



**Observatoire global
du Saint-Laurent**

**Programme sur les données
environnementales côtières
de référence
2020-2021**

Rapport de projet

Intégration de données

Contactez-nous

Email: info@ogsl.ca
Web: www.ogsl.ca

Suivez-nous

facebook.com/ogsl.slgo
twitter.com/ogsl_slgo

**Programme sur les données
environnementales côtières
de référence**

2020-2021

Rapport de projet
Intégration de données

Publié le 01 juin 2021

Rédaction

Gabriel Leblanc

Révision

Étienne Caxard
Stéphane Lapointe
Anne-Sophie Ste-Marie

Graphisme

Anne-Sophie Ste-Marie

Traduction

Jovette Taillefer

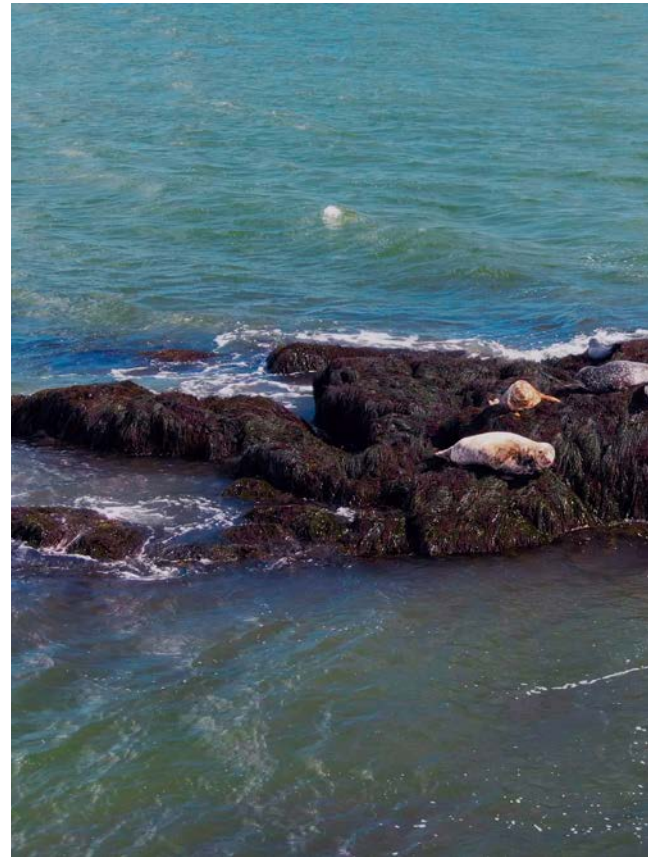


Contenu

1. Introduction
2. Les réalisations
3. Les données sensibles
4. Ouverture et discussions
5. Conclusion

01.

Introduction





Retour sur le travail entamé et pistes de développement futur

Ce rapport a été créé afin de présenter à l'ensemble des parties prenantes du Programme sur les données environnementales côtières de référence les jalons franchis grâce à votre constante collaboration avec l'OGSL. De plus, il suggère quelques pistes de réflexion émises par notre équipe que nous croyons pertinentes dans l'avancement du Programme. Or, il présente d'abord les principales réalisations de l'OGSL jusqu'à maintenant. Quelques éléments sont ainsi abordés, tels que les jeux de données intégrés, les identificateurs d'objets numériques, les normes de gestion de données et l'infrastructure informatique. Ensuite, nous avons jugé pertinent d'aborder la question des données sensibles et de proposer quelques réflexions. Finalement, ce rapport souhaite offrir quelques pistes potentielles de développement en entamant une discussion d'ouverture.

Le Programme fait partie du **Plan national de protection des océans**, financé par Pêches et Océans Canada.

