An open, efficient, and transparent spatial reproducible reporting tool for data discovery and science advice

Quentin Stoyel, Stephen Finnis, Catalina Gomez, Gordana Lazin, Rémi Daigle, Lindsay Brager, Adrian Hamer, Charlotte Smith, David Beauchesne, Kevin Cazelles, Sean Butler

Bedford Institute of Oceanography
Fisheries and Oceans Canada
1 Challenger Drive PO Box 1006
Dartmouth, Nova Scotia, B2Y 4A2, Canada

2022

Canadian Technical Report of
Fisheries and Aquatic Sciences 3495

Fisheries and Oceans Canada | Pêches et Océans Canada

Canada

Canadian Technical Report of Fisheries and Aquatic Sciences

Technical reports contain scientific and technical information that contributes to existing knowledge but which is not normally appropriate for primary literature. Technical reports are directed primarily toward a worldwide audience and have an international distribution. No restriction is placed on subject matter and the series reflects the broad interests and policies of Fisheries and Oceans Canada, namely, fisheries and aquatic sciences.

Technical reports may be cited as full publications. The correct citation appears above the abstract of each report. Each report is abstracted in the data base *Aquatic Sciences and Fisheries Abstracts*.

Technical reports are produced regionally but are numbered nationally. Requests for individual reports will be filled by the issuing establishment listed on the front cover and title page.

Numbers 1-456 in this series were issued as Technical Reports of the Fisheries Research Board of Canada. Numbers 457-714 were issued as Department of the Environment, Fisheries and Marine Service, Research and Development Directorate Technical Reports. Numbers 715-924 were issued as Department of Fisheries and Environment, Fisheries and Marine Service Technical Reports. The current series name was changed with report number 925.

Rapport technique canadien des sciences halieutiques et aquatiques

Les rapports techniques contiennent des renseignements scientifiques et techniques qui constituent une contribution aux connaissances actuelles, mais qui ne sont pas normalement appropriés pour la publication dans un journal scientifique. Les rapports techniques sont destinés essentiellement à un public international et ils sont distribués à cet échelon. II n'y a aucune restriction quant au sujet; de fait, la série reflète la vaste gamme des intérêts et des politiques de Pêches et Océans Canada, c'est-à-dire les sciences halieutiques et aquatiques.

Les rapports techniques peuvent être cités comme des publications à part entière. Le titre exact figure au-dessus du résumé de chaque rapport. Les rapports techniques sont résumés dans la base de données *Résumés des sciences aquatiques et halieutiques.*

Les rapports techniques sont produits à l'échelon régional, mais numérotés à l'échelon national. Les demandes de rapports seront satisfaites par l'établissement auteur dont le nom figure sur la couverture et la page du titre.

Les numéros 1 à 456 de cette série ont été publiés à titre de Rapports techniques de l'Office des recherches sur les pêcheries du Canada. Les numéros 457 à 714 sont parus à titre de Rapports techniques de la Direction générale de la recherche et du développement, Service des pêches et de la mer, ministère de l'Environnement. Les numéros 715 à 924 ont été publiés à titre de Rapports techniques du Service des pêches et de la mer, ministère des Pêches et de l'Environnement. Le nom actuel de la série a été établi lors de la parution du numéro 925.

Canadian Technical Report of
Fisheries and Aquatic Sciences 3495

2022

AN OPEN, EFFICIENT, AND TRANSPARENT SPATIAL REPRODUCIBLE REPORTING TOOL FOR DATA DISCOVERY AND SCIENCE ADVICE

by

Quentin Stoyel[1], Stephen Finnis[1], Catalina Gomez[1], Gordana Lazin[1], Rémi Daigle[1], Lindsay Brager[2], Adrian Hamer[1], Charlotte Smith[1], David Beauchesne[3], Kevin Cazelles[3], Sean Butler[3],

[1]Bedford Institute of Oceanography
Fisheries and Oceans Canada, 1 Challenger Drive PO Box 1006
Dartmouth, Nova Scotia, B2Y 4A2, Canada
[2]St. Andrews Biological Station
Fisheries and Oceans Canada, 125 Marine Science Drive
St. Andrews, New Brunswick, G1S 3M7, Canada
[3]inSileco
2-775 Avenue Monk
Quebec City, Quebec, G1S 3M7, Canada

Cat. No. Fs97-6/3495E-PDF ISBN 978-0-660-44314-0 ISSN 1488-5379

Correct citation for this publication:

Stoyel, Q., Finnis, S., Gomez, C., Lazin, G., Daigle, R., Brager, L., Hamer, A., Smith, C., Beauchesne, D., Cazelles, K., Butler, S. 2022. An open, efficient, and transparent spatial reproducible reporting tool for data discovery and science advice. Can. Tech. Rep. Fish. Aquat. Sci. 3495: vi + 27 p.

CONTENTS

ABSTRACT

Stoyel, Q., Finnis, S., Gomez, C., Lazin, G., Daigle, R., Brager, L., Hamer, A., Smith, C., Beauchesne, D., Cazelles, K., Butler, S. 2022. An open, efficient, and transparent spatial reproducible reporting tool for data discovery and science advice. Can. Tech. Rep. Fish. Aquat. Sci. 3495: vi + 27 p.

Open and reproducible research practices offer a means to keep pace with rapidly expanding knowledge as science becomes increasingly data-intensive. The Science branch of Fisheries and Oceans Canada (DFO) encompasses a range of research topics yet approaches to data governance are often impeded by siloed groups and outdated workflows. Using R for coding and Git for version control, we developed a tool that generates automated reports to enable data-discovery of DFO and non-DFO information within the Maritimes region. We focus our framework on co-creation between report users, data providers, and experts to document and identify datasets along with their caveats, uncertainties, or other disclaimers. We also proactively use this as an opportunity to increase collaboration and transparency within DFO by highlighting how reproducible methods can increase efficiency and modernize workflows. Reports currently summarize over thirty data sources, with approximately twenty Reports generated thus far. This tool has reduced time spent compiling and documenting data from weeks to several minutes, allowing more time for better science.

RÉSUMÉ

Stoyel, Q., Finnis, S., Gomez, C., Lazin, G., Daigle, R., Brager, L., Hamer, A., Smith, C., Beauchesne, D., Cazelles, K., Butler, S. 2022. An open, efficient, and transparent spatial reproducible reporting tool for data discovery and science advice. Can. Tech. Rep. Fish. Aquat. Sci. 3495: vi + 27 p.

Les pratiques de recherche ouvertes et reproductibles offrent un moyen de suivre l'évolution rapide des connaissances, dans un contexte où la science nécessite toujours plus de données. La Direction générale des sciences de Pêches et Océans Canada (MPO) englobe une série de sujets de recherche, mais les approches en matière de gouvernance de données sont souvent entravées par des groupes cloisonnés et des flux de travail dépassés. En nous servant de R pour le codage et de Git pour le contrôle des versions, nous avons développé un outil qui génère des rapports automatisés en vue de permettre la recherche de données du MPO et de l'extérieur dans la région des Maritimes. Nous avons axé notre cadre sur la co-création entre les utilisateurs des rapports, les fournisseurs de données et les experts afin de documenter et d'identifier les ensembles de données ainsi que leurs mises en garde, incertitudes ou autres avertissements. Nous profitons également de l'occasion pour améliorer la collaboration et la transparence au sein du MPO en soulignant comment les méthodes reproductibles peuvent accroître l'efficacité et moderniser les flux de travail. Les rapports résument actuellement plus de trente sources des données, et environ vingt rapports ont été générés jusqu'à présent. Cet outil a permis de réduire le temps passé à compiler et à documenter les données de plusieurs semaines à quelques minutes, ouvrant ainsi la voie à de meilleures données scientifiques.

1 INTRODUCTION

The ocean and environmental sciences are becoming increasingly interdisciplinary and technology-driven, which requires researchers to develop new approaches to summarize, handle, and disseminate the vast amounts of information being collected (Jasny et al. 2011; Sandve et al. 2013; Baumann et al. 2016; Farley et al. 2018). Despite these recent technological advances, the pressure to publish has led to an entrenched individual-focused work culture and reduced collaboration (Obradović 2019; Staples et al. 2019). This has resulted in a lack of time, and incentives, to create reproducible work, yet reproducibility is a key tenet of the scientific process (Leek and Peng 2015; Heesen 2018; Munafò et al. 2020). In particular, the digital transformation has drawn attention to the importance of computational reproducibility, or the ability to attain consistent results with a dataset using the same code and methods (Peng 2011; Leek and Peng 2015). This is critical for advancing scientific work by allowing researchers to more effectively build off existing knowledge while minimizing duplication of effort (Wolkovich et al. 2012; McKiernan et al. 2016; Boland et al. 2017). With mounting human pressures on the marine environment, scientists face a growing sense of urgency to quickly and accurately study these complex systems (Baumann et al. 2016; Lowndes et al. 2017).

