary record (eg: main record page)
- More importantly, the software harvester will query and harvest any geospatially relevant file objects (satellite imagery, geospatial vector files, etc)
- Once the data have been harvested, the software will create and normalize relevant geospatial data from the primary record (A) and from any associated digital objects (B) and extract all relevant metadata
- The extracted, normalized data will be deposited by the software into an existing geospatial data server, such as OpenGeoserver, which are stand alone-products capable of distributing geospatial data in a wide variety of formats to various services
- Data then will be harvested by a geospatial search interface such as GeoBlacklight - http://geoblacklight.org/, an open source geospatial search tool
- The user interface will be customized to the needs of the Federated Research Data Repository project (FRDR), providing a unified map-based search interface for research data in Canada.

Components:

Dataverse (DV) is an open-source data repository used for publishing and sharing research data. It offers researchers an easy-to-use deposit workflow to publish and share data across disciplines, and supports a range of research data standards, including ORCID, Data Documentation Initiative (DDI), file level DOIs, disciplinary metadata, and file versioning. Developed originally at the Institute for Quantitative Social Studies (IQSS) at Harvard, DV is being adopted globally as a repository for research data by national and institutional data services.

Portage is an initiative of the Canadian Association of Research Libraries (CARL) dedicated to advancing Research Data Management in Canada. The Dataverse North Working Group (DVN WG), an initiative of Portage, brings together close to 20 universities in Canada actively using DV. The University of British Columbia Library (UBC) hosts the Abacus Dataverse Network (http://dvn.library.ubc.ca/dvn/) for UBC researchers, and, for some institutions in BC including the University of Victoria, SFU and UNBC. Together there is tremendous interest in developing Dataverse for both the library and research communities in Canada. For instance, Scholars Portal (SP), a service of the Ontario Council of University Libraries (OCUL), located at and managed by the University of Toronto Libraries (UTL), provides Dataverse hosting services for institutions in Ontario (http://dataverse.scholarsportal.info). Both UBC and SP, with support from university libraries, have contributed to the Dataverse codebase and are active in the Dataverse developer community. UBC Library is also actively working with Portage and Compute Canada (CC) connecting to the national RDM infrastructure, including the Federated Research Data Repository project (FRDR) (http://frdr.ca/repo), which runs UBC Open Collections open source code as a discovery interface.

FRDR supports data discovery for approximately 30 data repositories in Canada. FRDR does this by harvesting metadata using the Open Archives Initiative (OAI) protocol, CKAN, CSW, and MarkLogic custom feeds, to gather meta(data) in its original form from each repository, including from Canadian DV instances. FRDR then stores and indexes this information (using Dublin Core and DataCite) to allow researchers to search across datasets from different disciplines. There is an increasing amount of research data being shared and deposited that contains geospatial information and data. At present, DV and FRDR have very limited support for handling geospatial data visualization and discovery.

Supporting researchers with managing and sharing their geospatial data requires extending repository tools such as Dataverse and FRDR, to improve overall data visualization and metadata indexing in order to promote new ways to discover Canadian research. Partnering with the Canadian Historical Geographic Information Systems Partnership (CHGIS) (http://geohist.ca/), a SSHRC funded interdisciplinary collaboration aimed at building a network of researchers and infrastructure to support visualization and sharing of historical geospatial data, will provide valuable input on the data management requirements, user interface design and behaviour, and community engagement that will be necessary for this project to be successful. The enhancements to DV and FRDR, as described above and below, would provide this and other communities with a robust platform for sharing, data management including file versioning, discovery tools, and the ability to reuse and preserve geospatial research data in Canada.

End Product:

The completed end product will consist of a federated map-based user interface for data. In addition to the traditional text-based entry for searching, end users will be able to visualize the research data *while* they search FRDR or other repositories. The intermediary geospatial server component will allow this same metadata and data to be disseminated to other platforms in a variety of formats, greatly enhancing access and usability.

Lead Contractor

Lead Contractor	
Organization Name	University of British Columbia Library
Organization Type	University
Anticipated role/s in the project	Software development, project management, metadata analysis and guidance, training support, service provider, hardware infrastructure

Principal Investigator	
Name	Eugene Barsky
Title	Research Data Librarian
Faculty/Department/Division/Program	UBC Library
Phone	604-822-9606
Email	eugene.barsky@ubc.ca

Co-Principal Investigator	
Name	Evan Thornberry
Title	GIS Librarian
Faculty/Department/Division/Program	UBC Library
Phone	(604) 822-8667
Email	evan.thornberry@ubc.ca

Co-Principal Investigator	
Name	Lee Wilson
Title	Service Manager
Faculty/Department/Division/Program	Portage Network
Phone	(902) 233-8927
Email	lee.wilson@ace-net.ca

Co-Principal Investigator	
Name	Amber Leahey
Title	Data & GIS Librarian
Faculty/Department/Division/Program	Scholars Portal, University of Toronto
Phone	416 978 7217

Email	amber.leahey@utoronto.ca

Co-Principal Investigator	
Name	Marcel Fortin
Title	Head, Map and Data Library
Faculty/Department/Division/Program	Map and Data Library, University of Toronto Libraries
Phone	416-978-1958
Email	marcel.fortin@utoronto.ca

Co-Principal Investigator	
Name	Jason Brodeur
Title	Head, Map, Data, GIS
Faculty/Department/Division/Program	McMaster University
Email	brodeujj@mcmaster.ca

Co-Principal Investigator	
Name	Paul Lesack
Title	Data/GIS Analyst
Faculty/Department/Division/Program	UBC Library
Phone	(604) 822-5587
Email	paul.lesack@ubc.ca

Project Manager/ Business Analyst CANARIE strongly recommends that your team include a dedicated Project Manager.	
Name	To be hired as a first priority as a part of this funding
Title	
Faculty/Department/Division/Program	
Phone	
Email	

Lead Software Developer If you do not currently have a Lead Software Developer, please leave this section blank.	
Name	To be hired as a part of this funding
Title	
Faculty/Department/Division/Program	
Phone	
Email	

Participants

Note:

- A Participant will incur eligible costs that will be claimed through the Lead Contractor.
- If you have multiple participants, please fill out the table below for each one.

