Polar Science xxx (xxxx) xxx

Contents lists available at ScienceDirect

Polar Science

journal homepage: www.elsevier.com/locate/polar

Polar federated search: New infrastructure to support the polar community

Chantelle Verhey *, Melinda Minch, Karen Payne

World Data System - International Technology Office, Ocean Networks Canada, University of Victoria, #100, 2474 Arbutus Road, Victoria, BC, V8N 1V8, Canada

ARTICLE INFO

Keywords:
Polar
Data
Federation
Semantics
Research infrastructure

ABSTRACT

This paper reports how the World Data System International Technology Office (WDS-ITO) has contributed resources in the Polar Research Data Management community. The office has focused on federated search and data enhancement practices that bridge the gap between data holdings in the Arctic and Antarctic polar research communities and has built a shared catalog characterized by harmonized "dialects'' of metadata standards. The portal, referred to as the Polar Federated Search (PFS) site, is part of a larger set of WDS-ITO polar support activities and allows researchers to find data that has been published in multiple repositories serving both Arctic and Antarctic data (World Data System – International Technology Office, 2022). The portal provides federated search capability through a single interface. The PFS indexes repository landing pages that have been enriched with the addition of semantic markup marrying the best practices from the web publishing world with domain specific metadata. This technique leads to broader discovery, notably by Google Dataset Search, and better reuse of data repository holdings. Increasingly, the base ontology that has been adopted for use in these markup activities is Schema.org (SDO) a simple ontology developed in the commercial sector. This article outlines the steps and progressions made throughout the first year of PFS development, its highlights, and describes how these infrastructures move far beyond metadata discovery to include data integration, analysis, visualization and advanced, reproducible workflows.

1. Introduction

This paper announces the launch of a new data portal (https://search.polder.info/), documentation (Minch and Working Group, 2022), and associated research data management activities and resources that serve the polar research community. The portal will be of interest to researchers who are looking for data as part of their inquiries, as well as software and infrastructure developers who support open source technologies. The portal, referred to as the Polar Federated Search (PFS) site, allows researchers to find data that has been published in multiple repositories publishing both Arctic and Antarctic data. The portal provides federated search capability through a single interface. Users can submit queries that search multiple databases and receive a single compiled response (Solomons and Hinton, 2022). Research datasets that are available through the PFS encompass many different types of observations include but are not limited to weather station records, human written stories, organism counts, and instrument cruises. Observations include measures of almost any variable that exists in nature. For this reason, there is no realistic prospect of standardizing and aggregating the data itself; therefore, federated metadata search is the only viable way to make these datasets easily discoverable, and maximize their value. In the case reported here, the federated search is built on top of semantically enriched metadata that utilizes extensions of Schema.org (Introducing schema.org: Search engines come together for a richer web, Introducing schema.Org: Search Engines Come Together for a Richer Web, 2011)).

This system was created in response to a need identified by the polar research community, in particular Polar Data Discovery Enhancement Research (POLDER - https://polder.info/). POLDER is a collaboration between the Southern Ocean Observing System, (SOOS) the Arctic Data Committee (ADC), and SCAR - Standing Committee on Antarctic Data Management (SCADM). The POLDER Data Management community has been working towards a federated search system since 2016. This work was conducted in two stages. First, the Polar Federated Search was launched in the form of a 'Pilot' to gauge its progress and receive community feedback about its functionality. The portal is now in its second stage, where we continue to solicit feedback and deploy desirable features identified by the community.

The World Data System (WDS) is a component of the International Science Council (ISC), which has a mandate to promote science for the

* Corresponding author.
E-mail addresses: cverhey@oceannetworks.ca (C. Verhey), ito-webdev1@oceannetworks.ca (M. Minch), ito-director@oceannetworks.ca (K. Payne).

https://doi.org/10.1016/j.polar.2023.100947
Received 19 July 2022; Received in revised form 6 April 2023; Accepted 21 April 2023
Available online 23 April 2023

public good. The WDS offices, the International Programme Office (WDS-IPO) in the US and the International Technology Office (WDS-ITO) in Canada (Gerlitz and Pirenne, 2019; Wilkinson et al., 2016), are committed to providing services and support to enhance access, quality and accessibility of data worldwide. The work reported here embraces the ISC objective to create an "ecosystem of resources that enable data to be Findable, Accessible, Interoperable and Re-useable (FAIR) for humans and machines, with effective terminologies and metadata specifications" (International Science Council ISC, 2021). In addition, the WDS supports the UNESCO Recommendation on Open Science (2021) that highlights the importance of "investing in open science infrastructures and services" and "promoting innovative approaches for open science at different stages of the scientific process; " In particular, the UNESCO recommendation recognizes the value of open source software and source code, and recommends that "Digital infrastructures for open science should be based, as far as possible, on open source software stacks." Moreover, the UNESCO recommendation recognizes that the "convergence between the various semantic artefacts (particularly vocabularies, taxonomies, ontologies and metadata schema) is essential for the interoperability and reuse of data for interdisciplinary research."

Toward that end, we have advanced the ability of researchers to find polar data, using open source software components and a well described ontology that is built on top of commonly used metadata terms found across the web, described further below. Creating the PFS using open source software has the added advantage of making it freely available to other communities to use in their own domain-specific search portals. The problem at hand is that the state of the data discovery is a complex landscape. As seen in Fig. 1, the POLDER group alone consists of 103 different data repositories. Globally, not all polar focused data repositories participate in the POLDER space. Moreover, some repositories are more generalist which further exacerbates the data discoverability problem. Some aggregate other repositories such as the Polar Data Catalogue, but there is no set of standards or recommendations to organize these harvesting structures. These relationships can become outdated and change without notice, and are generally not publicly known. For researchers who may be searching for data, this may mean that they are individually searching each repository and can result in a time-consuming hunt. A federated search helps reduce this dilemma by having repositories voluntarily complete the semantic mark-up in order to be included, have their participation publicly known for users to see, and it does not change their current harvesting structures that they may already have in place.

