Environmental Data Management Best Practices

BASIC CONCEPTS

St. Lawrence Global Observatory
July 2015

Document's Core Metadata:

TITLE	Environmental Data Management Best Practices – Basic Concepts
CREATOR	St. Lawrence Global Observatory (SLGO)
CREATION DATE	2015-06-02
LAST REVISION DATE	2015-07-09
TOPIC	Science Data Management
STATUS	Version 1
EDITOR	SLGO
TYPE	Text
DESCRIPTION	Overview of environmental data management principles, methods and resources intended for SLGO members and partners in order to provide basic notions, to contribute to a better understanding of the data life cycle and to foster the use of scientific data management best practices.
CONTRIBUTOR(S)	Original French Version: J. Hamel (Author); B. Robineau, C. Tremblay et K. Ratté (Reviewers) English Version: J. Hamel (Translator)
FORMAT	MS Word Document
SOURCE	©SLGO
RIGHTS	Public
IDENTIFIER	SLGODataManagement-2015.docx
LANGUAGE	English
RELATION	Non applicable
COVERAGE	Non applicable

This publication should be cited as follows:

St. Lawrence Global Observatory (SLGO). 2015. Environmental Data Management Best Practices – Basic Concepts. 36 p.

Financial support provided by:

Environment Canada Environnement Canada

Environmental Data Management Best Practices – Basic Concepts

Table of Content

Environmental Data Management Best Practices

Basic Concepts

1. Context

A large amount of data and information is produced on an ongoing basis by departments, research organizations, communities and citizens. However, all data are not readily understandable, accessible and usable by anyone or any organization. Basic notions of data management can contribute to improving data integrity, sustainability and accessibility.

The objective of the current document is to offer such basic concepts, to contribute to a better understanding of the data life cycle, to provide principles and to foster the adoption of environmental data management best practices. Such an endeavour is part of the St. Lawrence Global Observatory mandate as a service to its members, specifically targeted towards SLGO members without all the necessary infrastructures and means required to manage their data.

Each step of the data life cycle is presented and discussed. Additionally, three of the main themes of this cycle are further developed under specific titles: sustainability, metadata and data sharing.

Research and monitoring data represent an irreplaceable scientific heritage that must be:

- DOCUMENTED
- PRESERVED
- SECURED
- VALORIZED
- SHARED
- REUSED

2. Definitions

To organizations, research groups or project teams, data management means the development and implementation of the entire data life cycle architecture, processes and practices.

The current document provides a general overview of basic scientific data management concepts in order to help readers better understand technical and often complex notions, starting with the following definitions:

Data

Any information that can be stored electronically including text, numbers, images, video, audio, software, algorythms, equations, animations, models, simulations, etc. Data can be generated through various means such as observation, calculation or experimentation.[1]

Dataset

A dataset is a collection of structured data (often in table format) where the fields (columns) correspond to the different variables and the lines display the different values for those variables. Several file formats are used including structured formats (Ex. CSV – Coma Separated Values), geospatial (Ex. GeoTIFF) and XML (eXtended Markup Language) used for metadata.

Data Service

A data service makes data (including text, image, video, and audio) available via Internet. Ex.: RSS feed.

Georeferenced Data/ Geospatial Data/ Geodata

Geospatial/georeferenced data include geographical locations such as X-Y [Latitude-Longitude] coordinates or, at least, a reference to a site from which positions can be calculated. Geodata often includes a vertical component Z [depth/altitude].

GEOSS

The Global Earth Observing System of Systems is an international group of organizations combining their expertise across nine topics or «societal benefits ». GEOSS contributes to the monitoring, analysis and accessibility of data in these areas of interest.[2]

GEOSS SOCIETAL BENEFITS: disasters, health, energy, climate, agriculture, ecosystems, biodiversity, water and weather.

Harvesting

The process of collecting metadata descriptions from different sources/registries to facilitate data discovery.

[1] National Science Board. 2005. Long Lived Digital Data Collections: Enabling Research and Education in the 21st Century. 92 p. http://www.nsf.gov/geo/geo-data-policies/nsb-0540-1.pdf

[2] Fontaine, K.S. 2007. Architecture and Data Management Challenges in GEOSS and IEOS. 10 p. http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20070017998.pdf

Information

In general, the difference between « data » and « information » lies in the fact that data refers to raw observations acquired from research or monitoring activities while information is obtained by processing and/or interpreting data[3].

Interoperability

Interoperability is the capability of a product or a computer system to function with other existing products or systems without restrictions and independently of their own physical architecture and operating systems. Interoperability can be achieved through the use of Internet open standards. The mission of the World Wide Web Consortium – W3C[4] is to provide guidance and to contribute to the Web evolution by developing protocoles, standards and guidelines supporting interoperability.

ISO

The International Organization for Standardization (ISO) is the world's largest developer of voluntary international standards in a variety of areas ranging from currency codes, to water meters requirements, to date and time representation. ISO standards give world-class specifications for products, services and systems, to ensure quality, safety and efficiency[5].

Open Data

Governments are increasingly adopting an « open » approach in order to improve accessibility of publicly funded data and information.[6] The same concept is also guiding non-governmental initiatives aiming at fostering transparency, accountability and reuse of data.

Gartner, a world leader in information technology research, defines open data as "information or content made freely available to use and redistribute, subject only to the requirement to attribute it to the source"[7]. **Non-proprietary open data formats** allow producers to save data in a way that lets users access data without having to buy any specific software (or a particular version of software). Ex.: text files with the .ODF extention (Open Document Format).[8]

[3] International Oceanographic Data and Information Exchange (IODE) - Marine Data Management. http://www.iode.org/index.php?option=com_content&view=article&id=3&Itemid=33
[4] World Wide Web Consortium (W3C). http://www.w3.org/
[5] International Organization for Standardization (ISO). http://www.iso.org/iso/home/about.htm
[6] Gouvernement of Canada – Open Data. http://open.canada.ca/en/open-data
[7] Gartner – Open Data. http://www.gartner.com/it-glossary/open-data
[8] ISO. 2006. OpenDocument OASIS standard for data interoperability of office applications. http://www.iso.org/iso/home/news_index/news_archive/news.htm?refid=Ref1004

Metadata

Data that describes other data. The reference for metadata is the ISO 19115 international standard.

A world without metadata...

...would look like a video club without any thematic sections where hundreds of DVD boxes would be displayed without titles or covers and where even the disks would not be labelled.

How could anyone find any particular movie? How would it be possible to know what is available whithout having to open each box and play each DVD?

How would it be possible to find environmental data about the St. Lawrence if data producers did not document their datasets and did not publish the existence of their data in order to make them discoverable?

Registry / Catalogue

Catalogue services enable the publication and search of descriptive information (metadata) about data and data services. They can also harvest metadata from other catalogues[9]. The Open Geospatial Consortium (OGC) differentiates « catalogue » and "registry" by stating that a registry is a specialized catalogue that is maintained by an official entity in compliance with access procedures and policies and content management (ISO 19135, ISO 11179-6 standards).

Standard

Document that defines the specifications, characteristics, guidelines or requirements to ensure the consistent use of products, processes and services.

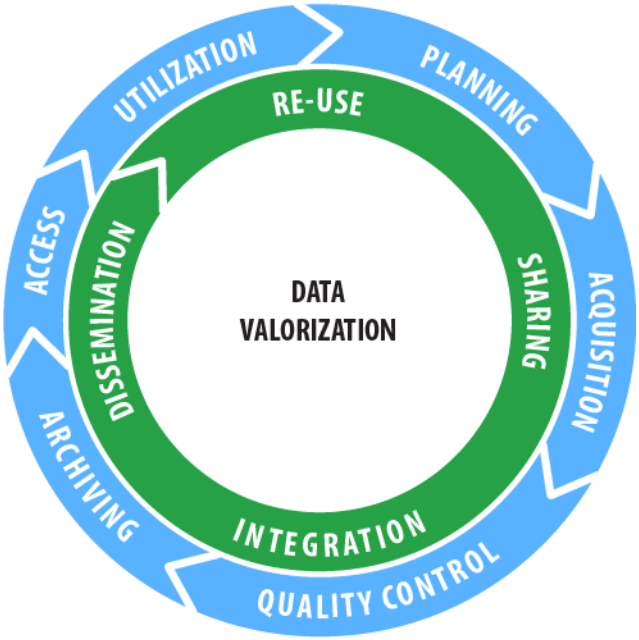
[9] Open Geospatial Consortium (OGC).2014. OGC I15 (ISO19115 Metadata) Extension Package of CS-W ebRIM Profile 1.0. 136 p. http://www.opengis.net/doc/ISx/csw-ebrim-i15/1.0

3. Data Life Cycle

3.1 SCOPE OF LIFE CYCLE

Data life cycle has been defined and described in many ways ranging from simple linear representations to more complex iterative processes[10]. The figure below summarizes the cycle major phases going from planning the acquisition to using the data.