The Science Branch within Fisheries and Oceans Canada (DFO) spans a range of diverse research topics and operations, yet is overwhelmed by many of the same issues related to transparent, transferable, and reproducible workflows affecting science globally. Developing reproducible tools, particularly those that focus on data discovery and reporting, were identified as a potential remedy to these siloed work environments (Edwards et al. 2018; Gomez et al. 2021). In response to these needs, different teams within DFO have developed open-source software tools to address these problems:

- The Pacific Region has created multiple R-based tools for reproducible reporting including ones to generate technical reports and Canadian Science Advisory Secretariat (CSAS) documents (i.e., the *csasdown* R package, Anderson et al. 2022, https://github.com/pbs-assess/csasdown). For example, *csasdown* was used to create an automated Research Document to model the populations of 113 groundfish species (Anderson et al. 2022, see links within https://github.com/pbs-assess).
- The Newfoundland and Labrador Region has developed interactive dashboards in support of stock assessment processes (Regular et al. 2020, https://github.com/PaulRegular/interactive-stock-assessment).
- The Maritimes Region developed a reproducible atlas technical report in 2012 to model the population status, important habitat, temperature and salinity preferences for 104 fish and invertebrate species (Ricard and Shackell 2013). This atlas is currently being updated (Ricard and Gomez 2021, https://github.com/dfo-gulf-science/Maritimes-SUMMER-Atlas) and a similar approach is currently underway in the Gulf Region. Further, a collaborative framework was developed to assess and monitor Marine Protected Areas, where all data assimilation and associated methods were encoded in R (Choi et al. 2018, https://github.com/jae0/aegis).

Projects and decisions within DFO can be controversial, complex, and resource-intensive (Doubleday et al. 1997; DFO 2018). Continued advancement of reproducible reporting tools

allows for increased efficiency, quality, and transparency by making workflows both repeatable and publicly available (Lowndes et al. 2017; Munafò et al. 2017).

The Strategic Science Planning and Program Integrity division in DFO Maritimes supports various projects and presents many opportunities to modernize data management and reporting practices. In 2018, requests to the division to identify and summarize the available DFO and non-DFO datasets within the Maritimes region were becoming increasingly frequent. Due to a need for swift and effective approaches, a team of self-proclaimed Strategic Reproducible Analytical Pipeline (RAP) Champions was formed to automate the creation of these reports. The primary objective of this initiative has been to develop a web-based tool to generate Reproducible Reports to identify and describe DFO and non-DFO datasets within a user-defined area. Specifically, we address internal requests that support processes that provide frequent and standardized advice, such as CSAS, Aquaculture Siting Responses, and Environmental Response, which typically focus on Species at Risk. We have encountered multiple challenges regarding data storage, access, and duplication of effort; therefore, a broad, secondary objective has been to spearhead discussions within DFO to advance reproducible workflows and improve data management practices, aligned with Open Government mandates to make information more accessible to everyone (PCO 2018; DFO 2020a; Gomez et al. 2021; SCC 2021). This technical report provides an overview of the work done so far to create this Spatial Reproducible Reporting Tool including a description of the workflow and code, lessons learned, and future directions. We present a snapshot of the progress made in the hopes that these efforts, while fulfilling a specific reporting need, will also facilitate increased collaboration and reproducibility in monitoring and research relevant to decision-making within DFO.

2 WORKFLOW

2.1 Brief overview

This technical report represents an overview of the processes employed in the Spatial Reproducible Reporting Tool, and each main box presented is described in a separate section (Fig. 1). This work was motivated by repeat requests for reports that identified relevant datasets, which previously took several weeks to prepare. In response, this data discovery tool was developed to generate automated reports to identify relevant DFO and non-DFO datasets within a user-defined area in the DFO Maritimes region. The resulting output is an HTML document outlining datasets, caveats, sources of uncertainty, and contacts to the relevant parties.

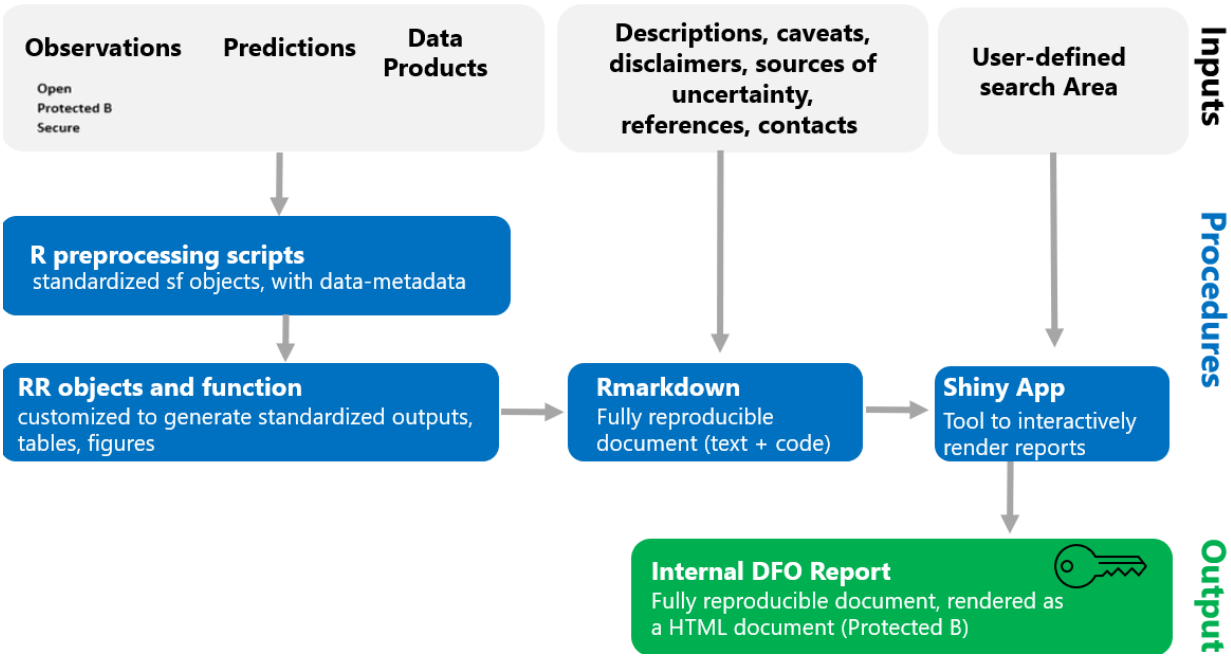

Figure 1. Overview of the generalized workflow used to create Spatial Reproducible Reports for DFO Science advice. Expert consultation and co-creation (not shown) are integral components of the entire workflow.

2.2 Collaboration

The development of this tool has been advanced and made possible by a core team of Strategic RAP Champions in close collaboration with various species experts, data providers, technicians, programmers, managers, and more, both within and external to DFO. Ultimately, the value of this tool is predicated on expert consultation and co-creation, and the input from these parties have been integral to the ideation and development process of this tool. For example, the end-users are involved to ensure the product is representative of their needs, and each section of the Report displays and summarizes the data in a way that is useful for decision-making.

Data providers and species experts have helped guide the creation of written descriptions, data visualizations, and documentation of the caveats and disclaimers of each dataset. Data collection can be challenging, difficult, expensive, and is often fit for a specific purpose; therefore, highlighting this information was essential to prevent misinterpretation or incorrect use of the data. As our work evolves and new funding streams are pursued, co-creation also ensures we continue to communicate with the relevant parties to verify that the goals and future ambitions remain reasonable and feasible with the information we are collecting, mining, and reporting.

2.3 System

We used R (R Core Team 2021) as the main programming language for data preparation, pre-processing, coding, and visualization. Specifically, we used the Shiny Application (App) (Chang et al. 2021) to create a user interface to define the search area and generate the Reproducible Reports with R Markdown (Xie et al. 2018). To avoid issues with R package updates and to ensure the versions were similar between all users, we used the dependency manager *renv* (Ushey 2022).

R was selected for the data pre-processing, user interface, and report generation because it is free, open-source, and is highly popular in biology (Lai et al. 2019; Wright et al. 2019; Jia et al. 2022), oceanography (Kelley 2018; Malde et al. 2020), and spatial data analysis (Kaya et al. 2019). R is widely-adopted within DFO (Choi et al. 2018; Edwards et al. 2018; Gomez et al. 2020, 2021), and its use presented an opportunity to maximize interoperability with tools, packages and workflows being developed by DFO Science (e.g., Ricard and Shackell 2013; Regular et al. 2020; Anderson et al. 2022). In addition, R has a large online help community, with multiple specialized packages to improve workflow (Boettiger et al. 2015; Tippmann 2015), and is frequently used in projects aiming to increase reproducibility in workflows and report creation (Lowndes et al. 2017; Xie 2017; Xie et al. 2018). Selecting R as a universal language for the tool also reduces the cognitive overhead for a small team and minimizes the technical knowledge required for collaboration. The modular workflows employed ensure that additional languages and tools can be incorporated with minimal overhead (e.g., replacing R Shiny with another user interface or using Python for data pre-processing of a particular data type) of the tool. In the future, R packages such as *reticulate* (Ushey et al. 2022) could be used to insert segments of Python code into R Markdown documents. In addition, Quarto (Quarto 2022), the next generation of R Markdown, could be used to render reports. Quarto incorporates a variety of other languages (i.e., R, Python, Julia, Observable JS), all while maintaining the literate programming advantages offered by R Markdown (Knuth 1984; Xie et al. 2018). The strength of this work is not linked to its current emphasis on R, but rather is based on the fundamentals of reproducibility and ease of collaboration among peers.

