

Bonnes pratiques en gestion de données environnementales

NOTIONS DE BASE



Métadonnées (de base) du présent document :

TITRE	Bonnes pratiques en gestion de données environnementales – Notions de base
CRÉATEUR	Observatoire global du Saint-Laurent (OGSL)
DATE DE CRÉATION	2015-01-14
DATE DE DERNIÈRE RÉVISION	2015-07-15
SUJET	Gestion des données scientifiques
ÉTAT	Version 11
ÉDITEUR	OGSL
TYPE	Texte
DESCRIPTION	Vue d'ensemble des principes, méthodologies et ressources en matière de gestion des données environnementales à l'intention des membres et des partenaires de l'OGSL afin de leur fournir les notions de base, de contribuer à une meilleure compréhension du cycle de vie des données et de favoriser l'adoption de bonnes pratiques pour la gestion des données scientifiques.
CONTRIBUTEUR(S)	J. Hamel (auteur), B. Robineau, C. Tremblay et K. Ratté (réviseurs)
FORMAT	MS Word Document
SOURCE	©OGSL
DROITS	Public
IDENTIFIANT	OGSLGestionDonnees-2015-V11-Final.docx
LANGUE	Français
RELATION	Sans objet
COUVERTURE	Sans objet

Référence pour cette publication :

Observatoire global du Saint-Laurent (OGSL.) 2015. Bonnes pratiques en gestion de données environnementales – Notions de base. 38 p.

Ce projet a été réalisé avec l'appui financier de :



Environnement
Canada

Environment
Canada

Bonnes pratiques en gestion de données environnementales - Notions de base

Table des matières

1. CONTEXTE	5
2. DÉFINITIONS	5
3. CYCLE DE VIE DES DONNÉES.....	9
3.1 PORTÉE DU CYCLE DE VIE	9
3.2 PLANIFICATION	10
3.3 ACQUISITION	11
3.3.1 VOCABULAIRE CONTRÔLÉ	11
3.3.2 SAISIE	13
3.3.3 INITIATIVES DE MONITORAGE ET D'ÉCHANTILLONNAGE	14
3.4 QUALITÉ	14
3.4.1 AQ/CQ (QA/QC)	15
3.4.2 PRÉCISION ET EXACTITUDE	16
3.4.3 VALIDATION	17
3.4.4 QUALIFICATION DES DONNÉES	17
3.5 ARCHIVAGE.....	18
3.5.1 FORMATS	18
3.5.2 PRÉSERVATION	19
3.5.3 DOCUMENTATION	20
3.6 ACCÈS.....	20
3.6.1 PROPRIÉTÉ INTELLECTUELLE (COPYRIGHT)	20
3.6.2 ACCÈS CONTRÔLÉ/RESTREINT	21
3.6.3 DÉCOUVERTE	22
3.6.4 DIFFUSION DE L'EXISTENCE DES DONNÉES - REGISTRES / CATALOGUES	22
3.7 UTILISATION	23
3.7.1 LICENCES ET DROITS	23
3.7.2 INTÉGRATION DES DONNÉES ET DES INFORMATIONS	23
3.7.3 VISUALISATION ET ANALYSE	24
3.7.4 PRODUITS ET SERVICES POUR LA PRISE DE DÉCISION	26
3.7.5 RÉTROACTION DES UTILISATEURS	26
4. PÉRENNITÉ DES DONNÉES	27
4.1 MISE EN ŒUVRE DE LA GESTION DES DONNÉES	27
4.2 DÉFIS ET ENJEUX.....	27
4.3 POLITIQUES DE GESTION DE DONNÉES	28
5. MÉTADONNÉES.....	30
5.1 NORMES	30
5.2 PUBLICATION ET RECHERCHE DE MÉTADONNÉES	31
6. VALORISATION DES DONNÉES PAR LEUR DIFFUSION	32
6.1 PRINCIPES	32
6.2 AVANTAGES	33
7. OUTILS.....	34
8. RÉFÉRENCES.....	36
ANNEXE 1. ÉLÉMENTS D'UN PLAN DE GESTION DES DONNÉES – LISTE DE CONTRÔLE	38

Bonnes pratiques en gestion de données environnementales

Notions de base

1. Contexte

Une quantité imposante de données et d'informations sur l'environnement est récoltée régulièrement par des ministères, des institutions de recherche, des groupes communautaires, par divers autres organismes et par des citoyens. Ces données recueillies depuis de nombreuses années ne sont cependant pas toutes sous une forme compréhensible, accessible et utilisable par d'autres individus ou par d'autres organisations. Afin de pouvoir améliorer l'intégrité, la pérennité et l'accessibilité des données, il importe d'abord de se sensibiliser à certaines notions en gestion de données.

Le présent document a été rédigé dans le but de fournir des notions de base, de contribuer à une meilleure compréhension du cycle de vie des données, de fournir des principes et de viser l'adoption de bonnes pratiques en matière de gestion de données environnementales. Cette démarche s'inscrit dans le mandat de l'Observatoire global du Saint-Laurent à titre de service à ses membres et s'adresse particulièrement aux partenaires de l'OGSL qui ne sont pas dotés des infrastructures nécessaires et des moyens requis pour la gestion de leurs données.

Chacune des étapes du cycle de vie des données est présentée et expliquée. Trois thèmes majeurs du cycle sont par la suite repris soit la pérennité, les métadonnées et le partage des données et font l'objet de sections distinctes du document.

Les données de recherche et de monitoring constituent un patrimoine scientifique inestimable qui doit être

- DOCUMENTÉ
- PRÉSERVÉ
- SÉCURISÉ
- VALORISÉ
- PARTAGÉ
- RÉUTILISÉ

2. Définitions

La gestion de données est un processus qui comporte de multiples ramifications autant au niveau du développement que de la mise en œuvre de l'ensemble de l'architecture, des processus et des pratiques de gestion de l'ensemble du [cycle de vie des données](#) d'une organisation, d'un groupe de recherche ou d'une équipe de projet.

La portée du présent document se voulant de nature générale, des notions de base sont présentées, afin de permettre au lecteur de s'y retrouver dans le domaine technique parfois complexe qu'est la gestion des données scientifiques, à commencer par les quelques définitions suivantes:

Données

Toute information pouvant être enregistrée en format numérique incluant texte, nombre, image, vidéo, audio, logiciel, algorithme, équation, animation, modèle, simulation, etc. Ces données peuvent être générées de diverses manières comme par observation, calcul ou expérimentation.¹

Données ouvertes

Les gouvernements adoptent de plus en plus une approche « ouverte » afin de rendre accessibles les données et les informations gouvernementales financées à l'aide des fonds publics.² Cette notion est également à la base d'initiatives non-gouvernementales favorisant la transparence, l'obligation de rendre compte et la réutilisation des données. Gartner³, un leader mondial en recherche sur les technologies de l'information, définit les données ouvertes comme de l'information ou du contenu rendu gratuitement disponible pour l'utilisation et la redistribution à condition d'en créditer la source. Les **formats de données ouverts non-propriétaires** conservent les données sous une forme qui permet à un utilisateur d'accéder librement aux données sans obligation de se procurer un logiciel (ou une version du logiciel) en particulier. Ex. : document texte avec extension .ODF (*Open Document Format*).⁴

Données géoréférencées / géospatiales / géodonnées

Les données géospatiales sont des données qui réfèrent à un lieu géographique et qui comportent une position X-Y [Latitude-Longitude] ou, au moins, une référence à un site dont on peut calculer la position. Souvent, les données possèdent également une composante verticale Z [profondeur/altitude].

Jeu de données

Un jeu de données est un ensemble de données organisées souvent sous une forme tabulaire où les champs (colonnes) correspondent aux diverses variables mesurées et les lignes aux valeurs de ces variables. Plusieurs formes existent incluant par exemple les formats structurés (ex. CSV – *Coma Separated Values*), géospatiaux (ex. GeoTIFF) et XML (*Extended Markup Language*) utilisé pour les [métadonnées](#) d'un jeu de données.

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 Gouvernement du Canada – Données ouvertes. <http://ouvert.canada.ca/fr/donnees-ouvertes>

³ Gartner – Open Data. <http://www.gartner.com/it-glossary/open-data>

⁴ 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

GEOSS

Global Earth Observing System of Systems: regroupement international d'organisations mettant en commun leurs expertises dans neuf domaines ou «bénéfices de société». GEOSS contribue au monitoring, à l'analyse et à l'accessibilité des données dans ces domaines.⁵



DOMAINES D'INTERVENTION DE GEOSS : désastres, santé, énergie, climat, agriculture, écosystèmes, biodiversité, eau et météo.

Information

De façon très générale, la distinction entre « données » et « informations » tient au fait que les données sont dites observables et brutes, résultant de recherches ou d'activités de monitoring alors que l'information fait référence à des données qui ont été traitées et/ou interprétées⁶.

Interopérabilité

L'interopérabilité est la capacité que possède un produit ou un système informatique à fonctionner sans restrictions avec d'autres produits ou systèmes existants et ce, indépendamment de leur architecture physique et de leurs systèmes d'opération respectifs, grâce à l'utilisation de normes ouvertes de l'Internet. Le *World Wide Web Consortium – W3C*⁷ s'est donné comme mission de guider et de contribuer à l'évolution du Web en développant des protocoles, des normes et des lignes directrices qui favorisent l'interopérabilité.

ISO

L'Organisation internationale de normalisation (ISO) est le premier producteur mondial de Normes internationales dans une multitude de domaines allant des codes pour la représentation des monnaies, aux normes pour les compteurs d'eau, en passant par celles décrivant la représentation de la date et de l'heure. Les Normes internationales ISO établissent des spécifications pour les produits, les services et les bonnes pratiques dans une optique de qualité, de sécurité et d'efficacité⁸.

