Dataverse North
Metadata Best Practices Guide

This guide was produced by the Metadata Subgroup of the Dataverse North Working Group on behalf of the Canadian Association of Research Libraries (CARL) with permission from Harvard for the use of definitions and the Texas Digital Library for basic design.

Version 2.0

February 2020

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

INTRODUCTION 2

FEATURES OF THE GUIDE 2
METADATA LANGUAGE 3
VERSIONS 3
QUESTIONS? 3

CITATION METADATA 4

CITATION METADATA BLOCK 4

GEOSPATIAL METADATA 12

INTRODUCTION 12
GEOSPATIAL METADATA BLOCK 13
EXAMPLES FROM REAL DATASETS 15

SOCIAL SCIENCE AND HUMANITIES METADATA 19

INTRODUCTION 19
CONTROLLED VOCABULARIES 19
SOCIAL SCIENCE AND HUMANITIES METADATA BLOCK 20

Introduction

One of the best features of the Dataverse repository platform is the large number of metadata fields it provides for describing research data. Standards-based for interoperability, Dataverse metadata supports both dataset and file-level descriptions, and is compliant with the DataCite[1] schema to support DOI registration. It draws principal influence from DDI Codebook[2], while incorporating metadata standards from other domains, making it well suited for describing data from the social, economic, behavioural, and health sciences; it is also easily adapted for use in the humanities, and in pure, applied, and environmental sciences. This flexibility can make Dataverse metadata seem complicated, especially to anyone new to Dataverse or Research Data Management (RDM) who may be wondering which fields to use or how to interpret them. This guide provides direction to both the novice and experienced user in creating metadata for datasets in a Dataverse repository.

Features of the guide

This guide:

- Provides definitions for each field and tips for clarification where needed. Fields are presented in the order in which they appear in the Dataverse interface. Note: Dataverse super-administrators can add tips to the properties files that will then display when users hover over field names.
- Distinguishes between required, recommended and optional fields.
    - Required fields are designated by the Dataverse system out-of-the-box. They are Title, Author's Name, Contact Email, Description Text and Subject.
      Note: Dataverse Administrators can adjust the settings to make additional fields required. See the Dataverse Users Guide for details.[3]
    - Recommended fields are considered best practice to improve data discovery and reuse. Two recommended fields (Contact Name and Producer) were deemed by the guide's authors to be as important as the required fields and could be amended as described above.
    - Optional fields are for additional information that may be available.

[1] "DataCite is a leading global non-profit organisation that provides persistent identifiers (DOIs) for research data and other research outputs." (https://datacite.org/mission.html)

[2] "DDI Codebook was the first version of the DDI specification to be published" (http://www.ddialliance.org/Specification/DDI-Codebook/). The Data Documentation Initiative (DDI) is an effort to create an international standard for describing data from the social, behavioral, and economic sciences. Expressed in XML (http://www.w3.org/XML/), the DDI metadata specification now supports the entire research data life cycle. DDI metadata accompanies and enables data conceptualization, collection, processing, distribution, discovery, analysis, repurposing, and archiving. The DDI Alliance (http://www.ddialliance.org/) oversees the development of the DDI metadata standard.

[3] See: http://guides.dataverse.org/en/latest/user/dataverse-management.html. Note: It is possible to make changes to the metadata blocks (e.g. make some fields required or remove domain-specific metadata blocks that are not relevant) on the General Information page; however, any inherited templates (e.g. those with Creative Commons licenses) are lost in the process. The solution is to first copy an inherited template and make it the default, and then make desired changes to the metadata blocks.

- Illustrates the use of each field with a made-up example.
    - A fictitious example is used to show fields in the Citation metadata block. The complete sample dataset can be found on the Demo Scholars Portal Dataverse platform at: https://doi.org/10.5072/FK2/TOXB6Q
    - Real examples from existing Dataverses are used to show the Geospatial and Social Science & Humanities metadata blocks.

Metadata language

The development of the Internationalization feature (added in January 2019) has provided Dataverse users around the world the ability to offer platforms in multiple languages. Scholars Portal Dataverse, for example, provides a bilingual interface with a French/English toggle. In addition to the platform, there is the language of the metadata to consider, a decision that rests, in most cases, with the dataset owner. It is recommended that the language be selected to maximize discovery by the intended audiences. In some cases, it will be beneficial to enter metadata in more than one language, taking advantage of the "Alternative Title" field, and repeatable fields such as Description, Keyword, and Geospatial Coverage.

Versions

- **Version 1, April 2019** - general citation metadata block for Dataverse 4.x
  Authors: Alexandra Cooper, Ève Paquette-Bigras, Martine Gagnon, Amber Leahey, Laure Perrier, Michael Steeleworthy, Sally Taylor
- **Version 1.1, June 2019** - general citation metadata block for Dataverse 4.x; updated to include corrections on the following fields - contact name and affiliation, ID type, ID number, producer name.
  Authors: Alexandra Cooper, Martine Gagnon, Mark Goodwin, John Huck, Amber Leahey, Michael Steeleworthy, Sally Taylor
- **Version 2, February 2020** - domain specific metadata blocks for geospatial, and social science and humanities added.
  Authors: Teresa Bascik, Philippe Boisvert, Alexandra Cooper, Martine Gagnon, Mark Goodwin, John Huck, Amber Leahey, Michael Steeleworthy, Sally Taylor
- **Future versions** - domain specific metadata blocks for life sciences, and astronomy and astrophysics.

Questions?

