Polar Pilot Federated Search - Year 1 in Review

March 2022

(Formerly PPFS Roadmap)

A REPORT BY THE
World Data System • International Technology Office
Funded under the Alliance

Chantelle Verhey
Research Associate for the International Technology Office
ito-ra2@oceannetworks.ca

Melinda Minch
Web Developer for the International Technology Office
ito-webdev1@oceannetworks.ca

Karen Payne
Associate Director for International Technology Office
ito-director@oceannetworks.ca

Introduction

The products described in this document were created as part of the Digital Research Alliance of Canada (the Alliance), identified by the University of Victoria (UVic) as fund number 51946, between Ocean Networks Canada (ONC) and the Alliance, for work conducted by the World Data System - International Technology Office (WDS-ITO) between 01 April 2021 and 31 March 2022.As part of this agreement, the ITO is contributing to the POLDER WG. POLDER is a collaboration between the Southern Ocean Observing System, Arctic Data Committee, and Standing Committee on Antarctic Data Management. Specifically, the ITO is supporting POLDER's PPFS project. This document is a roadmap for the continued support and development of this project to ensure they 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.

This pilot project utilized resources provided by the World Data System - International Technology Office in the form of a dedicated Web Developer and Research Associate to help the POLDER community work towards project milestones. The work completed by the ITO was guided by an Advisory team formed of community volunteer stakeholders who are interested in this project.

Background

A federated search enables users to utilize a single interface and through one query, search for data from multiple metadata catalogs. The system sends the search out to multiple data providers and waits for a response, then compiles and organizes the results for review by the user (source). The POLDER Data Management community has been working towards a federated search system for the last 5+ years (POLDER federated search drive). This pilot federated search is intended to serve the polar research community by creating a single user interface for researchers, community members and data managers to search across numerous polar repositories in a single query. For many data types, there is no realistic prospect of standardizing and aggregating the data itself at this time; therefore, federated metadata search is the only viable way to make these datasets easily discoverable, and so maximize their value.

We are aware of other federated search avenues that the community is able to participate in. These alternative platforms include the National Snow and Ice Data Center (NSIDC) and DataONE federated searches. We would like to acknowledge the advanced services that are available for repositories to participate/ subscribe to, although the DataONE services are provided in a proprietary manner, and the NDSIC had an established portal that is no longer being maintained. The POLDER community has elected to pursue this pilot project in order to provide a non-proprietary option for community members to engage with. By combining the gleaner harvester and the dataone harvesting infrastructure, this project has been able to establish a way to meet participating repositories half-way. We have identified that a lack of capacity and resources have restricted repositories from participating in similar projects in the

past, so by combining these strategies, we are able to take some of the burden off of them and still have their metadata included in a single searchable metadata index. To elaborate, the justification for this infrastructure is to support open science which in turn gives us the opportunity to develop the software in a way that lowers the barrier to participation, reducing the amount of metadata development work repository managers have to do to be included. The more technical information regarding this infrastructure is further outlined below. There are numerous projects that are currently being developed similar to this pilot federated search. Specifically, the Ocean Info Hub (OIH), Canadian Consortium for Arctic Data Interoperability (CCADI), and Blue Cloud are utilizing the open source Gleaner harvesting tool to create a single user interface for their respective communities. In order to align initiatives with the ongoing projects, we are taking a similar approach to the pilot by utilizing the Gleaner harvesting tool to pull from the participating repositories. By aligning this project with others, we will ensure to utilize their lessons learned, as well as work collaboratively to avoid siloing.

In our approach, participating data repositories enhance their metadata by adding SDO mark-up and publishing them in metadata landing pages. The Gleaner tooling then creates indices of the markup from those landing pages and stores them in an RDF compliant Blazegraph. The user can then search the indexed pages, as well as metadata indices provided by DataONE in a single interface. The advisory group determined the best host for the Pilot UI was to temporarily be hosted on DataONE servers.

