RDC

Research Data Canada

Guidelines for the deposit and preservation of research data in Canada

DOCUMENT CHANGE CONTROL

Revision number	Date issued	Authors*	Brief description of change
	2015-05-04	Claire Austin Susan Brown Chuck Humphrey Amber Leahey Peter Webster	Released by RDC-SINC and sent to RDC Steering Committee for approval.
1.0	2015-05-19		First official RDC version.
1.1	2015-06-17		Corrected & updated certification (p10, References)

** All authors contributed equally to the writing of this document. The authors declare no conflict of interest*

RDC-SINC Committee members and their affiliations

Claire Austin (Co-Chair)	Environment Canada
David Baker	CASRAI Consortia Advancing Standards in Research Administration Information
Mairi Best	EMSO European Multidisciplinary Seafloor and Water-Column Observatory
Alice Born	Statistics Canada
Susan Brown	University of Guelph / University of Alberta
Xinjie Cui	Alberta Center for Child Family and Community Research
James Doyle	Environment Canada
Isabel Fortier	McGill University
Julie Friddell	University of Waterloo
Chuck Humphrey	University of Alberta
Amber Leahey	Scholars Portal
Karen Morgenroth	National Research Council
Mathieu Ouellet	Fisheries and Oceans Canada
Jian Pei	Simon Fraser University
Peter Webster (Co-Chair)	St. Mary's University

EXTERNAL REVIEWERS

Jill Kowalchuk Shirley Mills	CANARIE Carleton University / Statistical Society of Canada

Feedback link:

https://www.surveymonkey.com/s/RDC-SINC_GuidelinesDataDeposit_SurveyAndApproval

Table of Contents

ABSTRACT Research Data Canada's Standards and Interoperability Committee (RDC-SINC) has developed guidelines to address a need for the implementation of standards with respect to research data and associated metadata, and data repositories. The underlying principles are "*independent understandability,*" long term preservation, and accessibility of research data. In addition to the guidelines, references and a non-exhaustive list of relevant standards are also provided.

INTRODUCTION

We offer the present guidelines as a starting point for the development of a common understanding between the various groups engaged with data in Canada. Especially for managers, researchers, and those new to the field of data management, we hope that this will contribute to moving us forward in concert toward collectively and collaboratively getting up to speed, filling the gaps, identifying roles and responsibilities, and meeting the challenges that face us.

Research data are a national resource and need to be treated as such; therefore, research data and associated metadata must be deposited in curated, open data repositories. Research data is understood to include 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. Research data may be experimental data, observational data, operational data, third party data, public sector data, monitoring data, processed data, or repurposed data (RDC Glossary of terms & definitions 2014).

The Standards and Interoperability Committee of Research Data Canada has surveyed practices related to the acquisition, handling, sharing, and preservation of research data by national and international research data repositories and archives (Austin et al. 2015; RDC-SINC 2015). The present guidelines represent current practice and "*best practice*" implemented by organizations that accept data for deposit in Canadian and international public data repositories. These guidelines extend to the preparation, description, appraisal, dissemination, and preservation of research data.

RDC uses the following definition of *Repository*, related to the deposit and preservation of research data: Repositories preserve, manage, and provide access to many types of digital materials in a variety of formats. Materials in online repositories are curated to enable search, discovery, and reuse. There must be sufficient control for the digital material to be authentic, reliable, accessible and usable on a continuing basis. Please see the Glossary for definitions of *Data centre*, *Data repository, Trusted Digital Repository (TDR)*, and other related terms (RDC 2014).

Facilities supporting data deposit have a responsibility to ensure that the data are discoverable, accessible, documented, useable by others than the data originators, and preserved for the future. However, the burden and responsibility for creating datasets that meet these criteria lie with the

creators and generators of the data provided to the data repositories. Creators of data include individuals, research teams, institutions, government agencies, and a variety of other organizations.