⁵ 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>

⁶ 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

⁷ World Wide Web Consortium (W3C). <http://www.w3.org/>

⁸ Organisation Internationale de Normalisation (ISO). <http://www.iso.org/iso/fr/home/standards.htm>

Métadonnées

Fichier de données décrivant un jeu de données. La référence en la matière est la norme internationale ISO 19115.



Un monde sans métadonnées...

...serait comme un club vidéo sans sections thématiques et où on trouverait des centaines de boîtes de DVD dont les couvertures et étiquettes ont été retirées des boîtes et des disques eux-mêmes.

Comment trouver un film en particulier ?
Comment connaître l'inventaire disponible sans devoir ouvrir chaque boîte et faire jouer chaque DVD ?

Comment pourrait-on trouver des données environnementales sur le Saint-Laurent si aucun producteur ne documentait ses données et n'en publiait l'existence de façon à les rendre découvrables ?

Moissonnage

Action de récolter des métadonnées dans divers catalogues. Un service de moissonnage fait en sorte que des données décrites dans un registre puissent être plus largement découvertes.

Norme

Document définissant des exigences, des lignes directrices, des spécifications, ou des caractéristiques à employer systématiquement pour assurer l'utilisation cohérente de produits, de processus et de services.

Registre / Catalogue

Les services de catalogues permettent la publication et la recherche de collections d'information descriptive (métadonnées) à propos de données et de services de données. Ils ont, de plus, la capacité de moissonner (*harvesting*) des métadonnées dans d'autres catalogues⁹. L'Open Geospatial Consortium (OGC) fait une distinction entre « catalogue » et « registre » en précisant qu'un registre est un catalogue spécialisé maintenu par une entité officielle et se conformant aux procédures et aux politiques d'accès et de gestion du contenu (normes ISO 19135, ISO 11179-6).

Service de données

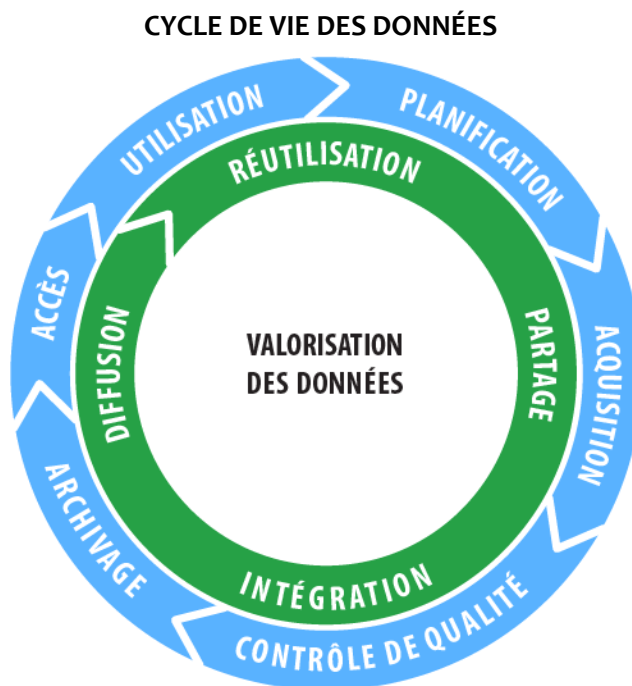
Un service de données rend des données (incluant texte, image, vidéo, audio) disponibles et les distribue via Internet. Ex. : Fil de nouvelles RSS.

⁹ Open Geospatial Consortium (OGC).2014. OGC 115 (ISO19115 Metadata) Extension Package of CS-W eBRIM Profile 1.0. 136 p. <http://www.opengis.net/doc/ISX/csw-ebrim-115/1.0>

3. Cycle de vie des données

3.1 PORTÉE DU CYCLE DE VIE

Il existe de nombreuses façons, plus ou moins complexes, d'illustrer les diverses étapes de la vie des données incluant des représentations linéaires, itératives et cycliques¹⁰. La figure ci-dessous en résume les grandes phases allant de la planification de la collecte de données jusqu'à leur utilisation.



CYCLE DE VIE DES DONNÉES : au cours des diverses étapes de leur transformation, la valeur intrinsèque des données est décuplée par leur réutilisation, par leur partage, par leur intégration à d'autres données ou par leur contribution au développement de produits et de services.

Tout au long du cycle, il incombe aux producteurs de données de bien documenter toutes les étapes, les procédures, les modifications et les analyses effectuées afin que les utilisateurs des données puissent bien les comprendre et en faire un usage le plus efficace et le plus éclairé possible.

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 PLANIFICATION

L'étape initiale de planification est sans aucun doute la plus cruciale du processus de gestion des données puisqu'elle doit nécessairement impliquer le développement d'une vision englobant l'ensemble des étapes subséquentes.

Avant de commencer le travail de collecte, il est essentiel de définir des objectifs, d'analyser les besoins, de prévoir les moyens, d'anticiper les risques et les enjeux, et de favoriser l'utilisation future des données. Le coût final de la gestion de données – incluant ressources humaines et équipement – représente toujours une portion significative des budgets d'opération et/ou de recherche. Définir une bonne stratégie de gestion et de conservation des données est donc une nécessité et représente un investissement qui s'avère un avantage indiscutable à long terme.

PRÉVOIR...

...d'inclure, dans les propositions de projets, des **ressources** qui seront **dédiées** à la gestion des données (de la collecte jusqu'au partage).



Les éléments présentés ci-dessous font partie de la préparation à une campagne de collecte de données. Ils seront développés de façon détaillée par les investigateurs/chercheurs responsables des projets selon leur contexte particulier (laboratoire, terrain, modélisation, activité opérationnelle, recherche, etc.). Cette liste pourra être utilisée à titre d'outil de planification et d'aide-mémoire.

□ Besoins

[pourquoi? pour qui?]

- Raisons de la récolte de données
- Objectifs d'affaires/de recherche visés
- Demandeurs/clients
- Auditoire visé
- Résultats attendus

□ Données visées

[quoi?]

- Identification et revue des données existantes
- Choix des variables à mesurer
- Nature des données à récolter : numériques, images, vidéo, audio, animations, simulations, modèles
- Adoption de formats standards

□ Méthodologie

[comment?]

- Plan d'échantillonnage
- Protocole expérimental
- Processus de récolte de données
- Mécanismes d'accès aux données

□ Ressources requises

[comment? qui? combien?]

- Moyens financiers : moyens existants et à solliciter

- Ressources humaines : expertises et compétences requises, disponibilité du personnel, besoins de formation, rôles et responsabilités
- Équipement : existant, à acquérir

Estimation des ressources adéquates pour la gestion des données recueillies : considérer les efforts requis au-delà du processus d'acquisition de données c.-à-d. pour tout le cycle de vie des données. Le fait d'estimer et d'inclure les coûts de la gestion de données en début de projet fait en sorte qu'elle ne devient pas un fardeau après coup.



On peut estimer les coûts de la gestion des données en évaluant les efforts requis pour la saisie, la validation, la structuration et le développement des mécanismes d'accès aux données. On parle dans la littérature de montants de l'ordre de 6 à 10% des budgets de projets de recherche.

□ Envergure du projet

[quand? où? combien?]

- Période : dates/durée, fréquence
- Territoire visé : situation géographique, définition et délimitation de la zone d'intérêt
- Livrables : clairs, gérables, atteignables, réalistes, définis dans le temps

□ Enjeux

[quoi? comment?]

- Analyse des risques
- Examen des opportunités

3.3 ACQUISITION

La quantité de données environnementales qui est récoltée à l'échelle mondiale lors d'expériences, d'observations, de simulations et de captation par divers senseurs et satellites est si énorme qu'elle va jusqu'à être qualifiée de « déluge de données »¹¹. Il importe donc, dans chaque cas, de s'assurer que les données soient documentées de façon appropriée afin qu'elles puissent être trouvées et utilisées par les moteurs de recherche et les outils de moissonnage de données (*data mining*, *data harvesting*).



3.3.1 VOCABULAIRE CONTRÔLÉ

Un des éléments clés à considérer est l'utilisation de vocabulaire contrôlé, c'est-à-dire d'une nomenclature reconnue et normalisée. Il est essentiel d'adopter une approche la plus près possible de ce qui se fait à l'échelle internationale de façon à ce que les données puissent être reconnaissables, intelligibles, compréhensibles et, éventuellement, retrouvables et accessibles par d'autres utilisateurs ou par divers systèmes d'information. Conséquemment, l'emploi de noms de variables reconnus et

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.

de dictionnaires de données établis et utilisés par la communauté internationale, permet d'éviter de la confusion. Pour des besoins spécifiques de certaines communautés, des organisations mettent également en place des serveurs qui fournissent aux utilisateurs la nomenclature appropriée.

Quelques exemples de ressources de référence sont présentés ci-dessous :

GCMD le **Global Change Master Directory (GCMD)**¹² de la NASA est une référence dans le domaine des données environnementales. GCMD fournit non seulement un vocabulaire contrôlé mais également des ressources de haute qualité pour la découverte, l'accès, l'utilisation et le partage de données scientifiques mondiales.

CIEM le **Conseil International pour l'Exploration de la Mer (CIEM)** met à disposition des codes de référence pour les données océanographiques, environnementales, les données de missions de recherche par chalutage et par pêche commerciale ainsi que les codes de plateformes d'échantillonnage.¹³

BODC le **British Oceanographic Data Center (BODC)**, centre de données désigné du **Natural Environment Research Council (NERC)**¹⁴, offre l'accès à des listes de termes normalisés couvrant un large éventail de disciplines pertinentes pour les communautés d'utilisateurs et fait partie du réseau **SeaDataNet**, une infrastructure pan-européenne de gestion de données marines et océaniques.¹⁵

ITIS **Integrated Taxonomic Information System (ITIS)**¹⁶ est un système intégré qui fournit de l'information taxonomique officielle sur les plantes, les animaux, les champignons et les microbes (noms d'espèces, codification taxonomique TSN et classification hiérarchique).

