UiT

THE ARCTIC UNIVERSITY OF NORWAY

Open Arctic Research Index 2019

Final report and recommendations

University library
Tamer Abu-Alam (OpenARI project manager, University library)
Tamer.abu-alam@uit.no

Filling a 60% findability gap of the open-access polar research data and documents

UiT The Arctic University of Norway – 2019

Septentrio Academic Publishing

http://septentrio.uit.no/

Septentrio Reports, number 3, 2019

ISSN: 2387-4597

DOI: https://doi.org/10.7557/7.4682

How to cite this report: https://doi.org/10.7557/7.4682

UiT Norges arktiske universitet
N-9037 Tromsø

Sentralbord: 77 54 40 00
Faks: 77 64 49 00

postmottak@uit.no
www.uit.no

EXECUTIVE SUMMARY

Access research data and research documents (e.g. publications) and make it more visible and findable through the internet is coming up as one of the major challenges for future development of the next generation of Digital Libraries. This challenge becomes more complicated when data producers (e.g. research institutes) are not aware by the needs of the scientific community for visibility and findability of their data or when the data producers lack the technology or the motivation to make their data available online.

Although the Open Arctic Research Index pilot project (here and thereafter as: OpenARI) focused only on the open-access research data and the open-access research documents published on Polar Regions, the OpenARI found 60% of these open-access records are unfindable through searchable platforms outside the institutional webpage itself. This raises an awareness sign of the need of the scientific community to harvest the metadata of these open-access records in a homogenous, seamless database and making this database available to researchers, students and publics through one search platform. At present, neither Google Scholar nor any other search platform provide this service.

Based on the obligations and the motivations of the University of Tromsø – the Arctic University of Norway (UiT) toward making the research data and research documents available to the scientific community, UiT launched the OpenARI as a pilot project to analyse the success opportunities and the challenges of creating a new service in which the metadata of all records openly available on issues of relevance to the Polar Regions will be collected, sorted and archived in a database. The final database will be available to researchers, students and publics through an easy, searchable user-interface.

Based on the fact that around 60% of the open-access polar records are unfindable through one search platform, we strongly suggest launching a full-scale management service. This new service will be built on existing experiences from High North Research Documents (i.e. an existing service at the UiT). In this final report, the need of new technical solution (based on open source technology) by which the OpenARI will be able to collect all the published material using algorithms that allow the best filtration processes will be clarified. Moreover, the OpenARI has mapped 115 major and relevant metadata providers that potentially can support and feed the full-scale model with content.

OpenARI has concluded fifteen needs that are required for the full-scale management model. In addition to the main service (i.e. make open-access polar records more visible and findable through one search platform), we suggest to add three new services: 1) hosting of original data from the Polar Regions; 2) creating a research platform; 3) creating an education platform. A new process including four stages of filtration is suggested in order to reduce the time and the overhead costs of using the UiT's server. End-users will be able to perform search using a map. In addition to the classical way of presenting the results of a search, the end-users will be able to see the search results on a map and/or as a timeline.

CONTENTS

EXECUTIVE SUMMARY 1
INTRODUCTION 3
Organization 4
TERMS AND ABBREVIATIONS 4
THE NEEDS OF THE SERVICE 6
60% findability gap of the polar records 6
Meet the strategic plans of UiT, NPI and Norwegian Research Council 11
Added values 12
TOWARD A UNIQUE SERVICE 12
A comprehensive service 12
Reference board 14
Editorial board 14
Clarify the need, solution and content framework of the service 14
Suggestions for technical solutions 17
Prototype 18
Outreach plan 19
RESOURCES AND TIME FRAME 19
ETHICAL CONSIDERATION 20
RECOMMENDATIONS 21
ACKNOWLEDGMENTS 21
REFERENCES 21
APPENDIXES 22
Appendix A 22
Appendix B 28
Appendix C 32
Appendix D 33
Appendix E 35
Appendix F 44

INTRODUCTION

The Open Arctic Research Index is the name of a pilot project, which aims to analyse and measure the success opportunity of a new planned service at the University of Tromsø – the Arctic University of Norway. The scope of the service is to collect, sort and archive metadata about open-access publications and datasets on the Arctic region. These records (i.e. metadata on the Arctic's publications and datasets) will be available to researchers, students and publics through an interactive searchable front-end website. The main goal of the pilot project was to assess the possibilities to build such a service and estimate the resources needed.

The main aims of the new service were singled out as follows: 1) to make results from arctic research (both research articles and research data) more visible and better retrievable through a common search index based on a standardized, interdisciplinary metadata set; and 2) support researchers and students in their research by making articles and datasets searchable in one and the same service. However, based on the progress of the pilot project, we suggest here additional services that in our view will make the planned service a unique platform for research and education in the polar sciences.

After the first discussion among the team members, we suggest to not restrict this service to only the Arctic region, but it should cover both the Arctic and the Antarctic regions (i.e. Polar Regions). The reasons behind this suggestion are: 1) the Arctic and the Antarctic regions are analogue to each other; 2) many researchers who are working in the Antarctic region are active in the Arctic region as well; 3) different research funding agents (including European/Scandinavian states funding agents) maintain and operate a large number of observatories and research stations both in the Arctic and Antarctic and as a result the research activities in both regions are linked and relevant (e.g. [1]); and 4) we will increase our audience by covering the all-Polar regions. Therefore, we suggest changing the name of the planned service to an expression that includes "Polar" instead of "Arctic".

It is worth mentioning here that the new planned service will be built on top of an existing service at the University of Tromsø – the Arctic University of Norway, namely the "High North Research Documents" – http://highnorth.uit.no/. The High North Research Documents is a service based on metadata collected from an international partner, "BASE" – Bielefeld Academic Search Engine. To reduce the overhead costs of the new planned service, we will rely on BASE as well; however international collaboration with other institutes/organizations will be required in order to have a better and wider database that covers the polar sciences in deep.

This final report is structured to show the conclusions and the recommendations of the OpenARI. In order to do that, the report will start with presenting the needs of the scientific community for such a service, followed with the consistency of the service with the vision of the University of Tromsø – the Arctic University of Norway, the Norwegian Polar Institute and the Norwegian Research Council. A technical section will clarify the needs, solution and content framework of the service. The technical section will present a prototype model of the service. The report will be concluded with a section discussing the resources and suggesting a timeframe. As we will use certain specialized terms, a section that summarizes and defines these terms will follow the introduction section.

Organization

The OpenARI pilot project is a joint project between the University of Tromsø – the Arctic University of Norway and the Norwegian Polar Institute. The University Library (UB) hosts the project.

TERMS AND ABBREVIATIONS

UiT: is the University of Tromsø – the Arctic University of Norway.

NPI: is the Norwegian Polar Institute.

BASE: is the database of the Bielefeld Academic Search Engine.

The service/new service/planned service: when it is used it means the name of the service that will be hosted by the University Library, UiT, if the proposed project is approved. The team of the OpenARI suggests using a name that includes "Polar" instead of "Arctic".

Admin of the service: member of the project's team who will take care of the IT part of the project and who are communicating with the metadata providers utilized by the service.

Records: means metadata about various document types like research articles, theses, maps, images, etc. and about research datasets that will be included in the service database. When the word "records" is used here it means only the open-access records.

End-user(s): means the researchers, students, teachers, public services as well as the interested public, who will use the user-interface of the planned service to search for records.

User-interface: is the platform / website that will be used by the end-users in order to perform a search.

Back-end stage: is the processes that occur on the UiT's server which include creating databases, read metadata from different data providers, filtering data, fill up the service database, sorting the records and filling gaps in the metadata of records.

Front-end stage: is all the processes that is linked to the end-users and the user-interface.

Service open-access database/service database: means a database that contains only metadata about records relevance to the Polar Regions.

Polar Regions: definition of Polar Regions may vary from a science to another e.g. social science and humanities will like to draw the line farther south for the Arctic region than a biology or a meteorologist will tend to do.

Based on the discussion with the steering group of the OpenARI, the OpenARI team suggests to define the Polar Regions to be:

Antarctic Region: Antarctic Polar Front (i.e. red line of Fig. 1) will define the Antarctic Region. The Antarctic Polar Front is a curve continuously encircling

Antarctica, varying in latitude, where cold, northward-flowing Antarctic waters meet the relatively warmer waters of the subantarctic.

Arctic Region: A combination line (i.e. red line of Fig. 1) of different definitions will be used to define the Arctic Region. This line is considered to cover the largest geographic area, which covers most of the cross-disciplinary definitions of the Arctic.

Based on a suggestion from an evaluation report commissioned by the Norwegian Research Council [2], we have included part of the mainland of Norway in order to include important terrestrial and social science research. Although this extension is not aligned to the official Norwegian definition of the Arctic, the polar research community is stressing to extend the boundary to include major parts of the mainland. Our definition of the Arctic region may extend to include other parts in north Sweden, Finland and Russia; however, we will rely on a reference board to discuss and modify these boundaries.

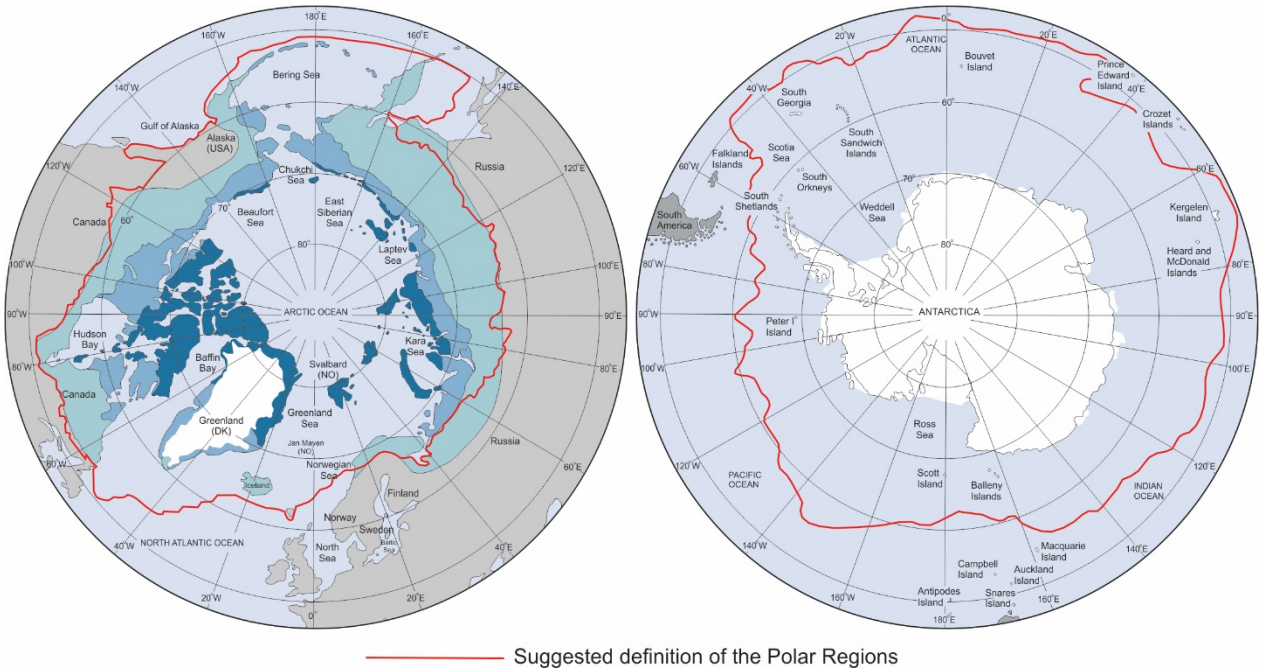


Fig. 1: The suggested definition of the Polar Regions. Note: the definition will be subjected to a discussion and modification according to suggestions from the reference board.

Reference Board: is a selected group from researchers and scientists who are active in polar sciences. The group should covers social, humanities, physical and life sciences.

Editorial Board: is a selected group represents the national and international partners. The editorial board will suggest new metadata providers from their countries. Moreover, the members of the editorial board will help to outreach the service in their countries.

Harvesting metadata: is an automated collection of metadata descriptions from different sources (i.e. pre-defined metadata providers) to create useful aggregations of metadata about open-access records on Polar Regions.

OAI-PMH: is Open Archives Initiative - Protocol for Metadata Harvesting. The OAI-PMH is a low-barrier mechanism for repository interoperability. According the definitions of the OAI-PMH, the metadata providers are repositories that expose structured metadata via OAI-PMH. Service providers then make OAI-PMH service requests to harvest that metadata. According to these definitions, the planned service at the UiT is classified as service provider.

API: Application Programming Interface (API) is a set of subroutine definitions, communication protocols, and tools for building software and websites. The API is a set of clearly defined methods of communication among various components of any website. In the suggested service, metadata on polar records can be harvested direct from the metadata providers' websites, if these websites are using API.

Our/we: when it is used it means the team of the OpenARI in some context but in other context, it means UiT.

THE NEEDS OF THE SERVICE

60% findability gap of the polar records

In order to test the needs of the scientific community to a new service that will provide them with access to the records published on the Polar Regions through one search platform, 115 major and trusted institutes/organizations/research units (here and thereafter as: metadata providers or only providers) were mapped (Appendix A). We have focused only on providers who are dealing with polar sciences. The different metadata providers were classified based on their relation to the BASE and High North Research Documents into three categories (Fig. 2):

1- Metadata providers not included in BASE and High North Research Documents (58 providers; Fig. 3).

2- Metadata providers included in BASE and High North Research Documents (21 providers; Fig. 3).

3- Metadata providers included (partially - as publisher) in BASE and High North Research Documents. These providers do not give full access to all their catalogs (34 providers; Fig. 3).