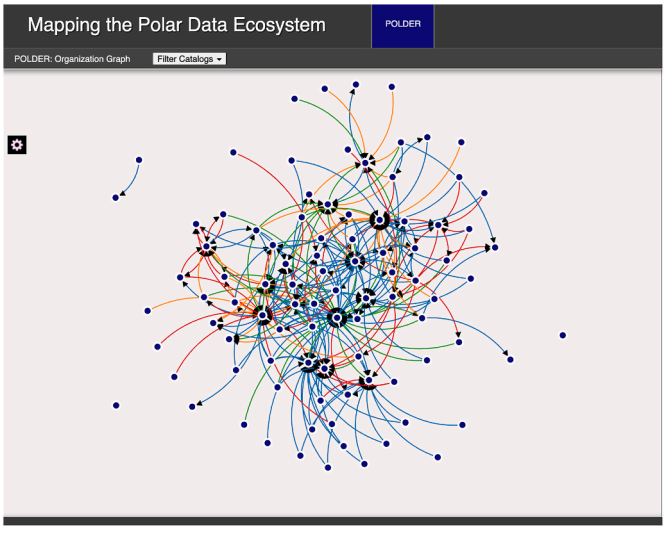

Fig. 1. This figure is a screenshot of the "Mapping the Polar Data Ecosystem Project", this graph shows the harvesting relationships among polar data catalogues. It shows the network of systems and how data is harvested from one system to another. Edges in the graph denote harvesting with arrows pointing toward the organization that consumes the harvest. (Pulsifer et al., 2020).

The current strategy for WDS-ITO's work with semantic alignment is to build on Schema.org (SDO, http://www.schema.org) successes, which are being replicated in other communities that are using SDO as a base ontology, and creating extensions to it that are relevant to their own communities (World Data System – International technology office, 2022). In particular, this strategy is being utilized by the Earth Science Information Partner's Science-on-Schema.org (SO–SO), the Ocean Data and Information System (ODIS) the ESIP soil cluster, and EarthCube via their P418 project, Magnetics Information Consortium (MagIC), and GeoCODES search engine (https://geocodes.earthcube.org). Far more impactful than data publishing and exposure, the long-term vision of this work is to support pathways for cross domain research. The initiative described in this paper is aided by the fact that partners in this work include repositories that are not solely polar; they have polar data but also data from other domains, including ocean, climate and biological data. Table 1 outlines the differences between federated search initiatives in the scientific research community compared to the Polar Federated Search project.

This paper reports how the WDS-ITO has contributed to resources in the Polar Research Data Management community by focusing on federated search and data enhancement practices that bridge the gap between data holdings in the Arctic and Antarctic polar research communities that ultimately allowed us to build a shared data catalog characterized by harmonized "dialects'' of metadata standards. We took a multi-pronged approach to this goal with activities in the following areas: Semantic alignment, Software development, and Infrastructure.

The current state of the activities in this article allowed for the development and publication of a single interface for discovering both Arctic and Antarctic datasets, and a description of the resulting infrastructure are described below.

2. Methods

2.1. Semantic alignment

As a preliminary step towards making repository holdings machine actionable, data managers have been exploring the utility of including semantic markup in their research metadata. These days, it is common practice at data repositories to build websites that display metadata about data holdings in metadata landing pages (Fenner et al., 2018). By adding semantic JSON-LD markup (https://json-ld.org/) to these webpages and making them available for indexing, managers marry the best practices from the web publishing world with domain specific metadata. This technique leads to broader discovery, notably by Google Dataset Search, and better reuse of their holdings. Increasingly, the base ontology that has been adopted for use in these markup activities is SDO (Benjelloun et al., 2020), a simple ontology developed in the commercial sector.

In this context, one of the roles of the WDS-ITO is to support the repositories that are interested in semantic markup, including helping to develop tools used in both implementations, as a steppingstone to bring assets from both communities under a single searchable ecosystem. To date the WDS-ITO contributions to tools and resources include (but are not limited to):

1. POLDER Schema.org implementation guidance (Bricher et al., 2023)
2. Guidelines for publishing structured metadata on the Web (in partnership with the Research Data Alliance Research Metadata Schemas Working Group) (Wu et al., 2021)
3. Schema.org for Research Data Managers: A Primer (Payne and Verhey, 2022)
4. A Collection of Crosswalks from 15 Research Data Schemas to Schema.org (Wu et al., 2022), and a Visualization of those crosswalks

Table 1
This table outlines the various federated search initiatives that the WDS-ITO is aware of and how it differs from the Polar Federated Search. This table is not all encompassing but rather a sample of projects for comparison purposes. These projects are all considered to be 'Federated Searches' but their scope, domain, and long term maintenance strategies vary, which was the major reason for the Polar community's interest in creating their own platform.

Project Name	Scope (Regional/ Global)	Domain	Polar Region	Publication/Meta (data)/Both/ Observation Assets	Shows results on a map	Links to dataset landing pages	Uses schema.org/ JSON-LD metadata	Currently Maintained	URL
Polar Federated Search	Global	Polar Science	Polar	Meta(data)	Yes	Yes	Yes	Yes	https://search.polder.info/
OIH	Global	Oceanographic data	Broader	Meta(data)	No	Yes	Yes	Yes	https://catalogue.odis.org/search
Earthcube	Global	Environmental	Broader	Meta(data)	No	Yes	Yes	No	https://geocodes.earthcube.org/#/landing
Polardex	Global	Polar Observation Assets	Polar	Observation Assets	Yes	No	No	Yes	https://polardex.org/
Arctic Observing Viewer	Regional	Arctic Observation Assets	Arctic	Observation Assets	Yes	No	No	Yes	https://arcticobservingviewer.org/
SOOS	Regional	Oceanographic Observation Assets	Antarctic	Observation Assets	Yes	No	Yes	Yes	http://www.soosmap.aq/
Norwegian Polar Data Centre	Global	Polar Research Science	Polar	Publication	No	Yes	No	Yes	https://data.npolar.no/home/
DataONE	Global	Multidisciplinary Data Centre	Broader	Meta(data)	Yes	Yes	Yes	Yes	http://dataone.org
FR-DR	Regional	Canadian Research Data	Broader	Meta(data)	Yes	Yes	No	Yes	https://www.frdr-dfdr.ca/repo

