SLGO.ca

St. Lawrence
Global Observatory

Coastal Environmental Baseline Program
2020-2021

Project Report

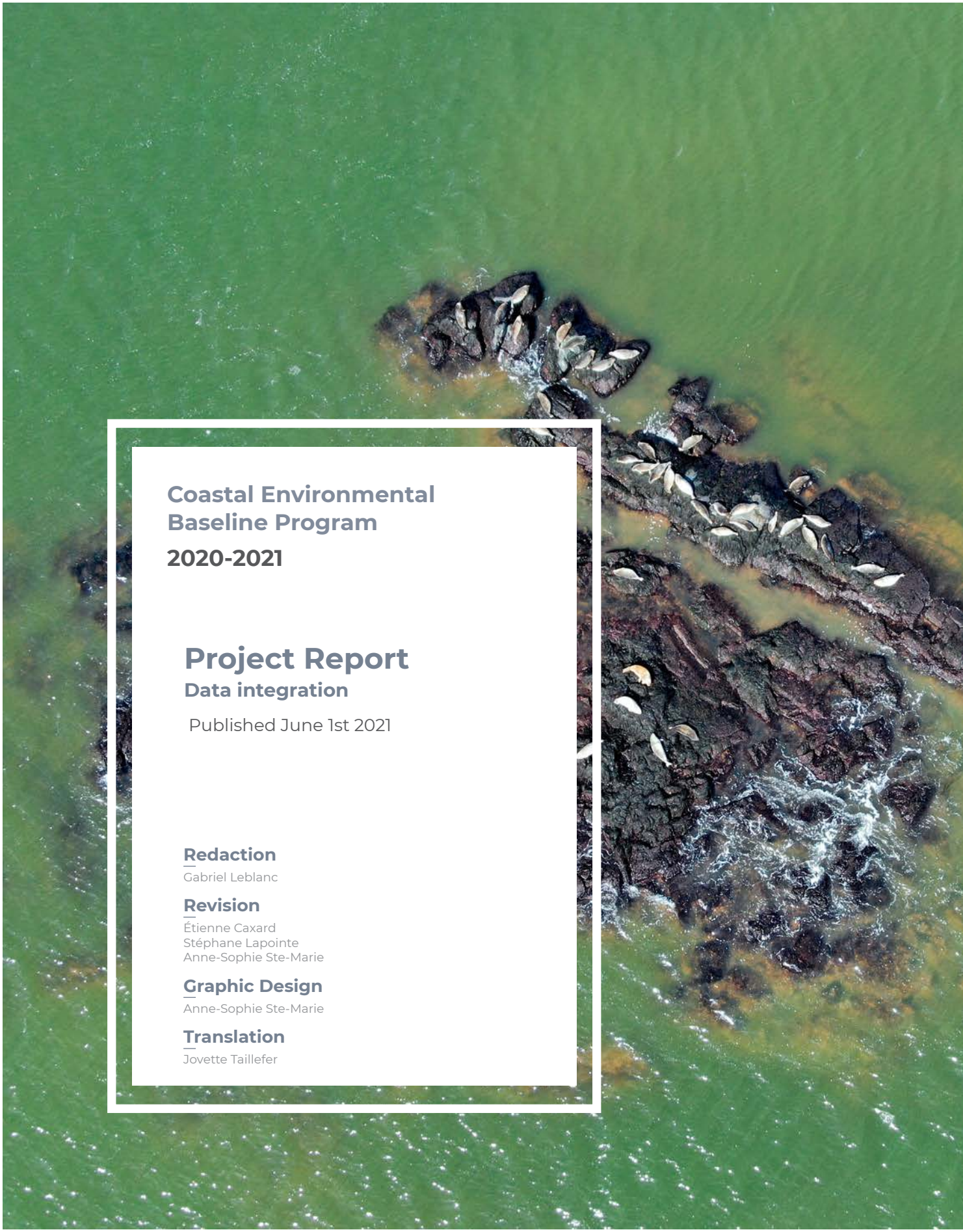
Data integration

Contact us

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

Follow us

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



Coastal Environmental Baseline Program

2020-2021

Project Report

Data integration

Published June 1st 2021

Redaction
Gabriel Leblanc

Revision
Étienne Caxard
Stéphane Lapointe
Anne-Sophie Ste-Marie

Graphic Design
Anne-Sophie Ste-Marie

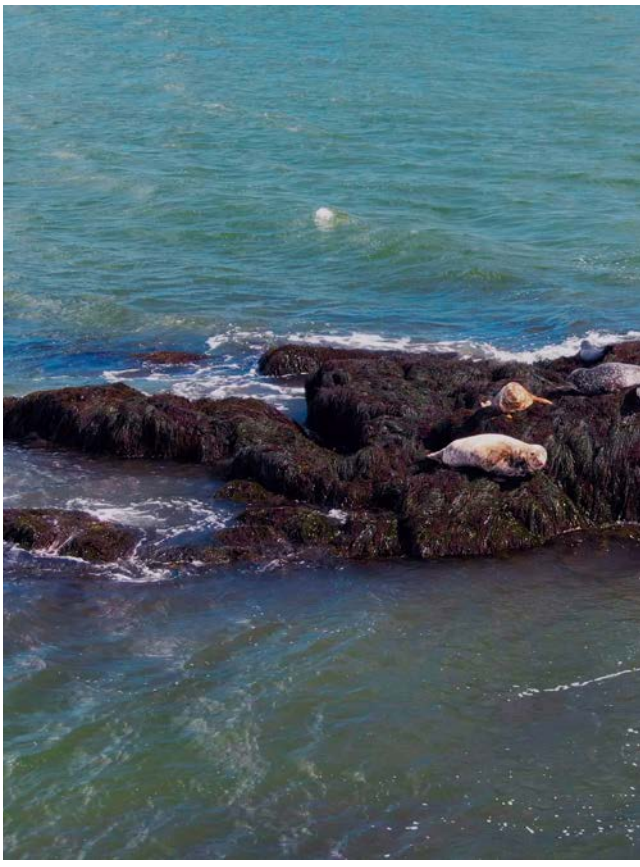
Translation
Jovette Taillefer

Picture: ACAP Saint John

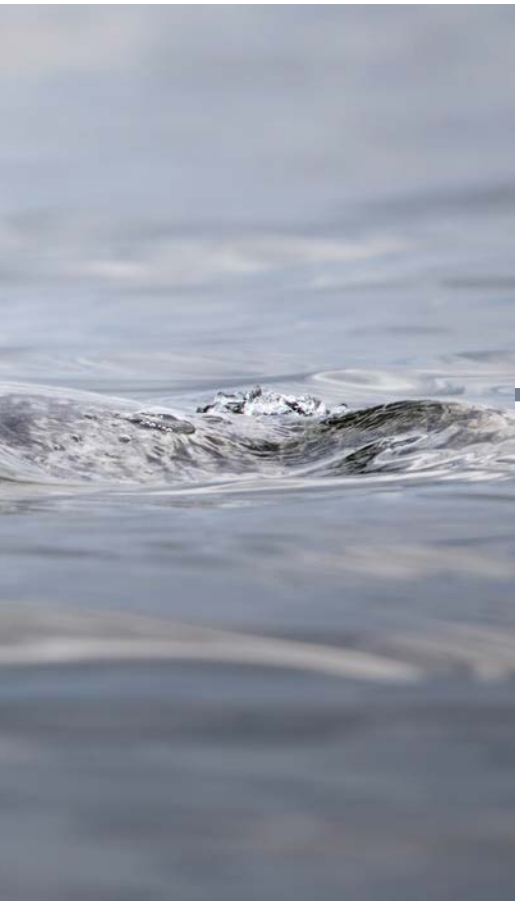
Content

01.

Introduction



Top picture: JC Lemay
Bottom picture: ACAP Saint John - Characterizing harbour s