Canada is in a period of rapid change as we transition from largely private, restricted, or specialized access datasets existing in a plethora of proprietary formats with varying degrees of documentation, quality control and interoperability to a world of data management planning (DMP), sharing and interoperability, big data, open data, open access, and open research. A number of Canadian initiatives have broken ground in this area (ARC; CPDN 2012; PDC 2011; Pinto 2014) and some very interesting projects are taking place in Europe (EUdat 2015). Canadian research organizations, including universities, government agencies, and discipline-specific research groups must be encouraged to develop fully functional data repositories as we move ever closer to data publication (Kratz & Strasser 2014). To be successful, this transition requires that the data providers and the facilities supporting data deposit work together in developing and implementing common standards that will guarantee the integrity of the data from the point of data planning and data collection through to its deposit and publication.

Criteria described in these guidelines for data and accompanying metadata explicitly refer to standards and commonly accepted "*best practices*" for data discovery and interoperability between heterogeneous datasets and repositories. We recognize that there may be cases where certain data and metadata cannot be made public for various policy or legal reasons. However, the default position should be that all research data and metadata are public. Metadata enabling data discovery should always be openly accessible. Where a case has been made to withhold research data from the public, the metadata and data guidelines described in the present document still apply.

Organizations and people who collect data must have a clear understanding of the metadata that need to be associated with their data before beginning to collect the data because it is highly inefficient, time consuming, error prone, and often impossible, to add the required metadata to a dataset after the fact. Preparation and implementation of a detailed data management plan at the beginning of a study is invaluable in ensuring that data collected will actually be usable by others and accepted for deposit in a repository (UKdataservice). In fact, agencies that fund research are beginning to require that data management plans be submitted with grant applications (NSF 2011; ESRC 2015; DOE 2014; RCUK 2014).

There is a broad recognition that better data authorship and data publishing standards are needed (NRC 2012). Recognizing that there are traditional cultural and technical barriers that discourage or prevent data sharing, the National Science Foundation (NSF) recently sponsored a workshop on supporting scientific discovery through norms and practices for software and data citation and attribution. "*The research community urgently needs new practices and incentives to ensure data producers, software and tool developers, and data curators are credited for their contributions*" (NSF 2015).

THE GUIDELINES

The Guidelines cover three broad areas of concern that must be addressed when considering the deposit of research data, and a number of essential considerations within each area.

METADATA

Without accompanying metadata, data are next to useless - or worse, if their use leads to incorrect or misleading conclusions. Metadata are structured information describing the data object or resource. The data must be described in sufficient detail that people other than the data originators can understand and make complete and valid use of the data without any need to contact the original data producers for additional information or additional associated data. This is known as the *independently understandable* principle in the Open Archival Information System reference model (OAIS Reference Model). Metadata are vital for digital preservation, ensuring that digital data and formats remain accessible and usable over the long-term. Documentation and metadata must be in compliance with commonly accepted open standards (see Appendix 1).

Metadata format

Dataset documentation must be provided in a standard, machine readable, openly accessible format to enable the effective exchange of information between users and systems. Acceptable formats include XML, RDF, and JSON.

Data citation

A formal data citation, or the necessary elements to produce a citation for reference identification and attribution, must be provided in the metadata using an appropriate standard. The preferred method is the use of persistent identifiers such as a digital object identifier, or DOI (Leadbetter et al. 2013).

Persistent identification (PID)

Data deposit facilities must register datasets using persistent identification (PID) technologies to ensure that data accessed through the Internet are independent of changeable repository file structures and storage locations. PIDs facilitate the discovery and identification of data (authority control) and ensure that information about the access location of the data persists. This is usually provided through a resolvable URL landing page from a dataset's associated metadata. PIDs are central to the proper citation and sharing of data.

A dataset must have at least one persistent identifier assigned at the point of deposit. PIDs should be assigned at the level of granularity that the data are likely to be cited or at the same level as the metadata. Decisions concerning the level of granularity at which a PID is assigned should be made by the data provider. PIDs should be assigned to 'original' research datasets only, not to duplicate copies of datasets. Best practice for registering data is the assignment of a digital object identifier (DOI). In Canada, DOI's are assigned through a national service (CISTI 2015a,b).