(Welcome to Schema.Org Crosswalks, 2020, https://rd-alliance.github.io/Research-Metadata-Schemas-WG/)

5. Recommending markup to data managers for their holdings based on community developed crosswalks

The SDO vocabulary has been extended by the Earth Science Information Partners (ESIP, https://www.esipfed.org/) to create science-on-schema.org (SO–SO, https://github.com/ESIPFed/science-on-schema.org). There are currently two implementations of SO–SO found in the Arctic and Antarctic communities: full and lightweight. The full ESIP implementation of SO–SO is necessary to be included in the DataONE (https://www.dataone.org/) community of repositories, while the lightweight implementation is being developed through POLDER (https://polder.info/). The latter has generated more interest than the full implementation amongst polar data managers, since it provides terms that are the most relevant for polar specific metadata. While the full SO–SO guidance recommends all terms be populated (Shepherd et al., 2022), the POLDER guidance selects a subset of mandatory metadata terms. The SO–SO guidance provides examples on how to mark-up all available fields, and with few exceptions the POLDER guidance involves a subset of those fields. The POLDER strategy involves decreasing the number of terms to be annotated and thereby reducing the barrier to implementation so that more repositories will see a path to adding semantic markup to their metadata.

2.1.1. *Community involvement in semantic alignment*

Community consensus and development was achieved through a series of 'Polar to Global hackathons' (https://p2g-data.org/). These hackathons are bi-monthly meetings between 2 and 3 h, where interested members of the community congregate and discussed which terms were to be included in the 'POLDER Schema.org implementation guidance' (Bricher et al., 2023). The hackathons began after they were identified as a need by the community during the 3rd Polar Data Forum (PDF) in Helsinki, Finland in 2019, and were organized by the Forum's leading organizations (SCADM, ADC, & SOOS). This space allowed for community discussion and collaborations, and reserved a space for progress on identified tasks by interested community members. This is the space where the group was able to go through the SO–SO recommendations, provide feedback and discuss which terms were most applicable for polar repository needs and further create the best practices documentation for implementation. After this stage was fleshed out, discussions surrounding infrastructure and the user interface had a starting point and was able to proceed.

The WDS-ITO had two goals in this project: help develop the aggregation and search tools for those repositories who have implemented either the lightweight or full version of SO–SO, and work with

Table 2
POLDER required and recommended metadata fields for polar data. To be indexed and included as part of Google Dataset Search, only the 'Title' and 'Description' are required. More information about the specifics of adding markup to these metadata fields can be found in the POLDER SO–SO implementation guidance (in development).

POLDER Required Metadata Fields	Optional (Recommended) Metadata Fields
Identifier	SameAs
Title (schema:name)	Keywords
Description (can be any length, although Google's dataset search requires it to be between 50 and 5000 characters)	Version
Temporal coverage	Date Published
Spatial coverage	Distribution
Parameters/Variables	
Citation	
Creator/Author	
Publisher	
License	

repositories who have yet to adopt any semantic markup to dip their toes in the water using the POLDER 'Best Practices'. Most of the work in getting a repository ready to be included in the Polar Federated Search is in getting its metadata to a state where the application can consume it and use it in searches. One option available to data managers, particularly those in the earth science community, is the adoption of SO–SO, an extension of SDO that includes twenty recommended terms and rich, detailed recommendations about how to use several more; in contrast, the POLDER implementation of SDO is lightweight, with only twelve mandatory terms. Table 2 below provides the current set of POLDER recommended metadata fields that should have SDO markup and in turn, can be discovered by users seeking polar data through search. polder.info. The PFS solution accepts both of these approaches resulting in a higher total number of datasets included in the federated search database into a single searchable archive over the more advanced markup options. However, even with the lightweight option, we recognize that many repositories that have polar data have limited experience with, and limited resources to develop their metadata and serve it so that it can be part of this project. In light of this, the WDS-ITO is running webinars and provides free consultations to repositories who are interested in adding mark up to their metadata, but need more support, described further below. The POLDER community has done a tremendous amount of difficult and time-consuming community building consensus work to define the fields that should be included in polar metadata. This bottom-up work has a direct impact on the resulting utility of the search portal for obvious reasons. To search on a field, that field must exist. If stakeholders want people to be able to find a data set using, say, a date search, they must attach temporal coverage information to their metadata in a consistent way to allow that search to be executed.

2.2. Infrastructure

The site https://search.polder.info/that harvests polar data is now operational. The site, as well as an additional deployment used for testing new features, is currently hosted by our partner DataONE, while the WDS-ITO is actively developing the application. Fig. 2 shows the components of the site, with the red circle representing the boundary of the site deployment itself. The PFS web app can be deployed with a single command using either Kubernetes (https://kubernetes.io/), via a set of Helm (https://helm.sh/) charts, or Docker Compose (https://docs.docker.com/compose/), although the current deployments are done with Kubernetes. In other words, everything inside the red circle in the diagram is inside a network that is private to the PFS deployment, with ingresses configured to allow the user interface to be shown on the internet, but to disallow internet-wide access to the individual components. This is a common practice for the deployment of web applications, and it is done to decrease the potential attack surface for hackers.