We have used the High North Research Documents and the BASE as a base of our analyses and as examples of the common search engines, however, major search engines (e.g. Google) were used to validate the results. The reasons of choosing the High North Research Documents as an example of the common search engines are that the database of the High North Research Documents contains more than 1,000,000 records from the Arctic Region. Most (82%) of the records in the High North Research Documents' database are publications and documents, and only 18% of the content are metadata about research data. In the modern day, using open-access research data becomes more significant. For this reason, we have focused here on the research data (i.e. most of the 115 metadata providers are hosting research data, but some of them host both research data and publications on the Polar Regions).

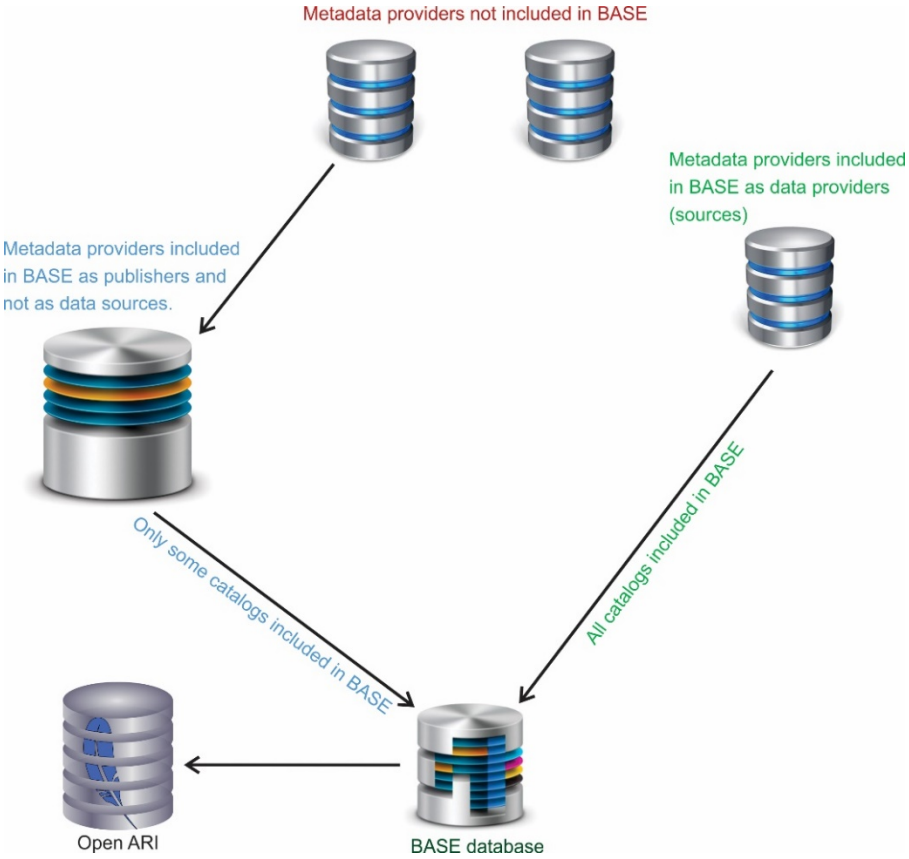


Fig. 2: The three different categories of metadata providers and their relation to the common search engines (e.g. BASE and High North Research Documents).

Numbers of records from metadata providers that are not included and partly included in BASE and High North Research Documents were used to indicate the findability gap of the polar records. Interestingly, the numbers of records that are unfindable through common search engines represent around 60% of the total numbers of the polar records (Fig. 4). This high percentage raises a series of questions. One of these questions is “why is such a high percentage unfindable although we targeted major and trusted data providers?”.

To answer such questions, we have performed a short survey of 7 questions (Appendix B) asking the metadata providers (i.e. the 115 providers of the Appendix A) of their knowledge of OAI-PMH service and if they allow harvesting their metadata through this service or through other common search engines. This survey run from October 2018 until February 15, 2019. We have received 52 responses, which represent about 46.02 % of the mapped providers (i.e. 115).

In the main text of the final report, we will focus on only three questions of this survey, which are: 1) are your metadata harvested by common search engines? 2) can your database be harvested via the OAI-PMH protocol? 3) does your database have its own API that allows to extract metadata?. Figures (5, 6 and 7) show a visual summary of these three questions, while appendix B shows the full survey results.

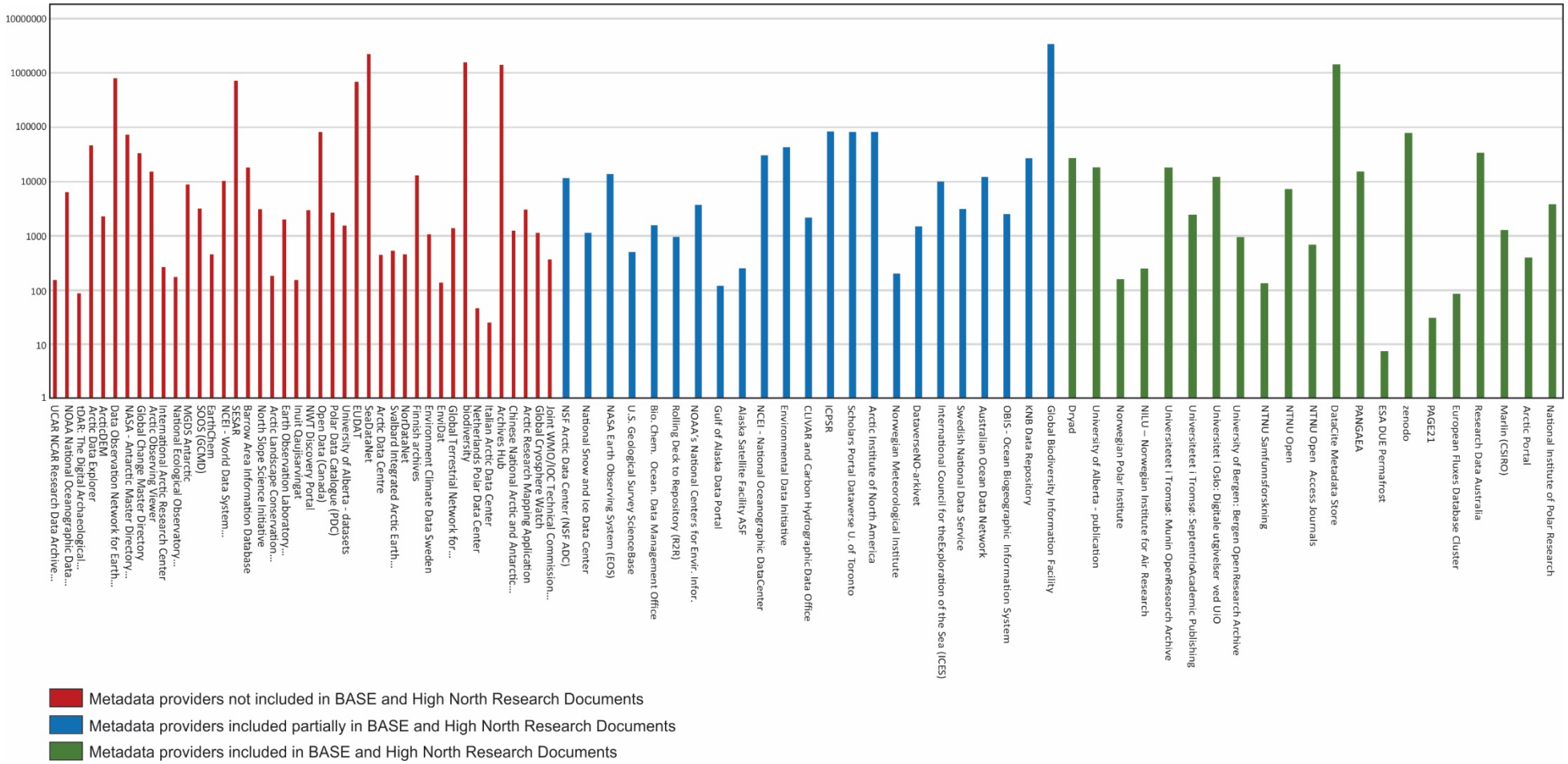


Fig. 3: A histogram shows the numbers of the polar records in the different metadata providers.

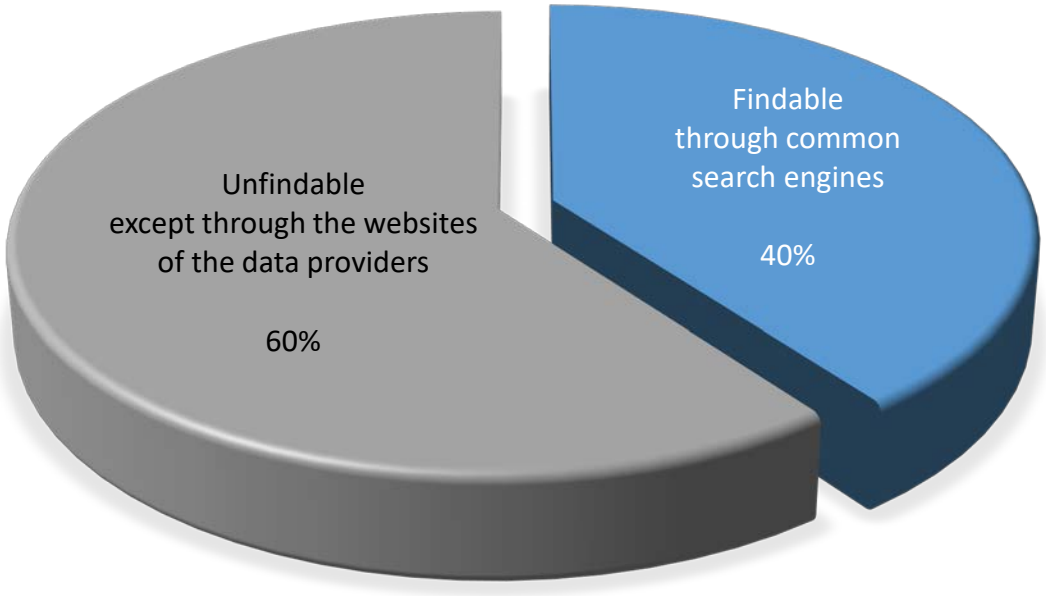


Fig. 4: 60% findability gap of the polar records based on the records numbers of 115 metadata providers.

Only 19.2% (Fig. 5) of the responded metadata providers allow common search engines to harvest their data. This means that the current search engines' databases are missing about 80.8% of records related to the polar sciences. Figure 6 shows a summary of the answers of the questions "Can your database be harvested via the OAI-PMH protocol?". 46.2% of the responded providers allow harvesting their metadata using OAI-PMH protocol, the rest (i.e. 53.8%) do not use OAI-PMH protocol or not aware of such protocol. 34.6% of the responded providers do not use API that allows extracting metadata from their websites.

These three questions show that 34.6% of the providers lack the technology that makes their records more findable, 53.8% of the providers have no time to adapt their contents to be compatible with the OAI-PMH protocol and 80.8% of them lack of the awareness of the needs to make their records searchable through common search engines. This may explain the reasons of the 60% findability gap of the polar records and reflects the importance and the needs of a new service, which will consider the unfindable records.

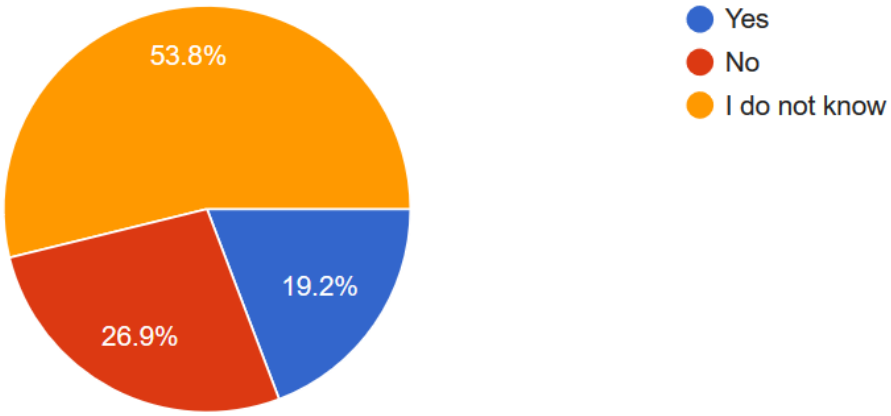


Fig. 5: a summary of the answers of the questions “Are your metadata harvested by common search engines?”.

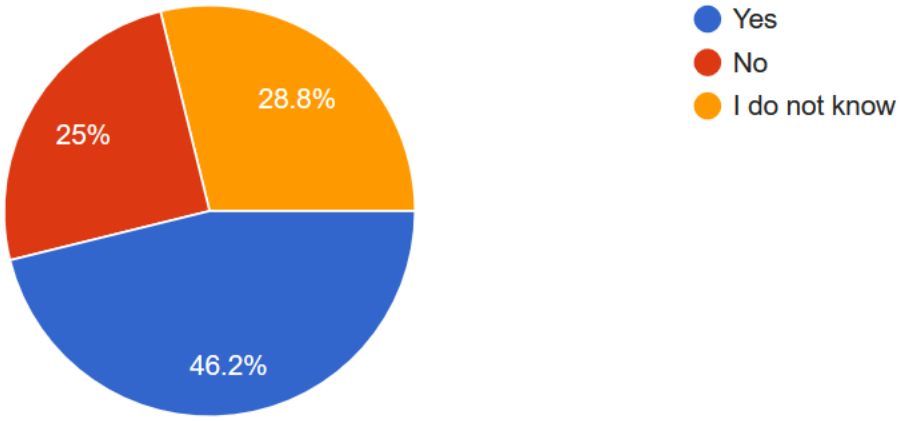


Fig. 6: a summary of the answers of the questions “Can your database be harvested via the OAI-PMH protocol?”. 46.2% of metadata providers allow harvesting their metadata using OAI-PMH protocol, the rest (i.e. 53.8%) do not use OAI-PMH protocol or not aware of such protocol.