DATA LIFE CYCLE



DATA LIFE CYCLE: throughout the transformation processes, the intrinsic value of data is amplified as a result of the various ways to use it, share it, integrate it and/or use it to develop products & services.

Throughout that cycle, data producers are responsible for documenting every phase, procedure, transformation and analysis in order to allow data users to better understand any dataset and make the most efficient and sensible use of the data.

[10] Committee on Earth Observation Satellites - Working Group on Information Systems and Services, Data Stewardship Interest Group. 2011. CEOS Data Life Cycle Models and Concepts. TN01, Issue 1.0. 78 p.

3.2 PLANNING

Planning is the first and most important step of the data management process because it entails the development of a vision that will encompass all subsequent steps.

Before collecting data, it is imperative to define objectives, to analyze needs, to identify means, to anticipate risks and challenges, and to envision future data use. The final cost of data management – including human resources and equipment – always represents a significant portion of operating and/or research budgets. Therefore, defining a sound data management and conservation strategy is an essential step that will become a definite investment in the long run.

PROJECT PROPOSALS...
... should include **resources** to be **dedicated** to data management (from data collection to sharing).

The components of data collection planning are described below. They will be developed by scientists/investigators according to their own context (laboratory, field, operational or research activities, etc.). This list can be used as a planning tool and check-list.

- **Needs** [why? for whom?]
  - Data collection requirements
  - Business/research objectives
  - Users/clients
  - Target audience
  - Anticipated benefits/results
- **Data** [what?]
  - Identification of existing data
  - Selection of variables
  - Type of data: digital, images, video, audio, animations, simulations, models
  - Standard formats
- **Methodology** [how?]
  - Sampling plan
  - Experimental/laboratory protocol
  - Data collection process
  - Data access mechanisms
- **Required resources** [how? who? how much?]
  - Financial means: existing / to be obtained
  - Human resources: expertises & skills required, staff availability, training needs, roles & responsibilities
  - Equipement: existing / to be acquired

Assessment of the necessary resources for data management: efforts required beyond the data collection process have to be considered e.g. for the entire data life cycle. Assessing and including data management costs in the early stages of a project ensure that it will not become a burden along the way.

> $ Data management costs can be assessed by looking at efforts that will be required to collect, validate and structure data and to develop data access mechanisms. Litterature often refers to amounts equivalent to 6 to 10% of research project budgets.

- **Scope of project** [when? where? how much?]
  - Period: dates/duration, frequency
  - Area of interest: geographical situation, definition et delineation of area of interest
  - Deliverables: Specific, Measurable, Achievable, Realistic, Timely – SMART
- **Issues** [what? how?]
  - Risk analysis
  - Identification of opportunities

3.3 ACQUISITION

The amount of environmental data collected worldwide using sensors and satellites during experimentation, observation, simulation and sampling activities is so enormous that it is often described as a "data deluge"[11]. It is therefore most important to make sure that data be thoroughly documented in order to facilitate their discovery and their use by search engines and data harvesting/data mining tools.

3.3.1 CONTROLLED VOCABULARY

One of the key elements to consider is the use of controlled vocabulary or recognized standard terminology. Adopting an approach that is closest to those internationally recognized is essential as it ensures that data will be identifiable, understood, discovered and accessed by users or by information systems. Consequently, using variable names and data dictionaries commonly used by the international community will prevent confusion. To support the needs of certain communities, servers are often setup by organizations willing to provide users with the appropriate terminology.

[11] Hey, A.J.G. and A.E. Trefethen. 2003. The Data Deluge: An e-Science Perspective. In, Berman, F., G.C. Fox and A.J.G. Hey (eds.). Grid Computing - Making the Global Infrastructure a Reality. Wiley and Sons, p. 809-824.

Below are a few examples of standard terminology resources:

GCMD NASA's ***Global Change Master Directory (GCMD)*** [12] is an environmental data reference. GCMD not only provides controlled vocabulary but also offers high quality resources for scientific data discovery, access and sharing.

ICES **International Council for the Exploration of the Sea (ICES)** provides reference codes for oceanographic data, trawling survey and commercial data as well as codes for sampling platforms.[13]

BODC ***British Oceanographic Data Center (BODC)*** of the Natural Environment Research Council (NERC) [14] is a designated data center and offers standard terminology in a wide range of topics. It is part of the ***SeaDataNet*** network, a marine European data management infrastructure.[15]

ITIS ***Integrated Taxonomic Information System (ITIS)***[16] is an integrated information system providing official taxonomic information about plants, animals, fungus and microbes (species names, TSN codes and hierarchical classification).

WoRMS ***World Register of Marine Species (WoRMS)***[17] is a registry of official lists of marine organisms and species name synonyms. The Canadian version is the ***Canadian Register of Marine Species (CaRMS)***[18].

CFMetadata ***CF (Climate and Forecast) Metadata***[19] is a NASA and Earth Science Data Systems Working Group convention that fosters interoperability between data producers, users and services using clear and non ambiguous standards for the representation of geolocations, time and quantities.

MMI ***Marine Metadata Interoperability (MMI)***[20] promotes the exchange, integration and use of marine data and fosters efficient publication, discovery, documentation and accessibility. It offers a semantic framework, vocabulary standards and metadata documentation tools.

[12] Olsen, L.M., G. Major, K. Shein, J. Scialdone, S. Ritz, T. Stevens, M. Morahan, A. Aleman, R. Vogel, S. Leicester, H. Weir, M. Meaux, S. Grebas, C.Solomon, M. Holland, T. Northcutt, R. A. Restrepo and R. Bilodeau. 2013. NASA/Global Change Master Directory (GCMD) Earth Science Keywords. Version 8.0.0.0.0. http://gcmd.nasa.gov/learn/keyword_list.html
[13] International Council for the Exploration of the Sea (ICES). Vocabulary Server. http://vocab.ices.dk
[14] Natural Environment Research Council (NERC). Vocabulary Server. http://www.bodc.ac.uk/products/web_services/vocab
[15] SeaDataNet - Pan-European Infrastructure for Ocean & Marine Data Management. Common Vocabularies. http://www.seadatanet.org/Standards-Software/Common-Vocabularies
[16] Integrated Taxonomic Information System (ITIS). http://www.itis.gov
[17] World Register of Marine Species (WoRMS). http://www.marinespecies.org
[18] Canadian Register of Marine Species (CaRMS). http://www.marinespecies.org/carms/index.php
[19] Climate and Forecasts (CF) Metadata. https://earthdata.nasa.gov/standards/climate-and-forecast-cf-metadata-conventions
[20] Marine Metadata Interoperability (MMI). https://marinemetadata.org/

3.3.2 COLLECTION

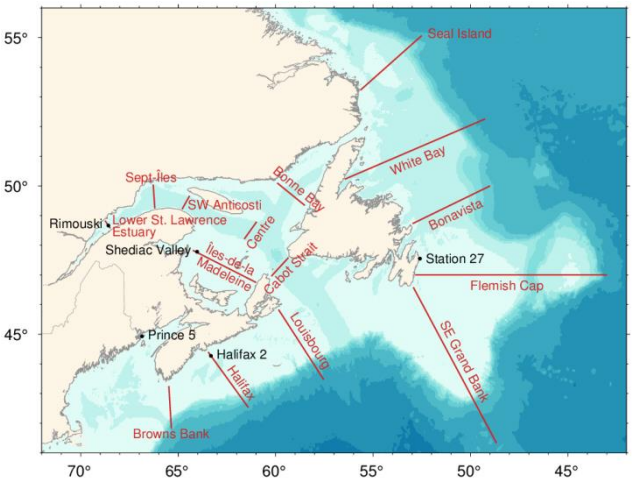
Data collection has greatly evolved during recent years. With new information technology advancements, sophisticated electronic data entry devices are replacing traditional tools and paper forms. Furthermore, tasks that used to take months to complete are currently being executed in record time as a result of the improved capabilities of computer systems.

Before going ahead with data collection, the following should be given consideration:

What?	Identification of variables to be measured, nature & format of data, litterature review, analysis of existing/published data, sampling plan, experimental protocols, documentation;
When?	Temporal coverage of the project, frequency of measurements/observations, work schedules;
Where?	Spatial coverage of the project, sampling sites, sensor networks, laboratory spaces;
How much?	Assessment of data volume to be produced and required archiving space;
How?	Type of measurements/observations (in the field, at sea), laboratory experimentation and simulations, equipment, procedures, data entry protocols, calibration, quality control/assurance processes;
Who?	Participants/staff: availability, skills, training. Data users: needs.