We also use Git (2022) and GitHub (2022) for version control. Git allows changes in plain text files (e.g., .R, .RData, or .Rmd file extensions containing the code), to be tracked and identified on a single computer, and GitHub is a web-based platforms to host the Git-tracked code online, and allow collaborative workflows with multiple team members (Blischak et al. 2016; Perez-Riverol et al. 2016; Git 2022; GitHub 2022). These tools are especially well-suited for team projects, since members can make additions to, or experiment with, the code without affecting the original project, and then merge it into the original workflow if it is deemed suitable (Blischak

et al. 2016; Perez-Riverol et al. 2016). Following these approaches, we can track our code with Git and revert back to previous versions if needed. Our code is made publicly available on GitHub (https://github.com/dfo-mar-odis/shinySpatialApp); however, due to issues related to Protected B information, the data are not publicly accessible. As a result, currently, only core members of our team at DFO can generate the reports in response to requests for information. The core team is not, and should not, be perceived as data providers or data custodians.

2.4 Species focus

Aquatic species (i.e., marine mammals, fish, reptiles, and mollusks) that have been listed under the Species at Risk Act (SARA 2002) or are under consideration for listing (i.e., assessed by the Committee on the Status of Endangered Wildlife in Canada, COSEWIC) are the current focus of this Reproducible Reporting framework. However, recognizing the importance of ecosystem-based approaches for management, this Report also includes information on ecosystem components (e.g., intertidal vegetation and habitat), and areas designated for spatial planning (e.g., Ecologically and Biologically Significant Areas). In addition, users of the Report have requested several additional species to be included due to their specific reporting needs.

2.5 Inputs

2.5.1 Data sources

Data are gathered from multiple DFO and non-DFO sources and are available in a variety of forms. We have adapted our workflows to accommodate these various data types. This has included but is not limited to:

- R packages (e.g., the *robis* R package for accessing the Ocean Biodiversity Information System (OBIS) data, Provoost and Bosch 2021);
- URL links to the data (e.g., the Government of Canada Open Data portal, Open Data 2022); and
- Emailed data files (e.g., csv files).

The data compiled and used by our core team are in multiple different formats including both vector (i.e., points, lines, polygons) and raster (i.e., gridded/cell-based) datasets. Although this work typically focuses on species presence (i.e., point data), several datasets are obtained from trawl surveys (i.e., line data) and habitat information (i.e., polygon data). Derived data products or species predictions are often in polygon or raster format.

We track the datasets that have been or will be incorporated into the Report in an Excel spreadsheet on SharePoint, and use this to state approximate timelines for completion, and to monitor which team member is responsible for the addition of each new dataset or section. This spreadsheet is continually updated as we progress, or as we are made aware of additional datasets that may be useful for the Report.

2.5.2 Data storage and updates

Copies of the processed data, metadata, and final Reports are stored on DFO's internal network, but we intend to move all datasets to the Azure cloud system once it can host Protected B information. Since new data or changes to datasets can bring challenges to our Report, we do not host "live" versions of the data. For example, the code may not be adapted to deal with potential issues such as new data formats or outliers. However, for Open Data records, automated GitHub actions were developed to periodically run tests and update the Open Data files. Our team is currently investigating methods and best practices for updating the datasets so the information remains current.

2.5.3 Data access

The datasets accessed by the Reproducible Reporting tool have differing Government of Canada security levels, which affect how we store and distribute the data. For each dataset, we identify its security level from the following classifications:

- None: data do not present any security risks or include sensitive government information and assets; and
- Protected B: Data may include information or assets that, if compromised, could cause serious injury to an individual, organization or government (PWGSC 2021).

The above classifications are then used to guide data use constraints, which specify how the data can be used by individuals. These include:

- None: data can be used and freely shared by anyone. This is typical of data that are made publicly available such as through the Canada Open Data Portal (Open Data 2022), or other online portals such as the Ocean Biodiversity Information System (OBIS) (2022) and the Global Biodiversity Information Facility (GBIF) (2022).
- DFO Internal Use Only: data are not to be shared externally to DFO. This is true for all Protected B data, and some DFO datasets that are not shared publicly.

Due to the presence of Protected B data in some Reports, we treat all Reports as Protected B, and they are not to be shared outside of DFO. We are currently considering approaches to create Reports that do not contain Protected B data, so they can be freely-shared with the public or other government departments or agencies.

2.5.4 Quality tiers

Tiers of data quality for each dataset were developed by our team and the data providers in order to provide guidance and urge caution when interpreting the data summaries (Table 1).

Considerable subjectivity remains in this classification; we welcome suggestions about how to improve this section to characterize information from a vast variety of sources.

Table 1. Quality tiers defined to describe sources of information queried and summarized in this Report.

Tier	Description	Usage
High	Records, products and/or outputs available from systematic surveys, and/or with qualified observers/personnel, including records vetted through peer-review processes.	These records are high quality, and come from reliable sources. Using this source is recommended.
Medium	Records, products and/or outputs available from a mix of opportunistic and systematic surveys. Some quality control has been applied, although additional work is required to verify information.	Use with caution and, where possible, validate inference with data assigned to the High Quality tier.
Low	Primarily opportunistic surveys that have not been through a quality control process, or results of invalidated models. Data quality protocols and validation are required.	Use only if data in High/Medium tiers confirms or validates information from this tier.

2.5.5 Writing descriptions

Written descriptions of each dataset are included with the data search outputs. Topics include an overview of the sampling strategy, survey locations, sampling year, sources of uncertainty, and references to the literature. For records from Open Data, we use the dataset description directly from the Open Data record. For all other records, dataset descriptions are created in collaboration with the experts to ensure the descriptions are correctly conveyed and communicated. Notably, the identification of any caveats and disclaimers is fundamental to our work to ensure the data is not misinterpreted or misused beyond its original intention.

In addition to disclaimers associated with each specific dataset, there are several disclaimers described below that are reflective of the entire Report including:

- This document is a tool to support, not to replace, science advice and peer-review processes.
- This Report does not endeavor to describe every source of information available; the data

presented does not represent all available data within the search area and additional information may be available from other sources, or more recent data may be available than what is presented here.

- This Report is not intended to provide the data itself, but to summarize what is available. Users are encouraged to:
  - Access the original data source if more information is required;
  - Circulate this document to all contacts outlined in the different sections of the Report for each data source, to ensure the veracity of inference drawn from the Report, and to provide any supplemental information that may support the provision of science advice.
- The focus of this Report is on available observations of species presence and not on absences, quantities (e.g., species abundance or biomass), frequency, or catch information.
- The absence of a species in this Report should be interpreted as an absence of observation or reporting of the species, not necessarily as a true absence of the species.
- Unless otherwise specified, data were queried from 2010 to present. This was selected as an arbitrary time range to present the most recent information. Please contact data providers if data outside of these ranges are required.
- Current outputs of species observations in the Report aggregate all years in summary plots and tables (i.e., 2010 to present). Annual summaries are not displayed. If you require annual summaries or other information beyond species presence, please contact relevant data providers/custodians for each data source.
- All maps in this Report are for informational purposes and are not suitable for legal or surveying purposes. Maps represent only the approximate location of boundaries.
- This Report is intended for internal DFO use only. Privacy screening (i.e., the rule of 5, see DFO 2020b) is not yet integrated into all data products and summaries generated by the Report. Outputs of this Report are not suitable for real-time, dynamic spatial planning.
- Because coastal areas of the Scotian Shelf bioregion are generally not adequately sampled to characterize fine-scale ecological characteristics (i.e., species, habitat composition and variability within a sub-regional focal area), the distribution of (some) species featured within the Report, when the search area is near to the coast, should be used as a first-order estimate limited by the spatial-temporal resolution of available data.

2.5.6 User-defined search area

The search area is determined by the Report user and often reflects dispersal of a substance such as Predicted Exposure Zones for aquaculture (DFO 2021) or the estimated area of dispersal from an oil spill. These search areas are provided to the Shiny App in several different ways including (Fig. 2):

- Draw from map: user can interactively draw a search area on the map in various formats including polygon, rectangle or circle;
- Bounding box: user provides minimum and maximum latitude and longitude coordinates;
- Individual point: user provides coordinates for a single location. In addition, a buffer, in meters, is typically also provided to search for data within a larger area; and
- Upload from file: user can upload a spatial data file. This tool supports a variety of vector formats (e.g., ESRI shapefiles, KMLs, etc.).