Funded Participant #1	
Organization Name	UBC Library
Faculty/Department/Division/Program	Library
Organization Type	University
Address	1961 East Mall Vancouver, BC Canada V6T 1Z1
Province	BC
Anticipated role/s and rationale for involvement	Software development support, project management support, storage and computing resources, training support, metadata support

Funded Participant #2	
Organization Name	UBC Advanced Research Computing
Faculty/Department/Division/Program	VP Research and Innovation Office
Organization Type	University
Address	UBC Advanced Research Computing Room 112A, Gerald McGavin Building 2386 East Mall Vancouver BC Canada V6T 1Z3
Province	British Columbia
Anticipated role/s and rationale for involvement	Hardware management, project management, financial management support in research accounting

Mandatory Criteria:

Please address each of the following mandatory criteria:

1. **How does the proposed software make use of Canadian digital infrastructure** (networking, compute, storage and/or sensors)**?**

DV is hosted and maintained by a number of universities and organizations in Canada, including UBC Abacus Dataverse, University of Alberta and Scholars Portal/UTL.

The Federated Research Data Repository (FRDR) is a joint partnership between Compute Canada (CC) and CARL/Portage built on CC infrastructure distributed throughout Canada, including on the Compute Canada Cloud service and national storage infrastructure. Its aim is to provide a scalable data repository solution for Canadian researchers that enables high speed data transfers of large files over CANARIE and other regional research networks. Additionally, FRDR functions as a national search and discovery layer by harvesting and aggregating metadata records for research data held across a variety of Canadian institutional, regional, government, and domain-specific repositories.

2. Explain how the proposed project contributes to one or more of <u>CANARIE's expected results:</u>

(a) enhance opportunities for collaborative knowledge creation and innovation within Canada's research and education communities through the maintenance and development of the CANARIE Network and related tools and services;

(b) expand the research and education community's access to and utilization of the CANARIE Network and the availability of tools and programming that increase the effectiveness of its use; and,

(c) enable the creation of innovative Information and Communications Technology (ICT) products and services and accelerate their commercialization in Canada.

Increasing amounts of research data are being created in geospatial formats or with real-world spatial attributes, from data collected by sensors and satellites to historical data derived from printed maps that have been digitized and georeferenced. Researchers are reusing this data for visualizations, data analysis, data integrations, and in countless other ways that link different datasets according to their shared locations. For example, a researcher in the medical sciences studying the habitat suitability of viruses on Vancouver Island requires a variety of geospatial data for a specific area. In this case, related data might be published by researchers in unrelated fields (forestry, climatology, biology, civil engineering, etc.), but are common by their location. Geospatial data support and discovery services that this proposal aims to enhance hugely benefit these types of place-based research.

DV currently supports a basic API that allows it to connect with a mapping application developed and hosted at Harvard's Center for Geographic Analysis called WorldMap. However, this connection does not support specialized standards for geospatial discovery and analysis provided by, for instance, the Open Geospatial Consortium (OGC) set of map service APIs. Extending the UBC Abacus DV platform to support OGC standards-based geospatial data deposit and map visualization, through the development of an integrated map server application, would better support existing use and creation of geospatial

research data, and also open the platform to a broad new community of researchers that have newly come to rely on GIS data for analysis and visualization, including economics, public health, and history.

In addition to providing the services to reuse and analyse geospatial data collaboratively across Canadian educational institutions and the general public, this project will enable researchers to easily share and discover geospatial data. This development will enhance the use of data repositories in Canada and build collections of reusable geospatial research data in Canada, including historical GIS data. Adoption of DV and FRDR for the management, sharing, and discovery of geospatial research data in Canada will expand the opportunities for researchers to utilize national RDM infrastructure to promote and share geospatial research data. By exposing research data as web-based map services, the project will allow researchers across Canada to access these resources over the CANARIE network for analysis and visualization.

This project will support enhanced research data publishing and sharing workflows in DV, and will support researchers from various sectors of academia and the private sector who collect, analyze, and reuse geospatial data. DV supports and encourages an "open data" model, and by integrating open APIs in the target software platforms, researchers in both academia and the private sector will be able to leverage that data to drive innovations in use of GIS sources in alternative platforms (e.g. mobile as well as desktop) and alternative representations (e.g. 3-D as well as 2-D representations).

3. Please provide information to support that the software development proposed is technologically and economically viable in the timeframe of the project plan.

A detailed project plan and task list is provided that outlines how the software development will progress throughout the project and in the given timeframe. The project plan includes sufficient time for each feature development and deliverable, including some contingency time in case certain unforeseeable issues arise that may impact the timeframe and economic viability of the project.

Much of the project will involve integration work using standard APIs, rather than development of entirely new systems from scratch. Existing familiarity with the tools and APIs of the core components of the proposed integration contributes to the technical viability of the project.

4. All funded work must be performed in Canada. Please identify the locations(s) at which the funded work would be performed.

Funded work will be performed at the University of British Columbia Library, Point Grey Campus, Vancouver, BC. Supportive work will also be contributed by partners, all located in Canada.

Partner locations: Scholars Portal at the University of Toronto (Robarts Library, 130 St George Street, Toronto, ON), Simon Fraser University, University of Saskatchewan (116 Science Place, Saskatoon, SK), Dalhousie University (Goldberg Computer Science Bldg, 6050 University Avenue, Halifax, NS) and CARL Office (309 Cooper, Suite 203, Ottawa, Ontario K2P 0G5).

5. The Lead Contractor must be a Canadian university, college, corporation or other legally recognized entity. Please provide the organization's URL.