Pictured in orange in Fig. 2 diagram, the web app itself is the Flask (https://flask.palletsprojects.com) application which hosts the user interface and performs the searches. User submitted search terms or facets are transformed into Solr (https://solr.apache.org/)/DataONE and SPARQL (https://www.w3.org/TR/sparql11-query/) queries that are then sent to the search endpoints. The endpoints return the results of the queries to the web application, which synthesizes them and displays them as both a ranked list and map extent in the user interface. Pictured as a black Gleaner logo and a blue database icon in Fig. 2, the data ingestion infrastructure consists of an installation of Gleaner (https://gleaner.io), which stores dataset metadata that it collects in an Resource Description Framework (RDF, https://www.w3.org/RDF/) store; in this case, an instance of GraphDB (https://graphdb.ontotext.com/), although it is also possible to deploy this stack with Blazegraph (https://blazegraph.com/). Other components outside the Kubernetes deployment, but in the diagram, include the DataONE index, which the PFS queries as part of getting search results, and data repositories, which are crawled by DataONE and Gleaner in order to retrieve dataset metadata.

2.3. Software development

The PFS is an open-source web application that was custom-built for the purpose of creating a federated search for research datasets with schema.org metadata from a particular community, with the hope that other research communities might adapt it to their own purposes. To that end, we chose tools that the community is already familiar with, and that would fit the needs of the project. Several components comprise the PFS, but the bulk of the software development work was done on what, in this paper, we refer to as the 'web app'. The PFS web app is built using Flask, a well-supported and reliable Python web framework. The WDS-ITO chose the Python programming language because it has support for RDF and SPARQL operations with RDFLib (https://rdflib.dev/), and it is a well-known and popular language in the scientific community. The user interface in the PFS is built in HTML, JavaScript and SCSS, using Parcel (https://parceljs.org/), all of which are languages and projects with broad adoption and good support. It has been tested for compliance with Web Content Accessibility Guidelines standards and usability with screen readers, browsers that do not support JavaScript, and mobile devices. The interface includes online maps, which were created using OpenLayers (https://openlayers.org/). OpenLayers offered the most flexibility when styling the maps, and, when coupled

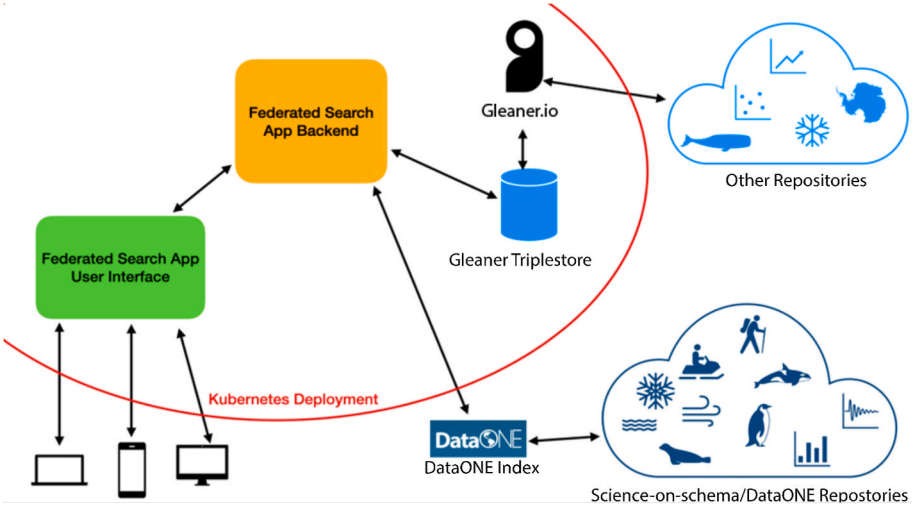

Fig. 2. Schematic of the Pilot Polar Federated Search deployment. The SO–SO/DataONE repositories have implemented the more rigorous version of SO–SO, while the Gleaner tools have been developed to harvest from other repositories who implement the less stringent SO–SO guidance. Pictured in green in Fig. 1 diagram, the user interface is the website that people interact with to make searches and see search results. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

with the JavaScript library proj4, made it relatively simple to create polar-projected views of dataset extents. A map view was identified by the POLDER community as a required feature, and a polar-projected view is a natural and user-friendly fit for our search domain.

In addition to creating the PFS web app and the software that coordinates the components that it relies upon, the WDS-ITO also made significant contributions to the open source metadata ingestion tool Gleaner. Gleaner collects metadata, in particular schema.org markup in JSON-LD format, that is contained in the landing pages of the data distribution websites deployed by data repositories that wish to be included in the PFS search. The WDS-ITO submitted bug fixes and added features that were necessary to support the federated search scenario required by polar data managers. We chose to use, adopt, and support Gleaner because it was developed for and by researchers who are involved with scientific data management, because it was specifically created to ingest schema.org metadata, which is what the PFS is targeting, and because we saw areas where our expertise could make a positive contribution in that open software project. WDS-ITO members are also writing automated unit tests for Gleaner, actively participating in code reviews, and attending meetings held for the community of Gleaner contributors. We continue to consult with the polar research community about what features they would like to see in this, or other, search tools, and conducted a gap analysis to identify the most pressing needs (Verhey, 2021).

Originally designed to work with the more rigorous implementation of SO–SO, we have, with the support of Gleaner's maintainers, chosen to submit patches to Gleaner so that it will optionally accept a looser interpretation of those standards by applying software logic to compensate for irregularities in metadata markup. This approach meets data managers where they are, and recognizes that they are working with limited resources, rather than asking them to create or fix their existing metadata markup. By working with the creators of Gleaner to adapt the tool that was originally optimized for the more stringent version of SO–SO to be able to function more broadly, we have opened the door for more data to be included in the polar federated search ecosystem, and other search portals in the future. Table 3 provides a list of all work done on different components of the Polar Federated search ecosystem by the WDS-ITO in 2022.

In practice, with the software and infrastructure mentioned above, the single user interface can be seen in Fig. 3. The home page displays the up to date search options, an 'about' page, and 'help' page. The user can use any of the available search options, singularly or combined (ie keyword and author, or just keyword – see Fig. 4), and the results from that search will be displayed as seen in Fig. 5. All search results provide a link to the original source of the metadata (hyperlinked in the title) and/or in a DOI. The user can access the data through these links. The 'About' page lists the current repositories in the interface, links to the up-to-date software GitHub repository (https://github.com/WDS-ITO/polder-federated-search/), and list of organizational partners.

