Canadian Integrated Ocean Observing System

Investigative Evaluations
Cyberinfrastructure
2017-11-30

Principal Investigators

Richard Kelly[1], Mike Smit[2]

Project Manager:

Shayla Fitzsimmons[2]

Contributors

Scott Bruce[1]
Craig Bulger[1]
Brad Covey[3]
Richard Davis[3]
Ryan Gosse[3]
Dwight Owens[4]
Benoit Pirenne[4]

Affiliations

[1]Fisheries and Marine Institute of Memorial University of Newfoundland
[2]School of Information Management, Dalhousie University
[3]Department of Oceanography, Dalhousie University
[4]Ocean Networks Canada, University of Victoria

Marine Institute
Memorial University
Dalhousie University
University of Victoria

Executive Summary

Numerous countries have employed a coordinated network of government agencies, research institutions, and private companies to establish national integrated Ocean Observing Systems (OOSes). Although Canada boasts a robust and diverse ocean economy, the country has implemented no such network.

To better adapt in the face of a changing environment and to assist the country in meeting national and international commitments, Fisheries and Oceans Canada (DFO) has commissioned investigative evaluations (IEs) to determine the cost and feasibility of creating a Canadian Integrated Ocean Observing System (CIOOS). This report contains the recommendations of the Cyberinfrastructure IE, and outlines three models, low, moderate and high, with varying levels of service.

To determine an appropriate cyberinfrastructure configuration for CIOOS, information was gathered from both national and international sources. Systems and standards were evaluated, stakeholders surveyed, and existing international OOSes consulted to identify potential limits or gaps to the implementation of CIOOS.

Low Service
To minimize the effort required from Canadian institutions to become interoperable with CIOOS, it is recommended that open standards be used instead of specific software when possible. The well-supported Open GeoSpatial Consortium standard Catalogue Service for the Web (CSW) is recommended for cataloguing, with Web Accessible Folders (WAFs) as a supplement to facilitate data harvesting and circumvent the deficiencies of CSW. No standard exists for catalogue aggregation; CIOOS may instead develop a tool in-house or employ an existing aggregator.

Data dissemination and aggregation is a situation for which a software is recommended over a standard. The ERDDAP implementation of OPeNDAP is the sole data server with native federation capabilities, and is recommended for use in CIOOS. ERDDAP also includes other useful features such as data serving for a wide range of file formats, eliminating the need for a standard file format.

Compute Canada is the preferred hardware provider for CIOOS. Given uncertainty regarding required resources, it is recommended that the organization's existing stack be utilized during the initial phases, and hardware purchased through Compute Canada once more is known about storage and processing needs. Several cores with 4 GB of RAM is sufficient for the low service model. Data ingestion assistance is not provided to data providers in this model.

Moderate and High Service
Moderate and high service models are difficult to distinguish, as the core system is necessary and remains unchanged. A model offering higher service may implement additional standards

as needed, but the main difference between models will be the amount of support given to data providers for submission of their data to CIOOS. Hardware at higher service models includes virtual machines to assist data providers with intensive data manipulation required to make their data CIOOS-compliant, and 256 to 512 GB of RAM to provide performant visualizations through caching.

The service model provided by CIOOS will be determined at a later date. But even the low service model is sufficient for a robust and flexible system which will allow Canada to become a rising star in the world of oceanography.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

Canada is an ocean nation. Its extensive coastline of 244 000 km, the largest of any nation on earth, spans from the temperate North Pacific Ocean, through the Arctic, and down the Atlantic seaboard to the USA. The Gulf of St. Lawrence and Hudson's Bay, two of the earth's great inland seas, are wholly contained within Canada's land mass. The Great Lakes are another shared coastline between Canada and the USA.

About 40% of the Canadian population lives within 100 km of these coastlines (Manson et al. 2005). Both historically and currently, Canadians have turned to the ocean for their livelihoods and well-being. Canada's ocean economy is diverse, and includes transportation, offshore energy, marine technology, defense, tourism, conservation and fisheries. The ocean economy accounts for about $26 billion, approximately 5% of Canada's annual GDP, and provides employment to more than 315 000 workers (DFO 2009). The marine environment and the Great Lakes are also critically important for Indigenous People's subsistence, social and ceremonial uses, and are the backbone of the socioeconomic well-being of Canada's coastal communities.

Advances in marine technology are providing unprecedented access to the ocean and are spawning a myriad of new economic and scientific activities. New, well-paying employment opportunities will bring many more people out to work on the ocean as this "Blue economy", or "ocean industrial revolution" (McCauley et al. 2015), accelerates. It will also add pressure to the ocean systems that provide essential ecosystemic services and that support the existing fisheries, tourism, and other sectors that are major engines of the Canadian economy. To understand and sustainably manage this development, Canada needs an ocean observing capacity that will provide integrated information needed for high-quality research as to inform policy management decisions.

The ocean drives planetary systems such as weather and water cycles, and while the environmental characteristics and fauna of the ocean may differ considerably among regions, the ocean is still an interconnected whole, as exemplified in the One Ocean concept (O'Dor et al. 2009). What happens in one part of the global ocean can have important impacts on other, distant regions. The species on which our fisheries depend are mobile and not constrained by national borders. Interconnectivity applies to environmental threats such as oil spills, invasive species, or rising sea level, temperature, and acidity. Consequently, humans have the shared global burden to provide the ocean science needed to plot a sustainable future. One mechanism that coastal nations are addressing this challenge is by signing international agreements to collect and exchange ocean data and knowledge, and to mutually address shared problems.

In Canada, ocean science is conducted by government, academia, industry, non-governmental organizations (NGOs), and the general public through citizen science. Fisheries and Oceans Canada (DFO) has by far the greatest investment and capacity for ocean science; while DFO's science sector pursues fundamental science, it is responsible of stock assessment for best

fisheries management, as well as providing advice in support of its other programs related to ocean protection, such as marine protected areas and species at risk and aquaculture. Its work also supports development of economic opportunities, such as aquaculture, and guides operations including search-and-rescue for the Coast Guard. The Canadian government is strongly committed to ocean science, as evidenced by the Mandate Letter issued by the government for the Minister of Fisheries and Oceans and the Canadian Coast Guard, which directs the Minister to:

> *(1) "Restore funding to support federal ocean science and monitoring programs",*
>
> *(2) "Ensure that decisions are based on science, facts, and evidence, and serve the public interest", and*
>
> *(3) "Work with the provinces, territories, Indigenous Peoples, and other stakeholders to better co-manage our three oceans."*

In parallel and supported by Canada's national academic funding agencies, our university and college sectors also have strong capabilities in ocean science. Academics undertake a variety of research, ranging from individual investigator, narrow-focus, short-term projects to large national networks that have the supporting infrastructure to sustain interdisciplinary research and the associated data management for longer (~ 5 years) periods.

Industry, Indigenous governments, NGOs and the public undertake more limited ocean research which is generally tied to specific interests or values of their organization. Many of these programs frequently address issues in which the public has a strong interest. With the advent of user-friendly ocean observing sensors, these groups can rapidly generate large volumes of high-quality data from geographic areas of great interest.

To meet efficiently Canada's needs in ocean science, it is essential that Canadian investigators from all sectors coordinate their data collection efforts to avoid duplication or lost opportunities, and ensure that data collected is discoverable, usable and shareable by Canadians to the benefit of all Canadians. This issue, highlighted in the Mandate Letter to the Minister of Fisheries and Oceans, was the subject of two reports commissioned by DFO and its partners (DFO 2010, OSTP 2011) and was a key finding of two reports issued by the Canadian Council of Academies (CCA) which examined the Canadian Ocean Science Sector (CCA 2012; Expert Panel on Canadian Ocean Sciences 2013). In addition, CCA (2013) identified 40 priority ocean sciences questions for Canada. Of those, two questions specifically addressed Canada's ocean information needs:

> *#24 How can a network of Canadian ocean observations be established, operated and maintained to identify environmental change, and its impacts?*

#25 What indicators are available to assess the state of the ocean, what is the significance of changes observed in those indicators, and what additional indicators need to be deployed?

The Expert Panel on Canadian Ocean Science (2013) examined how Canadian ocean science research is currently structured and concluded that the country faced three primary gaps:

Vision
Canada lacks a national vision and strategy for the oceans.

Coordination
We need to pool efforts from the local to the international scale to address our ocean science needs

Information
We lack information about the scale and scope of ocean research being carried out nationally, and on the availability and comparability of our existing research activity and of the data being generated.

Canada requires a coordinated integrated ocean observing system to meet the national ocean information needs of government, academia, industry, and the public. Such a system will directly support our international ocean commitments, and permit Canada to play a global leadership role in multidisciplinary ocean science. An ocean observing system it will help coordinate the collection of ocean data, be capable of adaptation in the face of changing needs and a changing environment, and will provide access to data currently not discoverable, especially the extensive holdings of the federal government.

The international and national context both offer favourable conditions for the establishment of a national ocean observing system in Canada. Internationally, a growing number of countries and organizations worldwide have well-established ocean observing systems. Canada's positive global reputation has us well-positioned to sustain our engagement in international efforts (UNCLOS, ESPOO, CBD, OSPAR, MARPOL 73/78, GEOOS, GOOS, etc.). Nationally, the amount of information and data generated by Canada's existing ocean observing assets distributed across the country (provincial and federal ministries, research organizations, universities, Indigenous Nations, NGOs, etc.), is already considerable and provides a solid foundation for establishing regional associations within an overarching Canadian Integrated Ocean Observing System (CIOOS) to address Canada's national priorities.

Such a system will require engaging in pan-Canadian efforts to achieve shared standards and practices among the existing organizations (Wilson et al., 2016). Each operates at its own level of sustainability, maturity, scope, and funding and will require investment and support in different areas. For example, the St. Lawrence Global Observatory, established in 2005 by a network of provincial and federal department and universities, integrates multidisciplinary

data from multiple partners, and in many ways, is a model for future regional associations. To this end, in 2016, Fisheries and Oceans Canada re-initiated a consultation process with stakeholder groups across the country to continue past discussions and move forward with the creation of CIOOS. In 2017, it commissioned three Investigative Evaluations (IEs) to make recommendations regarding the structure of a national observing system. The three IEs addressed issues within the topics of *Data and Observations*, *Visualization*, and *Cyberinfrastructure*.

To determine an appropriate cyberinfrastructure configuration for CIOOS, the Cyberinfrastructure (CI) Investigative Evaluation (IE) collected information, connected with stakeholders, and evaluated possible paths for building on extant platforms and expertise. Current best practices – both nationally and internationally – were established, and numerous systems and standards were evaluated to provide recommendations in the following key areas:

1. **Estimation of Software Requirements**: Software infrastructure was examined for strengths and deficiencies in light of the requirements provided by the other IEs. Necessary components were determined to be systems for data cataloguing, data aggregation and serving, metadata, and visualization. Compliance with national and international interchange standards was also considered as a key factor.

2. **Estimation of Hardware Requirements**: Consultations with existing hardware infrastructure providers were carried out to ascertain commercially available options and associated costs. Through collaboration with other IEs, the hardware requirements – including processing power and storage volume – were determined.

3. **Phased Approach**: Given the large undertaking required for implementation of CIOOS, a phased approach is recommended, and the necessary steps are elucidated.

These requirements and recommendations were informed by continued discussions with the other IEs. Given the complexity of ocean data, the IEs have developed three tiers of service in which there are variations regarding available tools and level of support. Even at the low service model, the listed recommendations are for a robust and flexible system which will allow Canada to meet future oceanographic challenges and adapt to the changing economic, societal, and research needs of the country.

2.0 National and International Information Gathering

To inform our recommendations, we undertook extensive information gathering. We used surveys, web content collection and analysis, interviews, and white papers to inform our understanding of the national and international experience with existing data, software, and interoperability standards. In addition to learning from international experience as we build a national system, this system must also comply with international standards.

Ocean observing systems currently exist in numerous countries. These observing systems were explored to determine their structure – with particular focus on the degree of centralization and data, metadata, and interchange standards utilized – in order to obtain insights into the various schemas available to CIOOS. Consultations with other ocean observing systems (OOSes) also provided lessons learned, which identified potential gaps, limits, and barriers to the implementation of a national integrated OOS in Canada.

The following sections will review the standards, the software systems, the structure of other ocean observing systems, how data are submitted to those systems, and lessons learned through consultations with the oceans community. The perspective will be both national and international.

2.1 Existing Data Exchange Standards

Exchange standards allow for the sharing of metadata in a standardized format and facilitate the sharing of interoperable data. There exist a number of standards and standards bodies which develop these, and many are focused on geospatial data. These include the Open Geospatial Consortium (OGC), the Internet Engineering Task Force (IETF), the International Standards Organization (ISO), and several others. Given the significant number of current standards, it is not feasible to provide an exhaustive overview. Instead, those which are either common or relevant to the ocean observing community were selected, and are explored within the following subsections.

2.1.1 Open Geospatial Consortium (OGC)

The Open Geospatial Consortium is a not-for-profit organization committed to the development of open standards. In recent years, OGC has collaborated with other standards bodies – such as ISO and the World Wide Web Consortium (W3C) – to increase the ease with which geospatial data may be shared. It is through the use of open standards that geospatial data can be interoperable between different software platforms from different vendors.

2.1.1.1 Catalogue Service for the Web (CSW)

CSW is a standard for the transportation of geospatial metadata records in eXtensible Markup Language (XML) over the internet. CSW is capable of transmitting metadata records in several different formats, including Dublin Core, Federal Geographic Data Committee (FGDC) and ISO 19115/19139. This standard is utilized worldwide to provide interoperability between metadata catalogues and to allow automatic harvesting of metadata from different catalogues. The use of CSW for transmitting metadata is considered a best practice; it is supported by all of the catalogue software discussed in Section 2.2.1.

2.1.1.2 Web Mapping Service (WMS)

WMS is a standard for serving georeferenced maps over the internet as a grid of static images. The full map is downloaded and visualized in client side Geographic Information System (GIS) software, data processing, or other visualization tools. WMS is a widely supported

OGC standard in both open source and commercial GIS software, and it allows interoperable map generation from a variety of sources.

Alone, WMS provides limited interactivity with the data it utilizes to generate maps. It is complemented by the Web Feature Service (WFS) and Web Coverage Services (WCS) to provide a deeper and more comprehensive view into the underlying data.

2.1.1.3 Web Feature Service (WFS)/Transactional Web Feature Service (WFS-T)

WFS and WFS-T are standards for querying and serving discrete feature data over the internet based primarily on spatial constraints. Unlike WMS, which returns pre-rendered images, WFS makes available the underlying data to be manipulated on the client side in the form of Geography Markup Language (GML) and tends to focus on vector data with well-defined boundaries. WFS-T is a transactional form of WFS that allows the creation, modification, and deletion of features using the WFS standard.

The provision of interoperable remote access through WFS(-T) does however place additional demands on the system infrastructure, as it must authenticate, authorize, scan, process, and deliver the query results. Software tools and expertise for proper delivery and management of these services, respectively, is also required.

2.1.1.4 Web Coverage Service (WCS)

WCS is a standard for querying and serving geospatial continuous feature data, also known as a coverage, over the internet based on any number of constraints, such as spatial or temporal. Unlike the feature data returned by WFS, WCS tends to focus on temporal and geospatial raster data without well-defined boundaries. It is subject to the same drawbacks as WFS.

2.1.1.5 Sensor Observation Service (SOS)

SOS is an XML-based standard to query sensor data and time series data in real time. Incorporated in SOS are additional standards such as the Observations and Measurements (O&M) standard for encoding sensor measurements and the Sensor Markup Language (SensorML), used to describe a sensor or collection of sensors. SOS is similar in operation when compared to WFS but, unlike the more general WFS standard, it is specifically designed to handle sensor information.

Although SOS provides reliable services, responsiveness is known to be sluggish and there is significant management overhead associated with its use (Section 2.5.3). Consultation with SOS users revealed it is not a preferred standard.

2.1.1.6 SensorThings Application Programming Interface (STA)

The SensorThings API is a standard based on JavaScript Object Notation (JSON). It is similar in function and purpose to SOS, but aimed at sensor devices in the Internet of Things (IoT). The standard was developed to address limitations of SOS, wherein the XML of SOS tends to be difficult for resource-constrained IoT devices to handle and is cumbersome for web developers to manage.

It has been demonstrated that the SensorThings API is interoperable with SOS,[1] although full interoperability has not been established for the reverse direction (SOS to SensorThings). Compared to SOS, STA is a lightweight means of sharing interoperable sensor data. However, it is a relatively new standard and support is currently limited.

2.1.1.7 Climate and Forecast Network Common Data Format (CF-netCDF)

The Network Common Data Form (netCDF) is a set of software libraries and data formats developed by Unidata to support the creation, access, and sharing of scientific data. The conventions for Climate and Forecast (CF) metadata have been designed to promote the sharing of files created with netCDF. Since it is possible to create a netCDF file without self-describing datasets, the CF-netCDF standard seeks to support the versatility of the netCDF format through use of CF metadata conventions to create self-describing datasets.

A dataset that does not require an external metadata record is more portable and can be meaningful to both humans and machines. It is a recommended standard for exchanging data.

2.1.2 Internet Engineering Task Force (IETF)

IETF is an open and international community of individuals who collaborate in the development and promotion of voluntary internet standards. The mission of the organization is to "make the internet work better" through the production of technical documents which influence how people "design, use, and manage the internet".[2]

2.1.2.1 Geographic JavaScript Object Notation (GeoJSON) [RFC 7946]

The Geographic JavaScript Object Notation (GeoJSON) is an open standard format, based on JSON, for representing simple features and their non-spatial attributes, such as dataset ID, name, description, species, and colour. GeoJSON is lightweight, well-supported, and human readable. It is a recommended standard for supporting visualization and an optional output format from WFS.

GeoJSON is a format for representing simple geographical features. As a derivative of JSON, it inherits the limitations of its parent – namely that it is difficult to represent complex formats, such as topology or multidimensional data, in a universally understood manner.

Website: http://geojson.org/

2.1.3 International Standards Organization (ISO)

The International Standards Organization (ISO) is a non-governmental organization. Composed of 162 national standards bodies, it has a goal of supporting and promoting innovation through the development of market-relevant international standards.

[1] http://www.opengeospatial.org/projects/initiatives/imisiot

[2] https://www.ietf.org/rfc/rfc3935.txt

2.1.3.1 19115/19139 Geospatial Metadata/XML Representation

ISO 19115 defines an extensible, interoperable metadata standard for geospatial data and services. ISO 19139 is the specification of how to represent, validate, and exchange this metadata in XML.