WoRMS le **World Register of Marine Species (WoRMS)**¹⁷ est un répertoire de listes officielles d'organismes marins et de synonymes des noms d'espèces. La version canadienne est le **Canadian Register of Marine Species (CaRMS)**¹⁸.

CFMetadata le **CF (Climate and Forecast) Metadata**¹⁹ est une convention de la NASA et du **Earth Science Data Systems Working Group**

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>

qui favorise l'interopérabilité entre les producteurs, utilisateurs et services de données grâce à une norme claire et non ambiguë pour la représentation de géolocations, du temps et de quantités physiques.

MMI Le **Marine Metadata Interoperability (MMI)**²⁰ fait la promotion de l'échange, l'intégration et l'utilisation de données marines et favorise la publication, la découverte, la documentation et l'accessibilité les plus efficaces. Il offre un cadre sémantique, des normes de vocabulaire et des outils de documentation de métadonnées.

3.3.2 SAISIE

La collecte de données a connu une évolution importante au cours des dernières années. Avec l'avènement des nouvelles technologies de l'information, les formulaires papier ont graduellement fait place aux outils de saisie électroniques des plus sophistiqués. De plus, ce qui prenait des mois de travail est présentement réalisable à une vitesse impressionnante grâce à la capacité des systèmes informatiques maintenant disponibles.

Avant de procéder à la récolte de données, il importe de prendre le temps de considérer les éléments suivants :

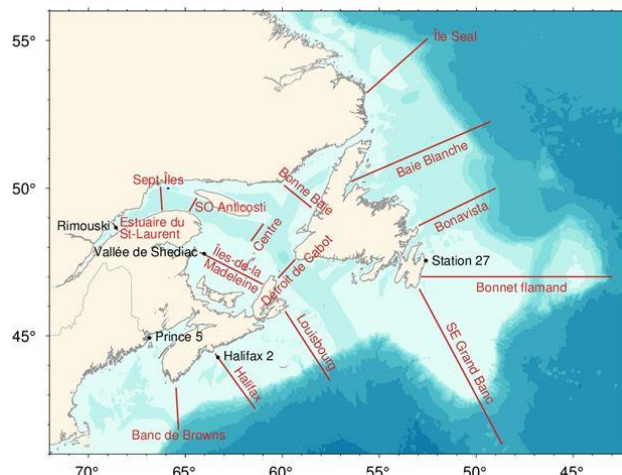
Quoi ?	Identification des variables à mesurer, revue de littérature, recherche de données publiées, examen de données existantes, nature et format des données à récolter, plan d'échantillonnage, protocoles expérimentaux, documentation;
Quand ?	Couverture temporelle du projet, fréquence des mesures/observations, horaires de travail;
Où ?	Couverture spatiale du projet, sites d'échantillonnage, réseaux de senseurs, espaces de laboratoire;
Combien ?	Estimation du volume de données qui seront produites et de l'espace d'archivage nécessaire;
Comment ?	Type de collecte, mesures et observations sur le terrain et en mer, expérimentation et simulations en laboratoire, appareillage requis, procédures d'utilisation, protocoles de saisie, calibration et étalonnage de l'équipement, processus de contrôle/assurance de la qualité;
Qui ?	Personnel participant : disponibilités, compétences, formation requise; utilisateurs des données : besoins.

²⁰ Marine Metadata Interoperability (MMI). <https://marinemetadata.org/>

3.3.3 INITIATIVES DE MONITORAGE ET D'ÉCHANTILLONNAGE

L'échantillonnage des ressources et le monitoring de l'environnement effectués sur une base régulière sont des activités principalement de responsabilité gouvernementale. Plusieurs programmes de monitoring sont en effet menés annuellement par divers ministères. Ceux-ci permettent de compiler des séries temporelles de données utilisées pour évaluer l'état des ressources (flore, faune, populations humaines, écosystèmes) et pour déceler les changements

environnementaux (air, eau, sols, contaminants).²¹ Le Programme de monitoring de la zone Atlantique (PMZA) est un exemple d'initiative de collecte de données océanographiques sur une grande échelle.



PMZA : Stations d'échantillonnage à haute fréquence et lignes de sections sélectionnées du Programme de monitoring de la zone Atlantique. *Source :* Observatoire global du Saint-Laurent - OGSL <http://ogsl.ca/fr/pmza/contexte.html>

De leur côté, les communautés sont de plus en plus intéressées aux enjeux environnementaux et s'impliquent également dans des activités de collecte de données. Les Comités ZIP en sont un bon exemple. Ces organismes régionaux de concertation et d'action ont le mandat de regrouper les principaux usagers du Saint-Laurent, sur leur territoire, et de favoriser leur concertation en vue de résoudre les problèmes locaux et régionaux touchant aux écosystèmes fluviaux et à leurs usages.²²

Au niveau du public en général, les citoyens sont pour leur part appelés à contribuer leurs observations grâce à des initiatives comme par exemple celles du Réseau des observateurs du capelan (ROC)²³ et du Réseau d'observation de mammifères marins – ROMM²⁴. L'OGSL a développé à cet effet une plateforme Web collaborative permettant aux individus intéressés de saisir eux-mêmes leurs données en ligne²⁵.

3.4 QUALITÉ

Qu'il s'agisse d'un avis scientifique, d'une découverte, d'une analyse ou d'un rapport, sa **crédibilité** dépend directement de la qualité des données qui ont servi à sa production.

²¹ Bureau du Vérificateur Général du Canada. 2011. Rapport du commissaire à l'environnement et au développement durable. Ch. 5. Une étude de la surveillance environnementale.

http://www.oag-bvg.gc.ca/internet/Francais/parl_cesd_201112_05_f_36033.html#appa

²² Stratégies Saint-Laurent. Les Comités ZIP. <http://www.strategiessl.qc.ca/les-organismes/les-comites-zip>

²³ Réseau des observateurs du capelan (ROC). <http://ogsl.ca/fr/biodiversite/poissons/mpo-capelan/reseau.html>

²⁴ Réseau d'observation de mammifères marins (ROMM). <http://www.romm.ca/>

²⁵ Observatoire global du Saint-Laurent (OGSL). Données du public. <http://ogsl.ca/fr/donnees-du-public.html>

La gestion de la qualité constitue donc un élément central du processus de gestion des données scientifiques.

3.4.1 AQ/CQ (QA/QC)²⁶

■ ASSURANCE DE LA QUALITÉ (AQ)

L'assurance de la qualité fait référence aux actions qui vont au-delà des tests et des contrôles et qui visent à identifier les problèmes de non-conformité et à les prévenir. AQ inclut également des aspects comme la formation et les audits.

L'acquisition de données en temps réel fait l'objet de procédures particulières puisque les données sont utilisées dès qu'elles sont produites. Il y a donc des risques de mauvaise interprétation de la part d'utilisateurs si ceux-ci ne peuvent distinguer les données inexactes. À titre d'exemple, on peut citer le travail du *Integrated Ocean Observing System - IOOS* qui a élaboré une série de procédures automatisées de contrôle des données océanographiques acquises en temps réel (QARTOD : *Quality Assurance for Real-Time Oceanographic Data*).²⁷



■ CONTRÔLE DE QUALITÉ (CQ)

En général, le contrôle de qualité représente l'ensemble des activités ayant pour but de vérifier la qualité de produits ou de services et ce, en trouvant des problèmes ou des éléments défectueux. En gestion de données, il s'agit d'examiner les jeux de données récoltées afin d'identifier les données erronées, manquantes ou résultant de problèmes dans le processus d'acquisition (Voir : Section 3.4.4 [Qualification des données](#)).

■ GESTION DE LA QUALITÉ

La gestion de la qualité se situe à un niveau plus large c.-à-d. organisationnel ou institutionnel. La Norme ISO 9001 décrit un modèle de gestion de la qualité qui inclut le système et ses procédures, la gestion des responsabilités, des ressources et du service ainsi que la gestion de la mesure, de l'analyse et de l'amélioration de la qualité.

Statistique Canada – une organisation dont la raison d'être est basée sur les données – définit son cadre de gestion de la qualité ainsi que la qualité de l'information de ses produits en fonction de leur adaptation à l'utilisation par les clients. Il s'agit d'un concept multidimensionnel englobant la **pertinence** des données en regard des **besoins de l'utilisateur**, ainsi que des caractéristiques

²⁶ Dans la littérature, on fait souvent référence à l'acronyme anglais QA/QC qui signifie *Quality Control / Quality Assurance*.

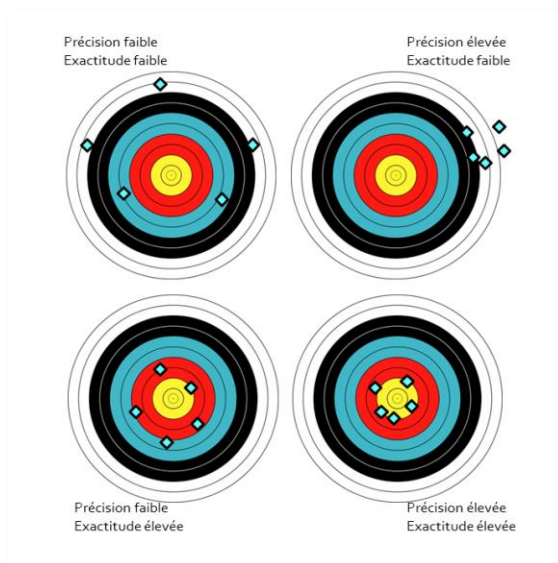
²⁷ Integrated Ocean Observing System (IOOS). 2014. Manual for the Use of Real-Time Oceanographic Data Quality Control Flags. 19 p. http://www.ioos.noaa.gov/qartod/temperature_salinity/qartod_oceanographic_data_quality_manual.pdf

influant sur la façon dont les données peuvent être utilisées, comme **l'exactitude**, **l'actualité**, **l'accessibilité**, **l'intelligibilité** et la **cohérence**.²⁸

3.4.2 PRÉCISION ET EXACTITUDE

Quoique souvent considérés comme des synonymes, ces deux termes ont des définitions bien différentes, telles qu'illustrées dans le schéma ci-contre.