The University of British Columbia: https://www.ubc.ca/
The University of British Columbia Library: http://www.library.ubc.ca/

6. No more than half of the membership and Board of Directors of the Lead Contractor can be composed of representatives or agents of the federal government. Please provide a list of Board members or provide the list via the relevant URL.

The Board of Governors of the University of British Columbia is responsible for the management, administration and control of property, revenue, and business affairs of the University. More information about the Board of Governors, including current membership, can be found at this URL: https://bog.ubc.ca/

7. In-Kind contributions must be at least 15% of total eligible project costs. This is to be demonstrated in the Preliminary Budget.

CONFIRMED X

In-kind contributions are outlined in the preliminary budget, and we anticipate significant support at all levels throughout the project.

8. Software developed under CANARIE funding must be made available for other researchers to use at no cost, through the CANARIE Research Software Registry at: https://science.canarie.ca/ for a period of 3 years from the end of the funding period.

CONFIRMED X

We agree to make the software developed available to other researchers at no cost. The code will also be distributed through an open Github repository available to any institution in Canada (and beyond) with its own DV installation. UBC Library is an active contributor to Open Source community on Github - https://github.com/ubc-library/

Scoring Criteria:

Applicants must clearly answer each of the following questions:

1. **What is the extent to which the project makes use of or contributes to digital research infrastructure?**

Digital Research Infrastructure (DRI) has been identified as a strategic priority for Canada, and a critical investment of $572.5 million over five years was allocated in the last federal budget to implement a DRI Strategy. This strategy will ensure that Canadian researchers have access to the advanced research computing infrastructure and services that they need to help drive Canada's future economic and social prosperity.

Building towards that goal, our project will create a Canadian solution for geospatial data discovery by integrating a number of existing federated DRI platforms, including Dataverse, FRDR, the national Advanced Research Computing (ARC) infrastructure managed by Compute Canada, and Canada's national research network managed by CANARIE. Connecting DV to a leading open source GIS server platform (GeoServer) will expose GIS data deposited in DV using OGC web mapping services. This will allow data to be visualized in DV and to be consumed and analyzed in other local, national, and international GIS platforms – server, desktop, and mobile. Geospatial resources published in DV will be harvested by FRDR to support data discovery at a national level. The geospatial search features of GeoBlacklight will be integrated into the national FRDR search portal to support location-based searching based on coordinate data derived from DV. Installing a GeoBlacklight instance as a part of FRDR's national search and discovery portal will also make better use of geospatial metadata already being harvested from a variety of Canadian open data portals (e.g. Open Data Canada).

Extending Dataverse to support geospatial data visualization and metadata packaging for discovery in the national FRDR platform (enhanced with GeoBlacklight services) will support new research applications including geo-enabled discovery features such as spatial searching, spatial data/attribute searching, and geographic map visualization of data deposited and published through DV.

Any researcher with spatial data or data that contains geographic information will be able to easily upload and publish research data using Dataverse, and researchers will be able to visualize their data on a map interface for sharing and exploration in the DV system. This metadata will then be published to FRDR for promotion and discovery, increasing the likelihood for reuse. The platforms will support open APIs that will allow for additional workflows and enhancements as needed. Additional repositories with geospatial metadata will be able to connect to this workflow for federated geo-enabled discovery with little effort, contributing towards the national digital infrastructure goals.

2. What is the extent to which the project creates or contributes to a national data service?

RDA National Data Services (NDS) Working Group defines national data service as: "A service that provides one or more data-related functions to applicable stakeholders and disciplines in a specific national context."

What is the problem that we are trying to solve and why is discovery of geospatial data is so critical in Canada?

An example may help to explain the need. While there are a number of geospatial data portals in Canada, including governmental, regional, and institutional portals, nowhere is this content aggregated, leaving it in silos defined (and limited) by discipline, region, or organization. A researcher interested in designing a land classification analysis study, for example, may require data for input into a model to predict soil erosion and flooding in a particular area. Searching only government geospatial databases, the researcher would uncover national-level contours and soil types, but might miss data created by researchers in other sectors who have shared that data in an open portal. This may include digitized historical air photos and vectorized historical data that indicate at one point the area had a significant fruit-bearing orchard. This kind of data, if available, might reveal to the researcher the existence of historical soil samples that would contribute to the researcher's land classification project.

Discovery of geospatial research data in Canada, then, requires national connections between data portals, repositories, and organizations, with a view enhanced forgeo-enabled data discovery features, in order to support interdisciplinary research workflows.

The primary challenge with a National Data Service (NDS) in the Canadian context is the lack of national oversight. International examples of successfully developing an NDS suggest that a centralized and coordinated effort is essential. However, given the nature of Canada's political and funding landscape, we believe that a federated model, based on the interoperability of many FAIR tools and platforms, makes the most sense for an NDS in the Canadian context. Our proposed development works within a federated framework by enhancing and strengthening connections between existing data repository and discovery platforms built on both local/regional and national infrastructure.

3. What is the extent to which the project supports FAIR principles?

In this proposal, we are working towards three (3) priority areas for CANARIE RDM funding consideration: Enriching (Meta)data and Discovery, Data Deposit and Curation, and Federated Repositories / Interoperability.

The development of a federated geospatial data discovery service would increase **findability** and **accessibility** of geospatial research data in Canada through metadata clean-up and crosswalking, guidance, and tools for the description of research data (e.g. Dublin Core, DDI, ISO 19115). Standard metadata accompanying geospatial research data provides the foundation for discovery and reuse, and support through a national service for the exchange of machine-readable metadata increases the

findability and accessibility of data through search engines (schema.org compatible), and, in a single, user-facing national data discovery platform to live alongside the existing FRDR search and discovery portal.

