

# Metadata for Discovery:

## Disciplinary Standards and Crosswalk Progress Report

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

Amber Leahey (Scholars Portal, Chair)  
Eugene Barsky (University of British Columbia)  
John Brosz (University of Calgary)  
Alex Garnett (Simon Fraser University)  
Vincent Gray (Western University)  
Joseph Hafner (McGill University)  
Kara Handren (Scholars Portal)  
Amanda Harrigan (University of Alberta)  
Christian Lacroix (Université Laval)  
Julienne Pascoe (Canadiana.org)  
Catelynne Sahadath (University of Ottawa)  
Dany Savard (York University)  
Andrew Senior (McGill University)  
Barbara Towell (University of British Columbia)  
Lee Wilson (Portage/ACENET)

SEPTEMBER 2017

Portage Network  
Canadian Association of Research Libraries  
[portage@carl-abrc.ca](mailto:portage@carl-abrc.ca)

[www.portagenetwork.ca](http://www.portagenetwork.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

Background.....	3
Progress Report - Phase 1: 2017 .....	4
List of Standards.....	4
FRDR Discovery Profile.....	7
Standards Evaluation and Mapping / Crosswalk .....	9
Recommendations for Improved Discovery in FRDR .....	13
Next Steps - Phase 2 2017-2018.....	15
Appendix 1 .....	16

## Background

Crossing a wide range of disciplines, the purpose of the Portage Data Discovery Expert Group is to support research data creators and curators in planning, producing, and managing descriptive metadata for effective discovery and reuse. The group facilitates discussions about data discovery in Canada, to promote the use of metadata standards for research data that support both machine-to-machine and human-to-machine discovery activities.

The Data Discovery Metadata Working Group (DDMWG)<sup>1</sup> works to define the scope of metadata standards and considerations for developing the Canadian [Federated Research Data Repository](#) (FRDR)<sup>2</sup>, a national data discovery and repository service (in beta stage as of July 2017). The main purpose of the DDMWG is to identify metadata standards used by Canadian data repositories and to develop detailed crosswalks making research data discoverable. The work includes gathering a list of descriptive metadata standards, evaluating, at a granular-level, metadata elements in use across disciplines and repositories, and making recommendations for a core set of elements for discovery in FRDR. The work scoped out by this group is not only for the benefit of the FRDR project, but also for general interest by anyone involved in the management of metadata for discovery systems and research data repositories. Some discussions have also focused on outreach with other national initiatives and organizations such as Research Data Canada<sup>3</sup>, Linked Open Data in Libraries, Archives, and Museums<sup>4</sup>, and the Association of Research Libraries (ARL) SHARE Project<sup>5</sup>, for broader understanding and knowledge sharing about open research data discovery systems and tools. Moving forward, closer ties to relevant national communities and linked data initiatives may be required.

---

<sup>1</sup> Portage Data Discovery Expert Group - Metadata Working Group- <https://portagenetwork.ca/working-with-portage/network-of-expertise/data-discovery/call-for-participation/metadata-wg>

<sup>2</sup> More information about this national infrastructure project to support discovery of research is provided online through CARL Portage (<https://portagenetwork.ca/frdr-dfdr>)

<sup>3</sup> Research Data Canada <https://www.rdc-drc.ca>

<sup>4</sup> LODLAM <http://lodlam.net>

<sup>5</sup> ARL SHARE Project <http://www.share-research.org>

# Progress Report - Phase 1: 2017

## List of Standards

DDMWG formed in February 2017 and started evaluating metadata standards used by a variety of Canadian research data repositories, including Open Data Canada, Oceans Network Canada, and Canadian Dataverses<sup>6</sup>. Since this kind of work is not unique, the group evaluated existing disciplinary metadata standards mappings, crosswalks, data models, ontologies, *et cetera* for research data. These were shared and discussed via our community meeting calls and e-mail listserv. Since the work of the FRDR had already begun several months before the formation of the DDMWG, a recommendation for a data model was determined to be out-of-scope for the group (at least in the initial stages of our work). The group focused on gathering standards, including looking at repositories listed in the CISTI National Gateway to Research Data registry<sup>7</sup> and the registry of global repositories Re3data.org<sup>8</sup>, to further identify metadata standards in use across disciplinary repositories. Ultimately, relying on the rich and varied expertise within the group (ranging from metadata expertise in the social sciences, geosciences, health sciences, cultural heritage and library sectors, etc.), led to a list of over twenty disciplinary metadata standards for further evaluation (see Figure 1).