seal populations in Saint John Harbour

Retrospective on accomplished work and avenues for future development

This report presents all Coastal Environmental Baseline Program partners with milestones achieved thanks to your constant collaboration with SLGO.

In addition, our team puts forward emerging ideas we believe are relevant to the advancement of the Program.

The report first outlines SLGO's main achievements so far. A few topics are reviewed, such as integrated datasets, digital object identifiers, data management standards, and IT infrastructure.

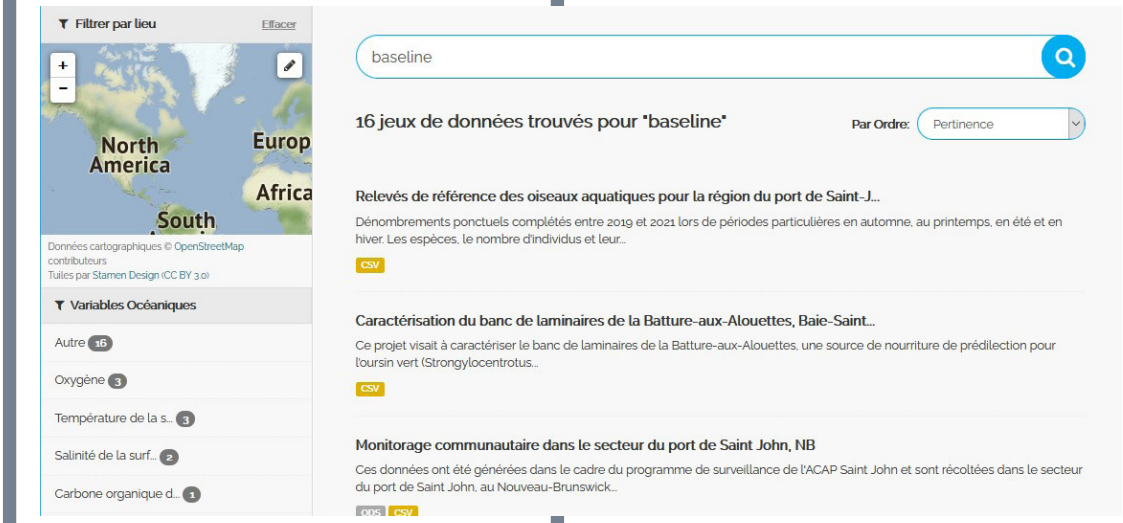
Further, we address the management of sensitive data and provide areas for reflection. Finally, we present some potential avenues for development by initiating a discussion.