Version control (revision control; source control)

Revision control tracks and controls changes to data files. Data must be timestamped and versioned; timestamps alone do not constitute version control. Datasets frequently change over time as new data are appended and older data are modified or subset. The associated metadata must provide citable information concerning the version of the dataset (e.g., a version number, version statement, or a Universal Numeric Fingerprint (UNF)), dates of creation and modification. When previously deposited data are modified, the metadata must also include information concerning the reasons for the changes and the party responsible for the changes. In order to maintain data traceability, previous versions and the differences between versions of the data must remain accessible and identifiable. Dynamic datasets pose a special challenge for maintaining a record of versions (Katz & Strasser 2014). A Research Data Alliance (RDA) working group has recently proposed a solution for citing dynamic data (RDA 2015a,b).

Ownership and data reuse

Metadata must include a description concerning ownership, licensing, and intellectual property rights of the data. Data providers may be required by a repository to agree to a set of data deposit terms upon submitting datasets. The terms of reuse must be clearly stated, in line with the relevant legal and ethical requirements where applicable (e.g., subject consent, permissions, restrictions, etc.), and the metadata must provide repository access information. Use of one of the Creative Commons licenses is a good practice (CC 2015).

The Canadian Tri-Agency open access policy requires that any peer-reviewed publications arising from agency funded research must be made freely available in an online repository or open access journal (GoC 2015). In addition, CIHR-funded bioinformatics, atomic, and molecular coordinate data must be deposited with the appropriate public database.

Privacy and confidentiality

Data providers and the data that they deposit must adhere to specific standards regarding the privacy of subjects studied, and other ethical obligations. Where required, "*need to know*" and "*highest anonymity*" principles must be applied. In such cases, the metadata must describe the privacy requirements and how they were met (e.g., encryption, anonymization, etc.). The metadata must also include information concerning other ethics issues (e.g., subject consent, ethics reviews, etc.) where applicable. Where there are restrictions on the use of the data, the reasons for these restrictions must be explained in the metadata, and also information must describe how to gain controlled access to the data if possible (e.g., contact information, access terms, access agreement templates, data request forms, etc.).

Authority

Dataset descriptions must include statements concerning the creators, depositors, curators, affiliations with a scholarly organizations, government agencies, funding bodies, etc. The metadata must identify the Principal Investigator (P.I.), where applicable. P.I.'s must have a permanent human identifier (e.g., ORCID, which resolves critical issues of identifying individuals. Contact information must also be provided. The P.I. has primary responsibility for the intellectual direction and integrity of the research or research-related activity, including data production, findings and results, and ensures ethical conduct in

all aspects of the research process including but not limited to the treatment of human and animal subjects, conflicts of interest, data acquisition, sharing and ownership, publication practices, responsible authorship, and collaborative research and reporting (RDC 2014).

Disciplinary coverage

Standard subject classifications or domain headings greatly facilitate predictable data discovery. Academic subject classification systems exist to describe standard subject or domain terms and several lists are available for use. Domain- and disciplinary-specific standards should be used when available and current with modern data management practices as described in the present Guidelines (e.g., DCC 2014).

Controlled vocabularies

Standardized terminologies are used to index, discover, and retrieve information from metadata content. Related to subject classifications, controlled vocabularies provide a standard set of terms to define the meaning of metadata elements (i.e., semantics) and acceptable content to be assigned to metadata elements. As such, controlled vocabularies play a critical role in metadata standards. The adoption of a controlled vocabulary should be followed in producing metadata (e.g., DDI Controlled Vocabularies).

Data type

The metadata must distinguish between types of research data such as primary (i.e., original), derived, dynamic, raw or aggregated data etc., as much as possible. In addition, known relationships or linkages to other data resources (e.g., parent data, derived data, data subsets, collection cycles, etc.) must be described.

Levels of description

Study-level and file-level documentation are required to understand the context of the data and to ensure intelligent reuse (Data-PASS 2007).

Study-level metadata

Study-level metadata describe the program or project from which the data come. These include key contextual elements about a study, including a study's purpose, the data collection process, and information describing aspects of the research design, such as, the objects being observed, the time period, geographic location, sampling methods, selection criteria, etc. Also commonly covered is a description of data processing practices, including coding, weighting, scaling or other normalization of the data, and information about the data quality. References to any known relationship to other datasets, including versions, derivations, cycles, related studies, publications, or established standards or code lists (e.g., classification of occupations, industries, etc.) that inform the use of the dataset, must also be provided. Some consideration about the granularity of the metadata is necessary since few standards cover both the study-level and file-level description for data. Dublin Core (DC Standard) provides a basic set of elements and is commonly used for study-level descriptions. Digital collections of all kinds, including research data, have been described using the DC metadata element set. For example, DataCite (DataCite 2014) has adopted DC to provide a high-level description (discovery-level) for datasets.