Please see the list of library contacts at your institution.
https://portagenetwork.ca/planning-managing-data/contacts-at-your-organization/

Citation Metadata

Citation Metadata Block

Field	Definition with tips	Required/Recommended/Optional	Example
Title	Full title by which the Dataset is known.	required	Social Media Use Among Teens, 2015 [Canada]
Subtitle	A secondary title used to amplify or state certain limitations on the main title. *Tip: subtitle is not included in generated citation. Include subtitle with title to be included in citation.*	recommended (if applicable)	Main Survey
Alternative Title	A title by which the work is commonly referred or an abbreviation of the title. *Tip: Acronym, short form or translation of full title.*	optional	Youth Social Media Survey
Alternative URL	A URL where the Dataset can be viewed, such as a personal or project website.	optional	http://youthsocialmedia.org
Other ID	Another unique identifier that identifies this Dataset (e.g., producer's or another repository's number). Consists of 2 subfields.		
Agency	Name of agency that generated this identifier.	optional	Youth Communication Development Project, Education Department, Queen's University
Identifier	Other identifier that corresponds to this Dataset.	optional	2202
Author	Person(s), corporate body(ies), or agency(ies) responsible for creating the work. Consists of 4 subfields.		
Name	The author's Family Name, Given Name or the name of the organization responsible for this Dataset.	required	Doe, Jane
Affiliation	The organization with which the author is affiliated.	recommended	Queen's University
Identifier Scheme	Name of the identifier scheme (ORCID, ISNI). *Tip: ORCID is a non-proprietary alphanumeric code to identify scientific and other academic authors and contributors uniquely.*	recommended	ORCID
Identifier	Uniquely identifies an individual author or organization, according to various schemes.	recommended	1111111

Contact	Contact(s) for this Dataset. Consists of 3 subfields.		
Name	The contact's Family Name, Given Name or the name of the organization.	recommended	Doe, Jane
Affiliation	The organization with which the contact is affiliated.	recommended	Queen's University
E-mail	The e-mail address(es) of the contact(s) for the Dataset. This will not be displayed.	required	jdoe@email.com
Description	Summary describing the purpose, nature, and scope of the Dataset. Consists of 2 subfields.		
Text	A summary describing the purpose, nature, and scope of the Dataset.	required	The Social Media Use Among Teens survey was conducted by the Youth Communication Development Project to understand social media communication behaviours among youth in Canada. The survey collected responses from Canadian youth using an online questionnaire that asks about social media use including, platform type, frequency of use, activity type, and location of use. This information is supplemented with the respondent’s demographic and household characteristics.
Date	In cases where a Dataset contains more than one description (for example, one might be supplied by the data producer and another prepared by the data repository where the data are deposited), the date attribute is used to distinguish between the two descriptions. The date attribute follows the ISO convention of YYYY-MM-DD.	optional	2018-01-18
Subject	Domain-specific Subject Categories that are topically relevant to the Dataset.	required	Social Sciences
Keyword	Key terms that describe important aspects of the Dataset. Consists of 3 subfields.		
Term	Key terms that describe important aspects of the Dataset. Can be used for building keyword indexes and for classification and retrieval purposes. A controlled vocabulary can be employed.	recommended	Social media, Communication

Vocabulary	For the specification of the keyword controlled vocabulary in use, such as LCSH, MeSH, or others. *Tip: controlled vocabulary is a standardized list of terminology for describing information (e.g. LCSH is Library of Congress Subject Heading, MeSH is Medical Subject. Heading).*	optional	Government of Canada Core Subject Thesaurus
Vocabulary URL	Keyword vocabulary URL points to the web presence that describes the keyword vocabulary, if appropriate. Enter an absolute URL where the keyword vocabulary web site is found, such as http://www.my.org.	optional	http://www.thesaurus.gc.ca/recherche-search/mtwdk.exe?k=these&l=60&w=4790&n=1&s=5&t=2
Topic Classification	Classification field indicates the broad important topic(s) and subjects that the data cover. Consists of 3 subfields.		
Term	Topic or Subject term that is relevant to this Dataset.	optional	Society and Culture
Vocabulary	*Provided for specification of the controlled vocabulary in use, e.g. LCSH, MeSH, etc.* *Tip: controlled vocabulary is a standardized list of terminology for describing information (e.g. LCSH is Library of Congress Subject Heading, MeSH is Medical Subject Heading).*	optional	Government of Canada Core Subject Thesaurus
Vocabulary URL	Specifies the URL location for the full controlled vocabulary.	optional	http://www.thesaurus.gc.ca/recherche-search/mtwdk.exe?k=these&l=60&n=0&s=cid&t=&w=97&h=SO%20Society%20and%20Culture
Related Publication	Publications that use the data from this Dataset. Consists of 4 subfields.		
Citation	Other identifier that corresponds to this Dataset. *Tip: The full bibliographic citation for any related publication.*	recommended (if applicable)	Doe, Jane. (2017). Teen use of social media: analysis of self-reported communication behaviours. Journal of Social Media Use. Vol 1. Iss. 1, 2017.
ID Type	The type of digital identifier used for this publication (e.g., Digital Object Identifier (DOI), handle, ISBN). *Tip: DOIs and handles are persistent identifiers used to identify digital objects uniquely.*	recommended (if applicable)	doi
ID Number	The identifier for the selected ID type.	recommended (if applicable)	10.0000/SP/TEST

URL	Link to the publication web page (e.g., journal article page, archive record page, or other).	optional	https://doi.org/10.0000/SP/TEST
Notes	Additional important information about the Dataset.	optional	This survey was administered online. Mode of interview has been found to impact results, therefore it is not recommended that these results are compared with other survey results where the interview mode was telephone based.
Language	Language of the Dataset	optional	English
Producer	Person or organization with the financial or administrative responsibility over this Dataset. Consists of 5 subfields.		
Name	Producer name	recommended	Youth Communication Development Project
Affiliation	The organization with which the producer is affiliated.	recommended	Queen's University
Abbreviation	The abbreviation by which the producer is commonly known. (ex. IQSS, ICPSR)	optional	YCDP
URL	Producer URL points to the producer's web presence, if appropriate. Enter an absolute URL where the producer's web site is found, such as http://www.my.org.	optional	http://youthsocialmedia.org
Logo URL	URL for the producer's logo, which points to this producer's web-accessible logo image. Enter an absolute URL where the producer's logo image is found, such as http://www.my.org/images/logo.gif.	optional	http://youthsocialmedia.org/image.png
Production Date	Date when the data collection or other materials were produced (not distributed, published or archived). *Tip: date when dataset was finalized and ready for analysis or distribution.*	recommended	2016-01-11
Production Place	The location where the data collection and any other related materials were produced.	recommended	Kingston, Ontario, Canada
Contributor	Organization or person responsible for either collecting, managing, or otherwise contributing in some form to the development of the resource. Consists of 2 subfields.		
Type	The type of contributor of the resource.	recommended	Researcher
Name	The Family Name, Given Name or organization name of the contributor.	recommended	Doe, Jane