2.1.4 Other/Protocols

2.1.4.1 Representational State Transfer (REST)

Representational State Transfer (REST) is a method of delivering interoperable web services over the internet. RESTful web services can return data in a variety of forms and are currently utilized by various software tools in the ocean observing community. REST is not a standard in itself, but an architectural style of interacting with web services. RESTful web services rely on a number of standards to function and provide a great deal of flexibility for development purposes.

As an architectural style which relies on other standards, a general use case does not exist for REST. Implementation is the purview of the author, and as such each interface is unique and may require custom handling, thus increasing management overhead.

2.1.4.2 Open-Source Project for a Network Data Access Protocol (OPeNDAP)

Open-Source Project for a Network Data Access Protocol (OPeNDAP) is a widely used protocol and data dissemination architecture. The protocol is maintained by a non-profit organization of the same name. Typically, an OPeNDAP implementation is employed as middleware to bridge the gap between a client program and the datasets that an OPeNDAP server has been configured to serve. OPeNDAP is capable of querying a dataset and returning the subset of requested data in a variety of data formats.

Implementations of OPeNDAP are extensible to allow for extra functionality. This necessitates a development effort; to create a robust and well-designed extension represents a significant investment. The decision to use an OPeNDAP install without native support for a required functionality must therefore be carefully considered.

Although versatile, the OPeNDAP protocol is not suitable for all data types. It is also middleware, and as such it is not suited to be public-facing. The onus for a strong user interface is thus placed on the implementing organization.

Website: https://www.opendap.org/

2.1.4.3 Web-Accessible Folder (WAF)

Web accessible folders (WAF) are merely a directory listing of files in a folder, as served by a web server. Files are visible and accessible to users once published in the directory, without need for further work or management. By itself, WAF is not special. But when used in conjunction with certain conventions – such as a predictable and pre-defined structure – it is capable of serving otherwise difficult to handle datasets and providing a harvestable directory

of XML formatted metadata records. The United States Integrated Ocean Observing System (US IOOS) has developed a number of best practices for using WAF that could be leveraged for CIOOS.

To ensure the directory is harvestable, it is necessary that WAFs adhere strictly to the relevant conventions; otherwise the harvesting software will not recognize the files as anticipated. Further, allowing open access to a directory presents a risk from a security perspective, as it can provide information about the system's internal structure and presents a vulnerability which may be exploited by malicious actors if not properly secured. When employing WAF it is important to restrict the potential for directory traversal and only allow the desired directories and files to be accessed.

Website: https://ioos.github.io/catalog/pages/registry/waf_creation/

2.1.4.4 Quality Assurance of Real-Time Oceanographic Data (QARTOD)

Quality Assurance / Quality Control of Real Time Oceanographic Data is a multi-agency effort to collaboratively address quality assurance and quality control issues of IOOS and the broader international community. To that end, QARTOD publishes QA/QC manuals for assessing the quality of particular types of data; these manuals are considered living documents and are periodically revised and updated as technologies and techniques evolve. Data that have undergone a QARTOD evaluation can have metadata accompanying data points or complete datasets to describe their quality. These can take the form of annotations.

Website: https://ioos.noaa.gov/project/qartod/

2.2 Existing Data Management Software

There is a plethora of software to manage, transform, index, and distribute data; the software discussed below represents a subset of all relevant software. The primary inclusion criterion was software that is already in use in the ocean observing community, resulting in the existence of a large body of expertise. Software that is experimental or still under development has not been considered because of the uncertainty it represents.

2.2.1 Metadata Catalogues

Essential to any data infrastructure is a well-organized metadata catalogue, which allows users to discover, evaluate, and access data. Many software packages exist for providing a metadata catalogue service, most possessing similar feature sets. We limit our review to those widely used in ocean sciences and other disciplines.

The standards which constitute a cataloguing service may be categorized into *metadata standards* and *exchange standards*. Metadata standards are used to identify datasets and describe their contents, and are described in detail in the report produced by the Data and Observations IE. Exchange standards allow for the sharing of metadata in a standardized

format, useful for accessing datasets programmatically and transferring datasets between catalogues. Catalogues which employ the same standards are interoperable.

Metadata catalogues explored below (Table 1) were selected because they support the Open Geospatial Consortium's (OGC) Catalogue Service for the Web (CSW) standard. This standard allows for interoperability between catalogue servers regardless of provider, is in widespread use internationally, and is considered a best practice use.

Table 1: Summary of the metadata catalogues researched.

Catalogue	Open Source	License	Standards Compliant
GeoNetwork OpenSource	✓	GPL 2.0[3]	✓
CKAN	✓	AGPL 3.0[4]	✓*
ESRI GeoPortal	✓	Apache 2.0[5]	✓
PyCSW	✓	MIT[6]	✓

2.2.1.1 Comprehensive Knowledge Archive Network (CKAN)

CKAN is an extensible, open source application for managing and publishing data and metadata collections. Although recent strides have been made to better support CKAN implementation on Windows machines, the catalogue is primarily Linux-based. It is a base platform with most functionality added through extensions, which allows it to support numerous standards and spatial data as well as OGC standards. It requires a plugin for CSW compatibility, which itself relies on implementation of PyCSW (Section 2.2.1.4). This presents two potential points of failure: the plugin itself and the external software on which the plugin relies.

CKAN is nonetheless a popular and well-established software, utilized by governments, research institutions, and other types of organizations worldwide. It is currently in use by the Canadian Government in the form of the Open Government (http://open.canada.ca) initiative.

Primary Development Language: Python
Website: http://ckan.org

2.2.1.2 GeoNetwork OpenSource

GeoNetwork is an open source metadata catalogue application designed to manage spatial data resources. It is mature, standards-based, and currently in use with numerous spatial data infrastructures and ocean observing systems around the world, such as the Australian Integrated Marine Observing System (IMOS).

[3] https://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html
[4] https://en.wikipedia.org/wiki/Affero_General_Public_License
[5] https://en.wikipedia.org/wiki/Apache_License
[6] http://docs.pycsw.org/en/latest/license.html

As compared to CKAN, however, GeoNetwork exhibits fewer capabilities and is less intuitive. Interface and management may also be daunting, as they require significant investment of time and resources.

Primary Development Language: Java
Website: https://geonetwork-opensource.org/

2.2.1.3 Environmental Systems Research Institute (ESRI) GeoPortal

ESRI GeoPortal server is an open source metadata catalogue released by ESRI under the Apache 2.0 license. It is a stand-alone, standards-compliant metadata catalogue server that also integrates easily with existing ESRI products, which are closed source and expensive.

GeoPortal appears to be less popular than alternative catalogue options, and though standards-compliant, documentation suggests it conforms primarily to ESRI practices and processes.

Primary Development Language: Java
Website: http://www.esri.com/software/arcgis/geoportal

2.2.1.4 Python Catalogue Service for the Web (PyCSW)

PyCSW is an open source, OGC CSW server that can be run standalone or integrated into other applications. Although the interface is rudimentary, it provides powerful capabilities through connections to third-party software. PyCSW is an official OGC reference implementation of the CSW standard.

PyCSW is the software that provides CKAN with its CSW capability as well as several other open data catalogues, such as GeoNode. It also supports the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), OpenSearch, and Search/Retrieval via URL (SRU).

Primary Development Language: Python
Website: http://pycsw.org/

2.2.2 Data Dissemination

Discussed within this section are the data access and dissemination software systems utilized by data servers to provide users with access to information located within a catalogue service.

Different types of data will require different means of representation and delivery, and as such there is no universal means of delivering data. Various OGC standards exist to assist with the distribution of interoperable data, but cannot cover every scenario. Although other protocols, such as OPeNDAP, are immensely helpful with the dissemination of structured data, unstructured data still present significant challenges. Summarized in Table 2 and discussed in detail are the three most common implementations of OPeNDAP: Hyrax, ERDDAP, and THREDDS.

Table 2: Summary of the features available through the three most common implementations of OPeNDAP: Hyrax, ERDDAP, and THREDDS. Priority refers to the importance of the feature to an ocean observing system, where 'high' is of greatest importance.

Feature	Priority	ERDDAP	THREDDS	Hyrax
Data Products via API	*High*	✓	✓	✓
Federation Capabilities	*High*	✓		
Intelligent Aggregation / Caching	*High*	✓		
GeoJSON	*High*	✓		
CSV	*High*	✓		✓
netCDF3/4	*High*	✓	✓	✓
ASCII	*High*	✓		✓
JSON	*Medium*	✓		✓
MAT	*Medium*	✓		
QARTOD Filtering	*Medium*	✓	✓	✓
WCS Server	*Medium*		✓	✓
GRIB	*Medium*	✓	✓	
NEXRAD	*Medium*		✓	
HDF4/5	*Medium*	✓	✓	✓
CDM	*Medium*		✓	
KML	*Medium*	✓		
XML	*Low*	✓	✓	
FITS	*Low*			✓
CEDAR	*Low*			✓
WMS Server	*Low*	✓	✓	✓
NcML	*Low*	✓	✓	✓

2.2.2.1 OPeNDAP Hyrax

Hyrax is an extensible, open source OPeNDAP server developed by the non-profit OPeNDAP organization, the same organization which developed the OPeNDAP protocol. Designed to be used primarily as middleware, it provides limited direct web functionality. This allows an organization to separate web access from data storage.

Hyrax is comprised of two servers working in tandem; one handles front end requests to the system and the other fulfills those requests by serving data. This design allows for a number of different configurations to address various use cases and can be leveraged to provide a rudimentary form of load balancing for higher stress environments.

Hyrax can provide data in many different formats as well as deliver data via OGC WMS and WCS standards. Hyrax is capable of generating a THREDDS catalogue of its internal holdings for better integration with THREDDS.

Primary Development Languages: C++, Java
Website: https://www.opendap.org/

2.2.2.2 Environmental Research Division Data Access Program (ERDDAP)

ERDDAP is an open source data server developed by the National Oceanic and Atmospheric Administration (NOAA) that implements the OPeNDAP protocol. ERDDAP is often used as middleware to provide viewers or other data servers with data as a source. The provided web interface is more extensive than is standard for OPeNDAP, and it is capable of federating with other instances of ERDDAP. This allows the server to function under very heavy load conditions by distributing demand across multiple ERDDAP installations. This distribution reduces strain at any one instance and provides a form of fault tolerance should an ERDDAP installation become overwhelmed. Implementation of load balancing through ERDDAP has been tested by Axiom Data Science in the US, and, though possible, requires additional development to ensure nodes are synchronized as datasets update.

ERDDAP provides an XML metadata catalogue which serves standard OPeNDAP metadata, FGDC, and ISO 19115-2 records via the WAF standard. It is widely used and excels at serving tabular and gridded data in many different formats. Its limitations are: visualizations tend to be basic and visually unpleasant; poor handling of particular types of multidimensional data, causing loss of functionality due to necessary restructuring of the datasets; and lack of multilingual capabilities.

Primary Development Language: Java
Website: http://coastwatch.pfeg.noaa.gov/erddap/

2.2.2.3 Thematic Real-time Environmental Distributed Data Services (THREDDS)

THREDDS is a data server developed and supported by Unidata. While it implements the OPeNDAP protocol, it also provides metadata support and other forms of connectivity that ERDDAP lacks, such as offering an integrated WCS server. In practice, THREDDS finds a role in bridging the gaps in capability that other OPeNDAP servers cannot. THREDDS provides a metadata catalogue in the form of an XML document which can be consumed by other DAP servers including other THREDDS installs, Hyrax and ERDDAP, allowing these data servers to point to THREDDS datasets.

THREDDS is often used in tandem with ERDDAP, or in place of it, as the situation demands. This is especially true in the cases of unstructured, gridded data which ERDDAP cannot handle.

It should be noted that the OPeNDAP organization and Unidata have been collaborating for a number of years in an effort to better align their respective products, Hyrax and THREDDS. The goal is to eventually integrate their software.

Primary Development Language: Java
Website: http://www.unidata.ucar.edu/software/thredds/current/tds/

2.2.2.4 Comprehensive Knowledge Archive Network (CKAN)

Although CKAN is generally intended for metadata management, it is also capable of publishing and distributing datasets. This is not the primary strength of the application.

2.2.3 Visualization

The visualization of geospatial data may be categorized into server-side software and client-side software. It is the former which is of primary concern to the Cyberinfrastructure IE. Server-side software will operate on CIOOS hardware, and as such the storage volume and processing power supplied must be sufficient to effectively serve data to the client-side software.

The primary type of server-side visualization software will be mapping servers. These servers take geospatial data and render it for use in viewers and interactive applications; such servers can be hosted on a node or leveraged by an external entity for use in their own applications through the use of standards.

More comprehensive information regarding client-side visualization services is detailed in the report produced by the Visualization IE.

2.2.3.1 MapServer

MapServer is a cross-platform, open source server originally developed in the 1990s. It supports many open standards as well as proprietary formats. It is one of the founding projects of the Open Source Geospatial Foundation (OSGeo) and has a broad base of support.

Website: http://mapserver.org

2.2.3.2 GeoServer

GeoServer is a java-based open source map server which focuses on the use of open standards to share, analyze, and edit geospatial data from a variety of sources. It is the reference implementation for the OGC Web Feature Service standard and exhibits performance

comparable to MapServer. Unlike MapServer, GeoServer includes a well-developed web interface for managing the server.

Website: http://geoserver.org

2.2.3.3 Environmental Systems Research Institute ArcGIS Server

ArcGIS Server is a commercial web-mapping platform which has deep integration with other ESRI Products. ArcGIS has grown to support a number of open standards in addition to ESRI's proprietary formats. It can be installed directly onto CIOOS hardware or run via ESRI's cloud services.

Website: http://server.arcgis.com

2.2.4 Data Management Tools

Server-side transformations or manipulations of data will have an impact on the resources available to the rest of the system. Some tools are meant to be run offsite or at the desktop level, while others will have a server component. Much like visualization tools, data management software is largely outside the scope of the Cyberinfrastructure IE.

In the moderate and high service models, software systems will be required to convert data contributed by local data providers into CIOOS-compliant forms, so that they are able to be discovered and visualized through the CIOOS portals. These conversions may need to take place upon data request, such as conversion from raw format to netCDF prior to being made available through an ERDDAP interface.

File conversion so that data is compliant with CIOOS standards is not without cost – both technical expertise and software tools are needed. The amount of work required will differ based on the specific circumstances of the data provider and the number of data providers who require assistance. Costs associated with providing resources to data providers is dependent on the service model adopted, and will be difficult to ascertain until more is known.

Automated tools in the form of compliance checkers can be leveraged to provide an initial inspection and validation of data before submission to CIOOS. Other OOSes, such as IOOS, the Australian Ocean Data Network (AODN), and SeaDataNet provide access to a number of their tools and open source projects; these may be adopted and/or modified for use in CIOOS. For example, IOOS provides an open source compliance checker to give their data providers a method of validating their data before submission, while AODN utilizes a modified version of this tool for the same purpose.

In addition to the tools available from other OOSes, there is a wide variety of data conversion and transformation software available commercially and in the open source community.

2.2.5 Database Platforms

Underpinning the public-facing CI infrastructure will be a database platform: a relational database management systems (RDBMS), or a NoSQL DBMS which is more flexible than its RDBMS cousins. The metadata catalogues discussed in Section 2.2.1 all use relational DBMSs to store metadata. Oceanographic data can be stored in either relational databases or as a collection of organized, standard-compliant files (e.g., netCDF files). Relational databases may not scale elegantly with the amount of data expected to be collected under CIOOS.

2.2.5.1 PostgreSQL with PostGIS

PostgreSQL is an open source RDBMS focused on standards compliance and extensibility. PostGIS is a popular geospatial extension for PostgreSQL which adds support for geographic data types and allows data to be queried against location. This combination is often the database platform of choice for GIS software and is very well supported.

Website: https://www.postgresql.org/

2.2.5.2 Oracle

Oracle is a popular, commercial RDBMS developed by a corporation of the same name. Like PostgreSQL, Oracle enjoys broad support and, when paired with Oracle Spatial and optional Graph component, powerful GIS capabilities. Licenses can be expensive, but are sought by governments and large-scale enterprise with high support and reliability requirements.

Website: https://www.oracle.com/

2.2.5.3 Structured Query Language (SQL) Server

SQL Server is a popular, commercial RDBMS developed by Microsoft. Although SQL Server does not enjoy the broad base of support that PostgreSQL and Oracle do, in recent years Microsoft has made great strides to compete in the geospatial arena.

Website: https://www.microsoft.com/sql-server/

2.2.5.4 MongoDB

MongoDB is an open source, NoSQL, document-oriented database which employs a JSON-like document structure capable of supporting schemas for said documents. MongoDB's strength lies in its flexibility as a platform.

The US IOOS metadata harvest registry makes use of MongoDB for centralized metadata harvesting. The registry is a means of allowing IOOS partners to add or update their information in the IOOS data catalogue, and is the entry point for partners to publish their datasets and services in the IOOS catalogue.

Website: https://www.mongodb.com/

2.2.5.5 Cassandra

Cassandra is an open source NoSQL database system designed for high reliability and performance. Cassandra adheres to a database model similar to the tabular model of traditional RDBMS database platforms. In addition, Cassandra implements a query language that is similar to SQL. Ocean Networks Canada (ONC) currently uses Cassandra to optimize the storage of and access to data that has a complex structure.

Website: https://cassandra.apache.org/

2.2.6 Security

Despite dealing only with open and publicly available data, security is not something that should be taken lightly. There are multiple security-related scenarios which may impact CIOOS such as: the computing resources of the CIOOS infrastructure could be hijacked and used for malicious ends; the data CIOOS is hosting could be subject to ransom; or user account information could be stolen.

It is important for security to be forefront at all stages of building CIOOS. Good security practices, implemented during creation, will help to mitigate many of the potential vulnerabilities associated with building a large, interconnected system. Specific recommendations for security are detailed in Section 6.2.

2.3 Current Cyberinfrastructure in Canada

A planning survey for CIOOS was completed mid-September 2017, garnering 18 respondents. Of those, 14 represented Canadian organizations. Included were questions concerning methods for data discovery and access, software systems and interoperability standards, storage and processing of data, and employment of technical staff.

Survey results were promising; a number of respondents indicated a willingness to contribute data to CIOOS and an enthusiasm for implementation of an ocean observing system in Canada. Results also revealed that the software tools used by existing organizations are many and varied, suggesting that care must be taken to ensure CIOOS maximizes interoperability with minimal effort.