3. Results

In addition to working on semantic guidance, software and infrastructure, the WDS-ITO is working directly with data managers to help them meet their metadata development goals and prepare them for inclusion into the Polar Federated Search portal. This includes adding SDO-based markup to their landing pages and ensuring that repositories provide a sitemap and/or robots.txt file so that harvesters like Gleaner can know which pages to crawl and include in the searchable index. Our latest work in this area has focused on support for the British Antarctic Survey. In addition, six other Arctic and Antarctic repositories have committed to implementing the POLDER SO–SO markup and open their metadata for harvesting in the PFS: two are currently included, and five are in development (see Table 4). An additional eight repositories from the Canadian Consortium for Arctic Data Interoperability (CCADI) consortium have agreed to participate in the POLDER federated index.

Table 3
Features developed as part of the Polar Federated Search ecosystem deployment.

Name	Component	Description/Functionality
Search interface and results	UI	A webpage with a search form - for full text and date range searches - that displays federated search results. It surfaces information about each data set, like title, keywords and DOI, and provides links to the data set landing pages.
Search result collation and federation	Python/Flask App, Solr Query Interface, SPARQL Query Interface	A web application, connected to the search interface, that handles user-initiated queries and presents results for those. It translates the user queries into SPARQL and Solr queries and gets results from both of those. It then processes and sorts those results so they can be presented to the user who made the original query. For the SPARQL queries in particular, care must be taken to write them so as to display good search results for the different data repositories, without making them unreasonably slow.
Accessibility	UI	An important part of this project is to make it useable by a broad range of people in a variety of circumstances. It was and continues to be regularly audited for accessibility for low-vision users and people who will be using it with assistive devices. In addition, a JavaScript-free version is available for people on slower or unreliable internet connections.
Search Pagination	UI, Backend Web app	To minimize the impact on DataONE's servers and make the website respond to user-initiated queries in a timely manner, searches are paginated (meaning that only a certain number of results are fetched and presented at a time).
Deployment Container	Kubernetes/Helm/ Docker	One of the goals of this project is for it to be easily portable and hostable almost anywhere. To that end, it can be deployed with a Docker Compose file or a Helm Chart.
Crawling/ Indexing	Gleaner and associated dependencies	The work done here is to configure an instance of Gleaner and its dependencies that the Python web app can talk to. It also involved identifying data repositories to index and working with repository owners to make sure that the metadata would provide good search results.

Fig. 2 reflects a set of use cases that are handled by the Polar Federated Search system faced by data managers listed in Table 4:

1. Harvesting the landing pages of repositories directly as is the case with the Greenland Ecosystem Monitoring portal;
2. Reading an index from an aggregator portal that harvests from other repositories (DataONE), and
3. Indexing data from a repository that is part of the DataCite consortium which provides markup on behalf of participating repositories, as is the case with the British Antarctic Survey.

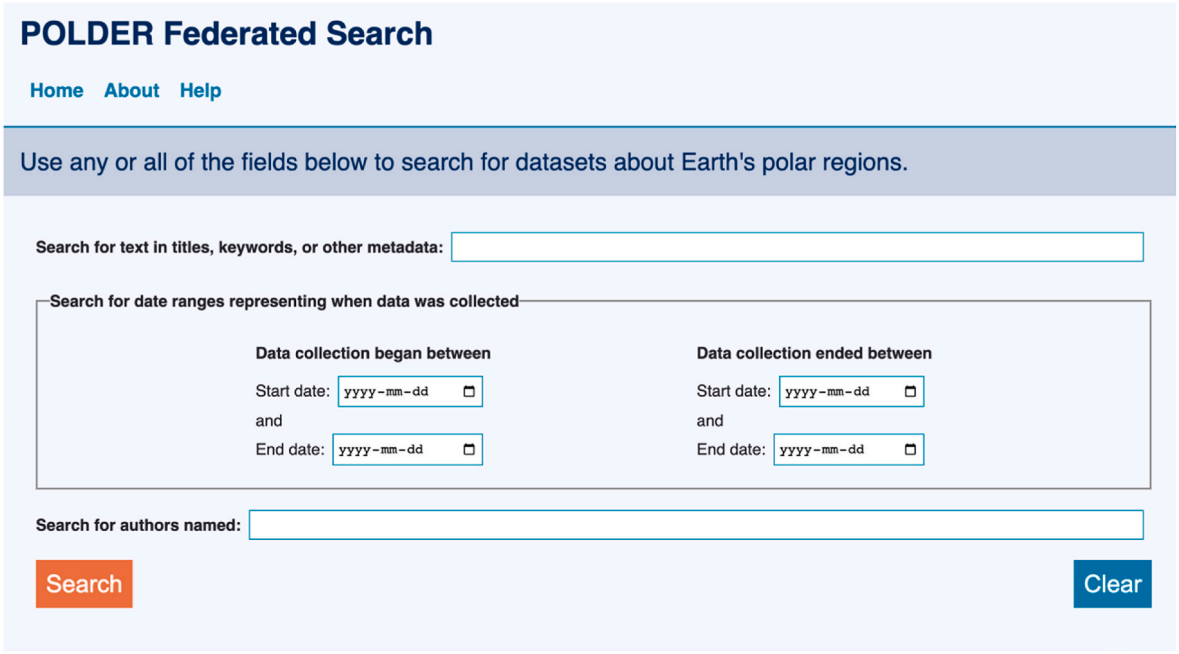

Fig. 3. Current search page for the Polar Federated Search (https://www.search.polder.info). The webpage includes options for a keyword search, spatial search, temporal search, and author search.