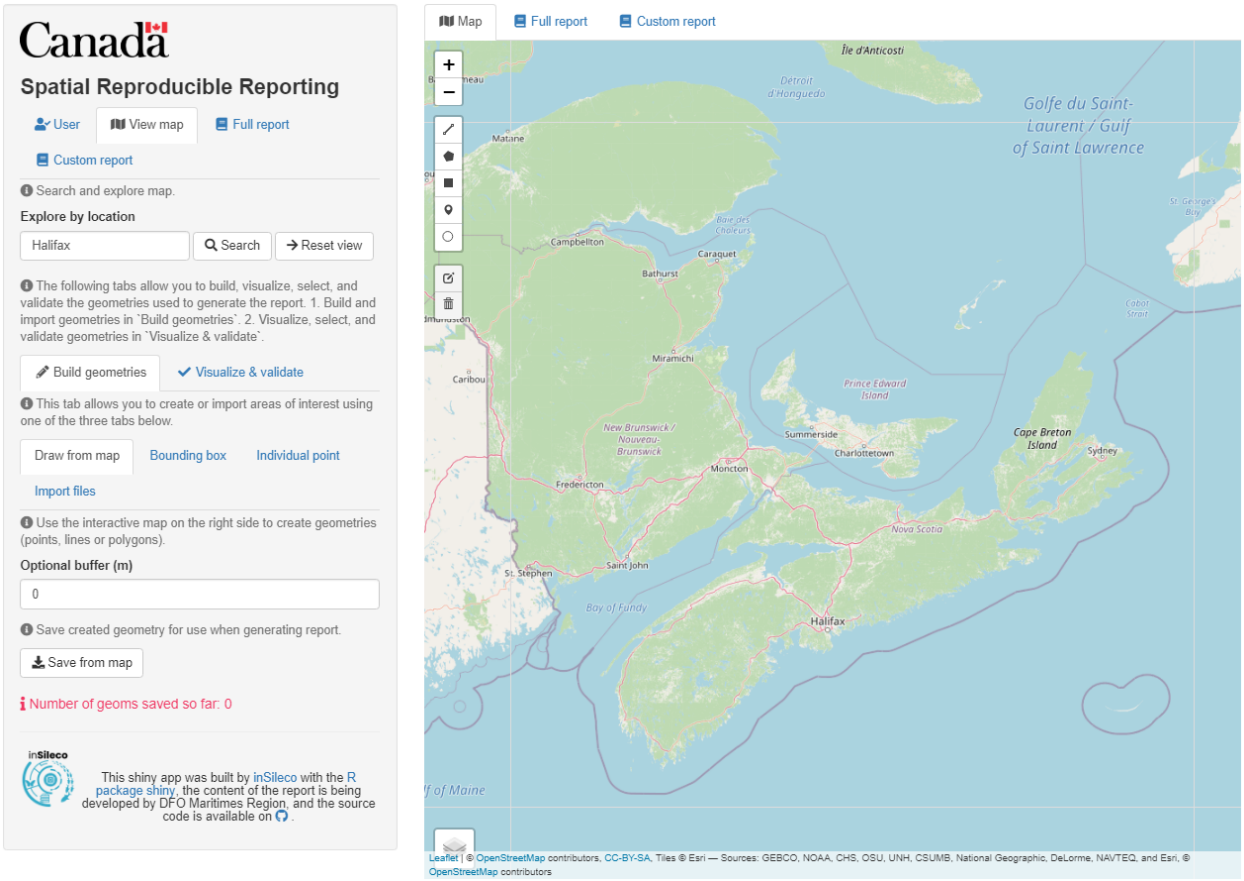

Figure 2. Screenshot of the R Shiny Application developed to draw an area of interest in various formats (draw from map, bounding box, individual point, or import files). Currently, only members of the core team are able to generate Reports due security concerns associated with Protected B information.

Typically, these Reports are generated to search for data within a single user-defined search area. However, the tool is flexible enough to allow a search for data within several areas. In these instances, the data within all areas are combined in the search outputs. If the data for each user-defined search area needs to be separated, different Reports could be created (e.g., one user-defined search area per report, or one report summarizing different components of multiple areas).

2.6 Procedures

2.6.1 Preprocessing

The provided datasets are subject to various preprocessing procedures to standardize the data files and ensure consistency between data objects before they are loaded into the R Markdown files. Each dataset is assigned dedicated preprocessing procedures, written as an R script, and includes (where applicable):

- Filtering years from 2010 until present (we exclude data prior to this date);
- Splitting the data (if necessary) for entry into the different sections of the Report using a spreadsheet list to assign them into these classes;
- Selecting required fields or columns and renaming columns to standardized headers (e.g., year, scientific name, species name, COSEWIC status, SARA status, latitude, longitude, etc.);
- Converting the spatial data into either a common data format. For most cases, standardized simple feature (*sf*) objects are used (explained more below), with raster data being converted into *raster* objects.
- Transforming any reference system to a common standard (WGS84);
- Clipping spatial data to the extent of the region; and
- Attaching the associated metadata.

The final output of the preprocessing steps is a named list with a format referred to as a Reproducible Reporting (*RR*) object which is saved into a .RData file.

2.6.2 Metadata

We refer to “metadata” as the information listed at the beginning of each dataset in the search outputs of the Report. These entries were determined by our team to include several key attributes of the datasets that Report users should be aware of. The metadata is stored within the *RR* objects created in preprocessing. Metadata entries include:

- Title: dataset title used in the report.
- Contact: name and email address of relevant contact for the dataset.
- URL: dataset URL to provide more information about the data.
- Last retrieved on: the date when the data were retrieved from the data provider or when the data provider sent us the data.

- Search years: the years of data collection. This is typically from 2010 until present, or as much as the dataset covers. Note that more data can be provided if needed, but this was selected as an arbitrary time range to present the most recent information relevant for Aquaculture Siting and Environmental Response.
- Security level: options are None or Protected B. Refer to the “Input data” section for a more detailed description of each option.
- Data use constraints: used to define how the data are allowed to be shared. Options are None or DFO Internal Use Only. Refer to the “Input data” section for a more detailed description of each option.
- Quality tier: options are High, Medium, or Low. Refer to the “Input data” section for a more detailed description of each option.

2.6.3 *RR* objects

The *RR* objects are created as a result of the preprocessing steps described above. These standardized objects are used as data inputs for a variety of common functions to develop the dynamic aspects of the Report (e.g., metadata, tables and maps). The spatial data in the *RR* objects are stored as simple feature (*sf*) objects, a formal standard for storing spatial data (ISO 19125-1:2004, Pebesma 2018). *sf* objects store both species information (e.g., common name, scientific name, SARA status, COSEWIC status, etc.) and a geometry column that contains the spatial information of the data. The most common geometries include point, linestring (i.e., line), and polygon, although others are available (Pebesma 2018). *sf* objects also include a coordinate reference system which describes how to place the data on the Earth’s surface (in this case, WGS84). Although *sf* objects could readily be exported as an ESRI shapefile or, more appropriately, another open data format (e.g., GeoJSON), saving these as .RData minimizes the read and write times, and prevents various formatting issues (e.g., ESRI shapefile field name character limits).

2.6.4 *RR* functions

Once in the *RR* format, custom functions were developed to further manipulate these data for visualization and search purposes. This includes listing the metadata, partitioning the data based on various spatial extents, and displaying the data in tabular (or other) format. These functions are then used in the R Markdown scripts to generate the Reports. The use of *RR* objects allows common functions to apply to most datasets, which reduces both coding effort associated with new datasets, and the total quantity and complexity of code. These functions increase the output quality by ensuring that figures, tables, and sections have a common formatting. Where output tables are necessary, functions were created to produce standardized tables of the search results. Typically, the total number of records of a Species at Risk located within the user-defined search area and search period (2010 to present) are tabulated, although based on sampling methods, sometimes only species presence is shown. Other information, such as the species’ COSEWIC and SARA statuses, are also included.

2.6.5 R Markdown

We used R Markdown to generate the final Reports. Each dataset within the Report contains an R Markdown file (extension .Rmd) which can be loaded into the main report as an R Markdown Child. Static information is written directly into the files and includes information such as the dataset descriptions, disclaimers, and caveats. The formatting of this text is also specified so features such as the headings and typographic emphasis (e.g., underline, bold, italics) are consistent. The dynamic information is called using the functions described above to display information such as the metadata, tables, references, and maps. In our case, the final document is rendered as an HTML file, although other formats are possible including PDFs or Microsoft Word Documents (Xie et al. 2018).

2.6.6 Shiny App

Built by inSileco (https://www.insileco.io/), the Shiny App supports the users in the process of interactively identifying a search area, and liaises with the R Markdown documents to generate the Reports. Because the code is fully open and available on GitHub, edits to the App can be made by the team at DFO without needing to contact the external contractors at inSileco. Due to security concerns associated with Protected B information, currently only the core team of this project are able to generate Reports.

The Shiny App, or Spatial Reproducible Reporting Tool, has various tabs on the left-hand side, with a Leaflet map display (Graul 2016) on the right to show spatial context and define the user's search area (Fig. 3). First, the user must enter their contact information and accept the Terms and Conditions of the Report (Fig. 3A). Next, the geometry of the search area is defined, either by directly drawing on the map, importing a spatial file, or coding numerically into the applicable box (Fig. 3B). These geometries are then validated to ensure topological consistencies, such as non-overlapping search areas or that the polygons are closed (Fig. 3B). The user can either generate a Full Report and select the relevant sections to be included (Fig. 3C), or generate a Custom Report which has fixed content for each end-user application (Fig 3D). The process for generating Custom Reports (Fig. 3D) is still in development. When specifying either Report type, the user can also define the filename, and select the language (English or French) to be used.