The program is part of the **Oceans Protection Plan**, funded by Fisheries and Oceans Canada.

Pêches et Océans Canada | Fisheries and Oceans Canada

02.

Our Realizations



Screenshot. Open data catalogue from SLGO.

Integration of DATASETS

Since the formalization of its participation in the Coastal Environmental Baseline Program in November 2019, SLGO set out to initiate contact with the various partners that make it up. So far, SLGO has officially published sixteen (16) datasets in its catalogue. A few other datasets are also in the process of being disseminated, and discussions with partners who share data are ongoing. As of May 2021, we soon expect the release of nine additional datasets.

Meanwhile, several achievements were completed, especially regarding the promotion of the Program and the resulting datasets. We added a specific mention in the dataset descriptions to the data ingestion process and enabled a quick search with the “Baseline” marking keyword in the catalogue’s search bar.

2.1 Digital Object Identifier (DOI)

SLGO has partnered up with the DataCite consortium to grant digital object identifiers (DOI) for its partners' datasets and enable increased data valuation. A testing phase began in the summer of 2020, but the official DOI assignment proceeded in the fall.

SLGO allocated six DOIs, two of which intended for Fisheries and Oceans Canada datasets related to the Coastal Environmental Baseline Program. Some other Program partners have already mentioned their interest in obtaining a DOI for their datasets.

We believe that granting DOIs will benefit our current and future partners. It makes data

The allocation of a unique persistent identifier (such as a DOI) follows the first concept of the FAIR principles :

Findable, Accessible, Interoperable, Reusable

Picture. AMIK - Characterisation of the intertidal marsh of the Betsiamites River estuary : Pessamit marsh

citation more reliable, simplifies and increases data access while allowing rigorous data usage monitoring. These DOIs can be granted to the datasets and notably make it possible to cite their data in scientific articles. Also, we can assign DOIs to resources, such as an image, a spreadsheet file, a report, and other types of documents.

2.2 Biodiversity Data

SLGO has investigated optimal standards and structures at the data level to promote the archiving of data for dissemination and increase their accessibility.

Darwin Core standards and the Ocean Biodiversity Information System

The Darwin Core (DwC) is a glossary of terms intended to facilitate sharing information on biodiversity by providing identifiers, labels, and standardized definitions. The DwC is mainly based on the occurrences of specimens and suggests terms that allow documentation by adding information. In addition, it “provides a stable, simple and flexible framework for collecting data on biodiversity from a variety of disparate sources.”

SLGO has opted to comply with this standard to structure the biodiversity data shared by partners. This standard plays a fundamental role in the sharing, use and reuse of open-access biodiversity data. The Ocean Biodiversity Information System (OBIS) uses the standardized terms proposed by the DwC glossary. Together, OBIS and DwC form an internationally recognized biodiversity database structure.

OBIS, meanwhile, is an online platform for disseminating and sharing data. It displays more than 70 million records, 77 million extended measurements or facts, 3,800 datasets and 154,000 observed species. It is an evolving strategic alliance of people and organizations with a shared vision to make marine biodiversity data from around the world freely available online. It is a database rich in information on marine biodiversity and accessible by various tools (mapper, R packages, API).

Together with DwC, OBIS has determined that

Picture. AMIK - Green sea urchin biomass assessment at the mouth of the Saguenay Fjord in 2019

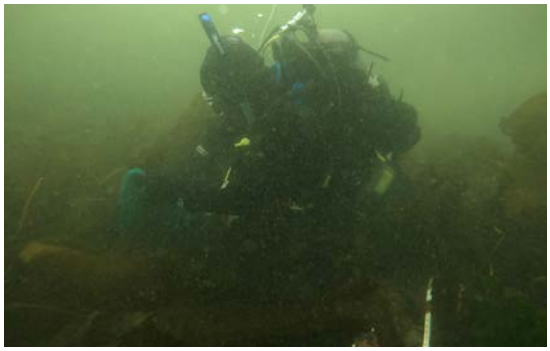
eight terms are required to make the data valid. Thus, when we begin work on standardization and data structuring, we ensure that these variables are indeed present. This information is crucial for our sensitive data management protocol (see next section).