À cet effet, la Norme ISO 5725²⁹, qui porte sur les principes et méthodes de mesure et d'estimation, parle plutôt de « justesse » (*trueness*) d'un test que d'exactitude. La justesse fait ici référence au degré de concordance entre la moyenne arithmétique d'un grand nombre de résultats de tests et la valeur de référence acceptée; la précision porte sur la concordance entre les résultats des tests.



PRÉCISION ET EXACTITUDE : il est possible d'être très précis mais... en même temps, très inexact.

Les erreurs inhérentes à la collecte de données sont généralement définies selon leur source :

- **humaine** : erreur de lecture, de saisie ou de transcription, de jugement;
- **environnementale** : erreur due aux conditions ambiantes pouvant influencer les mesures. Ex. : température de l'air;
- **relative aux équipements** : erreur liée aux limites et aux capacités des appareils de mesure.

BONNES PRATIQUES

POUR LIMITER LES ERREURS :

- S'assurer que le personnel possède la **formation adéquate**;
- Effectuer des tests de contrôle. Ex. : comparaison des mesures et observations par rapport à une **référence**.

De plus, les erreurs peuvent être dites « systémiques » lorsqu'elles présentent une similitude dans leur occurrence ou dans leur amplitude (on peut aussi parler de biais) ou, au contraire, dites « aléatoires » lorsqu'elles ne suivent aucune tendance particulière.

²⁸ Statistique Canada. 2002. Le cadre d'assurance de la qualité de Statistique Canada.

<http://www5.statcan.gc.ca/olc-cel/olc.action?ObjId=12-586-X&ObjType=2&lang=fr&limit=0>

²⁹ ISO 5725-1:1994. Accuracy (Trueness and Precision) of Measurement Methods and Results.

<https://www.iso.org/obp/ui/#iso:std:11833:en>

3.4.3 VALIDATION

La validation est l'étape où les données sont examinées afin de détecter les erreurs résultant de la collecte et de s'assurer que les données recueillies sont complètes, exactes, fiables et prêtes à être utilisées.

Il s'agit d'identifier les données manquantes, invalides ou incohérentes à l'aide de règles qui sont établies selon le contexte et le type de données. Par exemple, une règle pourrait stipuler qu'une donnée acceptable de température de l'eau d'un cours d'eau donné doit se situer entre des valeurs $T_{\min}$ et $T_{\max}$, celles-ci étant le minimum et le maximum observés pour un endroit et une période en particulier.

3.4.4 QUALIFICATION DES DONNÉES

Lors du processus de validation, l'examen des données permet une certaine qualification des données. Ces informations peuvent être documentées et incluses dans la base de données. Ainsi les données sont codifiées selon leur état, ce qui permet par la suite à un utilisateur de comprendre davantage les données et leur pertinence.

À titre d'exemple, l'UNESCO (2010) a établi une série de codes de qualité pour les données de température et salinité de l'eau (*Global Temperature-Salinity Profile Program – GTSP*)³⁰. Le code de qualité est ici un entier compris entre 0 et 9.

Code	Signification
0	aucun contrôle de qualité
1	la donnée semble correcte
2	la donnée semble probablement bonne
3	la donnée semble douteuse
4	la donnée semble erronée
5	la donnée a été modifiée
6 à 8	réserve à un usage futur
9	la donnée est manquante

Il est important de préciser quelles sont les unités associées à chacune des variables mesurées. À titre d'exemple, la section Références du Système de gestion des données environnementales sur le portail de l'OGSL fournit une description des unités utilisées pour un ensemble de variables environnementales allant de la pression atmosphérique à la concentration en oxygène dissous.³¹

³⁰ UNESCO-IOC. 2010. GTSP 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

³¹ Système de gestion des données océanographiques (SGDO). Variables et unités. http://ogsl.ca/app-sgdo/fr/docs_reference/variables-unites.html

3.5 ARCHIVAGE

3.5.1 FORMATS

Le format fait référence à la façon dont les données sont organisées dans un fichier électronique. La sauvegarde des données dans des formats standards et durables est un enjeu majeur en gestion de données. En effet, les données recueillies et préservées adéquatement pourront être consultées à long terme par les utilisateurs, peu importe l'évolution des outils de consultation utilisés.

BONNES
PRATIQUES

POUR L'ARCHIVAGE DES DONNÉES :

- enregistrer les données dans des **formats ouverts non-propriétaires** afin qu'elles puissent être lues à long terme indépendamment des logiciels et de leur évolution;
- établir une **nomenclature claire** pour l'identification des fichiers de données;
- établir une **stratégie d'entreposage** incluant deux différentes formes ex. : sur disque dur et sur CD et deux endroits physiques différents;
- **numériser** la documentation papier et la conserver en format portable (ex. : PDF/A);
- s'assurer que le lieu d'**entreposage** des données (numériques et non-numériques) est **approprié et sécuritaire**;
- copier les données sur un **nouveau support** (ruban magnétique, disque, etc.) de 2 à 5 ans après leur création puisque certains médias optiques et magnétiques sont sujets à la dégradation physique;
- vérifier régulièrement l'**état de l'intégrité** des données enregistrées.

À titre d'exemple, la Norme ISO 19005 propose le PDF/A comme format de conservation à long terme des données³². ISO et IEC (*International Electrotechnical Commission*) approuvent également l'utilisation de formats ouverts comme ODF³³ afin d'assurer l'interopérabilité et l'accès efficace aux données et lui ont conféré la désignation ISO/IEC 26300. Ce format est un standard notamment pour les documents de type « suite Office » incluant texte, chiffrier, présentation et éléments graphiques.

De son côté, la *Library of Congress* du Gouvernement américain, dresse un tableau complet de la démarche adoptée et du processus de décision en ce qui concerne la préservation des données et les formats des fichiers.³⁴

En ce qui concerne les fichiers images et vidéos, Nozères (2011) a effectué un examen approfondi de l'archivage et des bonnes pratiques



³² 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

³³ Open Document Format (ODF). <http://www.opendocumentformat.org/aboutODF>

³⁴ US Government – Library of Congress. Digital Preservation. http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml

en la matière³⁵. Les caractéristiques des divers formats d'imagerie, le géomarquage, les métadonnées et les options de stockage de données y sont présentés de façon détaillée.



NOMENCLATURE DE FICHIERS :

- donner des noms de fichiers **significatifs**
Ex. : **2012-IVerte-Hareng009.xls**
[Par exemple, pour un chiffrier de données biologiques d'un échantillon no. 009 de hareng pris à l'Ile-Verte en 2012]
- éviter d'utiliser des **espaces**, des **accents** et des **caractères spéciaux** tels \$/%& @!
- éviter les noms de fichiers **trop longs**
Ex. : **2012/08/19 pointe Est Ile Verte-échantillons JLambert.xls**

3.5.2 PRÉSERVATION

La préservation des données vise à assurer que les données sont copiées et sécurisées. Évidemment nécessaire mais souvent mise au second plan, elle requiert une planification à court et à long terme.

À court terme, il faut s'assurer que des copies de sécurité sont faites régulièrement, de façon à toujours avoir une version des données disponibles advenant par exemple une corruption de fichiers ou une perte de données. Il est important de vérifier régulièrement l'intégrité des copies de sécurité et de s'assurer qu'elles sont conservées dans un endroit sûr. Il faut, de plus, clairement identifier les données brutes, les données modifiées ou traitées et les différentes versions des fichiers.

À long terme, il faut penser à conserver les données dans un dépôt ou un centre de données. La plupart des institutions possèdent des archives de données et du personnel est dédié à leur entretien et préservation. Un des enjeux de l'archivage est relié à la pérennité des supports physiques c.-à-d. du matériel utilisé incluant les disques durs, les logiciels, les processeurs, etc. Étant donné l'évolution rapide des technologies de l'information et des communications, il faut s'assurer régulièrement de la fonctionnalité de l'infrastructure informatique utilisée et effectuer les migrations de systèmes d'exploitation lorsque nécessaire. Outre les centres de données institutionnels, il existe aussi des dépôts de données organisés autour de communautés d'intérêt. Par exemple, l'infrastructure d'information GBIF pour les données de biodiversité.

Une solution de plus en plus populaire est l'archivage virtuel de type « cloud ». Dans ces environnements virtuels, de l'espace sur des serveurs distants est mis à disposition des utilisateurs et ce, parfois même gratuitement. Dans ces cas, il faut s'assurer non seulement de la fiabilité du service mais également de la sécurité, des contrôles d'accès et du respect de la confidentialité des données qu'on y entrepose.

³⁵ 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>

Pour plus d'information, voir la Section 4. [Pérennité des données](#).

3.5.3 DOCUMENTATION

La documentation relative à un jeu de données contribue à la compréhension du contexte et des méthodes utilisées pour la collecte et l'archivage. Elle explique comment les scientifiques ont organisé, décrit, codé, formaté, calculé et validé ces données. Les métadonnées font également partie intégrante de la documentation des jeux de données.

**BONNES
PRATIQUES**

UNE BONNE DOCUMENTATION COMPREND :

- le **contexte**, l'historique, les objectifs et les hypothèses du projet de collecte de données;
- les **méthodes** utilisées : collecte, analyse, instrumentation, logiciels, calibration;
- la **structure** du (des) jeu(x) de données;
- les **procédures** de contrôle, de vérification, de validation des données;
- les **changements** apportés aux données et l'identification des différentes versions des fichiers;
- les **conditions** d'utilisation et les aspects de confidentialité associés aux données;
- les **métadonnées**.

3.6 ACCÈS

Alors qu'on s'accorde à dire que l'accès aux données doit être facilité, il semble qu'il ne soit pas si aisé de partager librement ses données. Plusieurs enjeux, dont la propriété intellectuelle et la reconnaissance, entrent en ligne de compte. De l'avis de plusieurs, il faut que s'opère un changement culturel dans la communauté scientifique afin, non seulement que les publications soient valorisées mais, que le soient autant les données qui sont la **base** de ces publications.