3.3.3 SAMPLING & MONITORING INITIATIVES

Regular resource sampling and environmental monitoring initiatives are mainly part of governmental mandates. Several ongoing monitoring programs carried out by various departments allow scientists to compile time series that are used to assess resource status (flora, fauna, human populations, ecosystems) and to detect environmental changes (air, water, ground, contaminants).[21] The Atlantic Zone Monitoring Program (AZMP) is an example of a large



AZMP: Atlantic Zone Monitoring Program sampling stations and transects. Source: St. Lawrence Global Observatory - SLGO http://ogsl.ca/en/azmp/context.html

[21] Office of the Auditor General of Canada. 2011. Report of the Commissioner of the Environment and Sustainable Development. Ch. 5. A Study of Environmental Monitoring. http://www.oag-bvg.gc.ca/internet/English/parl_cesd_201112_05_e_36033.html#appa

scale oceanographic data collection initiative.

Communities are also increasingly interested in environmental issues and get involved in data collection activities. For example, the *Comités ZIP* in Quebec are regional organizations bringing together the main users of the St. Lawrence while fostering a coherent approach to solving local and regional issues affecting ecosystems.[22]

Similarly, the general public can be involved. Citizens are progressively welcome to contribute their own observations through initiatives such as the Capelin Observation Network (CON)[23] and the Marine Mammal Observation Network (MMON)[24]. SLGO has developed a collaborative Web platform to support such programs, allowing individuals to enter their own data online[25].

3.4 QUALITY

Whether it is a scientific advice, a discovery, an analysis or a report, its credibility is directly linked to the quality of the data used to produce it. Quality management is therefore a key component of the science data management process.

3.4.1 QA/QC

- **QUALITY ASSURANCE (QA)**
  Quality assurance refers to actions beyond tests and controls, and aims at identifying and preventing non conformity problems. QA also includes training and audits.

  Real-time data acquisition involves specific procedures due to the fact that data is being used as soon as it is produced. This presents a level of risk for users who could misinterpret data or could not recognize inaccuracies. As an example, the Integrated Ocean Observing System – IOOS has developed a series of automated procedures for real-time ocean data control (QARTOD: Quality Assurance for Real-Time Oceanographic Data). [26]

[22] Stratégies Saint-Laurent. Les Comités ZIP. http://www.strategiessl.qc.ca/les-organismes/les-comites-zip
[23] Capelin Observation Network (CON). http://slgo.ca/en/biodiversity/fish/dfo-capelin/network.html
[24] Réseau d'observation de mammifères marins (ROMM). http://www.romm.ca/
[25] St. Lawrence Global Observatory (SLGO). Crowdsourcing. http://ogsl.ca/en/crowdsourcing.html
[26] Integrated Ocean Observing System (IOOS). 2014. Manual for the Use of Real-Time Oceanographic Data Quality Control Flags. 19 p. http://www.ioos.noaa.gov/qartod/temperature_salinity/qartod_oceanographic_data_quality_manual.pdf

- **QUALITY CONTROL (QC)**
  Quality control generally refers to activities aiming at products & services quality verification by looking at problems and abnormal elements. In the area of data management, data is examined in order to identify errors, missing data or acquisition problems (see: section 3.4.4 Quality flagging).

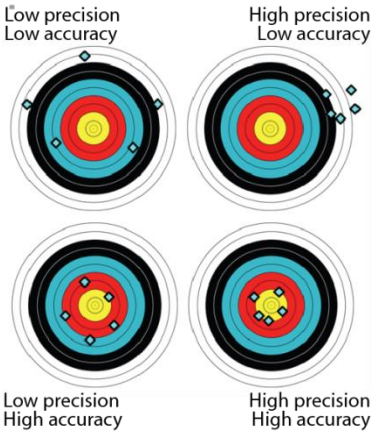
- **QUALITY MANAGEMENT**
  Quality management represents a higher level of management e.g. organizational or institutional. ISO standard 9001 describes a quality management model including procedures, management of responsibilities, resources and services, as well as management of quality assessment, analysis and improvement.

  Statistics Canada – an organization based on data – defines its quality management framework and the quality of its information products in terms of their fitness for use by clients. It is a multidimensional concept including the **relevance** of the information to **users' needs**, as well as its **accuracy**, **timeliness**, **accessibility**, **interpretability** and **coherence**.[27]

3.4.2 PRECISION & ACCURACY

Often incorrectly used as synonyms, these two terms have very different meanings as illustrated by this diagram.

ISO standard 5725[28] refers to trueness as a better word for describing the principles and methods used for measuring and assessing the accuracy of a test. Trueness is defined as the degree of agreement between the arithmetic mean of a large number of test results and an accepted reference value; precision refers to the agreement between test results.



PRECISION & ACCURACY: one can be very precise but… also very inaccurate.

Data management errors are generally defined by their source:

- **human**: errors in reading, typing, transcribing or in judgement;
- **environmental**: errors due to ambiant conditions having an influence on measurements. Ex.: air temperature;
- **equipment**: errors due to equipment limitations and/or capacities.

BEST PRACTICES

TO LIMIT ERRORS:

- Ensure staff is **appropriately trained**;
- Perform quality control tests. Ex.: comparing measurements and observations to **references**.

[27] Statistics Canada. 2002. Statistics Canada's Quality Assurance Framework. http://www5.statcan.gc.ca/olc-cel/olc.action?ObjId=12-586-X&ObjType=2&lang=en&limit=0

[28] ISO 5725-1:1994. Accuracy (Trueness and Precision) of Measurement Methods and Results. https://www.iso.org/obp/ui/#iso:std:11833:en

Errors can also be called « systemic » when they present similarities in occurrence and pattern ("bias" is also used) or, on the contrary, called « random » when no particular trend is detected.

3.4.3 VALIDATION

During the validation phase, datasets are examined in order to find errors, to verify completeness and to ensure they are trustworthy and ready to be used.

Rules are put in place for specific data types and contexts to help identify missing data, inconsistencies and errors. For example, a rule could state that an acceptable water temperature value for a specific stream has to be between T_{min} and T_{max}, corresponding to the minimum and maximum observed at a specific site during a specific period.

3.4.4 QUALITY FLAGGING

During the validation process, data are examined and can be flagged. For instance, data can be assigned certain attributes to describe their quality. This information can be documented and included into the database. Data are then associated to codes allowing users to further understand the data and their relevance.

As an example, UNESCO (2010) has defined a series of quality flags for water temperature and salinity data (Global Temperature-Salinity Profile Program – GTSPP)[29]. In this case, quality codes are integers between 0 and 9.

Code	Description
0	No quality control has been assigned to this element
1	The element appears to be correct
2	The element appears to be probably good
3	The element appears doubtful
4	The element appears erroneous
5	The element has been changed
6 to 8	Reserved for future use
9	The element is missing

Units associated with each variable should be clearly noted. As an example, the Oceanographic Data Management System references section on the SLGO portal provides a description of units used for a wide range of environmental variables from athmospheric pressure to dissolved oxygen concentration.[30]

[29] UNESCO-IOC. 2010. GTSPP Real-Time Quality Control Manual, First Revised Edition. IOC Manuals and Guides No. 22, Revised Edition. IOC/2010/MG/22Rev. 145 p. http://iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=6437

[30] Oceanographic Data Management System (ODMS). Variables and Units. http://slgo.ca/app-sgdo/en/pdf/docs_reference/variables-unites.pdf

3.5 ARCHIVES

3.5.1 FORMATS

Format refers to the way data is organized in an electronic file. Saving data in sustainable standard formats is a major data management issue. Well preserved data will always be available to users regardless of the way accessing tools are evolving.

BEST PRACTICES

DATA ARCHIVING:

- save data using **non proprietary open formats** to allow data to be accessible independantly of software types and versions;
- use **clear identifiers** for data files;
- define an **archiving strategy** including the use of two different physical formats Ex.: hard disk and CD stored in two different physical locations;
- **scan** paper documentation and save it in portable file format (ex.: PDF/A);
- ensure data **archiving location** (numerical and non-numerical) is **appropriate** and **safe**;
- copy data to **new support** (magnetic tape, disk, etc.) 2 to 5 years after their creation as certain media can deteriorate;
- check archived data **integrity** regularly.