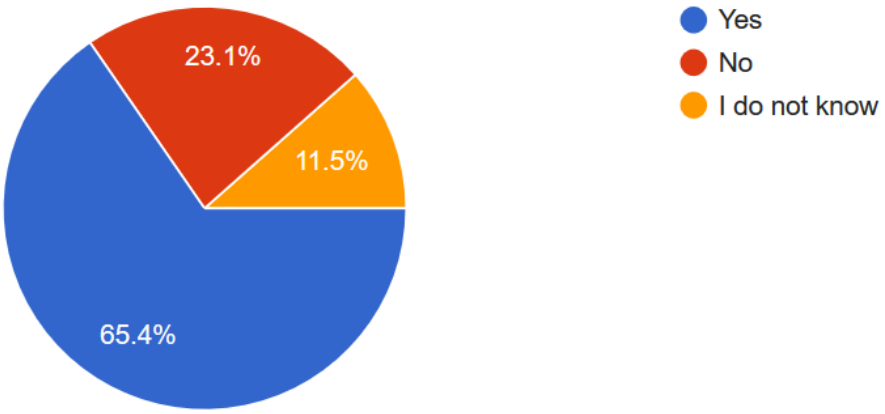


Fig. 7: a summary of the answers of the questions “Does your database has its own API that allows to extract metadata?”. 65.4% of metadata providers allow extracting their metadata using API. 34.6% do not use API.

Meet the strategic plans of UiT, NPI and Norwegian Research Council

The new suggested service at the university library lies directly at the core of the strategic plans of both the University of Tromsø – the Arctic University of Norway 2014-2022 [3] and the Norwegian Polar Institute 2019-2024 [4]. Moreover, the new service meet the main vision of the Norwegian Research Council on the Polar Regions [2].

The UiT aims to help promote economic, cultural and social development in the north including the Arctic region through building knowledge and human capital. This will help to promote people's quality of life. For example, understanding what happens in the Arctic and the Antarctic regions is a key to understanding global climate and environmental changes. Studying Arctic life and how it is affected by global environmental changes is another challenge facing UiT's researchers. Societal and business development in the north depends upon future-oriented sustainable management of natural resources including conventional and renewable energy production. All of these research points and others (which are linked directly or indirectly to the polar research) require researchers to do extensive searches for existing knowledge, research data and research documents. This is true not only for the UiT but also for NPI as a responsible governmental institute to do research, taking care and running the infrastructures at both Svalbard and Dronning Maud Land.

When it comes to educational activities, it is a priority to UiT to develop and implement new educational tools and student active learning methods including developing digital technology competence and web-based teaching methods. In the suggested new service at the university library, we aim to develop an educational platform in order to collect teaching materials related to polar sciences (e.g. online lectures and simplification of research data to meet the requirement of teaching) as an additional subservice to the main service (i.e. make open-access polar records more visible and findable through one search platform). Moreover, we have the ambition to develop a simple way of virtual teaching; this will help and will reduce the costs of the too expensive fieldworks in Polar Regions. These education and research activities of UiT in the Arctic will definitely consolidate the collaboration between UiT and the UNIS (the University Centre in Svalbard), which is another target of the vision of the UiT.

At the international level, a comprehensive evaluation report commissioned by the Norwegian Research Council has highlighted the importance of Norway to take a leading role in scientific and political affairs relevant to both the Arctic and Antarctic [2]. According to the expert committee, Norway should enhance the quality and impact of its polar research by establishing and promoting a national open data policy, which is the core idea behind the current proposal. The policies of Science Europe and European Research Council encourage different universities and research units to go more toward the open sciences (e.g. Plan S), the suggested service will work to broadcast this policy by making all the open-access research data and publications on Polar Regions more visible. Moreover, the evaluation report on Norwegian Polar Research stressed the needs of developing and implementing a plan for recruitment and retention of a diverse next generation of polar researchers. Providing current students with a tool that helps them getting access to knowledge published on Polar Regions through an interactive search platform will help reaching this goal. The suggested service will cover the first two principles of FAIR data guidance [5; i.e. findable and accessible], by making 60% of open-access research data more visible and findable, and thus more accessible to researchers, students and publics.

Although the visions of UiT, NPI and the Norwegian Research Council are concreted around the education and the research activities in the Polar Regions, these visions are challenged by

60% findable gap of the data and the research documents that were published on Polar Regions. The UiT in collaboration with NPI has to fill this gap in order to allow researchers and students to take advanced steps in their research. Our priority of the suggested service is to cover this gap.

Added values

Although our target is clear which is making the open-access polar records more visible and findable through one search platform, our service should be a unique service attracting a wide range of audiences (e.g. researchers, students, publics).

Since the suggested service will cover cross-disciplinary sciences and will focus mainly on Polar Regions, it will be the best helping tool for politicians and decision makers. Many common search engine (e.g. Google Scholar) can help researchers and students to perform search on publications, but our service will focus on both publications and research data (i.e. very few search engines consider the research data as a target for their services). Moreover, we will provide the end-users with quick access to open publications and research data through presenting links for each record that allow end-users to get what they are looking for in the shortest time. In addition, the end-users will be able to perform search on map (and see the results plotted on map), which will give them visual focusing on certain geographic locations of their interest.

Our proposal will lead to a high-quality service since it will be based on continuous recommendations from references and editorial boards (as will be explained in the next section of the report). The references board will drive the content of the service to meet the needs of the scientific community, while the editorial board will suggest new metadata providers and will help to broadcast the service internationally.

We aim to add three additional subservices to the main one. These subservices will cover 1) hosting original data, 2) creating research platform, and 3) helping in teaching activities of polar subjects by creating an education platform. These three subservices in additional to the main search service will add a great value for researchers, students and teachers.

In contrast to most search engines, we come here with a solid outreach plan. For example, the service will have a friendly, easy to use user-interface, which will allow end-users to share their results among different channels of social media. This will attract more audiences. Moreover, we will use our wide connections to different institutes, research units and universities focused on Polar Regions (e.g. The University of the Arctic - UArctic) to promote the new service. Attending international conferences on polar sciences and climate changes by posters and flyers will attract the attention to the service.

TOWARD A UNIQUE SERVICE

A comprehensive service

The main service of the current proposal is to make metadata on open-access polar records more visible and findable. However to be a comprehensive and unique service that attract more researchers and students to use the service website, we suggest to add three additional subservices which will be discussed in detail in the next paragraphs. In the main service (i.e.

make open-access polar records more visible and findable), we plan to maximize the numbers of relevant records in our database. Therefore, metadata from different providers are needed to be added. The team of the OpenARI has mapped 115 major metadata providers nationally and internationally (Appendix A). By adding the records of these metadata providers, the numbers of the records in our database will be extended to the maximum (as possible), covering a wide range of subjects. We plan to use Dublin Core Metadata schema (Appendix C). Although the Dublin Core Metadata is a standard schema used in the communication among different metadata providers, many providers do not use this schema. Even some of the providers who are using the schema, they do not use it in a proper way (i.e. some information are entered in wrong fields). We are suggesting here to use a combination of MySQL and Solr technology together with GO and php programming codes (see the suggestions for technical solutions section for more detail). This combination will allow us to read, organize and re-sort the metadata fields. Moreover, we plan to add geographic location information (i.e. latitude and longitude) for each record using a pre-existing geo-database. The new organized database will be available to be harvested by other search engines (as a contribution from the UiT to the international scientific community).

In additional to be a search engine makes open-access polar records more visible and findable, we suggest three new subservices (Fig. 8): 1) archiving original data from the Polar Regions; 2) creating a research platform; 3) creating an education platform.

As the 13th Munin Conference on Scholarly Publishing (UiT – 28-29 November 2018) has highlighted the importance of making the research data openly accessible to researchers, we suggest here hosting of original data from the Polar Regions. This service will be limited to 1) researchers who have not access to alternative data repositories for archiving, 2) researchers from UiT by linking the suggested service to the UiT's archive of the Open Research Data; and 3) other partners who have agreements with UiT.

Creating an open-access research platform where the researchers can create teams to discuss data and publications related to the polar sciences is another subservice that can be added to our main service. Several existing platforms aim to connect researchers (e.g. Open Science Framework; Connected Researchers). It is our target to make our database more visible and easy to be handled by such existing platforms. This suggested subservice will help discussing the existing data and therefore finding gaps in the knowledge on the Polar Regions.

The existing service at the UiT (i.e. High North Research Documents) is used by some universities for educational purposes. However, the current way of presenting the records in this service is made mainly for research and not for educational purposes. Here, we suggest to develop an educational platform helping teachers to find suitable material for their lectures. In this context, we suggest that this subservice can be performed in three steps: 1) collect a series of online lectures and talks on the polar sciences and make it available and searchable to students; 2) simplify some selected research data to be used for the educational purposes; 3) using virtual/gaming technology to prepare online courses and/or field training.

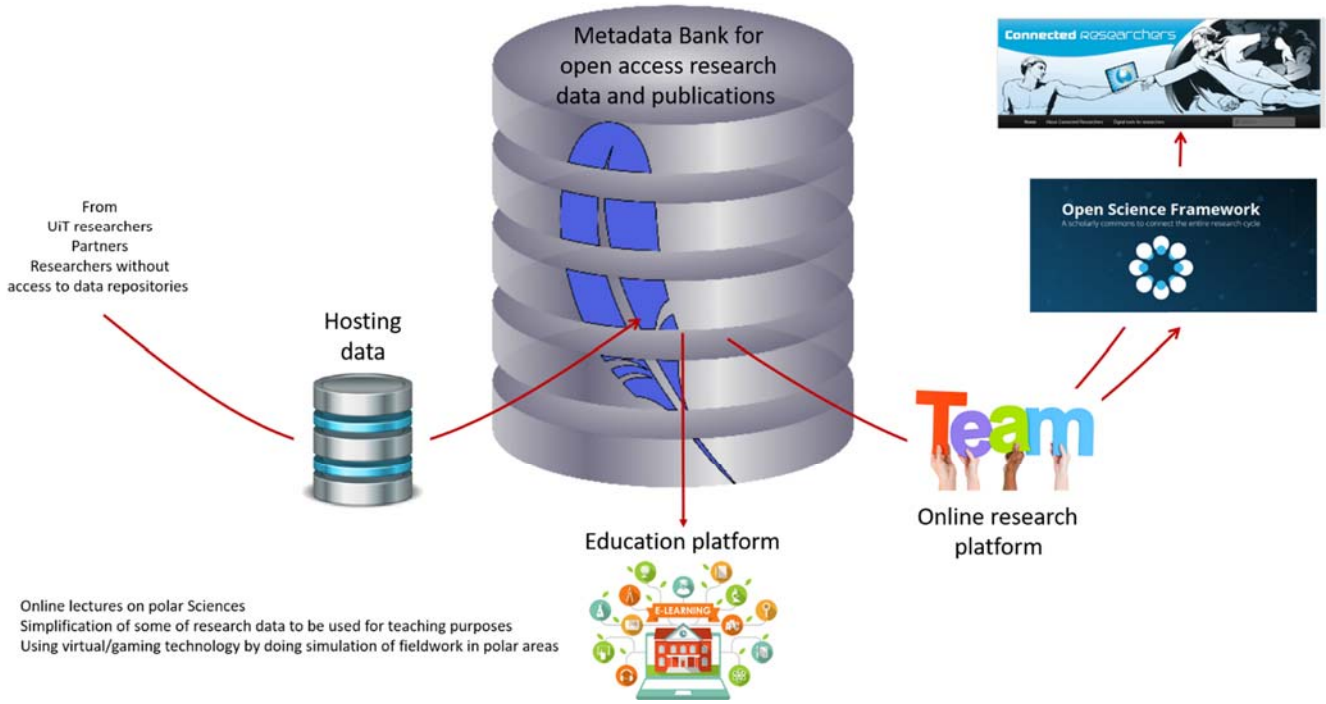


Fig.8: a chart shows the integration between the main service (i.e. make open-access polar records more visible and findable through one search platform) and the suggested three new services.

Reference board

To be sure, that the content of the service is up-to-date and meets the requirement of the scientific community, a reference board, which is a selected group of researchers and scientists who are active in polar sciences, is needed. The reference board will give advices on the content, and how to present it. Since the proposed service is multi-disciplinary, the reference board should cover social, humanities, physical and life sciences.

The reference board will be an asset to the service and make it a unique service in comparison with similar common search engines that do not rely on experts researchers.

Editorial board

An editorial board, represents different international/national partners and different polar nations, is suggested. The editorial board can highlight metadata providers from their countries that are not covered by the service to be included in our considerations. Moreover, the members of the editorial board will help to outreach the service in their countries. The international partners in the editorial board can bring more resources to the services (e.g. in term of providing technical help based on their experiences). The leading role of the UiT and NPI of the board will strengthen UiT's and NPI's position as a nationally leading knowledge center on polar research.

Clarify the need, solution and content framework of the service

The team of the OpenARI have mapped fifteen needs that are required in order to launch the new planned service. Some of these needs already exist in the current version of the High North

Research Documents, however we will mention these needs here in order to be kept in mind while the UB's team is setting up the new service.

1- Maximize the numbers of relevant records:

   The current database of the High North Research Documents has around 1,000,000 records all of them were collected after a filtration process of the records hosted by BASE. However, in the new planned service, we need to maximize the numbers of relevant records by getting records from the major metadata/data providers covering polar research either through the BASE or by direct contact to these providers.