When crafting CIOOS cyberinfrastructure recommendations, the survey results were considered in conjunction with other factors, such as the structure of international OOSes and lessons learned through consultations with existing ocean observing systems. Survey results relevant to cyberinfrastructure are summarized in Appendix A.

2.4 Structure of Other OOSes

Ocean observing systems currently exist in numerous countries. Several of these were investigated to provide insights as to how CIOOS may be structured – in terms of governance, data products, and technologies. The former is of importance to potential CIOOS

cyberinfrastructure as it affects the selection of cataloguing and data access software, and also impacts the requirement for federation of data.

Data products and technologies utilized by existing OOSes provide an overview of the tools currently employed internationally. CIOOS should leverage such information to select software tools and standards which maximize interoperability – thus allowing Canada to cooperate with our international partners, contribute to GOOS, and fulfill international obligations.

The systems examined were: US IOOS, EuroGOOS, IMOS, EMODNet, SeaDataNet, CMEMS, PANGAEA, BODC, and EUMIS (see Appendix B for definitions). Although significant differences exist between each organization, most consist of a national or international body with regional or national partners.

It should be noted that the information contained within this section represents an overview of these organizations, and is not intended to represent an exhaustive review. A standard or tool not mentioned within this overview means no indication was found that it is used; lack of mention is not definitive proof that a tool is not employed by an organization. Data access and standards found to be employed by the surveyed organizations are summarized in Appendix C.

2.4.1 United States Integrated Ocean Observing System (US IOOS)

2.4.1.1 Structure

US IOOS is structured into a national office with eleven regional associations (RAs).[7] Prior to the formation of IOOS, there existed disparate units, primarily academic, with sometimes diverging goals. The RAs were built atop these organizations, and as such the structures differ between regions. The emphasis of each RA is also region-dependent, with some focusing primarily on the integration and delivery of data as required by both individuals and organizations (e.g., NERACOOS). Others focus more on the underlying science and work towards the collection and analysis of data (e.g., MARACOOS).

Common between them is a focus on addressing the unique needs of the local research community, determined through stakeholder engagement. This process has brought into focus the diverse requirements of each region for oceanographic data, resulting in disparate data products and data portals across the regional associations. For example, the Alaska RA (AOOS) provides data regarding sea ice,[8] information not relevant for other regions and not provided elsewhere.

There do exist data assembly centres (DACs) within US IOOS at the national level, such as for glider and high-frequency radar data.[9,10] These may be termed *thematic nodes*, as they

[7] https://ioos.noaa.gov

[8] http://www.aoos.org/historical-sea-ice-atlas/

[9] https://gliders.ioos.us

center on a data type as opposed to a region. The data originates primarily from the regional associations, and is often available through both the RA and the DAC portals. Private consultant companies, such as Axiom Data Science and RPS ASA, are employed to manage data storage for such nodes.[11]

2.4.1.2 Software Tools

The software tools utilized within US IOOS vary by regional association and governing body. The national governing body employs a harvesting tool which aggregates metadata from WAFs maintained by each regional association into a CKAN catalogue product. Data serving does not occur nationally, and the catalogue instead links to the data access services provided by the regional associations.

The regional associations do not employ a catalogue product, instead producing harvestable WAFs containing data in netCDF format and metadata in XML format, as is required by the national program office. Data serving software varies by region, and the services utilized are outlined in Table 3. The exact structure of data storage at the regional associations is region-dependent, but consists of using hardware at partnering institutions such as universities, and employing private consulting companies.[12]

Table 3: The software and standards employed by US IOOS regional associations.

Software and Standards	
THREDDS	LAS
ERDDAP	GeoServer
OPeNDAP	WCS Client
KML Feeds	WMS
SOS	WFS

2.4.1.3 Data Acquisition

Regional associations are responsible for the integration of oceanographic data within a prescribed region, and through consultations with US IOOS RAs, it was discovered that this takes on both *push* and *pull* characteristics. *Push* refers to situations wherein data providers approach the RA and request their data be included in the regional data products, and, by extension, the national catalogue. In these circumstances it is primarily the responsibility of the data provider to ensure the data meets US IOOS standards for quality and format.

Pull refers to conditions wherein the regional association identifies a potentially beneficial dataset, often due to its relevance to stakeholder interests, and requests that the data provider allow for the inclusion of said data. The onus falls on the RA to ensure that data meets US IOOS standards for quality and format.[13]

[10] https://ioos.noaa.gov/project/hf-radar/

[11] Consultation with US IOOS National Office

[12] Consultation with US IOOS National Office

[13] Consultation with NERACOOS

Although much of this process would be the domain of either governance or data management teams, it is necessary that CIOOS cyberinfrastructure include tools which enable this process.

2.4.2 European Global Ocean Observing System (EuroGOOS)

2.4.2.1 Structure

EuroGOOS is a pan-European non-profit association which brings together oceanographic research and operational services from 41 members in 19 European countries. Responsibilities of the network also include the coordination of five regional operational systems covering the Arctic, the Baltic, the European Northwest Shelf, the Ireland-Biscay-Iberian area, and the Mediterranean. Member institutions for the regional systems consist of subsets of the 41 EuroGOOS member organizations.[14]

The focus of EuroGOOS is the identification of priorities and enhancement of communication amongst oceanographic researchers within Europe.[15] Working groups establish strategies and priorities to ensure a cohesive continent-wide approach to oceanographic research,[16] while task teams promote cooperation and collaboration amongst the member organizations.[17]

EuroGOOS members are drawn from existing organizations within the participating countries, primarily universities and national institutions.[18] As such, each are unique in terms of structure, priorities, and standards.

2.4.2.2 Software Tools

EuroGOOS is an international coordinating body and does not offer data cataloguing or data serving services. Data storage and hardware requirements, which may be handled in-house or contracted to an external organization, are the responsibility of the member institutions. EuroGOOS instead works to coordinate international organizations and feed oceanographic data to pan-European portals such as EMODnet and CMEMS.[19]

2.4.3 Australian Integrated Marine Observing System (IMOS)

2.4.3.1 Structure

IMOS is a national collaborative research infrastructure – a fully integrated national system which collects data on physical, chemical, and biological oceanographic variables. Operations are carried out through the aggregation of observation and data management capabilities from eight different institutions, consisting of a lead agent and numerous partners, which together form an unincorporated joint venture.

[14] http://eurogoos.eu/about-eurogoos/overview/
[15] http://eurogoos.eu/about-eurogoos/goals/
[16] http://eurogoos.eu/working-groups/
[17] http://eurogoos.eu/task-teams/
[18] http://eurogoos.eu/about-eurogoos/list-of-eurogoos-member-agencies-and-contact-persons/
[19] http://eurogoos.eu/about-eurogoos/eurogoos-strategy/

The eight institutions are considered IMOS Facilities, and undertake the deployment of oceanographic equipment and delivery of the associated data streams. Because the Australian ocean community is large, diverse, and dispersed, science planning and priority setting is performed through a thematic node and five regional nodes. These cover the areas of Bluewater and Climate (thematic), Queensland, New South Wales, Southern Australia, Western Australia, and South East Australia (regional). While priorities differ based on regional needs, activities in each feed into the five major themes that make up the unified IMOS science plan (Table 4).

Table 4: The five major themes that make up the unified IMOS science plan.

IMOS Major Themes	
Long-Term Ocean Change	Climate Variability and Weather Extremes
Boundary Currents	Continental Shelf and Costal Processes
Ecosystem Responses	

Data generated through IMOS-funded activities are collected via ten technology platforms, and all streams, including near real time and/or delayed mode, are discoverable through the Australian Ocean Data Network (AODN). The AODN Portal comprises a metadata catalogue, a search interface, a map interface, and data downloads. It does not host information, but instead serves as a portal, a single access point, for Australian oceanographic data.[20]

2.4.3.2 Software Tools

The AODN Portal is a stateless web application which utilizes an instance of the AODN Open GeoSpatial Portal (an open source project owned by IMOS). Although not employing data serving software itself, it has been noted that the OPeNDAP protocol and specifically the THREDDS implementation is commonly used by IMOS data providers.[21]

Developed primarily on ExtJS[22] and operating a Grails[23] backend, AODN has no database of its own – and thus minimal hardware requirements – but is instead a portal for external applications.[24] The software tools used and their associated functions are summarized in Table 5.[25]

Communication between the AODN Portal and other software platforms is based on OGC standards, such as WMS, WFS and WPS. They are employed to deliver map layers, serve non-gridded data, and standardize inputs and outputs for geospatial processing services.[26] Communication with external platforms is also supported through use of SOS.[27]

[20] http://imos.org.au/about/

[21] https://help.aodn.org.au/contributing-data/data-storage/

[22] https://www.sencha.com/products/extjs/

[23] https://grails.org

[24] http://imos.org.au/facilities/aodn/imos-data-management/imos-information-infrastructure/

[25] https://dnahodil.files.wordpress.com/2014/11/nahodil_poster_eresearch2014.pdf

[26] https://help.aodn.org.au/user-guide-introduction/aodn-portal/information-infrastructure/

[27] http://imos.org.au/fileadmin/user_upload/shared/IMOS%20General/ACOMO/ACOMO_DAY_2/06.Roger_Proctor_1.pdf

Table 5: Software tools used in AODN and their associated functions.

Software Tool	Use of Tool
Open GeoSpatial Portal	Web Portal Access to IMOS Data
GeoNetwork	Metadata Catalogue and Search Functionality
GeoServer	Maps, Subsets and Data Downloads
ncWMS	Maps for Satellites and Land-Based Radar Data
AODAAC	Data Subset and Download for Satellite Data
GoGoDuck	Data Subset and Download for Gridded Data
Ocean Depth Service	Provides Ocean Depth Information

AODN does not require that specific guidelines be followed for data storage, although the website notes that netCDF is a format commonly used by data contributors. As such the network has provided a recommendation for IMOS netCDF format. Also available is a MATLAB Toolbox wherein certain raw instrument data may be converted to IMOS compatible files.[28]

2.4.4 European Marine Observation and Data Network (EMODnet)

2.4.4.1 Structure

The European Marine Observation and Data Network (EMODnet) is a web portal which aggregates oceanographic data from numerous sources within Europe. Cooperation is promoted amongst participating organizations to minimize the effort involved in collecting, processing, and making freely available a variety of oceanographic data across several thematic nodes (Table 6).[29]

Table 6: The thematic nodes supported by EMODnet, and through which data is made available.

EMODnet Thematic Nodes			
Bathymetry	Geology	Seabed Habitats	Chemistry
Biology	Physics	Human Activities	

Currently, data management experts and more than 160 data repositories are involved in the aggregation of said data sources. This work involves ensuring long-term stewardship of the datasets, creation of data products such as digital terrain models, and ensuring interoperability between systems.[30] Datasets from local to international organizations are included, such as the EuroGOOS regional data portals.

EMODnet offers a Data Ingestion Portal as a pathway for data providers to submit their data. Metadata fields and links to dataset locations are submitted via the EMODnet Data Submission Service, and checked for quality by an Ingestion Portal Manager. Data is then assigned to a data centre for processing, where data centres are a subset of those organizations

[28] https://help.aodn.org.au/contributing-data/data-storage/

[29] http://www.emodnet.eu/what-emodnet

[30] https://www.emodnet-ingestion.eu/about/why

which coordinate the thematic nodes. Assignment is based on relevance to the work performed at the centre.[31]

Finalized data are included in the data management system of the data centre, but also populated into the appropriate European infrastructure (e.g., SeaDataNet) and the EMODnet thematic node. Each thematic node – developed in a consultative manner involving a collaborative network of several organizations[32] – possesses a gateway to data archives wherein users may access oceanographic data from a number of diverse sources. The exact offerings are dependent on the thematic node, but include the following services:

- Data discovery and access
- Composite product discovery and access
- Viewing and download
- Dynamic map facilities for viewing and downloading
- Dashboard reporting
- Machine-to-machine communication

Data services offered by the national EMODnet portal are a metadata catalogue, map viewer, and query tool – the latter of which aggregates and makes discoverable data from each of the seven thematic nodes.[33]

2.4.4.2 Software Tools

Several Open GeoSpatial Consortium Standards, primarily WMS and WFS, but also including WMTS, WCS, and REST, are utilized by the thematic nodes for data visualization, as is GeoServer. The Physics node also utilizes GODIVA2 and THREDDS to provide visualization, with data dissemination performed through THREDDS and OPeNDAP.[34]

Beyond this, little information is readily available regarding the software tools behind the national EMODnet portal, the data ingestion process, or the thematic nodes.

2.4.5 SeaDataNet

2.4.5.1 Structure

SeaDataNet is a distributed pan-European infrastructure for the management of large and diverse oceanographic datasets, and is operated in conjunction with National Oceanographic Data Centres and data focal points from 35 countries. The system aggregates information from approximately 600 European data providers,[35] resulting in unified access to large volumes of marine data from European seas and oceans.

[31] https://www.emodnet-ingestion.eu/data-submission

[32] http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52012SC0250

[33] http://www.emodnet.eu/dataservices/

[34] http://www.emodnet-physics.eu/hfradar/docs/confirmed/3.%20RITMARE_eurogoos.pdf

[35] https://www.seadatanet.org/content/download/1754/10447/file/SeaDataCloud_EGU2017_4377_April2017_Schaap_presentation.pdf

Upon submission to SeaDataNet, data is directed to one of over 100 professional SeaDataNet data centres, all of which are connected to the SeaDataNet portal through the Common Data Index service. Quality control procedures are undertaken, and once the information has been confirmed to be in compliance with SeaDataNet standards, it is made available through numerous metadata portals – including the SeaDataNet portal, EDMED, EDIOS, and CSR.[36]

Although the SeaDataNet portal facilitates data discovery and download, it does not host datasets. All downloads occur directly from the data centre at which the information is hosted.

2.4.5.2 Software Tools

The overarching body of SeaDataNet contains a technical task group with a mandate to develop software tools which facilitate common means for data analysis across organizations. Several such software tools are currently available via the SeaDataNet website, including a file format converter and visualization software package.[37]

The specific cataloguing software and visualization tools utilized for the SeaDataNet portal are not readily discoverable, although it is known that the former is CSW-compliant.[38] Nonetheless, posters and presentations from the International Marine Data and Information System (IMDIS) Conference indicate that SeaDataNet Data Centres employ various software setups, including some combination of the tools indicated in Table 7.[39]

Table 7: Examples of software and standards utilized at SeaDataNet Data Centres. The exact configuration is dependent on the specific Data Centre.

Software and Standards			
THREDDS	ERDDAP	FTP	HTTP
GIS	OPeNDAP	CSW	GeoServer

There also exists some indication on the SeaDataNet website that a future goal is implementation of the Sensor Observation Service as a means of facilitating access to near real-time data. This is a major challenge for the organization, and there is no information available regarding progress.

2.4.6 Copernicus Marine Environment Monitoring Service (CMEMS)

2.4.6.1 Structure

The Copernicus Marine Environment Monitoring Service (CMEMS) is a thematic node under the umbrella of Copernicus, a European earth observing system. Coordinated by Mercator Ocean, a non-profit company funded by five major French institutions,[40] the service offers numerous oceanographic observation and modelling products. These cover physical and

[36] https://www.seadatanet.org/content/download/1754/10447/file/SeaDataCloud_EGU2017_4377_April2017_Schaap_presentation.pdf
[37] https://www.seadatanet.org/Software
[38] http://seadatanet.maris2.nl/newsletter.asp?v0=8
[39] https://www.seadatanet.org/Events/IMDIS-Conferences
[40] http://marine.copernicus.eu/about-us/about-mercator-ocean/

biogeochemical states of the global ocean and the six European basins, with temporal periods extending from the 1990s (historical) to near real-time (present) to multi-day forecasts (future).[41] Currently 152 data products are offered by CMEMS.

Near real-time satellite data is provided to the service through instruments operated by ESA, EUMETSAT, CNES, and NASA. Also utilized are historical satellite observations from past missions. *In situ* observations are not collected by CMEMS itself, but are instead aggregated from data providers such as EuroGOOS Regional Alliances and JCOMM Networks. The service also collaborates with SeaDataNet and EMODnet Physics to involve new partners with the network.[42]

2.4.6.2 Software Tools

CMEMS employs THREDDS servers with the ncWMS extension.[43] This interactive OGC-compliant online web-GIS tool allows users to visually discover data while also facilitating data serving.[44,45] THREDDS also serves as part of the back-end for Motu, the web interface employed for extraction and data download. This tool utilizes a queue server to manage incoming requests and balance the processing load.[46]

The exchange format employed by CMEMS is that of netCDF. The service does not appear to include a catalogue, which indicates that metadata may be regulated and built into the netCDF files. The specific services through which data are available vary by product, but include CSW, WMS, FTP, MFTP, Subsetter, and DirectGetFile – where the latter two options are CMEMS service tools for downloading either a subset of or a full netCDF file, respectively.[47]

2.4.7 PANGAEA

2.4.7.1 Structure

PANGAEA is an open access data library that is jointly hosted by the Alfred Wagner Institute, Hemholtz Centre for Polar and Marine Research and the Center for Marine Environmental Sciences, University of Bremen.[48] Services provided include long-term archival, publication, distribution, and management of quality-controlled scientific data. As a permanent facility, PANGAEA guarantees that archived information is available in formats that are both secure, accessible, and consistent.[49]

Any earth or life sciences data, and in any format, are accepted into the system, whether coming from an institution or an individual scientist. Once submitted via the provided Ticket

[41] https://www.mercator-ocean.fr/en/solutions-expertise/off-the-shelf-oceanographic-services/
[42] http://marine.copernicus.eu/training/education/observation/
[43] http://forum.marine.copernicus.eu/discussion/498
[44] http://marine.copernicus.eu/services-portfolio/technical-faq/
[45] http://marine.copernicus.eu/newsflash/cmems4200-system-maintenance-saturday-june-18th-6hutc-13hutc/
[46] https://github.com/clstoulouse/motu
[47] http://marine.copernicus.eu/faq/how-can-i-access-the-documentation-associated-to-a-product/
[48] https://www.pangaea.de/about/
[49] https://www.marum.de/en/Infrastructure/PANGAEA-Data-Publisher-for-Earth-and-Environmental-Science.html

System,[50] data and metadata are assigned a curator who ensures quality standards are met for consistency and completeness. Data are then converted into the publication format and uploaded to PANGAEA and other relevant systems, such as the Electronic Publication Information Center (ePIC). [51]

After archival, a Digital Object Identifier is generated and provided to the author of the data. They are asked to proofread the information to ensure errors were not introduced during the archival process. Upon author approval, the final version of the data is published.[52]

2.4.7.2 Software Tools

PANGAEA employs a three-tiered client / server architecture which controls the flow of information within and outside of the system. At the backend is the PANGAEA Data Warehouse, which mirrors all archived information. The system utilizes relational database management software (RDBMS), specifically Sybase's Adaptive Server Enterprise (ASE) and Sybase IQ, on a multiprocessor computer to provide efficient data retrieval and compilation.[53]

Middleware used includes: a server component which creates flat files for serving through the website; the search interface PANGAEA Search, which is powered by Elasticsearch; and panFMP for the near real-time indexing of metadata. Interoperability of these services is ensured through the use of standard interfaces for communication.[54] The web service employs the Simple Object Access Protocol (SOAP) standard.[55]

The frontend of PANGAEA supports access to the system through numerous clients. The graphical user interface employs 4th Dimension software (ACI), while the various domains and services for data retrieval, download, and harvesting are run by a web server. The specific clients which provide data access and download are elucidated in Table 8.[56]

Table 8: List of the clients which provide data access and download for PANGAEA.