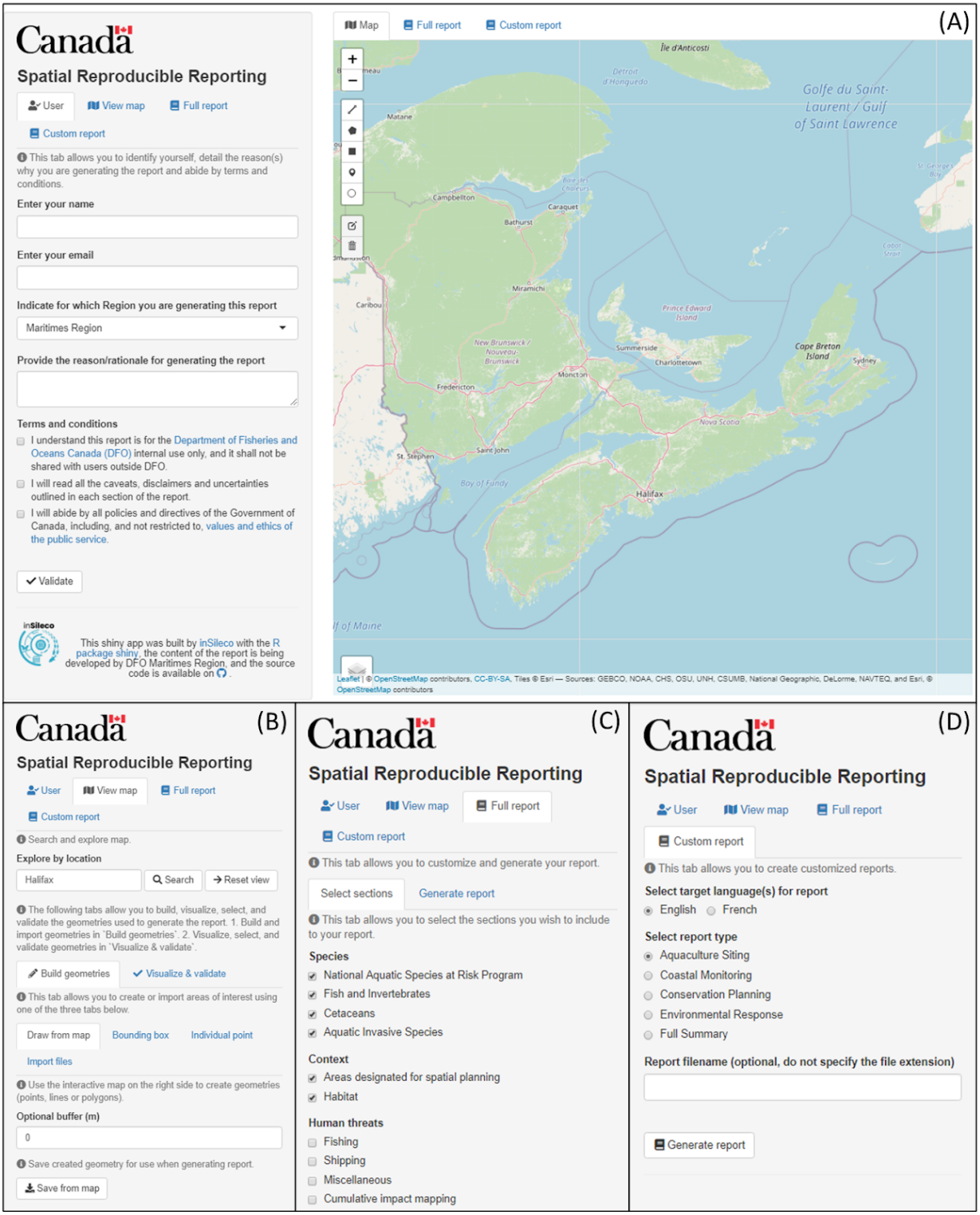

Figure 3. Overview of the R Shiny Application that allows users to (A) state their contact information and accept the Terms and Conditions, and (B) enter their region of interest and ensure these geometries are valid (e.g., polygons are closed). The user then either (C) generates a Full Report and selects the relevant modules to be included, or (D) creates a Custom Report that is tailored for various applications. Currently, only members of the core team are able to generate Reports due security concerns associated with Protected B information.

2.6.7 Output

The final HTML document takes approximately four minutes to generate. The Report has various sections and is constantly being adapted as new datasets are added, or as the goals and objectives of this work expand. This section describes the outputs for Full Reports (Fig. 3C), rather than Custom Reports (Fig. 3D), which are still in development. Currently, these sections are as follows:

1. **Background information**

   This section includes a multi-paragraph description of the Report including an overview of the intent of the document, the system (i.e., the programming behind the Report), disclaimers for Report use, data access, and quality tiers.

2. **Search results**

   Within this section, the available data within the user's search area are presented. If no data for a specific dataset are observed, an automatically-generated message will appear stating this. When data are present, typically both a map depicting the data observations, and a table with a data summary are presented. For point data, tables usually document the number of occurrences for each species within the search area. For polygon data (e.g., Ecologically and Biologically Significant Areas), tables typically state which regions intersect with the search area, although this may vary depending on the dataset. Currently, datasets are presented within three main modules, including:

   i. **Species**: Results are currently categorized into (a) Information from the National Aquatic Species at Risk Geodatabase, (b) Fish and Invertebrates, (c) Cetaceans, and (d) Areas designated for spatial planning (Fig. 3C). We recognize these are not perfect divisions, but were selected as a method for grouping similar datasets and were primarily based on our users' needs. Datasets with information in more than one of these listed modules may be listed more than once. For example, data from OBIS and GBIF contain both "Fish and Invertebrate" (b) and "Cetacean" (c), and data from these portals are separated and summarized within each submodule.

   ii. **Context**: Provides information on species habitat such as rockweed presence/absence, and EBSAs.

   iii. **Human Threats**: This section is currently under development and primarily focuses on data products of human activities (i.e., Fishing, Shipping, Miscellaneous, and Cumulative Impact Mapping; Fig. 3C) within the search area. This section also contains summary tables (under the Miscellaneous submodule; Fig 3C) which show whether or not a species was present in any of the available datasets, and which datasets they were observed in. A literature review of threats and threat mitigation for Species at Risk is also provided. Because users can interactively select the modules and submodules to be included from the Shiny App, not all sections and subsections may be included in the Full Reports (Fig. 3C).

3. **Contributors**

   This section lists and recognizes the contributions from the approximately sixty individuals who have provided data, context, and ideas for the Reproducible Reporting initiative.

4. **References**

   Here, the citations for references and datasets that were included in the Report are presented. Citation information was stored in BibTeX format and called in the R Markdown documents.

2.6.8 Community engagement

As Strategic RAP Champions, we aim to bring reproducible workflows to the forefront within DFO. We have led multiple engagement sessions to bridge siloed units, which has involved broad discussions and topics such as making data more available, and challenges associated with data security, storage and upgrades. Procedures developed as part of this project support national efforts to encourage Open Data publication, and ensure consistency and quality control of Open Data products, by providing validation and feedback to fix any errors that may be detected. For example, an error was detected and reported for the newly-incorporated Passamaquoddy Open Data record (DFO 2022a). We have also used our work as an opportunity to move data holders from outdated storage approaches to managed solutions, and in many cases, this work has served as an opportunity for data rescue. The means of communicating our work and messages have varied over time. The networking capacity in the department at a national level dramatically increased in the past years with the mobilization of the workforce to the Microsoft Teams platform, in response to the COVID-19 pandemic. This shared virtual platform has provided a unique opportunity to collaborate and deliver this project via formal and informal meetings, discussions, presentations, and chats.

The following includes specific examples where we have actively contributed and shared knowledge related to our work:

- We have coordinated and participated in activities related to the R Learning & Development community to improve approaches, share lessons learned, and transfer reproducibility skills to the broader community. For example, this has included a series of five presentations and focused break-out groups presented by inSileco to introduce interested members of DFO to Git, GitHub, R Markdown, and *csasdown*.
- Multiple presentations and discussions (30+) have been held for the past three years with various sectors (e.g., Science, Aquaculture Management, Marine Planning and Conservation, Integrated Planning, and the Species at Risk Program), including a DFO Maritimes Coffee Chat, a Fed Talk, and an hour-long presentation to the Population Ecology Division as well as the National R Learning & Development community at DFO.
- We led a two-day hackathon in February 2022 with members of the Ocean Tracking Network and DFO-Maurice Lamontagne Institute (MLI). The DFO-MLI team is interested in using reproducible reporting tools to generate general outputs from their annual at-sea missions. Further, a post-event analysis of lessons learned contributed new insights on how to improve our approach to engagement going forward.

3 REPORT USE: AUDIENCE OF THE SPATIAL REPRODUCIBLE REPORTING TOOL

3.1 Integrated Marine Response Planning (IMRP)

The DFO Science Branch provides support to lead agencies during ship source marine oil spill incidents and exercises in a number of ways. In particular, this includes the provision of scientific and technical information on environmental sensitivities present in the incident area. During the initial emergency phase of an incident response, the Integrated Marine Response Planning (IMRP; formerly Planning for Integrated Environmental Response, PIER) initiative works to develop a list of biological and ecological sensitivities in the spill area. This list of sensitivities is then incorporated into the initial departmental Resources at Risk list, used to assist in planning response strategies to mitigate environmental impact. The dynamic nature of oil spill response is such that information must flow quickly and efficiently to the appropriate responders, enabling an effective response in changing incident conditions. The Reproducible Reporting initiative has assisted the IMRP team in incorporating science information from a variety of authoritative sources.

The different phases of environmental response require varying degrees of information; the initial emergency phase requires high-level information quickly, and once the response is underway information provided can become more in-depth. With this in mind, the IMRP team may use all sections of the Reproducible Report throughout a response, and depending on trajectory modeling, will likely require a new Report to be generated. In the initial phase of the response, the Report is useful for synthesizing information on Species at Risk (or species under consideration for listing), Critical Habitats, Ecologically and Biologically Sensitive Areas (EBSAs) and Marine Protected Areas (MPAs). As the response evolves, additional focus can be placed on other species within the search area, while incorporating species group vulnerability to oil products, and considering the fate and behaviour of the spilled product in the marine environment.

Marine oil spill incidents come in many shapes and sizes, from minor to major geographic scales, and the Spatial Reproducible Reporting Tool provides the flexibility required to meet this challenge. Typically, IMRP has used the tool at a relatively small scale (approximated 5 - 50 km^2) in the near shore environment (e.g., vessel grounding), though Reports have also been requested for offshore areas (e.g., sunken vessel).