SLGO has started a collaboration with OBIS Canada (the Canadian regional OBIS node) to share biodiversity data according to the best international standards. Our organization has now completed the testing phase for data sharing via the Integrated Publishing Toolkit (IPT), and we are ready to begin integrating data into OBIS Canada. Our reflections on our strategic positioning vis-à-vis OBIS await completion. However, we view our link with OBIS as an additional service offered to partners. Thus, if they have the ambition to push their data towards this service, we can provide full support.

OBIS standardization allows the creation of an automated data ingestion pipeline to the Biodiversity application. To date, all biodiversity data from the Program is standardized according to OBIS and DwC practices. The standardization allows us to consider migrating biodiversity data from the Program to the Biodiversity application efficiently.

2.3 Infrastructure & Migration

The Program has enabled SLGO to make some advances in infrastructure, particularly with our Environmental Data Management System (EDMS) application. In addition, we considered it relevant to transmit certain information concerning the migration of our servers and our storage space, considering the possibilities implied for the Program and its partners.

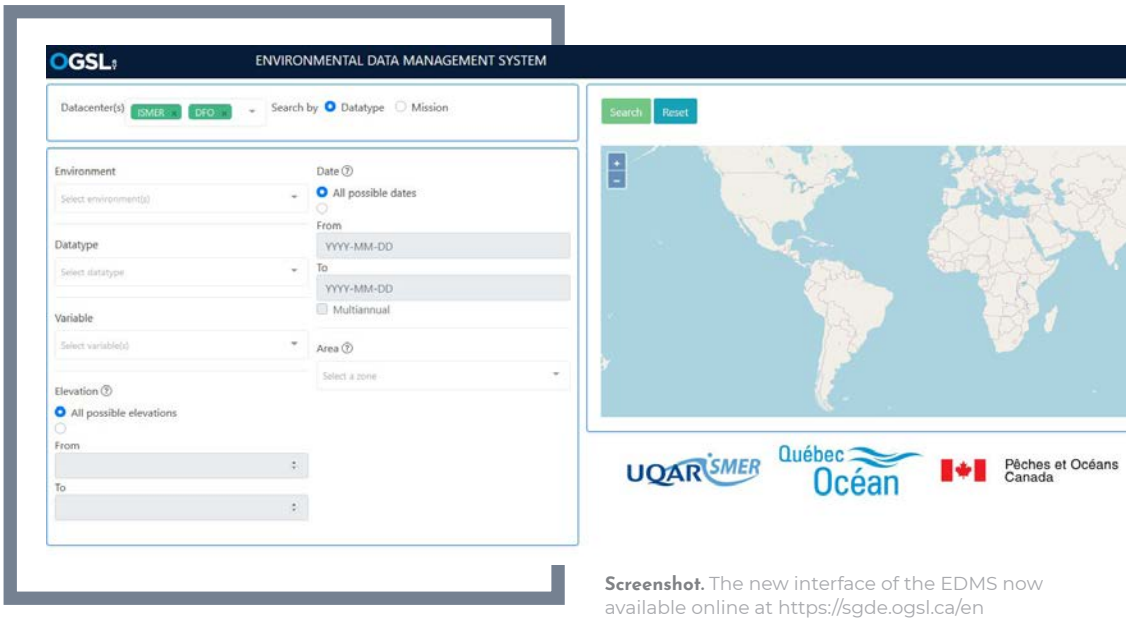


Picture. Fisheries and Oceans Canada - Characterization of the Batture-aux-Alouettes kelp bed, Baie-Sainte-Catherine, Quebec

The EDMS application was rewritten entirely to query the Maurice-Lamontagne Institute (IML) services and facilitate maintenance and future developments. The IML datasets are now accessible and downloadable from this application. It is also possible to directly access the datasets and missions via a hyperlink, which facilitates referencing and sharing, primarily through the catalogue. This change not only allows access to the data collected within the Program's framework but also to access all the physical-chemical data of the IML.

A modern and adapted app is available to improve rendering on different digital devices (i.e., mobile devices, tablets, and others).

Then the migration to cloud servers is officially complete. The change will meet the growing needs in data management by offering a greater capacity to adapt to significant storage needs, such as audio and video files. The infrastructure is now fully hosted by



Screenshot. The new interface of the EDMS now available online at https://sgde.ogsl.ca/en

the service provider Microsoft Azure. This platform offers the advantage of providing a very reliable service and resources on demand. We can therefore expand our infrastructure according to identified needs and our budget. In addition to the change of servers, the migration consisted of using containerization technology for all our applications, making it possible to automate most of our deployment processes. It also allows us to have a more flexible infrastructure and consider facilitated migrations to other hosting services if relevant opportunities arise.