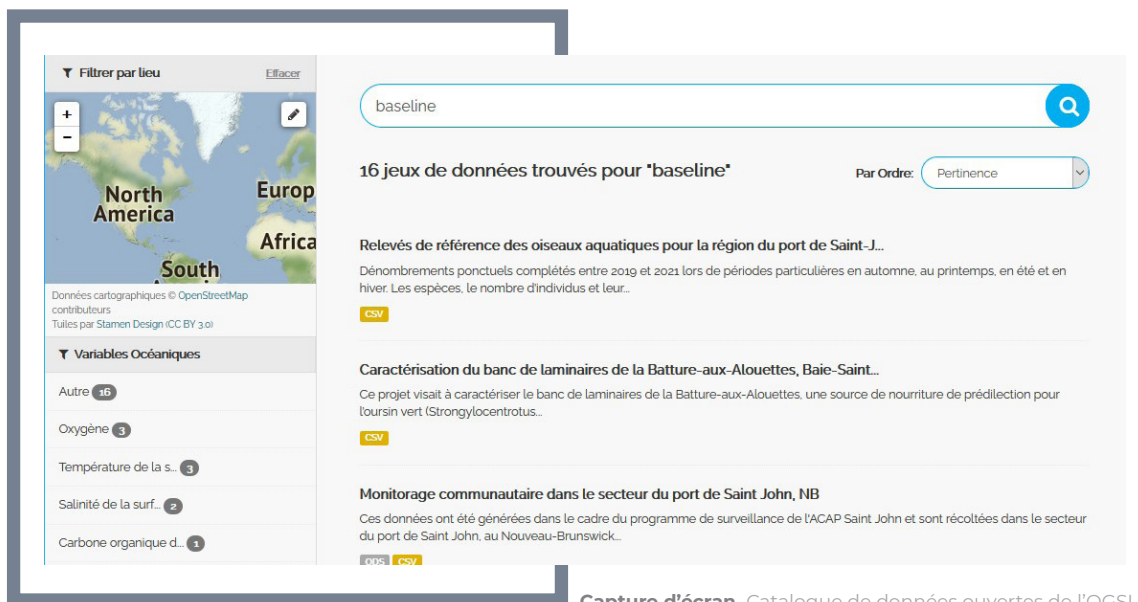


Pêches et Océans
Canada

Fisheries and Oceans
Canada

02.

Les Réalisations



Capture d'écran. Catalogue de données ouvertes de l'OGSL

Intégration de JEUX DE DONNÉES

Depuis la formalisation de la participation de l'OGSL au Programme sur les données environnementales côtières de référence en novembre 2019, maints efforts ont été déployés pour amorcer la prise de contact avec les différents partenaires. Jusqu'à présent, seize (16) jeux de données sont officiellement publiés dans le catalogue de l'OGSL. Quelques autres jeux de données sont également sur le point d'être diffusés et des discussions avec les partenaires se poursuivent. En date de mai 2021, ce sont neuf jeux de données additionnels qui sont sur le point d'être publiés. Parallèlement, plusieurs réalisations ont été complétées, notamment en ce qui a trait à

la valorisation du programme et des jeux de données y étant issus. Ainsi, afin de donner une reconnaissance au Programme sur les données environnementales côtières de référence, nous avons intégré à notre processus d'ingestion des données l'ajout d'une mention particulière dans la description des jeux de données. Cela permet notamment de les retrouver rapidement en marquant le mot-clé "Baseline" dans la barre de recherche du catalogue.



2.1 Identificateur d'objets numériques (DOI)

Pour permettre une valorisation accrue des données, l'OGSL s'est associé au consortium DataCite afin de pouvoir octroyer des identificateurs d'objets numériques (DOI) pour les jeux de données de ses partenaires. Une phase de test a été lancée lors de l'été 2020, mais le processus officiel d'attribution des DOI

a débuté à l'automne.

À l'heure actuelle, l'OGSL a alloué un total de six DOI, dont deux ont été créés pour des jeux de données de Pêches et Océans Canada liés au Programme sur les données environnementales côtières de référence.

L'attribution d'un identifiant persistant unique (comme un DOI), permet de s'inscrire dans le premier concept des principes FAIR :

Facile à trouver/Découvrable, Accessible, Interopérable et Réutilisable



Certains autres partenaires du programme ont déjà mentionné leur intérêt à obtenir un DOI sur leurs jeux de données.

Nous pensons que l'octroi de DOI sera un avantage supplémentaire pour nos partenaires actuels et futurs puisqu'il rend la citation des données plus fiable. De plus, il simplifie et augmente l'accès aux données, tout en

permettant un suivi rigoureux de leur utilisation. Ces DOI peuvent être accordés aux jeux de données et permettent notamment de citer ses données dans des articles scientifiques, par exemple. Aussi, il nous est possible d'attribuer des DOI à des ressources particulières, comme une image, un fichier tableur, un rapport, etc.

2.2 Données de biodiversité

Au niveau même de la donnée, l'OGSL s'est penché sur l'étude des normes et structures optimales pour favoriser l'archivage des données en vue de leur diffusion et pour accroître leur accessibilité.

Les normes du Darwin Core et le Ocean Biodiversity Information System

Le Darwin Core (DwC) constitue un glossaire de termes destinés à faciliter le partage d'informations sur la diversité biologique en fournissant des identifiants, des étiquettes et des définitions normalisées. Le DwC est principalement basé sur les occurrences de spécimens et suggère des termes qui permettent de les documenter par ajout d'informations. En outre, il « offre un cadre stable, simple et flexible pour colliger des données sur la biodiversité à partir de sources variées et hétéroclites. »

L'OGSL a opté de se conformer à cette norme afin de structurer les données de biodiversité qui lui sont partagées. Cette norme joue un rôle fondamental dans le partage, l'utilisation et la réutilisation des données de biodiversité en accès libre. Les termes normalisés proposés



par le glossaire DwC sont utilisés par le Ocean Biodiversity Information System (OBIS). Ensemble, OBIS et DwC forment une structure de bases de données de biodiversité internationalement reconnue.