For instance, ISO standard 19005 proposes the use of PDF/A as a long term data conservation format[31]. ISO and the IEC (International Electrotechnical Commission) have also approved the use of open formats such as ODF[32], designated ISO/IEC 26300, in order to ensure interoperability and efficient data access. This standard is used namely for "Office Suite" type of documents including texts, spreadsheets, presentations and graphical elements.

The US Government Library of Congress has adopted a comprehensive approach to format issues including the development of a detailed decision process with respect to data preservation and file formats.[33]

Image and video files have also been examined by many. For instance, Nozères (2011) has done an in-depth analysis providing detailed descriptions of best practices in terms of image/video file formats, geotagging, metadata and archiving options[34].

[31] ISO. 2009. ISO 19005-1:2005 Document management - Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1). http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920

[32] Open Document Format (ODF). http://www.opendocumentformat.org/aboutODF

[33] US Government – Library of Congress. Digital Preservation. http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml

[34] Nozères, C. 2011. Gestion des données d'images en sciences aquatiques: une introduction aux bonnes pratiques et aux flux de travail. Rapp. Tech. can. Sci. halieut. Aquat. 2962F: xiv + 195 p. http://www.dfo-mpo.gc.ca/Library/345814.pdf

FILE NAMING:

- use **meaningful** file names [Ex.: ***2012-Tadoussac-beluga009.xls*** for beluga biological data file 009 collected in 2012 in the Tadoussac area]
- avoid using **spaces**, **accents** and **special characters** such as $/%& @!
- avoid **very longs** file names
  [Ex.: ***2012-08-19 Tadoussac bio samples beluga whales JeanLambert.xls]***

3.5.2 CONSERVATION

Data conservation ensures that data are secured and backed-up. It is an essential step, often postponed or forgotten, that requires short-term and long-term plans.

First, data back-ups have to be made on a regular basis in order to keep a current version of all data collected. With back-ups, one can react to events such as corrupt data files or loss of data. Back-up integrity should be checked regularly and copies should be kept in a safe place. It is also important to clearly identify all raw data, modifications, computations as well as the various file versions.

Eventually, consideration should be given to using a data warehouse or data center. Most institutions have such facilities and they have staff dedicated to data conservation. One of the main issues related to data conservation is the sustainability of physical supports e.g. type of media used for storage such as hard drives, software, processors, etc. As information and communication technologies evolve rapidly, verifications of the computer infrastructure become essential in order to keep up with technological advancements. System functionality should be tested regularly and operating system migrated as required. Aside from institutional data centers, there are those organized around the specific interests of communities, for example the GBIF information infrastructure for biodiversity data.

Another type of storage that has become quite popular is "cloud" archiving where virtual storage space is made available to users on remote servers, often free of charge. In this case, users should ensure that the service is reliable, safe, with proper access controls and will respect data confidentiality.

For more information, see Section 4. Data Sustainability.

3.5.3 DOCUMENTATION

Data documentation helps users understand the context and methods used during the data collection and archiving processes. It explains how scientists have organized,

GOOD DOCUMENTATION INCLUDES:

- **context**, data collection project history, objectives and hypothesis;
- **methods** used: for collection, analysis, instrumentation, software, calibration;
- **structure** of dataset(s);
- **procedures** for data control, verification and validation;
- **changes** to data and file version history;
- **data use conditions** and confidentiality;
- **metadata**.

described, coded, formatted, computed and validated data. Metadata are also part of dataset documentation.

3.6 ACCESS

Although facilitating data access seems widely acceptable, it appears that freely sharing one's data is not always that easy. In fact, several issues are to be considered including intellectual property and recognition. According to many, cultural changes are required within the scientific community, not only to valorize publications but also the data used as the **basis** for those publications.

> "We need to change the culture of science to one that equally values publications and data."
>
> Nelson, B. 2009. Empty Archives. NATURE, Vol. 461. p. 160-163.

3.6.1 INTELLECTUAL PROPERTY (COPYRIGHT)

Data ownership has to be clearly stated whether it belongs to a scientist or to the institution where he/she works. This becomes a key element when defining data access parameters.

Contexts differ across research institutions, departments and non profit organizations. Generally, the state owns data produced by government scientists. In universities, sometimes scientists have rights to their data, sometimes, the institution owns the data.

As more alliances and partnerships are established between various organizations, intellectual property rights and recognition definitely have to be clearly stated.

Suggested readings:

- *Gouvernement du Québec*. 2002. *Gestion de la propriété intellectuelle dans les universités et dans les établissements du réseau de la santé et des services sociaux où déroulent des activités de recherche.* 29 p.
  http://www.frsq.gouv.qc.ca/fr/ethique/pdfs_prop_int/plan_pi.pdf
- National Research Council Canada (NRCC). 2014. NRC's Intellectual Property (IP) policy.
  http://www.nrc-cnrc.gc.ca/eng/about/intellectual_property/index.html
- Penn State University. Acquisition, Management, Sharing and Ownership of Data.
  http://www.research.psu.edu/training/sari/teaching-support/data-management
- Natural Sciences and Engineering Research Council of Canada (NSERC). 2009. Policy on Intellectual Property.
  http://www.nserc-crsng.gc.ca/NSERC-CRSNG/Policies-Politiques/ip-pi_eng.asp
- Côté, M.-È. et J.M. Hébert. 2002. *Sommaire des politiques et règlements relatifs à la propriété intellectuelle de certaines universités.* ROBIC. 28 p.
  http://www.robic.ca/admin/pdf/701/294-MEC.pdf

3.6.2 CONTROLLED/RESTRICTED ACCESS

Who will be allowed to find, read, use, analyze, transform, disseminate, and share the data? If this is not an open access context, all these questions are most relevant.

It is important to determine who will have data access privileges throughout the data production process and to clearly define the boundaries with respect to data edition/modification.

When data are ready for publication, decisions have to be made regarding the conditions of use. It is possible to limit data use to a specific user group, a research team or a community of practice, especially when dealing with sensitive or confidential data. In such cases, user authentication can be mandatory (with user identification and password requirements) in order to determine who the data consumers are. However, as soon as data use is controlled or restricted, an efficient data access management process has to be in place.

3.6.3 DISCOVERY

As mentioned earlier (See Section 2. Definitions), metadata means data about data. They are used to document datasets and to help discover, describe and promote the existence of data. For more details, see Section 5. Metadata.

In a document about imagery management in aquatic science, Nozères (2011) proposes a set of core metadata for image and video files:

- File name (highly recommended)
- Creation Date (automatic, default)
- Creator - Author (highly recommended)
- Keywords (highly recommended)
- Copyright Status (optional)
- Title (optional)
- Description - Légende (optional)
- Location – sub-location (optional)
- GPS (latitude, longitude, altitude)

3.6.4 FINDING DATA - REGISTRIES / CATALOGUES

Registries and catalogues are used both to publish and search through metadata collections. They help document and find descriptive information about existing data and services.

DISCOVERY

Publising the existence of data increases:

- project visibility;
- number of citations;
- data re-use;
- collaboration opportunities;
- contribution to other research;
- data valorization.

These directories can also harvest metadata from other registries. The Open Geospatial Consortium (OGC) makes a distinction between "catalogue" and "registry" by stating that a registry is a specialized catalogue maintained by an official entity in compliance with access and content management procedures and policies (ISO 19135 and ISO 11179-6 standards).

For more information, see Section 6. Data Sharing.

3.7 UTILISATION

3.7.1 LICENCES AND RIGHTS

Similar to the data access framework, data utilisation requires clearly defined rules and boundaries. A licence is a tool that can be used to let users know how the data owner wishes to manage the exploitation of its data. There is a wide range of licence types, which often complicates the situation for both producers and data users[35]. However, there is a trend towards simplification and openness as can be seen with GEOSS's efforts to facilitate access to earth observation data[36].

As an example, a licence can include the following elements:

- Letting users copy and distribute data by stating:

  « Anyone can, without authorization and free of charge, copy, publish, adapt, translate and communicate by any means, any document, data and information available on this site »

- Demanding that credit be given to author(s) or data owner according to a specific format:

  « … conditional to citing the source as follows:
  SOURCE: PRODUCER XYZ, HTTP://XYZ.CA, (YEAR). »

- Allowing (or not) commercial use of data:

  « … use of data for personal or commercial use… ».

3.7.2 DATA AND INFORMATION INTEGRATION

Increasing efforts dedicated to improving access and data sharing prompt issues related to integrating[37] multiple sources of data. In a perfect world, all data producers would adopt international standards and would use accessible and interoperable computer environments. But what is really the current situation?