We intend to start working with this volume to assess the optimal architecture to put in place. In the Program's context, this additional storage space translates to added service, allowing us to meet the needs of big data owners more adequately. It also makes it possible to foresee the storage of specific data in the longer term. Finally, we should soon share our automated metadata entry form, accessible directly from our website. Therefore, we will leave our Google form behind and offer a modernized and more reliable service capable of producing files meeting the ISO 19115 standard.

This new infrastructure also allows us to access more storage space. We estimated our requirement to be between 10 and 50 terabytes (TB) and reserved 35 TB with Calcul Québec.

Picture. JC Lemay

03.

Management of sensitive data

The Program encouraged our reflection on the management of sensitive data and the implementation of a protocol. Data can be considered sensitive by partners for different reasons.

For example, we are sometimes asked not to disseminate data currently used in a scientific paper still in progress. Data can sometimes contain confidential information that could compromise the integrity of the organization collecting it or harm the community to which it belongs, which justifies the need not to make it publicly accessible.

Contextualization of reflection

In addition, "making all data accessible because public funds finance it" does not seem to respond to different realities because this idea, in its ill-advised application, can raise sensitive issues and erode relationships of trust. First and foremost, the purpose of sharing data is to communicate its existence, avoid duplication of collection efforts (and consequently, inefficient public spending), and possibly allow informed collective decision-making. However, the resource may not necessarily be disseminated, mainly if it contains information that may compromise its integrity in the eyes of those who produced it.

Moreover, this vision aligns with the FAIR principles (Findable, Accessible, Interoperable and Reusable), four guiding concepts to which the SLGO mission is linked. We take this specification as a reminder of our role as a data broadcaster to let end users know that the resource exists without the obligation to release it publicly. However, many ways of preserving the integrity of data producers are officially in place at SLGO.

Picture. AMIK - Characterisation of the intertidal marsh of the Betsiamites River estuary: Pessamit marsh

Show only metadata

Metadata is the information used to describe data. They indicate who produced the data, under what circumstances, where and when. Concretely, they can also constitute the title and the description of a dataset. By displaying the metadata only (in our catalogue, for example), we let end-users know that the data exists. We can then indicate that to have access to the data, the user must contact the data producer. This way, the data producer determines whether to share the data with the requester or not.

Disseminate only partial data

As mentioned previously, OBIS determined that a minimum of eight variables were required to allow the data to be released. However, it is possible to partially transmit data insofar as these variables are in the shared database. Information like a species count, size, and life stage, do not have to be disclosed if considered sensitive. In addition, although it is required to submit geographic coordinates to locate an observation, it is possible to transmit coordinates in the form of bounding boxes or polygons. This way, it becomes impossible to pinpoint the exact location where the sighting was made, thus preventing malicious use of sensitive information. Similar logic also applies to dates, which are mandatory variables according to OBIS. Although usually required, it is possible to transmit only a year, or a year and a month, rather than specific dates and times.

It is essential to mention that SLGO does not have the arbitrary power to determine what information or data is deemed too sensitive. We consider that only the data provider can determine how sensitive the disclosure of information may be. Our role remains to assist our partners in standardization, structuring and dissemination, neutrally and professionally.

04.