OBIS, quant à lui, est une plateforme en ligne de diffusion et de partage de données. Il affiche plus de 70 millions d'enregistrements d'occurrences, 77 millions de mesures ou de faits étendus, 3 800 jeux de données et 154 000 espèces observées. Il s'agit d'une alliance stratégique évolutive de personnes et d'organisations partageant une vision visant à rendre les données sur la biodiversité marine du monde entier disponibles gratuitement en ligne. C'est une base de données riche en informations sur la biodiversité marine et accessible par divers outils performants (mapper, packages R, API).

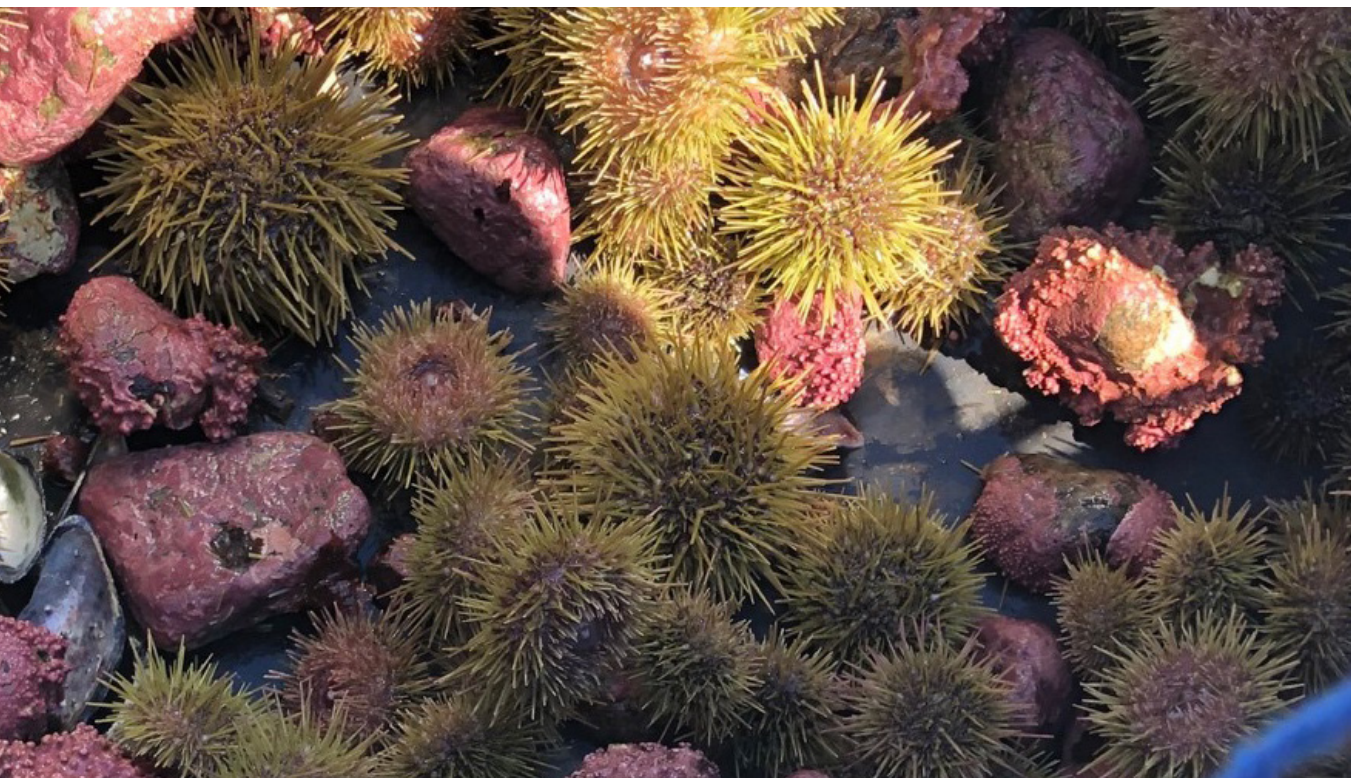


Photo. AMIK - Évaluation de la biomasse d'oursins verts à l'embouchure du fjord du Saguenay en 2019

Conjointement au DwC, OBIS a déterminé que huit termes sont requis pour rendre la donnée valide. Ainsi, lorsque nous amorçons des travaux de normalisation et de structuration de données, nous nous assurons que ces variables sont effectivement présentes. Cette information est particulièrement importante pour ce qui concerne notre protocole de gestion des données sensibles (voir prochaine section).

L'OGSL a entamé une collaboration avec OBIS Canada (le nœud OBIS régional canadien) pour partager des données sur la biodiversité selon les meilleures normes internationales. Notre organisation a maintenant terminé la phase de test pour le partage de données via la boîte à outils de publication intégrée (IPT), et nous sommes prêts à commencer l'intégration des

données dans OBIS Canada. Nos réflexions sur notre positionnement stratégique vis-à-vis d'OBIS ne sont pas encore finalisées. Cependant, nous voyons notre lien avec OBIS comme un service supplémentaire qui peut être proposé aux partenaires. Ainsi, s'ils ont l'ambition de pousser leurs données vers ce service, nous pouvons offrir un accompagnement complet.