**Figure 1 - List of General and Disciplinary Metadata Standards (considered)**

Metadata Standard	General/Discipline	Organization and URL
Audio-MD		Library of Congress <a href="https://www.loc.gov/standards/amd/vmd">https://www.loc.gov/standards/amd/vmd</a>
CIDOC		ICOM International Committee for Documentation <a href="http://network.icom.museum/cidoc">http://network.icom.museum/cidoc</a>
Darwin Core	Biodiversity information	Biodiversity Information Standards (TDWG) <a href="http://rs.tdwg.org/dwc/terms/simple/#simplifiedwcaxml">http://rs.tdwg.org/dwc/terms/simple/#simplifiedwcaxml</a>
Data Documentation Initiative (Codebook and Lifecycle versions)	General / Social Science	DDI Alliance <a href="https://www.ddialliance.org">https://www.ddialliance.org</a>
Data Tag Suite (DATS)	General / Health Sciences	BioCADDIE <a href="https://biocaddie.org">https://biocaddie.org</a>

<sup>6</sup> Dataverse Project, installations map: <https://dataverse.org>

<sup>7</sup> <https://dr-dn.cisti-icist.nrc-cnrc.gc.ca/eng/home/collection/Gateway%20to%20Research%20Data>

<sup>8</sup> <http://www.re3data.org>

DataCite	General, issuing DOIs	DataCite <a href="https://www.datacite.org">https://www.datacite.org</a>
EAD3	“International metadata transmission standard for hierarchical descriptions of archival records.”	Society of American Archivists <a href="https://www2.archivists.org/sites/all/files/TagLibrary-VersionEAD3.pdf">https://www2.archivists.org/sites/all/files/TagLibrary-VersionEAD3.pdf</a>
Ecological Markup Language (EML)	Ecology	The Knowledge Network for Biocomplexity <a href="https://knb.ecoinformatics.org/#external/emlparser/docs/index.html">https://knb.ecoinformatics.org/#external/emlparser/docs/index.html</a>
FGDC (CSDG)	Geographic	Federal Geographic Data Committee <a href="https://www.fgdc.gov/metadata/csdgm-standard">https://www.fgdc.gov/metadata/csdgm-standard</a>
ISAAR/ EAC-CPF	“General rules for the standardization of archival descriptions of records creators and the context of records creation.”	International Council on Archives <a href="http://www.ica.org/sites/default/files/CBPS_Guidelines_ISAAR_Second-edition_EN.pdf">http://www.ica.org/sites/default/files/CBPS_Guidelines_ISAAR_Second-edition_EN.pdf</a>
ISAD(G)	General guidance for the preparation of archival descriptions.	International Council on Archives <a href="http://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition">http://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition</a>
ISO 19115 (NAP)	Geographic	Natural Resources Canada <a href="http://nap.geogratis.gc.ca/metadata/napMetadata-eng.html">http://nap.geogratis.gc.ca/metadata/napMetadata-eng.html</a>
MARC 21	Bibliographic records description.	Library of Congress <a href="https://www.loc.gov/marc/marcdocz.html">https://www.loc.gov/marc/marcdocz.html</a>
MODS	Bibliographic element set	Library of Congress <a href="http://www.loc.gov/standards/mods">http://www.loc.gov/standards/mods</a>
NetCDF CF Metadata Conventions 1.6	Climate and Forecast / Geospatial	Climate and Forecast Conventions <a href="http://cfconventions.org">http://cfconventions.org</a>
Open Data Canada Profile (DCAT)	General	W3C <a href="https://www.w3.org/TR/vocab-dcat">https://www.w3.org/TR/vocab-dcat</a>
Protocol Data Element Definitions	Health Sciences. Used to describe interventional studies (clinical trials) and observational studies	U.S. National Institutes of Health <a href="https://clinicaltrials.gov">https://clinicaltrials.gov</a>
RAD	General. Description of archives.	Canadian Committee on Archival Description <a href="http://www.cdncouncilarchives.ca/archdesrules.html">http://www.cdncouncilarchives.ca/archdesrules.html</a>
Sensor Model Language (SensorML)	Multidisciplinary / Geoscience	Open Geospatial Consortium <a href="http://www.opengeospatial.org/standards/sensorml">http://www.opengeospatial.org/standards/sensorml</a>
Video-MD		Library of Congress <a href="https://www.loc.gov/standards/amdvmd">https://www.loc.gov/standards/amdvmd</a>
VRA-CORE	General. Used for the “description of works of visual culture as well as the images that document them.”	Library of Congress <a href="https://www.loc.gov/standards/vracore">https://www.loc.gov/standards/vracore</a>

Working from the definition of research data<sup>9</sup> provided by the Data Discovery Collections Development Working Group (DDCDWG)<sup>10</sup>, a related sub-group of the Portage DDEG, the DDMWG decided to focus on just those standards that describe digital research data. Some mapping work was completed for the additional metadata standards listed above, however this report focuses on research data standards. Working group members signed-up for standards and provided detailed field-level notes and analysis in a shared spreadsheet. As the standards covered a range of disciplines and data types, such as social science, health and life science, and geoscience, it was important that the group took the time to review and learn about each of the disciplinary standards before any metadata mapping or cross walking began. In many situations, members had deep knowledge and experience with a particular standard, making for very interesting discussion and cross-domain knowledge exchange.