“Nous devons changer la culture scientifique en une culture qui valorise les données autant que les publications.”

3.6.1 PROPRIÉTÉ INTELLECTUELLE (COPYRIGHT)

Il est important que la propriété des données soit clairement déterminée, que ce soit celle d'un scientifique ou de l'organisation à laquelle il appartient. C'est en effet un élément de premier plan, particulièrement lors de la définition des conditions d'accès aux données.

Les situations diffèrent selon qu'il s'agit d'une institution de recherche, d'un ministère ou d'un organisme sans but lucratif (OSBL). Généralement, l'état est propriétaire des données produites par les scientifiques gouvernementaux. Dans les universités, parfois ce sont les chercheurs qui détiennent les

[Nelson, B. 2009. Empty Archives. NATURE, Vol. 461, p. 160-163.](#)

droits sur leurs données, parfois, c'est l'université qui possède les données.

De plus en plus, on observe des alliances et des partenariats entre diverses institutions, incluant la participation d'OSBL. La clarté en matière de reconnaissance des droits de propriété intellectuelle devient donc primordiale.

Lectures suggérées :

- 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
- Conseil National de Recherches Canada (CNRC). 2014. Politique de gestion de la propriété intellectuelle du CNRC.
http://www.nrc-cnrc.gc.ca/fra/apropos/proprieete_intellectuelle/index.html
- Penn State University. Acquisition, Management, Sharing and Ownership of Data.
<http://www.research.psu.edu/training/sari/teaching-support/data-management>
- Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG). 2009. Politique sur la propriété intellectuelle.
http://www.nserc-crsng.gc.ca/NSERC-CRSNG/Politiques-Politiques/ip-pi_fra.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 ACCÈS CONTRÔLÉ/RESTREINT

Qui pourra trouver, consulter, utiliser, analyser, transformer, diffuser, partager les données? S'il ne s'agit pas d'accès libre à des données ouvertes, toutes ces questions sont des plus pertinentes.

Il est d'abord important de déterminer qui possèdera les privilèges d'accès aux données à toutes les étapes de leur production et de mettre des balises claires en matière de droit d'édition et de modification des données.

Une fois les données prêtes à être publiées, on doit décider des conditions de leur utilisation. Il est possible de limiter celle-ci à un auditoire en particulier, à une équipe de recherche ou à une communauté de pratique, spécialement lorsqu'il s'agit de données sensibles ou confidentielles. Il est alors approprié d'exiger l'authentification de l'utilisateur à l'aide d'un compte avec mot de passe afin de pouvoir identifier les consommateurs des données. Quoiqu'il en soit, dès que l'on décide de contrôler ou de restreindre l'utilisation des données, il est nécessaire d'être en mesure d'effectuer une gestion efficace des accès.

3.6.3 DÉCOUVERTE

Tel que mentionné précédemment (Voir Section 2. [Définitions](#)), les métadonnées sont des données à propos des données. Celles-ci servent à documenter les données et aident à découvrir, à faire connaître et à promouvoir leur existence. Plus de détails sont présentés à la Section 5. [Métadonnées](#).

Dans son document sur la gestion des données d'images en sciences aquatiques, Nozères (2011) propose les champs de métadonnées minimales pour les images et les vidéos :

- Nom de fichier (fortement recommandé)
- Date de création (automatique, par défaut)
- Créateur - Auteur (fortement recommandé)
- Mots-clés (fortement recommandé)
- Statut de droit d'auteur (facultatif)
- Titre (facultatif)
- Description - Légende (facultative)
- Emplacement - Sous-emplacement (facultatif)
- GPS (latitude, longitude, altitude)

3.6.4 DIFFUSION DE L'EXISTENCE DES DONNÉES - REGISTRES / CATALOGUES

Les services de catalogues permettent la publication et la recherche de collections d'information descriptive (métadonnées) à propos de données et de services. Ils servent non seulement à trouver des données existantes mais à faire connaître l'existence de ses propres données. Ils ont, de plus, la capacité de moissonner (*harvesting*) des métadonnées dans d'autres catalogues. L'*Open Geospatial Consortium (OGC)* fait une distinction entre « catalogue » et « registre » en précisant qu'un registre est un catalogue spécialisé maintenu par une entité officielle se conformant aux procédures et aux politiques d'accès et de gestion du contenu (normes ISO 19135, ISO 11179-6).

DÉCOUVERTE

++++

Publier l'existence des données augmente :

- la visibilité du projet;
- le nombre de citations;
- la réutilisation des données;
- les possibilités de collaboration;
- la contribution à d'autres recherches;
- la valorisation des données.

Pour plus d'information, voir la Section 6. [Partage des données](#).

3.7 UTILISATION

3.7.1 LICENCES ET DROITS

Tout comme pour l'accès, lorsqu'il est question de l'utilisation des données, il faut en déterminer clairement les conditions et les limites. La licence d'utilisation est un outil qui sert à informer les utilisateurs sur la façon dont le propriétaire des données souhaite en encadrer les usages. Une myriade de types de licences existe, ce qui rend la situation complexe autant pour le producteur que pour l'utilisateur de données³⁶. Il y a cependant une tendance vers la simplification et l'accès ouvert aux données tels que tente de le faire GEOSS pour les données d'observation de la terre³⁷.

À titre indicatif, une licence peut inclure les éléments suivants :

- Donner à l'utilisateur la possibilité de copier et de distribuer les données, en indiquant par exemple que :
« Quiconque peut, sans autorisation ni frais, reproduire, publier, adapter, traduire, communiquer par quelque moyen que ce soit les documents, données et informations disponibles sur ce site »
- Exiger que le crédit à l'auteur ou au propriétaire des données soit mentionné par l'utilisateur selon un format spécifique :
« ... à la condition de mentionner la source de la façon suivante :
SOURCE : PRODUCTEUR XYZ, HTTP://XYZ.CA, (ANNÉE). »
- Permettre (ou interdire) l'utilisation commerciale des données :
« ... que ce soit pour des fins personnelles ou commerciales... ».

3.7.2 INTÉGRATION DES DONNÉES ET DES INFORMATIONS

Une des conséquences de l'accroissement de l'accès et du partage des données est sans contredit la problématique de l'intégration³⁸ des données provenant de sources multiples. Idéalement, tous les producteurs de données se conformeraient à des normes internationalement reconnues et adopteraient des environnements informatiques interopérables et accessibles mais, qu'en est-il vraiment?

HÉTÉROGÉNÉITÉ

Prenant un exemple du quotidien, il n'y a qu'à penser à la façon dont les gens notent la date pour comprendre que l'hétérogénéité des procédures, des formats, de la syntaxe et des systèmes constitue un des enjeux majeurs pour l'intégration de données. En effet, bien que la norme internationale ISO 8601 dicte l'emploi du format AAAA-MM-JJ,

³⁶ 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/o/advpub_14-043/_pdf

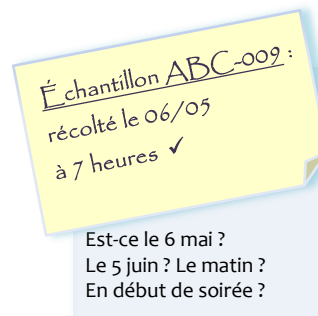
³⁷ Onsrud, H., J. Campbell and B. van Loenen. 2010. Towards Voluntary Interoperable Open Access Licenses for the Global Earth Observation System of Systems (GEOSS). International Journal of Spatial Data Infrastructures Research, 2010, Vol. 5, 194-215. <http://ijdir.jrc.ec.europa.eu/index.php/ijdir/article/view/168/203>

³⁸ 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>

on peut retrouver des dizaines de formes pour la seule date du 2015-02-12 tel qu'illustré dans le tableau ci-dessous :

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.

De façon globale, on peut voir que, si on décide de combiner des jeux de données où les diverses variables sont représentées dans des formats différents, les résultats de l'intégration seront affectés. Il en est de même pour les unités choisies, la précision des mesures effectuées ou les types de projections cartographiques utilisées. La rigueur et la cohérence sont évidemment de mise.



ASSIMILATION DANS DES MODÈLES

Un des exemples les plus connus en matière de modélisation est le travail qu'effectuent les météorologistes. En effet, ceux-ci utilisent une variété de paramètres environnementaux qui alimentent leurs modèles climatiques de façon à pouvoir produire les prévisions météorologiques les meilleures possibles. L'assimilation des données est un processus cyclique qui représente l'inclusion de données d'observation *in situ* dans les modèles de façon à raffiner les prévisions. C'est de cette façon que, par exemple, le modèle couplé eau-atmosphère développé par Saucier et al.^{39 40} est utilisé entre autres pour la production des prévisions de courants de surface dans l'estuaire et le golfe du Saint-Laurent.⁴¹

3.7.3 VISUALISATION ET ANALYSE

Les outils disponibles sont très nombreux et permettent aux producteurs de données de littéralement voir leurs données afin de les comprendre et de les analyser. Que ce soit une carte, un graphique ou un tableau, les outils de visualisation offrent la possibilité de combiner des informations diverses dans une même vue d'ensemble, ce qui aide à identifier des tendances ou des anomalies.

³⁹ Saucier, F.J., F. Roy, S. Senneville, G. Smith, D. Lefavre, 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

⁴⁰ Gouvernement du Canada. Environnement Canada, Modélisation. https://meteo.gc.ca/model_forecast/model_f.html

⁴¹ Observatoire global du Saint-Laurent (OGSL). Prévisions océaniques. <http://ogsl.ca/ocean>

Prévisions océaniques

46° 44' 45.06" N - 71° 17' 12.42" O

Choisir une période

Dates disponibles : 2015-03-02 - 2015-04-01

Début 2015-03-02

Fin 2015-03-03

Afficher

Niveau d'eau

lat: 46.7458 | long: -71.2868 | généré le: 2015-03-02 07:43:05 HNE



VISUALISATION D'UNE SÉRIE TEMPORELLE de prévisions de niveaux d'eau dans l'estuaire du Saint-Laurent permettant de se déplacer dans les données selon la période sélectionnée. *Source* : Observatoire global du Saint-Laurent (OGSL). <http://ogsl.ca/ocean>

Les outils de visualisation permettent parfois de raffiner le processus de validation des données. Par exemple, en illustrant de façon cartographique la distribution des données observées, il sera possible de voir si des positions géographiques sont erronées c.-à-d. placées à des endroits non échantillonnés ou impossibles.