Client	Purpose
PangaVista	Search Engine
Advanced Retrieval Tool (ART)	Data Mining
Direct Download Interface (DDI)	Dynamic Query
PanCore	Metadata Search
Digital Object Identifier (DOI)	Persistent Link / Identifier
OAI-PMH	Metadata Harvesting

[50] https://pangaea.de/submit/
[51] https://www.pangaea.de/about/
[52] https://wiki.pangaea.de/wiki/Data_submission
[53] https://wiki.pangaea.de/wiki/Sybase
[54] https://wiki.pangaea.de/wiki/Technology
[55] http://www.copernicus.eu/sites/default/files/library/PANGAEA_Archiving_and_Publication_of_Scholarly_Data_for_the_Long_Tail_of_Science_01.pdf
[56] https://wiki.pangaea.de/wiki/PANGAEA

PANGAEA has also engaged in software development to produce freeware tools for the visualization, exploration, and interpretation of data, while a well-developed interoperability framework allows for data and metadata dissemination to other portals and services.[57]

2.4.8 British Oceanographic Data Centre (BODC)

2.4.8.1 Structure

The British Oceanographic Data Centre (BODC) is a national institution dedicated to the storage and dissemination of oceanographic data. Data management objectives for the organization encompass three areas: storage of quality-controlled data, open and online distribution of data, and data management services for active projects.

Currently, the BODC databases contain measurements for approximately 22,000 different variables concerning the biological, chemical, physical, and geophysical properties of marine environments, along with multiple numerical model datasets.[58] Web services deployed are the NERC Vocabulary Server, the Marsden Square translator service, and the GEBCO Web Map Service.[59]

Data submission to BODC is a process which involves numerous steps, with the specific procedure being dependent on whether the data are time series or discrete samples. In general, submitted information – both data and its associated documentation – are archived in their original form, ahead of conversion into a standard format. Metadata is then compiled and loaded into the database. Several quality checks are employed, including the use of in-house interactive visualization software to flag suspect values. Full documentation regarding the dataset is produced so as to minimize ambiguity or uncertainty for future users. [60]

Upon completion of processing, data is loaded into the appropriate database and is subject to an additional audit – involving different BODC staff – to ensure no errors were made. Only after a successful audit is the information made available, either to project participants on request or through the BODC website.[61] Data are also harvested from BODC and made available via partner organizations such as SeaDataNet.[62]

2.4.8.2 Software Tools

BODC databases include the Project Database and the Web Database, which respectively manage data for multi-national data projects and meet the needs of web applications for online delivery of information. Each employ a relational database design, and utilize the Oracle Relational Database Management System. The National Oceanographic Database

[57] https://www.pangaea.de/about/services.php
[58] https://www.bodc.ac.uk/about/what_is_bodc/
[59] https://www.bodc.ac.uk/resources/products/web_services/
[60] https://www.bodc.ac.uk/submit_data/what_do_we_do_with_your_data/data_processing_steps/moored_instrument_data_processing/
[61] https://www.bodc.ac.uk/submit_data/what_do_we_do_with_your_data/data_processing_steps/
[62] https://www.bodc.ac.uk/about/outputs/brochures_and_posters/documents/ycsec2015.pdf

(NODB) – which indexes metadata for all datasets hosted by BODC – was developed using a Conference on Data System Languages (CODASYL).[63]

Within BODC there exists a software development team which utilizes numerous languages to develop and maintain code internally and for web applications (Table 9). Software applications developed include visualization platforms, the BODC Transfer System for conversion of files into the standard in-house format, and the BODC Explorer Package. Written in Delphi, the latter provides various tools for the querying, visualization, and export of data.[64]

Table 9: The languages utilized by the BODC software development team to develop and maintain code internally and for web applications.

Internal Development	Web Applications
MATLAB	arcIMS
C++	JSP
Java and Java Server Pages (JSP)	Perl
Python	JavaScript
Microsoft Access	HTML
Oracle SQL Developer	XHTML
Linux Scripting Languages	
Delphi	

Little information is readily available regarding open source technologies utilized by BODC, although it was found that OPeNDAP technology is employed for subsetting of numerical model datasets,[65] and the NERC Vocabulary Server is compliant with W3C standards.[66]

2.4.9 European Marine Information System (EUMIS)

2.4.9.1 Structure

The European Marine Information System (EUMIS) was developed as a pilot for the Open Service Network for Marine Environmental Data (NETMAR) programme. The system aggregates various data types – satellite, *in situ*, and model – into a single user-configurable portal which provides search, download, and integration functions for historical, near-real time, and forecast data. EUMIS also integrates four case studies developed via NETMAR, and allows for the generation of composite products through additional data processing. [67,68]

[63] https://www.bodc.ac.uk/submit_data/what_do_we_do_with_your_data/database_design/

[64] https://www.bodc.ac.uk/submit_data/what_do_we_do_with_your_data/software_engineering/

[65] https://www.bodc.ac.uk/about/news_and_events/numerical_models.html

[66] https://www.bodc.ac.uk/about/outputs/presentations_and_papers/documents/sdn2_wp8_vocabulary.pdf

[67] https://netmar.nersc.no

[68] https://netmar.nersc.no/content/pilots

EUMIS is a pilot project, and it is likely that the current configuration will evolve over time. Nonetheless, its configuration provides insight as to what the ocean community requires of a large integrated system, which is highly relevant to the development of CIOOS.

2.4.9.2 Software Tools

To effectively provide flexible search, download and integration functions, EUMIS utilizes open source standards for three major software components:

- Visualization: OGC (WMS, WFS, WCS)
- Data Serving: OPeNDAP, WPS
- Cataloguing: CSW

Also employed are semantic technologies to support "Smart Discovery" of data, wherein searches may return datasets with keywords that are different from but semantically linked to the search term(s), such as returning a dataset labeled *rainfall* for a search on *precipitation*.[69]

2.5 Lessons Learned from Other OOSes

Establishing a national ocean observing system is an enormous undertaking, requiring significant collaboration between government and oceanographic data providers. There are numerous factors to consider, many of which may remain undetectable until systems and procedures are well-established. However, the existence of established OOSes in other regions presents an opportunity to avoid those items which have historically created unanticipated challenges during implementation.

The Cyberinfrastructure IE consulted with existing OOSes and private consultants specializing in oceanographic data management – specifically US IOOS, NERACOOS, MARACOOS, Axiom Data Science, and RPS ASA. The lessons learned are elucidated in the following sections, and the information was considered when forming CI recommendations.

2.5.1 Centralization

Centralization of cyberinfrastructure – such as hardware infrastructure and data formats – is a highly desirable goal. Such harmonization across regions would increase interoperability and minimize duplication of effort; regions could then focus primarily on scientific endeavors. But it is not realistic. Existing organizations have proven methods and connections to the broader community – they understand regional needs and are able to tailor information and data products to their user base. Often significant resources have been devoted to the development of such a system and network, an investment which must be respected. To do otherwise may deter participation in CIOOS.

Complete centralization is not without drawbacks. For example, a single interface to serve multiple user types may lead to excessive complexity and a poor user experience. US IOOS

[69] http://eumis.nersc.no:80/web/guest/wiki/-/wiki/Main/Technologies+and+tools/

has experienced such issues, and stressed the importance of clearly identifying the target user and system purpose prior to development.

Without complete centralization – where the national level federates data from the regions in real-time – a regional data outage would affect the national portal, and may adversely affect the reputation of CIOOS. Communication between the national and regional levels thus becomes very important, so that issues are addressed quickly and efficiently. Axiom Data Science suggested that it may be worthwhile to develop an automated monitoring system which notifies the regions of data outages.

Centralization also refers to the creation of central data products, such as the US IOOS Glider DAC or HF Radar DAC. The regional associations that were consulted recommended against a highly prescriptive approach for implementation of such products. Co-development was instead favored, wherein the national level and regional associations work together to create products which are widely applicable but also consider the needs and resources of the RAs.

2.5.2 Interoperability

US IOOS revealed that, due to the existence of regional data providers prior to the implementation of the national association, interoperability has been challenging. The existing regional organizations were accustomed to working independently, and have sometimes been resistant to accepting centralized tools or adopting new procedures which would harmonize the regions. To minimize the potential for such complications, US IOOS highly recommends that CIOOS structure be carefully considered in the initial phase.

2.5.3 Software and Standards

Open source software such as THREDDS, ERDDAP, and open standards such as OGC are prevalent in the structures of existing OOSes, although their use has not been without obstacles. Such software can be complicated and is not always as robust as hoped, necessitating the development of in-house solutions which may be incompatible with future builds, as management and control of the technology is via a democratic community and not the OOS team.

One software tool common among OOSes is that of a data harvester, which automates the data ingestion process – a process which can be complicated due to the multiple legitimate locations within a file for specific data. Axiom Data Science recommended that a standard metadata profile be created and the exact fields be clearly specified, so as to avoid situations wherein the harvester does not recognize fields within files due to their alternative configuration.

Consultations with the national body and regional associations within US IOOS revealed Sensor Observation Service to be a problematic standard. Although it generally performs as expected, there have been situations wherein an unanticipated behaviour has caused difficulties, such as server downtime. Further, responsiveness is sluggish and there is

significant management overhead. At the time of inception there were no mature alternatives for the necessary functionality provided by SOS. The intervening years, however, have yielded technologies with similar features – such as the OPeNDAP standard, used by ERDDAP and THREDDS specifically. As such, SOS was not recommended for implementation of a new system.

Software obsolescence has also been experienced by existing OOSes, such that established software becomes outdated and difficult to maintain; the window before this becomes a concern is approximately five years.

2.5.4 Software Development

Software development is an inevitability with the implementation of a national and integrated ocean observing system, whether it be for a national data product, to align the regional associations, or for an alternative purpose. An important consideration during development is the expected maintenance costs, including management overhead. Sensor Observation Service is one such example; the software performs as expected, but requires a significant time investment.

Any software or standard developed at the national level and required in the regions must also be thoroughly tested, lest the regions be required to spend limited resources, which would be better utilized elsewhere, to ensure successful adoption. NERACOOS experienced such a situation, and as a preventative measure has recommended that the most complex dataset from each region be examined with the new software or standard to ensure it is compliant.

2.5.5 Hardware

Within US IOOS, hardware for data processing and storage is often provided by either a regional node or private consultant. The exact specifications vary based on the needs of the region and the number of copies or backups that are kept. Axiom Data Science, which hosts and manages data for three regional associations within US IOOS, retains multiple copies – raw data, data with improved metadata, computed data (e.g., averages), and data chunked for visualization purposes.

With multiple copies and data from three RAs, Axiom filled a 200 TB drive array in approximately one year – and expects an exponential growth in data volume.

2.5.6 Submission (Interoperable)

Regional associations within an ocean observing system are in the best position to aid individual researchers and other data sources as they strive to submit data to the system. It is critical to the success of any OOS that the barriers to data submission be as low as possible. Regional associations and their embedded nodes should make technical expertise available to anyone submitting data to the system.

2.5.7 Looking Forward

US IOOS has discovered that for many aspects, including software and standards, there exists a balance between allowing regional associations to act independently and having the national governing body be highly prescriptive. Interoperability must be achieved while allowing the RAs to satisfy regional requirements. Consultation with existing organizations, during both planning and implementation, is highly recommended.

3.0 Storage Requirements for Core Variables

The cost of hardware is directly correlated to the storage and processing requirements of a system. For the purposes of estimation, the various ocean parameters have been broadly categorized into two types: tabular and non-tabular. Tabular data is formatted as either rows and columns, or a geospatial grid. Tabular data typically fits nicely into standard scientific data management tools. In contrast non-tabular data cannot be represented using rows or a grid. Audio and video data are examples of non-tabular data.

Storage requirements in CIOOS are a function of the volume of data input into the system, consisting primarily of those core variables which have been detailed in the report provided by the Data and Observations Investigative Evaluation. Based on the recommendations therein and on current data volumes at a number of institutions – including Oceans Network Canada (ONC), the Marine Institute (MI), the St. Lawrence Global Observatory (SLGO), and private consultant Axiom Data Science – CIOOS data volumes were estimated for both tabular and non-tabular data (Table 10). Typically the storage requirements for non-tabular data significantly outweigh the requirements for tabular data.

Table 10: Estimated data volumes, per day and per year, for organizations which manage oceanographic data.

MI Buoy Data	6 MB / Day	2.5 GB / Year
MI Multi-Beam Data	--	1 TB / Year
ONC Tabular Data	20 GB / Day	7 TB / Year
ONC All Data	250 GB / Day	100 TB / Year
Axiom	--	200 TB / Year
SLGO All Data[70]	--	3.5 TB / Year

As discussed in Section 2.5.5, consultation with Axiom Data Science revealed that a 200 TB drive array was required to host one year's worth of data for three US IOOS regional associations, and exponential growth in required storage volume is expected.

Long-term storage of all data will increase costs, regardless of the storage solution recommended; the solution to historical archiving requires great consideration. For initial implementation it is not recommended that CIOOS provide archival services, although the system should guarantee data for a minimum of five years after end of project.

[70] Much of the data managed by SLGO is hosted on servers owned by other organizations, such as DFO and Institut des sciences de la mer de Rimouski (ISMER). The storage volumes indicated in Table 10 reflect on that data which is hosted directly by SLGO.

Table 11: Potential storage volumes required to for five years of oceanographic data. Assumes data volume in year one of 150 TB or 200 TB, and annual growth of 25% or 50%.

Growth	Total Volume (Year 1 = 150 TB)	Total Volume (Year 1 = 200 TB)
25%	1.25 PB	1.70 PB
50%	2.00 PB	2.70 PB

Based on data volumes from Canadian organizations, the estimated storage requirements are 50 TB per year per region – given three regions, an estimated total of 150 TB is required per year. Based on data volumes managed by Axiom, an estimated 200 TB are required per year. This provides a range of 150-200 TB per year for the first year, with expected growth ranging from 25-50%. The possible total storage volume required by the end of the first five years, for both starting scenarios and both growth scenarios, is illustrated in Table 11.

4.0 Consultation with Hardware Infrastructure Providers

The hardware necessary for implementation of CIOOS may be managed directly by a data provider or regional association, or contracted to an organization which specializes in such systems. The former would require the purchase and maintenance of hardware, employment and training of specialized staff, and future upgrades. The longevity of a cyberinfrastructure system is discussed in Section 5.0.

Table 12: Costs for storage and processing associated with several major cloud hosting providers.

Service	CPU Cost	Memory Cost	Storage Cost
AWS	36 core 60 GB / $8,293 USD/Year		$600 USD/TB/Year
Azure	64 core 256 GB / $28,032 USD/Year		$312 USD/TB/Year
Bluemix	32 core 32 GB / $13,122 USD/Year		$2,400 USD/TB/Year
Compute Canada	Per 1 core + 4 GB / $155 CAD / Year		$55 CAD/TB/Year
Google Cloud	$203 USD/vCPU/Year	$28.2 USD/GB/Year	$490.8 USD/TB/Year

As such, employment of an organization (commercial or not-for-profit) specializing in hardware infrastructure was considered to be of greater cost effectiveness and efficiency, particularly given the expertise and economies of scale available. Table 12 identifies the costs associated with some of the major cloud hosting providers. Information was compiled from the respective websites of the organizations and via a conference call with Compute Canada.

We engaged in ongoing conversations with ACENET / Compute Canada during this IE, and discussed a range of hardware solutions and scenarios. They indicated their commitment to working with an eventual CIOOS implementation to identify solutions that meet CIOOS needs, and as the table shows, they are doing so at highly competitive pricing.

5.0 System Longevity

Major Science Infrastructure, such as astronomical observatories, large vessels, and nuclear reactors, is typically designed to last between 25 and 50 years, after which potentially costly upgrades and modernization would be required. The case of an ocean observing system is no exception, and the design of an infrastructure capable of supporting an integrated ocean observing system must take a number of considerations into account. Computer technology tends to have a shorter lifespan than other science infrastructure.

Table 13: The expected lifetimes for major cyberinfrastructure components.

Elements of a Data Management System	Duration (Years)
High-Level Design, Topology, External Environment	Lifetime
Software Architecture	10-15
Programming Language	10+
Operating System	10
Storage Technology	8-10
Main Software Element Design	7
Computers Running Software	4-5
Storage System	3-5

Given the expected duration of cyberinfrastructure components, based on experience from other Major Science Infrastructure (Table 13), managers must upgrade both hardware and software to new specifications. The system will necessarily be in a constant state of change, thus clearly demonstrating the need for continuous funding. Failure to support this technology refresh cycle will result in early obsolescence and ultimately an increase in the cost of operations:

> (1) Cost of maintenance for old hardware will increase; for example, operation of numerous small disk drives as opposed to a few large disk drives.
>
> (2) Continued operation of legacy software may be problematic when hiring developers, as finding an individual with knowledge of an outdated programming language or operating system may be difficult.
>
> (3) Novel instrumentation design or radical changes to instrument methodology may present difficulties to the continued operations of the system, as they may be incompatible with the assumptions which led to the construction of the system at that date.

It is necessary to excogitate the system over time, such that its capabilities are aligned with the specifications of state-of-the-art software and hardware. Annual operational costs are

maintained at 10-20% of overall operations and maintenance costs, which provide sufficient funds to maintain the system at a level competitive with similar international facilities and also sustains the larger initial investment.