File-level or variable-level metadata

This level of metadata describes the structure and format of the data, enabling application tools to input the data appropriately. File-level metadata must identify the file name, file location (if access is being provided online), format type, file version statement, date of deposit, access conditions, the record layout, etc. Variable-level metadata describe the attributes of the data, including variable names, missing value assignments, coding schemas, types of measurements, derived variables, value labels, etc. A description of the logical groups of observations and variables, concepts, questions and categories must also be included.

A note concerning readme files

Metadata handling standards and best practice, such as described in the present guidelines, must be implemented when collecting data. *Readme* files definitely do not fall in the category of "*best practice*." Information that can be handled via the implementation of standards and best practice that do not involve *readme* files must not be included in *readme* files. Where metadata standards and best practice have been properly implemented, *readme* files may still have some legitimate utility, but they should be avoided or eliminated as much as possible. It must be recognized that researchers and data providers need appropriate tools, infrastructure, and training to implement metadata standards and best practice. In the absence of these tools and training, although far from ideal, we recognise that the "*practice*" of using *readme* files is a huge improvement over the other common practice of no such information being provided at all.

DATA FILES

Data repositories need to declare which file formats they will and will not accept (InterPARES 2006; ICPSR 2015). Data files must be in compliance with applicable standards (see Appendix 1). The format in which data are stored is directly related to reuse and sharing of data, now and in the future. While data may have been collected or processed using proprietary formats, the data provider must submit data in the preferred format of the intended repository. Ideally, a discussion about the formats of files to be deposited will be held between the data provider and a curator from the intended repository at the beginning of a project. Such discussions are valuable because some repositories will convert proprietary formats to a preferred preservation format as part of its archival processing. It must be recognized that there is an inherent risk of introducing errors during conversion from one format to another, which is compounded if performed by a third party lacking the data creator's familiarity with the data. Such risks are mitigated when the data creator stores and/or converts the data to a non-proprietary format at source. As a principle, data in a repository must be immune to software, software version, and technology changes. Data repositories must, therefore, store non-proprietary, interoperable, lossless file formats. If the data provider uses the services of a repository or some other third party to convert data files to another format or data structure, the responsibility for ensuring the accuracy of the result, i.e., the content of the data files stored by the repository, must lie with the data provider who remains accountable for the data.

Acceptable data file formats include ASCII text or character-delimited text files, XML, or JSON with a tabular data structure (column names in the first row and data in subsequent rows). Whether files should be deposited in unencrypted and uncompressed at the dataset level is a discussion to be had with the intended repository. Acceptable multimedia file formats include .wav, .tif, and .jpg. Unacceptable file formats include .xls, .xlsx, .doc, .ppt, .wpd, .pdf.

A detailed discussion of data structures, file formats, data standards, and information that must be included in data files, and how it must be represented, is beyond the scope of the present document.

REPOSITORIES

Repositories must implement preservation processing standards, policies, best practices, and workflows to ensure the integrity, authenticity, and usability of the data and its associated metadata (DCC 2009; CRL 2007; NESTOR 2006). These requirements involve systems that document and audit the handling of files over time and that implement safeguards to prevent file degradation and loss. Preservation copies of datasets must be maintained in lossless, open, non-proprietary file formats than can be migrated to new formats as standards and technology change over time. To ensure appropriate preservation processes are enacted, the metadata accompanying a dataset must include the appropriate technical information about file structure and format.

Data providers should inform themselves about a repository's operations to understand what preservation formats are used, what preservation processing is implemented, what ongoing evaluation and assessment of deposited data takes place, and what succession plans exist in the event a repository is no longer sustainable (IHSN/ICPSR 2009).