Grant Information	Grant information. Consists of 2 subfields.		
Grant Agency	Grant Number Agency	recommended (if applicable)	Social Sciences and Humanities Research Council (SSHRC)
Grant Number	The grant or contract number of the project that sponsored the effort.	recommended (if applicable)	CCB123456
Distributor	Organization designated by the author or producer to generate copies of the particular work including any necessary editions or revisions. Consists of 5 subfields.		
Name	Distributor name	recommended	Data Services
Affiliation	The organization with which the distributor contact is affiliated.	recommended	Queen's University Library
Abbreviation	The abbreviation by which this distributor is commonly known (e.g., IQSS, ICPSR).	optional	QUL
URL	Distributor URL points to the distributor's web presence, if appropriate. Enter an absolute URL where the distributor's web site is found, such as http://www.my.org.	optional	http://library.queensu.ca/data/services
Logo URL	URL of the distributor's logo, which points to this distributor's web-accessible logo image. Enter an absolute URL where the distributor's logo image is found, such as http://www.my.org/images/logo.gif.	optional	http://www.queensu.ca/encyclopedia/sites/webpublish.queensu.ca.qencwww/files/images/l/logo/QueensLogo_colour.png
Distribution Date	Date that the work was made available for distribution/presentation. *Tip: This field may be the same as the Deposit Date. Use the field if data was previously distributed.*	optional	2018-01-22
Depositor	The person (Family Name, Given Name) or the name of the organization that deposited this Dataset to the repository. *Tip: The name of the person/institution who provided the dataset(s) to the archive (i.e. not necessarily the person doing the submission into DV).*	recommended	Doe, Jane

Deposit Date	Date that the Dataset was deposited into the repository. *Tip: Date is pre-populated with the date of upload into Dataverse. It can be edited to reflect the date when the data was received by an external or mediated data repository service.*	recommended	2018-01-15
Time Period Covered	Time period to which the data refer. This item reflects the time period covered by the data, not the dates of coding or making documents machine-readable or the dates the data were collected. Also known as span. Consists of 2 subfields.		
Start	Start date that reflects the time period covered by the data, not the dates of coding or making documents machine-readable or the dates the data were collected.	recommended	2015-03-20
End	End date that reflects the time period covered by the data, not the dates of coding or making documents machine-readable or the dates the data were collected.	recommended	2015-06-21
Date of Collection	Date(s) when the data were collected. Consists of 2 subfields.		
Start	Date when the data collection started.	recommended	2015-03-20
End	Date when the data collection ended.	recommended	2015-06-21
Kind of Data	Type of data included in the file: survey data, census/enumeration data, aggregate data, clinical data, event/transaction data, program source code, machine-readable text, administrative records data, experimental data, psychological test, textual data, coded textual, coded documents, time budget diaries, observation data/ratings, process-produced data, or other.	recommended	Survey data
Series	Information about the Dataset series. Consists of 2 subfields.		
Name	Name of the dataset series to which the Dataset belongs.	recommended (if applicable)	Social Media Use Among Teens
Information	History of the series and summary of those features that apply to the series as a whole.	recommended (if applicable)	Established in 2005, the Youth Communication Development Project aims to gather key research and data about youth development and social media use through a series of independent, annual, cross-sectional surveys titled Social Media Use Among Teens. The overall objectives of the program is to gather data

			on youth and social media trends in order to monitor changes in the well-being of young Canadians, and to provide information on specific social policy issues.
Software	Information about the software used to generate the Dataset. Consists of 2 subfields.		
Name	Name of software used to generate the Dataset. *Tip: useful for specialized software or instruments.*	optional	SPSS
Version	Version of the software used to generate the Dataset.	optional	24
Related Material	Any material related to this Dataset.	optional	Youth Social Media Trends: 2015 Report [Canada]. YCDP, Queen's University, 2016. Access URL: http://dataverse.scholarsportal.info/queensu/2016report.pdf
Related Datasets	Any Datasets that are related to this Dataset, such as previous research on this subject.	optional	Social Media Use Among Teens, 2010 [Canada]. YCDP, Queen's University, 2011. DOI. Access URL: http://dataverse.scholarsportal.info/queensu/2010data.xhtml
Other References	Any references that would serve as background or supporting material to this Dataset.	optional	Social Media Use Among Teens: Survey Questionnaire, 2015 [Canada]. YCDP, Queen's University, 2016. DOI. Access URL: http://dataverse.scholarsportal.info/queensu/2016questionnaire.pdf
Data Sources	List of books, articles, serials, or machine-readable data files that served as the sources of the data collection.	optional	Statistics Canada. National Household Survey, 2011: Median Household Income by Census Tracts, Census Metropolitan Areas. NHS 2011, Statistics Canada. Access URL: https://www12.statcan.gc.ca/nhs-enm/2011/dp-pd/prof/index.cfm?Lang=E
Origin of Sources	For historical materials, information about the origin of the sources and the rules followed in establishing the sources should be specified.	optional	National Household Survey, 2011. http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5178
Characteristic of Sources Noted	Assessment of characteristics and source material. *Tip: describes noteworthy aspects of the data collected.*	optional	

Documentation and Access to Sources	Level of documentation of the original sources. *Tip: can be used to explain any restrictions or access to source data documentation.*	optional	Open

Geospatial Metadata

Introduction

Geospatial metadata can describe maps, GIS files, or other location-based data. Any dataset associated with a location should include geospatial metadata in addition to the general citation metadata block. At a minimum, provide place names to describe locations in your data and use GeoNames.org to confirm these terms. Alternate names (e.g., in other languages) may be added. If applicable, enter bounding box[4] coordinates to allow the data to be findable with map-based search tools.

The table below:

- Provides definitions for each field and tips for clarification where needed.
- Distinguishes between strongly recommended, recommended and optional fields.
  Note: No geospatial fields are required in a basic installation of Dataverse; however, Dataverse Administrators can adjust the settings to make additional fields required (e.g. Country). See the Dataverse Users Guide for details.[5]
- Illustrates the use of each field with a made-up example.