6.0 Recommendations

The recommendations for the cyberinfrastructure of a Canadian Integrated Ocean Observing System (CIOOS) are detailed in the following sections. When applicable, standards and tools are delineated into a low, moderate, or high service model. Each tier builds upon the previous one, and as such includes all recommendations from lower service models.

6.1 Canadian Integrated Ocean Observing System (CIOOS) Structure

CIOOS itself will consist of several layers. Topmost is the national portal, which governs the underlying regional associations – which themselves overlay the regional or thematic nodes, to which data is fed by the data provider. The structure of CIOOS is consistent across all service models. The CIOOS component hierarchy is illustrated in Figure 1.

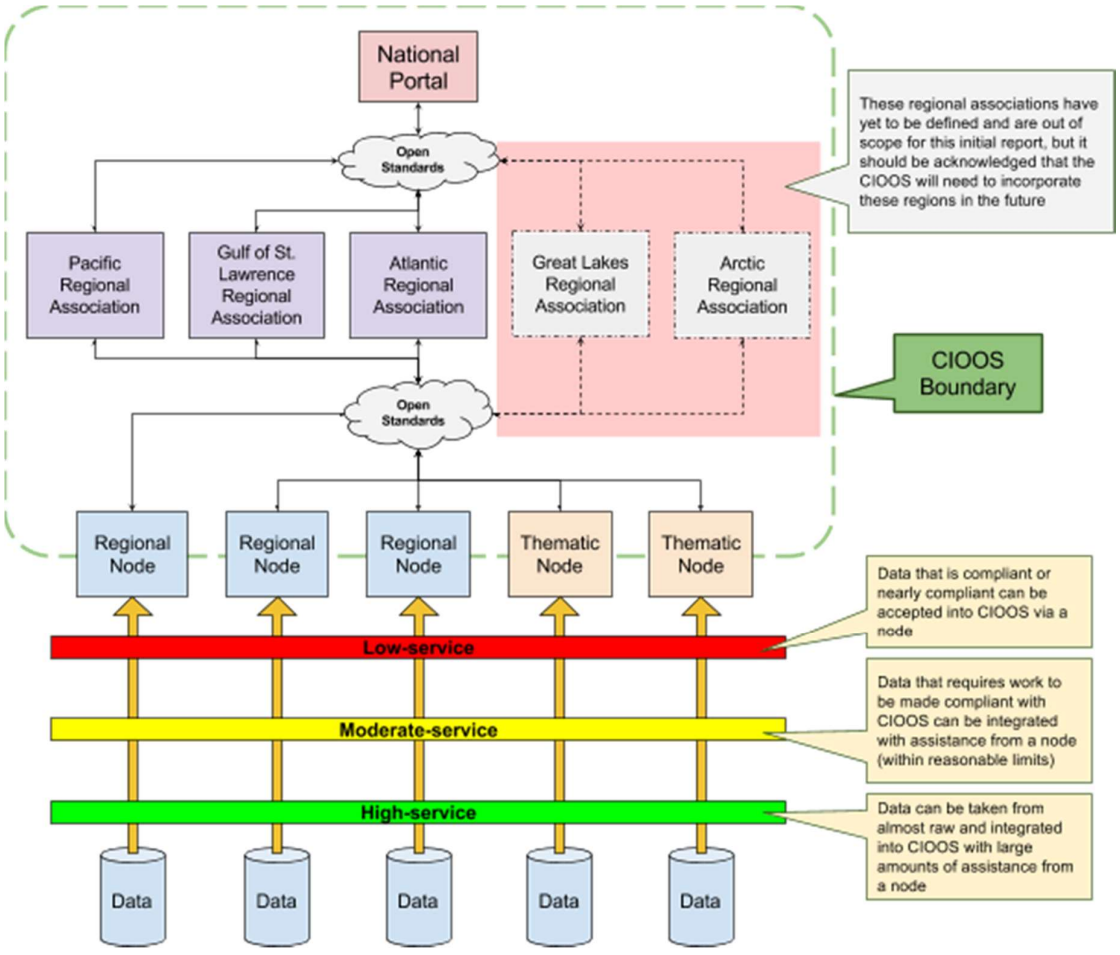


Figure 1: The CIOOS component hierarchy. Data feeds into regional and thematic nodes and as aggregated by the regional association corresponding to the region from which the data originates. The national portal also aggregates data for the purposes of visualization.

6.1.1 National Portal

The CIOOS National Portal will not be a source for data ingestion; that is the role of the regional and thematic nodes. It will instead be a federated collection from the Regional Associations (RAs), capable of storing aggregated data, metadata, and visualizations[71] for the purpose of national discovery, improved visualizations, and load balancing. With all CIOOS data accessible through this single portal, it may be considered the hub for the overall system.

6.1.2 Regional Associations (RAs)

Regional associations will interface with and / or manage Regional Nodes (RNs), providing governance for those RNs under their purview and ensuring that all datasets provided meet CIOOS standards. RAs are responsible for managing relationships with local stakeholders, and represent regional interests at the national level.

This structure is recommended because the maintenance of relationships with regional nodes is vital for facilitating engagement with CIOOS, and the regional associations are better equipped – as compared to the national portal – to understand local needs and successfully work with RNs. This is also consistent with the best practice of keeping data close to source.

Regional associations will host their own regional metadata catalogues, visualizations, interoperable services, and data dissemination capabilities. Although data hosted by RAs will typically be confined to information applicable to their region, datasets which cross boundaries, because they contain information pertaining to multiple regions, should be aggregated and discoverable in each applicable region.

Methods will need to be developed to identify and replicate inter-regional datasets. Interoperability between regional associations, the national portal, and international portals will be achieved through the use of open standards.

6.1.3 Regional Data Nodes (RNs)

Regional data nodes are the primary point of ingest for data and metadata into CIOOS, and all data ingested must be compliant with CIOOS standards. Whenever possible, any data transformation and manipulation required to become compliant with CIOOS should be performed at the node level. The amount of support provided to assist a data provider in reaching compliance will vary by service model (Section 6.4).

A node may be tightly coupled with a regional association, wherein it is difficult to make a distinction between them, or a separate entity working with or under the governance of an RA. In the latter case, the relationship may be a loose one, with only level-of-effort support to the RN. The history of US IOOS has illustrated that the relationship between RAs and regional nodes is prone to fluctuation – they tend to evolve and change over time, and varies between regional associations.

[71] Aggregation of data, metadata, and services will be handled via open standards to promote interoperability.

6.1.4 Thematic Nodes

Thematic nodes will exist to handle and house data that is outside the scope of any one regional node. Thematic nodes may apply to specialized datasets (e.g., gliders, high-frequency radar) and/or the incorporation of national datasets (e.g., federal government datasets). Management of a thematic node will be situation-dependent; it may be handled by a regional association with a strong background in that type of data, or as a collaboration between multiple RAs and their respective nodes.

6.1.5 Other

6.1.5.1 Restricted Data

It is strongly recommended that CIOOS not incorporate restricted data into its system because of the expense and complications this kind of data presents. CIOOS should only deal with datasets that are open and available for public consumption.

6.1.5.2 Attribution

Our US counterparts have made compelling arguments regarding the importance of proper attribution of data sources, as they have a direct impact on funding for data providers and the proper recognition of the contributing organizations. In many cases it may be necessary to credit several organizations for a single dataset. For example, different organizations could each be responsible for different steps in the acquisition, transformation, dissemination and visualization of data. That chain of attribution should be accounted for within both the metadata and CIOOS overall.

6.1.5.3 Bilingualism

Interfaces to CIOOS RAs and the national portal should be made available in French and English, as well as select metadata fields, to enable the bilingual discovery of data. It is important, however, not to impose language constraints on the data itself or on data-associated information. Standardized international controlled vocabularies would remain unchanged, for example, and contextual information provided in a separate document would not need to be translated.

6.1.5.4 Interoperable Design

Interoperable design is the concept that systems, all the way from sensors to RAs, are planned from the start to be able to communicate freely with each other. CIOOS should promote a ground-up approach in interoperable design. Such activities may include working with sensor manufacturers to educate them on CIOOS infrastructure, which may lead to sensors that are ‘plug-in ready’.

6.2 Security

No system is impregnable, and it is necessary to secure the infrastructure and endpoints of CIOOS as much as possible to discourage potential hackers. There are numerous factors to consider, and it is impractical to discuss them all within this document. Nonetheless, those

issues which are expected to either create a significant risk or be a common concern are discussed.

1. Administration interfaces are a common component across software systems. These should either be disabled or hidden from external exposure unless absolutely necessary.

2. All default administrator passwords for software systems should be changed after installation and, if possible, administrator accounts should be removed from the attack surface.

3. Software and servers should be updated and patched to the degree it is reasonable. Because patches may adversely affect the system, allow for a cool-down period on all but critical patches to avoid unplanned outages. Deleterious effects of updates can be mitigated by applying them first to software operating in a sandbox and not to the operational systems.

4. Only services, ports, and endpoints which are in use should be enabled.

5. Some software may require a login from external users, such as those users who need to maintain their metadata in the metadata catalogue. It is important that these users are not granted permissions for tasks beyond what is needed for their authorized activities.

6. Transactional services offered by some Open Geospatial Consortium (OGC) standards, which allow the creation, updating and deletion of content, should be disabled by default. If a compelling reason exists for these services to be enabled, precautions must be taken to secure the end points and data which could be affected by either a careless mistake or a malicious attack.

7. Any traffic which uses or requires authentication should implement Transport Layer Security (TLS) / Secure Sockets Layer (SSL) encryption to prevent electronic eavesdropping. Outdated and compromised ciphers and encryption suites should be disabled to prevent the subversion of these encryption protocols.

8. Internal service accounts should implement meaningful usernames and strong passwords or passphrases. Service accounts should be granted only the minimum level of access required to perform their necessary tasks, and root or administrator accounts should never be used as service accounts.

9. The amount of data an individual user can consume at any one time should be throttled, such that no one user can choke the system. Without such limits, denial of service attacks – whether intentional or unintentional – would be difficult to defend.

10. The aforementioned security practices should be considered during development of any software meant for use with CIOOS.

Security training and education for CIOOS employees is important. Security starts and ends with the people who manage the system.

6.3 Software Development

While constructing the regional associations and national portal, any code developed should be tracked using a revision control system such as GitHub, and made public so others may make use of the tool. GitHub is a web-based version control system providing the *git* version control system as-a-service, and is free-to-use for open source projects. It may fulfill this need; it is used by US IOOS for this purpose, for example. GitLab, open source software for a self-hosted version control code repository similar to GitHub, is also an option.

Development of implementation guides and white papers to help educate data providers would be beneficial. Although standards provide excellent interoperability, there are few resources which allow individuals to bridge the gap between knowing and doing. Open Geospatial Consortium (OGC) is eager to collaborate on such initiatives, which could be fruitful for both parties.

Software development may occur in-house, or be contracted out to private consulting companies. The report produced by the Data and Observations Investigative Evaluation provides a list of companies which may hired to perform such work. It should be noted, however, that software development is expensive. As such, public tools should be utilized where possible and software developed to aid data ingestion should minimize 'one-off scripts' for data parsing.

6.4 Service Models

There are several service models which may be offered by CIOOS. The differences between each relate to: the amount of support available to data providers to achieve compliance with CIOOS standards; the software and standards employed; and the hardware necessary for storage volume and processing power. Each service model builds off the previous.

6.4.1 Recommendations Overview

Three tiers of service for CIOOS cyberinfrastructure have been described, and are summarized in Table 14. Detailed explanations are provided in the subsequent sections. The cost associated with each service model is detailed in Appendix D.

6.4.1.1 Software and Standards

To ensure interoperability with international Ocean Observing Systems (OOSes) and the ability to contribute to the Global Ocean Observing System (GOOS), it was considered that standards and software infrastructure must comply with national and international data standards, while also supporting those requirements specified in the reports produced by the

Table 14: Summary of the low, moderate, and high service models proposed for CIOOS, wherein each successive tier provides additional functionality to the system.

Portal	Feature	Standard / Software	Low Service	Moderate Service	High Service
National	Catalogue	CSW + WAF	✓	✓	✓
	Catalogue Aggregator	WAF Harvester	✓	✓	✓
	Data Server (Internal)	ERDDAP	✓	✓	✓
	Data Aggregator	ERDDAP	✓	✓	✓
		Custom Tools	As Needed	As Needed	As Needed
	Visualization Platform	See Vis IE	✓	✓	✓
	Standards Requiring Additional Software	WFS		As Needed	As Needed
		WCS		As Needed	As Needed
		SensorThings API			As Needed
		SOS			As Needed
		Other		As Needed	As Needed
	Hardware Requirements		CPUs HDD/SSD RAM	RAM+ Hybrid RAID	As Needed
Regional	Catalogue	CSW + WAF	✓	✓	✓
	Catalogue Aggregator	WAF Harvester	✓	✓	✓
	Data Server (Public)	ERDDAP	✓	✓	✓
	Data Aggregator	ERDDAP	✓	✓	✓
		Custom Tools	As Needed	As Needed	As Needed
	Visualization Platform	See Vis IE	✓	✓	✓
	Standards Requiring Additional Software	WFS		As Needed	As Needed
		WCS		As Needed	As Needed
		SensorThings API			As Needed
		SOS			As Needed
		Other		As Needed	As Needed
	Hardware Requirements		CPUs HDD/SSD RAM	VMs RAM+ Hybrid RAID	As Needed
	Support for Data Submission		Low	Moderate	High

Data and Observations and Visualization Investigative Evaluations. Both open source and commercial solutions were considered.

Five major software components are required for an ocean observing system to effectively aggregate, search, and serve data: a catalogue service; a catalogue aggregation service; a data serving system; a data aggregation service; and one or more visualization platforms. Table 15 illustrates the manner in which the five necessary components are distributed between the national portal and regional associations.

Table 15: The distribution of necessary components for an ocean observing system between the national portal and the regional associations.

	National Portal	Regional Associations
Cataloguing Service	✓	✓
Catalogue Aggregation Service	✓	
Data Serving System (Internal)	✓	
Data Serving System (Public)		✓
Data Aggregation Service	✓	✓
Visualization Platform(s)	✓	✓

The Cyberinfrastructure IE recommends the use of open standards as opposed to specific software when possible. This will minimize the effort required from Canadian institutions to become interoperable with CIOOS and also provide the interoperability required for the national portal to harvest relevant metadata catalogues and communicate with international partners. The relationships between the recommended standards are illustrated in Figure 2.

6.4.1.2 Hardware

To ensure CIOOS is capable of ingesting and storing data from regional nodes,[72] sufficient hardware must be utilized to support the core variables and visualization needs as outlined in the reports by the Data and Observations and Visualization Investigative Evaluations. Both private and public cloud providers were considered. Given the costs outlined in Section 4.0, the recommended provider is Compute Canada.

Compute Canada services include multi-site backups, maintenance and operation of the hardware, and user support. Given the costs associated with employing technical staff, the ability to outsource is of great benefit. An additional important consideration is that Compute Canada's storage sites allow high bandwidth speeds, multiple uplinks, and do not meter connections.

[72] Only when necessary. If a data provider is already compliant with CIOOS standards, there might not be a need to ingest its data into a different repository; a metadata catalogue may be sufficient. There is the possibility that an independent data production node or observatory is a much bigger organization than its regional association, in which case the regional association might leverage the capabilities of this pre-existing data provider.

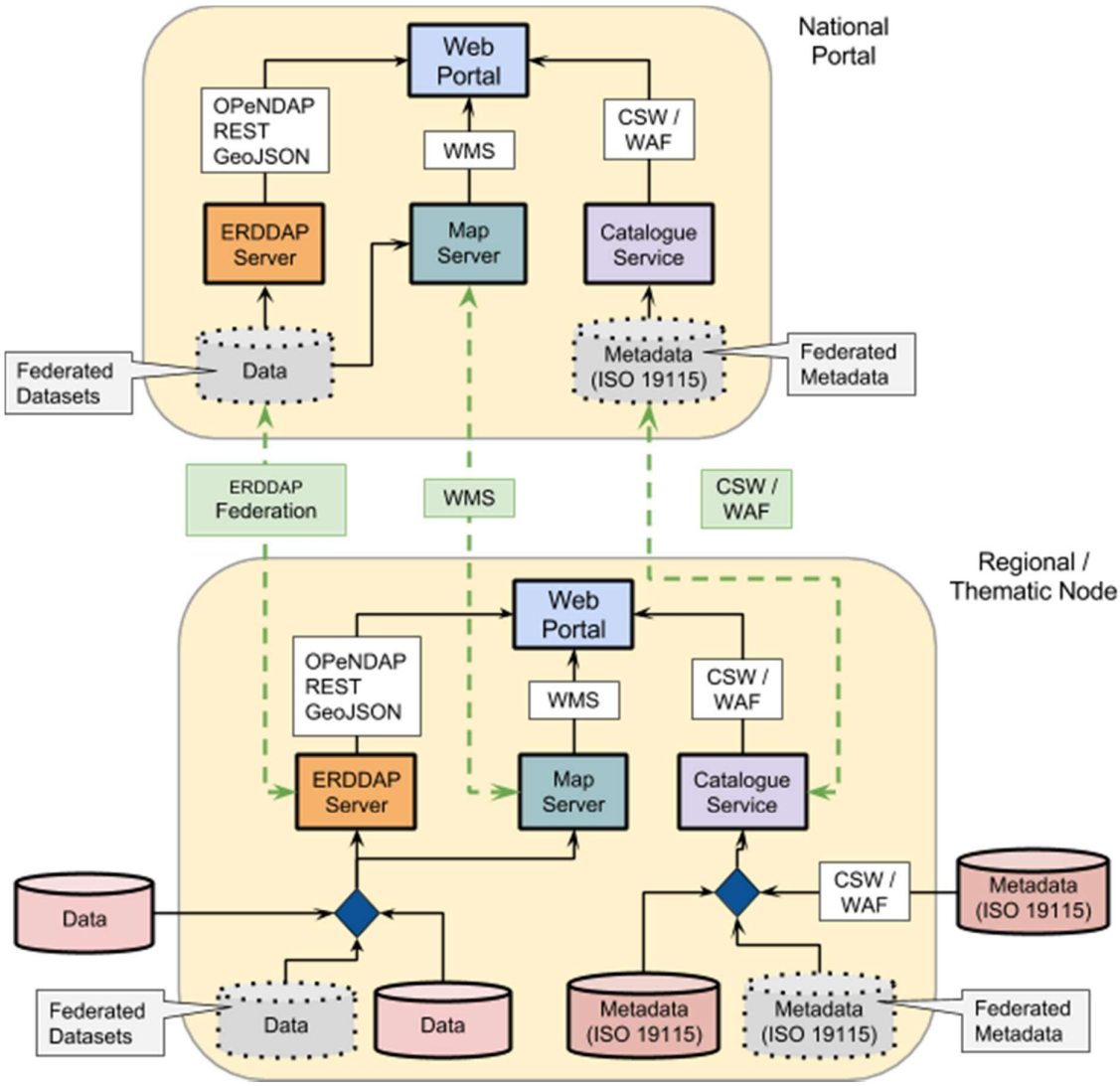


Figure 2: Relationships between the open standards recommended for use in CIOOS, from data providers to regional nodes to the national portal.