A federated geospatial data discovery service would enable **interoperability** by investing in efforts to support open geospatial metadata exchange and interoperability through disciplinary metadata standards and open APIs. Moreover, the integration of preservation processing and curation tools (e.g. Archivematica pipeline for Dataverse) would help to translate files into interoperable formats for migration purposes and use across different platforms. Furthermore, our proposed developments would support **re-usability** by providing an open, fit-for-purpose interface for discovering and exploring geospatial data that would encourage researchers, where appropriate, to apply open licenses to their data (e.g., Creative Commons Zero or CC BY), support standard citation of research data, and, ensure appropriate administrative metadata and documentation for reproducibility and reuse.

4. **What is the extent to which the project integrates with international digital research infrastructure?**

Supporting open geospatial research data management workflows will increase capacity for geospatial research work in Canada and beyond, supporting new use cases and partnerships between governments and academia as well as enhanced engagement with commercial software projects. For example, support at the national level for geo-enabled research data discovery can open up opportunities for other subject or national repositories and geospatial data platforms, including commercial platforms such as CARTO and ESRI's ArcGIS Online, to integrate with these research tools in the future.

Moreover, we will work closely with the Harvard Dataverse project and make all our code open source and available on Github, so it could be implemented in Dataverse 5.X and beyond for other institutions to benefit from this spirit of open source code development. Also, increased sharing of geospatial research data will provide new sources of data for use in research software development, to drive reuse in interdisciplinary research areas including environmental science, physical geography, land use planning, education, and transportation.

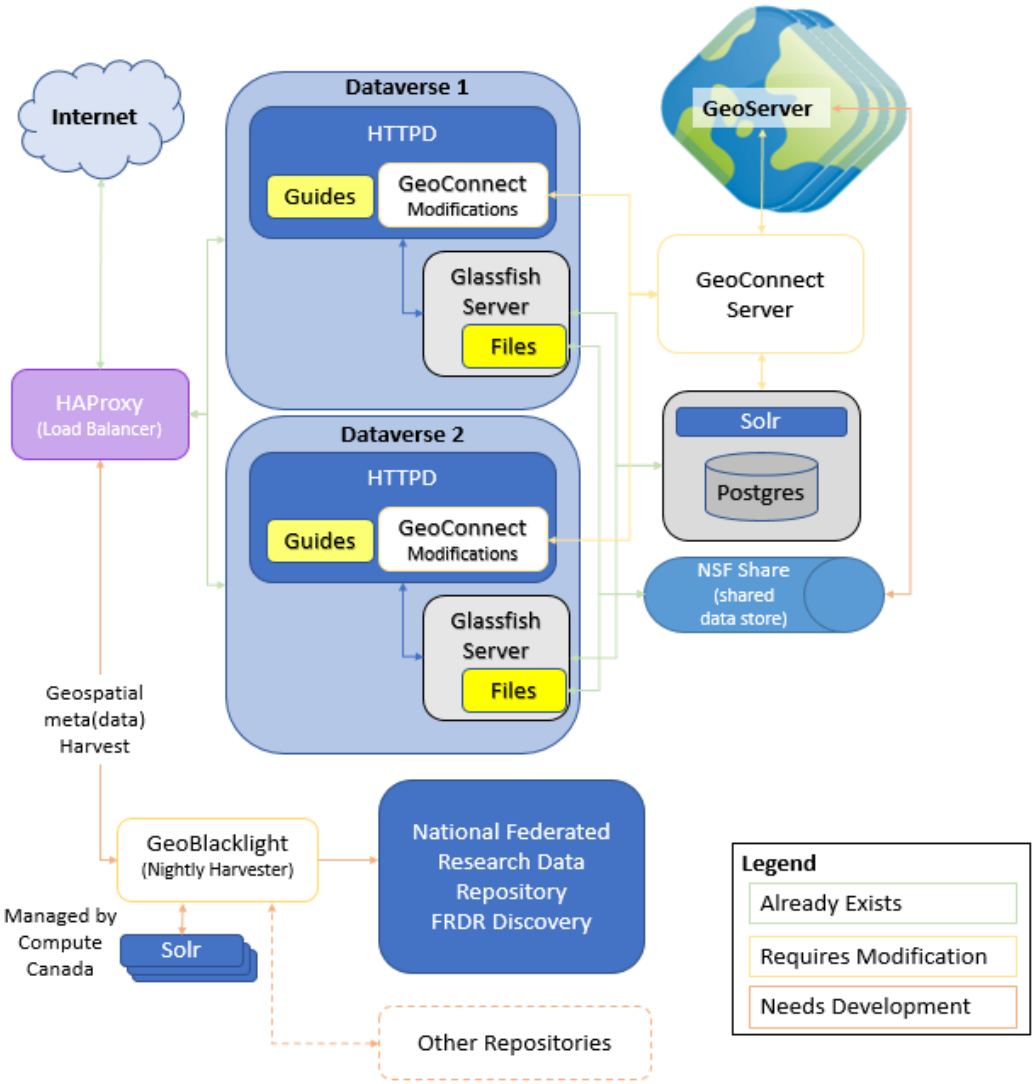
Proposed enhancements to the GeoConnect code will include additional features for defining geographic data in a spatial database, defining symbology and stylization for the map visualization, and ensuring feature and attribute-level elements are stored as part of the dataset metadata in DV. Geospatial metadata harvesting from the enhanced DV to a national instance of GeoBlacklight will support geo-enabled discovery of research data in Canada, but also by other national and subject discovery platforms, e.g. Australian ANDS or DataverseEU. The GeoBlacklight stack (http://geoblacklight.org/) is among the best-maintained and most widely used open source geographical data discovery applications; it is library- developed (mostly in Stanford) and supported by several large US research universities. GeoBlacklight can support harvesting from multiple data

repositories, and can easily support harvesting from other repositories with geospatial metadata in the future. The national GeoBlacklight discovery instance will be a complement to FRDR, and will be integrated into the same system architecture of FRDR for metadata harvesting and indexing.