Scope

Members within the Polar community have expressed important challenges that this project will help to address. We documented these challenges to ensure creative solutions and lessons learned are captured in every step of the development process. The first difficulty noted for this project is that the pilot federated search will require quite a lot of developer time in order to develop a User Interface. Similar projects have spent years and countless web development hours to create a robust federated search engine. This pilot project's scope is to establish a basic user interface/ federated search and establish the basis for future development if the interest from the community is present. Additionally, the Polar community is composed of diverse organizations with complicated relationships. A large majority of repositories/catalogues that host polar data are not strictly Polar repositories, rather they host multidisciplinary data not restricted by geographic boundaries. Below, we will define the spatial ‘Polar’ region to be included in the search, although it is still to be determined which repositories will be included in the search once the pilot turns into a fully functioning federated search. For example, additional functionality developed after the pilot project phase would include the user's ability to filter repositories that are strictly Polar versus a more generalist repository that happens to host some Polar data.

The following list is the name of repositories that have committed to being included in the Pilot Federated Search:

1. Australian Antarctic Data Centre
2. NASA/GCMD

3. Arctic Data Center
4. Netherlands Polar Data Center
5. National Snow and Ice Data Center
6. BCO-DMO
7. USAP-DC
8. CLIVAR and Carbon Hydrographic Data Office (CCHDO)
9. British Antarctic Survey (BAS)
10. Greenland Ecosystem Monitoring
11. CCADI
    a. Canadian Watershed Information Network
    b. Polar Data Catalogue
    c. Nordicana D
    d. Committee on Earth Observing Satellites (CEOS)
    e. Arctic Science and Technology Information System (ASTIS)
    f. ArcticConnect
    g. Arctic Spatial Data Infrastructure (ASDI)
    h. INTERACT
    i. Inuvialuit Regional Corporation (IRC)

The Advisory Team

The advisory group was established during the 6th Polar to Global Data Interoperability Hackathon on the 8th of June 2021. At this specific hackathon, the project proposal was presented to community members and the support/viability of the project was discussed. The overall consensus was that this project would further benefit the polar community and enhance current interoperability initiatives. The advisory group had separate meetings quarterly, meanwhile additional feedback and guidance was provided during the bi-montly Polar to Global Hackathons. The following are the individuals who volunteered to participate in an advisory role for this pilot project.

Members:

- Pip Bricher (SOOS/POLDER)
- Doug Fils (ODIS)
- Adam Shepard (BCO-DMO),
- Matt Jones (ADC - DataONE)
- Alice Fremand (UKPDC)
- Taco De Bruin (IODE)
- Shannon Christoffersen (AINA)

The Audience

Early consultations with members from groups such as the Southern Ocean Observing System (SOOS), Standing Committee for Antarctic Data Management (SCADM), Canadian Consortium for Arctic Data Interoperability (CCADI), POLDER, NSF Arctic Data Center (ADC), and the Polar Data Catalogue (PDC) created a consensus that a polar federated search project

should include an array of repositories from across the polar community. The WDS-ITO then led a round-table at the 6th Polar to Global Data Interoperability Hackathon, which is a platform for all Polar organizations to have the opportunity to come together to discuss polar RDM initiatives.

The pilot consulted with the first 10 repositories (identified above), who had volunteered from the POLDER working group, to assess the current state of their landing pages and schema.org. The POLDER WG has been diligently working on a 'Best Practices' guide which utilizes the Schema.org - science-on-schema (SO-SO) guidelines, and the POLDER community selected a subset of the SO-SO recommendations with a tailored focus for polar repositories. Following the Pilot, the WDS-ITO and POLDER WG will work with interested repositories to implement the POLDER Best Practices.

For this project, the Polar region will be defined as above 50 Degrees latitudinal north in North America, Scandinavia, Asia; 40 degrees latitude north in Europe and below 40 degrees latitude south in the southern hemisphere. Additionally, metadata with the keyword tags 'arctic', 'antarctic', 'polar', will show up in the results. This approach is to ensure that the search identifies all relevant data that may otherwise be excluded due to geographical locations of these regions varying across scientific domains.