   In order to increase the records in our database, but keeping the costs at the minimum level, we suggest that a list of the potential metadata providers should be delivered to BASE in order to get their data harvested by BASE and then by our new planned service. However if BASE does not succeed to harvest these new providers, we will need to start a negotiation process with these metadata providers. Some metadata providers are willing to send us their records, but do not use OAI-PMH protocol nor allow extracting their metadata through API. As a result, we suggest developing a code that will be able to collect their metadata directly by reading the HTML of their websites (see the suggestions for technical solutions section).

2- Better filtration process of the records; and
3- Reduce the time of the filtration process

   We suggest here to perform the filtration process with a slightly different way than that used for High North Research Documents in order to reduce the filtration time to the minimum and to maximize our benefits by not letting irrelevant records pass through the filters. The suggested filtration process will be presented in detail in a later section.

4- Give the end-users the option to perform his/her search on a map;
5- Showing the search results on a map; and
6- Add location value for each record (e.g. latitudes and longitudes)

   Searching by map and/or showing the search results on a map is an new suggested options to be added to the new planned-service which will allow the end-users choose the records based on the geographic distributions of these records. However these two options need to add new information to our database which is the location value. Most of the records come from BASE are missing the location value. We suggest to write a program code that will be able to go through the metadata of each record and getting out geographic words (e.g. name of location; Tromsø, Longyearbyen, Ny-Ålesund). Then the program will search an existing database (e.g. https://www.geonames.org/) that will provide us with the latitudes and longitudes data of these geographic words. The returned latitudes and longitudes data will be added and saved together with the metadata of each record.

7- A proper map projection:

   Points 4 and 5 need a map to show the result and/or to perform search. As the new planned service targets the Polar Regions, we need to use a proper map projection. We think it does not make sense when Africa or other equator areas are presented in the center of the map. We suggest to use two map projections one for the North Pole and another for the South Pole (e.g. Fig. 1).

8- Show the results as a timeline

As it is important to the researchers to see the progress that occurs within a scientific subject on an area, we suggest to give the end-users the ability to see the results as a timeline (from the oldest to the youngest).

9- Rank the research results on the user-interface according to the most relevant records;
10- Add advanced search options to the user-interface; and
11- Multi-languages user-interface

These options already exist in most of the recent user-interfaces. We just need to activate "rank records by relevance" and "Advanced Search" during the setup stage of the new service. For the multi-languages options, we need to translate the different buttons of the user-interface to those additional languages and save the translation in a configuration file of the user-interface. We suggest translating the user-interface to Norwegian, Sami, German, French, Russian in addition to the English. However, to work in the resources frame of the project, we suggest to start with English, Norwegian and Sami as a first stage. Other languages will be added at later stages as upgrade versions of the service.

12- Add to the existing records:

To reduce the time of the filtration process, some fields of the existing records (meaning the records already in the High North Research Documents or the records of the previous filtration process) will not be modified during any future filtration process. These fields are for example: 1) the title of the record; 2) original ID; 3) the geographic information (see point 6 of the needs). So searching for geographic information will be performed only for the newly added records.

13- Use metaphone function (or) autocomplete search words algorithm

This option will help the end-users to perform search even if they misspelt the search word. Metaphone function needs creating an extra database, while the autocomplete search words is an algorithm that comes with most of recent user-interfaces. As a result, we suggest choosing the autocomplete search words algorithm instead of using the metaphone function.

14- An easy to use user-interface

Most of our needs are linked to the user-interface. So it is our target to look for an easy to use user-interface that will come with the some (or all) of our needs. Although we need to increase the functionality of the user-interface, we would like to keep it as simple as we can (e.g. all the functions in one interface with a nice structure). Moreover, we are looking for an open source user-interface that is supported by a large community of developers and programmers in order to keep our user-interface up-to-date at the minimum cost level. We have mapped 10 user-interfaces (Appendix D) to choose one of them. OpenARI team has chosen Vufind and DSpace user-interfaces to be tested. We suggest to use the Vufind since it gives us most of our needs and is supported by a large community of developers.

15- Give the end-users the ability to comment and share different records:

This will increase the visibility of the new planned service by being cited and shared through different social media channels.

Suggestions for technical solutions

To achieve our main goal and the sub-goals of the project, a modification on the programing flow of the High North Research Documents is required. In general, the technical solutions can be divided into two groups: a) technical solutions related to the back-end stage and b) technical solutions related to the front-end stage. The back-end stage is mainly linked to how we will perform the filtration process. Filling gaps (e.g. geographic information) in the metadata of each record is also part of the back-end stage. The front-end stage includes all the processes and the functions of the user-interface. Technical solutions related to the front-end stage will be presented in the next section "Prototype".

Back-end stage – collecting the metadata from different sources and metadata providers:

Three ways are suggested to collect the metadata. These are: 1) metadata collected from metadata providers who allow using OAI-PMH service. As we explained early that the BASE will be our main metadata provider but maybe we need to consider making the harvesting process by the team of the new planned service at the UiT. 2) Metadata collected from providers who are willing to collaborate but they do not offer their data through the OAI-PMH service. 3) Metadata collected direct from the end-users through a form on the front-end user interface.

The first way of collecting metadata will result in XML files that contain metadata from different providers. It is worth mentioning here that different providers may have different file structures. A computer program is needed to harvest metadata from providers who are willing to collaborate but do not offer OAI-PMH (i.e. way 2 of metadata collections). The OpenARI team has the competences required to write the code of this automatic harvesting program. A temporary database will be created in order to save the metadata collected from way 1 and 2. Different fields of the temporary database will be linked to a form on the user-interface in order to collect metadata from the end-users (i.e. way 3) which should be approved from the admin of the new service before saving it in the temporary database.

The next step of the back-end stage is to filter the metadata of the temporary database.

Back-end stage – filtration process:

It is our ambition to get most of the relevant records on polar sciences through a process including four stages of filtration. The flow of these four stages aims to reduce the number of the records that transferred from a stage to the next (i.e. minimize the number of the records that being filtered). Reducing the number of the records will definitely reduce the time required for the filtration and therefore reduce the overhead costs of using UiT's server. The four stages filtration process includes:

First stage: filtered all records from most relevant journals and source institutes (pre-defined list e.g. journal of Polar Research).

Second stage: filtered out the broken records (e.g. records missing links, text, irrelevant....).

Third stage: compare the records in the temporary database with those that we have in the final database (from the previous filtration process). Removing the duplications and keep only the most updated version.

Fourth stage: check for certain keywords.

Appendix E shows a detail programing flow of the all processes (i.e. all the back-end stage) which includes a four stage filtration process. We suggest to do the filtration process using both MySQL and Solr technology. The reasons of using combination of MySQL database (to store the metadata) and the Solr technology (to filter these metadata) are:

- We cannot rely on Solr/Lucene as a secure database. Solr/Lucene databases are: missing recovery options, lack of acid transactions, possible complications etc.
- Solr is perfect for storing documents that include data from several tables and relations that would otherwise require complex joins to be constructed.
- Solr/Lucene provides mind blowing text-analysis / stemming / full text search scoring / fuzziness functions.
- MySQL is very trivial and limited. Weighting fields, boosting documents on certain metrics, score results based on phrase proximity, matching accuracy, etc are very hard to work or almost impossible.

All the records that will successfully pass the filters during the filtration process will be stored in a semi-final database waiting to fill gaps and a final stage of manual checking.

Back-end stage – filling gaps:

The most important gap that we need to take care is the geographic information. Most of the records that come from different metadata providers are missing the geographic information as latitudes and longitudes, but they contain other geographic information (e.g. name of city or location).

It is our aim to fill this gap by running a program code searching the metadata for certain geographic words (i.e. name of city, location, .. etc.). These geographic words will be compared to a pre-existing database (e.g. https://www.geonames.org) that contains location information (i.e. latitudes and longitudes). The returned geographic information will be stored as metadata in each record in order to give the end-users access to search and see the results on map. Records with no clear geographic information need to go through a manual check.

At this stage, the database becomes ready to be used by the end-users through the use-interface.

Prototype

Two (i.e. Vufind and DSpace) out of 10 user-interfaces (Appendix D) were tested in order to see which of them can provide the end-users the maximum performance. The eight un-tested user-interfaces were tried using existing websites. We were looking for a user-interface with the following features: 1) open sources, 2) supported by a large community of developers, 3) import mechanisms that makes it possible to populate the system in an automatic and efficient way, 4) the possibility to make search using a map interface, 5) the ability to show the results on a map, 6) has nice look, 7) lets the end-users interact with social media, 8) lets the end-users

add comments to the records, 9) has advanced search option, 10) narrows the results, 11) ability to give the user the full citation of the records, 12) ability to save the results for later reviewing by the end-user, 13) lets the end-users add tags to the records; and 14) has multi-language user-interface.

Example of the DSpace can be found at https://highnorth.uit.no/. The Vufind prototype can be found at http://oliver2.ub.uit.no/vufind/. The map search of the Vufind is at: click here, while the advanced search can be found here: http://oliver2.ub.uit.no/vufind/Search/Advanced?edit=106.

As a result of our assessment, the team of OpenARI is strongly suggesting the use of Vufind, since it meets all of features that we are looking for. However, slight modifications are needed in order to include the suggested subservices to the main user-interface. As an open source user-interface with a large community of developers, it will be easy to incorporate such modifications without using any additional resources.

Outreach plan

Although we are planning for the most comprehensive service that will provide the end-users with access to the openly published research data and documents on the Polar Regions, an outreach plan is required in order to expand our audiences. We have already started to distribute the idea by attending conferences on open data (e.g. Appendix F). We are planning to use the following strategy to attract more users to the service:

1- Make our service available to be harvested by common search engines.
2- Use our connections to our partners (national and international) through the editorial board to distribute the service in their countries.
3- Provide our partners and the data providers with a quick search box on their websites that will drive the users to search our database.
4- Provide our partners and the data providers with our logo to be highlighted on their websites.
5- Use and activate different social media channels.
6- Attending conferences and regional meetings on polar and environmental sciences, and global warming (e.g. SCAR Meetings - The Scientific Committee on Antarctic Research; meeting of the International Arctic Science Committee, Arctic Frontiers, Polar Libraries Colloquy).
7- Use the network of UiT with other educational and research units (e.g. University of the Arctic) to broadcast the service.
8- Preparing posters and flyers on the service and distribute it widely (digitally) to research institutes/unites dealing with the polar sciences.

RESOURCES AND TIME FRAME

One of our targets during the pilot project is to use open source programs and reduce the costs of the project as possible. All the programs, code, databases and user-interface that will be used in the service will be free of charge. However, a slight fund is required for the manpower. Table (1) shows the current members of the pilot project. We suggest that the same or a similar team will be needed for the full-scale model. In addition, we suggest that two more members (i.e. a member from the IT department of UiT and a PhD student with 25% of his/her tasks toward the

project) will be necessary to implement the service. As a collaboration project between UiT and NPI, we suggest the PhD student be based on NPI.

Table 1: Participants of the pilot project.

Name	Role	Unit
Tamer Abu-Alam	Project manager	UiT-UB
Faggruppeleder Per Pippin Aspaas	Professionally responsible	UiT-UB
Seniorrådgiver Stein Høydalsvik	Project member	UiT-UB
Universitetsbibliotekar Leif Longva	Project member	UiT-UB
Senioringeniør Karl Magnus Nilsen	Project member	UiT-UB
Senioringeniør Obiajulu Odu	Project member	UiT-UB
Seksjonsleder Roy Dragseth	Professionally responsible	UiT-ITA
Seksjonsleder miljødata Stein Tronstad	Project member	NPI
Senioringeniør Conrad Helgeland	Project member	NPI

We estimate three years as a period for the full-scale management project (Table 2). These three years can be subdivided into two phases; phase A) developing the service (i.e. a two years phase) and phase B) initial normal management phase of the service using the internal resources of the UB (i.e. one year phase).

Table 2: Suggested timetable and targets of the project.

Project phases		Targets	Year
A	i.	Establish technical solution for the service.	Year 1
	ii.	Get metadata from sources that support automatic harvesting..	
	iii.	Standardize and enrich metadata.	Year 2
	iv.	Integration with non-standard data sources.	
	v.	Establish collaboration networks.	
	vi.	Present and promote the service.	
B	vii.	Establish a management model using the internal resources of the UB.	Year 3

ETHICAL CONSIDERATION

The proposed project does not involve any experiments with animals or risk of injury to people, however, part of the project (i.e. hosting original research data) may involve data that contains personal information. The team of the project will design the database to make sure that the personal information will be treated according to the guidelines for research ethics in science and technology [6] provided by the Norwegian National Research Ethics Committees. Ethics guidelines of dealing with personal information will be presented clearly to the authors (owners) of the data and they must accept the guidelines prior to submitting their data. Moreover, such data will not be available to the public unless it reviewed and approved by the team of the project.

RECOMMENDATIONS

Based on the presented facts that 60% of open-access records on Polar Regions are unfindable through one common search engine, the responsibility of Norway (e.g. UiT and NPI) to undertake, supervise and disseminate research in the Polar Regions, and the existing competences at UB, UiT, the OpenARI strongly recommends launching a new service at the UB in order to make the open-access records on the Polar Regions more visible and easy to be retrieved from one search platform. The project will strengthen UiT's and NP's position as an international leading knowledge center on polar research. We suggest implementing the full-scale management model during a three years project which will include two phases; phase A) developing the service and phase B) initial normal management phase of the service using the internal resources of the UB.

We suggest to support the main service (i.e. make the open-access records on the Polar Regions more visible and findable) of the project by adding additional three subservices: 1) hosting original data from the Polar Regions; 2) creating a research platform; 3) creating an education platform. We have identified fifteen needs that are required to build a unique service. We have mapped 115 major and relevant metadata providers that potentially can support and feed the service with metadata. By filling this 60% findability gap, the University of Tromsø – the Arctic University of Norway will support researchers and students by providing them a quick access to polar records through one single search engine.