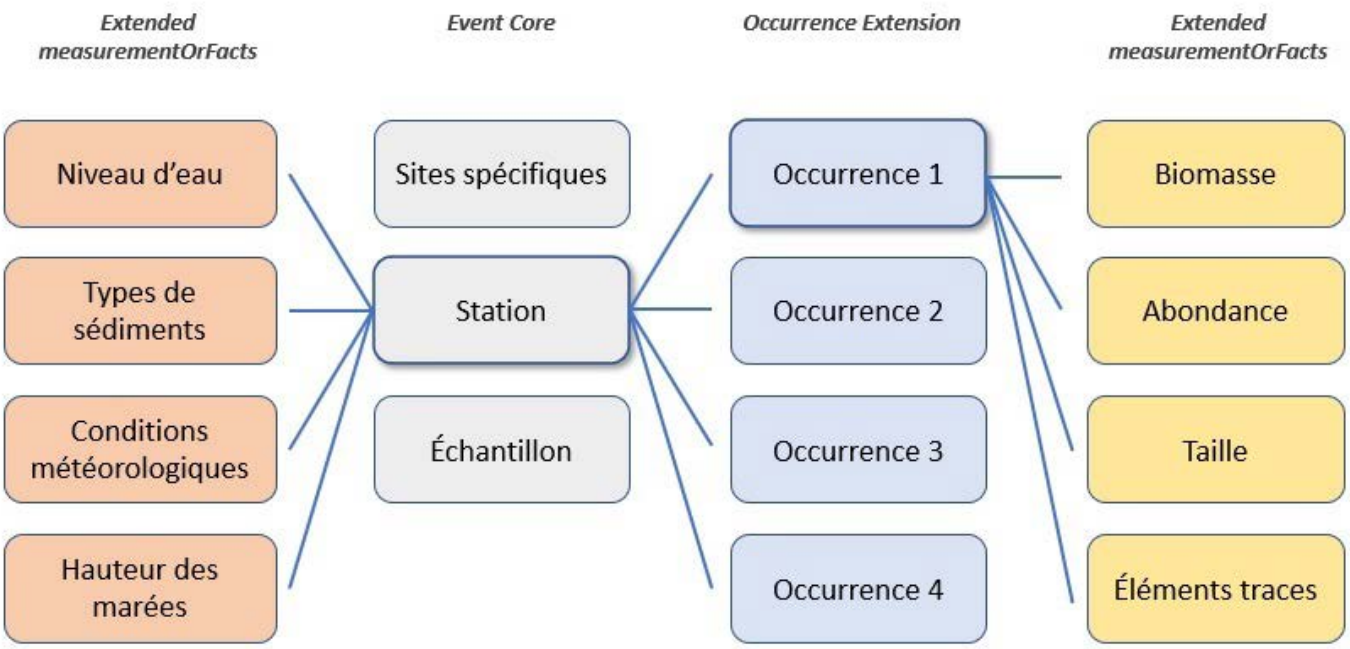
Discussions

Picture. JC Lemay

During the Program's second year of the first phase, SLGO will modernize its Biodiversity application. We wish to maximize its potential for use and integrate the biodiversity data collected under the Program.

To date, SLGO has spoken with many partners to introduce them to the leading data management standards for archiving and dissemination. The objective of these meetings was to suggest a standardized data structure to harmonize data coming from partners according to rigorous and internationally recognized standards.

Regarding biodiversity data, SLGO has deemed it wise to match its practices with those of OBIS. Indeed, OBIS has developed a set of data management practices for two decades now. They serve as a globally recognized standard in biology. Using the terms created by the Darwin Core (DwC),



OBIS offers a data structure based on occurrences, which can then be associated with environmental characterization data and with measurements taken on observations (by its principle of measurements and facts extended (EMoF) and unique identifiers).

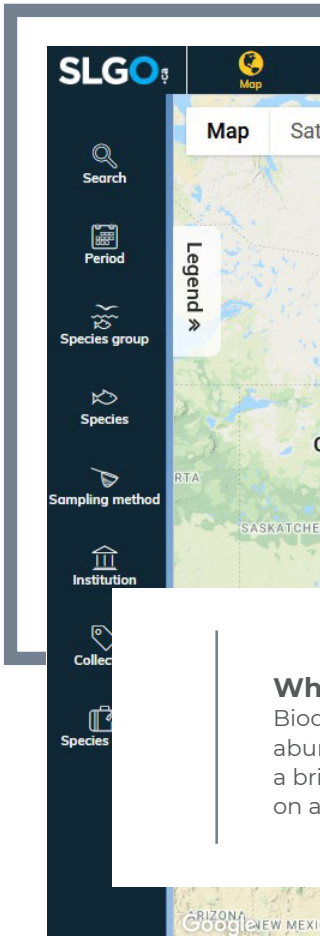
This standardization of biodiversity data allows us to consider the continuation of our work, promote data accessibility, and ensure their impact. We believe that accessibility can be increased by integrating the resources shared with us through the Biodiversity visualization tool. This integration would constitute a new gateway from which the data could be discoverable.

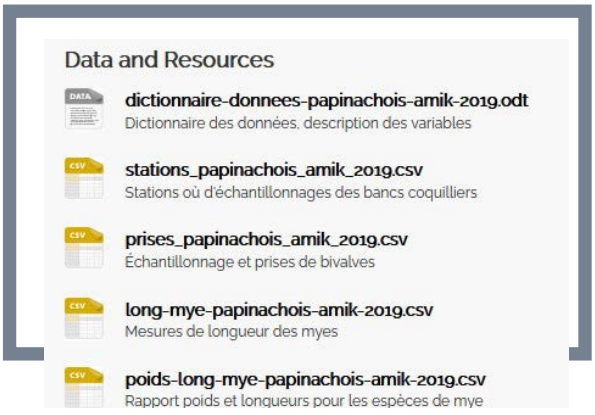
Biodiversity (continued)

The structured and standardized data (according to the OBIS principles) within the Program framework allow us to consider creating an automated ingestion pipeline towards the Biodiversity application, favouring a greatly facilitated integration of the data shared with us and those to come. In the short term, we envision the genuine possibility that the biodiversity data from the Program will all be integrated into the application.