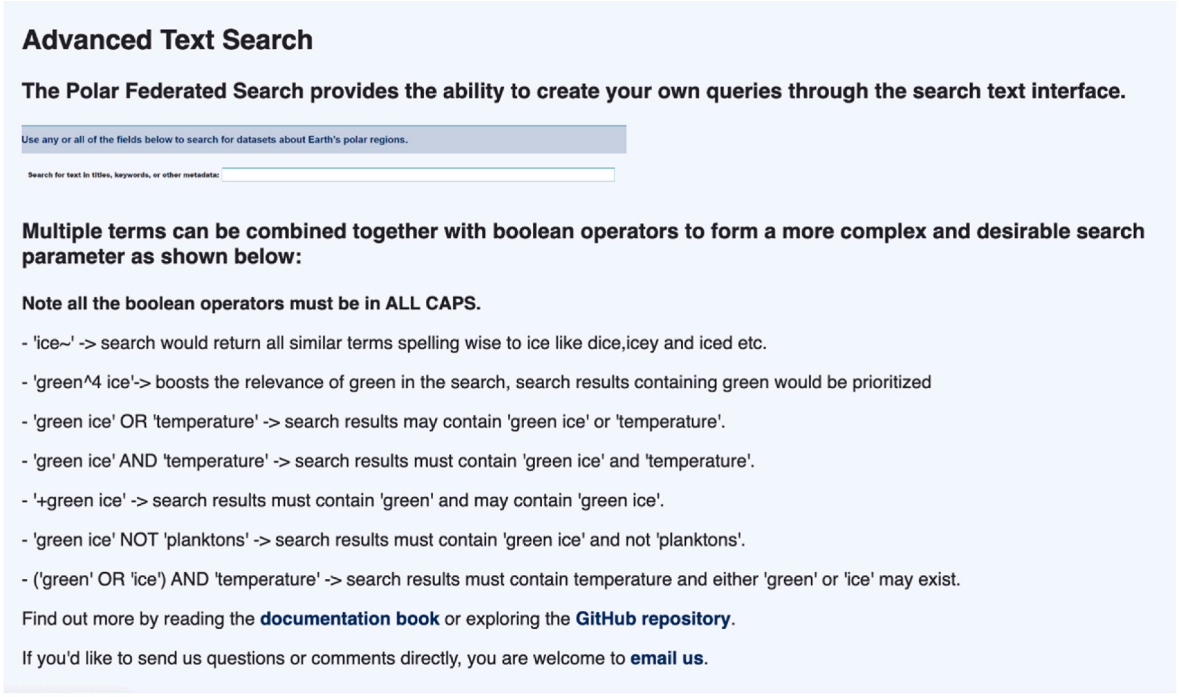

Fig. 4. This figure is a screenshot from the 'Help' page on the Polar Federated Search page. It outlines how to use the advanced search options and gives examples on the various combinations that can be used in the search interface (https://search.polder.info/help/).

The community strategy is that as repositories continue to receive support from the WDS-ITO, and see the progress in the search results that allow researchers to explore newly discovered datasets, in turn, more repositories will wish to be included in the PFS.

3.1. Enhanced data discovery

As outlined throughout this article, polar data discovery can be a lengthy endeavor. As of February 2023, the PFS in full production harvests a total of 19 repositories, and four aggregators. This results in stakeholders being able to search across the participating repositories simultaneously, in the case of PFS - across 19 sources opposed to the full 103, saving time and resources. We are aware of other similar but distinct projects in this space. Namely, repositories that publish their own data holdings that include observations from both poles, such as the Japanese National Institute of Polar Research (NIPR) (Kadokura et al., 2022)., the Polar Data Catalogue (re3data.org, 2022), and other initiatives, such as the Norwegian Global Open Access Portal which harvests

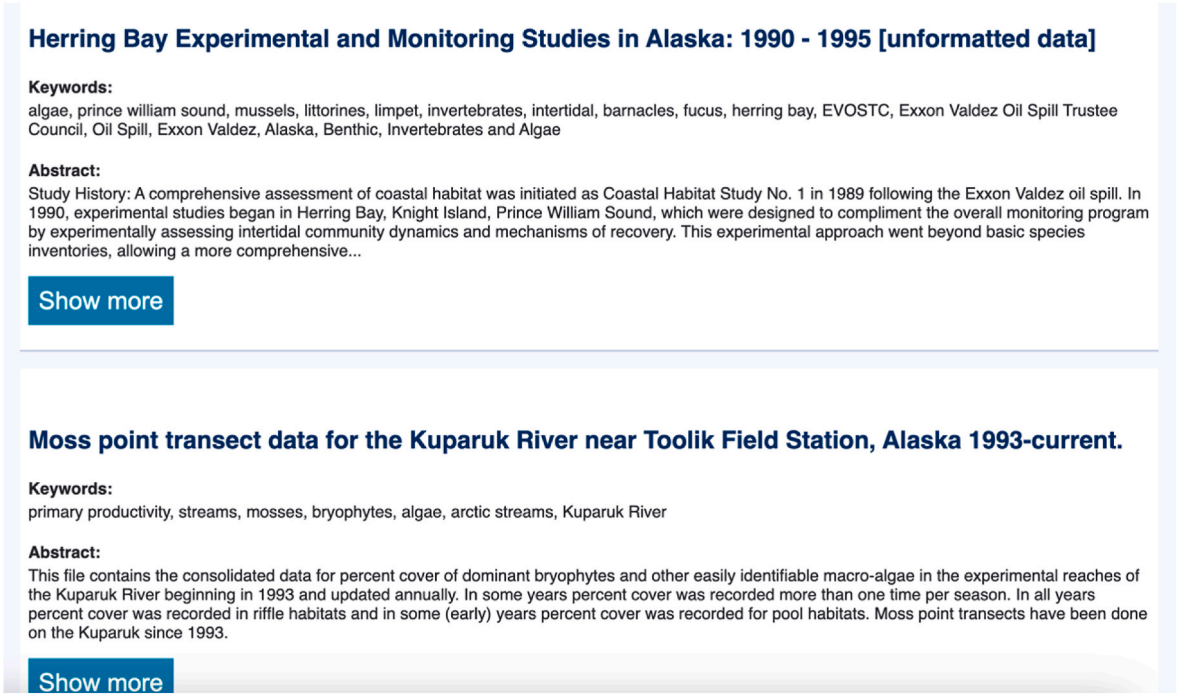

Fig. 5. An example of search results for keyword: 'algae'. The search displays metadata that matches the keyword. Results display the metadata: 'Title', 'Keywords', and 'Abstract', with the option to view more such as 'Timespan' and 'Author(s)'.