Discovery

A data repositories must publicly state its collection development policy (domain subjects and scope of materials) and provide a description of the designated communities served. Data repositories must ensure that metadata are openly available and searchable online (e.g., discoverable through major online search engines). Metadata must be listed with a central registry or directory to increase the likelihood of discovery. Regional and subject domain repository directories also improve cross-disciplinary visibility (CISTI 2015a,b,c; DCC 2014; R3Data 2015).

Security

Data repositories must have policies and processes in place that address the integrity and security of research data. They may participate in widely distributed backup and preservation systems including LOCKSS, CLOCKSS, or other systems to ensure the long term integrity of datasets. Controls must be in place to prevent unauthorized alteration, misuse, or deletion of the data and associated metadata.

Maintenance

Data repositories must be able to demonstrate a sustainability plan for ongoing operations, including maintenance as collections, software and other requirements change over time. Preferably, the

institution hosting a repository has a policy declaring its continued commitment to the repository and to its data stewardship responsibilities.

Curation

Curation is the activity of managing and promoting the use of data from the point of creation to ensure that the data are fit for contemporary purpose and available for discovery and reuse (RDC 2014). For dynamic datasets this may mean continuous enrichment or updating to maintain fitness for purpose. Higher levels of curation also involve links with annotation and with other published materials. Curation includes archiving to ensure that data are properly selected, stored, and can be accessed, and that the data's logical and physical integrity is maintained over time, including security and authenticity. Repositories must provide clear statements describing the curation processes used.

Succession plan

A repository must have a publicly accessible succession plan that clearly states how the repository's data and associated metadata will be dealt with in the event that the repository or organization that maintains the repository ceases to operate.

Certification

Data repositories must be in compliance with applicable standards (see Appendix 1) and should obtain certification to validate their practices, including curatorial, organizational and administrative processes, technical functions, and quality control (CCSDS 2014; CRL/OCLC 2007; CRL 2014; CRL 2015; Donnelly et al. 2009; DSA 2013). Certification is possible through self-evaluation (DSA 2013), peer-review (APA/WP33), or external evaluation (CLR 2014; WDS 2012). Examples of certified trusted repositories include: Chronopolis, CLOCKSS, Hathitrust, Portico, Scholars Portal (CRL 2014), and World Data System members (WDS 2015). The Digital Curation Centre has developed a Digital Repository Audit Method Based on Risk Assessment – DRAMBORA (DCC 2009). The Research Data Alliance Repository Audit and Certification DSA–WDS Partnership Working Group aims to develop a common standard, then a framework for certification and a service of trusted data repositories (RDA 2015c). See, also, the RDA/WDS Certification of Digital Repositories Interest Group (RDA 2015d).

Repository platforms

A variety of software is emerging to support underlying repository operations, both commercial and open source. Examples of open source repository software include DataVerse, DSpace, and CKAN, while examples of commercial platforms include Ex Libra's Rosetta, CONTENTdm, and Preservica's Cloud. A data provider should be informed about the platform being used by the intended repository and should inquire about the platform's state of development and available options. The choice of platform by a repository may be an indicator about the sustainability of its service.

FEEDBACK

Readers are invited to provide feedback to the authors at the following link:
https://www.surveymonkey.com/s/RDC-SINC_GuidelinesDataDeposit_SurveyAndApproval

REFERENCES

All hyperlinks accessed on 2015-04-25

Austin CC; Brown S; Fong N; Humphrey C; Leahey L; Webster P (2015). *Research Data Repositories: Review of current features, gap analysis, and recommendations for minimum requirements*. Presented at the IASSIST Annual Conference, Minneapolis MN, June 2-5, and expected to be published in the IASSIST Quarterly (IQ, Summer 2015). International Association for Social Science, Information Services, and Technology.