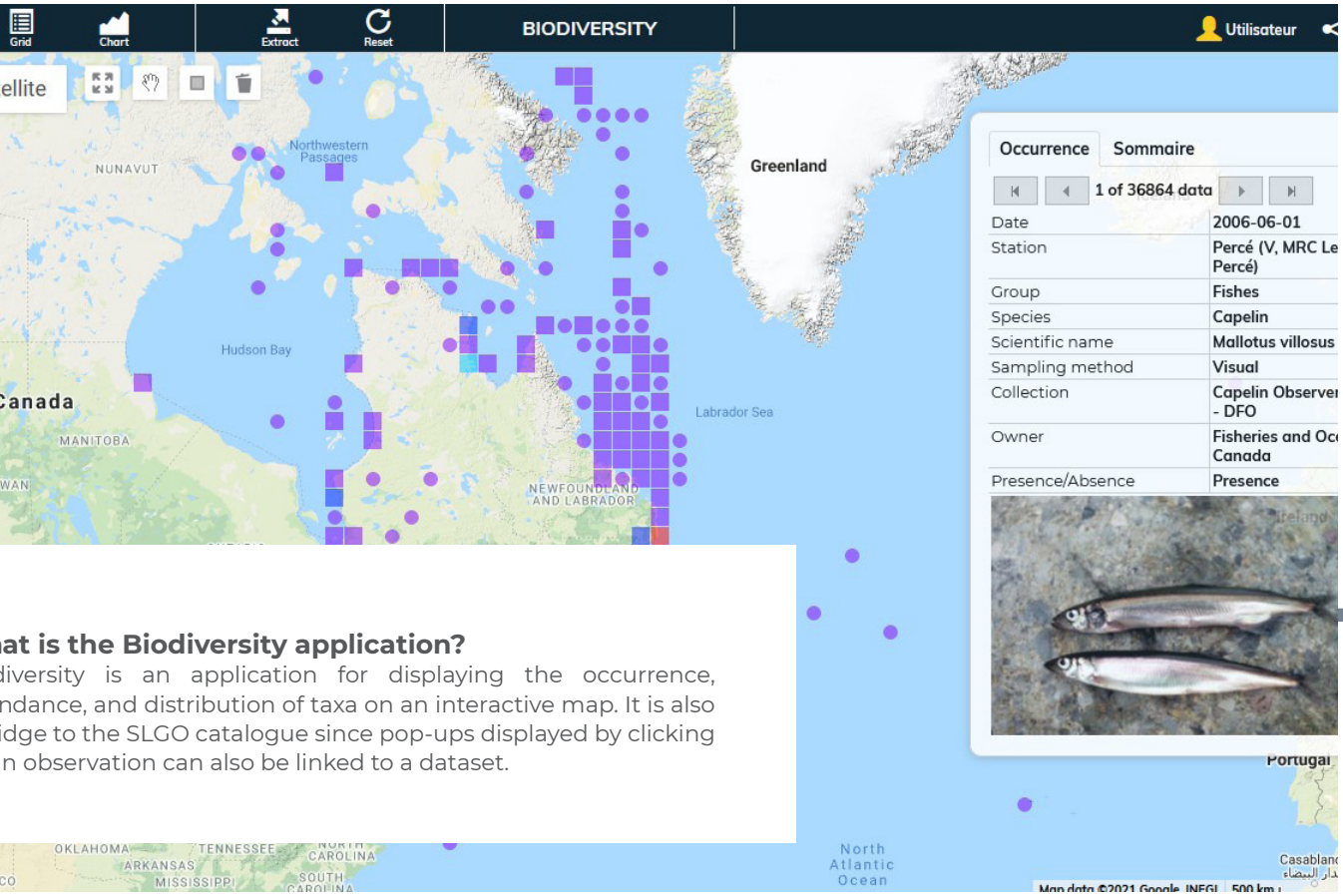
We consider other developments in our reflections concerning Biodiversity. Ultimately, we would like to adapt this visualization tool's infrastructure to generate data collections that maintain attribution on the data. In other words, we would like to add a filter to highlight the data collected as part of the Program while ensuring that the credit for entering the data is attributed to the partners who shared it. To date, it is difficult for us to measure the scale of the task, which we will verify during an analytical phase.

In addition, we want to develop the possibility of generating personalized URLs from the Biodiversity application to link specific resources to an address. These personalized URLs could make it possible to link defined species, particular collections and more to an address. This development would offer various opportunities. For example, an organization with data





Screenshot. Available data resources for a dataset from the Program in SLGO's Catalogue. To see all the Program's dataset, visit https://catalogue.ogsl.ca/en/dataset?q=baseline



hat is the Biodiversity application?

diversity is an application for displaying the occurrence, ndance, and distribution of taxa on an interactive map. It is also idge to the SLGO catalogue since pop-ups displayed by clicking n observation can also be linked to a dataset.

in the Biodiversity application could take this URL link and display it on its website to direct its users directly to a dynamic map showing its resources. Otherwise, a URL link could be inserted on the Baseline page of the SLGO website, which would allow direct access to the Program data in the application.