ACKNOWLEDGMENTS

The OpenARI pilot project has a team of eight professionals who I would like to acknowledge for their support and contributions; Aspaas, Per Pippin; Dragseth, Roy; Høydalsvik, Stein; Helgeland, Conrad; Longva, Leif; Nilsen, Karl Magnus; Odu, Obiajulu; Tronstad, Stein (Alphabetically according to the family name).

REFERENCES

[1] European Research in the Polar Regions: Relevance, strategic context and setting future directions in the European Research Area. Edited by the ESF European Polar Board (2011).

[2] Norwegian Polar Research: An Evaluation. Prepared and edited by the Division for Energy, Resources and the Environment, Norwegian Research Council (2017).

[3] Strategic plan for UiT - The Arctic University of Norway 2014-2022. Retrieved from https://en.uit.no/om/art?p_document_id=377752&dim=179033

[4] Norsk Polarinstitutt, strategi 2019 – 2024. Retrieved from http://www.npolar.no/npcms/export/sites/np/files/vedlegg/strategi.pdf

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

[6] Guidelines for Research Ethics in Science and Technology, the National Committee for Research Ethics in Science and Technology. ISBN: 978-82-7682-075-1 (2016).

APPENDIXES

Appendix A

Shows the mapped 115 metadata providers with their websites and number of records.

ID	Country	Institute/Data repository	Included in BASE	Included in Other Data repositories	Website	Number of datasets/records
1	USA	UCAR NCAR Research Data Archive (UCAR NCAR RDA)	Not included	Arctic Data Explorer	https://rda.ucar.edu/	155
2	USA	NOAA National Oceanographic Data Center (NOAA NODC)	Not included	Arctic Data Explorer	https://www.nodc.noaa.gov/	6355
3	USA	tDAR: The Digital Archaeological Record (tDAR)	Not included	Arctic Data Explorer, DataOne	https://www.tdar.org/about/	86
4	USA	Arctic Data Explorer	Not included		http://arctic-data-explorer.labs.nsidc.org/	46680
5	USA	ArcticDEM	Not included		https://www.pgc.umn.edu/data/arcticdem/	2300
6	USA	Data Observation Network for Earth (Dataone)	Not included		https://www.dataone.org	802632
7	USA	NASA - Antarctic Master Directory (GCMD/AMD)	Not included		https://gcmd.gsfc.nasa.gov/KeywordSearch/Home.do?Portal=amd&MetadataType=0	73129
8	USA	Global Change Master Directory	Not included		https://gcmd.nasa.gov/learn/staff.html	33075
9	USA	Arctic Observing Viewer	Not included		https://www.arcticobservingviewer.org/	15000
10	USA	International Arctic Research Center	Not included	DataOne	http://climate.iarc.uaf.edu/geonetwork/srv/en/main.home	267
11	USA	National Ecological Observatory Network	Not included	DataOne	https://www.neonscience.org/data	177
12	USA	MGDS Antarctic	Not included		http://www.marine-geo.org/index.php	8685
13	USA	SOOS (GCMD)	Not included		http://www.soos.aq/	3176
14	USA	EarthChem	Not included		https://www.earthchem.org	458
15	USA	NCEI - World Data System Paleoclimatology Data	Not included		https://www.ncdc.noaa.gov/data-access/paleoclimatology-data	10187
16	USA	SESAR	Not included		http://www.geosamples.org	714300
17	USA	Barrow Area Information Database	Not included		http://barrowmapped.org/	18200
18	USA	North Slope Science Initiative	Not included		https://northslopescience.org/catalog#catalog-news	3090
19	USA	Arctic Landscape Conservation Cooperative	Not included		http://arcticlcc.org/products/	182
20	USA	Earth Observation Laboratory UCAR/NCAR	Not included		https://www.eol.ucar.edu/	2004
21	USA	NASA ABoVE	Not included		https://above.nasa.gov/	Not available

ID	Country	Institute/Data repository	Included in BASE	Included in Other Data repositories	Website	Number of datasets/records
22	Canada	ArcticConnect	Not included		https://arcticconnect.org/	Not available
23	Canada	Inuit Qaujisarvingat	Not included		http://www.inuitknowledge.ca	154
24	Canada	NWT Discovery Portal	Not included		http://nwtdiscoveryportal.enr.gov.nt.ca/geoportal/catalog/search/search.page	2983
25	Canada	Open Data (Canada)	Not included		https://open.canada.ca/data/en/dataset?portal_type=dataset	81,146
26	Canada	Polar Data Catalogue (PDC)	Not included	Arctic Data Explorer	https://www.polardata.ca/	2637
27	Canada	University of Alberta - datasets	Not included		https://www.library.ualberta.ca/	1541
28	Canada	Centre Etude Nordique	Not included		http://ccadi.ca	Not available
29	Canada	Dept. Fisheries and Oceans Canada	Not included		http://www.dfo-mpo.gc.ca/	Not available
30	EU	EUDAT	Not included		https://eudat.eu/	693,563
31	EU	SeaDataNet	Not included		https://www.seadatanet.org/Data-Access	2262108
32	EU	ECOMET	Not included		http://www.ecomet.eu/	Not available
33	Norway	Arctic Data Centre	Not included		https://pm.met.no/arctic-data-centre	440
34	Norway	Svalbard Integrated Arctic Earth Observing System	Not included		https://sios-svalbard.org/metadata_search	531
35	Norway	Institute of Marine Research	Not included	Arctic Data Centre	https://www.imr.no/en/hi	Not available
36	Norway	Norwegian Satellite Earth Observation Database for Marine and Polar Research	Not included	Arctic Data Centre	https://normap.nersc.no/	Not available
37	Norway	INTAROS	Not included		https://www.nersc.no/project/intaros	Not available
38	Norway	NORSAR	Not included		https://www.norsar.no	Not available
39	Norway	NorDataNet	Not included		https://www.nordatanet.no/	457
40	Norway	Norwegian Marine Data Centre (NMDC)	Not included	Arctic Data Centre	https://nmdc.no/nmdc/datasets	Not available
41	Finland	Finnish archives	Not included		https://www.finna.fi/	**13,103**
42	Finland	ARCTIC SPACE CENTRE	Not included		http://fmiarc.fmi.fi/	Not available
43	Sweden	Environment Climate Data Sweden	Not included		https://ecds.se/dataset	1061
44	switzerland	EnviDat	Not included		https://www.envidat.ch/ui/#/	138
45	Switzerland	Group on Earth Observations System of Systems	Not included		https://www.earthobservations.org/geoss.php	Not available
46	Germany	Global Terrestrial Network for Permafrost (GTN-P)	Not included	Arctic Data Explorer	http://gtnpdatabase.org/	1389

ID	Country	Institute/Data repository	Included in BASE	Included in Other Data repositories	Website	Number of datasets/records
47	Belgium	biodiversity	Not included		http://www.biodiversity.aq/	1,558,071
48	Netherlands	Netherlands Polar Data Center	Not included		https://npdc.nl/	46
49	Italy	Italian Arctic Data Center	Not included		http://mainnode.src.cnr.it/cnr/index.php	25
50	UK	Archives Hub	Not included		https://archiveshub.jisc.ac.uk/search/	1411197
51	Russia	Russian Federation Service's Arctic and Antarctic Research Institute	Not included		http://www.aari.ru/	Not available
52	China	Chinese National Arctic and Antarctic Data Centre	Not included		http://www.chinare.org.cn/en/index/	1250
53	International	Arctic Research Mapping Application	Not included		http://armap.org/	3000
54	International	Global Cryosphere Watch	Not included		https://globalcryospherewatch.org/	1120
55	International	Joint WMO/IOC Technical Commission for Oceanography and Marine Meteorology	Not included		https://www.jcomm.info/	365
56	International	Ocean Data and Information System	Not included		https://www.iode.org/index.php?option=com_content&view=featured&Itemid=89	Not available
57	International	GEOTRACES	Not included		http://www.geotraces.org/	Not available
58	International	Arctic Spatial Data Infrastructure	Not included		https://arctic-sdi.org/	Not available
59	USA	NSF Arctic Data Center (NSF ADC)	Included as publisher	Arctic Data Explorer, DataOne	https://arcticdata.io/	11771
60	USA	National Snow and Ice Data Center (NSIDC)	Included as publisher	Arctic Data Explorer	https://nsidc.org/	1116
61	USA	NASA Earth Observing System (EOS) Clearing House (ECHO) (NASA ECHO)	Included as publisher	Arctic Data Explorer	https://earthdata.nasa.gov/about/science-system-description/eosdis-components/common-metadata-repository	14040
62	USA	U.S. Geological Survey ScienceBase (USGS ScienceBase)	Included as publisher	Arctic Data Explorer	https://www.sciencebase.gov/catalog/	499
63	USA	Biological and Chemical Oceanography Data Management Office (BCO-DMO)	Included as publisher	Arctic Data Explorer	https://www.bco-dmo.org/	1571
64	USA	Rolling Deck to Repository (R2R)	Included as publisher	Arctic Data Explorer, DataOne	http://www.rvdata.us/about/technical	944
65	USA	NOAA's National Centers for Environmental Information, World Data Service for Paleoclimatology (NOAA WDS Paleo)	Included as publisher	Arctic Data Explorer	https://www.ncdc.noaa.gov/data-access/paleoclimatology-data	3703

ID	Country	Institute/Data repository	Included in BASE	Included in Other Data repositories	Website	Number of datasets/records
66	USA	Gulf of Alaska Data Portal	Included as publisher		https://goa.nceas.ucsb.edu/#data	118
67	USA	Alaska Satellite Facility ASF	Included as publisher		https://www.asf.alaska.edu	250
68	USA	NCEI - National Oceanographic Data Center	Included as publisher		https://www.nodc.noaa.gov/	30448
69	USA	Environmental Data Initiative	As publisher (66812 records)	Datacite, DataOne	https://environmentaldatainitiative.org/data/	42882
70	USA	CLIVAR and Carbon Hydrographic Data Office	Included as publisher		https://cchdo.ucsd.edu/	2151
71	USA	ICPSR	Included as publisher		https://www.icpsr.umich.edu/icpsrweb/ICPSR/index.jsp	86719
72	Canada	Scholars Portal Dataverse U. of Toronto	Included as publisher		https://scholarsportal.info/	82,484
73	Canada	Arctic Institute of North America	Included as publisher		https://arctic.ucalgary.ca/about-astis	82,000
74	Canada	Yukon Research Centre	Included as publisher		https://www.yukoncollege.yk.ca/search/node	Not available
75	Norway	Norwegian Meteorological Institute (Met.no)	Included as publisher	Arctic Data Explorer, Arctic Data Centre	https://www.met.no/	200
76	Norway	The University Centre in Svalbard	Included as publisher		https://www.unis.no/library/	Not available
77	Norway	DataverseNO-arkivet	Included as publisher (1,131 records)	DataCite Metadata Store	https://dataverse.no/	1500
78	Denmark	International Council for the Exploration of the Sea (ICES)	Included as publisher	Arctic Data Explorer	http://ecosystemdata.ices.dk/	10123
79	Denmark	The Geological Survey of Denmark and Greenland	Included as publisher		http://www.eng.geus.dk/	Not available
80	Sweden	Swedish National Data Service	Included as publisher	DataCite Metadata Store	https://snd.gu.se/en/catalogue/search	3131
81	UK	UK Polar Data Centre	Included as publisher		https://data.bas.ac.uk/	Not available
82	UK	British Oceanographic Data Centre	Included as publisher		https://www.bodc.ac.uk/	Not available
83	Australia	Australian Ocean Data Network	As publisher (11861 records)		https://catalogue.aodn.org.au/geonetwork/srv/eng/main.home	12212
84	Australia	Australian Antarctic Data Centre - AADC	Included as publisher		https://data.aad.gov.au/	Not available
85	Japan	Arctic Data archive System	Included as publisher		https://ads.nipr.ac.jp/	Not available
86	Iceland	Conservation of Arctic Flora and Fauna	Included as publisher		https://www.abds.is/	Not available

ID	Country	Institute/Data repository	Included in BASE	Included in Other Data repositories	Website	Number of datasets/records
87	China	Polar Research Institute of China	Included as publisher		http://www.pric.org.cn/	Not available
88	Korea	Korea Polar Research Institute	Included as publisher		https://kpdc.kopri.re.kr/	Not available
89	International	OBIS - Ocean Biogeographic Information System	Included as publisher		http://www.iobis.org	2533
90	International	KNB Data Repository	Included as publisher	DataOne	https://knb.ecoinformatics.org/	27,358
91	International	Global Biodiversity Information Facility	As publisher (11000 records)	DataOne	https://www.gbif.org/	3,511,491
92	International	Scientific Committee on Antarctic Research	Included as publisher		https://www.scar.org/data-products/data/	Not available
93	USA	Dryad	Included in BASE		https://datadryad.org/	132468
94	Canada	University of Alberta - publication	Included in BASE		https://www.library.ualberta.ca/	83500
95	Norway	Norwegian Polar Institute	Included in BASE	PANGAEA	http://www.npolar.no/en/	339
96	Norway	NILU – Norwegian Institute for Air Research	Included in BASE	Arctic Data Centre	https://www.nilu.no	571
97	Norway	Universitetet i Tromsø: Munin Open Research Archive	Included in BASE		https://munin.uit.no/	13251
98	Norway	Universitetet i Tromsø: Septentrio Academic Publishing	Included in BASE			3.799
99	Norway	Universitet i Oslo: Digitale utgivelser ved UiO (DUO)	Included in BASE		https://www.duo.uio.no/	52096
100	Norway	University of Bergen: Bergen Open Research Archive (BORA-UiB)	Included in BASE		http://bora.uib.no/	2711
101	Norway	NTNU Samfunnsforskning (Norwegian University of Science and Technology / Norges teknisk-naturvitenskapelige universitet)	Included in BASE		https://brage.bibsys.no/xmlui/handle/11250/299124	276
102	Norway	Norges teknisk-naturvitenskapelige universitet: NTNU Open / Norwegian University of Science and Technology	Included in BASE		https://brage.bibsys.no/xmlui/handle/11250/223328	28790
103	Norway	Norges teknisk-naturvitenskapelige universitet, Trondheim: NTNU Open Access Journals	Included in BASE		https://www.ntnu.no/ojs/	1865
104	Germany	DataCite Metadata Store	Included in BASE		https://search.datacite.org	13,394,962
105	Germany	PANGAEA	Included in BASE		https://www.pangaea.de/	69000