Following the table are examples taken from real datasets to illustrate how fields relate to each other.

According to Dataverse documentation, Geospatial Metadata fields are compliant with DDI Lite, DDI 2.5 Codebook, DataCite, and Dublin Core. The Country / Nation field uses ISO 3166-1 controlled vocabulary.[6]

[4] A bounding box is an area defined by two longitudes and two latitudes. https://wiki.openstreetmap.org/wiki/Bounding_Box

[5] See: http://guides.dataverse.org/en/latest/user/dataverse-management.html

[6] See: http://guides.dataverse.org/en/latest/user/appendix.html

Geospatial Metadata Block

Field	Definition with tips	Strongly Recommended/ Recommended/ Optional	Example
Geographic Coverage	Information on the geographic coverage of the data. Includes the total geographic scope of the data. Consists of 4 subfields. *Tip: for consistency, use the Geonames database to check the form and spelling of place names: https://www.geonames.org/*		
Country / Nation	The country or nation that the Dataset is about. *Tip: select from drop-down list of names from ISO-3166. If dataset covers multiple countries, list all of them.*	Strongly recommended	Canada
State / Province	The state or province that the Dataset is about. Use GeoNames for correct spelling and avoid abbreviations. *Tip: if using this field, also include Country to disambiguate.*	Recommended	British Columbia
City	The name of the city that the Dataset is about. Use GeoNames for correct spelling and avoid abbreviations. *Tip: if using this field, also include State/Province AND Country to disambiguate.*	Recommended	Vancouver
Other	Other information on the geographic coverage of the data. *Tip: use for geographical names that are not a country, state/province or city, e.g. regions, water bodies, astronomy names. If applicable, disambiguate by including City AND/OR State/Province AND/OR Country.*	Optional	Jericho Beach Park Musqueam Park Pacific Spirit Regional Park Stanley Park Vanier Park
Geographic Unit	Lowest level of geographic aggregation covered by the Dataset, e.g., village, county, region. *Tip: use when the lowest geographic level that can be analyzed in the dataset is different from the dataset's entire area. (e.g. when a dataset about parks in Vancouver can be faceted by the individual parks)*	Optional	park

Geographic Bounding Box	The fundamental geometric description for any Dataset that models geography is the geographic bounding box. It describes the minimum box, defined by west and east longitudes and north and south latitudes, which includes the largest geographic extent of the Dataset's geographic coverage. This element is used in the first pass of a coordinate-based search. Inclusion of this element in the codebook is recommended but is required if the bound polygon box is included. Consists of 4 subfields. *Tip: to determine bounding box, use: http://boundingbox.klokantech.com/*		
West Longitude	Westernmost coordinate delimiting the geographic extent of the Dataset. A valid range of values, expressed in decimal degrees, is -180,0 <= West Bounding Longitude Value <= 180,0.	Recommended	-123.265
East Longitude	Easternmost coordinate delimiting the geographic extent of the Dataset. A valid range of values, expressed in decimal degrees, is -180,0 <= East Bounding Longitude Value <= 180,0.	Recommended	-123.115
North Latitude	Northernmost coordinate delimiting the geographic extent of the Dataset. A valid range of values, expressed in decimal degrees, is -90,0 <= North Bounding Latitude Value <= 90,0.	Recommended	49.314
South Latitude	Southernmost coordinate delimiting the geographic extent of the Dataset. A valid range of values, expressed in decimal degrees, is -90,0 <= South Bounding Latitude Value <= 90,0.	Recommended	49.226

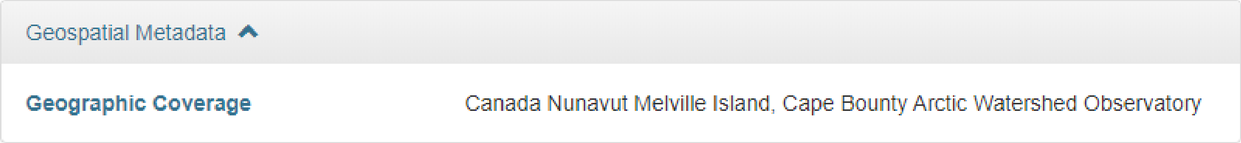
Examples from real datasets

Geographic Coverage: Other

1. In this example, “Other” indicates the Cape Bounty Arctic Watershed Observatory on Melville Island.

   Beamish, Alison; Scott, Neal; Wagner, Ioan; Neil, Allison, 2016, "Impact of active layer detachments on carbon exchange in a high-Arctic ecosystem, Cape Bounty, Nunavut, Canada (2010)", https://hdl.handle.net/10864/11825, Scholars Portal Dataverse, V2

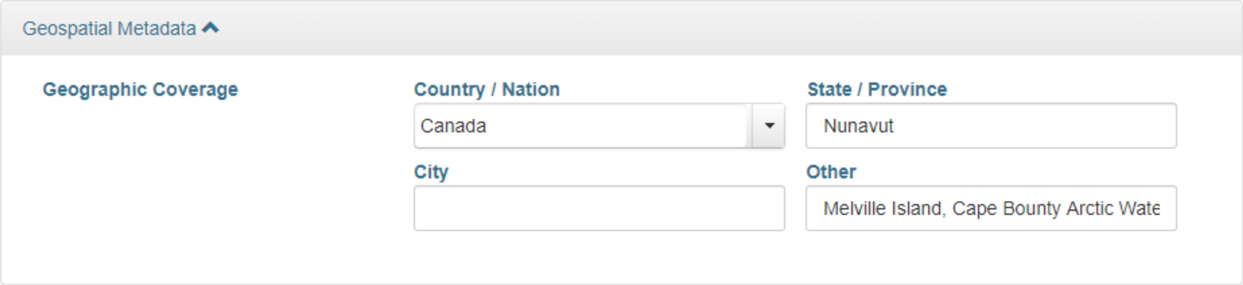
   Public view



Geospatial Metadata

Geographic Coverage: Canada Nunavut Melville Island, Cape Bounty Arctic Watershed Observatory