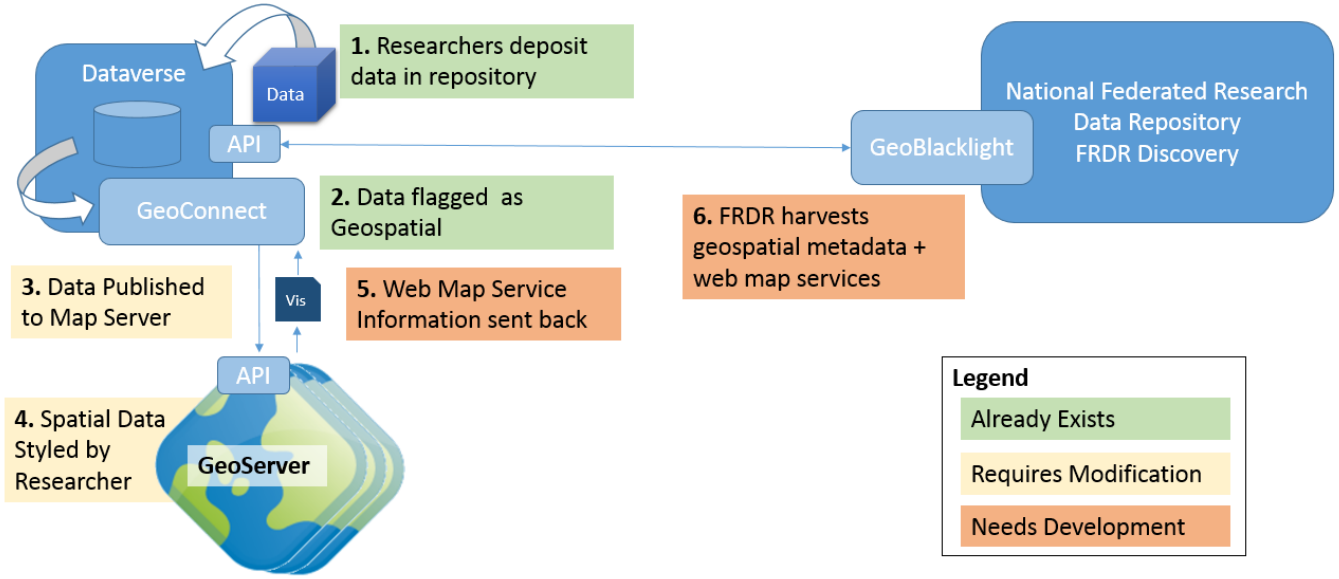
In addition to answering the questions above, applicants will also be scored on the following criteria:

1. **Completeness/quality/sustainability of the project/design.**
2. **Overall assessment of the project**

System Architecture



Software Architecture



Software Development Summary

Our proposal will integrate Dataverse with a leading open source GIS server platform (GeoServer) to expose GIS data deposited in DV using OGC web mapping services. This will allow the data to be visualized in DV and to be consumed and analyzed in other GIS platforms – server, desktop and mobile. Geospatial resources loaded into DV will be harvested into FRDR to support discovery at a national level. The geospatial search features of GeoBlacklight will be integrated into the FRDR search portal to support location-based searching based on coordinate data derived from DV and other sources already being harvested by FRDR.

In order to support geospatial data discovery, however, FRDR will need to accommodate additional metadata, including spatial coordinates and spatial features and attributes. Furthermore, to support visualization and exploration of the data, FRDR will need to connect to repository-hosted web map

services. Building a national geospatial data discovery service requires appropriate connections to data repositories in order to support extended disciplinary metadata harvesting and reuse of map visualization services presented on an open map interface.

There are number of open source software parts for this project, including:

Dataverse

DV (http://dataverse.org) is an open-source research data repository software platform that is used by many institutions and organizations, including many in Canada (e.g. Scholars Portal and the Ontario Council of University Libraries (OCUL), Abacus Dataverse at UBC, and more), for delivering library supported research data repository services to academic researchers. Increasingly, DV is being adopted by researchers and research teams as an easy-to-use and feature rich data repository, especially for those researchers that do not already have disciplinary or domain repositories. There is a clear need to enhance the existing DV instances to support the publishing and sharing of geospatial research data, including the ability to create reusable and embeddable map visualizations.

GeoConnect - WorldMap

DV currently has support for processing and packaging spatial data formats that are deposited in the system, including processing vector data in shapefile format (.shp). The DV GeoConnect application works as a middle layer, allowing geospatial data files in DV to be visualized with the Harvard University WorldMap portal (http://worldmap.harvard.edu). Harvard WorldMap provides the geospatial database and map server application component that supports the map visualization of spatial data passed through to it from DV - GeoConnect. GeoConnect can parse through the related database files (.dbf) that are packaged with shapefiles, thus allowing for metadata and data deposited in DV to be prepared, edited, and packaged in such a way that it can be systematically published to the Harvard WorldMap. While GeoConnect can presently support vector data for processing and publishing to Harvard WorldMap, there is tremendous potential to expand this middle API layer to support raster data, including georeferenced images, such as aerial photographs and digitized maps.

GeoServer

The installation and hosting of GeoServer would support geospatial visualization functionality integrated with the UBC Abacus DV. Installing and connecting DV to the open-source GeoServer map server application would extend support beyond vector data and include support for raster data types. GeoServer would be installed and configured to make API calls to and from DV, thus supporting the same functionality of GeoConnect presently, with some minor enhancements for symbolization and custom styling of map visualizations, while storing and indexing spatial data features and attribute metadata for searching and discovery. Hosting GeoServer and connecting it to UBC Abacus DV also supports an “in-Canada” solution that ensures research data are hosted on Canadian servers which can function in both of Canada’s official languages.