ID	Country	Institute/Data repository	Included in BASE	Included in Other Data repositories	Website	Number of datasets/records
106	Germany	The Arctic-HYDRA Program	Included in BASE	Arctic Portal	https://arctichydra.arcticportal.org/	Not available
107	Austria	ESA DUE Permafrost	Included in BASE	PANGAEA	http://geo.tuwien.ac.at/permafrost/	10
108	EU	zenodo	Included in BASE		https://zenodo.org/	460401
109	EU	PAGE21	Included in BASE	PANGAEA	https://www.page21.eu/	52
110	EU	European Fluxes Database Cluster	Included in BASE	PANGAEA	http://www.europe-fluxdata.eu/	167
111	Australia	Research Data Australia	Included in BASE		https://researchdata.ands.org.au/	170,356
112	Australia	Marlin (CSIRO)	Included in BASE		http://marlin.csiro.au/geonetwork/srv/eng/search	3864
113	Iceland	Arctic Portal	Included in BASE		https://arcticportal.org/	1,008
114	Japan	National Institute of Polar Research	Included in BASE		https://www.nipr.ac.jp/english/database/	13717
115	Germany	Bielefeld Academic Search Engine (BASE)			https://www.base-search.net/	136,160,738

Appendix B

Metadata providers Survey

In orders to get a deep view of the causes of the 60% findability gap in the polar records, we have performed a short survey of 7 questions asking the providers of their knowledge of the OAI-PMH service and if they allow harvesting their metadata through this service or other. The following are the questions of the survey followed by the statistics of the answers that were collected.

Survey questions:

1- What is the name of your institution? and your database?

2- How will you describe the content in your database? Select all relevant choices
   2-1 Mainly open-access and/or open data
   2-2 Mainly metadata
   2-3 Both metadata records and open-access/open data
   2-4 Harvested or imported from other databases
   2-5 Others

3- How many records do you have in your database related to Polar, Arctic and Antarctic issues? (An estimate is OK, if you do not have the exact numbers).

4- Are your metadata harvested by common search engines?
   4-1 Yes
   4-2 No
   4-3 I do not know

5- Can your database be harvested via the OAI-PMH protocol?
   5-1 Yes
   5-2 No
   5-3 I do not know

6- If your answer of the previous question is "Yes", please add your OAI-PMH-url

7- Does your database has its own API that allows to extract metadata?
   7-1 Yes
   7-2 No
   7-3 I do not know

We have received 52 responses, which represent about 46.02 % of the mapped providers. The answers of the above 7 questions are:

What is the name of your institution? and your database?	Number of Polar records
NASA Common Metadata Repository	6,000
Polar Geospatial Center, ArcticDEM	270,000
Centre d'Ã©tudes nordiques (CEN)	40
Polar Data Catalogue	2,700
Columbia University. EarthChem Library (ECL) and EarthChem DB (ECDB)	11,000
U.S. National Center for Atmospheric Research/Earth Observing Laboratory	1,907
SeaDataNet	146,000
the NASA ORNL DAAC	300
World Data System for Paleoclimatology	500
Southern Ocean Observing System (SOOS) Portal on GCMD	4,500
Royal Belgian Institute for Natural Science and Antarctic Biodiversity Portal	430
Royal Belgian Institute for Natural Science and Antarctic Biodiversity Portal	3,000,000
Biological and Chemical Oceanography Data Management Office (BCO-DMO)	900
National Council of Research of Italy Italian Arctic Data Center	100
Royal Belgian Institute of Natural Sciences, Antarctic biodiversity portal	430
National Center for Ecological Analysis and Synthesis. NSF Arctic Data Center	5,400
SNAP Data Portal	81
IARC Data Archive	91
National Snow and Ice Data Center/CIRES/University of Colorado	1,000
University of Gothenburg. Swedish National Data Service	15
UNIS - The University Centre in Svalbard	2,308
Conservation of Arctic Flora and Fauna (CAFF); Arctic Biodiversity Data Service (ABDS)	400,000
GBIF	7,000,000
Universitetet i Tromsø¸: Munin Open Research Archive	13,251
Netherlands Polar Data Center	46
ICPSR	500
Swiss Federal Institute for Forest, Snow and Landscape Research - WSL; Environmental Data Portal EnviDat	3
Environmental Data Initiative Data Repository	1,800
International Council for the Exploration of the Sea (ICES)	10,000
NCEAS, the Knowledge Network for Biocomplexity (KNB)	27,386
Alaska Ocean Observing System Ocean Data Explorer	3,700
Alaska Satellite Facility, NASA's Common Metadata Repository	Millions
Swedish National Data Service	200
University Library UiT; Munin	1,000
British Antarctic Survey. Repository is UK Polar Data Centre	400
Columbia University, US Antarctic Program Data Center (USAP-DC)	500
APGC	144
University of Bergen, and BORA (https://bora.uib.no/)	200
Norwegian Marine Data Centre (hosted by Institue of Marine Research)	1,000
Dryad	400
Institution: Interdisciplinary Earth Data Alliance (IEDA) at the Lamont-Doherty Earth Observatory of Columbia University, Database: System for Earth Sample Registration (SESAR)	230,000

DataCite	100,946
Australian Antarctic Data Centre	2,797
Dongchan Joo, Korea Polar Reach Institute, Korea Polar Data Center	1,141
SIOS	1,000
GCW Data Portal	1,400
NORMAP	50
WMO Arctic Data Centre	2,500
NorDataNet	2,500
Archives Hub (Jisc)	2,000
Universitetet i Tromsø,: Septentrio Academic Publishing	3,799
UiT The Arctic University of Norway-DataverseNO	400

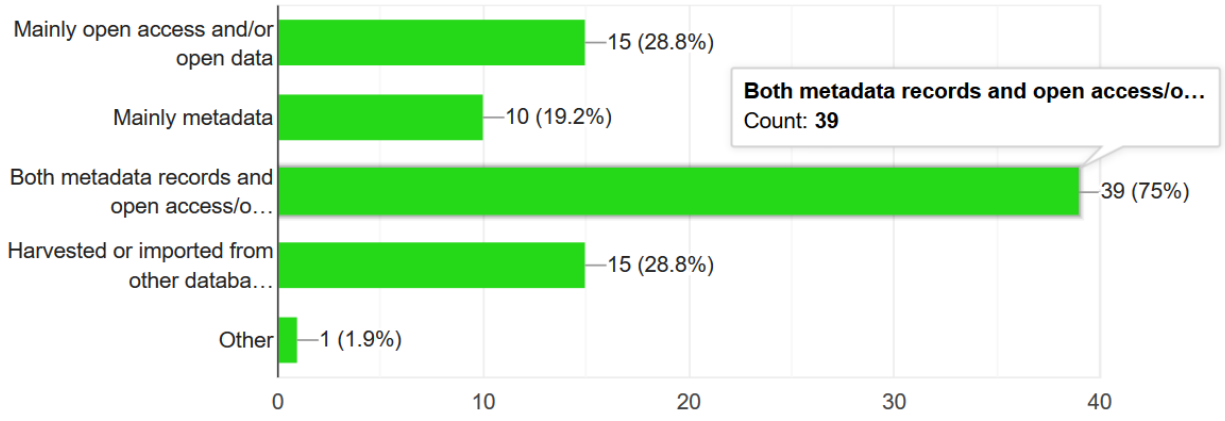


Fig. B1: a summary of the answers of the questions "How will you describe the content in your database?". 28.8% are mainly open-access and/or open data; 19.2% are mainly metadata; 75% are both metadata records and open-access/open data; 28.8% are harvested or imported from other databases and 1.9% are other.

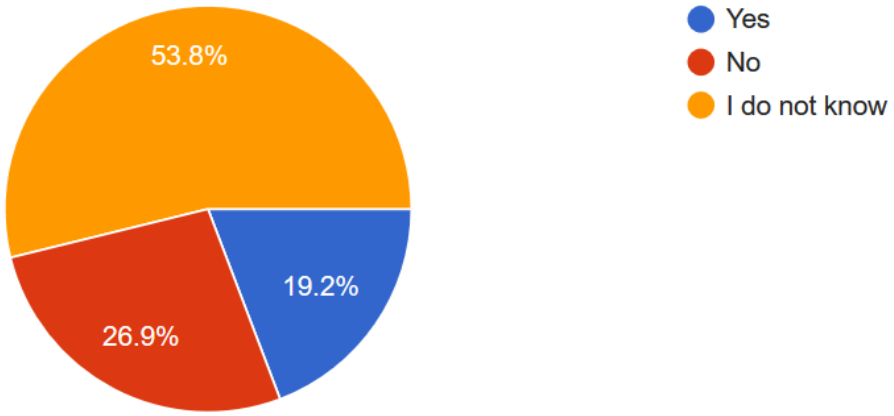


Fig. B2: a summary of the answers of the questions “Are your metadata harvested by common search engines?”.

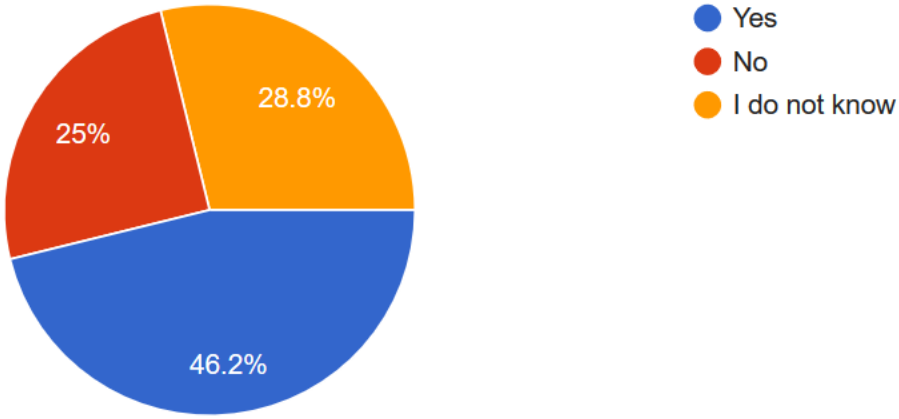


Fig. B3: a summary of the answers of the questions “Can your database be harvested via the OAI-PMH protocol?”. 46.2% of metadata providers allow harvesting their metadata using OAI-PMH protocol, the rest (i.e. 53.8%) do not use OAI-PMH protocol or not aware of such protocol.

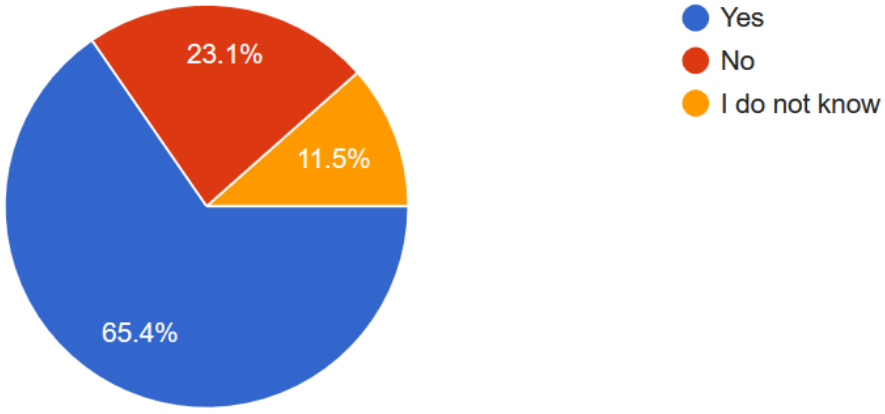


Fig. B4: a summary of the answers of the questions “Does your database has its own API that allows to extract metadata?”. 65.4% of metadata providers allow extracting their metadata using API. 34.6% do not use API.

Appendix C

Shows the metadata fields of the database. Note the metadata fields and structure are following the Dublin Core Metadata schema.

Elements	Dublin Core Elements	Suggested service	High North Research Documents
Database name		polardata	highnorth
Number of Tables		1	1
Table name		records	base_records
Record_ID		Record_ID	record_oai
Title	Title	Record_title	title
Creator	Creator	Creator	Creator
Authors	contributor	contributor	contributor
Publishing_date	Date	Publishing_date	date
Publishing_year		Publishing_year	year
Publisher	Publisher	Publisher_Journal	publisher
Originating_Center		Originating_Center	
Country of data generation		Country	
Volume/page_numbers		Volume_Page	
Identifier	Identifier	Identifier	Identifier
Description/Abstract	Description	Description_Abstract	description
Link to Graphical Abstract		Graph	
Temporal Coverage	Coverage	Temporal_Coverage/Start_Date	coverage
		Temporal_Coverage/End_Date (if the data is still collected - this can be "Continuous")	
Last_Revision_Date		Last_Revision_Date	
Spatial Coverage	Coverage	Spatial_Coverage/S_Latitude	coverage
		Spatial_Coverage/N_Latitude	
		Spatial_Coverage/W_Longitude	
		Spatial_Coverage/E_Longitude	
Source	Source	Source	source
Link to Source		Link_Sourse	
Language	Language	Language	Language
Language			Lang
Record_Format	Format	Record_Format	format
Data_Type	Type	Data_Type	type
Typenorm		Typenorm	typenorm
Relation (original papers)	Relation	Relation	relation
Citation (works cited the data)		Citation	
Keywords	Subject	Keywords	subject
Geo_Keywords		Geo_Keywords	
Copy Rights	Rights	Rights	rights
Datestamp		hdate	hdate
			collection

Appendix D

Shows different user interfaces that were studied during the pilot project.