HETEROGENEITY

On a daily basis, everyone is using some form of calendar. However, looking at how people write the date, one can easily understand that heterogenous procedures,

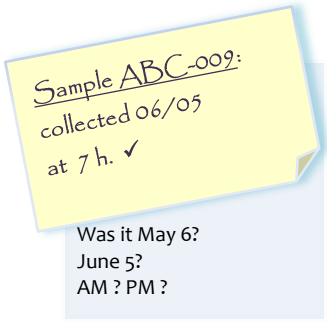
[35] Campbell, J. 2015. Access to Scientific Data in the 21st Century: Rationale and Illustrative Usage Rights Review. Advance Publication, Data Science Journal. 28 p. https://www.jstage.jst.go.jp/article/dsj/advpub/0/advpub_14-043/_pdf

[36] Onsrud, H., J. Campbell and B. van Loenen. 2010. Towards Voluntary Interoperable Open Access Licenses for the Global Earth Observation System of Systems (GEOSS). International Journal of Spatial Data Infrastructures Research, 2010, Vol. 5, 194-215. http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/168/203

[37] Ludäscher, B., K. Lin, S. Bowers, E. Jaeger-Frank, B. Brodaric and C. Baru.2005. Managing Scientific Data: From Data Integration to Scientific Workflows. 21 p. http://users.sdsc.edu/~ludaesch/Paper/gsa-sms.pdf

formats, syntax and systems are in fact major issues with respect to data integration. For example, although the international ISO standard 8601 defines YYYY-MM-DD as the official date format, a wide variety of ways to write 2015-02-12 can be found as shown in the table below:

12/02/15	15/02/12	02/12/15	15/12/02
12-02-15	15-02-12	02-12-15	15-12-02
12-02-2015	2015-02-12	02-12-2015	2015-12-02
Feb. 12/15	February 12, 2015	12 fév. 2015	12 février 2015
12.02.2015	2015.02.12	12 de febrero	Etc. etc. etc.



Generally speaking, we can see that combining datasets where variables are represented in different formats can cause problems. The same goes for the units, measurement precision or cartographic projections used. Rigor and consistency are therefore essential.

ASSIMILATION IN MODELS

The work of meteorologists illustrates well the use of models: weather experts feed a variety of environmental parameters into climate models in order to produce the best forecasts possible. The cycle of data assimilation in this process adds *in situ* observations into the models as a way to fine-tune the forecasts. For instance, this is how the coupled water-athmosphere model developed by Saucier *et al* [38] [39] can produce surface current forecasts for the Estuary and Gulf of St. Lawrence.[40]

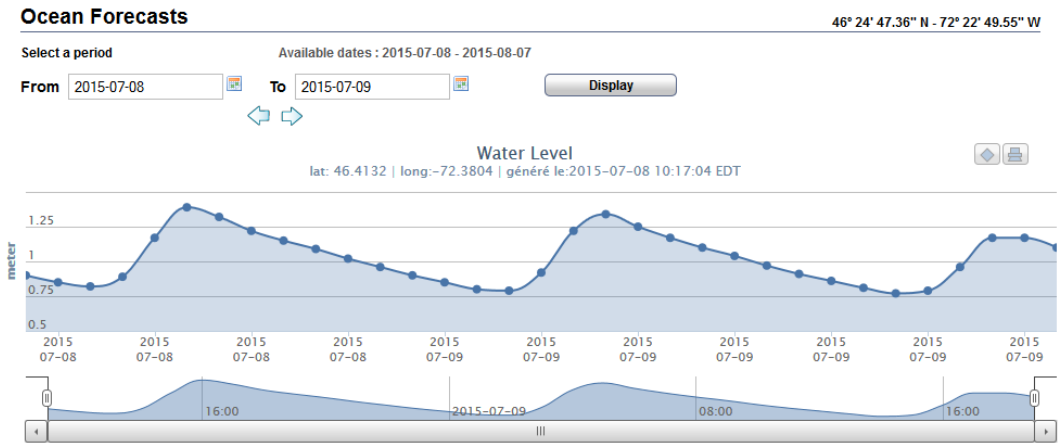
3.7.3 VISUALIZATION AND ANALYSIS

Data producers can use a variety of tools in order to see their data and better understand and analyse them. Whether it is a map, a chart or a table, vizualisation tools provide options for combining various informations into a single view which helps to identify trends or anomalies.

[38] Saucier, F.J., F. Roy, S. Senneville, G. Smith, D. Lefaivre, B. Zakardjian et J.-F. Dumais. 2009. Modélisation de la circulation dans l'estuaire et le golfe du Saint-Laurent en réponse aux variations du débit d'eau douce et des vents. Revue des sciences de l'eau / Journal of Water Science, vol. 22, n° 2. p. 159-176. http://www.ismer.ca/IMG/pdf/Saucier_et_al_2009_RSE.pdf

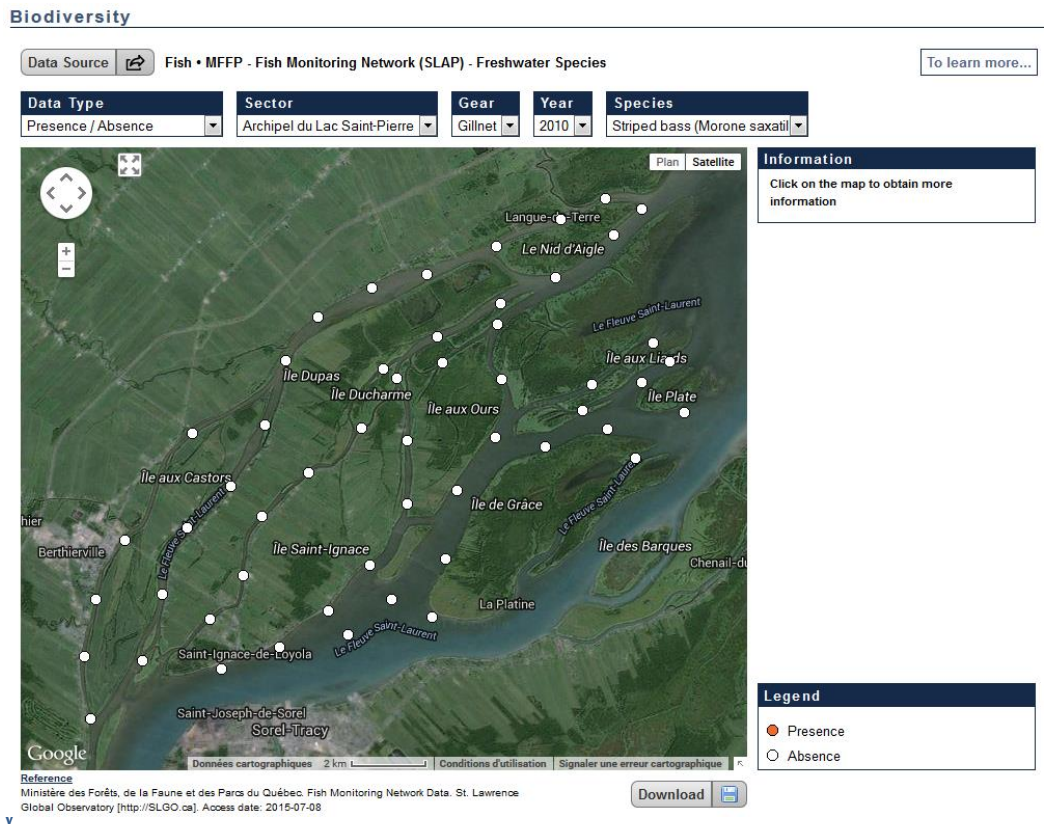
[39] Gouvernement du Canada. Environnement Canada, Modélisation. https://meteo.gc.ca/model_forecast/model_f.html

[40] St. Lawrence Global Observatory (SLGO). Ocean Forecasts. http://slgo.ca/ocean



TIME-SERIES VISUALIZATION: Web tool allowing users to move through series of water level forecasts for the Estuary and Gulf of St. Lawrence for selected time periods. Source: St. Lawrence Global Observatory (SLGO). http://slgo.ca/ocean

Data visualization tools can help refine the data validation process. For example, using a map to view data point distribution and sampling locations can help identify possible errors in geographical coordinates.



SPATIAL VISUALIZATION: striped bass sampling stations showing presence/absence of specimens. The tool helps users move throughout the dataset by selecting sectors, fishing gear, year and species. Source: St. Lawrence Global Observatory (SLGO). http//slgo.ca/bio

3.7.4 DECISION MAKING PRODUCTS AND SERVICES

Decision making is a process that requires analyzing the context, identifying stakeholders, and assessing risks, issues, uncertainties and consequences. Science is often a key component of decision making; therefore it has to be accessible, worthy and understandable.