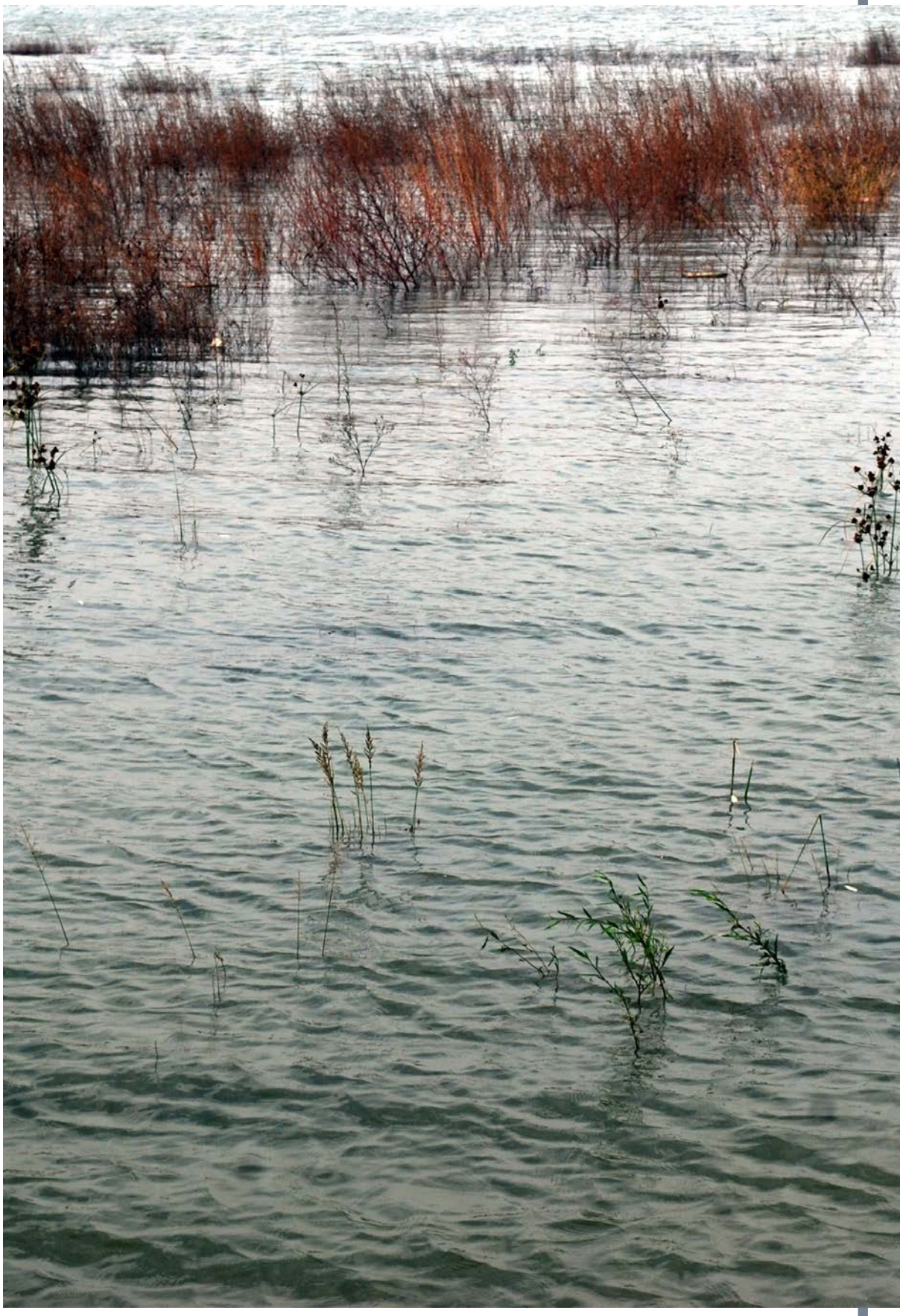
Upstream, we believe that an adaptation of the database that powers the application is necessary. Integrating Program data, for example, should add a few hundred taxa (of species, genus, families), which may cause some tweaking to keep the user experience enjoyable and easy: add drop-down lists, filters based on taxonomic ranks, etc.

Conclusion

Coastal Environmental Baseline Program

2020-2021

SLGO would like to thank the many partners and data providers who contribute to its mission, make data accessible and enable new knowledge. Our partners are our raison d'être and drive our work. The efforts promote their influence and allow us considerable progress. These achievements and reflections have only been possible through this collaboration.



SLGO.ca
St. Lawrence
Global Observatory

Project Report

Data integration

Coastal Environmental Baseline Program
2020-2021