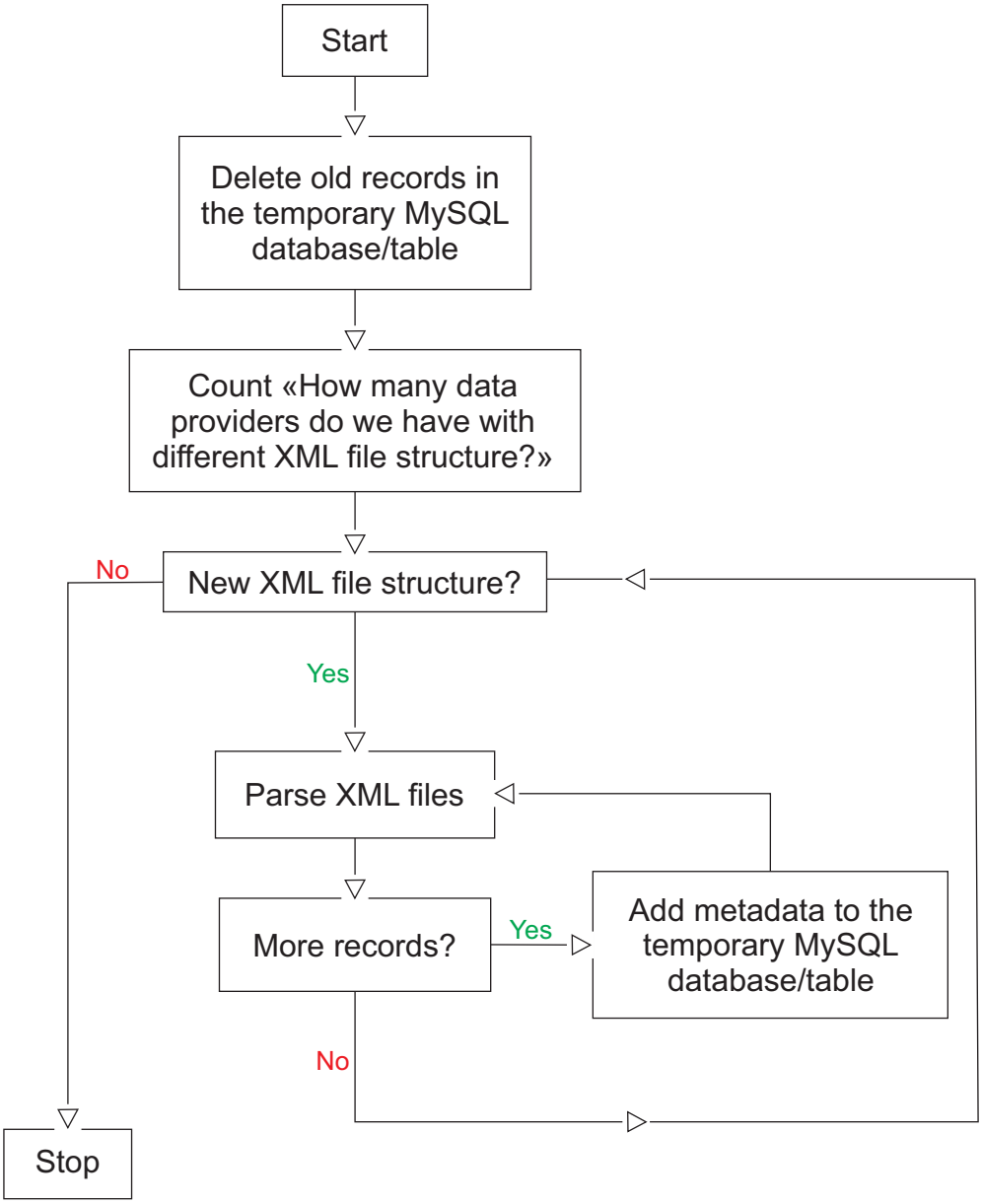
Feature	DSpace	Vufind	Koha	Evergreen ILS	Greenstone Digital Library	Blacklight	Invenio	Atoll Digital Library	BibLivre	user interface of Polar Data Catalogue
Open-access	Y	Y	Y	Y	Y	Y	Y			
Nice outlook	Y	Y	Y	N/Y	N/Y	Y	Y			
Let the user interact with social media	N (sharing only)	Y	Y			Y	Y			
Let the user adding comments to a record	N	Y	Y			Y				
The possibility to make search using a map interface	N	Y				Y				
Plotting the results on a map	N					Y				
Authors Biography	N	Y								
Advanced search	N/Y	Y	Y			Y				
Narrow the results	Y	Y	Y	Y	Y	Y	Y			
Ability to give the user the full citation of a record	N	Y					Y			
Ability to save the results for later reviewing by the user	N	Y	Y			Y				
Let the user adds tags to a record	N	Y								
Multi-Language user interface	Y	Y	Y			Y	Y			
End_user account	Y		Y	Y		Y	Y			
Examples	Link	Link			Link	Link	Link			
Software website	Link	Link	Link	Link	Link	Link	Link			Link
Recommentation by team members		Recommended						Not recommend	Not recommend	
Metadata import technology	XML (SAF), text file (BME), OAI-PMH, SWORD, ++	SolrMarc, XSLT/XML, OAI-PMH	MARC-file	MARC records	XML (eget format), OAI-PMH	Open (just get it into solr)	BibConvert/MARCXML, OAI-PMH			
Internal metadata format	Dublin Core	MARC	MARC	MARC	Dublin Core ++	MARC	MARC 21			
Can be harvested with OAI-PMH	Y	Y	Y	Y	Y	Y	Y			

Feature	DSpace	Vufind	Koha	Evergreen ILS	Greenstone Digital Library	Blacklight	Invenio	Atoll Digital Library	BibLivre	user interface of Polar Data Catalogue
Technology/platform/language	Java, XSL/Angular(JS), postgres	PHP, MySQL, Java	XHTML/CSS/JS, REST, MySQL	Perl, XHTML/JS, PostgreSQL	Java/tomcat	Ruby on Rails, JSON API,	Flask WDF (Python/JS), REST, JSON			
Support for full text records	Y	? (full text indexing from external URLs)			Y		Y			

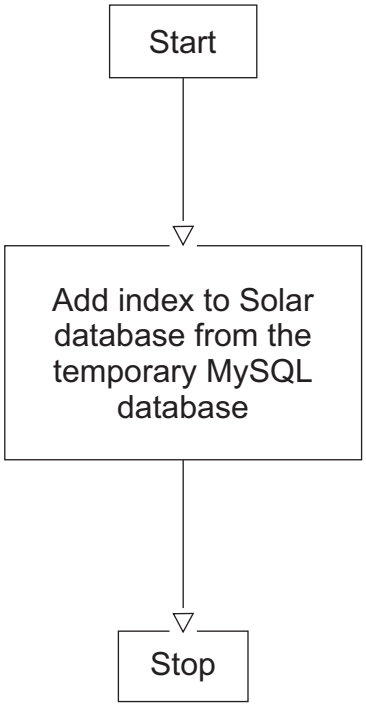
Appendix E

Shows a detail programing flow of the all processes (i.e. all the back-end stage) which includes a four stage filtration process.

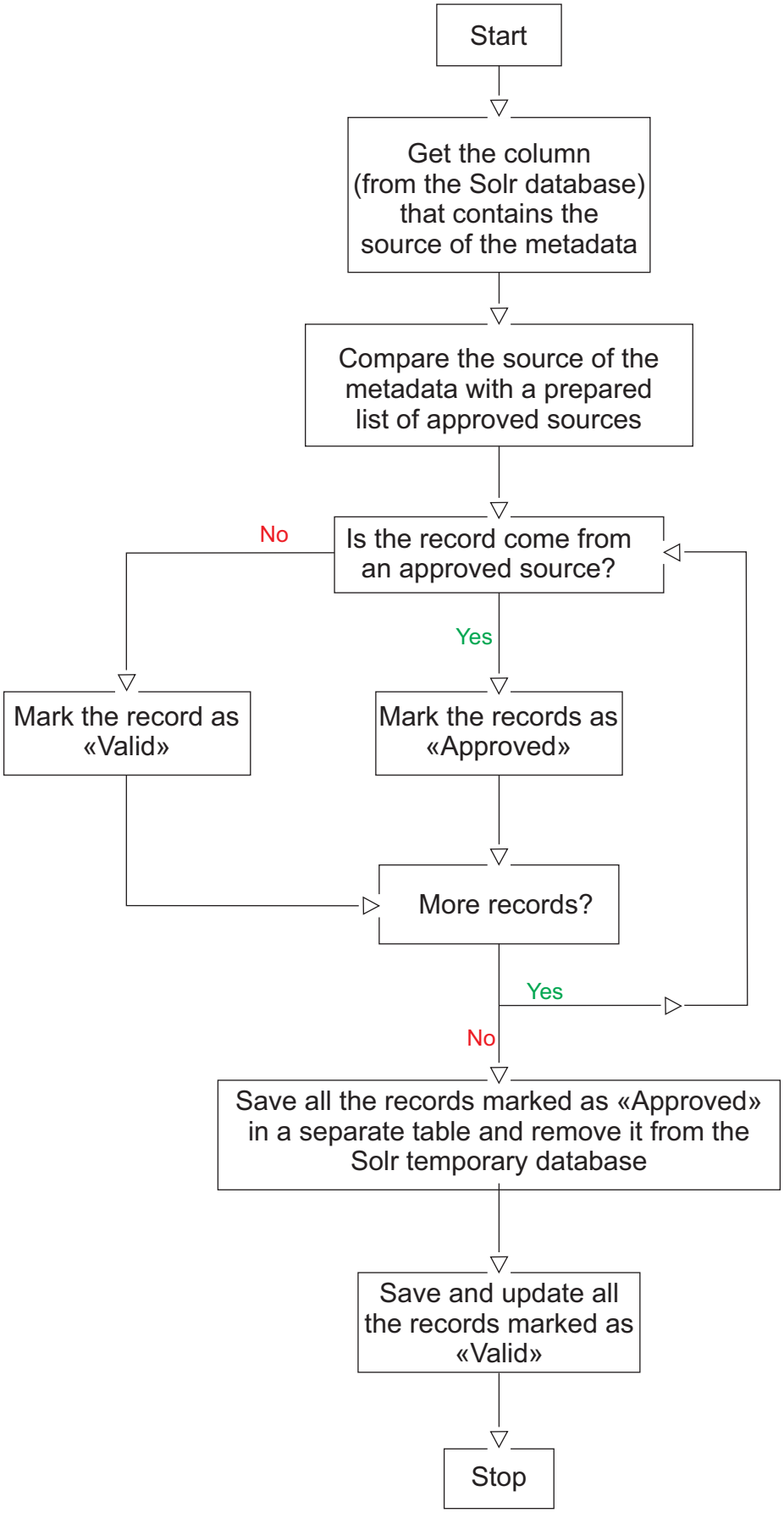
1. **Populate a temporary database (MySQL) with the metadata from the different data providers.**



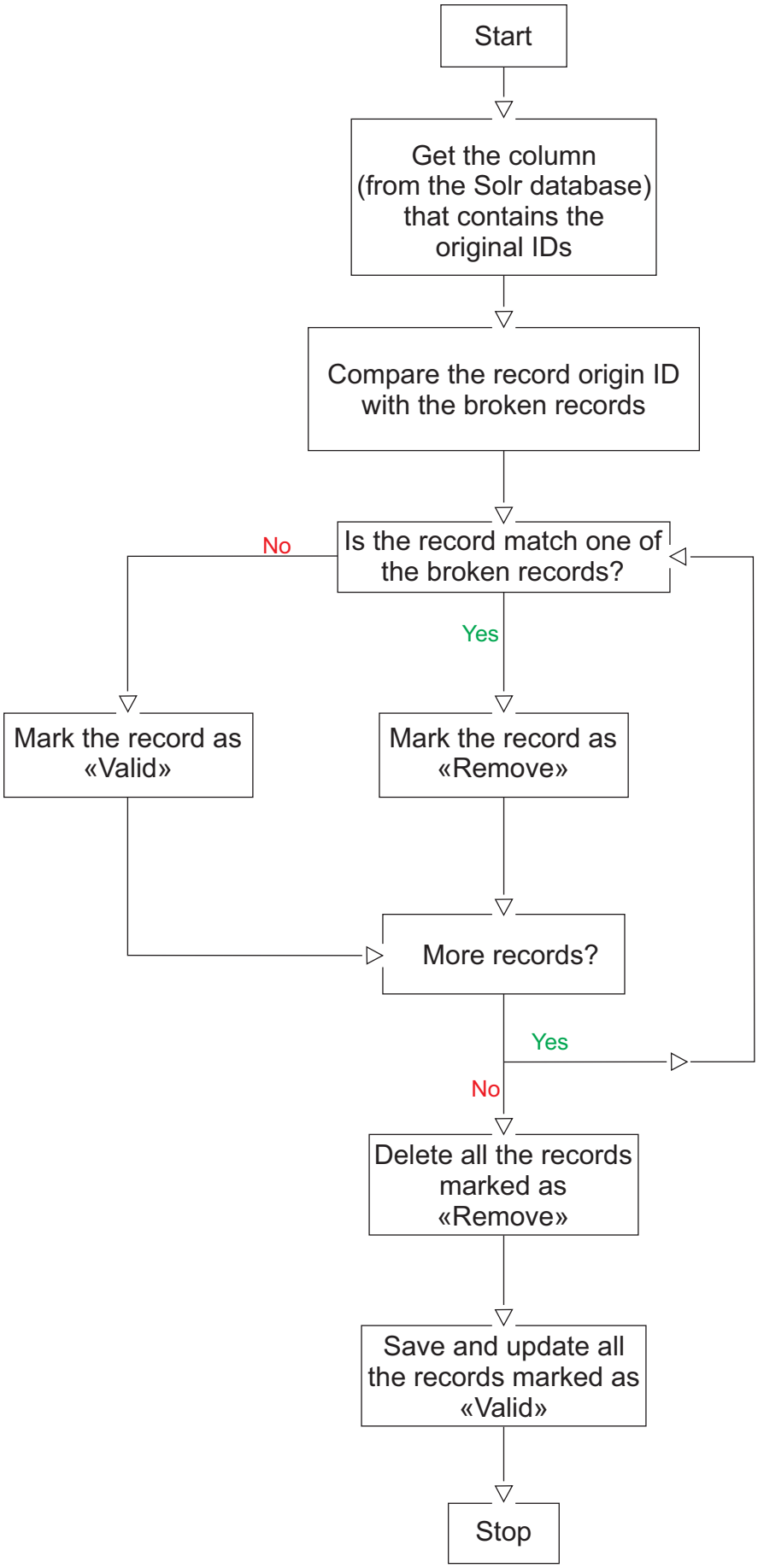
2. Index the temporary MySQL database/table into a Solr search engine (Solr database).