   Edit view



Geospatial Metadata

Geographic Coverage

Country / Nation: Canada

State / Province: Nunavut

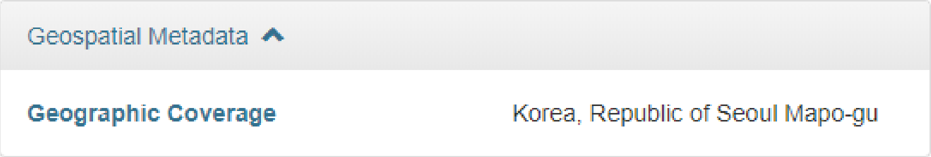
City:

Other: Melville Island, Cape Bounty Arctic Wate

2. In this example, “Other” indicates the specific district of Mapo-gu within the city of Seoul.

   Da In Choi, 2015, "Korean War Interviews, 2015", https://hdl.handle.net/10864/11174, Scholars Portal Dataverse, V5

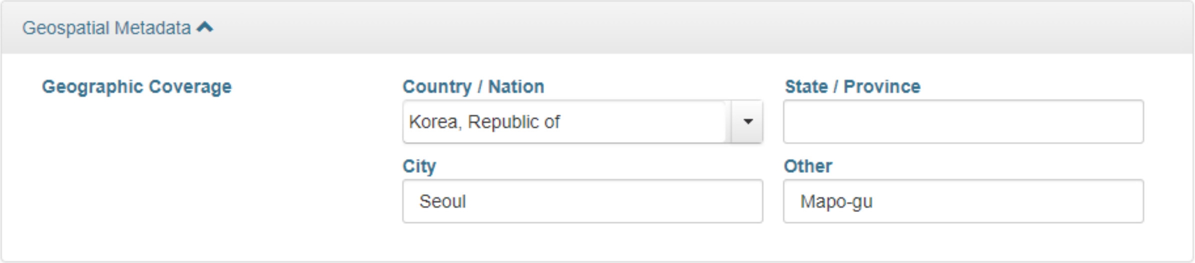
   Public view



Geospatial Metadata

Geographic Coverage: Korea, Republic of Seoul Mapo-gu

Edit view


Geospatial Metadata

Geographic Coverage

Country / Nation: Korea, Republic of

State / Province:

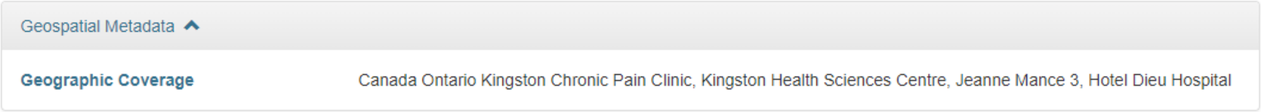
City: Seoul

Other: Mapo-gu

3. In this example, “Other” indicates a specific health clinic where the study occurred.

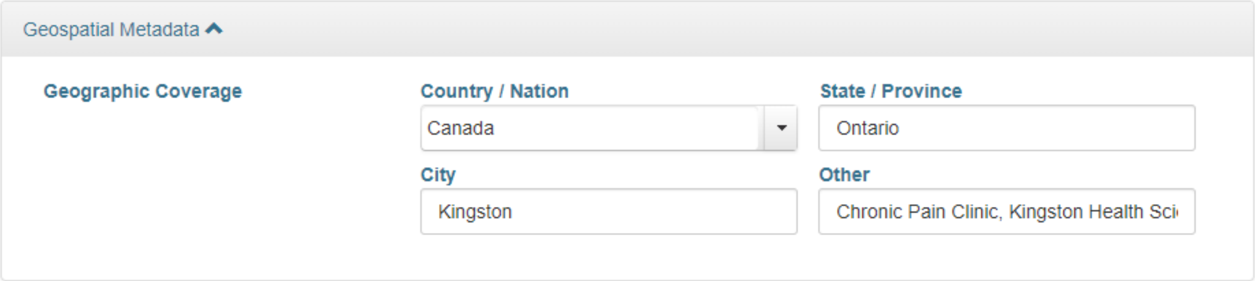
   Wilson, Rosemary A.; VanDenKerkhof, Elizabeth G.; Duggan, Scott; Gilron, Ian; Good, Mary Anne; Henry, Richard; Carley, Meg, 2018, "Chronic Pain Surveillance at Queen's, 2013-2017", https://doi.org/10.5683/SP2/GAPNRM, Scholars Portal Dataverse, V1, UNF:6:d+jC5YYQZO7ERTS1Y37v0Q== [fileUNF]

   Public view


Geospatial Metadata

Geographic Coverage: Canada Ontario Kingston Chronic Pain Clinic, Kingston Health Sciences Centre, Jeanne Mance 3, Hotel Dieu Hospital

   Edit view


Geospatial Metadata

Geographic Coverage

Country / Nation: Canada

State / Province: Ontario

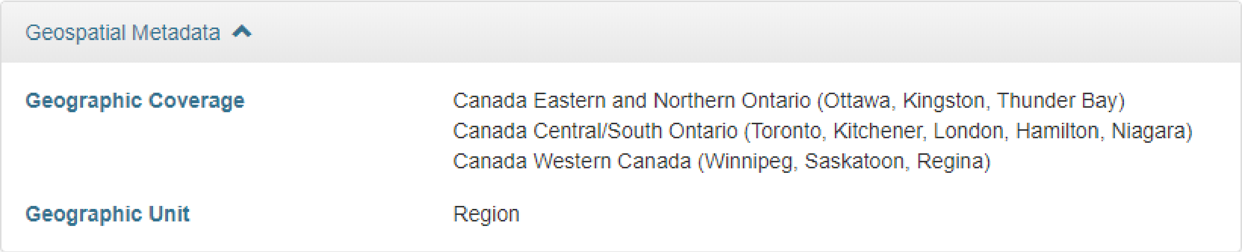
City: Kingston

Other: Chronic Pain Clinic, Kingston Health Sci

Geographic Unit

1. In this example, the Region is the lowest level of geographic aggregation covered by the Dataset.

   Margaret B. Harrison; Practice and Research in Nursing Group: Wound Care Collaborative; Ian Graham; E. Andrea Nelson; Elizabeth VanDenKerkhof; Karen Lorimer; Connie Harris; Meg Carley; The Canadian Bandaging Trial Group, 2013, "Practice and Research in Nursing (PRN) Wound Studies, 1999-2009 [Canada]", https://hdl.handle.net/10864/CORX8, Scholars Portal Dataverse, V6