Biodiversité

Source de données

Poissons • MFFP - Réseau de suivi ichtyologique (PASL) - Poissons d'eau douce

En savoir plus...

Type de données

Présence / Absence

Secteur

Archipel du Lac Saint-Pierre

Engin

Filet maillant

Année

2010

Espèce

Bar rayé (Morone saxatilis)



Information

Cliquez sur un point de la carte pour obtenir plus d'information

Légende

● Présence
○ Absence

Données cartographiques 2 km L Conditions d'utilisation Signaler une erreur cartographique

Télécharger

VISUALISATION DE LA DISTRIBUTION de stations d'échantillonnage de bar rayé avec indication de la présence ou de l'absence de l'espèce. L'outil permet de se déplacer dans la base de données et de choisir des secteurs, des engins de pêche, des années et des espèces différentes. *Source* : Observatoire global du Saint-Laurent (OGSL). <http://ogsl.ca/bio>

3.7.4 PRODUITS ET SERVICES POUR LA PRISE DE DÉCISION

La prise de décision est un processus qui implique d'analyser une situation, d'examiner les enjeux et les risques, de peser les conséquences, de considérer les incertitudes et les intervenants. La science étant souvent à la base de décisions, elle se doit d'être accessible, de qualité et compréhensible.

L'exemple du projet Old Harry, dans le golfe du Saint-Laurent, illustre bien le rôle de la science dans le contexte d'une région soumise à une forte pression pour de l'exploration gazière et pétrolifère. Les travaux de Bourgault et al (2014)⁴² ont permis, d'une part, de montrer le manque de considérations scientifiques dans les processus décisionnels et dans les politiques encadrant une telle industrie. D'autre part, la modélisation de la circulation des masses d'eau a permis d'identifier les zones les plus sensibles dans l'éventualité de déversements de pétrole. L'ensemble de ces résultats, et de ceux de travaux multidisciplinaires subséquents, servira de base de connaissances qui pourra éclairer les prises de décisions dans le cadre de la mise en œuvre de travaux d'exploitation pétrolifère dans cette région.

3.7.5 RÉTROACTION DES UTILISATEURS

Pour illustrer l'importance de la gestion de données, il n'y a qu'à regarder la haute considération des États-Unis envers la science des données. En effet, le 18 février 2015, le gouvernement américain s'est doté d'un « champion » des données, Dr D.J. Patil, le premier « *US Chief Data Scientist* » à la Maison Blanche. En créant ce poste clé au sein de son administration, l'état démontre ainsi qu'il donne une place prépondérante aux données et à leur utilisation afin de redonner des bénéfices aux citoyens.⁴³ À cet effet, Patil considère la rétroaction des utilisateurs comme étant la plus importante composante des processus de développement de la science des données.⁴⁴ Il s'agit toutefois d'un aspect souvent négligé qui a été inclus ici à l'étape finale, « Utilisation », du cycle de vie des données tel que présenté au début de la [Section 3](#). Cependant, en plus de fermer la boucle du cycle, la rétroaction des utilisateurs de données pourrait en fait être un intrant à toutes les étapes de la gestion des données. Les commentaires et les suggestions des clientèles utilisatrices deviennent matière à amélioration et à ajustement des processus et procurent des avantages aux producteurs de données qui prennent le temps de les considérer.

⁴² 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>

⁴³ 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>

⁴⁴ 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. Pérennité des données

Tel qu'évoqué précédemment, il est nécessaire d'assurer un accès efficace et durable aux données en mettant en place des procédures et des mécanismes d'archivage et de conservation des données. L'importance de ces actions est très éloquemment exprimée dans un document des Archives de France sur la conservation à long terme des documents scientifiques et techniques : « ... la perte des données créées à grands frais et avec les avancées technologiques les plus à la pointe du progrès serait une sorte de régression, une perte irrémédiable de ce capital scientifique, lui-même à l'origine de nouvelles découvertes. »⁴⁵

En fait, à une époque où les produits de la science sont de plus en plus numériques, la gestion de l'ensemble des étapes du cycle de vie des données devient nécessaire plutôt que simplement souhaitable. La demande pour ces données est conséquemment en croissance et requiert la capacité d'y répondre efficacement.

4.1 MISE EN ŒUVRE DE LA GESTION DES DONNÉES

La bonne gestion et le partage des données procurent de nombreux avantages aux producteurs et aux utilisateurs de ces données et permettent notamment :

- l'accélération du progrès scientifique;
- l'accroissement de la visibilité et des retombées de la recherche;
- le respect des politiques des organismes subventionnaires;
- l'évitement de duplication des travaux de recherche;
- la reproduction et la validation des résultats de la recherche;
- l'intensification de la collaboration.

4.2 DÉFIS ET ENJEUX

À l'échelle nationale, beaucoup d'efforts sont encore requis pour qu'il soit possible d'affirmer que l'ensemble des données est bien géré. C'est le constat du Groupe de travail sur la stratégie des données de recherche de l'Institut canadien de l'information scientifique et technique qui a mis en lumière en 2008 d'importantes lacunes en matière de gestion des données de recherche au Canada⁴⁶. L'examen d'indicateurs clé a permis de mesurer les écarts par rapport à un état dit « idéal ». Ces constats demeurent aujourd'hui des enjeux et des défis auxquels les institutions font toujours face, soient :

- **Politiques** : peu d'organisations ont des politiques claires de gestion des données;
- **Financement** : la gestion des données est rarement financée à long terme c.-à-d. au-delà de la durée de vie des projets de recherche;

⁴⁵ 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>

⁴⁶ Institut canadien de l'information scientifique et technique - Groupe de travail sur la stratégie des données de recherche. 2008. Gérance des données de recherche au Canada : une analyse des écarts. 36 p. http://publications.gc.ca/collections/collection_2009/cnrc-nrc/NR16-123-2008F.pdf

- **Rôles et responsabilités** : les responsabilités des intervenants au long du cycle de vie des données ne sont pas clairement définies et certains rôles ne sont pas attribués;
- **Dépôts de données** [numériques fiables] : il existe peu de dépôts de données pour tous les domaines de recherche, la grande majorité des données de recherche sont encore hébergées sur les disques durs des chercheurs;
- **Normes** : les normes pour assurer l'interopérabilité et celles relatives aux métadonnées ne sont pas toujours connues et/ou respectées;
- **Compétences et formation** : on retrouve peu de personnel ayant les compétences et étant dédié à la gestion de données;
- **Récompense et reconnaissance** : peu de mécanismes de récompense et de reconnaissance pour le partage des données de recherche;
- **Recherche et développement** : les activités de R-D au Canada ne sont pas coordonnées et orientées par des priorités nationales;
- **Accessibilité** : les chercheurs sont réticents à partager leurs données; certaines politiques sont contradictoires (« respect de la vie privée et éthique » et « accès et conservation »);
- **Conservation** : peu d'engagement des établissements envers la conservation à long terme, les organismes de recherche n'exigent que rarement des plans de gestion des données.

À l'heure actuelle, les principaux organismes fédéraux accordant des subventions de recherche tels que le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG)⁴⁷, le Conseil de recherches en sciences humaines du Canada (CRSH)⁴⁸ et les Instituts de recherche en santé du Canada (IRSC)⁴⁹, ont élaboré des politiques et/ou suivent les principes directeurs sur l'accès aux données de recherche énoncés par le Gouvernement du Canada dans le domaine des sciences⁵⁰. Ces principes visent notamment à :

- faire avancer les connaissances,
- réduire le dédoublement de la recherche,
- maximiser les retombées de la recherche et
- faire connaître les réalisations des chercheurs.

4.3 POLITIQUES DE GESTION DE DONNÉES

Au niveau organisationnel, on retrouve des exemples de politiques de gestion des données complètes comme celle développée et mise en application par le Secteur des Sciences du Ministère des Pêches et des Océans du Canada (MPO).⁵¹ L'objectif principal de cette politique est d'assurer la protection des bases de données scientifiques et de

⁴⁷ Programme de subventions de réseaux stratégiques. Conseil de recherches en sciences naturelles et en génie du Canada.

http://www.nserc-crsng.gc.ca/Professors-Professeurs/RPP-PP/SNG-SRS_fra.asp

⁴⁸ Conseil de recherches en sciences humaines du Canada (CRSH). Politique sur l'archivage des données de recherche.

http://www.sshrc-crsh.gc.ca/about-au_sujet/policies-politiques/statements-enonces/edata-donnees_electroniques-fra.aspx?

⁴⁹ Instituts de recherche en santé du Canada (IRSC). Politique de libre accès des IRSC. <http://www.cihr-irsc.gc.ca/f/32005.html>

⁵⁰ Gouvernement du Canada – Sciences. Accès aux résultats de la recherche : principes directeurs.

<http://www.science.gc.ca/default.asp?Lang=Fr&n=9990CB6B-1>

⁵¹ Ministère des Pêches et des Océans du Canada (MPO). Politique sur les données scientifiques.

<http://www.dfo-mpo.gc.ca/science/data-donnees/policy-politique-fra.htm>

maximiser leur utilité grâce à l'interconnectivité, l'accessibilité et le respect des normes. Le MPO vise ainsi à obtenir les résultats suivants :

- des données sécurisées maintenant et pour le futur,
- des données interconnectées et interoperables,
- des données utiles découvrables et accessibles de façon standardisée, et
- une gestion rentable des données.