Measuring Success

During the 2021-2022 funding year the Advisory team has identified and advanced some technical issues around a discovery process and worked with pilot members for proof of concept; this was determined to be sufficient for the current funding cycle. The WDS-ITO dedicated Web developer is to contribute time and space along with text search over SDO to the gleaner tool - spatial search is being implemented by many at the ESIP SDO SO-SO cluster. The following are the ideal features to be included in the UI. These features were identified to be

- A search tool that can do the following:
    1. Text search
    2. Basic date search
    3. Basic spatial search
    4. A few other fields identified by community feedback (ie Author, Institution, keywords)
- A portable or containerized distribution mechanism so that this project can be deployed and moved easily

Milestones

- 2021-**May**
    - Establish timeline/viability/support
    - Meeting with Polar Leads
- 2021-**June**:
    - Project Discussion @ PtoG Hackathon → brainstorm/organization
    - Web Dev Posting

- 2021-**July**:
  - First Meeting
  - Advisory Team Roadmap-draft 1 Review
- 2021-**August**:
  - Completed interviews for Web Developer
  - Web Developer offer letter accepted by successful candidate
- 2021-**September**:
  - PDF4 Hackathon
    - Pilot Repositories identified
  - Developer comes on board
    - Roadmap- Draft 1 refined
- 2021-**October**
  - Advisory Team Meeting #2
  - Beginning of Repository implementation sessions
  - Supporting work on Gleaner
  - Initial simple version of the web app that federates text-based searches from a Gleaner repository and DataOne
- 2021-**November**:
  - Easy and flexible deployment at https://api.test.dataone.org/polder/
  - Continued supporting work on Gleaner
  - Better user interface
  - Roadmap Draft 2
- 2021-**December**:
  - Best Practices Updated
  - PtoG Hackathon
  - Include more repositories
  - Start helping some data repositories get their metadata up to standards
  - Add Gleaner support for common data repository problems, if applicable
- 2022-**January**:
  - Finish Roadmap
  - Date search
  - Including more repositories
  - Continuing Gleaner support
- 2022-**February**:
  - PtoG Hackathon
  - Including more repositories
  - Continuing Gleaner support
  - Search tuning and refinement
  - UI tuning and refinement
  - Meeting with CCADI Team to start the process of including their consortium of repositories
  - Deployment at http://search-dev.polder.info
- 2022-**March**:
  - Devise a method to build a sitemap from a DataCite query, and crawl it

- Continuing Gleaner support
- Best Practices V1 Completed
- Final Advisory Team Meeting for Pilot
- Complete Pilot/ Launch Party! At AOS

Documentation

Throughout this project we have thoroughly documented all processes in order to provide a complete picture of the federated search to the final hosting repository. The following are a list of documents that describe the project's process.

- Federated Search App Documentation
    - Instructions for deployment and development
- Gleaner Tooling Documentation
- PPFS Journal Article
    - Describing the technicalities of the Infrastructure/software

The following diagram depicts the PPFS architecture and workflows. The beta site search.polder.io that harvests polar data is now operational. The site is currently hosted by our partner DataONE and the ITO is simultaneously preparing for deposits from additional repositories, fixing bugs and enhancing the front end. As noted above, we have developed the gleaner tool and deployed it using POLDER best practices on DataONE servers (beta site). Figure 1 shows the components of the site, with the ITO development on DataONE servers inside the red circle.