CC (2015). *Creative Commons licenses*. http://creativecommons.org

CCSDS (2014). *Requirements for bodies providing audit and certification of candidate trustworthy digital repositories (TDR)*. The Consultative Committee for Space Data Systems. http://public.ccsds.org/publications/archive/652x1m2.pdf

CISTI (2015a). *DataCite Canada Datasets and DOIs.* Canada Institute for Scientific and Technical Information. http://cisti-icist.nrc-cnrc.gc.ca/obj/cisti-icist/doc/datacite/datasets.pdf

CISTI (2015b). *DataCite Canada Services*. Canada Institute for Scientific and Technical Information. https://www.cisti-icist.nrc-cnrc.gc.ca/eng/services/cisti/datacite-canada/index.html

CISTI (2015c). *Gateway to Research Data*. Canada Institute for Scientific and Technical Information. http://dr-dn.cisti-icist.nrc-cnrc.gc.ca/eng/home/collection/Gateway%20to%20Research%20Data/

CLOCKSS (2015). *Controlled LOCKSS*. http://www.clockss.org/clockss/Home

CPDN (October 2012). *Guidelines for the implementation the North American Profile (NAP) of ISO 19115:2003 – Geographic Information – Metadata*. Canadian Polar Data Network (CPDN). http://polardatanetwork.ca/wp-content/uploads/CPDNBestPracticeForNapMetadata1.pdf

CRL (2007). *Core Requirements for Digital Archives – Ten Principles* http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re

CRL (2014). *Certification & Assessment of Digital Repositories*. Center for Research Libraries. http://www.crl.edu/archiving-preservation/digital-archives/certification-assessment

CRL (2015). Other digital repository assessment tools. Center for Research Libraries. http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/other-assessment-tools

CRL/OCLC (February 2007). *TRAC – Trustworthy Repositories Audit & Certification: Criteria and Checklist v 1.0.* Center for Research Libraries & Online Computer Library Center. http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf

Data-PASS (2007). *Data-PASS Metadata Requirements Version 1.2*. Data Preservation Alliance for the Social Sciences, http://www.data-pass.org/sites/default/files/metadata.pdf

DCC (2009). *Digital Repository Standards for Any Lifecycle Action*. Digital Curation Centre, Digital Curation Standards, DIFFUSE – Dissemination of InFormal and Formal Useful Specifications and Experiences. http://www.dcc.ac.uk/resources/standards/diffuse/standards?framework_id=1&lifecycle_id=0

DCC (2014). *Disciplinary Metadata – Metadata Standards, Profiles and extensions, Use cases, Tools*. Digital Curation Centre. http://www.dcc.ac.uk/resources/metadata-standards

Donnelly M; Innocenti P; McHugh A; Ruusalepp R(2009). *Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) Interactive User Guide*. Digital Curation Centre (DCC) & Digital Preservation Europe (DPE). http://www.dcc.ac.uk/sites/default/files/DRAMBORA_Interactive_Manual%5B1%5D.pdf

DSA (July 2013). *Repositories – Data Seal of Approval Guidelines, v2*. Data Seal of Approval. http://datasealofapproval.org/media/filer_public/2013/09/27/guidelines_2014-2015.pdf

EUdata (2015). Research Data Services, Expertise & Technology Solutions. http://www.eudat.eu

GOC (February 27, 2015). *Tri-Agency Open Access Policy on Publications*. Government of Canada http://www.science.gc.ca/default.asp?lang=En&n=F6765465-1

ICPSR (2015). *Details on Appraisal Criteria*. Inter-University Consortium for Political and Social Research. http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/details.html

IHSN/ICPSR (December 2009). *Principles and Good Practice for Preserving Data.* International Household Survey Network & Interuniversity Consortium for Political and Social Research. http://ihsn.org/home/sites/default/files/resources/IHSN-WP003.pdf

InterPARES (December 2006). *Selecting digital file formats for long-term preservation*. International Research on Permanent Authentic Records in Electronic Systems. http://www.interpares.org/display_file.cfm?doc=ip2_gs11_final_report_english.pdf

Kratz J; Strasser C (2014). *Data publication consensus and controversies, v3*. http://f1000research.com/articles/3-94/v3

Leadbetter A; Raymond L, Chandler C; Pikula L; Pissierssens P; Urban E (2013) Ocean Data Publication Cookbook. Paris: UNESCO, 41 pp. & annexes. (Manuals and Guides. Intergovernmental Oceanographic Commission, 64), (IOC/MG/64). http://www.iode.org/components/com_oe/oe.php?task=download&id=20171&version=1.0&lang=1&format=1