In some cases, it was straightforward to gather full field-level metadata for the different standards, while in other cases it required additional investigation including retrieval of sample metadata from known repositories, standards organization's websites, and even conceptual documents referencing the metadata elements. In all, we ended up with thirteen complete sets of descriptive metadata standards to evaluate and then crosswalk. A simple Dublin Core set (model) was chosen for the FRDR data model<sup>11</sup> before the initial WG formed, so this was the common standard used for the mapping.

---

<sup>9</sup> Definition: *Data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results.* (based on the CASRAI definition [http://dictionary.casrai.org/Research\\_data](http://dictionary.casrai.org/Research_data))

<sup>10</sup> <https://portagenetwork.ca/wp-content/uploads/2017/04/DDEG-CollectionsWG-TOR-EN.pdf>

<sup>11</sup> Although this was chosen with the understanding that any domain metadata that could be harvested natively (or provided by depositors) would always be preserved and indexed in the FRDR backend with its original namespace, so that discovery would not be limited to Dublin Core -- rather, it is used as a minimum baseline.

## FRDR Discovery Profile

The FRDR baseline Discovery Profile is based on Simple Dublin Core and is defined using the Open Archives Initiative (OAI) Dublin Core (OAI\_DC) standard which is widely used for harvesting metadata from digital repositories. Some elements from the DataCite Schema are also used to define geographic information relating to the data, such as place or location. Together these elements make up the FRDR Metadata Profile.

The full profile contains 18 elements that are searchable in FRDR:

dc:title	dc:date	dc:relation
dc:creator	dc:type	dc:coverage
dc:subject	dc:format	dc:rights
dc:description	dc:identifier	DataCite_geolocationPlace
dc:publisher	dc:source	DataCite_geolocationPoint
dc:contributor	dc:language	DataCite_geolocationBox

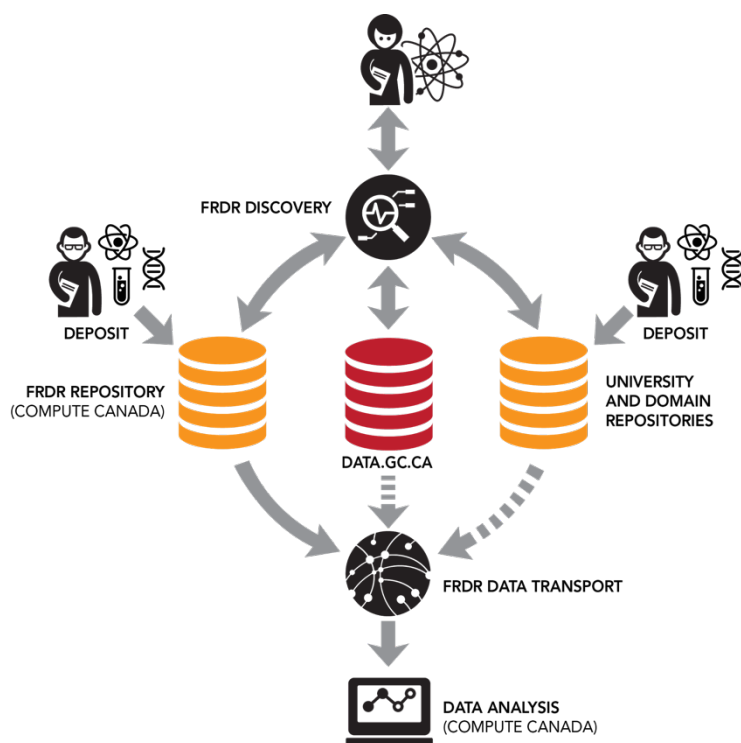
FRDR contains a metadata harvesting system (see Figure 2) that works in the backend to harvest metadata from research data repositories that use a variety of disciplinary metadata standards. This is primarily done using open metadata APIs such as CKAN and OAI-PMH<sup>12</sup>. Custom harvesters rely on the conceptual metadata crosswalk to the FRDR Discovery Profile and JSON schematic mapping<sup>13</sup>, in order to display consistent metadata for searching conducted in the FRDR discovery platform.

---

<sup>12</sup> <https://www.openarchives.org/pmh>

<sup>13</sup> OAI JSON schema mappings stored in GitHub [https://github.com/axfelix/globus\\_oai](https://github.com/axfelix/globus_oai)

Figure 2 - FRDR Harvesting<sup>14</sup>



In the initial discussions about the mapping to the FRDR Discovery Profile, the group had several concerns related to how each member would map disciplinary standards to a simple Dublin Core set of elements. Concerns about consistency in the intellectual mapping process across all standards was raised, since individual members of the group were performing different mappings. The second concern was the interpretation of the Dublin Core fields and in what way we would be using them in terms of the discovery interface in the FRDR. This included discussions about access links back to the original data provider / repository, contributor definitions and roles, and relationships and linkages between research data and other related resources. Figure 3 provides information about how certain Dublin Core elements were interpreted and used by the working group members.