La normalisation OBIS permet la création d'un canal d'ingestion automatisé de données vers l'application Biodiversité. À ce jour, toutes les données portant sur la biodiversité issues du Programme sont normalisées selon les pratiques d'OBIS et du DwC. Cela nous permet d'envisager facilement la migration de ces données vers l'application Biodiversité.

2.3 Infrastructure & Migration

Le Programme a permis à l'OGSL de faire quelques avancées en matière d'infrastructure, notamment avec notre application Système de Gestion des Données Environnementales (SGDE). De plus, nous avons jugé pertinent de transmettre certaines informations concernant la migration de nos serveurs et de notre espace de stockage en raison des possibilités que cela peut ouvrir pour le Programme et ses partenaires.

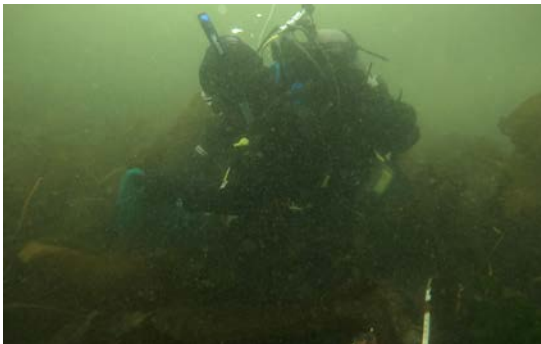


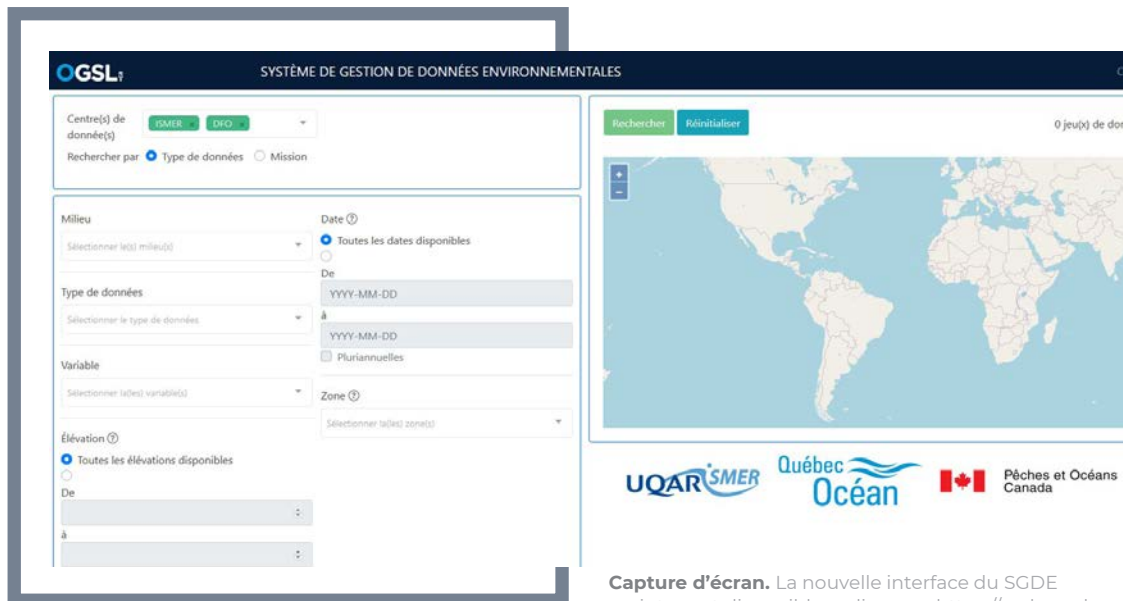
Photo. Pêches et Océans Canada - Caractérisation du banc de laminaires de la Batture-aux-Alouettes, Baie-Sainte-Catherine, Québec

D'abord, l'application SGDE a été entièrement réécrite pour interroger les services de l'Institut Maurice-Lamontagne (IML) et pour faciliter la maintenance et les développements futurs. Les jeux de données de l'IML sont désormais accessibles et téléchargeables depuis cette application. Il est également possible d'accéder directement aux jeux de données et aux missions via un lien hypertexte, ce

qui facilite le référencement et le partage, notamment par l'entremise du catalogue. Ce changement permet non seulement d'accéder aux données collectées dans le cadre du Programme, mais aussi d'accéder à toutes les données physico-chimiques de l'IML.

De plus, l'application a été modernisée et adaptée pour améliorer le rendu sur différents types d'appareils numériques (c'est-à-dire les appareils mobiles, les tablettes, etc.).