Metadata enhancements in Dataverse - connection to national FRDR

Much of our work will be specifically focused on metadata. The metadata that is processed by GeoConnect to support editing and packaging for publishing to the map server application can be stored as part of the dataset-level metadata in DV. FRDR can then harvest this additional geospatial metadata for those datasets that are flagged as 'geospatial' and connect these to the national GeoBlacklight discovery search catalogue for indexing spatial data features and attributes for enhanced discovery.

Who would be responsible for the Dataverse - Geoconnect - GeoServer developments?

The Lead Contractor - UBC Library, will work with Portage and CC to ensure that connections to the national FRDR discovery service from DV are operational. This will entail developments to DV and GeoConnect, the installation and configuration of a GeoServer instance at UBC Abacus Dataverse, any customizations required to the workflow for integrating these pieces, and any modifications or metadata packaging enhancements required to connect this to the national FRDR discovery service.

About FRDR, FRDR - GeoBlacklight & national discovery enhancements

FRDR is a joint initiative between CARL and CC to develop research data management infrastructure, including a national discovery service platform, for all of Canada. FRDR harvests metadata from repositories across Canada, including Canadian DV instances. It supports searching across data repositories using a core set of metadata based on the Dublin Core and DataCite metadata schemas. In order for FRDR to provide effective geospatial data discovery, enhancements in the platform will be necessary. These include, with a help of our metadata analyst, written in this proposal, extending the discovery metadata schema in FRDR to support additional geospatial metadata, including the point of access to the data's map visualization (map service URL), and feature and attribute-level metadata, which can be defined using Open Geospatial Consortium (OGC) web map service standards. The installation and configuration of GeoBlacklight, an open-source geospatial data catalogue system platform, will extend the FRDR platform to support geo-enabled discovery of the data from DV and, in the future, support other data repositories that can share geospatial metadata. The GeoBlacklight stack (http://geoblacklight.org/) is among the best-maintained and most widely used open source geographical data discovery applications. It is library-developed and supported by several large US research universities (including MIT Libraries, New York University Libraries, Princeton University Library, Stanford University Libraries, and University of Minnesota Libraries). Its architecture is largely comprised of a Ruby on Rails web framework and a Solr index, and it is designed to be deployed directly from Docker with minimal customization of the metadata schema or the web frontend CSS.

A FRDR GeoBlacklight would connect to a metadata database and be setup with a single large Solr backend to simultaneously address scalability and customizability for searching. Specific considerations will need to be made about the FRDR GeoBlacklight interface customizations, as well as optimization of the workflow for ingesting and indexing spatial metadata in the Solr index and database. This would complement the existing FRDR Open Collections discovery interface (https://frdr.ca), providing more flexibility to users while adding a relatively small

maintenance burden, and still achieving the goal of making more Canadian data discoverable from the same centralized authority.

Who would be responsible for FRDR development and discovery enhancements?

The Portage – Compute Canada FRDR team, working with the Lead Software Developer and Metadata Analyst, is the logical match for development of the discovery enhancement pieces, coordinated by the Project Manager.

Future Customization and/or Extension of Functionality

Connection to other data repositories that have geospatial data and metadata for harvesting would be an obvious extension of this system.

In a number of provinces, for example, there are several geospatial data portals, including government and library-hosted portals for licensed and open datasets, that could be harvested and integrated into FRDR. Scholars GeoPortal (http://geo.scholarsportal.info) is a geospatial data repository of licensed and government datasets maintained by SP and the OCUL. In Quebec, the Geoindex + (http://geoindex-plus.bibl.ulaval.ca/) portal provides Laval University researchers with geospatial data resources that could be made available to researchers outside of that institution through the national FRDR-GeoBlacklight discovery service. The proposed development of DV and FRDR will support future integrations between geospatial data repositories in Canada and offer researchers a robust and open place to deposit, share, discover, and reuse geospatial research data.

Project Plan

KEY TASKS LIST

Key Task		Start Date yyyy-mm-dd	End Date yyyy-mm-dd	% of total effort required for the project	Task Owner (Lead Contractor or Participant)	Milestones / Deliverables - must be tangible and measurable
1.	Hire Project Manager/Business Analyst	2018-11-01	2018-12-15	3.3%	Lead Contractor	Work with UBC HR to develop posting and hire Project Manager/Business Analyst position at UBC
2.	Hire Lead Developer	2019-01-05	2019-02-14	3.3%	Lead Contractor	Work with UBC HR to develop posting and hire Lead Developer position at UBC
3.	Hire Metadata Analyst	2019-01-05	2019-02-14	3.3%	Lead Contractor	Work with UBC HR to develop posting and hire Metadata Analyst position at UBC
4.	Setup Dataverse development environment with GeoConnect application running	2019-03-01	2019-04-01	10%	Lead Contractor	GeoConnect application available for testing and development for connecting to map service application component. (testing workflow requires current application to be setup in development environment)
5.	Setup and configure map server application (GeoServer) to connect to GeoConnect (Dataverse)	2019-04-15	2019-06-01	20%	Lead Contractor	Map server application installed and connected to development Dataverse environment

6.	Optimize the workflow to ensure scalable web map services environment and connections (e.g. setup on cloud storage / computing, application synchronization, load balancing, packaging etc.)	2019-06-15	2019-08-01	10%	Lead Contractor/ Participants	Map server application will be setup on scalable server and storage environment; Dataverse / GeoConnect / Map server application will support processing of spatial data up to 2GB in file size to render map visualizations; Map visualizations will be reliable and present visualized data without any interruptions or delays (other than those experienced by slow network connections).
7.	Develop workflows, guidelines and protocols for packaging data and metadata for harvesting by national discovery service	2019-02-15	2019-08-01	10%	Lead Contractor / Participants	Package metadata including geospatial metadata into standard metadata schema; Develop cross-walk between Dataverse / GeoConnect and GeoBlacklight schema Test enhanced metadata harvesting process Prepare metadata protocols and an ingestion manual
8.	Setup GeoBlacklight development environment / search database	2019-06-01	2019-08-01	10%	Lead Contractor / Participants	GeoBlacklight application available for testing and development Search database setup
9.	Optimize harvesting from Dataverse / GeoConnect (e.g. scheduling, testing metadata crosswalk, map visualization)	2019-08-01	2019-09-15	10%	Lead Contractor / Participants	Setup harvesting system Schedule harvesting to run on regular intervals Metadata harvested and mapped appropriately to search database
10.	Optimize search indexing for geo meta(data)	2019-09-15	2019-10-15	5%	Lead Contractor / Participants	Metadata are findable through search interface, indexed according to best practices, fast, reliable search results retrieved