The IMRP team began using the Reports regularly in March 2021, and have relied on their information on at least ten occasions (two spill incidents, five spill exercises, and three planning processes) since then. With the development of the Spatial Reproducible Reporting Tool, IMRP is able to combine the Report with other sources of information to provide more in-depth sensitivities lists within four to five hours of a pollution report. Improvements have been noted in not only speed but also in the depth and breadth of data available to synthesize.

From an Environmental Response perspective, the Spatial Reproducible Reporting Tool could be improved to allow broader DFO individual user access and customizable Reports, as well as continued database inclusion. The inclusion of oil spill vulnerability scores in the Reports would be highly useful in environmental response, as would access to authoritative species fact sheets on species present within the search area. Overall, the continued evolution of the tool will

improve the ability to meet IMRP and DFO objectives related to Environmental Response.

3.2 Aquaculture Siting

Fisheries and Oceans Canada (DFO) Maritimes Region participates in the review of applications for proposed new and amended marine finfish aquaculture sites. These applications are submitted by industry proponents to the provinces (Nova Scotia and New Brunswick) who then engage with other provincial and federal network agency partners for review and comment. As per the Canada-New Brunswick and Canada-Nova Scotia Memorandums of Understanding on Aquaculture Development, DFO reviews submitted applications and provides advice in relation to DFO's legislative mandate.

DFO undertakes a multi-sectoral review process that especially focuses on potential impacts to fish and fish habitat. To help inform DFO's review of each application, DFO Science advice is requested on the Predicted Exposure Zones (PEZs) associated with the range of aquaculture activities, and the predicted impacts on susceptible fish and fish habitat, including sensitive Species at Risk (SAR), susceptible fishery species, and the habitats that support them.

DFO receives baseline information in the application that is collected by the proponent as required by the Aquaculture Activities Regulations (DFO 2022b). This includes a current meter record of at least thirty days from within the proposed lease. DFO Science estimates PEZs using this proponent-collected current data. PEZs are precautionary overestimates used as a tool for identifying, albeit at a larger spatial scale, areas of potential overlap with species and habitats that are sensitive to exposures of organic loading, and any fish health treatment products, if used.

The size of a PEZ is site-dependent, but has typically been on the order of hundreds of meters to kilometers away from the proposed site. While information from a fish and fish habitat survey is also submitted by the proponent as part of the baseline requirements, there are limitations to using this information as the requirements are focused on the immediate vicinity of the lease area at one point in time.

Because of this, DFO Science's participation in the review of aquaculture site applications was one of the earliest motivations behind development of the Spatial Reproducible Reporting Tool. The tool is used to search the PEZs for an efficient and consistent way of gaining an indication of species and habitats that have been observed within the PEZs, and therefore may be exposed to organic matter and/or fish health treatment products from the proposed site. To provide the advice requested of DFO Science, there is a focus on information related to SAR listed under Schedule 1 of the Species at Risk Act, fishery species, Ecologically Significant Species (ESS), and their associated habitats, as well as EBSAs. The records returned from the tool are then used to guide further consultation with species and subject matter experts about aspects such as spatial and temporal distribution in the area, uniqueness to the area, and whether or not they may be susceptible to exposures from the proposed aquaculture site.

4 DISCUSSION

In a direct response to repeat requests for data reports, we have developed a Spatial Reproducible Reporting Tool to display species data within a user-defined area of interest. By coding with R and R Markdown, in conjunction with Git and GitHub for version control, we have created this tool to increase the efficiency and transparency for science responses. This tool is intended for data-discovery, and currently summarizes over thirty sources, primarily focusing on Species at Risk, from DFO and non-DFO data. So far, approximately twenty Reports have been generated to serve these purposes. These can be rendered in under ten minutes, resulting in a quantifiable reduction of time and resources spent acquiring and manipulating data into the required format. Ultimately, this aligns with the goals of reproducible data science globally, within the Canadian Government, and DFO.

Open data tools have underpinned the success of this work, and allowed us to drastically decrease the amount of time spent creating each Report. Specifically, R and R Markdown have been revolutionary by allowing us to move from outdated approaches of data collection, management, and visualization, to more modernized timesaving “digital” methods (Lowndes et al. 2017). Previously, data were stored in multiple locations, and we were often unaware of important datasets. Spending time compiling this data, and making this code available for repeat requests, helps mitigate against the continued challenges of shifting or evolving teams (Wilson et al. 2014), and instead helps preserve institutional memory (Wilson et al. 2017). A supportive online community and other coders within DFO have greatly contributed to the advancement of this work, as we are made aware of new methods and approaches to improve our workflows. The adoption of Git and GitHub was also key to continue growing and expanding this work. Throughout the years, the Spatial Reproducible Reporting Tool has been through multiple adaptations, both large and small. Using a version control system allowed team members to code more collaboratively, make changes, test new algorithms, explore better visualization, and optimize data storage methods. Importantly, the version control system allows the team to revert to a previous version if necessary. Continued advancement of this initiative and a more widespread adoption of reproducible tools will lead to increased time-saving opportunities within DFO, and ultimately adds to the cumulative knowledge in the scientific process (Wolkovich et al. 2012; McKiernan et al. 2016; Boland et al. 2017).

Computing skills are increasingly important in science (Boettiger et al. 2015; Baker 2016), yet many scientists are not formally trained in these practices (Lowndes et al. 2017; Daniel 2019). A lack of exposure and confidence in their abilities can often cause individuals to be resistant or hesitant to incorporate these tools into their work (Lowndes et al. 2017). This has several important consequences for DFO. In the workplace, scientists often do not have the time to learn in-depth computing skills, beyond focusing on immediate problems (Ram 2013). DFO is increasingly offering more opportunities for professional development, and we see significant value in building a culture where these skills are valued, taught and practiced. For example, in 2022, the R Learning and Development Series organized lectures for DFO Maritimes employees, run by inSileco, to cover topics of open data science, which included smaller breakout groups for hands-on practice with these tools. Often, there are many tangible benefits of using a few simple commands for tasks such as transferring code, organizing versions of files, and collaborating in groups (Blischak et al. 2016). Introducing reproducible and open workflows will likely not yield immediate results or benefits (Janssen et al. 2012), but instead is an incremental practice

requiring patience, motivation, and sustained effort (Peng 2011; Wilson et al. 2017).

While we have achieved our initial objective of creating automated data-discovery Reports, we have faced several obstacles in developing the tool, and expanding this initiative. In particular, our team has frequently confronted challenges related to scope creep, which is routinely identified as one of the most common reasons for failure in software project management (Bjarnason et al. 2012; Kumari and Pillai 2013; Komal et al. 2020). We have been highly encouraged by the feedback and willingness to collaborate, yet each discussion brings new questions and ideas, and we are often pulled in multiple different directions to accommodate these requests. To address these concerns and ensure that our work is truly helpful to the DFO community, our focus on co-creation remains key for prioritization and accountability. That said, we also value the continued advancement of new skills, intrinsic curiosity, and the pursuit of new opportunities, and this has been a major factor in our success in overcoming various obstacles. We acknowledge that there is no perfect balance for managing these competing and conflicting tasks and interests, but this can be at least partially remedied by continuous scope prioritization and close cooperation within our team (Bjarnason et al. 2012). We believe there are multiple opportunities to both expand the capacity of this tool and broadly encourage reproducible workflows, yet acknowledging and addressing issues related to scope creep remains important when pursuing this initiative.

The data revolution is drastically transforming government (Abiteboul and Stoyanovich 2019; SCC 2021), yet data management often remains a hindrance to productive use (Tenopir et al. 2011; Daniel 2019). Gomez et al. (2021) noted there are many challenges within DFO for spatial data analysis such as no central repository for spatial data, a lack of infrastructure, and limited data documentation. This tool helps us achieve some of these objectives in support of data discovery, and align efforts for open data policies globally (Murray-Rust 2008; Kitchin 2014; Culina et al. 2018), within the Canadian Government (PCO 2018; SCC 2021), and with DFO's Data Strategy (DFO 2020a). However, despite the various political, economic, social, operational, and technical benefits to making data open and reproducible (Murray-Rust 2008; Janssen et al. 2012), the barriers and challenges of this warrant further attention. For example, Janssen et al. (2012) reviewed the myths of open data and open government, and listed over fifty potential barriers to publicizing data including task complexity, institutional make-up, and technical challenges. While there are data managers and custodians to pursue much of this work, such as the Marine Spatial Planning program and the Enterprise Data Hub (EDH; built as part of the Target Architecture for Data and Application Platform, TADAP), as the scientists who use, manage, and collect these data, we realized it was increasingly important to take a proactive approach to reduce silos and increase communication between groups. We have attempted to lead discussions to make data sharing more interoperable, free and open, connect various groups, and avoid duplication of effort to address shared concerns. Ultimately, data management is a large task and requires participation at all levels to ensure we are making the best use of the available resources.