The Old Harry project, taking place in the Gulf of St. Lawrence, is a good example of how science plays a role in a context of oil & gaz exploration and pressure placed upon a region. On the one hand, Bourgault *et al* (2014) [41] have shown that industry's decision making process and policies were lacking scientific considerations. On the other hand, ocean circulation models have shown the most sensitive areas to oil spills. These results, and those of upcoming studies, will become the knowledge base for any future implementation of oil & gaz exploitation projects in this region.

3.7.5 USER FEEDBACK

A good example of how important data management can be is how highly the United States considers science. On February 18, 2015, the US Government has appointed Dr D.J. Patil to the White House as its first US Chief Data Scientist. By creating this "data champion" key position within its administration, the state sends a clear signal with respect to the significance of data and the many ways they can be used to serve citizens.[42] Patil has even stated that user feedback is the most important component of data science development.[43] In the current document, it is discussed in the last step of the data life cycle (Ref. Section 3. under "Utilization"). Although it is shown here closing the loop, user feedback can however be included at every step of the data management process. Comments and suggestions from user groups can be very useful when trying to improve and adjust processes. They are a definite benefit to data producers who agree to give them consideration.

[41] Bourgault, D., F. Cyr, D. Dumont and A. Carter. 2014. Numerical simulations of the spread of floating passive tracer released at the Old Harry prospect. Environ. Res. Lett. 9 054001. 14 p. http://dx.doi.org/10.1088/1748-9326/9/5/054001

[42] White House. 2015. The White House Names Dr. DJ Patil as the First U.S. Chief Data Scientist. http://www.whitehouse.gov/blog/2015/02/19/memo-american-people-us-chief-data-scientist-dr-dj-patil

[43] Weathington, J. 2015. Start Tackling Data Science Inefficiencies by Properly Defining Waste. TechRepublic. http://www.techrepublic.com/article/start-tackling-data-science-inefficiencies-by-properly-defining-waste/

4. Data Sustainability

As discussed earlier, data archiving and conservation procedures and mechanisms are key to efficient and sustainable data access. This is very eloquently described in an *Archives de France* document on the importance of scientific and technical documents long-term conservation. It is stated that while data collection is highly expensive and technological breakthrough are essential, the loss of data would be a regression, an irreversible loss of scientific capital which is in turn critical to any new discovery.[44]

Currently, as products and services are becoming increasingly numerical, managing the entire data life cycle is no longer optional but turns out to be essential. The growing demand for data requires the capacity to efficiently respond.

4.1 DATA MANAGEMENT IMPLEMENTATION

Good data management and sharing provide benefits for both data producers and users, namely by:

- contributing to scientific progress;
- increasing exposure of scientific research;
- helping to comply to funding agency policies;
- reducing duplication of research efforts;
- allowing to reproduce and validate results;
- increasing cooperation opportunities.

4.2 CHALLENGES AND ISSUES

At the national level, there is still a lot of work to be done before one can say that all data are well managed. This is what the Canada Institute for Scientific and Technical Information (CISTI) Research Data Strategy Working Group has concluded in 2008 while uncovering important gaps with respect to research data management in Canada[45]. The examination of key indicators allowed the Committee to quantify gaps between the current situation and an ideal state. As of today, these findings are still relevant and represent challenges and issues institutions are presently facing:

- **Policies**: few organizations have clear data management policies;
- **Funding**: long-term data management funding is deficient e.g. large gaps in funding all stages of data life cycle;
- **Roles and responsibilities**: stakeholders responsibilities throughout the data life cycle are not clearly defined, certains roles are not assigned, lack of data management planning;

[44] Ministère de la Culture et de la Communication. 2003. La valorisation et la pérennisation des données scientifiques et techniques. Bulletin des Archives de France sur la conservation à long terme des documents électroniques. N° 10. 2003. 8 p. http://www.archivesdefrance.culture.gouv.fr/static/1674

[45] Canada Institute for Scientific and Technical Information- Research Data Strategy Working Group. 2008. Stewardship of Research Data in Canada: A Gap Analysis. 31 p. http://publications.gc.ca/collections/collection_2009/cnrc-nrc/NR16-123-2008E.pdf

- Trusted Digital **Data Repositories**: few existing data repositories for all research subject areas, a large portion of research data are still stored on scientists' hard drives;
- **Standards**: interoperability and metadata standards are not always known and/or respected;
- **Skills and Training**: there are few competent scientists dedicated to data management;
- **Reward and Recognition Systems**: few mechanisms are in place for the recognition of contributions with respect to research data management and sharing;
- **Research and Development**: lack of national priorities to support the coordination and orientation of R&D activities in Canada;
- **Accessibility**: researchers are reluctant to share data; certain contradictions are found between policies (« privacy and ethics » and « access and preservation»);
- **Preservation:** lack of committment towards long-term preservation, research organizations are rarely demanding data management plans.

Currently, the main federal funding agencies such as the Natural Sciences and Engineering Research Council of Canada (NSERC)[46], the Social Sciences and Humanities Research Council (SSHRC)[47] and the Canadian Institutes of Health Research (CIHR)[48] have developed and/or have adopted policies with respect to access to scientific research data in Canada[49]. The underlying principles are:

- advancement of knowledge,
- reduction of duplicated research efforts,
- increase of research benefits and
- promotion of researchers' results.

4.3 DATA MANAGEMENT POLICIES

At the organizational level, there are good examples of complehensive data management policies such as the one developed and implemented by the Science sector of the Department of Fisheries and Oceans Canada (DFO).[50] The main objective of this policy is to ensure the preservation of science data and therefore to ensure long-term use through interoperability, accessibility and standards. DFO's expected results are:

- secure data now and in future,
- interconnected and interoperable data;
- useful data discoverable and accessible through standard means; and

[46] Natural Sciences and Engineering Research Council of Canada. Strategic Partnership Grants. http://www.nserc-crsng.gc.ca/Professors-Professeurs/RPP-PP/SPG-SPS_eng.asp
[47] Social Sciences and Humanities Research Council (SSHRC). Research Data Archiving Policy. http://www.sshrc-crsh.gc.ca/about-au_sujet/policies-politiques/statements-enonces/edata-donnees_electroniques-eng.aspx
[48] Canadian Institutes of Health Research (CIHR). Tri-Agency Open Access Policy on Publications. http://www.cihr-irsc.gc.ca/f/32005.html
[49] Gouvernment of Canada – Sciences. Access to Research Results: Guiding Principles. http://www.science.gc.ca/default.asp?Lang=En&n=9990CB6B-1
[50] Department of Fisheries and Oceans Canada (DFO). Policy for Scientific Data. http://www.dfo-mpo.gc.ca/science/data-donnees/policy-politique-eng.htm

- cost effective data management.

Another example is the International Polar Year (IPY) 2007-2008 data management policy, an international interdisciplinary observation and research program for the advancement of the global knowledge of polar processes. "The overarching objective of IPY 2007-2008 data management is to ensure the security, accessibility and free exchange of relevant data that both support current research and leave a lasting legacy."[51]

Generally speaking, scientific data management underlying policies include:

- recognizing the **value** (priceless and irreplaceable) of scientific data and the need to ensure they are well managed to guarantee their **conservation** and **sustainability**;
- making sure data are **available**, **accessible, relevant** and **reliable** by managing the entire data life cycle, from acquisition to dissemination;
- implementing **structured** and **secured** data repositories to ensure long-term preservation;
- acknowledging that recent data are the most critical for decision making and making sure they are **timely** and accessible;
- fostering **open access** while respecting the confidential and/or sensitive nature of data;
- promoting data **exchange** and **sharing** with the international scientific community to enhance knowledge.

[51] International Polar Year (IPY) 2007-2008. Canadian IPY 2007-2008 Data Policy. http://www.api-ipy.gc.ca/pg_IPYAPI_055-eng.html

5. Metadata

"A metadata record is an information file that captures the basic characteristics of a geographic data or information resource, and represents the **who**, **what**, **when**, **where**, **why** and **how** of the geodata resource." [52] Metadata allow producers to thoroughly describe datasets in order to help users understand their context and limitations and be able to assess their usefulness. Metadata are also used in catalogues, helping data discovery and repurposing to the benefit of users and the scientific community.

5.1 STANDARDS

ISO 19115[53] is the international metadata standard. It includes hundreds of fields describing datasets characteristics. For a long time, the Federal Geographic Data Committee (FGDC) [54] standard was implemented by many organizations but FGDC has decided to follow the international trend and to facilitate the transition towards a common ISO standard.

PROFILES

A profile is a subset of a standard which combines relevant metadata fields for a specific context.