11.	Customization, Training, and Promotion	2019-10-15	2019-12-15	5%	Lead Contractor / Participants	Branding applied to applications User Guide development Training development Promotional and marketing strategy development
12.	Demo release for testing	2019-12-15	2020-03-01	5%	Lead Contractor/ Participants	Demo release available for public use Ongoing testing feedback / consultation
13.	Final production release	2020-03-01	2020-03-30	5%	Lead Contractor / Participants	Production release available for public use Regular updates as scheduled

FEATURES LIST

Deliverable	Feature #	Feature Description
Dataverse setup with GeoConnect	1	Geospatial data is deposited in Dataverse and users are presented with a "View Map" feature; GeoConnect form allows researcher to add additional details about the data, and the system extracts geospatial coordinates and attributes from data for indexing in the search.
Map server application setup and connected to GeoConnect / Dataverse	2	Web map services will be produced upon deposit and confirmation by the researcher using the Dataverse GeoConnect feature "View Map" and "Publish"; Web Services will allow for spatial data to be presented on a map interface and embedded in the Dataverse application for simple, exploration, viewing, and reuse by others.

Optimization of web server application and connection to GeoConnect / Dataverse	3	Support for scalable deployment of processing of geospatial data to create web map services will support faster more reliable service for researchers uploading and publishing data to the system.
Develop/support standard schema and protocols for geospatial metadata packages	4	Using metadata standards like those supported by Dataverse, including FGDC and ISO 19115, as well as Open Geospatial Consortium (OGC) for web map services, develop cross-walks between these systems and packages to support their reuse by other systems thus developing interoperability between systems. The systems will support Open APIs and standard metadata at all stages.
GeoBlacklight and search database setup	5	(out-of-the-box) GeoBlacklight allows for geospatial data discovery and searching through the indexing of geospatial metadata and display of map visualization; Search database will be setup and connected to GeoBlacklight to support metadata store, as well as, indexing features
Harvesting system for geospatial meta(data)	6	Harvesting system will use OAI-PMH and APIs for the exchange of geospatial metadata including descriptive and attribute-level metadata. Metadata cross walking between the system schemas will be required and delivered as part of this project. Input from relevant experts including librarians, researchers, and technical experts will be consulted at all stages.
Search indexing setup for geospatial meta(data)	7	Search indexing on a number of metadata fields will improve search performance and user experience when using the discovery search features. Indexing attributes found within geospatial data will increase discovery of research data and drive reuse of the data.
Customization, training, promotion	8	Certain customizations such as search fields, facets, as well, as interface design customizations including branding, local service connections, translation, etc. will all be considered. Training and promotion will include the development of open training materials such as User Guides, tutorials, etc. as well as a promotion and

		outreach strategy and plan that make include conference presentations, webinars, and workshops to communicate about the project to researchers.
Beta / Demo GeoShare Portal release	9	A demo / beta release of the product will increase promotion and usability of the service. It will allow for timely consultation and feedback with the research community to ensure an effective release. Usability testing.
Production GeoShare Portal release	10	Final production release of the GeoShare Portal (name tbd later) will be incorporated into existing application / software releases (including a release of the SP Dataverse and FRDR discovery platform) and the GeoShare Portal will be available for all researchers in Canada.

Risk Assessment and Mitigation Plan

Risk	Mitigation Plan
Hiring could be delayed by a lack of qualified candidates applying / selected during the hiring process	Start right away if we are successful with this application, develop posting early enough to ensure timely cooperation with HR. Develop innovative strategies for promotion and outreach beyond traditional hiring channels.
System and developer environments at two different locations (UBC, and Compute Canada) and with different technologies (the number of systems to learn)	Project managers / Business analyst will use her skills to coordinate team efforts between various physical locations. We will also Introduce Lead Developer to all technologies and system environments from the beginning. This will ensure enough time to become familiar and work through any issues (connectivity or otherwise).
Lack of buy-in or understanding from researchers	First, we will recruit small **faculty advisory board** from this project with at least one representative from the West and from the East. We will also develop and share open work plans, engaging with researchers and research team from the beginning. Ensure Pilot / Demo release phase is promoted and there are sufficient training materials and opportunities for researchers to get started with the new tool. In Phase 2, ensure there is a focus on promotion and training sessions to outreach to new researchers and groups that may benefit from using this tool.

Software Provenance

Lead contractor - UBC Library, working with the Principal Investigators, and Partners – Compute Canada, Scholars Portal and Portage, would authorize software releases for the new portal, consulting the faculty advisory board. Using an Agile Development Model, smaller incremental developments would be released to the development Dataverse and FRDR environments for testing at all stages of the development/project plan. A testing workflow would be established early in the project to ensure researcher needs are met and would capture the majority of activities and interactions in the system. Testing would help to fix any bugs and ensure quality control on the code development at all stages.

As part of the release package, documentation including open source GitHub code releases, GitHub documentation, and user guides, would be provided to ensure effective reuse of the new tool. We also plan to release a full project report to communicate the project to key stakeholders including researchers, libraries, and national partners.

UBC, Portage and Compute Canada would manage software releases including upgrades and patches in accordance with the software platform community releases, on a regular and as needed basis. We will ensure that any developments to the open-source platforms - Dataverse and/or GeoServer and/or GeoBlacklight and/or FRDR - get into the core community codebase of the relevant platforms to support the sustainability of these enhancements.