Compute Canada is adopting a cloud-service model in addition to the long-standing batch-processed high-performance computing capacity for which it is best known. Virtual machines, local and object storage, and on-demand provisioning provide a flexible hardware infrastructure that can expand and contract to meet changing demands.

While Compute Canada is best known for providing compute infrastructure to Canadian researchers at no cost to them, this model requires annual (or tri-annual) re-applications, and there is no guarantee that requests for storage and processing power will be met. We therefore recommend a pay-for-service model, which guarantees ongoing access to a CIOOS-specified hardware service.

It is recommended that CIOOS utilize Compute Canada's existing stack and work with the organization as a pay-for-service partner during the pilot phase (Section 8.1), as lessons learned at this stage will enable informed decisions for subsequent phases. It is anticipated that the best route forward after the pilot phase be that CIOOS purchase hardware through Compute Canada and purchase management services from the organization on an ongoing basis.

This purchase enables custom hardware configurations to meet anticipated CIOOS-specific requirements, such as high-speed reads from persistent storage to support visualization products, which might require hybrid HDD/SSD storage, in-memory caching, RAID10, and other non-standard elements in the hardware stack. If Compute Canada is able to meet requirements with existing or planned hardware purchases, a simple pay-for-service model may still be appropriate.

It must be noted that because Compute Canada is funded by the Canada Foundation for Innovation (CFI),[73] the project must involve CFI-eligible researchers from its inception.

6.4.1.3 Data Submission

Data contributors to CIOOS must make their data available using interoperability methods compatible with published CIOOS standards. Metadata should be as complete as possible and adhere to CIOOS metadata standards. Metadata must pass compliance checks prior to submission. CIOOS will provide the compliance checking tools required. Observation data must be submitted in a standardized format.

6.4.1.4 Support Staff

With regards to staffing, both the regional associations and national portal will require technical and administrative staff. The cyberinfrastructure is not merely a collection of hardware and software required to run CIOOS, but also the people required to ensure smooth operation.

Technical staff are required to maintain the software and develop specialized tools for data ingestion and data visualization as needed; these may include full-featured data serving tools for challenging data types such as audio and video. Staff specialized in data curation of metadata and standard compliance are also necessary, as are user support staff to ensure barriers to submission and use are as low as possible.

The effort required to ingest data from a given region will be dependent on the scope of the regional association, which includes: the number of regional nodes supported, level of data maturity and metadata completeness at each node, the number of local stakeholders requiring support, and more.

[73] https://www.innovation.ca/

6.4.2 Low Service

6.4.2.1 Software and Standards

The *Core Interchange Standards* are those baseline tools and standards required for successful implementation of an integrated ocean observing system. They are summarized in Table 16, and explained in detail in the subsequent sections.

In edge-case scenarios where the *Core Interchange Standards* are insufficient – they cannot handle the data, or do not provide the desired functionality – a regional association may recommend additional standards to be used alongside the core standards. Prior to selection of a new standard, all regional associations should be consulted and allowed to provide input.

Table 16: Core Interchange Standards for CIOOS.

Cataloguing Service	WAF & CSW supported catalogue
Catalogue Aggregation	IOOS WAF harvester
Data Dissemination	ERDDAP, WAF, and specialized tools as needed
Data Aggregation	ERDDAP, WAF, and specialized tools as needed
Visualization Tools	See Visualization report

It is anticipated that there will be specific data types for which there is no existing interchange standard, or situations wherein all existing standards lack a desired functionality. In such cases, consultation with existing standards bodies, to find the optimum path to fulfill the need, is recommended (Appendix E).

6.4.2.1.1 Cataloguing Service

Catalogue services which provide capabilities to index and search a collection of dataset metadata records and to direct users to relevant resources for each dataset are an integral component of an integrated ocean observing system.

Since numerous catalogues employ the same standards, it is not essential to recommend a specific catalogue to fulfill the cyberinfrastructure requirements. The CI IE instead recommends that all institutions and associations involved with CIOOS support the catalogue exchange standards Catalogue Service for the Web (CSW) and Web Accessible Folder (WAF), thus providing interoperability between nodes and regions while allowing organizations to select the software most beneficial to their individual needs. Results from the CIOOS Planning Survey support this recommendation (Section 2.3).

CSW is recommended because it is a well-supported OGC standard. It is specialized for geospatial datasets and includes native support for the OGC visualization standards Web Map Service (WMS) and Web Feature Service (WFS), satisfying requirements put forth by the Visualization Investigative Evaluation. Deficiencies in CSW include lack of support for non-geospatial datasets and limited support for the exchange of some metadata standards.

As a supplement to CSW, all metadata should be published to a well-structured WAF to facilitate data harvesting amongst nodes and with external CIOOS partners and collaborators. It may also be employed to circumvent the limitations of CSW; it can deliver difficult-to-manage datasets such as acoustic and video,[74] and places few restrictions on the metadata standards which may be exchanged.

The use of WAFs should be securely configured and governed by well-structured conventions. Metadata and data hosted in WAFs should have logical separation and adhere to the folder and file naming conventions outlined in the report generated by the Data and Observations IE.[75]

To properly function, it is necessary that a database underlie the catalogue. The database software recommended is dependent on the catalogue utilized; Comprehensive Knowledge Archive Network (CKAN) documentation instructs the use of PostgreSQL with the PostGIS extension, for example, while GeoNetwork ships with H2 but can use others. However, when handling a large number of records, the use of PostgreSQL with PostGIS is recommended. The Cyberinfrastructure IE is therefore not proposing a specific database, but instead recommending that the catalogue documentation be followed.

6.4.2.1.2 Catalogue Aggregation Service

An integrated ocean observing system requires that the national portal pull from the regional associations to create a central listing of all datasets within the system. The resultant aggregate catalogue allows for creation of national-level visualizations and provides users a means of searching datasets across all regions.

It is recommended that CIOOS adopt a tool which is capable of aggregating CSW and WAF enabled catalogues. And because open standards are recommended for exchange, one tool is capable of aggregating many different catalogues. CIOOS may develop a tool in-house, or employ an existing aggregator – such as that available from US IOOS.[76] The IOOS harvester aggregates from multiple WAF listings and builds a central CKAN catalogue.

It should be noted that automation of catalogue aggregation requires that a common metadata profile is used. This is explored further in the report produced by the Data and Observations Investigative Evaluation.

6.4.2.1.3 Data Dissemination System

A required component of an ocean observing system is a data serving system; to provide users a method for accessing data directly from the hosting regional association and to programmatically feed data into both regional and national visualizations. The system must also support GeoJSON to satisfy the requirements of the Visualization IE.

[74] Although specialized tools from other disciplines may be used to serve audio and video data, they are outside the initial scope for CIOOS.
[75] https://ioos.github.io/catalog/pages/registry/waf_creation/
[76] https://github.com/ioos/catalog-harvesting

The ERDDAP implementation of OPeNDAP is recommended. As a specific software, ERDDAP breaks from the Cyberinfrastructure IE convention of recommending standards, but as the only data server which provides aggregation and automated dataset caching (Section 2.2.2) it was deemed essential.

Not only does ERDDAP provide the necessary data dissemination features and native GeoJSON support, it also supplies users with a method to query data subsets, provides an Application Programming Interface (API) which may be used in the development of data products such as web applications and visualization tools, and supports data serving for a wide range of file formats (Section 2.2.2).

Because numerous file formats are well-supported within ERDDAP, a standard file format is not required for CIOOS. This reduces the effort required from data providers to convert data into CIOOS-compliant forms, and has the potential to encourage interaction with and contribution to the system.

Certain data types will be unable to fit well within the ERDDAP standard. Such data include ambient sound and video; these may be handled using WAFs (Section 2.1.4.3), presuming datasets are of a size where transfer over HTTP is possible. The THREDDS implementation of OPeNDAP may be used to complement ERDDAP in special cases (Section 2.2.2.3).

ERDDAP is also capable of filtering on the quality control flags which Quality Assurance / Quality Control of Real Time Oceanographic Data (QARTOD) – described in Section 2.1.4.4 and recommended by the Data and Observations IE for both the moderate and high service models – adds to each data column, thus providing users a means to search based on quality.

Unlike the metadata catalogue, it is not necessary for a database to underlie ERDDAP; in fact, use of a database is actively discouraged. Instead, the Cyberinfrastructure IE suggests that implementation follow the recommendations in ERDDAP documentation, which specifies the use of netCDF files for data storage:

> "NetCDF-3 .nc files are a good, general recommendation because they are binary files that ERDDAP can read very quickly. For tabular data, consider storing the data in a collection of .nc files that use the CF Discrete Sampling Geometries (DSG) Contiguous Ragged Array data structures and so can be handled with ERDDAP's EDDTableFromNcCFFiles. If they are logically organized (each with data for a chunk of space and time), ERDDAP can extract data from them very quickly."[77]

[77] http://coastwatch.pfeg.noaa.gov/erddap/download/setupDatasetsXml.html#ServingTheDataAsIs

6.4.2.1.4 Data Aggregation Service

The fourth component necessary for an integrated ocean observing system is a data aggregation service: to aggregate datasets which contain information relevant to multiple regions, and to aggregate data nationally so as to guarantee efficient and reliable national visualizations. The ERDDAP implementation of OPeNDAP is recommended.

ERDDAP possesses federation capabilities, wherein the protocol can federate from any DAP-compliant server and intelligently aggregate and cache their data to minimize load on the system and maximize reliability (Section 2.2.2.2). As an example the datasets hosted at Dalhousie University and Marine Institute were federated with only a half-days' worth of effort, solely because both institutions used ERDDAP as their data servers. Because ERDDAP is capable of both data aggregation and data serving, the complexity associated with implementing an additional software or standard is avoided.

Although WAF is recommended as a catch-all for serving datasets unsupported by OPeNDAP, the WAF standard is not capable of data aggregation. If it becomes necessary to aggregate such datasets, development of a WAF-based aggregator would be required.

6.4.2.1.5 Visualization Platform(s)

The final necessary software component is one, or more, visualization platform to ease data discovery and to advertise the ocean observing system. Recommended visualization platforms and supporting tools and standards are described in the report produced by the Visualization Investigative Evaluation.

6.4.2.2 Hardware

The Data and Observations Investigative Evaluation has outlined numerous core variables which are necessary to successfully implement CIOOS (Table 17). Those highlighted in grey have the potential to be data-intensive, requiring significant storage volume.

Table 17: Core Variables for CIOOS as recommended by the Data and Observations IE.

Ice	Seagrass Cover	Surface Heat Flux	Phytoplankton
Nutrients	Live Coral	Surface Stress	Oxygen
Ambient Sound	Sea Surface Height	Zooplankton	Bottom Type
Temperature (Surface & Subsurface)	Salinity (Surface & Subsurface)	Inorganic Carbon	Sea State
Currents (Surface & Subsurface)	Fish Abundance and Distribution	Marine Mammals	

It is recommended that the pilot phase begin with 16 Compute Canada cores with 64 GB of RAM. Because Compute Canada's existing stack is being utilized (Section 6.4.1.2), any additional resources required may be added seamlessly and with little or no downtime. Should a common sandbox be deployed as per Section 8.1, it is recommended that a separate Virtual

Machine (VM) with similar specifications be used, with on-going load balancing as needed. Hard disk drives with a Redundant Array of Independent Disks (RAID) controller is recommended to create redundancy for post-pilot phases.

The specific hardware and processing requirements for post-pilot phases of CIOOS will be informed by the lessons learned during the pilot phase. Estimations for storage volumes are elucidated in Section 3.0, but are based on mature organizations and as such are not applicable to the developing system that is CIOOS.

Hardware to provide performant visualization services for large datasets is not included in the low service model. The pilot phase hardware is sufficient to support the recommendations of the Visualization Investigative Evaluation, although load times for large datasets may be slow.

6.4.2.3 Data Submission

For the low service model only data and metadata which are compliant or nearly compliant can be accepted. A compliance checker will be supplied to assist data providers. Compliant data corresponding to the supported core variables will be ingested directly into the CIOOS framework with minimal effort.

In the low-service model there is little support available to assist data providers in the transformation and manipulation of data. They are responsible for ensuring that the data are made available in the applicable relevant formats and adhere to CIOOS terminology, coordinate reference systems, and units of measure. It is assumed the data provider will have some technical capacity to perform the necessary actions.

6.4.3 Moderate and High Service

6.4.3.1 Software and Standards

The *Core Interchange Standards* may be extended at higher service models to complement and expand the ability of CIOOS to share and interoperate with ocean observing data. Standards at models above low service will be implemented as needed, and as such are not separated by tier. Examples of obvious standards which may be implemented are illustrated in Table 18; high service will enable all listed standards (and more), whereas moderate service will enable only some.

Table 18: Additional standards which may be implemented in the moderate and high service models of CIOOS.

Moderate Service	OGC Web Feature Service (WFS)
	OGC Web Coverage Service (WCS)
High Service	OGC SensorThings API
	OGC Sensor Observation Service (SOS)

WFS and WCS both expand upon the capabilities provided by the *Core Interchange Standards*. They increase the interoperability of CIOOS by expanding the capabilities for sharing feature data and enabling the sharing of coverage data, respectively.

Sensor Observation Service (SOS) is a mature standard for sharing interoperable sensor data with other ocean observing systems which also make use of the standard, such as US IOOS. Consultations with other OOSes have indicated that SOS is a terminal system, and is expected to fall out of use in the coming years. Implementation would primarily be for interoperability with US IOOS, which currently employs the standard. As such, SOS is recommended only if resources are available after implementation of other standards.

The SensorThings API is a standard for interoperability with geospatially-enabled Internet of Things (IoT) sensors, data, and applications. It is capable of creating an SOS-compliant service.[78] However, as discussed in Section 2.1.1.6, it is a relatively new standard and support is limited. Given the uncertainty associated with this standard, caution is recommended. Mature standards should be explored before implementation of the SensorThings API is considered.

6.4.3.2 Hardware

Although the recommended software and standards are not significantly different between the moderate and high service models, such is not the case for the Visualization and Data and Observations recommendations, and it is on these which the hardware requirements depend.

6.4.3.2.1 Moderate Service

The Data and Observations Investigative Evaluation has outlined seven additional variables to be included in the moderate service model, in addition to those outlined in the low service model (Section 6.4.2.2). None are expected to produce significantly large volumes of data, and will have minimal impact on the required storage volume.

However, given the additional support available to data providers (Section 6.4.3.3.1), the moderate service model is expected to encounter data which requires intensive manipulation. It is necessary that CIOOS have computing power sufficient to meet this need.

Specific actions which are computationally expensive include the formatting and converting of large datasets and recalibration of months of data from large datasets, often necessary after identification of a sensor calibration issue. To manage the load from such processes, it is recommended that CIOOS employ several virtual machines; a node may require 6 to 8 VMs with 8 cores and 64 GB of RAM each, along with significant space for temporary data storage.

The moderate service model of CIOOS will also provide performant visualization services for large datasets. Through consultations with numerous web service and visualization providers

[78] http://www.opengeospatial.org/projects/initiatives/imisiot

it was discovered that the primary bottleneck for ocean observing systems is data querying and retrieval. As such, 256 to 512 GB of RAM is recommended. This would significantly improve responsiveness of the web portal and minimize the load times experienced by users.

If necessary a hybrid caching RAID controller may also be employed. This controller would intelligently cache heavily used storage segments on a small handful of SSDs to improve performance. Note that the numbers are dependent on the number of end-users.

6.4.3.2.2 High Service

A further seven variables – in addition to those included in the low and moderate service models – have been outlined by the Data and Observations IE for inclusion in the high service tier, along with a recommendation to review and develop additional variables on an ongoing basis. None are expected to be data-intensive, and will therefore have minimal impact on the required storage volume.

The hardware requirements specified in the moderate service model are sufficient to support the high service models specified by the Data and Observations and Visualization IEs. Hardware requirements for the high service model in phase one of CIOOS are therefore the same as those outlined for the moderate service model (Section 6.4.3.2.1). In subsequent phases, however, the high service model will allow CIOOS to provide exceptional service to the marine sector.

In phase two, it is recommended that several Graphics Processing Units (GPUs) be added to aid in the development of solutions which are well served by parallel computing, such as ocean simulations for research or the output of weather models. Computer simulations for wave state, wind speed, and water currents have numerous applications which include ship design, breakwater design, and the design of oil platforms. Water current and wind speed simulations also have applications in oil spill response and search and rescue.

The supercomputer employed for creation of the COSMO-1 model – employing 96 NVIDIA Tesla K80 GPU accelerators and 24 Intel Haswell CPUs[79] – is an example of an exceptional system for weather modeling. Using published prices at the time of this writing, an estimate of hardware cost for such a system is approximately $500,000 CAD. It is therefore recommended that GPUs be employed from the Compute Canada pool as needed. This may necessitate that initial simulations be run at lower grid resolutions; those which exhibit success may be tested at higher resolution during periods where Compute Canada's resources experience less demand. Dedicated hardware may be purchased once a simulation has proven to have significant merit.

6.4.3.3 Submission

Unlike the software and standards recommendations, the ingestion support available to data providers differs between the moderate and high service models.

[79] https://www.hpcwire.com/2016/04/01/swiss-supercomputer-weather-forecasting-now-fully-operational/

6.4.3.3.1 Moderate Service

Regional associations will have the resources to provide tools and training to non-CIOOS compliant data providers, assisting them in bringing their data and metadata into the system. This could involve data transformations or conversions that are readily available with trusted software, such as those discussed in Section 2.2.4.