Ensuite, la migration vers les serveurs infonuagiques est officiellement terminée. Cela permettra notamment de répondre aux besoins croissants en termes de gestion de données en offrant une plus grande capacité d'adaptation aux grands besoins



Capture d'écran. La nouvelle interface du SGDE maintenant disponible en ligne au <https://sgde.ogsl.ca>

de stockage, comme les fichiers audio et vidéo. L'infrastructure est dorénavant entièrement hébergée par le fournisseur de services Microsoft Azure. Cette plateforme a l'avantage de fournir un service très fiable et des ressources à la demande. Nous pouvons donc étendre nos infrastructures

à travailler avec ce volume afin d'évaluer l'architecture la plus optimale à mettre en place. Dans le contexte du Programme, l'ajout de cet espace de stockage supplémentaire peut être compris comme un service additionnel nous permettant de répondre plus adéquatement aux besoins

Cette nouvelle infrastructure nous permet également d'accéder à plus d'espace de stockage. L'estimation de nos besoins se situant entre 10 et 50 téraoctets (TO), une réservation de 35 TO a ainsi été faite auprès de Calcul Québec.

en fonction des besoins identifiés et de notre budget. Outre le changement de serveurs, la migration a consisté à utiliser la technologie de conteneurisation pour l'ensemble de nos applications, ce qui a permis d'automatiser la plupart de nos processus de déploiement. Cela nous permet également d'avoir une infrastructure plus flexible et d'envisager des migrations facilitées vers d'autres services d'hébergement si des opportunités pertinentes se présentent.

Notre approche est donc de commencer

des propriétaires de données volumineuses. Elle permet également d'entrevoir le stockage de certaines données à plus long terme.

Finalement, nous devrions prochainement être en mesure de partager notre formulaire automatisé de saisie de métadonnées, qui sera accessible directement depuis notre site Web. Nous laisserons donc derrière nous notre formulaire Google en offrant un service modernisé et plus fiable, capable de produire des fichiers répondant à la norme ISO 19115.



Photo. JC Lemay

03.

Gestion des données sensibles

Le Programme a donné une impulsion à notre réflexion sur la gestion des données sensibles et sur la mise en place d'un protocole visant à les traiter adéquatement. Avant toute chose, mentionnons que les données peuvent être jugées sensibles par les partenaires pour différentes raisons.

Par exemple, il nous est parfois demandé de ne pas diffuser les données parce qu'une rédaction d'article scientifique les utilisant est en cours. Aussi, les données peuvent parfois contenir des informations confidentielles qui pourraient compromettre l'intégrité de l'organisation productrice de ces données ou de sa communauté d'appartenance, ce qui justifie la nécessité de ne pas les rendre publiquement accessibles.

Contextualisation de la réflexion

En outre, "rendre accessibles toutes données parce qu'elles sont financées par des fonds publics" ne nous semble pas répondre aux réalités particulières, notamment parce que cette idée, dans son application malavisée, peut provoquer certains enjeux et effriter des relations de confiance. Il convient sans doute de mentionner que le partage des données vise avant tout à laisser savoir que celles-ci ont été récoltées, de manière à éviter la duplication des efforts de collecte (et conséquemment, une dépense inefficace de fonds publics) et éventuellement permettre la prise collective de décisions éclairées. Or, la ressource n'a pas nécessairement à être diffusée, surtout si elle contient des informations qui, aux yeux de celles et ceux qui l'ont produite, peuvent compromettre son intégrité.

Cette vision est d'ailleurs en adéquation avec les principes FAIR (Découvrable, Accessible, Interopérable et Réutilisable), quatre concepts directeurs auxquels s'accolle la mission de l'OGSL. Nous nous permettons cette spécification pour rappeler notre rôle en tant que diffuseur de données, c'est-à-dire de laisser savoir aux utilisateurs finaux que la ressource existe, sans obligatoirement devoir la diffuser publiquement. Or, maintes façons de préserver l'intégrité des producteurs de données sont officiellement en place à l'OGSL.



Photo. AMIK - Caractérisation du marais intertidal de l'estuaire de la rivière de Betsiamites : Le marais de Pessamit

N'afficher que les métadonnées

Les métadonnées sont des informations qui servent à décrire la donnée. En outre, elles indiquent qui a produit la donnée, dans quelles circonstances, à quel endroit et à quel moment. Concrètement, elles peuvent aussi constituer le titre et la description d'un jeu de données. En affichant uniquement les métadonnées, dans notre catalogue par exemple, on laisse savoir aux utilisateurs finaux que la donnée existe. Il nous est alors possible d'indiquer que pour avoir accès à la donnée, il faut contacter son producteur. De cette manière, la discrétion lui revient de déterminer s'il veut partager les données au demandeur.