Testing Plan

Under the leadership of our Project Manager / Business Analyst, we will be using Agile and behavior-driven development (BDD) methods. Smaller incremental development of deliverables will be tested using a standard behaviour and task-oriented testing workflow. The development team will develop a set list of testing procedures that will be developed early-on and in conjunction with project partners and researchers. As part of the ongoing development of the DV and FRDR systems, regular standardized testing is incorporated into the development process.

Automated software testing and QA/QC tools, such as unit tests scripts used by the open-source platform communities, as well as metadata protocols and in-house scripts, will also be used to ensure optimal testing using automated processes.

We also plan to engage a usability and user experience consultant to create a usable and clear interface for geospatial data searching in Canada.

User Training Plan

Working closely with partners and the faculty advisory board, we intend to showcase new enhancements to DV and FRDR during a demo / pilot phase of the project, before a final production release at the end of the 18-month development period. This three months demo period will allow for sufficient time for end-user feedback and consultation about workflows and feature development. Part of this demo / beta phase will also include the development of end-user / researcher focused training and promotional materials including user guides, software documentation, and workshop material to be delivered online and in person at any number of related RDM and research events. Github documentation will also provide training for the installation of the software and use of APIs for data reuse.

There will also be an opportunity to showcase the DV and FRDR enhancements and provide training directly to the researchers' community through conferences such as IASSIST, ACMLA CARTO, Open Repositories, International Digital Curation Conference, Canadian Cartographic Association, and Congress. We also plan to share workshop materials through a registry of RDM training maintained by the Portage Network Training Expert Group (TEG). Working closely with the TEG to support DV

and FRDR training will provide additional training resources as needed throughout this cross-country project.

Maintenance and Support Plan

We will spend the last six months of this project to focus on ongoing maintenance and support of the work, thus ensuring we can effectively integrate code developments into the core system architecture(s) of each organization's respective systems. Support in year two will also ensure an effective amount of time and resources to provide training and facilitate adoption by the researchers and librarians' communities and other new research teams as they are identified.

Since DV and FRDR are existing platforms supported by UBC, and Portage – Compute Canada respectively, we anticipate these enhancements and integrations to be incorporated into the regular maintenance load of the existing development support teams once the funding runs out. UBC Library is an active developer in the open source community and as such will anticipate merging code enhancements into the Dataverse core codebase to facilitate wider adoption and sustainability for ongoing maintenance.

Intellectual Property

Dataverse and FRDR are both open-source software projects, there is no intention to commercialize or attach IP to these software enhancements. In fact, we plan to open our software developments to the community on Github to build upon!

Appendix A – Bios

Eugene Barsky is UBC's Research Data Services Librarian, a faculty position. He has expertise in research data management, including data discovery and curation, and has authored 37 publications and 45 conference presentations. Importantly for this application, he plays a leadership role in various national research data initiatives, chairing the Portage Data Discovery Expert Group and co-chairing the Business Models Group for the soon-to-be-launched Dataverse North. He is the recipient of numerous awards in library science, including the Physics-Astronomy-Mathematics Award (Special Libraries Assoc., 2015), Innovation in Access to Engineering Information Award (American Soc. of Engin. Ed., 2011), and Emerging Leader Award (Can. Health Libraries, 2007). Eugene is an adjunct faculty member at the iSchool at UBC, teaching courses in science librarianship and research data management, and is an active member of the Pacific Northwest data curators group.

Lee Wilson is the Service Manager for Portage, a national, library-based network that builds capacity and coordinates RDM activities in Canada. He is on secondment from ACENET, where he works as a Research Consultant specializing in Research Data Management. Prior to taking on this role, Lee worked with the Marine Environmental Observation Prediction and Response (MEOPAR) network's Data Management Project solving issues related to the storage, discovery, and accessibility of ocean data.

Jason Brodeur is the Manager of Maps, Data, GIS in the McMaster University Library. He received his Ph.D. at McMaster University, where he studied the uncertainty associated with estimates of forest-atmosphere greenhouse gas exchanges. In his work, he is involved in a number of initiatives that relate well to this proposal: As chair of the Portage Data Curation Expert Group, he is leading the development of a national approach to data curation in Canada. As previous moderator of the Ontario Council of University Libraries (OCUL) Geo Community, he was a co-lead in a collaborative map digitization project that released of over 1000 historical maps of Ontario as discoverable and accessible geospatial data--this project was highlighted in 2017 to mark the 50th Anniversary of OCUL.

Marcel Fortin is the Head of the Map and Data Library at the University of Toronto. He is responsible for data, Research Data Management, GIS, and maps services and collections at the University of Toronto Libraries where he has worked for the past 20 years. He is a graduate of the University of Ottawa, Western University, and the University of British Columbia. In 2014, along with Historian Jennifer Bonnell of York University, he produced the edited work *Historical GIS Research in Canada*, published by the University of Calgary Press. For the past three years, he was also the principal investigator for the SSHRC funded Partnership Development Grant titled the Canadian Historical GIS Partnership (http://geohist.ca), whose aim was to build an HGIS community of researchers in Canada and beyond, to share best practices for developing historical data and research methods.

Amber Leahey is the Data & GIS Librarian at Scholars Portal, a shared digital library project of the Ontario Council of University Libraries (OCUL) based at the University of Toronto. She is responsible for managing the data and GIS services team at SP to deliver library services in Ontario and beyond. In addition to her work at SP, Amber chairs and co-chairs various RDM related groups, including the Federated Research Data Repository (FRDR) Discovery Working Group and the Data Documentation Initiative (DDI) Training Group. She is an active member of the Dataverse repository community, including Dataverse North community.

Appendix B – Letter(s) of Commitment

UBC Library

UBC ARC

CARL Portage

Compute Canada

Scholars Portal

Canadian Historical Geographic Information Systems Partnership (CHGIS)

McMaster