The ability for regional associations to provide curation efforts benefits all data providers, but particularly those with limited technical support who would otherwise face great difficulty when attempting this task.

6.4.3.3.2 High Service

Regional associations will have the resources to ingest raw or nearly-raw data. Data providers can expect significant assistance in ensuring their data and metadata is compliant with CIOOS standards. Custom software may be required. Staff will be available to assist researchers in accessing and manipulating data. Principal Investigator's will be given a workspace with adequate processing power to aid in their research.

For the purpose of this IE, both data curation and data rescue are supported in the high service model.

7.0 Out of Scope Items

Throughout the course of the cyberinfrastructure evaluation, numerous concerns and considerations were raised which are out of scope for the Cyberinfrastructure IE, or out of scope for the initial implementation of the system. They are nonetheless important aspects to note, and are discussed below.

7.1 Cross-Cutting Activities

CIOOS is an ocean observing system which integrates oceanographic data from organizations across Canada. Although each organization and region has specific interests and needs, there exist elements which are common across these institutions and associations. The need for cross-cutting activities arises, wherein particular activities are joint efforts involving individuals from all relevant organizations and regions.

Cross-cutting activities expected to be of relevance to CIOOS include regular design meetings between regional associations and with other national organizations, as a means to help unify disparate groups and provide knowledge cross-pollination. Attendance at national and international conferences and workshops, such as the Canadian OGC Summit,[80] is recommended so CIOOS members can stay abreast of the latest technological advancements within Ocean Observing Systems.

[80] http://www.opengeospatial.org/event/170626canadian

To preserve regional expertise while maintaining centralization at the national portal, it has been suggested that a centralized IT group be developed and that it be democratically assigned to work on projects that benefit most or all of the regions. Expertise from the regions would contribute to the work done by the centralized group. This cross-cutting requirement should be covered under overall governance of the system.

Meaningful metrics of success for CIOOS as a whole must also be determined, with consideration given to activities of the various regions.

7.2 User Engagement

Although user engagement is outside the scope of the investigative evaluations, it is nonetheless an important aspect. Done well, user engagement has the potential to increase participation in CIOOS and increase overall utility of the system. Strong engagement from the community may assist in the acquisition of funding at both the regional and national levels, and would provide benefit to numerous organizations, including industry, research institutions, and government departments.

User engagement may occur at the national or regional level. Within US IOOS, regional associations have existing ties to local and regional stakeholders, along with knowledge of region-specific users and technologies, which may facilitate engagement. Given the numerous perspectives arising from a broad user base, national efforts have focused primarily on common data products, such as gliders and high-frequency radar.

There exist two perspectives when considering user engagement – the engagement of end users and the engagement of data providers.

7.2.1 Engagement of End Users

Expected end users of CIOOS include both the general population – where engagement may be considered representation of success when determining funding – and those key users who will support the system, financially or otherwise (Table 19). To maximize the probability of strong engagement, efforts must occur early and often, and it is important that CIOOS identify and meet user needs, both regionally and nationally.

Table 19: Examples of key users and general users for CIOOS.

Key Users	General Users
DND	Researchers
GOOS / International Communities	Industry
DFO Internal Stakeholders	Public Citizens

There are numerous methods which may be employed to determine user needs, such as: an annual meeting with partners and stakeholders; region-specific workshops to discuss relevant issues; attendance at national and international oceanographic conferences, with presentations

to spark interest; and attendance at or participation in general stakeholder meetings. The latter may be most beneficial in ascertaining end user needs, as they may provide information about the importance of each issue to the community writ large.

Successful engagement of end users may also encourage the participation of data providers, as strong uptake of CIOOS within the community may assist such organizations in demonstrating knowledge mobilization to private sector or policy influence, and thus positively impact funding opportunities.

7.2.2 Engagement of Data Providers

To engage data providers with CIOOS, it would be beneficial to create a data ingestion process which is both easy and rewarding. Examples of features which may provide such an experience are elucidated in Table 20.

Table 20: Features of CIOOS which may make the system both easy and rewarding for users.

Easy	Rewarding
Tools that Gather all Necessary Information	Archiving
Support / Funds to Assist	Ability to Access and Share Data
Compliance with Research Funding	Citation and Attribution Capabilities

7.2.3 Metrics for Success of User Engagement

When selecting those metrics which indicate successful user engagement, it is necessary to ensure that the information provided is meaningful – that it is not considered merely because it exists. Such metrics require further consideration, but may include a formal cost-benefit analysis,[81,82] evidence-based policy, or analytics such as volume of downloads or number of citations.

7.3 Recommended Features for Subsequent Phases

The recommendations put forth by the investigative evaluations relate only to the initial phases of implementation for CIOOS. Beyond such suggestions there exist elements which are currently too complex for inclusion, though they would increase the robustness of the overall system. Such elements are considered here as potential options for later phases of CIOOS.

7.3.1 User Login

A user login feature – replete with an optional, but encouraged, user profile, individual dataset download history, and the capability to save customized dashboards and filters – would be a valuable supplement to CIOOS. Beyond the general advantages of additional functionality, there exist benefits for both users and system administrators.

[81] http://www.iooc.us/wp-content/uploads/2010/09/IOOS_Report_Volume_I_120503.pdf
[82] https://ioos.noaa.gov/project/ocean-enterprise-study/

Profiles would provide information about CIOOS consumers beyond visitor counts and geolocation via IP. Who is CIOOS attracting, for example – an unexpected set of users, or an expected set of users? Such information may be used to target user engagement activities.

The aforementioned user information, in conjunction with CIOOS dataset download history, may also provide insights regarding which data is being used, and by whom. This may assist data providers with funding applications, as it provides additional details regarding dataset use. Download history may further be employed to inform users when changes are made to a dataset, so they may obtain the updated information if desired.

7.3.2 Improved Visualization of Map Layers

Visualization is an efficient method for data discovery; it provides an overview of available data, allowing users to easily determine potential datasets of interest. Much of oceanographic data is geographic and is visualized via a georeferenced map, generated when a map server pulls data from a geographic information system database.

Map layers are utilized to organize dataset features. Each layer references a specific set, or subset, of datasets, and specifies the visualization parameters such as colors, symbols, and labels. Too many layers creates complexity within the visualization systems, and as such the initial phases of CIOOS will focus on the most common data formats. It is recommended that additional layers, such as visualizations for hydrophone data, be done only after initial implementation is complete.

7.3.3 Online Discovery for Audio/Video

Without options for online preview of audio and video data, these formats present challenges for data discovery. Although the files may be annotated, it can be difficult to determine the dataset(s) of greatest relevance. Often users must resort to the ‘hit-or-miss’ method, wherein all potentially relevant datasets are downloaded and examined.

As the download size of these files is significant, this presents a strain on the system’s bandwidth. And although throttling may be employed to ensure no one user overloads the system, the increased download times may negatively impact user opinions of CIOOS. With an online video and audio preview system, users may examine only those portions of the file which are of interest. The volume downloaded will thus be more reasonable and will present less strain to the system’s bandwidth.

To provide users with enhanced discovery of audio and video datasets, it is recommended that an embedded audio and video player be implemented for a subsequent phase of CIOOS. Should video someday become part of, or enable, a supported variable, such as corals, it will be essential to generalize the use of such preview tools throughout CIOOS.

7.3.4 Alternative Acronyms for Canada's Ocean Observing System

Currently this project is known by the acronym CIOOS: Canadian Integrated Ocean Observing System. However, other alternatives exist that are not as derivative. Although CANOOS appears to be employed for the Canadian Network of Operational Oceanography Systems, there is indication that the program may be obsolete and the acronym currently unused. CANGOOS also remains an option: the Canadian Geospatial Ocean Observing System, or alternatively the Canadian Global Ocean Observing System.

7.4 Modelled Data

Modelled data, such as model outputs, may provide numerous benefits to users. They are also a complex entity requiring significant additional cyberinfrastructure considerations for both storage and management. As such, this data type is out of scope for the initial phase of CIOOS. It is strongly recommended that modelled data be included in subsequent phases of CIOOS.

7.5 CIOOS Compliance Standards

It is recommended that a CIOOS-specific standard compliance matrix be implemented for regional and thematic nodes. The matrix would be necessary to ascertain that the services offered by the data provider comply with CIOOS standards for data access, visualization, and data and service quality. Compliance with the matrix would be used as a condition for receiving on-going funds for continued participation in CIOOS.

Compliance requirements could be inspired by the existing World Data System (ICSU-WDS) membership conditions or by the Data Seal of Approval of DataOne. Those CIOOS membership conditions will cover the gamut of technology used, service level, support of a minimum set of data curation, quality, access, and visualization standards. The specific conditions applicable to a regional node will depend on the features that node might be interested in contributing to or supporting on behalf of the wider CIOOS community.

For CIOOS, such a compliance determination and certification should be managed by the Governance entity.

8.0 Steps to a Phased Approach

8.1 Pilot Phase

The purpose of the pilot phase is to illustrate the proof of concept of the CIOOS framework. In this phase, it is not anticipated that regional associations will be established, but only that a few test nodes will be deployed. A sample national node would connect to one or more regional nodes. The nodes would be harvesting a small number of easy-to-handle core variables from some of its member sites.

Step 1:
Engage with standards bodies to seek partnerships for developing pilot nodes. OGC has previously expressed interest in creating a Canadian sandbox wherein developers may test their applications. Other organizations may have similar interests if approached. A sandbox would be beneficial in the pilot phase by allowing multiple developers to interact with the developing system, maximizing the possibility that major deficiencies will be identified prior to roll out of a production site.

Step 2:
Use the services of Compute Canada to set up a virtual machine containing a catalogue service, an ERDDAP server, a mapping server and Web Accessible Folders as per the recommendations of Section 6.4.2. The VM may be cloned for each organization participating in the pilot phase, simplifying the initial setup. The national node will also employ a VM. This node will be used to federate datasets from the regional nodes and offer sample visualizations.

Step 3:
Begin ingestion of ocean data into the system. A simple subset of core variables, such as surface temperature, would be selected for cataloguing, data serving, and visualization. Although more complex core variables, such as ambient sound, will not be visualized, it may nonetheless be beneficial to make available complex data types on a limited scale, to provide developers an opportunity to tackle them in the sandbox.

Step 4:
Deploy a sandbox as described in Step 1. Put out an open call to developers.

Step 5:
Create a sample national web portal. Although site layout should be complete, the web portal may not be fully functional at this stage. Some form of web analytics, such as Google Analytics, should be deployed to track usage information. Implement visualizations of a simple subset of the core variables identified in Step 3.

Step 6:
Provide access to the pilot national web portal to a group of representative users. Seek user feedback, and analyze system performance and site analytics. Compile a list of lessons learned from the pilot phase, and recommend any necessary design changes to be incorporated in phase one.

8.2 Phase 1

Building on lessons learned and frameworks from the pilot phase, the first phase would implement nodes for the regional associations and a national node. The initial phase would expand on the variables used for the pilot nodes and use secure data sources to minimize the amount of data curation.

Step 1:
Form the regional and national associations with their associated Board of Directors and Principal Investigators, as per recommendations set forth in the Governance report.

Step 2:
Create one or more nodes for each regional association, building on what was learned from the pilot phase. Nodes will federate at the regional association level so that each association has one endpoint from which the national node will pull.

Step 3:
Following recommendations from the Data and Observations report, data from the various organizations will be curated. From a technical perspective this will be one of the more labour intensive tasks of phase one and will require a significant upfront investment in technical staff.

Step 4:
Further develop the national web portal, and develop web portals for each regional association. It is at this stage that design decisions surrounding the web portal – such as web server selection, web content management systems, database selection, and site analytics software – will be made. Creation of the web portals will be guided by recommendations from the Visualization report. It is recommended that this step be completed concurrently with Step 3 so as to ensure compatibility. This step will also require a significant upfront investment.

Step 5:
It is anticipated that each regional association will have a board of directors, which will collectively provide a broad representation of ocean observing activities across Canada. As such, there should be a regional and national review process to ensure the web portals are sufficient to meet both the specific needs of each region and overarching CIOOS goals. Additionally, complete a security audit to ensure recommendations from Section 6.2 are met.

Step 6:
Upon completion of Step 5, go live with the web portals and associated services at both the regional and national level.

8.3 Phase 2

Phase two would build on the frameworks for phase one. Additional core variables would be incorporated and additional data sources, with increasing levels of curation, would be brought on line. Analysis of the hardware selected in the first phase would be evaluated based on performance and usage to determine if a change or upgrade is required. User flows from site analytics will be reviewed to further critique the design. More complex recommendations from the higher service models can be added to CIOOS' infrastructure on a project-by-project basis.

9.0 Conclusion

Within this report, recommendations for the cyberinfrastructure of a Canadian Integrated Ocean Observing System are detailed. Such a system will provide numerous benefits to Canada, such as: national coordination in data collection efforts; the ability to minimize unnecessary duplication and lost opportunities; and the provision of discoverable and usable data to the world. The latter is particularly important, as it will assist Canada in addressing national priorities and meeting international commitments to the ocean science community as well as improving our understanding of the oceans.

Given its integrated nature, CIOOS may allow Canada to better adapt in the face of changing needs and a changing environment. Provision of open oceanographic data will also benefit the country by assisting in the development of ground-breaking research – thus providing an opportunity for Canada to become a global leader in multidisciplinary ocean science.

CIOOS may be considered as an overview of Canadian ocean science resources, providing industry and researchers with a holistic picture of the interconnected Canadian oceans, and assisting in answering one of the 40 priority ocean sciences questions identified by the Canadian Council of Academies (CCA, 2013):

> *What indicators are available to assess the state of the ocean, what is the significance of changes observed in those indicators, and what additional indicators need to be deployed?*

Implementation of a Canadian Integrated Ocean Observing System will: elevate Canada in the international ocean science community; provide new and exciting avenues for both the nation as a whole and for the individuals within, including research communities, private industry, and public citizens; and present an answer to another of the priority ocean sciences questions identified by the Canadian Council of Academies (CCA, 2013):

> *How can a network of Canadian ocean observations be established, operated and maintained to identify environmental change, and its impacts?*

References

CCA (Canadian Council of Academies) 2012. 40 priority questions for ocean science in Canada; Ottawa, ON; CCA. The Core Group on Ocean Science in Canada.

DFO 2009. Our oceans, our future: Federal programs and activities. Ottawa, Department of Fisheries and Oceans.

DFO 2010. Canadian Survey of Atlantic, Pacific, Arctic and Great Lakes Observing Systems Fisheries and Oceans Canada Ottawa, ON in association with the Ocean Science and Technology Partnership (OSTP).

Expert Panel on Canadian Ocean Science 2013. Ocean Sciences in Canada: Meeting the challenge; seizing the opportunity. Canadian Council of Academies, Ottawa, Canada.

Manson, G. K. 2005. On the Coastal Populations of Canada and the World. *Proceedings of the Canadian Coastal Conference 2005.* Canadian Coastal Science and Engineering Association. Ottawa, ON.

McCauley, D. J., M. L. Pinsky, S. R. Palumbi, J. A. Estes, F. H. Joyce, and R. R. Warner 2015. Marine defaunation: Animal loss in the global ocean. Science 347(2015): 1255641-1 – 7. doi: 10.1126/science.1255641

OSTP 2011. Lessons Learned from OOS in Canada: Preliminary Assessment of OOS Value. Ocean Science Technology Partnership Report prepared for Fisheries and Oceans Canada, and the Canadian Space Agency.

O'Dor, R., K. Fennel, and E. V. Berghe. 2009 A one ocean model of biodiversity. Deep Sea Research Part II: Topical studies in Oceanography 56(19-20): 1816-1823. Doi: 10.1016/j.dsr2.2009.05.023

Wilson, L., Smit, M., and Wallace, D. W. (2016). Towards a unified vision for ocean data management in Canada: Results of an expert forum. MEOPAR. https://dalspace.library.dal.ca/handle/10222/72192

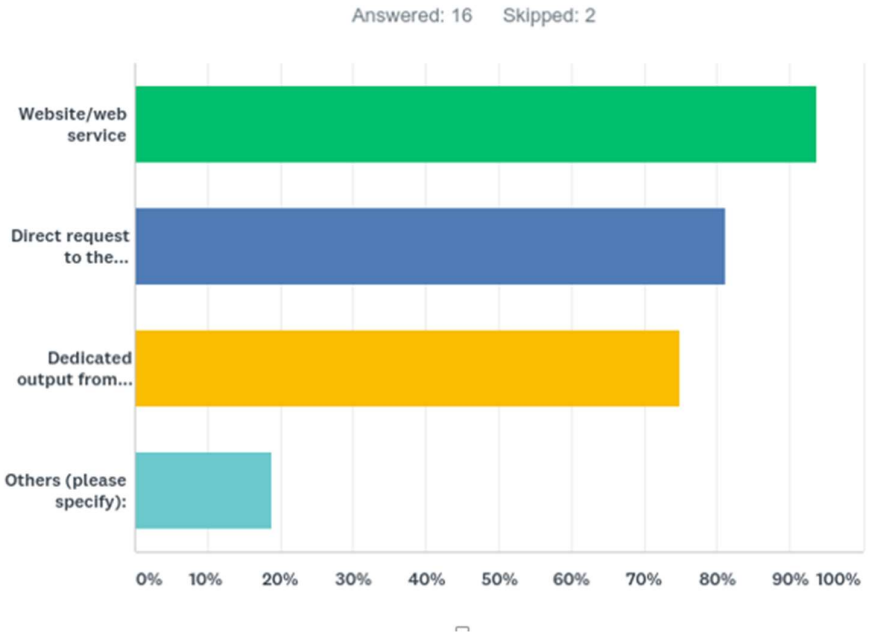
Appendix A: Results of CIOOS Planning Survey

Completed mid-September 2017, the CIOOS planning survey garnered 18 responses from both Canadian and international organizations. The information gathered was considered when forming recommendations for CIOOS infrastructure. Each question was optional, and as such some questions were not answered by all respondents.

Access Methods

What are the methods used to access data/information? (Choose all that apply.)

Respondents: 16 of 18



The methods employed by the surveyed organizations are primarily website/web service, direct request to the organization's operator, and dedicated output from systems and/or sensors. Given the high response levels for each option, it is clear that most organizations employ multiple access methods.