Ne diffuser que partiellement la donnée

Tel que mentionné, OBIS a déterminé qu'un minimum de huit variables identifiées étaient requises pour permettre la diffusion de la donnée. Or, il est possible de transmettre partiellement une donnée dans la mesure où ces variables sont contenues dans la base de données partagée. Des informations comme le compte d'une espèce, sa taille, son stade de vie, etc. n'ont pas à être divulguées si elles sont jugées sensibles. De plus, bien qu'il soit demandé de soumettre des coordonnées géographiques pour localiser une observation, il est possible de transmettre des coordonnées sous forme de boîte de limitation (ou bounding box) ou de polygones. De cette manière, il devient impossible de localiser avec exactitude l'endroit où l'observation a été réalisée, prévenant ainsi une utilisation malveillante d'une information sensible. Une logique similaire s'applique aussi pour les dates, qui constituent une variable

obligatoire selon OBIS. Bien qu'elle soit requise, il est possible de ne transmettre qu'une année, ou une année et un mois, plutôt que des dates et heures précises.

Il est important de mentionner que l'OGSL ne dispose pas du pouvoir arbitraire de déterminer quelle information ou donnée doit être jugée sensible. Nous considérons que ce rôle est tenu par le fournisseur de données, qui est le seul à même de déterminer comment la divulgation d'une information peut s'avérer sensible. Notre rôle demeure d'assister nos partenaires dans la normalisation, la structuration et la diffusion, de manière neutre et professionnelle.



04.

Ouverture et discussions

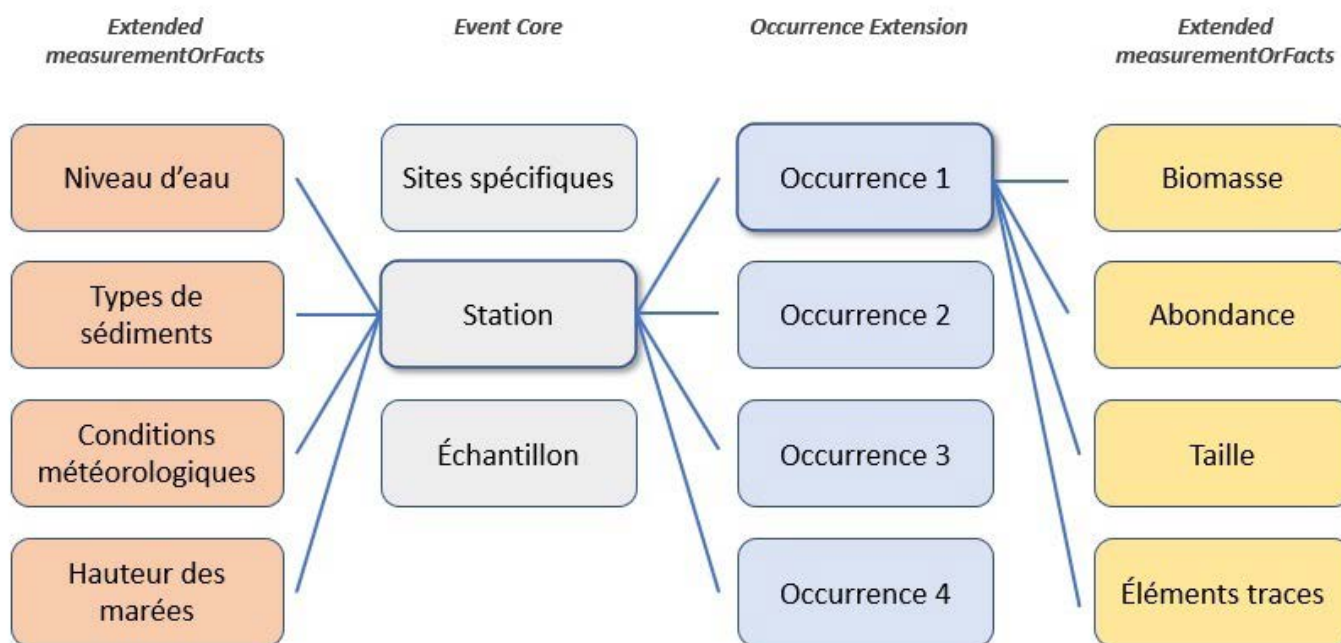


Photo. JC Lemay

Lors de cette deuxième année de la première phase du programme, l'OGSL souhaite moderniser son application Biodiversité afin de maximiser son potentiel d'utilisation et d'y intégrer les données de biodiversité récoltées dans le cadre du programme.

À ce jour, l'OGSL s'est entretenu avec de nombreux partenaires afin de les introduire aux principales normes de gestion de données concernant l'archivage et la diffusion. L'objectif de ces rencontres était d'émettre la suggestion d'une structure de données uniformisée afin d'harmoniser les données provenant d'horizons différents selon des normes rigoureuses et internationalement reconnues.

En ce qui concerne les données de biodiversité, l'OGSL a jugé judicieux d'accorder ses pratiques avec celles de OBIS. Comme indiqué précédemment, OBIS développe depuis maintenant deux décennies un ensemble de pratiques en gestion de données. Elles font dorénavant office de normes reconnues à l'échelle planétaire en biologie. Reprenant les termes développés par le Darwin Core (DwC), OBIS propose une structure de données



basée sur les occurrences, qui peuvent ensuite être associées à des données de caractérisation environnementale et à des mesures prises sur les observations (par son principe des mesures et faits étendus (EMoF) et les identifiants uniques).