The **North American Profil (NAP)** is a Canadian standard developed to suit the needs of georeferenced data users and producers in Canada and in the US. "The goal of this Profile is to provide a mechanism for organizations producing geographic information to describe datasets in detail. The Profile helps users to better understand geographic metadata, the assumptions and limitations of geographic information, and facilitates the search for proper information to fit users' needs." [55]

The **Marine Community Profile of ISO 19115** was developed by the Australian Ocean Data Centre Joint Facility (AODCJF) as a subset of the ISO 19115 standard. It defines additional components, codes and vocabulary to support marine data description.[56]

Generally, profiles define elements (mandatory, conditional, optional) to be used to describe datasets including data type, topic, geographical & temporal boundaries, language, contact.

[52] Natural Resources Canada - Digital Geospatial Metadata. http://www.nrcan.gc.ca/earth-sciences/geomatics/canadas-spatial-data-infrastructure/standards-policies/8912
[53] ISO Standard 19115. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=53798
[54] Federal Geographic Data Committee (FGDC). http://www.fgdc.gov/metadata/geospatial-metadata-standards
[55] North American Profil (NAP) of ISO 19115:2003 — Geographic Information — Metadata. http://nap.geogratis.gc.ca/metadata/napMetadata-eng.html
[56] Marine Community Profile (MCP) of ISO 19115. https://marinemetadata.org/references/marineprofile19115

5.2 PUBLICATION AND METADATA SEARCH

Several initiatives allow data producers to create and publish their own metadata records and also to search available directories. Below are a few examples:

Canadian Geospatial Data Infrastructure (CGDI) – Discovery Portal [57]
The GeoConnections Discovery Portal is a metadata catalogue that enables users, developers and data suppliers to find, assess, access, visualize and publish Canadian geospatial and geoscience data products and Web services.

Infrastructure for Spatial Information in the European Community (INSPIRE) [58]
The INSPIRE metadata editor is a user friendly interface available in 22 languages. A user guide[59] is provided explaining how to enter and register metadata records.

Polar Data Catalog (PDC) [60]
The Polar Data Catalogue includes data and metadata generated by researchers in the Arctic and Antarctic in compliance with international standards. The public search tool provides access to various themes including social, health and environmental sciences.

BlueNet Metadata Entry and Search Tool (MEST) [61]
The Australian Marine Science Data Network metadata entry and search tool supports archiving and accessing scientific research data.

Open Archives Initiative – (OAI) [62]
The OAI develops and promotes interoperability standards in support of efficient Web content dissemination and numerical resources accessibility especially by education networks (*eScholarship, eLearning, eScience*).

GeoNetwork OpenSource [63]
GeoNetwork OpenSource is a type of catalogue application, downloadable and open source. It is available in various languages, it can search and generate metadata and it also offers a user manual.

GeoDoc [64]
The GeoDoc metadata editor helps create, import and export metadata according to various standards.

[57] Canadian Geospatial Data Infrastructure (CGDI) – Discovery Portal. http://geodiscover.cgdi.ca/web/guest/home
[58] Infrastructure for Spatial Information in the European Community (INSPIRE). Metadata Editor. http://inspire-geoportal.ec.europa.eu/editor/
[59] Infrastructure for Spatial Information in the European Community (INSPIRE). Metadata Editor User Guide. http://www.eurogeoss.eu/Documents/EuroGEOSS_D_2_2_3.pdf
[60] Polar Data Catalogue (PDC). https://www.polardata.ca/
[61] Australian Ocean Data Network (AODC) BlueNet MEST. http://bluenet.aodn.org.au/geonetwork/srv/en/main.home
[62] Open Archives Initiative (OAI). http://www.openarchives.org
[63] GeoNetwork OpenSource. http://geonetwork-opensource.org
[64] GeoDoc. http://www.gogeo.ac.uk/gogeo/metadata/geodoc.htm

6. Data Valorization

New information technologies help accessing, reusing and valorizing data. They help create opportunities for collaborations, support innovation and foster the development of new research initiatives. Scientific data dissemination and sharing are key elements facilitating data valorization; they help democratize data access and also provide exposure for scientific research. Such benefits motivate governments, organizations and scientists across the world and encourage the adoption of data management best practices throughout the data life cycle.

6.1 PRINCIPLES

The **Gouvernment of Canada** has established a series of open data principles[65]. These can easily apply to any dataset:

1. **Completeness:** datasets should be as complete as possible and should reflect the entirety of what is recorded about a particular subject including metadata explaining raw data as well as details about calculation methods.
2. **Primacy:** datasets should come from a primary source, including original data collected and details about how data was collected in order to allow users to verify that data was collected and recorded properly and accurately.
3. **Timeliness:** data should be made available to users in a timely fashion without delays.
4. **Ease of Physical and Electronic Access**: datasets should be as accessible as possible without complicated access conditions ou requirements to using complex technologies.
5. **Machine Readability**: datasets should be stored in widely-used file formats that easily lend themselves to machine processing (e.g. CSV, XML). These files should be accompanied by documentation related to the format and how to use it in relation to the data.
6. **Non-discrimination**: there should not be obstacles to accessing datasets by anyone, at any time, nor should there be a need to identify oneself and to provide justifications.
7. **Use of Commonly Owned Standards**: datasets should be in freely available file formats as often as possible; users should not need to get a particular application to read the data.
8. **Licencing**: the use of an open licence increases openness and minimizes restrictions on the use of the data.
9. **Permanence**: For optimal use, online information should remain online, with appropriate version-tracking and archiving over time.
10. **Usage Costs**: open data is free of charge.

[65] Gouvernment of Canada – Open Data. http://open.canada.ca/en/open-data-principles#toc95

The **US Gouvernment** has a similar approach. The US White House Project Open Data[66] states that data will be documented, complete, reusable, public and timely. An action plan about governmental open data was published in 2014.[67]

On the **international scene**, the Global Earth Observation System of Systems (GEOSS) has also adopted such principles in the context of data sharing being essential to supporting societal benefits[68]. Complete, timely and open accessibility of metadata and data is a key element of this strategy. The International Oceanographic Data and Information Exchange (IODE) of the Intergovernmental Oceanographic Commission (IOC) of UNESCO is a good example of an oceanographic data and information exchange program.[69]

The large number of international initiatives has prompted GEO to setup a working group tasked with the examination of data production and management worldwide. Among its conclusions, the group has identified a need to harmonize the various approaches, terminologies and definitions.[70]

6.2 BENEFITS

A study by RIN/NESTA[71] about UK researchers' practices has demonstrated that those who had adopted an open approach had a definite advantage over those who did not. Benefits included:

- Increased **efficiency** (or decreased duplication of efforts), sharing good practices and tools, decreased data collection costs;
- **Promotion** of scientific **rigour** and **quality** of research among peers by sharing methods and protocoles as well as results (including negative ones);
- Increased **visibility** of research and **collaboration opportunities** amongst scientists and with various communities (including public engagement);
- Enhanced **scope** of research facilitating data re-use and exchange and sharing expertise amongst researchers and institutions;
- Increased **socio-economic impact** and **innovation opportunities** and better cooperation beyond the scientific community.

Several studies show a correlation between sharing research data and the number of citations. [72] Litterature shows that citations can increase by 69% in cases with open

[66] US White House Project Open Data. https://project-open-data.cio.gov/
[67] US Government. 2014. U.S. Open Data Action Plan. 21 p. http://www.whitehouse.gov/sites/default/files/microsites/ostp/us_open_data_action_plan.pdf
[68] Global Earth Observation System of Systems (GEOSS). http://www.earthobservations.org/geoss.php
[69] International Oceanographic Data Exchange (IODE). http://www.iode.org/
[70] GEO Data Sharing Working Group. 2013. Interpretation of the "Full and Open" Access to and Use of (Geographic) Data: Existing Approaches. 13 p. https://www.earthobservations.org/documents/dsp/201310_full_and_open_principle_interpretation_living_paper.pdf
[71] Research Information Network (RIN) and National Endowment for Science, Technology and the Arts (NESTA). 2010. Open to all? Case Studies of Openness in Research. 52 p. http://www.rin.ac.uk/our-work/data-management-and-curation/open-science-case-studies

research data.[73] Such recognition for research results proves to be a definite benefit for researchers.

Openness, sharing and data valorization are becoming the basis of a trend within the scientific community. Increasingly, it can be observed throughout multidisciplinary projects, the inclusion of citizen input and the development of numerous data access, visualization and dissemination tools.

7. Tools

Although several initiatives and references have been presented throughout the document, the following are offered as additional suggestions:

PLANNING

The data life cycle as described earlier requires a series of procedures and frameworks specific to each one of its phases. The overall process is documented as a **data management plan**.