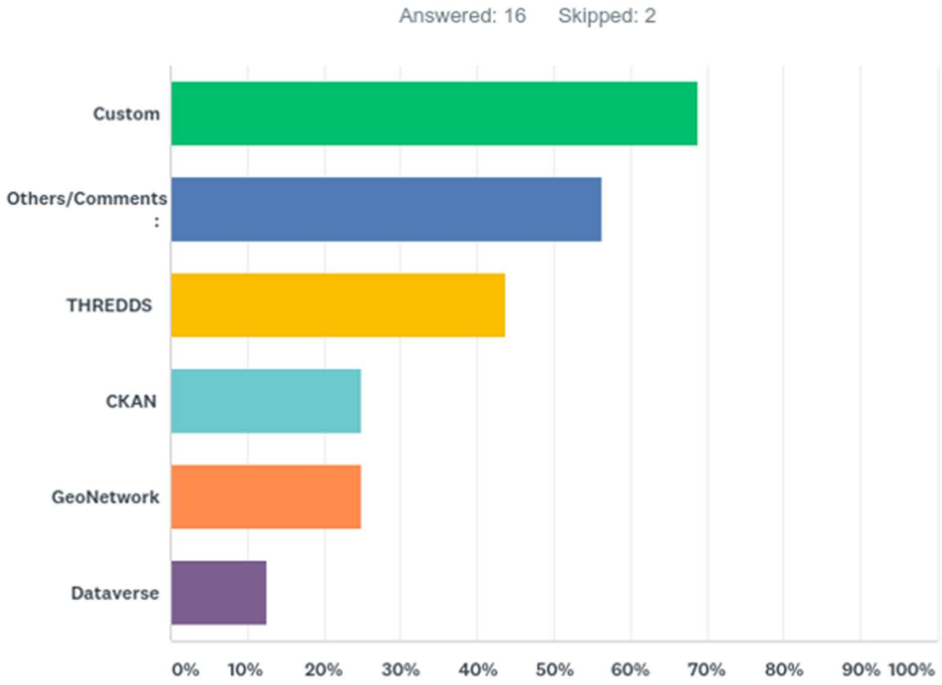
Other access methods used include an online data store and database, historic data which is manually accessed, and a phone digital speech device for water levels.

Tools for Search, Catalogue, Preview

For variable search, cataloguing, and preview, what are your current/planned tools? (Choose all that apply.)

Respondents: 16 of 18

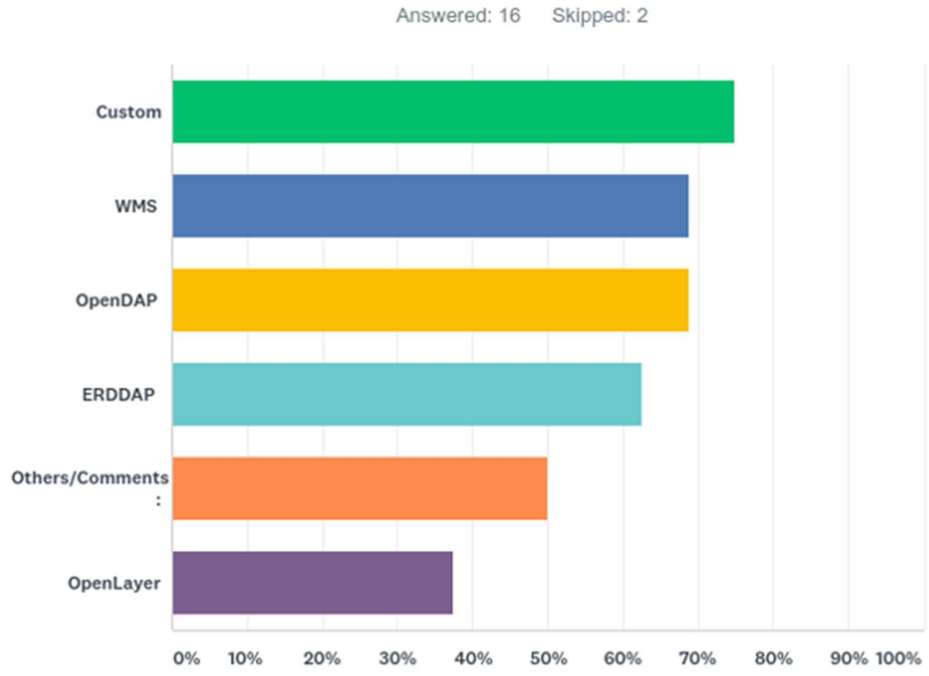
The surveyed organizations primarily employ custom tools or 'other' for data discovery, although THREDDS is also relatively common. Of the nine respondents who selected 'Others/Comments', five discussed custom software while four described public tools. These included ERDDAP, GeoPortal, CSW-enabled GeoServer, CKAN, and GeoNetwork.



Variable Access and Delivery Tools

For variable access and delivery, what are your current/planned tools? (Choose all that apply.)

Respondents: 16 of 18



The most common access and delivery tools utilized by the survey respondents are Custom, WMS, OPeNDAP, and ERDDAP. Minimal discussion of custom tools was provided in the

comments, though respondents also described the use of tools such as OGC SOS, JSON, ESRI, REST, OAI-PMH, ArcGIS REST, GeoServer, and CSW-enabled GeoNetwork. Given the high response levels for each option, it is clear that most organizations employ multiple access methods.

Current and Planned Product Types

If your organization has any datasets that are not currently available to the public, would it be willing to share them if proper infrastructure and associated technical support were available?

Respondents: 13 of 18

This question was not applicable for one of the organizations which responded. Of the twelve remaining respondents, one was willing but unable due to constraints from data contributors and partner programs. Three were uncertain, and commented that agreement from data providers was needed before the data may be made publically available.

Eight responses were positive, indicating that the organizations would be willing to publically share data if the necessary infrastructure and technical support were made available.

Metadata Software Systems

Does your organization make use of any software systems for tracking metadata?

Respondents: 11 of 18

The software systems utilized by survey respondents for tracking metadata are varied; there is little overlap between organizations. The table below summarizes the tools listed in each response.

CKAN	52North SOS	netCDF	THREDDS
GeoNetwork	Sharepoint	Postgres	Jira
Confluence	Research Workspace	Inventory System	Custom Software
ncISO CF Compliance Checker	Database-Driven Web Interface		

Interoperability

What interoperability challenges does your organization face?

Respondents: 14 of 18

Many of the survey respondents face the same interoperability challenges. Those relevant when considering CIOOS cyberinfrastructure are: non-compliant metadata, lack of standards, existing systems which are not readily interoperable, and lack of resources such as staff and funds.

Centralization

Is data storage centralized in one location or distributed across regions?

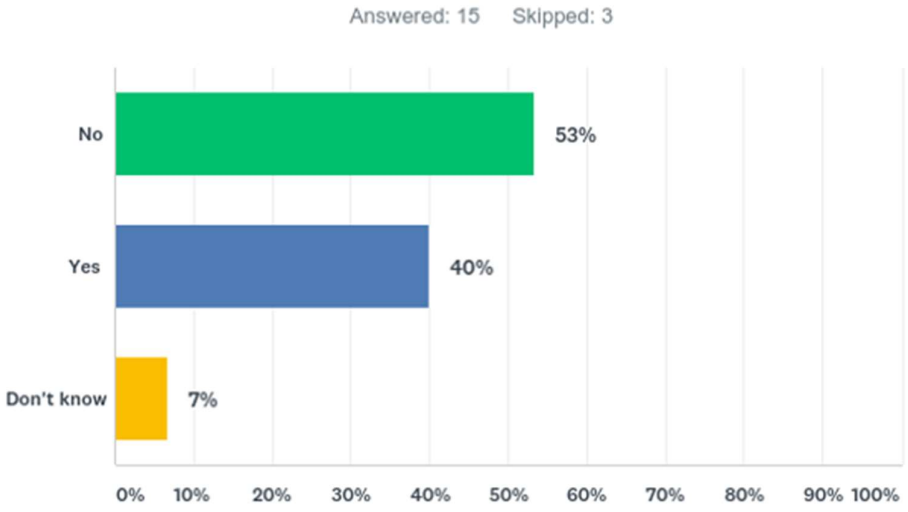
Respondents: 15 of 18

Survey respondents were almost evenly split regarding centralization of data storage, with eight organizations employing distributed storage and seven employing centralized. Several organizations with centralized data utilize off-site backups for redundancy.

3rd Party Hosting

Does your organization work with a paid or unpaid third party (e.g., Compute Canada) for the hosting of data and information?

Respondents: 15 of 18



Although surveyed organizations have a slight tendency to host their own data, a significant portion of respondents use third party providers such as Amazon Web Services, Compute Canada, Axiom Data Science, and ISMER. These cover the categories of commercial vendors, not-for-profit organizations, data management consultants, and research institutions.

If you answered 'yes' to the previous question, what services does the 3rd party provide? (Choose all that apply.)

Respondents: 6 of 12

All respondents who answered 'yes' to the previous question also replied to this. All six respondents use third party providers for data storage and archiving, while three also use said providers for high-performance computing, and one for analytics and visualization. The surveyed organizations also engage third party providers for web hosting, S3, EC2, and limited system administrator support.

Appendix B: List of Abbreviations

ACI	4th Dimension Software
AGPL	Affero General Public License
AODN	Australian Ocean Data Network
AOOS	Alaska Ocean Observing System
API	Application Programming Interface
ART	Advanced Retrieval Tool
ASE	Sybase's Adaptive Server Enterprise
BODC	British Oceanographic Data Centre
CDM	Common Data Model
CF	Climate and Forecast
CFI	Canada Foundation for Innovation
CF-netCDF	Climate and Forecast Network Common Data Form
CI	Cyberinfrastructure
CIOOS	Canadian Integrated Ocean Observing System
CKAN	Comprehensive Knowledge Archive Network
CMEMS	Copernicus Marine Environment Monitoring Service
CODASYL	Conference on Data System Languages
CSV	Comma Separated Variables
CSW	Catalogue Service for the Web

DAC	Data Assembly Centre
DDI	Direct Download Interface
DFO	Fisheries and Oceans Canada
DOI	Digital Object Identifier
DSG	Discrete Sampling Geometries
EMODnet	European Marine Observation and Data Network
ESRI	Environmental Systems Research Institute
ePIC	Electronic Publication Information Center
ERDDAP	Environmental Research Division Data Access Program
EUMIS	European Marine Information System
EuroGOOS	European Global Ocean Observing System
FGDC	Federal Geographic Data Committee
GeoJSON	Geographic JavaScript Object Notation
GIS	Geographic Information System
GML	Geography Markup Language
GNU	GNU's Not Unix!
GPL	GNU General Public License
GPU	Graphics Processing Unit
GRIB	GRIdded Binary
IE	Investigative Evaluation

IETF	Internet Engineering Task Force
IMDIS	International Marine Data and Information System
IMOS	Australian Integrated Marine Observing System
IoT	Internet of Things
ISO	International Standards Organization
JSON	JavaScript Object Notation
JSP	Java and Java Server Pages
MEOPAR	Marine Environmental Observation Prediction and Response Network
MI	Marine Institute
netCDF	Network Common Data Format
NETMAR	Open Service Network for Marine Environmental Data
NGO	Non-Governmental Organization
NODB	National Oceanographic Database
O&M	Observations & Measurements
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OGC	Open Geospatial Consortium
OGSL	Observatoire Global du Saint-Laurent
ONC	Ocean Networks Canada
OOS	Ocean Observing System
OPeNDAP	Open-Source Project for a Network Data Access Protocol

OSGeo	Open Source Geospatial Foundation
PyCSW	Python Catalogue Service for the Web
QARTOD	Quality Assurance of Real-Time Oceanographic Data
RA	Regional Association
RAID	Redundant Array of Independent Disks
RAM	Random Access Memory
RDBMS	Relational Database Management System
REST	Representational State Transfer
RN	Regional Node
SensorML	Sensor Markup Language
SLGO	St. Lawrence Global Observatory
SOAP	Simple Object Access Protocol
SOS	Sensor Observation Service
SQL	Structured Query Language
SRU	Search/Retrieval via URL
SSD	Solid State Drive
SSL	Secure Sockets Layer
STA	SensorThings Application Programming Interface
TLS	Transport Layer Security
THREDDS	Thematic Real-Time Environmental Distributed Data Services

US IOOS	United States Integrated Ocean Observing System
WAF	Web Accessible Folder
VM	Virtual Machine
W3C	World Wide Web Consortium
WCS	Web Coverage Service
WFS	Web Feature Service
WFS-T	Transactional Web Feature Service
WMS	Web Mapping Service
XML	eXtensible Markup Language

Appendix C: Software and Standards of Existing OOSes

The software and standards utilized by the ocean observing systems examined in Section 2.4 are summarized within the following table. A checkmark indicates that the tool is utilized. Lack of a checkmark indicates that no evidence could be found to suggest it is a primary standard or software employed by the organization.

	US IOOS	**EuroGOOS***	**IMOS**	**EMODnet**	**SeaDataNet**
CSW	✓		✓	✓	✓
WAF	✓				
THREDDS	✓		✓	✓	✓
ERDDAP	✓				✓
OPeNDAP	✓		✓	✓	✓
KML Feeds	✓				
SOS	✓		✓		
LAS	✓				
GeoServer	✓		✓	✓	✓
WCS	✓			✓	
WMS	✓		✓	✓	✓
WFS	✓		✓	✓	
In-House	✓		✓		✓
WMTS					
REST				✓	
Relational DB					
WPS					

	CMEMS	**PANGAEA**	**BODC**	**EUMIS**	
CSW	✓	✓		✓	
WAF					
THREDDS	✓				
ERDDAP					
OPeNDAP			✓	✓	
KML Feeds		✓			
SOS					
LAS					
GeoServer	✓	✓			
WCS				✓	
WMS	✓	✓		✓	
WFS				✓	
In-House	✓	✓	✓		
WMTS					
REST	✓				
Relational DB		✓	✓		
WPS				✓	

*EuroGOOS functions only as an international coordinating body; it does not employ standards or software.

Appendix D: Resources for Implementation

Although this report details the recommended cyberinfrastructure for a Canadian Integrated Ocean Observing System, the resources required for implementation are not solely technical; staff and support personnel are necessary to ensure successful setup and continued operation of the system.

The costs associated with cyberinfrastructure are outlined below, including a cost breakdown for the pilot phase and each service model. Because each service model builds off the previous, hardware and software costs are cumulative.

Capital Investments: During the pilot phase it is recommended that Compute Canada's existing stack be utilized, and no capital investment is required. Subsequent phases may require purchase of hardware through Compute Canada, with cost distributed over a five-year amortization period. Prices quoted for Compute Canada include cost of hardware and ongoing management services.

Software: All recommended software is open source, and as such has no associated cost. It will be necessary to employ individuals to both setup and maintain the software systems.

Software Development: Because software packages are easily shared, major development need only occur once. Tweaks necessary to accommodate regional needs will be performed in the relevant region on an as-needed basis. The overall effort required for development is dependent on whether a tool must be developed from scratch, or if an existing tool may be modified. This uncertainty makes software development costs difficult to ascertain at this stage.

In the low service model, a compliance checker will be supplied to assist data providers in ensuring their data is CIOOS-compliant. US IOOS provides an open source compliance checker which may be modified for use in CIOOS. Provision of additional tools in higher service models will necessitate further software development.

Step	Service Model	Item	Cost (CAD) per Unit	# per Region	# at National
Hardware	Pilot	HDD / SSD Cores (4 GB RAM)	55 / TB / Y 150 / Core	TBD 16+	TBD TBD
	Low	As Needed	–	TBD	TBD
	Moderate	VMs (8 Cores / 64 GB RAM) 256 to 512 GB RAM	4,000 / VM 4,000 / 128 GB	6 to 8 2 to 4	– 2 to 4
	High	Hardware costing in this service model to be based on augmentation of the system on a project-by-project basis.			

Personnel: Regardless of the service model, each regional association and the national portal require a technical director and a developer. This 'skeleton crew' would be sufficient for CIOOS to function, but provides little redundancy and causes difficulty in succession planning. Use of a skeleton crew may also negatively impact system uptime and reliability.

To provide additional services, such as those recommended in the moderate and high service models, additional personnel are required. Costs associated with providing resources to data providers is dependent on the service model adopted, and will be difficult to ascertain until more is known about the state of data held by Canadian researchers and technical expertise that resides within existing organizations.

Personnel costs were estimated based on project pricing practices at the Marine Institute, and include both salaries and associated overhead expenses. Costs also represent the 'effort' involved, and may be distributed over more or less personnel; a specific salary may cover one employee at full-time or two at half-time.

Personnel requirements are not cumulative. The requirement for non-technical staff is discussed in the Data and Observations report.

Step	Position	Cost (CAD) per Person	# at National	# per Region			
				Pilot	*Low*	*Mod.*	*High*
Personnel	Technical Director	90,000	1	1	1	1	1
	Sandbox Coordinator	90,000	–	1	–	–	–
	Software Developer	90,000	1+	1	1	1+	2+
	Web Developer	90,000	1	1	1	1	1

Due to uncertainty regarding overall development effort, the estimates provided for hardware, software, and personnel should be considered a best estimate based on current pricing and knowledge of the state of ocean observing; a more precise estimate should accompany any proposals submitted for the formation of the regional associations. Costing provided below assumes that a prototype node is constructed in the pilot phase, and that results from that effort carry forward to inform the nascent RA in whichever service model is adopted.

Appendix E: Consultation with Standards Providers

CIOOS will work with existing standards providers to establish new standards for data types which are not currently covered, but are important for Canada. As an example, ambient sound is identified as a core variable from within the Data and Observations report. While these data can be served using WAF, there are certain limitations with the standard. WAFs do not support streaming of content nor does it support real-time updates. Because of this WAF is not suitable for serving very large, uncompressed hydrophone data files. To the best knowledge of the Cyberinfrastructure IE, there currently does not exist a standard to analyze hydrophone data in a standardized way.

The CI IE did reach out to ISO Technical Committee 43, Sub Committee 3 which focuses on underwater acoustics. They were also unaware of any standard for analyzing ambient sound, but suggested such a standard would be useful. This sub-committee is currently active, and already has a number of Canadian experts. For these reasons, ISO TC 43/SC3 may be a good place to develop such a streaming standard for ambient sound. If CIOOS were to supply live ambient sound recordings as a service, consultation with the Department of National Defence would be needed to ensure such records are safe for public consumption.

Acoustics is one example of where CIOOS could play an important role, particularly as the CFI-funded MERIDIAN project spins up. MERIDIAN has a focus on ocean acoustics and will be housed at Dalhousie University. It is anticipated that other standards will need to be enhanced or developed and CIOOS should be prepared to tackle this.