Table 4
Participants in the Polar Federated Search. In column one: two Repositories are currently included in the deployed Federated Search Portal. One of the two repositories (DataONE) is an aggregator site. The PFS is indexing DataOne, not using the gleaner to harvest them; the DataONE index includes data from the four other repositories listed. Column two: five Repositories that have agreed to implement the POLDER guidance and become part of the Pilot Federated Search Portal with assistance from the ITO (in progress). Column three: eight CCADI Repositories who have agreed to be part of the pilot project and have already implemented the full SO–SO specification, but have not been harvested yet.

Indexed Repositories (completed implementation)	Repositories in the process of implementation	Repositories working on implementation through an API
1. Greenland Ecosystem Monitoring 2. DataONE a. Netherlands Polar Data Center b. Arctic Data Center c. BCO-DMO d. USAP-DC	1. Australian Antarctic Data Centre 2. NASA/GCMD 3. National Snow and Ice Data Center 4. CLIVAR and Carbon Hydrographic Data Office (CCHDO) 5. British Antarctic Survey (BAS)	1. Polar Data Catalogue 2. Nordicana D 3. Committee on earth observing satellites (CEOS) 4. Arctic Science and Technology Information System (ASTIS) 5. ArcticConnect 6. Arctic Spatial Data Infrastructure (ASDI) 7. INTERACT 8. Canadian Watershed Information Network (CWIN)

metadata from Arctic and Antarctic repositories. The PFS portal announced in this paper is distinct from these projects. In the first case, unlike other bi-polar repositories, we are harvesting metadata from multiple outside providers. Moreover, our site is distinct from the Norwegian Portal because they are designed for traditional metadata harvesting, as compared to our SDO-based indexing. We see the PFS being a feeder for more complex infrastructures in development.

4. Discussion

Similar initiatives for polar federated search have come to fruition but unfortunately fallen by the wayside due to lack of funding and/or lack of capacity. The polar community is 'small but mighty', with many individuals stretched thin while they try to manage their own workload and contribute to the broader research data management community. This is not a problem that is unique to the polar community, and researchers and managers from many other domains can sympathize. Considering this, one of our goals in this work is to develop community endorsed, highly functionally, easily deployable and easily portable tools that could assist with this problem long term.

4.1. Roadmap

The WDS-ITO had created a 'PPFS Roadmap: Year 1 in Review' which recaps activities and decisions conducted during the pilot stage of the portal and outlines the future development of this project to ensure it continues to stay relevant and valuable to the research data management (RDM) community. Specifically, the document begins by briefly describing the pilot federated search, and it lays out a plan for the disposition and development of the project. In particular the steps to success, milestones, hosting documentation, and next steps (Verhey et al., 2022).

The PFS project operates under the guidance of an advisory team that provides recommendations on project development. The advisory team consists of Biological and Chemical Oceanography - Data Management Office, DataONE, Arctic Institute Observing System, British Antarctic Survey, ODIS, and IODE. Moreover, the WDS-ITO actively collects community feedback during POLDER working group meetings, regularly scheduled hackathons and requests submitted to the office via email that are maintained on a Trello board in order to determine the future direction for this work.

As of the time of this writing, the PFS priorities are as follows:

1. Adding more data repositories
2. Adding support for geographic search

3. Displaying more or better information in the search results, such as spatial location, authors, organizational information and affiliations, and the like
4. User experience improvements (things like making the date fields able to handle just a year)
5. Supporting faceted searches, like searches by data file type or research funding organization

Using the community feedback mechanisms listed above, this work will be prioritized as deemed appropriate, according to what the community feels that they need the most. As the web interface is changed and added to, it is periodically tested for compliance with W3C accessibility guidelines, as well as usability on a variety of web browsers and mobile devices (Lawton Henry and Spellman, 2022).

In conjunction with the PFS advisory team, we also intend to create a 'SWAT' team to conduct outreach to individual repositories and assess the steps needed for them to comply with the POLDER Best practices in order for their repositories' meta(data) to be included in the PFS.

5. Conclusion

The pilot PFS project came to an end on 31 March 2022, and the pilot moved to production as of 1 April 2022. To date, the project has successfully used open source software and community defined vocabularies that allows researchers to discover both Arctic and Antarctic data through a single search interface. The project will continue to be developed in the next year through collaborations with the WDS-ITO, POLDER WG, and DataONE developers. Moving forward, we will work with the POLDER community to further determine the next hosting space of the portal and associated infrastructure after the first year of production. In addition to its portability, it is important that this tooling is developed in a way that best serves the Polar community, without re-inventing existing or hindering significant investments in much more complicated infrastructures. Towards this end, we are collaborating with, and closely following projects such as the Canadian Consortium for Artic Data Interoperability and the Virtual Research Environments currently being contemplated by the polar community (Theocharides et al., 2021). These infrastructures move far beyond metadata discovery to include data integration, analysis, visualization and advanced, reproducible workflows. Our work here demonstrates that the WDS-ITO is committed to being a trusted partner in these projects, and to supporting data managers, especially WDS members, so that they are well represented in these integrated ecosystems.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank Devin Prater for conducting accessibility testing and providing valuable feedback on the search webpage's user interface. We would also like to thank DataONE for providing support and server space for the duration of the Pilot Project, and finally to the entirety of the POLDER WG for their continuous support and encouragement.