<sup>14</sup> Figure reproduced with permission from Compute Canada.



**Figure 3 – Interpretation and Application of Dublin Core Elements for Mapping**

Dublin Core Element	Interpretation and Application
dc:contributor	A contributor is defined as a person or organization that has contributed, either directly to the intellectual content, or, in providing the dissemination and access to the resource (e.g. distributor role, etc. as defined in some standards). Repeatable.
dc:coverage	Coverage is used to define the temporal coverage or time period that this dataset may cover or occur within, and can include other types of coverages not defined here, but not spatial coverage which is managed using a set of DataCite GeoLocation elements, in conjunction with the FRDR Discovery Profile.
dc:source	<p>Source is defined as the original source metadata that has been presented for discovery and reuse in the FRDR repository. This is a mandatory field that can be derived and also auto generated as part of the harvesting / metadata exchange process. The conditions for this field are that it must be a URL that is a resolvable unique resource location for the system to point users to access the original source metadata.</p> <p>Additional note: This could be the same as the dc:identifier field in some cases, however dc:source must be web resolvable.</p>
dc:relation	Relations include references, citations, in either structured or unstructured form, to related materials, publications, documentation, studies, etc., that may be related to this item.

## Standards Evaluation and Mapping / Crosswalk

First, the group developed criteria to codify the elements to assist with the mapping process. A colour coding scheme was used during the evaluation of elements, and was entered directly into the metadata spreadsheet. This included the identification of elements thought to be closely related to Dublin Core, unique identifiers or related resources linkages that may provide linkages to sources and other related studies, as well as, elements considered to be discipline-specific.

Figure 4 - DATS Metadata Mapping to Dublin Core (example selection)

FRDR-MD (OAI_Dublin Core) <a href="http://www.openarchives.org/OAI/2.0/oai_dc.xsd">http://www.openarchives.org/OAI/2.0/oai_dc.xsd</a>	Refinement	DATS	
dc:title		entity:Dataset property:title	"title" : "Recurrent somatic mutations in POLR2A define a distinct subset of meningiomas [RNA-seq]"
dc:title	Alternative	entity:Dataset property:alternateIdentifiers	
dc:creator		entity:Dataset property:creators	
dc:subject		entity:Dataset property:keywords	"keywords" : "functional genomics"
dc:description	Abstract	entity:Dataset property:description	"description" : "RNA polymerase II mediates the transcription of all protein-coding genes in eukaryotic cells, a process that is fundamental to life..."
dc:publisher		entity:Person property:fullName or entity:Organization property:name	
dc:contributor	Distributor	entity: DataRepository property:name	
dc:contributor	Contact	entity:Person property:fullName	
dc:date		entity:Dataset property:dates (The type of date is specified in the dateType field, following the DataCite practice.)	"dateModified" : "09-12-2016"
dc:type		entity:Dataset property:type	"types" : "gene expression"

Close mapping to Dublin Core	Linkage potential	Discipline specific
------------------------------	-------------------	---------------------

Evaluating the thirteen data standards required discussion among group members to gain understanding about the inherent differences between the standards, such as the granularity and discipline-specific language used to define elements. A problem identified for the group was weighing the benefits of including parts of

various standards that did not map to Dublin Core, with some degree of flexibility. For example, if a repository uses the DDI metadata standard (which provides descriptive fields down to the dataset variable and value level), it would be a shame to lose this information in the discovery interface. At the same time, establishing a standard model for discovery to fit all disciplines and to assume that such metadata will always be included is challenging. Variations in the metadata received would make searches on fields incomplete and potentially misleading.

Next, after the initial mapping had been performed for each of the standards, the group combined these mappings into a super-mapping that provided a high-level crosswalk of all the disciplinary and general metadata standards, to the FRDR Discovery Profile (see Figure 5; see Appendix 1 for full-size version). This high-level crosswalk provides a comparative view of the conceptual mapping across disciplinary standards. It is helpful for determining what elements are core across all standards, and how we may fill in the gaps or look to developing a flexible approach for displaying metadata across disciplines in a single discovery interface.