LOCKSS (2015). *Lots of Copies Keep Stuff Safe*. http://www.lockss.org

NESTOR (December 2006). *Catalogue of Criteria for Trusted Digital Repositories*. Network of Expertise for Long-Term Storage and Long-Term Availability of Digital Resources in Germany, Working Group on Trusted Repositories Certification Nestor Working Group, 2008. http://files.d-nb.de/nestor/materialien/nestor_mat_08-eng.pdf

NRC (2012). Developing data attribution and citation practices and standards: Summary of an international workshop. National Academy of Sciences. http://www.nap.edu/catalog/13564/for-attribution----developing-data-attribution-and-citation-practices-and-standards

NSF (April 2015). *Workshop on supporting scientific discovery through norms and practices for software and data citation and attribution. National Science Foundation.* https://softwaredatacitation.org/Workshop%20Report/SoftwareDataCitation_workshop_report_2015_April_20_without_logo.pdf

ORCID (2015). *Open Research and Contributor ID*. ORCID Inc. http://orcid.org/content/about-orcid

PDC (October 2011). *Best Practices for Sharing and Archiving Datasets*. Polar Data Catalogue. https://polardata.ca/pdcinput/public/PDC_Best_Practices_FULL.pdf

Pinto J (2014). *Research Data Canada Federated Preservation Pilot*. https://docs.google.com/document/d/1WopZ3CrOBak0oSuMy_kQLygwC6KT26PDSNn8tiwfS0I/edit

RDA (2015a). *Scalable Dynamic Data Citation*. Research Data Alliance Data Citation Working Group (WG). https://rd-alliance.org/sites/default/files/attachment/Scalable%20Dynamic%20Data%20Citation_v1.pdf

RDA (2015b). *Scalable Dynamic Data Citation - RDA-WG-DC Position Paper*. Research Data Alliance Data Citation Working Group (WG). https://www.rd-alliance.org/groups/data-citation-wg/wiki/scalable-dynamic-data-citation-rda-wg-dc-position-paper.html

RDA (2015c). Repository audit and certification DSA–WDS partnership working group. Research Data Alliance, Data Seal of Approval/World Data System. https://www.rd-alliance.org/groups/repository-audit-and-certification-dsa%E2%80%93wds-partnership-wg.html

RDA (2015d). RDA/WDS Certification of Digital Repositories Interest Group. Research Data Alliance/World Data System. https://rd-alliance.org/groups/rdawds-certification-digital-repositories-ig.html

RDC (2014). *Glossary of terms and definitions*. Research Data Canada. http://www.rdc-drc.ca/glossary

RDC-SINC (February, 2015). *Research Data Repository Requirements and Features Review*. Research Data Canada, Standards and Interoperability Committee http://hdl.handle.net/10864/10892

Re3Data (2015). *Global registry of research data repositories*. http://re3data.org

WDS (2012). Certification of WDS members. International Council for Science (ICSU), World Data System. https://www.icsu-wds.org/files/wds-certification-summary-11-june-2012.pdf

WDS (2015). World Data System members. https://www.icsu-wds.org/community/membership/regular-members

APPENDIX 1 – STANDARDS FOR METADATA, DATA, AND DATA REPOSITORIES

Author identification

Open Research and Contributor ID (ORCID)

Country identification

ISO 3166 (Parts 1-3) - Codes for the representation of names of countries and their subdivisions

Data management and curation

DAMA - Data management body of knowledge

Date and time format

ISO 8601 - Data elements and interchange formats, Information interchange, Representation of dates and times

IANA (Internet Assigned Numbers Authority) time zone database

Digital object identifiers (DOI)

ISO 26324 - Information and documentation -- Digital object identifier system (DOI)

ISO/TR 23081 (Parts 1-3) - Information and documentation -- Records management processes -- Metadata for records

Digital Preservation

ISO 14721 - Open Archival Information System (OAIS) Reference Model

Metadata Encoding and Transmission Standard (METS)

Preservation Metadata Maintenance Activity (PREMIS)

Lots of copies keep stuff safe (LOCKSS)/ Controlled LOCKSS (CLOCKSS)

Environmental

ISO 6107 - Water quality -- Vocabulary

World Meteorological Organization (WMO) – Guide to climatological practices

Guide to Agricultural Meteorological Practices (GAMP)