Cette normalisation des données de biodiversité nous permet d'envisager la poursuite de notre travail, c'est-à-dire de favoriser l'accessibilité

des données et d'assurer leur rayonnement. Nous croyons que cette accessibilité puisse être accrue par l'intégration des ressources nous ayant été partagées dans l'outil de visualisation Biodiversité. Cette intégration constituerait une nouvelle porte d'accès à partir de laquelle les données pourraient être découvrables.

Biodiversité (suite)

Les données structurées et normalisées (selon les principes OBIS) dans le cadre du Programme nous permettent d'envisager la création d'un canal d'ingestion automatisé vers l'application Biodiversité, ce qui favoriserait une intégration grandement facilitée des données nous ayant été partagées et celles à venir. À court terme, nous envisageons cette possibilité très réaliste que les données de biodiversité en provenance de Baseline soient toutes intégrées à l'application.

D'autres développements concernant Biodiversité sont envisagés dans nos réflexions. À terme, nous aimerions

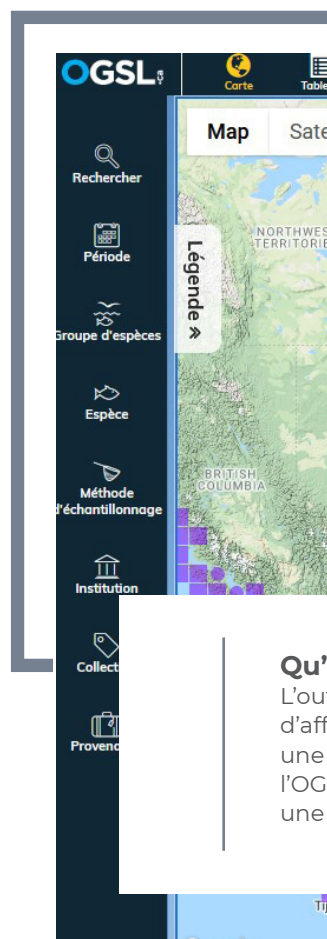
adapter l'infrastructure de cet outil de visualisation de manière à pouvoir générer des collections de données qui permettent de conserver l'attribution sur les données. Autrement dit, nous aimerions pouvoir ajouter un filtre permettant de mettre en exergue les données récoltées dans le cadre du programme, tout en s'assurant que le crédit de la saisie des données soit attribué aux partenaires qui les partagent. À ce jour, il nous est difficile de mesurer l'ampleur de la tâche, ce qui sera vérifié lors d'une phase analytique.

De plus, nous désirons développer la possibilité de générer des URL personnalisées à partir de l'application Biodiversité, de manière à pouvoir lier des ressources en particulier à une adresse. Ces URL personnalisées pourraient

Données et ressources

	dictionnaire-donnees-papinachois-amik-2019.odt Dictionnaire des données, description des variables
	stations_papinachois_amik_2019.csv Stations où d'échantillonnages des bancs coquilliers
	prises_papinachois_amik_2019.csv Échantillonnage et prises de bivalves
	long-mye-papinachois-amik-2019.csv Mesures de longueur des myes
	poids-long-mye-papinachois-amik-2019.csv Rapport poids et longueurs pour les espèces de mye

Capture d'écran. Données disponibles pour un jeu de données du Programme dans le Catalogue de l'OGSL. Pour voir l'ensemble des données du Programme, visitez <https://catalogue.ogsl.ca/dataset?q=baseline>



est-ce que l'application Biodiversité?

Le outil de visualisation Biodiversité est une application permettant d'afficher l'occurrence, l'abondance et la distribution de taxons sur une carte interactive. Il s'agit aussi d'un pont vers le catalogue de l'OGSL, puisque les fenêtres contextuelles s'affichant en cliquant sur une observation peuvent aussi être liées à un jeu de données.

Occurrence | **Sommaire**

1 de 424 données

Date	2013-06-15
Station	Petite-Rivière-Saint-François (MÉ, MRC Charlevoix)
Groupe	Poissons
Espèce	Capelan
Nom scientifique	<i>Mallotus villosus</i>
Méthode d'échantillonnage	Visuel
Collection	Réseau des observateurs du capelan - MPO
Propriétaire	Pêches et Océans Canada
Présence/Absence	Absence

notamment permettre de lier des espèces définies, des collections particulières, etc. à une adresse. Ce développement offrirait diverses opportunités. Par exemple, une organisation ayant des données dans l'application Biodiversité pourrait reprendre ce lien URL et l'afficher sur son site web de manière à diriger ses utilisateurs directement vers une carte dynamique affichant ses propres ressources; ou encore, un lien URL pourrait être inséré sur la page Baseline du site de l'OGSL, ce qui permettrait aux intéressés d'accéder directement aux données du programme

étant diffusées dans l'application; etc.