Figure 5 - High-level Disciplinary Metadata Crosswalk (FRDR)<sup>15</sup>

FRDR-ND (OAI_Dublin Core) <a href="http://www.openarchives.org/OAI/2.0/oai_dc.xsd">http://www.openarchives.org/OAI/2.0/oai_dc.xsd</a>	Dataverse	DCATS	Open Data Canada	Darwin Core	EML	DATS	Protocol Data Element Definitions	SensorML	CF 1.4	DDI 3.2	DDI 2.5	FODC	ISO19115
dc:title	title	title	10.65 resource_name_en dec:datasetName	title	entity:Dataset property:title entity:Dataset property:alternateIdentifiers	Official Title: Brief Title	<gmt:name> title; long_name	<Title> <FullTitle> <AltTitle>	<Title> <AltTitle>	<Title> <AltTitle>	<Title> <AltTitle>	title	<gmt:title> <gmt:creatorResponsibilityParty> <gmt:creator> <gmt:investigator> or <gmt:author>
dc:creator	creator	10.10 creator	dec:creatablyBy; dec:identifiably	creator	entity:Dataset property:keywords	Overall Official	<gmt:contacts> <gmt:contact> <gmt:responsibilityParty> <gmt:classification> <gmt:classification> <gmt:keywords>	institution	<Creator> <ResearcherID> <Topic Coverage> <Keywords> <Subject> <Keywords>	<Authority>	<Authority>	originator	<gmt:creatorResponsibilityParty> <gmt:creator> <gmt:investigator> or <gmt:author>
dc:subject	subject	dc:subject dc:theme	10.82 topic_category 10.87 subject	dec:genus; dec:subgenus	keyword	entity:Dataset property:description	Study Purpose	comment; url; methods; source; history	<Keywords> <Subject> <Keywords>	<Keywords> <Subject> <Keywords>	<Keywords> <Subject> <Keywords>	subject	<gmt:topicCategory>
dc:description	description	dc:publisher	10.16 notes_en; 10.17 notes_fr 10.56 owner_org 10.59 org_title_at_publicatio n_3 10.69 org_title_at_publicatio n_3	dc:description	abstract	entity:Person property:fullName; entity:Organization property:name	N/A	<gmt:description>	<Abstract>	<Abstract>	<Abstract>	abstract	comment; purpose; process; contents reference <gmt:abstract>
dc:publisher	publisher		dec:institutionCode	publisher	entity:Person property:fullName; entity:DataRepository property:name	Collaborator	<Contributor>	<Publisher>	<Producer>			publisher	<gmt:creatorResponsibilityParty> <gmt:creator> "Publisher" <gmt:creatorResponsibilityParty> <gmt:creator> "Collaborator" or <gmtdistributor>
dc:contributor	contributor	dc:issued	10.8 contributor_en 10.8 contributor_fr 10.11 date_captured 10.59 resource_date_publish ed	dc:contributor	entity:Dataset property:dates	First Received; Last Updated; Last Changed Date	<Contributor> <Contributor> <Contributor>	<PublicationDate> <Date> <SingleDate> <MultipleDate>	<PublicationDate> <Date> <SingleDate> <MultipleDate>	<PublicationDate> <Date> <SingleDate> <MultipleDate>	<PublicationDate> <Date> <SingleDate> <MultipleDate>	date	abstract; process; contents reference <gmt:creatorResponsibilityParty> <gmt:creator> "Publisher" <gmt:creatorResponsibilityParty> <gmt:creator> "Collaborator" or <gmtdistributor>
dc:date	publicationyear	dc:mediaType	dec:modified; dec:modified	pubdate	entity:Dataset property:type	Available Study Data/Documents: Type [Item list - Individual Participant Data Set, Study Protocol, Statistical Analysis Plan, Informed Consent Form, Clinical Study Report, Analytic Code, Other (specify)]	<gmt:validTime>	<Date> <Date> <Date> <Date>	<Date> <Date> <Date> <Date>	<Date> <Date> <Date> <Date>	<Date> <Date> <Date> <Date>	date	<gmt:date>
dc:type	resourcetype	dc:format	10.70 resource_type 10.70 resource_format	dc:resourceType	dataset; citation; protocol; software	NCT ID	<gmt:characteristics name> <gmt:characteristics>	NatureType; char. type; short int. feat. real; double	<Type> <KindOfData>	<DataFormat>	<DataFormat>	resource; format	<gmt:spatialRepresentationType>
dc:format	size	dc:format	dec:collectorCode; dec:categoryHeader; dec:recordNumber; dec:openitemID	physical	entity:Dataset property:identifier; entity:Dataset property:alternateIdentifiers entity:Dataset property:relatedIdentifier entity:Access property:landingpage	NCT ID	<gmt:identifier>	<FileFormat> <UserDir> <InternationalIdentifier>	<FileFormat> <UserDir> <InternationalIdentifier>	<FileFormat> <UserDir> <InternationalIdentifier>	<FileFormat> <UserDir> <InternationalIdentifier>	format; formname	<gmt:fileType>; <gmt:resourceFormat>
dc:identifier	identifier	dc:identifier dc:downloadURL	10.71 resource_urlpath_id 10.71 resource_url 10.77 resource_url 10.78 10.78 object_identifier	dc:identifier	data:source		<gmt:identifier>	stand_name	<Identifier>	<Identifier>	<Identifier>	source	<gmt:identifier>
dc:source	language	dc:language	10.82 resource_language	dc:resourceLanguage	language	N/A	<gmt:language>	<Language>	<Language>	<Language>	<Language>	language	<gmt:language>
dc:relation	RelatedIdentifier	dc:landingPage dc:landingPage	10.27 program_page_url_en 10.28 program_page_url_fr 10.34 resource_related_keyword; 10.63 resource_related_type	dec:associatedReferences	citation	entity:Dataset property:dates	Study Start Date; Primary Completion Date	<gmt:documentation>	<Relationship>	<Relationship>	<Relationship>	crossref	<gmt:documentation>
dc:coverage	rights	dc:rights	10.88 time_period_coverage_start 10.89 time_period_coverage_end 10.35 license_id	dec:rightsHolder	entity:Dataset property:licenses	coverage; temporalCoverage	calendar (T); timeSeries	<TemporalCoverage> <SpatialCoverage> <TemporalCoverage> <SpatialCoverage>	<TemporalCoverage> <SpatialCoverage> <TemporalCoverage> <SpatialCoverage>	<TemporalCoverage> <SpatialCoverage> <TemporalCoverage> <SpatialCoverage>	<TemporalCoverage> <SpatialCoverage> <TemporalCoverage> <SpatialCoverage>	temporal	<gmt:extent>
dc:rights	rights	dc:rights	dec:rightsHolder	intellectualRights	entity:Dataset property:licenses	Available Study Data/Documents: Comments City, State/Province, Country	<gmt:legalConstraints>	<Copyright> <GeographicLocation> <SpatialCoverage>	<Copyright> <GeographicLocation> <SpatialCoverage>	<Copyright> <GeographicLocation> <SpatialCoverage>	<Copyright> <GeographicLocation> <SpatialCoverage>	access constraints; use constraints	<gmt:resourceConstraints> <gmt:keyword>; <gmt:MSC_KeywordTypeCode_TypeCode>
DataCite_geolocationPlace	geolocationPlace	dc:spatial	10.24 geographic_region	dec:locationID; dec:continent; dec:country; dec:stateProvince; dec:locality	geographicCoverage	N/A	<gmt:location>	region	<GeographicLocation> <SpatialCoverage>	<GeographicLocation> <SpatialCoverage>	<GeographicLocation> <SpatialCoverage>	place keyword	<gmt:resourceConstraints> <gmt:keyword>; <gmt:MSC_KeywordTypeCode_TypeCode>
DataCite_geolocationPoint	geolocationPoint	dc:verticalCoordinate	dec:verticalCoordinate	N/A	N/A	N/A	<gmt:point>	point; coordinates; url; URI (Y); (X)	<Point>	<Point>	<Point>	horizays	<gmt:EX_Extent>
DataCite_geolocationBox	geolocationBox	dc:horizontalCoordinate	dec:horizontalCoordinate	N/A	N/A	N/A	<gmt:boundary>	boundary; url; URI (Y); (X)	<GeographicBoundary> <BoundingBox>	<GeographicBoundary> <BoundingBox>	<GeographicBoundary> <BoundingBox>	bounding	<gmt:EX_GeographicBoundingBox>