À titre d'autre exemple, on retrouve la politique de gestion des données pour l'Année polaire internationale (API) 2007-2008, un programme international de recherche et d'observation interdisciplinaire visant à mieux comprendre les processus polaires et leurs liens avec la planète. « L'objectif premier de la gestion des données de l'API 2007-2008 consiste à garantir la sécurité, l'accessibilité et la gratuité de l'échange des données pertinentes qui facilitent la recherche actuelle et qui demeureront accessibles aux générations à venir ».⁵²

De façon générale, les principes suivants sont sous-jacents à l'élaboration de politiques de gestion des données scientifiques :

- la **valeur** inestimable et le caractère irremplaçable des données scientifiques engendrent la nécessité de les gérer et d'en assurer la **conservation** et la **pérennité**;
- l'utilisation des données et leur pertinence impliquent que celles-ci soient librement **accessibles**, **compréhensibles** et de **qualité**, ce qui nécessite une gestion de l'ensemble du cycle de vie des données, de l'acquisition jusqu'à la dissémination;
- la préservation à long terme des données requiert des archives et des dépôts de données **structurés** et **sécurisés**;
- les données récentes étant les plus critiques pour la prise de décision, elles doivent être accessibles en **temps opportun**;
- le **libre accès** aux données est privilégié sous réserve d'enjeux de confidentialité et/ou de sensibilité;
- l'**échange** et le **partage** de données avec la communauté internationale sont encouragés afin de favoriser la mise en commun des connaissances et l'avancement de la science.

⁵² Année Polaire Internationale 2007-2008 (API). http://www.api-ipy.gc.ca/pg_IPYAPI_055-fra.html

5. Métadonnées

« Un enregistrement de métadonnées est un fichier d'information qui comporte les caractéristiques élémentaires d'une ressource d'information ou de données géographiques. Il représente le **qui**, le **quoi**, le **quand**, le **où**, le **pourquoi** et le **comment** de la ressource de géodonnées. Les métadonnées permettent à un producteur de décrire entièrement un ensemble de données pour qu'on puisse en comprendre le contexte et les limites, et d'évaluer l'applicabilité de l'ensemble de données par rapport à son utilisation prévue. »⁵³ De plus, les métadonnées sont utilisées dans les catalogues et facilitent la découverte et la réutilisation des données au bénéfice de la communauté scientifique et des utilisateurs.

5.1 NORMES

La norme internationale en matière de métadonnées est la norme ISO 19115⁵⁴. Elle compte plusieurs centaines de champs décrivant les caractéristiques des jeux de données. La norme du *Federal Geographic Data Committee (FGDC)*⁵⁵ a été pendant longtemps la norme en vigueur pour plusieurs organisations. Cependant, le FGDC a décidé de suivre la tendance internationale en adoptant et en facilitant la transition vers la norme commune ISO.

PROFILS

Un profil est un sous-ensemble de la norme qui regroupe un sous-ensemble de métadonnées pertinentes dans un contexte spécifique.

Le **Profil nord-américain (PNA)** représente une norme nationale canadienne qui comble les besoins des producteurs et des utilisateurs de données géoréférencées au Canada et aux USA. « L'objectif de ce Profil est de fournir aux organismes qui produisent des données géographiques, un mécanisme pour décrire en profondeur les ensembles de données. (...) De plus, il facilite la recherche d'information propre aux besoins de l'utilisateur. »⁵⁶

Le « **Marine Community Profile of ISO 19115** » est un profil développé par l'*Australian Ocean Data Centre Joint Facility (AODCJF)* suivant les normes internationales. Il est également un sous-ensemble de la Norme ISO 19115 qui inclut toutes les métadonnées de base et qui, en plus, définit des éléments additionnels, une codification et un vocabulaire qui facilitent la description des données du milieu marin.⁵⁷

De façon générale, les profils définissent un ensemble d'éléments obligatoires, conditionnels et optionnels qui permettent de décrire un jeu de données c.-à-d. type de données, sujet, zone géographique, limites temporelles, langue, contact.

⁵³ Ressources naturelles Canada - Métadonnées géospatiales numériques. <http://www.rncan.gc.ca/8913>

⁵⁴ Norme ISO 19115. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=53798

⁵⁵ Federal Geographic Data Committee (FGDC). <http://www.fgdc.gov/metadata/geospatial-metadata-standards>

⁵⁶ Profil Nord-Américain (PNA) de la norme ISO 19115:2003 — Information géographique — Métadonnées.

<http://nap.geogratis.gc.ca/metadata/napMetadata-fra.html>

⁵⁷ Marine Community Profile (MCP) of ISO 19115. <https://marinemetadata.org/references/marineprofile19115>

5.2 PUBLICATION ET RECHERCHE DE MÉTADONNÉES

Plusieurs initiatives permettent aux producteurs de données à la fois de créer leurs métadonnées, de les publier et d'effectuer des recherches dans les registres disponibles. Quelques exemples sont présentés ci-dessous :

Infrastructure canadienne de données géospatiales (ICDG) – Portail de découverte⁵⁸

Le Portail de découverte de GéoConnexions est un catalogue de métadonnées qui permet aux utilisateurs, aux développeurs et aux fournisseurs de données de trouver, d'évaluer, de visualiser et de publier des produits et des services Web de données géospatiales et géoscientifiques canadiennes.

Infrastructure for Spatial Information in the European Community (INSPIRE)⁵⁹

L'éditeur de métadonnées de l'infrastructure d'information spatiale de la Communauté Européenne - INSPIRE offre une interface conviviale disponible en 22 langues. Un guide d'utilisation⁶⁰ (en anglais) explique la façon de saisir et d'enregistrer des métadonnées.

Polar Data Catalog (PDC)⁶¹

Le *Polar Data Catalogue* est une base de données et de métadonnées générées par des chercheurs en Arctique et en Antarctique selon les normes internationales. L'outil de recherche public permet l'accès à des données de diverses disciplines dont les sciences environnementales, sociales et de la santé.

BlueNet Metadata Entry and Search Tool (MEST)⁶²

L'outil de création et de recherche de métadonnées de l'*Australian Marine Science Data Network* soutient l'archivage et l'accès aux données de recherche scientifique.

Open Archives Initiative – (OAI)⁶³

L'OAI développe et fait la promotion des normes d'interopérabilité en soutien à la diffusion efficace du contenu Web et à l'accessibilité des ressources numériques particulièrement par les réseaux éducatifs (*eScholarship*, *eLearning*, *eScience*).

GeoNetwork OpenSource⁶⁴

GeoNetwork OpenSource est une application de type catalogue, téléchargeable et du domaine du logiciel libre. L'application, disponible en plusieurs langues, permet de chercher et de générer des métadonnées et offre un manuel de l'utilisateur.

GeoDoc⁶⁵

L'éditeur de métadonnées *GeoDoc* permet de créer, de valider, d'éditer, d'importer et d'exporter des métadonnées dans plusieurs formats selon plusieurs normes en vigueur.

⁵⁸ Infrastructure canadienne de données géospatiales (ICDG) – Portail de découverte. <http://geodiscover.cgdi.ca/web/guest/home>

⁵⁹ Infrastructure for Spatial Information in the European Community (INSPIRE). Metadata Editor. <http://inspire-geoportal.ec.europa.eu/editor/>

⁶⁰ Infrastructure for Spatial Information in the European Community (INSPIRE). Metadata Editor User Guide. http://www.eurogeoss.eu/Documents/EuroGEOSS_D_2_2_3.pdf

⁶¹ Polar Data Catalogue (PDC). <https://www.polardata.ca/>

⁶² Australian Ocean Data Network (AODC) BlueNet MEST. <http://bluenet.aodn.org.au/geonetwork/srv/en/main.home>

⁶³ Open Archives Initiative (OAI). <http://www.openarchives.org>

⁶⁴ GeoNetwork OpenSource. <http://geonetwork-opensource.org>

⁶⁵ GeoDoc. <http://www.gogeo.ac.uk/gogeo/metadata/geodoc.htm>

6. Valorisation des données par leur diffusion

L'utilisation des nouvelles technologies de l'information favorise l'accessibilité, la réutilisation et la valorisation des données, crée de nombreuses opportunités de collaboration, soutient l'innovation et facilite la réalisation de nouvelles recherches. La diffusion et le partage des données scientifiques soutiennent la valorisation et la démocratisation des données et font rayonner les travaux de recherche en science. Ces valeurs animent de plus en plus de gouvernements, d'organisations et de scientifiques à travers le monde et stimulent l'adoption de meilleures pratiques de gestion de l'ensemble du cycle de vie des données.

6.1 PRINCIPES

Le **Gouvernement du Canada** a établi une série de principes en matière de données ouvertes⁶⁶. Ces principes généraux peuvent aisément s'appliquer à tout jeu de données :

- 1 **Exhaustivité** : les jeux de données devraient être aussi exhaustifs que possible et contenir l'ensemble de l'information colligée sur un sujet donné incluant les métadonnées qui définissent et expliquent les données brutes, de même que des explications sur les méthodes de calcul des données.
- 2 **Primauté** : les données devraient provenir d'une source primaire, comprenant les renseignements originaux et toutes précisions disponibles sur la manière dont les données ont été recueillies afin de permettre aux utilisateurs de vérifier que l'information a été recueillie adéquatement et consignée avec exactitude.
- 3 **Rapidité** : les données devraient être mises à la disposition des utilisateurs dans les meilleurs délais.
- 4 **Facilité d'accès physique et électronique** : les jeux de données devraient être aussi facilement accessibles que possible en évitant de compliquer les conditions d'accès ou en exigeant des technologies ou applications trop contraignantes.
- 5 **Lisibilité par machine** : les jeux de données devraient être conservés dans des formats de fichiers largement utilisés qui se prêtent aisément à un traitement par machine (ex. CSV, XML). Ces fichiers devraient être accompagnés de la documentation relative au format et à son mode d'utilisation en lien avec les données.
- 6 **Non-discrimination** : les jeux de données devraient présenter le moins possible d'obstacles à l'utilisation par toute personne, à tout moment, sans qu'elle ait à s'identifier ou à fournir une justification.
- 7 **Utilisation de formats communs** : dans la mesure du possible, les jeux de données devraient être conservés dans des formats de fichiers accessibles et non propriétaires afin d'éviter que l'utilisateur doive se procurer une application particulière pour lire les données.
- 8 **Licence d'utilisation** : une licence ouverte est conçue de manière à limiter au minimum les restrictions à l'utilisation des données.