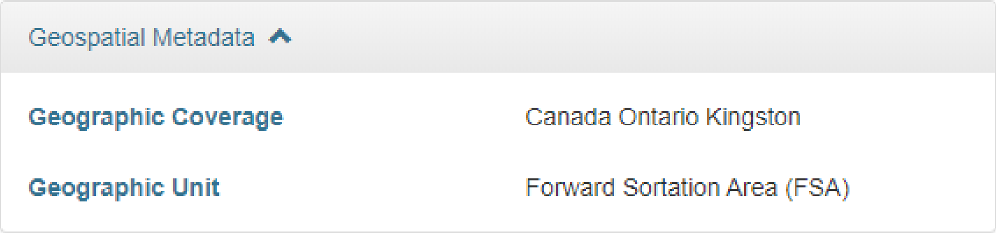


Geospatial Metadata

Geographic Coverage	Canada Eastern and Northern Ontario (Ottawa, Kingston, Thunder Bay) Canada Central/South Ontario (Toronto, Kitchener, London, Hamilton, Niagara) Canada Western Canada (Winnipeg, Saskatoon, Regina)
Geographic Unit	Region

2. In this example, the Forward Sortation Area (FSA) is the lowest level of geographic aggregation covered by the Dataset.

   Hird, Myra J.; Lougheed, Scott C.; Kuyvenhoven, Cassandra; Rowe, R. Kerry, 2016, "Perspectives on Municipal Waste Management in Kingston, Ontario, 2012", https://hdl.handle.net/10864/11926, Scholars Portal Dataverse, V1



Geospatial Metadata

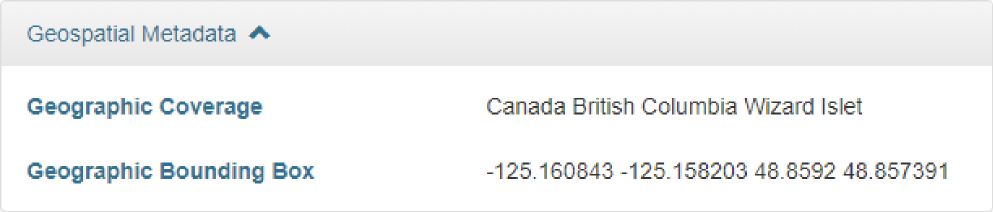
Geographic Coverage	Canada Ontario Kingston
Geographic Unit	Forward Sortation Area (FSA)

Geographic Bounding Box

1. In this example, the two longitude and two latitude coordinates define the geographic extent of the dataset.

   Anderson, Lauren; Beasley, Barb; Flumerfelt, Sidney-Rae; Fox, Caroline; Friesen, Sarah; Macfarlane, Gemma; McKay, Taesagh, 2019, "Replication data for: Long-term monitoring in Barkley Sound: a temporal analysis of intertidal biodiversity on Wizard Islet, British Columbia from 1997 to 2017", https://doi.org/10.5683/SP2/C8G480, Scholars Portal Dataverse, V1, UNF:6:mBVVVVwtuVbcT4h2Au8RXQ== [fileUNF]

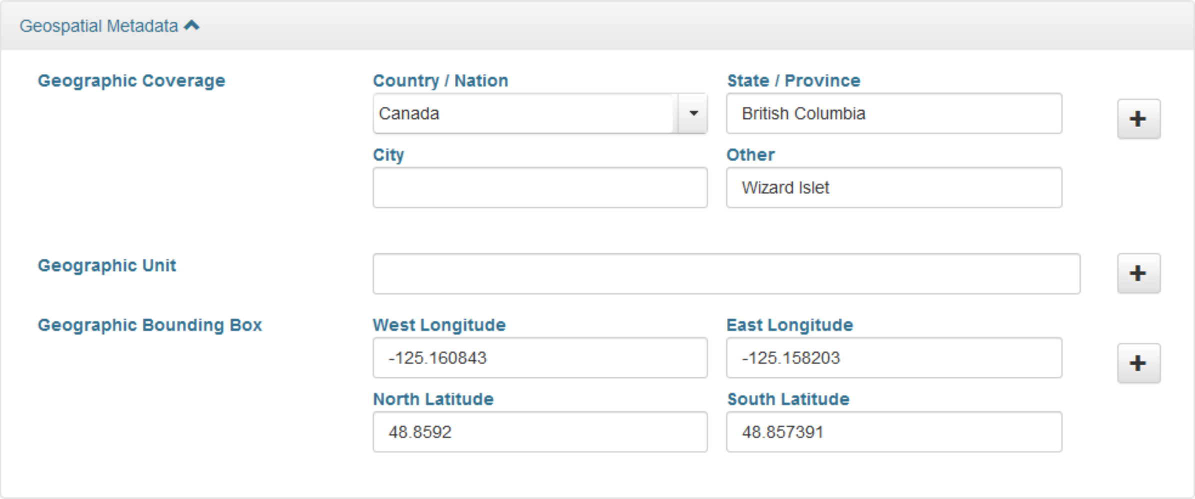
   Public view



Geospatial Metadata

Geographic Coverage	Canada British Columbia Wizard Islet
Geographic Bounding Box	-125.160843 -125.158203 48.8592 48.857391

Edit view



Geospatial Metadata

Geographic Coverage

Country / Nation: Canada

State / Province: British Columbia

City:

Other: Wizard Islet

Geographic Unit:

Geographic Bounding Box

West Longitude: -125.160843

East Longitude: -125.158203

North Latitude: 48.8592

South Latitude: 48.857391

Social Science and Humanities Metadata

Introduction

The Social Science & Humanities Metadata section builds on the general citation metadata block.

The table below:

- Provides definitions for each field and tips for clarification where needed.
- Distinguishes between recommended and optional fields. At a minimum, metadata should be included for all recommended fields, when applicable.
  Note: No social science and humanities fields are required in a basic installation of Dataverse; however, Dataverse administrators can adjust the settings to make additional fields required. See the Dataverse Users Guide for details.[7]
- Provides multiple examples to illustrate how the fields can be used. For the most part, these examples have been taken from existing Dataverse datasets, and links to the source datasets have been provided.