DFO collects a wealth of data, yet in various instances, datasets not initially intended for spatial analysis contain useful information for our Report to strengthen decision-making. Our work focuses on automation, and these datasets require significant data wrangling to identify, extract, and integrate this information into a usable form (Kandel et al. 2011; Furche et al. 2016). For example, funding agencies often focus on threats facing Species at Risk as part of their research priorities. There is great potential for this tool to identify and provide this threat information

alongside the Species at Risk observed within a search area, and we have started this work within our Human Activities module. However, many of the documents mentioned in the SARA recovery planning process contain threat information with various levels of detail. Threat information can be mentioned in paragraphs, tables, sentences, or point form, and range from highly qualitative to quantitative, and these documents have gone through multiple generations of changes to further refine and define how to characterize threats (DFO 2007, 2010, 2014). We are currently in discussion with our collaborators for ways to best summarize threat information so it is comprehensive, current, and accurate, to achieve these goals. We have also encountered additional challenges with non-standardized formats when undergoing an additional data mining and wrangling exercise to extract Species at Risk data from Section 73 permits which are used when a human activity affects listed wildlife species (e.g., developments, research purposes; SARA s.73, 2002). They often contain a mix of digital and hand-written documentation, have coordinate information in multiple formats, or in several instances, provide limited to no spatial context. Other sectors have recognized these challenges and are in discussion to standardize these approaches in future years. Advancements in big data and reproducible approaches cause excitement, yet there are stark warnings that irresponsible data wrangling within reproducible workflows can be misleading or draw incorrect conclusions (Leek and Peng 2015). Aware of these challenges, we continue to be careful, measured, and consistently consult species experts to guide our work, and provide credible information without losing important context.

There are many potential future directions and opportunities within DFO to pursue this work through the continued development of the Spatial Reproducible Reporting Tool, and its corresponding influence on data practices in the department. We have presented only a single application of the reproducible analytical pipelines developed in this work. The highly agile workflows created for data wrangling and processing could allow more complex issues to be addressed with reduced overhead, such as multi-species and multi-threat approaches, and more tailored aquaculture siting or environmental response reporting. We have other grand aspirations such as expanding this work into other regions in DFO, creating version(s) of the App that can be made available to the public, introducing these workflows to freshwater ecosystems, adding Aquatic Invasive Species data, and incorporating the widespread co-creation and code review of the Reproducible Reports. These endeavors become increasingly feasible as the data pipelines become more streamlined; however, this project is still heavily dependent on funding streams each year, which will influence our future directions. Beyond working within the confines of the existing workflows, their modularity allows individual components to be rapidly exported into other projects. Ongoing development of R packages to encompass widely applicable sections of this work can enhance the scope of pipelines across DFO and beyond.

5 CONCLUSIONS

This technical report described a successful, innovative and cross-sectoral initiative for providing consistent and efficient DFO and non-DFO information relevant to decision-making. Our motivation is driven by tangible examples of peer-support, collaboration and workflow efficiency: reports assembled in the past using traditional approaches took weeks to be created, in comparison, this tool reduced the time to minutes, with a broader set of contextual information. This work has significantly increased reporting quality and efficiency for staff who generate/provide information, assemble reports with information from a variety of sources, and are accountable for providing information and advice in a timely fashion. In support of traceable, open science, and digital open government, this spatial tool includes a record of the underlying data analytics and provenance through curated version control (i.e., Git). The growing partnership of Strategic RAP Champions, collaborators, and advisors will continue to break silos and will expand as long as we can continue to foster growth, opportunities, peer-support, and training. We are committed to continue bringing partners together to support the strategic vision and priorities of DFO, aligned with individual research interests.

6 ACKNOWLEDGMENTS

The Marine Spatial Planning (2019-2021), Nature Legacy (2020-2022), and SARA programs (2021-2022) have supported the Spatial Reproducible Reporting initiative in the Maritimes Region. This initiative would not have been possible without the inspiration, advice, and support from multiple colleagues that have contributed to different stages of this project, including (listed in alphabetical order): Amanda Babin, Andrea Morden, Angelia Vanderlaan, Caroline Bakelaar, Brent Law, Brian Bower, Christine Stortini, Clark Richards, Colin O'Neil, David Fishman, Dan Ricard, Diane Amirault-Langalis, Emma Marotte, Gregory Puncher, Heath Stone, Heather Bowlby, Hilary Moors-Murphy, Javier Murillo, Koren Spence, Laura Feyrer, Nancy Shackell, Nick Jeffery, Pamela Emery, Paul Regular, Phil Greyson, Ryan Stanley, Sarah Tuziak, Shelley Lang, Susan Heaslip, and Tanya Pelrine. Great appreciation is addressed to (in alphabetical order) Biljana Narancic, Bruce Delo, Caroline Vanier, Claude Nozères, Jonathan Pye, and Virginie Roy for the helpful discussions and ideas generated during the various hackathon sessions. Additional thanks are addressed to reviewers Jake Coates and Mike McMahon for their time, insightful edits, and suggestions for this Technical Report. We would also like to specifically thank Tana Worcester for her guidance, encouragement, and unending support throughout all stages of this initiative.

7 REFERENCES

Abiteboul, S., and Stoyanovich, J. 2019. Transparency, fairness, data protection, neutrality: Data management challenges in the face of new regulation. J. Data Inf. Qual. 11(3): 1–9.

Anderson, S.C., Grandin, C., Edwards, A.M., Grinnell, M.H., Ricard, D., and Haigh, R. 2022. Csasdown: Reproducible CSAS reports with bookdown.

Baker, M. 2016. 1,500 scientists lift the lid on reproducibility. Nature 533(7604): 452–454.

Baumann, P., Mazzetti, P., Ungar, J., Barbera, R., Barboni, D., Beccati, A., Bigagli, L., Boldrini, E., Bruno, R., Calanducci, A., Campalani, P., Clements, O., Dumitru, A., Grant, M., Herzig, P., Kakaletris, G., Laxton, J., Koltsida, P., Lipskoch, K., Mahdiraji, A.R., Mantovani, S., Merticariu, V., Messina, A., Misev, D., Natali, S., Nativi, S., Oosthoek, J., Pappalardo, M., Passmore, J., Rossi, A.P., Rundo, F., Sen, M., Sorbera, V., Sullivan, D., Torrisi, M., Trovato, L., Veratelli, M.G., and Wagner, S. 2016. Big Data Analytics for Earth Sciences: the EarthServer approach. Int. J. Digit. Earth 9(1): 3–29.

Bjarnason, E., Wnuk, K., and Regnell, B. 2012. Are you biting off more than you can chew? A case study on causes and effects of overscoping in large-scale software engineering. Inf. Softw. Technol. 54(10): 1107–1124.

Blischak, J.D., Davenport, E.R., and Wilson, G. 2016. A Quick Introduction to Version Control with Git and GitHub. PLoS Comput. Biol. 12(1).

Boettiger, C., Chamberlain, S., Hart, E., and Ram, K. 2015. Building software, building community: lessons from the rOpenSci project. J. Open Res. Softw. 3(1).

Boland, M.R., Karczewski, K.J., and Tatonetti, N.P. 2017. Ten Simple Rules to Enable Multi-site Collaborations through Data Sharing. PLoS Comput. Biol. 13(1): e1005278.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. 2021. shiny: Web Application Framework for R.

Choi, J.S., Vanderlaan, A.S.M., Lazin, G., McMahon, M., Zisseron, B., Cameron, B., and Munden, J. 2018. St. Anns Bank Framework Assessment. DFO Can. Sci. Advis. Sec. Res. Doc. 2018/066. vi + 65 p.

Culina, A., Baglioni, M., Crowther, T.W., Visser, M.E., Woutersen-Windhouwer, S., and Manghi, P. 2018. Navigating the unfolding open data landscape in ecology and evolution. Nat. Ecol. Evol. 2(3): 420–426.

Daniel, B.K. 2019. Big Data and data science: A critical review of issues for educational research. Br. J. Edu. Techol. 50(1): 101–113.

DFO. 2007. Revised Protocol for Conducting Recovery Potential Assessments. DFO Can. Sci. Advis. Sec. Sci. Advis. Rep. 2007/039.

DFO. 2010. Guidelines for Terms and Concepts Used in the Species at Risk Program. DFO Can. Sci. Advis. Sec. Sci. Advis. Rep. 2010/065.

DFO. 2014. Guidance on Assessing Threats, Ecological Risk and Ecological Impacts for Species at Risk. DFO Can. Sci. Advis. Sec. Sci. Advis. Rep. 2014/065.

DFO. 2018. FINAL EVALUATION REPORT - Evaluation of the Ocean Management Program. Project number 6D014.

DFO. 2020b. Evaluation of Economic Analysis and Statistics: Final Report. Project number 96261.

DFO. 2020a. Data Strategy: Fisheries and Oceans Canada.

DFO. 2021. DFO Maritimes Region Review of the Proposed Marine Finfish Aquaculture Boundary Amendment, Whycocomagh Bay, Bras d'Or Lakes, Nova Scotia. DFO Can. Sci. Advis. Sec. Sci. Resp. 2021/041.

DFO. 2022b. Aquaculture Activities Regulations: SOR/2015-177.

DFO. 2022a. Passamaquoddy Bay Biodiversity Trawl [Dataset].

Doubleday, W.G., Atkinson, D.B., and Baird, J. 1997. Comment: Scientific inquiry and fish stock assessment in the Canadian Department of Fisheries and Oceans. Can. J. Fish. Aquat. Sci. 54(6): 1422–1426.

Edwards, A.M., Duplisea, D.E., Grinnell, M.H., Anderson, S.C., Grandin, C.J., E. A. Keppel, D.R. abd, Anderson, E.D., Baker, K.D., Benoît, H.P., Cleary, J.S., Connors, B.M., Desgagnés, M., English, P.A., Fishman, D.J., Freshwater, C., Hedges, K.J., Holt, C.A., Holt, K.R., Kronlund, A.R., Mariscak, A., Obradovich, S.G., Patten, B.A., Rogers, B., Rooper, C.N., Simpson, M.R., Surette, T.J., R. F. Tallman, R.F., Wheeland, L.J., Wor, C., and Zhu, X. 2018. Proceedings of the Technical Expertise in Stock Assessment (TESA) national workshop on 'Tools for transparent, traceable, and transferable assessments,' 27-30 November 2018 in Nanaimo, British Columbia. DFO Can. Sci. Advis. Sec. Res. Doc. 3290: v + 10 p.