This work was made possible through funding provided by The Digital Research Alliance of Canada and through the hosting partnership with Ocean Networks Canada.

References

Benjelloun, O., Chen, S., Noy, N., 2020. Google dataset search by the numbers. In: The Semantic Web – ISWC 2020. ISWC 2020. Lecture Notes in Computer Science, 12507. Springer. https://doi.org/10.1007/978-3-030-62466-8_41.

Bricher, P., Verhey, C., Duerr, R., Ingram, R., Pulsifer, P., Collins, J., Jones, M., de Bruin, T., Manley, W., Christoffersen, S., Budden, A., Van de Putte, A., Gaylord, A., Tacoma, M., Fremand, A., Aulicino, G., Gao, A., Minch, M., Tronstad, St, et al., 2023. POLDER Best Practice Guide to Implementing schema.Org for Data Discovery (1.0). https://doi.org/10.5281/zenodo.7787161. Zenodo.

Fenner, M., Crosas, M., Durand, G., Wimalaratne, S., Gräf, F., Hallett, R., Bernal, L.M., Schindler, U., Clark, T., 2018. Listing of Data Repositories that Embed schema.Org Metadata in Dataset Landing Pages (1.1.2). https://doi.org/10.5281/zenodo.1263942 [Data set]. Zenodo.

Gerlitz, L., Pirenne, B., 2019. Facilitating Stewardship of Data on a Global Scale: the World Data System's New International Technology Office. Research Data Canada, Blog 7 Mar 2019 https://www.rdc-drc.ca/facilitating-stewardship-of-data-on-a-global-scale-the-world-data-systems-new-international-technology-office/. (Accessed 23 May 2022).

Kadokura, A., Kanao, M., Yabuki, H., Tanaka, Y., Nishimura, K., 2022. Activities of the polar environment data science center of ROIS-DS, Japan. Data Sci. J. 21 (1), 12. https://doi.org/10.5334/dsj-2022-012.

Lawton Henry, S., Spellman, J., 2022. W3C Accessibility Guidelines (WCAG) 3.0 Working Draft. W3C,5 Apr 2022. (Retrieved from https://www.w3.org/WAI/standards-guidelines/wcag/wcag3-intro/). (Accessed 7 April 2022).

Payne, K., Verhey, C., 2022. Schema.org for research data managers: a primer. Int. J. Big Data Management 2 (2), 95–116.

Welcome to Schema.Org Crosswalks, 2020. https://rd-alliance.github.io/Research-Metadata-Schemas-WG/. (Accessed 13 June 2022).

Minch, M., POLDER Working Group, 2022. POLDER Polar Federated Search Documentation 31 Mar 2022 GitHub Repository (Retrieved from https://polder-crew.github.io/Federated-Search-Documentation/intro.html). (Accessed 7 April 2022).

Pulsifer, P.L., Kontar, Y., Berkman, P.A., Taylor, D.F., 2020. Information ecology to map the arctic information ecosystem. In: Governing Arctic Seas: Regional Lessons from the Bering Strait and Barents Sea. Springer, Cham, pp. 269–291 re3data.org: Polar Data Catalogue; editing status 2022-01-03; re3data.org - Registry of Research Data Repositories doi:10.17616/R3QK86. (Accessed 1 June 2022).

Shepherd, A., Jones, M., Richard, S., Jarboe, N., Vieglais, D., Fils, D., Duerr, R., Verhey, C., Minch, M., Mecum, B., Bentley, N., 2022. Science-on-Schema.org v1.3.0 (1.3.0). https://doi.org/10.5281/zenodo.6502539. Zenodo.

Solomons, T., Hinton, E., 2022. Federated searches: why a one-stop shop approach to literature searching falls short for evidence synthesis. JBI Evidence Synthesis: June 2021 19 (6), 1259–1262. https://doi.org/10.11124/JBIES-21-00177.

Theocharides, S., Payne, K., Van de Putte, A., 2021. VRE Hackathon: Defining the Needs of the Polar Data Community. 4th Polar Data Forum, Conference, 22 Sept 2021. https://polar-data-forum.org/workshops-hackathons/. (Accessed 31 May 2022).

UNESCO, 2021. UNESCO Recommendation on Open Science. Programme and Meeting Document. https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en. (Accessed 12 December 2022).

Verhey, C., 2021. Semantics and Vocabularies. 4th Polar Data Forum, Conference, 22 Sept 2021. https://polar-data-forum.org/workshops-hackathons/. (Accessed 31 May 2022).

Verhey, C., Minch, M., Payne, K., 2022. Polar Pilot Federated Search Roadmap: Year 1 in Review. Zenodo. https://doi.org/10.5281/zenodo.6344534.

International Science Council (ISC), 2021. Science and Society in Transition - ISC Action Plan: 2022–2024. https://council.science/wp-content/uploads/2020/06/202110_ISC-Action-Plan_ONLINE.pdf. (Accessed 31 May 2022).

Introducing schema.Org: Search Engines Come Together for a Richer Web, 2011. Google blog. https://developers.google.com/search/blog/2011/06/introducing-schemaorg-search-engines. (Accessed 31 May 2022).

Wilkinson, M.D., et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018. https://doi.org/10.1038/sdata.2016.18.

World Data System - International Technology Office, 2022. About. 29 Jan 2022. (Retrieved from https://wds-ito.org/about/). (Accessed 7 April 2022).

Wu, M., Juty, N., RDA Research Metadata Schemas, W.G., Collins, J., Duerr, R., Ridsdale, C., Shepherd, A., Verhey, C., Castro, L.J., 2021. Guidelines for Publishing Structured Metadata on the Web (3.1). https://doi.org/10.15497/RDA00066.

Wu, M., Hagan, P., Cecconi, B., Richard, S.M., Verhey, C., RDA Research Metadata Schemas, W.G., 2022. A Collection of Crosswalks from Fifteen Research Data Schemas to Schema.Org (1.0). https://doi.org/10.15497/RDA00069.