NASA Global change master directory

Ocean data Standards

File Formats

JPEG 2000

OpenDocument (ODF)

ISO 19005-1: Document Management - Electronic Document File Format for Long-term Preservation (PDF/A/A-1)

Tagged Image File Format (TIFF)

Network Common Data Form (netCDF)

Flexible Image Transport System (FITS)

American Standard Code for Information Interchange (ASCII)

Geospatial
ISO 19115 - Geographic information, Metadata
ISO 19119 - Geographic information -- Services
ISO 19128 - Geographic information, Web map server interface
ISO/TS 19139 - Geographic information -- Metadata -- XML schema implementation
Open Geospatial Consortium Standards
Geospatial metadata Standards

Glossaries
Research Data Canada - Terms & definitions
Glossary of meteorology (American Meteorological Society)
TeD-t Term Definition Tool (Research Data Alliance)

Health
ICD - International classification of disease
Canadian coding standards for ICD-10-CA and CCI

Information management
CAN/CGSB-72.34 - Electronic records as documentary evidence
ISO 15489 - Information and documentation -- Records management

Information security
ISO 27000 - Information security standards

Interoperability
Protocol for Metadata Harvesting - Open Archives Initiative (OAI-PMH)
Object Reuse and Exchange - Open Archives Initiative (OAI-ORE)

Language codes
ISO 639-1 - Codes for the representation of names of languages (Parts 1-5)

Languages, computer
Extensible markup language (XML)
Unified modeling language (UML)

Measurement Units
The unified code for units of measure

Metadata
ISO 15836/ANSI Z39.85 (NISOZ3985) - Dublin Core Metadata Element Set

ISO/IEC 11179 - Information technology -- Metadata registries (MDR)
DDI - Data Documentation Initiative
EML - Ecological Metadata Language
Metadata Object Description Schema (MODS)
See Disciplinary Metadata – Metadata Standards, Profiles and extensions, Use cases, Tools. (DCC Digital Curation Centre)

Repositories
Digital Repository Standards for Any Lifecycle Action. (DCC Digital Curation Centre)
ISO 16363 - Space data and information transfer systems -- Audit and certification of trustworthy digital repositories (TDR)
Trustworthy Repositories Audit & Certification (TRAC)
International Council for Science (ICSU) – World Data System (WDS)
Data Seal of Approval (DSA)

Semantic Web
Rule Interchange Format (RIF)
Resource Description Framework (RDF)

Semantics
Semantics Of Business Vocabulary And Rules (SBVR)

Subject classification
LCSH - Library of Congress Subject Headings
CSH - Canadian Subject Headings, Library and Archives Canada
BISAC subject headings
Dewey Decimal Classification (DDC) (OCLC)
Universal decimal classification (UDC) system
Controlled Vocabularies - Data Documentation Initiative (DDI)

Version Control
Dataverse Universal Numeric Fingerprint (UNF)

Government of Canada Standards and Documents

Treasury Board of Canada Standards
Electronic Documents and Records Management Solutions (EDRMS), Standard for
Geospatial Data, Standard on
Government of Canada Open Data Metadata Element Set (based on Dublin Core)
Metadata, Standard on
Optimizing Websites and Applications for Mobile Devices, Standard on
Privacy and Web Analytics, Standard on

TBITS 23: Information Technology Vocabulary - Implementation Criteria
TBITS 3: Coded Character Set for Information Interchange - Implementation Criteria
TBITS 36: All-Numeric Representation of Dates and Times -Implementation Criteria
Web Interoperability, Standard on
Web Usability, Standard on

Treasury Board of Canada Policy Documents

Access to information
Framework for Information and Technology
Information management
Information Management Guidelines
Information Management Roles and Responsibilities
Open Government Directive
Recordkeeping
Strategic Direction for Government: Information Management

Other Government of Canada Standards

MANAB – Environment Canada Manual of Word Abbreviations
MANCLIM – Environment Canada Manual of Climatological Observations
Guide to harmonize ISO 19115:2003 / North American profile metadata for government of Canada geospatial data, v2.0, October 2, 2014. Not available online.