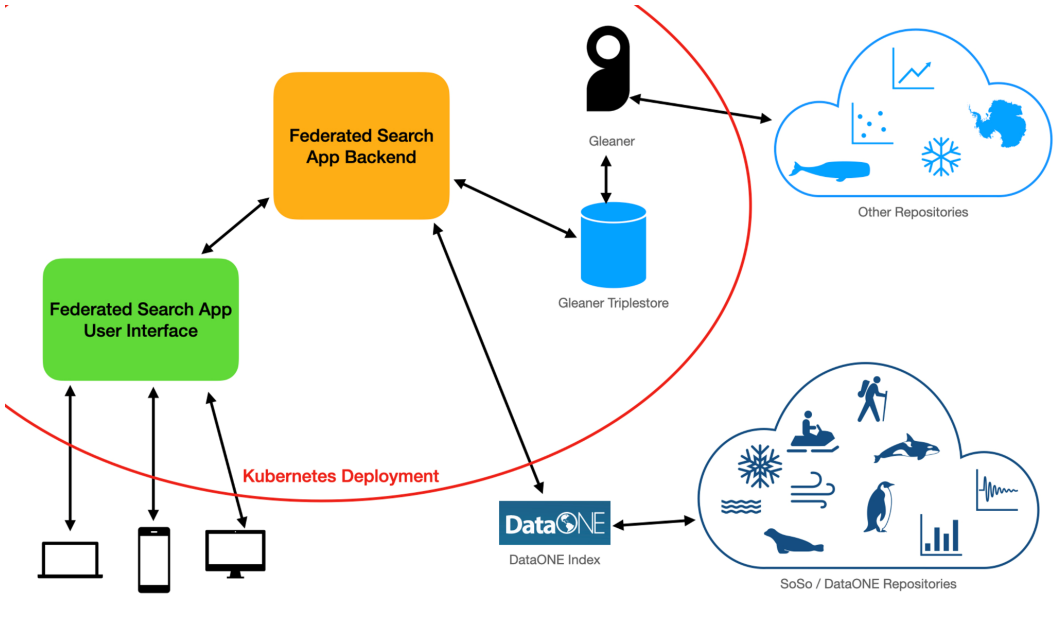

Figure 1: This diagram visualizes the relationships between the repositories, aggregate services, federated search app, and final user interface/consumption. The PPFS is unique from other similar initiatives in its aggregation because it was designed to utilize both the dataone harvesting services and the gleaner infrastructure to incorporate both into a single umbrella.

Features developed during the PPFS

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 usable 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	In order 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.

Next Steps

Once the Pilot concludes in March of 2022, the ITO intends to dedicate resources for taking the federated search into full production for a full year (April 2022 - March 2023). This includes activities such as establishing usage data analytics, assisting more repositories with their SDO mark-up to be included in the search, and responding to feedback from the community.

Anticipated Enhancements

We are in the process of collecting community feedback about which direction this work should take next, but in broad terms, our priorities are as follows:

1. Add a polar map representation / search for the data
2. Adding more data repositories
3. Displaying more or better information in the search results
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
6. Supporting advanced searches, like text searches that include boolean operations

Through community feedback, the dedicated development team intends to continue enhancing the search -- the exact nature of this work is to be determined, but it is likely to include tuning, scoping, traditional knowledge (TK) label support/ search faceting, and possibly language translations. 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.

The following list of documents are not limited to the Pilot, but are intended to be included in the larger picture of the project. We will deliver documentation that supports users, developers and data managers who want more information about the project, such as:

1. Relevant background and context-setting information, or links to such information
2. Instructions for use / searching (ie User guide)
3. Instructions for linking data repositories(ie ITO SDO primer, POLDER Best practices doc link)
4. Common problems and how to solve them

SWAT Team Development

Moving forward, we are assessing whether the Advisory team will continue to meet separately or merge the meetings into the Polar to Global bi-monthly hackathons. In that space, the Advisory team would assist the volunteer pilot repositories to implement SDO-POLDER Best Practices and be integrated into the pilot search. The first largest lift that the Advisory team will

need to complete is to help repositories set up their SDO so that the search appliance can be successful.

Specifically identified by the advisory team is the creation of a ‘SWAT’ team. This team would host additional meetings as needed and its goal would be to complete 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 effectively included.

The polar community is working on securing more resources for the next phase of the Polar Federated Search through BELSO. If they are funded, they have committed to providing resources to the project, in particular, assistance and resources dedicated to EU inclusion

Conclusion

At the end of February 2022, the WDS-ITO deployed a survey to help determine immediate needs, priorities and feedback on the current state of the federated search. We are committed to ensuring that we are addressing the wants and needs of the community, and dedicating time to the features that matter most for them. The results will be presented to the 2022 Arctic Observing Summit as well as a ‘Launch Party’ for the kick off of the deployment, which will occur in the Data Sharing working session.

Finally, at the conclusion of the full production year, once the federated search is up and functional, the final homing site of the search will be established. At this time, the pilot is to be hosted on available DataOne servers, and the advisory team will reassess if it shall remain at its current location or if it shall be moved to another repository. The first year of production will have a roadmap for its initiatives and developments, including where it will be promoted, new features, and it will be available on the production site.