Controlled vocabularies

The DDI Alliance has created a set of controlled vocabularies (http://www.ddialliance.org/controlled-vocabularies) that can be used with some fields within the Social Science and Humanities section.

Controlled vocabularies are available for the following fields:

- Unit of Analysis - http://www.ddialliance.org/Specification/DDI-CV/AnalysisUnit_1.0.html
- Time Method - http://www.ddialliance.org/Specification/DDI-CV/TimeMethod_1.2.html
- Sampling Procedure - http://www.ddialliance.org/Specification/DDI-CV/SamplingProcedure_1.1.html
- Collection Mode - http://www.ddialliance.org/Specification/DDI-CV/ModeOfCollection_3.0.html
- Type of Research Instrument - http://www.ddialliance.org/Specification/DDI-CV/TypeOfInstrument_1.1.html
- Type of Note - http://www.ddialliance.org/Specification/DDI-CV/TypeOfNote_1.0.html

According to Dataverse documentation, Social Science and Humanities fields are compliant with DDI Lite, DDI 2.5 Codebook, and Dublin Core.[8]

[7] See: http://guides.dataverse.org/en/latest/user/dataverse-management.html
[8] See: http://guides.dataverse.org/en/latest/user/appendix.html

Social Science and Humanities Metadata Block

Field	Definition with tips	Recommended/ Optional	Example
Unit of Analysis	Basic unit of analysis or observation that this Dataset describes, such as individuals, families/households, groups, institutions/organizations, administrative units, and more.	Recommended	• Individual • Family • Household
Universe	Description of the population covered by the data in the file; the group of people or other elements that are the object of the study and to which the study results refer. Age, nationality, and residence commonly help to delineate a given universe, but any number of other factors may be used, such as age limits, sex, marital status, race, ethnic group, nationality, income, veteran status, criminal convictions, and more. The universe may consist of elements other than persons, such as housing units, court cases, deaths, countries, and so on. In general, it should be possible to tell from the description of the universe whether a given individual or element is a member of the population under study. Also known as the universe of interest, population of interest, and target population.	Recommended	• Canadians aged 12-30 Source: https://doi.org/10.5683/SP/HY2H1A • Queen's University 2nd year medical students who were part of the 2016 Critical Enquiry Course in the School of Medicine and agreed to participate in the study. Source: https://doi.org/10.5683/SP/D6NISS
Time Method	The time method or time dimension of the data collection, such as panel, cross-sectional, trend, time- series, or other.	Optional	• Longitudinal • Time series • Longitudinal: Panel
Data Collector	Individual, agency or organization responsible for administering the questionnaire or interview or compiling the data.	Recommended	• Trained student interviewers, both anglophone and francophone. Source: https://hdl.handle.net/10864/ZJ17A
Collector Training	Type of training provided to the data collector	Optional	• The interviews were conducted by professional interviewers under the supervision of the Institute for Social Science Research. Source: https://doi.org/10.7910/DVN/SRVIO4

			• From the documentation: "Each staff member was thoroughly trained prior to beginning work on the survey. Interviewers received about three days of classroom training plus self-training materials. Additional study materials and classroom training were planned throughout the interviewing period. Quality control measures, such as editing returns, observing interviews and re-interviewing selected households were employed throughout the survey." Source: https://doi.org/10.7910/DVN/YT09KD
Frequency	If the data collected includes more than one point in time, indicate the frequency with which the data was collected; that is, monthly, quarterly, or other.	Optional	• Annual • Data was collected at baseline and at one month follow-up. Source: https://doi.org/10.7939/DVN/10889 • Hourly Source: https://doi.org/10.5683/SP/KYKL9M
Sampling Procedure	Type of sample and sample design used to select the survey respondents to represent the population. May include reference to the target sample size and the sampling fraction.	Recommended	• Canadians adults randomly selected from Angus Reid Forum panel members. Source: https://hdl.handle.net/10864/11510 • Telephone recruitment from random sample of 1300 telephone numbers from Utilities directory. Additional recruitment through posters, media releases, social media posts. Source: https://hdl.handle.net/10864/11926
Target Sample Size	Specific information regarding the target sample size, actual sample size, and the formula used to determine this. Consists of 2 subfields.		
Actual	Actual sample size. *Tip: The research study's actual sample size may be stated in this numeric field for reference purposes.*	Optional	• 1015
Formula	Formula used to determine target sample size. *Tip: A plain-text, general description of a sample size formula may be stated here for reference purposes. This*	Optional	• Eligible employees who lived within the following FSAs: K6V, K7A, K7C, K7G, K7H, K7K, K7L, K7M, K7N, K7P, K7R, K8N, K8P, K8R, K8V, K0E, K0G, K0H, K0K.

	may include particular methodologies, practices, and outcomes from existing scholarly literature.		
Major Deviations for Sample Design	Show correspondence as well as discrepancies between the sampled units (obtained) and available statistics for the population (age, sex-ratio, marital status, etc.) as a whole.	Optional	• The suitability of Ohio as a research site reflected its similarity to the United States as a whole. The evidence extended by Tuchfarber (1988) shows that Ohio is representative of the United States in several ways: percent urban and rural, percent of the population that is African American, median age, per capita income, percent living below the poverty level, and unemployment rate. Although results generated from an Ohio sample are not empirically generalizable to the United States, they may be suggestive of what might be expected nationally. Source: http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation.html • Oversample of persons 50 and older Source: https://doi.org/10.7910/DVN/FGTJGO
Collection Mode	Method used to collect the data; instrumentation characteristics (e.g., telephone interview, mail questionnaire, or other).	Recommended	• Interview • Paper and online questionnaire • Coded from psychiatric hospital files, court records, and police agencies. Source: https://hdl.handle.net/10864/12053
Type of Research Instrument	Type of data collection instrument used. Structured indicates an instrument in which all respondents are asked the same questions/tests, possibly with precoded answers. If a small portion of such a questionnaire includes open-ended questions, provide appropriate comments. Semi-structured indicates that the research instrument contains mainly open-ended questions. Unstructured indicates that in-depth interviews were conducted.	Recommended	• Questionnaire • Structured • Technical instrument: Static Chamber,Vaisala Humicap HM70 relative humidity/ temperature probe, Vaisala Carbocap GMP343 infrared analyzer, Hobo Pro v2 U23-003 temperature logger, Kestrel 3500 weather meter, Taylor 9878 thermometer. Source: https://hdl.handle.net/10864/11825