<sup>15</sup> See Appendix 1 for full-sized version.

The Phase 1 recommendations from the DDCDWG are to include ten Canadian data repositories in FRDR for discovery. These repositories cover a range of academic disciplines. In some cases, further evaluation of the repositories standards and subjects will be required to assess whether the initial metadata mappings and crosswalk provided by the DDMWG will be sufficient for harvesting from these ten repositories.

**Figure 6 - Repository Shortlist with Subject and Standard**

Research Data Repository	Top-level Subject (using <a href="http://re3data.org">re3data.org</a> )	Standard(s)	Included in FRDR Beta Y/N
Canadian Opinion Research Archive ( <a href="http://www.queensu.ca/cora">http://www.queensu.ca/cora</a> )	Humanities and Social Sciences	OAI / DDI-Codebook	No
Ocean Networks Canada ( <a href="http://www.oceannetworks.ca">http://www.oceannetworks.ca</a> )	Life Sciences - Natural Sciences - Engineering Science	N/A - still investigating	No
Polar Data Catalogue ( <a href="https://www.polardata.ca">https://www.polardata.ca</a> )	Humanities and Social Sciences - Life Sciences	OAI / FGDC	Yes
Canadian Dataverses ( <a href="http://dataverse.scholarsportal.info">http://dataverse.scholarsportal.info</a> , <a href="https://dataverse.library.ualberta.ca">https://dataverse.library.ualberta.ca</a> , <a href="http://dvn.library.ubc.ca/dvn">http://dvn.library.ubc.ca/dvn</a> )	Humanities and Social Sciences - Life Sciences - Natural Sciences - Engineering Sciences	OAI /DC / DDI	Yes
Mouse Atlas of Gene Expression ( <a href="http://www.mouseatlas.org/mouseatlas_index.html">http://www.mouseatlas.org/mouseatlas_index.html</a> )	Life Sciences	N/A	No
World Ozone and Ultraviolet Radiation Data Centre ( <a href="http://woudc.org">http://woudc.org</a> )	Natural Sciences	OAI / ISO 19115	No
Ocean Tracking Network (OTN) ( <a href="http://oceantrackingnetwork.org">http://oceantrackingnetwork.org</a> )	Life Sciences - Natural Sciences	CSW	No
Hakai Institute ( <a href="https://www.hakai.org">https://www.hakai.org</a> )	Natural Sciences	N/A	No
BC Conservation Data Centre ( <a href="http://www2.gov.bc.ca/gov/content/environment/plants-animals-ecosystems/conservation-data-centre">http://www2.gov.bc.ca/gov/content/environment/plants-animals-ecosystems/conservation-data-centre</a> )	Life Sciences	OAI / CKAN API	Yes - through Open Data B.C.