Farley, S.S., Dawson, A., Goring, S.J., and Williams, J.W. 2018. Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions. BioScience 68(8): 563–576.

Furche, T., Gottlob, G., Libkin, L., Orsi, G., and Paton, N. 2016. Data wrangling for big data: Challenges and opportunities. *In* Advances in database technology-EDBT 2016: Proceedings of the 19th international conference on extending database technology. pp. 473–478.

GBIF. 2022. What is GBIF? The Global Biodiversity Information Facility.

Git. 2022. Git Version Control System.

GitHub. 2022. A Collaborative Online Platform To Build Software.

Gomez, C., Nephin, J., Lang, S., Feyrer, L., Keyser, F., and Lazin, G. 2021. Spatial Data, Analysis and Modelling Forums: An initiative to broaden the collaborative research potential at DFO. Can. Tech. Rep. Aquat. Sci. 3416: v + 36 p.

Gomez, C., Sameoto, J.A., Keyser, F., Regular, P., Beazley, L., Layton, C., Richards, C., Stanley, R., Tam, J., Ferguson, K., and Kraska, P. 2020. Proceeding of the Interactive Tools for Science Advice Workshop Series in Maritimes Region, November-December, 2019. Can. Tech. Rep. Aquat. Sci. 3369: v + 20 p.

Graul, C. 2016. leafletR: Interactive web-maps based on the leaflet JavaScript library.

Heesen, R. 2018. Why the reward structure of science makes reproducibility problems inevitable. J. Philos. 115(12): 661–674.

Janssen, M., Charalabidis, Y., and Zuiderwijk, A. 2012. Benefits, Adoption Barriers and Myths of Open Data and Open Government. Manag. Inf. Syst. 29(4): 258–268.

Jasny, B., Chin, G., Chong, L., and Vignieri, S. 2011. Again, and Again, and Again .... Science 334(6060): 1225.

Jia, L., Yao, W., Jiang, Y., Li, Y., Wang, Z., Li, H., Huang, F., Li, J., Chen, T., and Zhang, H. 2022. Development of interactive biological web applications with R/Shiny. Brief. Bioinformatics 23(1): bbab415.

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., Ham, F. van, Riche, N.H., Weaver, C., Lee, B., Brodbeck, D., and Buono, P. 2011. Research directions in data wrangling: Visualizations and transformations for usable and credible data. Inf. Vis. 10(4): 271–288.

Kaya, E., Agca, M., Adiguzel, F., and Cetin, M. 2019. Spatial data analysis with R programming for environment. Hum. Ecol. Risk Assess. 25(6): 1521–1530.

Kelley, D.E. 2018. Oceanographic analysis with R. Springer.

Kitchin, R. 2014. The data revolution: Big data, open data, data infrastructures and their consequences. Sage.

Knuth, D.E. 1984. Literate programming. The Computer Journal 27(2): 97–111.

Komal, B., Janjua, U.I., Anwar, F., Madni, T.M., Cheema, M.F., Malik, M.N., and Shahid, A.R. 2020. The impact of scope creep on project success: An empirical investigation. IEEE Access 8: 125755–125775. IEEE.

Kumari, S.N., and Pillai, A.S. 2013. A survey on global requirements elicitation issues and proposed research framework. *In* 2013 IEEE 4th int. Conf. Soft. Eng. Serv. sci. pp. 554–557.

Lai, J., Lortie, C.J., Muenchen, R.A., Yang, J., and Ma, K. 2019. Evaluating the popularity of R in ecology. Ecosphere 10(1): e02567.

Leek, J.T., and Peng, R.D. 2015. Opinion: Reproducible research can still be wrong: Adopting a prevention approach. Proc. Natl. Acad. Sci. 112(6): 1645–1646.

Lowndes, J.S.S., Best, B.D., Scarborough, C., Afflerbach, J.C., Frazier, M.R., O'Hara, C.C., Jiang, N., and Halpern, B.S. 2017. Our path to better science in less time using open data science tools. Nat. Ecol. Evol. 1(6): 160.

Malde, K., Handegard, N.O., Eikvil, L., and Salberg, A.-B. 2020. Machine intelligence and the data-driven future of marine science. ICES J. Mar. Sci. 77(4): 1274–1285.

McKiernan, E.C., Bourne, P.E., Brown, C.T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B.A., Ram, K., and Soderberg, C.K. 2016. Point of view: How open science helps researchers succeed. eLife 5: e16800.

Munafò, M.R., Chambers, C.D., Collins, A.M., Fortunato, L., and Macleod, M.R. 2020. Research Culture and Reproducibility. Trends Cogn. Sci. 24(2): 91–93.

Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J.J., and Ioannidis, J.P.A. 2017. A manifesto for reproducible science. Nat. Hum. Behav. 1(1): 21.

Murray-Rust, P. 2008. Open data in science. Nat. Preced.: 1.

OBIS. 2022. Intergovernmental oceanographic commission of UNESCO. Ocean Biodiversity Information System.

Obradović, S. 2019. Publication pressures create knowledge silos. Nat. Hum. Behav. 3(10): 1028.

Open Data. 2022. Government of Canada: Open Data.

PCO. 2018. Report to the clerk of the privy council: A data strategy roadmap for the federal public service. Privy Council Office.

Pebesma, E. 2018. Simple Features for R: Standardized Support for Spatial Vector Data. R J. 10(1): 439–446.

Peng, R.D. 2011. Reproducible Research in Computational Science. Science 334(6060): 1226–1227.

Perez-Riverol, Y., Gatto, L., Wang, R., Sachsenberg, T., Uszkoreit, J., Leprevost, F. da V., Fufezan, C., Ternent, T., Eglen, S.J., Katz, D.S., Pollard, T.J., Konovalov, A., Flight, R.M., Blin, K., and Vizcaíno, J.A. 2016. Ten Simple Rules for Taking Advantage of Git and GitHub. PLoS Comput. Biol. 12(7): e1004947.

Provoost, P., and Bosch, S. 2021. Robis: Ocean biodiversity information system (OBIS) client.

PWGSC. 2021. Security levels for sensitive government information and assets. Public Works and Government Services Canada.

Quarto. 2022. Welcome to Quarto.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ram, K. 2013. Git can facilitate greater reproducibility and increased transparency in science. Source Code Biol. Med. 8(1): 7.

Regular, P.M., Robertson, G.J., Rogers, R., and Lewis, K.P. 2020. Improving the communication and accessibility of stock assessment using interactive visualization tools. Can. J. Fish. Aquat. Sci. 77(9): 1592–1600.

Ricard, D., and Gomez, C. 2021. Maritimes-SUMMER-atlas.

Ricard, D., and Shackell, N.L. 2013. Population status (abundance/biomass, geographic extent, body size and condition), important habitat, depth, temperature and salinity of marine fish and invertebrates on the Scotian Shelf and Bay of Fundy (1970-2012). Can. Tech. Rep. Aquat. Sci. 3012: viii + 180 p.

Sandve, G.K., Nekrutenko, A., Taylor, J., and Hovig, E. 2013. Ten Simple Rules for Reproducible Computational Research. PLoS Comput. Biol. 9(10): e1003285. Public

Library of Science.

SARA. 2002. An act respecting the protection of wildlife species at risk in canada. Species at Risk Act.

SCC. 2021. Canadian data governance standardization roadmap. Standards Council of Canada.

Staples, T.L., Dwyer, J.M., Wainwright, C.E., and Mayfield, M.M. 2019. Applied ecological research is on the rise but connectivity barriers persist between four major subfields. J. Appl. Ecol. 56(6): 1492–1498.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., and Frame, M. 2011. Data Sharing by Scientists: Practices and Perceptions. PLoS One 6(6): e21101.

Tippmann, S. 2015. Programming tools: Adventures with R. Nature 517(7532): 109–110.

Ushey, K. 2022. Renv: Project environments.

Ushey, K., Allaire, J., and Tang, Y. 2022. Reticulate: Interface to 'python'.

Wilson, G., Aruliah, D.A., Brown, C.T., Chue Hong, N.P., Davis, M., Guy, R.T., Haddock, S.H.D., Huff, K.D., Mitchell, I.M., Plumbley, M.D., Waugh, B., White, E.P., and Wilson, P. 2014. Best Practices for Scientific Computing. PLoS Biol. 12(1): e1001745.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., and Teal, T.K. 2017. Good enough practices in scientific computing. PLoS Comput. Biol. 13(6): e1005510.

Wolkovich, E.M., Regetz, J., and O'Connor, M.I. 2012. Advances in global change research require open science by individual researchers. Glob. Change Biol. 18(7): 2102–2110.

Wright, A.M., Schwartz, R.S., Oaks, J.R., Newman, C.E., and Flanagan, S.P. 2019. The why, when, and how of computing in biology classrooms. F1000research 8: 1854.

Xie, Y. 2017. Dynamic Documents with R and knitr. Chapman; Hall/CRC.

Xie, Y., Allaire, J.J., and Grolemund, G. 2018. R markdown: The definitive guide. Chapman; Hall/CRC.