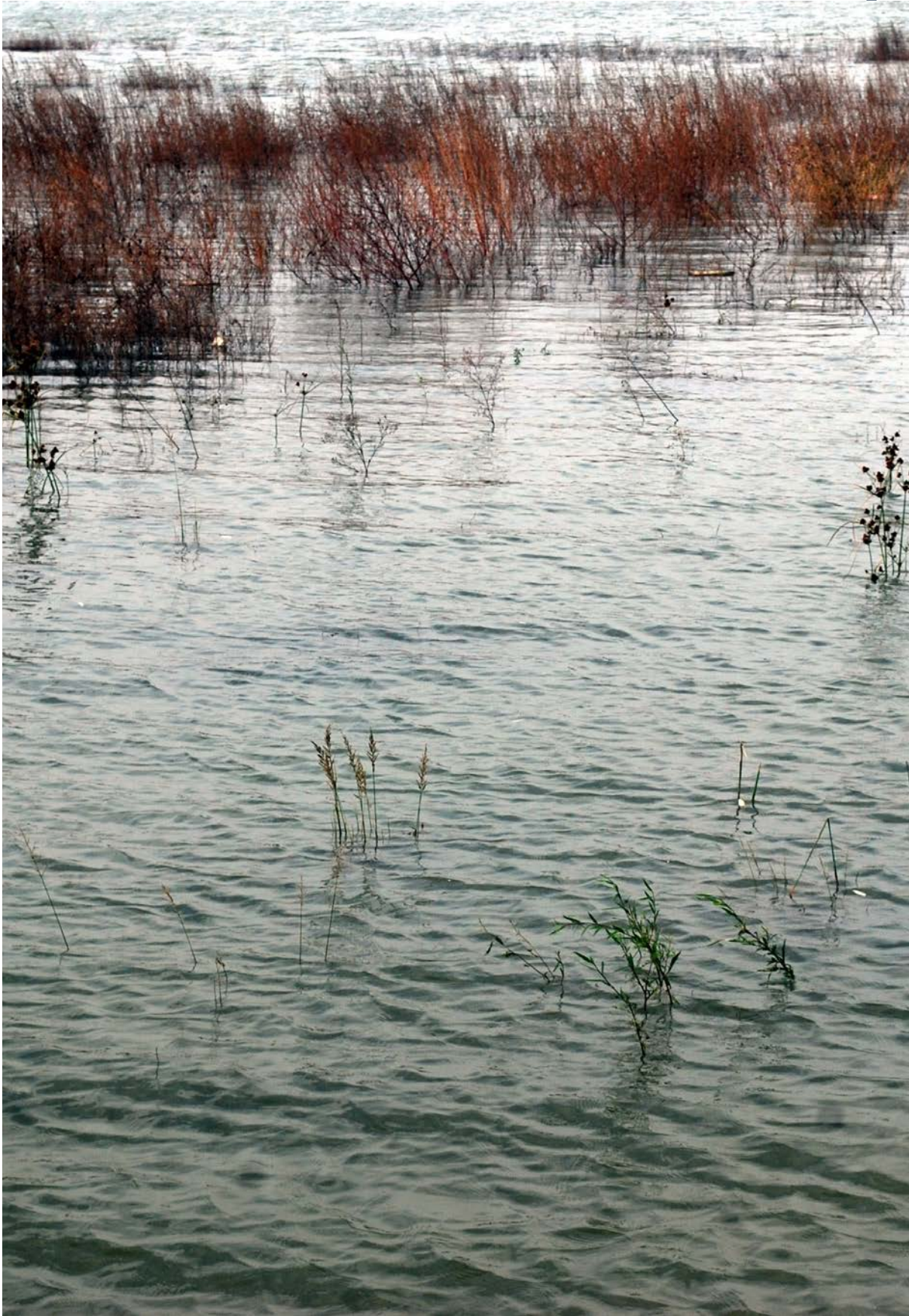
En amont du travail d'intégration, nous jugeons qu'une adaptation de la base de données qui alimente l'application est nécessaire. L'intégration des données du programme, par exemple, devrait ajouter quelques centaines de taxons (d'espèces, de genres, de familles, etc.), ce qui peut provoquer quelques ajustements pour que l'expérience de l'utilisateur demeure agréable et facile : ajouts de listes déroulantes, de filtres en fonction des rangs taxonomiques, etc.

Conclusion

Programme sur les données environnementales côtières de référence

2020-2021

En terminant, l'OGSL souhaite remercier les nombreux partenaires et fournisseurs de données qui contribuent à sa mission et à rendre accessible des données pour permettre la génération de connaissances nouvelles. Nos partenaires constituent notre raison d'être et stimulent notre travail. Les efforts que nous mettons actuellement en place visent à favoriser leur rayonnement et nous permettent des avancées considérables. Ces réalisations et réflexions n'ont été possibles que par cette collaboration à laquelle vous vous prêtez.





Rapport de projet

Intégration de données

Programme sur les données environnementales
côtières de référence
2020-2021