Mapping new repositories will continue into the fall of 2017 as we may need to consider if a custom schema is used by the repository. If so, additional mapping to the FRDR Profile will be required.

## Recommendations for Improved Discovery in FRDR

In order to support the granularity and disciplinary specific elements required for rich discovery in the FRDR discovery interface, the DDMWG recommends additional support be developed in the FRDR system.

Recommendations for improved data discovery in FRDR:

- Support for subject faceting / browsing of search results with faceting
- Flexible and granular disciplinary metadata support (e.g. dataset variable-level)
- Enhanced linkages to related resources and source datasets

There may be several ways to achieve these recommendations for improved discovery in FRDR, some of which may also require additional investigation and research by the DDMWG and others, in close collaboration with the FRDR development team. So far, the DDMWG has evaluated approaches that may provide solutions for improved discovery in FRDR, including the use of a standard subject classification such as the re3data.org subjects accessible via an open metadata API for assigning subjects to all repositories' metadata (for those listed in re3data.org)<sup>16</sup>, and/or, the use of data services such as the OCLC FAST Service<sup>17</sup> for programmatic access to Library of Congress Subject Headings. Nevertheless, there are significant challenges associated with maintaining additional enhanced metadata for external metadata resources. The sustainability of enhanced metadata would not be likely, since each repository may use different standards for assigning subjects to metadata / data sources, thus requiring additional and resource intensive maintenance of subject mappings to FRDR core subjects (if adopted).

The group has discussed the need for FRDR to support 'Disciplinary Views' for different standards in order to be flexible and offer up metadata elements to the search interface as needed. For example, some repositories include rich metadata down to the dataset's variable / element level, which can be very helpful to researchers searching for particular data elements. Storing original source

---

<sup>16</sup> Re3data.org API <http://www.re3data.org/api/doc>

<sup>17</sup> OCLC FAST Search <http://fast.oclc.org/searchfast>

metadata in its entirety and developing flexible metadata viewers to be called upon from the FRDR discovery interface will be necessary to support this kind of discovery in the system.

Linking data to related publications and resources on the web is a growing area of concern within libraries, archives, and for researchers. It is important to present research data within context of other related research outputs available in repositories, and with citations and resource links on the web. Discovery of research data across a variety of systems will be a requirement for understanding the full research lifecycle and tracking research outputs for evaluation of tenure and other academic achievement. To support a model where research data are linked to related research outputs, including publication, structured linkages must exist between resources on the web. This includes structured approaches to including linkages to publications, related studies, and source data, within adopted data models and discovery frameworks. Currently, FRDR does not support a linked data model such as the Dublin Core Resource Description Framework (RDF) schema<sup>18</sup>, which describes the semantic relationships between elements defined in Dublin Core, and, to other related resources and ontologies found on the web.

The group also discussed the use of other well documented linked data models such as DCATS<sup>19</sup> and the Portland Common Data Model<sup>20</sup>. While it may take considerable resources to achieve a comprehensive linked data model for research data in FRDR, we recommend considering worthwhile partnerships with other organizations aiming to achieve similar goals. The ARL-OSF SHARE Project, is one such organization that has demonstrated interest in working closely with Portage, and beginning in 2017, will be addressing linked data with its current data model review.

---

<sup>18</sup> DC-RDF <http://dublincore.org/documents/dc-rdf>

<sup>19</sup> DCATS <https://www.w3.org/TR/vocab-dcat>

<sup>20</sup> PCDM <https://github.com/duraspace/pcdm/wiki>

## **Next Steps - Phase 2 2017-2018**

Over the next few months and into early 2018 the DDMWG will work closely with the FRDR development team to provide support with metadata mapping and crosswalking for the remaining repositories for inclusion. This will include identifying metadata standards in use where there is not identified standards currently. We will also work to achieve better understanding and pathways forward about best approaches for providing subject faceting and browsing in FRDR, granular and flexible discovery of disciplinary metadata in the system, and approaches to linked data, as they take shape within the broader metadata community.