Characteristics of Data Collection Situation	Description of noteworthy aspects of the data collection situation. Includes information on factors such as cooperativeness of respondents, duration of interviews, number of call backs, or similar.	Optional	• There were 1,419 respondents who answered questions in telephone interviews lasting approximately 35 minutes each. Clarifications to survey questions were limited and respondents were directed to provide a response based on the information provided as to not allow interviewer bias/assumptions to influence the survey results. Source: https://hdl.handle.net/10864/ZJ17A
Actions to Minimize Losses	Summary of actions taken to minimize data loss. Include information on actions such as follow-up visits, supervisory checks, historical matching, estimation, and so on.	Optional	• Reminder e-mails were distributed to target population. Source: https://doi.org/10.5683/SP/L1H3SS • Cards reminding parents about the follow-up visit were given out. Source: https://doi.org/10.7939/DVN/10889
Control Operations	Methods to facilitate data control performed by the primary investigator or by the data archive.	Optional	• Field validation is built into REDCap data collection forms. Source: https://doi.org/10.7939/DVN/10907 • Blinded double data entry and third person cross-validation were used. Source: https://doi.org/10.7939/DVN/10900
Weighting	The use of sampling procedures might make it necessary to apply weights to produce accurate statistical results. Describes the criteria for using weights in analysis of a collection. If a weighting formula or coefficient was developed, the formula is provided, its elements are defined, and it is indicated how the formula was applied to the data.	Recommended	• Rim weighting is used with this file. By region, the file was weighted to census targets on sex (wtsex), age (Wtage), and education (Wtedu) using the 2011 census. For this file, a religion weight (wtreligion) was also included based on the 2011 National Household Survey (NHS). The wtg2 variable includes all of these weights within it. Source: https://doi.org/10.5683/SP/78RONJ • The final sample obtained for each area is not proportional to the Alberta population it makes up. For instance, Edmonton is over-sampled as shown by TABLE 1. Edmonton makes up only 24% of the Alberta population but has 43% of the interviews.

			Therefore, in order to combine the samples for a provincial sample weighting is necessary. The weighting factors used for the 1987 survey are as follows: Edmonton 0.558439, Calgary 1.151521, and Other Alberta 1.471173. Source: https://doi.org/10.7939/DVN/10567 • wtx used to correctly weight respondents against Stats Canada Alberta population estimates Source: https://doi.org/10.7939/DVN/10813
Cleaning Operations	Methods used to clean the data collection, such as consistency checking, wildcode checking, or other.	Recommended	• For income data, all respondents are matched to the tax data file unless they refuse to have their information linked. Data obtained from the tax file are complete and do not require imputation. Income figures are imputed only in the absence of tax data. Donor imputation by the nearest neighbour method is generally used and is performed primarily with Statistics Canada’s Census Edit and Imputation System (CANCEIS). However, amounts received through certain government programs such as the universal child care benefit and child tax benefits are derived from other information (e.g. number of children in the household) using a deductive imputation method. Source: http://hdl.handle.net/11272/10619 • Physiological data was reviewed for outliers. Individual breaths with tidal volume (VT), respiratory rate (RR) or minute ventilation (VE) that lay outside the 95% confidence interval for all infants were removed as outliers; 99.7% of all measured breaths were included in the final analyses. Source: https://doi.org/10.7939/DVN/10910
Study Level Error Notes	Note element used for any information annotating or clarifying the methodology and processing of the study.	Optional	• The computerized questionnaire contains many features designed to maximize the quality of the

			data collected. Many edits are built into the questionnaire to compare the reported data with unusual values and detect logical inconsistencies. When an edit fails, the interviewer is prompted to correct the information (with the respondent's help, if necessary). Once the data are transmitted to Head Office, a comprehensive series of processing steps are undertaken for the purpose of detailed verification of each questionnaire. Invalid responses are corrected or flagged for imputation. Edits were applied at a micro level. Deterministic edits and consistency edits were also performed at the micro level. Data was checked for outliers and extreme values, and were corrected at a micro level when required. Source: http://hdl.handle.net/11272/10619
Response Rate	Percentage of sample members who provided information.	Recommended	• Based on 100km radius, the survey response rate is 1874/3994 (46.9%), and the survey completion rate is 1732/3994 (43.4%). Based on FSAs for locations served by Kingston Transit, the survey response rate is 1469/3151 (46.6%), and the survey completion rate is 1356/3151 (43.0%). Source: https://doi.org/10.5683/SP/CNXSVN • At one-month follow-up: 60.2% (n=136/226). Source: https://doi.org/10.7939/DVN/10889
Estimates of Sampling Error	Measure of how precisely one can estimate a population value from a given sample. *Tip: Examples include confidence intervals, non-response, response bias.*	Recommended	• In SFS 2016, the 95% confidence interval for the average net worth of Canadian families had a width of $38,500. Source: http://hdl.handle.net/11272/10619 • + or - 2.5%; design effect of weighting not calculated Source: https://doi.org/10.7910/DVN/FGTJGO
Other Forms of Data Appraisal	Other issues pertaining to the data appraisal. Describe issues such as response variance, nonresponse rate and	Optional	• OSBD is subject to interpretation since it is an indirect behavioral measure of perceived distress.

	testing for bias, interviewer and response bias, confidence levels, question bias, or similar.		Source: https://doi.org/10.7939/DVN/10841
Notes	General notes about this Dataset. Consists of 3 subfields.		
Type	Type of note.	Optional	• Processing note
Subject	Note subject.	Optional	• Variable corrections
Text	Text for this note.	Optional	• Info (Misc) v2 note: Corrections were made to variables: PAS1MRG1, PAS1MRG2, PASRDPO1, PASRDPO2, PASRDPO3, PASRDPO4, PASRDPO5 and VERDATE. Source: http://hdl.handle.net/11272/10619