Appendix 1 is presented as a data management planning tool and also as a checklist. It will serve as a reminder of the entire data management process and will help monitor all elements of the plan as they were detailed throughout this document.

DISSEMINATION AND SHARING

Generally, ocean observation systems such as the US Integrated Ocean Observing System use Web platforms as data dissemination tools. In Quebec, a large number of data producers benefit from using the St. Lawrence Global Observatory as their official data dissemination mechanism.

Another example is the PANGAEA[74] information system which is a data sharing tool covering a wide selection of geoscientific and environmental datasets. It has search functions, allows for data archiving and provides a DOI (Digital Object Identifier) to help reference data and give credit to authors.

DATA MANAGEMENT SYSTEM

During regular SLGO members annual meetings (*Rencontres annuelles d'orientation des membres - ROM*) as well as those of the SLGO Scientific Advisory Committee (SAC), the needs of universities, NGOs and government partners have been widely discussed. One of the main issues identified was the need to increase the organizations' internal data

[72] Piwowar, H.A. and T.J. Vision. 2013. Data reuse and the open data citation advantage. PeerJ 1:e175. http://dx.doi.org/10.7717/peerj.175

[73] Piwowar, H.A.,R.S. Day and D.B. Fridsma. 2007. Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000308

[74] PANGAEA. Policy for the Data Library PANGAEA - Publishing Network for Geoscientific & Environmental Data Open Access Data Archive and Publishing System for Earth System Research. 4 p. http://www.pangaea.de/curator/files/pangaea-data-policy.pdf

management capabilities. One of the proposed solutions was the implementation of an environmental data management system (EDMS) as an open source application. Although that type of development requires a lot of efforts, this a relatively low-cost option for organizations and it can also serve as an incentive for them to adopt best practices and a sustainable approach to data management. Along with experts from DFO and UQAR-ISMER, SLGO has developed such an EDMS Web application (to be available in 2015)[75].

METADATA DOCUMENTATION

As a national solution, DFO has also developed an application called JMetaWriter available in both French and English. This application can be deployed on SLGO partners' infrastructures to help them document their datasets and consequently feed international data infrastructures.

[75] To be confirmed.

8. References

Recommended documents and useful links about data management:

Association des bibliothèques de recherche du Canada (ABRC). 2009. Les données de recherche: un potentiel insoupçonné. 16 p. http://www.carl-abrc.ca/uploads/pdfs/data_mgt_toolkit-f.pdf

Boston University. Research Data Management - Data Life Cycle.
http://www.bu.edu/datamanagement/background/data-life-cycle/

Canadian Science Policy Conference (CSPC). 2014. 6th Canadian Science Policy Conference Proceedings. 36 p.
http://www.sciencepolicy.ca/sites/default/files/cspc-proceedings-commercial-printer-v1_0.pdf

Creative Common Licenses. http://creativecommons.org/licenses/by-nc-nd/3.0/

Données de recherche Canada. http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/fra/index.html

Glushko, R.J. 2014.The Discipline of Organizing: Core Concepts Edition. O'Reilly Media. Ebook ISBN:978-1-4919-1203-4 | ISBN 10:1-4919-1203-0. 420 p.

Gouvernement du Canada. Canada numérique 150. http://www.ic.gc.ca/eic/site/028.nsf/fra/accueil

Gouvernement du Canada. 2014. Plan d'action du canada pour un gouvernement ouvert 2014-2016. ISBN: 978-1-100-25318-3. No de catalogue BT22-130/2014F-PDF
http://ouvert.canada.ca/fr/contenu/plan-daction-du-canada-gouvernement-ouvert-2014-2016

Humphrey, C. Preserving Research Data in Canada.http://preservingresearchdataincanada.net/

Infrastructure for Spatial Information in the European Community (ISPIRE). http://inspire-geoportal.ec.europa.eu/

Integrated Ocean Observing System (IOOS). http://www.ioos.noaa.gov/

Integrated Ocean Observing System (IOOS). Quality Control of Real-Time Oceanographic Data - QUARTOD.
http://www.ioos.noaa.gov/qartod/welcome.html

Interagency Working Group on Digital Data. 2009. Harnessing the Power of Digital Data for Science and Society. Report to the Committee on Science of the National Science and Technology Council. 60 p.
https://www.nitrd.gov/About/Harnessing_Power_Web.pdf

International Organization for Standardization (ISO). L'ISO et l'eau. Des solutions mondiales aux enjeux mondiaux. http://www.iso.org/iso/fr/iso_and_water.pdf

International Organization for Standardization (ISO). 2012. Principes de management de la qualité. 12p.
http://www.iso.org/iso/fr/qmp_2012.pdf

International Oceanographic Data and Information Exchange (IODE). Ocean Data Portal - Seamless Access to Ocean Data. http://www.oceandataportal.org/

Joint Technical Commission for Oceanography and Marine Meteorology (JCOMM). Catalogue of Practices and Standards. http://bestpractice.iode.org/all_records.php

Jones, S. 2011. How to Develop a Data Management and Sharing Plan. DCC How-to Guides. Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides/develop-data-plan

Nelson, B. 2009. Empty Archives. NATURE, Vol 461|10. p. 160-163.
http://www.nature.com/news/2009/090909/pdf/461160a.pdf

Stevens, J., J. M. Smith, and R. A. Bianchetti. 2012. Mapping Our Changing World. Editors: Alan M. MacEachren and Donna J. Peuquet, University Park, PA: Department of Geography, The Pennsylvania State University. https://www.e-education.psu.edu/geog160/node/1922

Tenopir C., S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff and M. Frame. 2011. Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6(6): e21101. doi:10.1371/journal.pone.0021101 http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021101

Therriault, J.-C., B. Petrie, P. Pepin, J. Gagnon, D. Gregory, J. Helbig, A. Herman, D. Lefaivre, M. Mitchell, B. Pelchat, J. Runge, and D. Sameoto. 1998. Proposal for a northwest Atlantic zonal monitoring program. Can. Tech. Rep. Hydrogr. Ocean Sci. 194: vii+57 p http://www.dfo-mpo.gc.ca/Library/224076.pdf

US Geological Survey (USGS). Data Management. http://www.usgs.gov/datamanagement/why-dm/lifecycleoverview.php

Van den Eynden, V., L. Corti, M. Woollard, L. Bishop and L. Horton. 2011. Managing and Sharing Data - Best Practices for Researchers. UK Data Archive. ISBN 1-904059-78-3. 40 p. http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Wiggins, A., R. Bonney, E. Graham, S. Henderson, S. Kelling, R. Littauer, G. LeBuhn, K. Lotts, W. Michener, G. Newman, E. Russell, R. Stevenson and J. Weltzin. 2013. Data Management Guide for Public Participation in Scientific Research. DataONE: Albuquerque, NM. 15 p. https://www.dataone.org/sites/all/documents/DataONE-PPSR-DataManagementGuide.pdf

Appendix 1. Data Management Plan Elements – Check-list.

PROJECT TITLE: _____		PROJECT NO.: _____		
ELEMENT	**DESCRIPTION**	**REFERENCE DOCUMENTS**	**ELEMENT VERIFIED BY**	**DATE**
❑ **Data description**	Data description (nature, scope and volume).	⇩ ⇩ ⇩ Record file names and storage locations in this column.		
❑ **Existing data**	Review of existing data relevant to the project, their usefulness, how they can be integrated.			
❑ **Roles & responsibilities**	Identification of project participants, their roles & accountability, their tasks throughout the data life cycle.			
❑ **Format**	Data formats when generated, archived, distributed with emphasis on their relevance to ensure sustainable use and archival of data.			
❑ **Data organization**	Data file management, file naming procedures, version control, etc.			
❑ **Metadata**	Metadata description, identification of standard to be used.			
❑ **Quality**	Data quality control/assurance procedures throughout the project.			
❑ **Archiving & preservation**	Archiving methods, backups, resources and spaces used (physical & virtual). Long-term conservation procedures.			
❑ **Security**	Data protection procedures including sensitive and confidential data, description of data access controls restrictions, privileges and methods.			
❑ **Intellectual property**	Identification of intellectual property (IP) owner(s) (individual or organization), IP protection methods, when necessary.			
❑ **Audience**	Description of data users.			
❑ **Access & sharing**	Description of data sharing methods, access mechanisms, procedures and conditions (open or restricted), estimated date of data publication.			
❑ **Ethics & privacy**	If necessary, ethics and privacy protection procedures.			
❑ **Budget**	Costs of managing all data life cycle phases, funding sources, funding requests.			
❑ **Regulatory issues**	Funding agencies' requirements.			