# Appendix 1

## High-level Disciplinary Metadata Crosswalk (FRDR)

FRDR-MD OAL (Dublin Core) http://www.openarchives.org/OAI/2.0/OAI-Record	Dataverse	DCATS	Open Data Canada	Darwin Core	EML	DATS	Protocol Data Element Definitions	SemanticML	CF 1.6	DDI 3.2	DDI 2.5	FODC	ISO15915
dc:title	title	file	10.65 resource_name_en	dwc:datasetName	title	entity:Dataset property:file; entity:Dataset property:alternateIdentifiers	Official Title; Brief Title	<gml:name>	file;	<<Title>> <<SubTitle>> <<AlternateTitle>>	<<Title>> <<SubTitle>> <<AltTitle>>	file	<gml:title>
dc:creator	creator	dc:subject	10.10 creator	dwc:recordedBy; dwc:identifiedBy	creator	entity:Dataset property:creators	Overall Official	<gml:contact>; <gml:responsibleParty>	institution	<<Creator>> <<ResearcherID>> <<TopicalCoverage>> <<Keywords>>	<<Author>> <<PrincipalInvestigator>> <<Author>>	originator	<gml:responsibleParty>
dc:subject	dc:subject	dc:subject	10.07 subject	dwc:taxon; dwc:subgenus	keyword	entity:Dataset property:keywords	Conditions; Keywords	<gml:keywords>					
dc:description	description		10.16 notes_en; 10.17 notes_fr	dwc:description	abstract	entity:Dataset property:description	Study Purpose	<gml:comment>; <gml:method>; <gml:source>; <gml:history>					<gml:abstract>
dc:publisher	publisher		10.56 owner_org			entity:Person property:fullName; entity:Organization property:name	N/A						
dc:contributor	contributor	dc:publisher	10.59 contributor_en; 10.60 contributor_fr	dwc:contributor	publisher	entity:Person property:fullName; entity:DataRepository property:name	Collaborators						<gml:contributor>
dc:date	publicationYear	dc:date	10.69 resource_date_published	dwc:resourceDate; dwc:identifiedDate; dwc:modified	pubdate	entity:Dataset property:dates	First Released; Last Updated; Last Changed Date						<gml:date>
dc:type	resourceType	dc:format	10.76 resource_type	dwc:taxon; dwc:taxonType	dataset citation, protocol, software	entity:Dataset property:types	Available Study Data/Documents: Data Set, Study Protocol, Statistical Form, Clinical Study Report, Analytic Code, Other (specify)						
dc:format	size	dc:format	10.71 resource_size; 10.72 digital_object_identifier	dwc:collectionCode; dwc:catalogNumber; dwc:recordNumber; dwc:organismID	physical	entity:Dataset distribution property:formats	N/A	<gml:characteristics>					
dc:identifier	identifier	dc:identifier	10.73 resource_url	dwc:accessionNumber; dwc:accessionNumber; dwc:accessionNumber	packaged; identifier; URL in EML Document	entity:Dataset property:identifier; entity:Dataset property:relatedIdentifier	NCT ID	<gml:identifier>					<gml:fieldIdentifier>
dc:source	language	dc:language	10.82 resource_language	dwc:language	dataSource	N/A	N/A						<gml:source>
dc:language	language	dc:language	10.27 program_page_url_en; 10.27 program_page_url_fr; 10.64 resource_abstract_text; 10.63 resource_record_type	dwc:language	language	entity:Dataset property:primaryPublications; entity:Dataset property:relatedIdentifier	Publication Citation						<gml:language>
dc:relatedIdentifier	RelatedIdentifier	dc:temporal	10.88 related_coverage_start; 10.89 related_coverage_end	dwc:associatedReferences	citation	entity:Dataset property:dates	Study Start Date; Primary Completion Date						
dc:coverage	rights	dc:rights	10.35 license_id	dwc:rightsHolder	coverage; temporalCoverage	entity:Dataset property:licenses	Available Study Data/Documents; Comments						
dc:rights	rights	dc:rights	10.24 geographic_region	dwc:taxonID; dwc:stateProvince; dwc:locality	intellectualRights	entity:Dataset property:licences	City, State/Province, Country						<gml:resourceConstraints>
DataCite_geolocationPage	geolocationPage			dwc:geographicRegion	geographicCoverage	entity:DataAcquisition (subclass of Activity) property:locations	N/A						<gml:keyword>
DataCite_geolocationPoint	geolocationPoint			dwc:verbatimCoordinates		N/A	N/A						<gml:place>
DataCite_geolocationBox	geolocationBox			dwc:verbatimCoordinates		N/A	N/A						<gml:EX_Extent>