⁶⁶ Gouvernement du Canada – Données ouvertes. <http://ouvert.canada.ca/fr/labc-des-donnees-ouvertes#toc6>

9 Permanence : pour être utilisée de manière optimale, l'information publiée en ligne devrait rester en ligne avec un suivi des différentes versions et un archivage adéquats au fil du temps.

10 Frais d'utilisation : les données ouvertes sont publiées gratuitement.

Le **Gouvernement américain** abonde dans le même sens avec le *US White House Project Open Data*⁶⁷ en affirmant que ses données seront documentées, complètes, réutilisables, publiques et disponibles en temps opportun. Un plan d'action pour les données ouvertes gouvernementales a été publié à cet effet en 2014.⁶⁸

À l'échelle mondiale, le *Global Earth Observation System of Systems (GEOSS)* base également son approche sur des principes visant le partage de données sans lequel les bénéfices de société visés ne peuvent être réalisables⁶⁹. L'accessibilité complète, ouverte et en temps opportun des métadonnées et des données est l'un des fondements de cette stratégie. À titre d'exemple, on peut citer le programme *International Oceanographic Data and Information Exchange (IODE)* du *Intergovernmental Oceanographic Commission (IOC)* de l'UNESCO qui vise à faciliter l'échange de données et d'informations océanographiques.⁷⁰

Devant la multiplicité des initiatives à travers le monde, un groupe de travail du GEO a examiné la situation qui prévaut chez les principaux acteurs mondiaux en matière de production et de gestion de données. Le groupe a soulevé particulièrement le besoin d'harmonisation des diverses approches, des terminologies et des définitions.⁷¹

6.2 AVANTAGES

Une étude de RIN/NESTA⁷² portant sur les pratiques de groupes de chercheurs au Royaume-Uni a montré que ceux ayant adopté une approche ouverte bénéficiaient d'avantages importants par rapport aux autres travaillant de façon plus fermée. Ces avantages incluent :

- Augmentation de l'**efficacité** soit la réduction de la duplication d'efforts, le partage de bonnes pratiques et d'outils et la diminution des coûts de collecte de données;

⁶⁷ US White House Project Open Data. <https://project-open-data.cio.gov/>

⁶⁸ 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

⁶⁹ Global Earth Observation System of Systems (GEOSS). <http://www.earthobservations.org/geoss.php>

⁷⁰ International Oceanographic Data Exchange (IODE). <http://www.iode.org/>

⁷¹ 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

⁷² 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>

- **Promotion** de la **rigueur** et de la **qualité** de la recherche auprès des pairs en faisant connaître les méthodes et protocoles utilisés ainsi que les résultats (incluant ceux négatifs);
- Augmentation de la **visibilité** de la recherche en fournissant des **opportunités de collaboration** avec la communauté scientifique et d'implication auprès des collectivités;
- Élargissement de la **portée** de la recherche en facilitant la réutilisation des données et les échanges et le partage d'expertises entre chercheurs et entre institutions;
- Augmentation de **l'impact socio-économique** et des **opportunités d'innovation** en facilitant les collaborations au-delà de la communauté scientifique.

Plusieurs études démontrent une corrélation entre le fait de partager les données de recherche et le nombre de fois que ces recherches sont citées.⁷³ Certains parlent même d'augmentation de l'ordre de 69% du nombre de citations pour les cas où les données sont ouvertes.⁷⁴ Une telle reconnaissance des résultats de leur recherche représente un avantage tangible pour les chercheurs.

Bref, les principes d'ouverture, de partage et de valorisation des données, représentent la base d'une approche que la communauté scientifique s'approprie de plus en plus. Cette approche s'observe à l'intérieur de projets collaboratifs interdisciplinaires, on la reconnaît dans la volonté d'inclure la participation citoyenne et elle est source d'inspiration pour le développement de nombreux outils d'accès, de visualisation et de diffusion des données.

7. Outils

Outre les initiatives et les nombreuses références présentées tout au long de ce document, les outils suivants sont proposés :

PLANIFICATION

Le cycle de vie des données décrit précédemment implique nécessairement de nombreuses procédures et démarches spécifiques à chacune de ses phases. L'ensemble de celles-ci constitue le **plan de gestion des données**.

À titre d'outil de planification et également d'aide-mémoire, l'**Annexe 1** présente une liste de contrôle des éléments d'un plan de gestion des données. Cette liste permet de garder en

⁷³ 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>

⁷⁴ 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>

tête l'ensemble du processus de gestion des données et de suivre le développement de chaque élément du plan, tel que présenté de façon détaillée au long du présent document.

DIFFUSION ET PARTAGE

De façon générale, les systèmes d'observations des océans tel « *l'Integrated Ocean observing system* » aux États-Unis, sont des organismes possédant des plateformes Web comme outil de diffusion de données. Au Québec, un grand nombre d'organismes producteur de données bénéficient du portail Web de l'Observatoire global du Saint-Laurent comme outil de diffusion de données reconnu.

Comme autre outil de partage, on peut citer le système d'information PANGAEA⁷⁵ qui regroupe un ensemble de données géoscientifiques et environnementales. Il permet de rechercher mais également de déposer des jeux de données et d'obtenir un identifiant unique (DOI – *Digital Object Identifier*) qui permet le référencement des données et les crédits aux auteurs.

SYSTÈME DE GESTION DES DONNÉES

Lors de Rencontres annuelles d'orientation des membres (ROM) ainsi qu'au cours des réunions régulières du Comité aviseur scientifique (CAS) de l'OGSL, les besoins identifiés par les universités, les ONG et les ministères partenaires ont fait l'objet de plusieurs discussions. Un des principaux enjeux identifiés est la nécessité d'accroître la capacité interne de gestion des données des organisations. Une des solutions proposées est l'implantation chez les organismes partenaires de l'OGSL d'un système de gestion de données environnementales (SGDE) disponible en logiciel libre. Bien qu'elle nécessite beaucoup d'efforts de développement, cette solution présente un coût relativement faible d'installation pour les organismes hôtes et peut servir d'incitatif à poursuivre à l'interne, de façon durable, les activités et les bonnes pratiques de gestion de données. En collaboration avec les experts du MPO et de l'UQAR-ISMER, l'OGSL a développé l'application Web SGDE qui devrait être disponible au cours de 2015.

DOCUMENTATION DES MÉTADONNÉES

De son côté, le MPO a également développé comme solution nationale une application nommée *JMetaWriter* qui a notamment été traduite en français par la région du Québec. Celle-ci peut être déployée chez les partenaires de l'OGSL et permet de faciliter la documentation de leurs jeux de données et, par conséquent, d'alimenter les infrastructures de données internationales.

⁷⁵ 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>

8. Références

Bien que non exhaustive, la liste ci-dessous propose quelques références et liens supplémentaires sur le thème de la gestion des données :

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_o.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 (INSPIRE). <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 Standards Organisation (ISO). L'ISO et l'eau. Des solutions mondiales aux enjeux mondiaux. http://www.iso.org/iso/fr/iso_and_water.pdf

International Standards Organisation (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. Lefavre, 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>

Annexe 1. Éléments d'un plan de gestion des données – Liste de contrôle.

TITRE DU PROJET : _____		PROJET NO. : _____		
ÉLÉMENT	DESCRIPTION	DOCUMENTS DE RÉFÉRENCE	ÉLÉMENT VÉRIFIÉ PAR	DATE
☐ Description des données	Description (nature, étendue et volume) des données qui seront produites ou récoltées.	↓ ↓ ↓ inscrire dans cette colonne les noms et emplacement des fichiers contenant les informations.		
☐ Données existantes	Revue des données existantes pertinentes au projet, examen de la façon dont elles peuvent être intégrées.			
☐ Rôles et responsabilités	Identification des personnes responsables, des participants au projet, de leurs tâches tout au long du cycle de vie des données.			
☐ Format	Formats dans lesquels les données seront générées, conservées et distribuées considérant la pertinence de ces formats en fonction de l'archivage, de la pérennité et de l'utilisation des données.			
☐ Organisation des données	Procédures de gestion des fichiers de données, nomenclature, contrôle des versions des fichiers, etc.			
☐ Métadonnées	Description des métadonnées qui accompagneront les données, identification de la norme utilisée.			
☐ Qualité	Procédures de contrôle et d'assurance de la qualité des données tout au long du projet.			
☐ Archivage et préservation	Méthodes d'archivage et de copies de sécurité, ressources et espaces physiques et virtuels utilisés. Procédures de préservation à long terme.			
☐ Sécurité	Procédures de protection des données incluant données sensibles et confidentielles, description des permissions, restrictions, privilèges et méthodes de contrôle de l'accès.			
☐ Propriété intellectuelle	Identification de l'entité ou de (des) individu(s) possédant les droits de propriété intellectuelle (PI) des données, méthodes de protection de la PI, si nécessaire.			
☐ Auditoire	Description des utilisateurs éventuels des données.			
☐ Accès et partage	Description des méthodes de partage des données, procédures et mécanismes d'accès, conditions d'accès (accès ouvert ou restreint), estimation de la date de publication des données.			
☐ Éthique et confidentialité	Si nécessaire, procédures de protection de la confidentialité et considérations éthiques.			
☐ Budget	Coûts de la gestion de toutes les étapes du cycle de vie des données, sources de financement, demandes de subventions.			
☐ Aspects réglementaires	Obligations de rendre compte aux organismes subventionnaires.			