3. First filtration stage: Check the sourses of the metadata. If the source of the metadata is an approved sourse (e.g. Polar Data Catalogue, NPI, or any other polar institute), then extract the record and save it to a separate table.

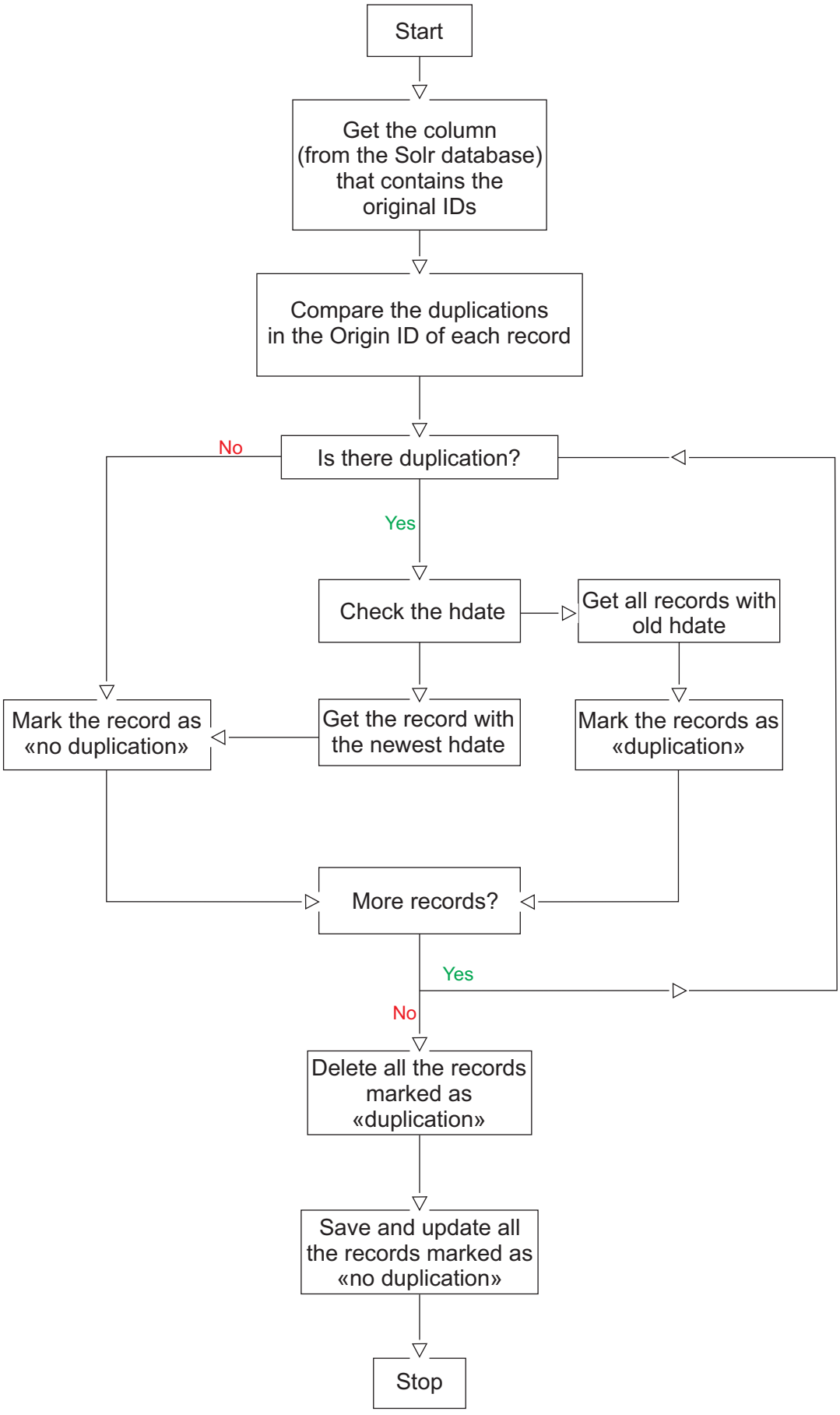


4. Second filtration stage: Check the records and compare it to the broken records (e.g. records missing links, text, unreleavent, ...). Then remove the broken records.

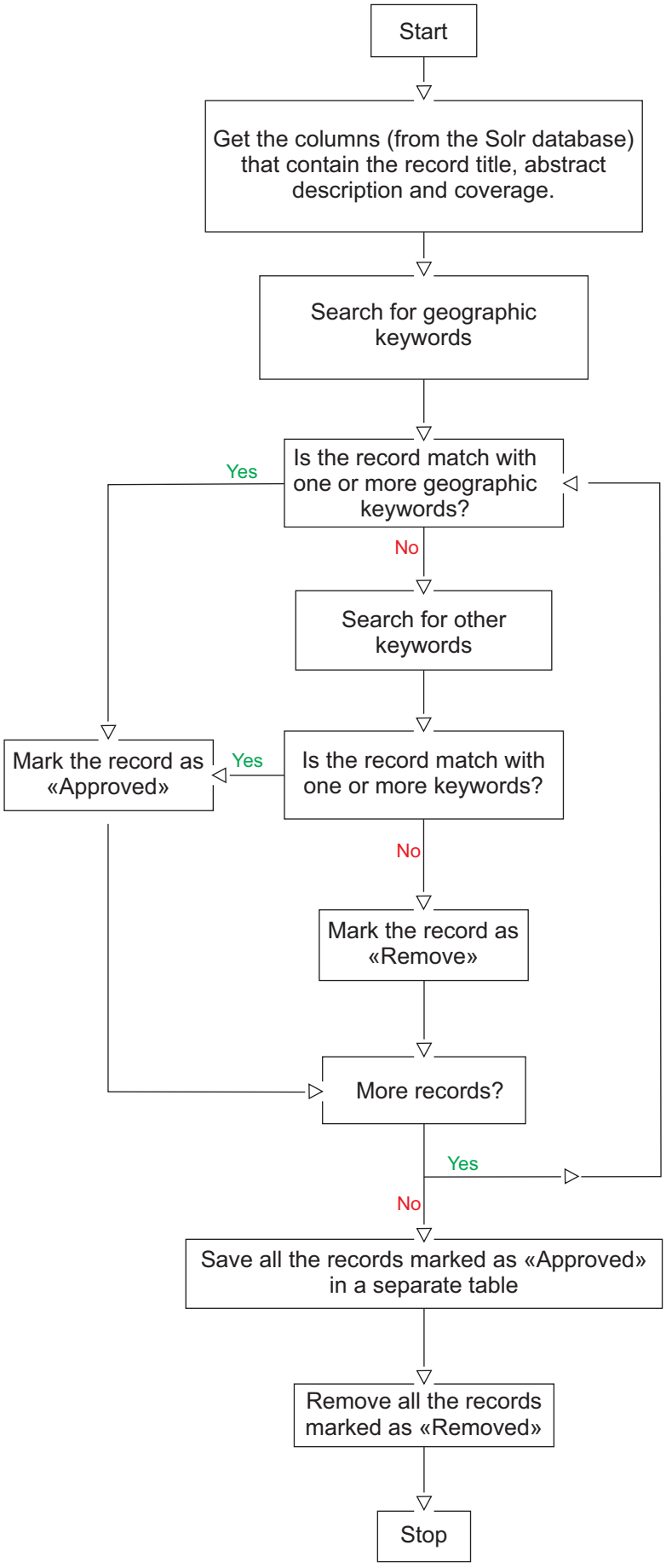


5. Third filtration stage (remove the duplicated records based on the original ID).

The aim from this stage is to check the original ID for each record and compare them, then remove the duplications and keep only a record from a certain data provider (or based on the recent hdate).

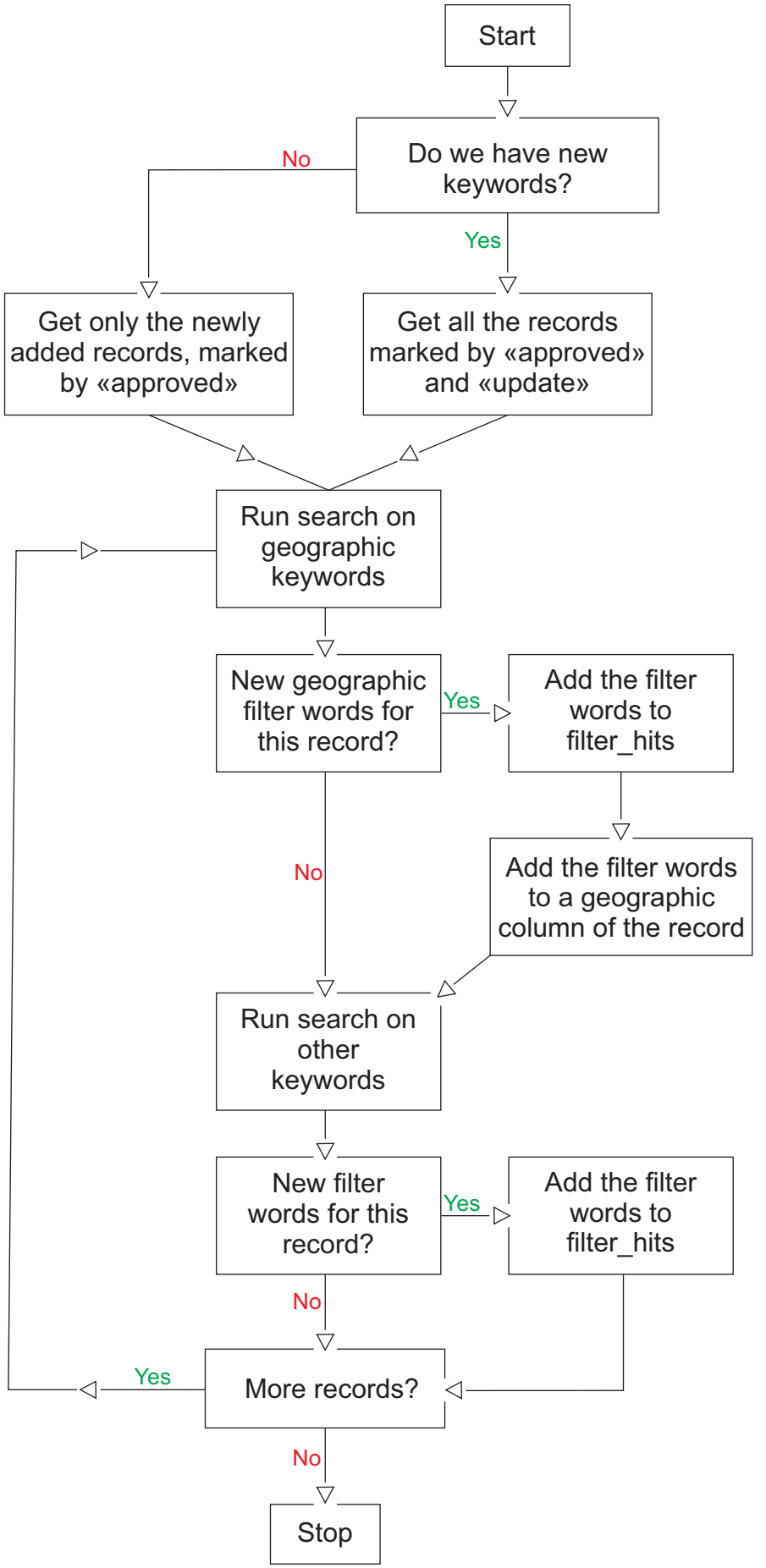


6. Fourth filtration stage: Check the title, the abstract and the coverage and search for keywords (geographic or other approved keywords). Use combination of one or more keywords in order to get the best out of this filtration stage.



7. Extract and list geographic keyword for each new record and link it to the location table.

Find how many keywords hit by each record. Get all the approved records and run search on keyword

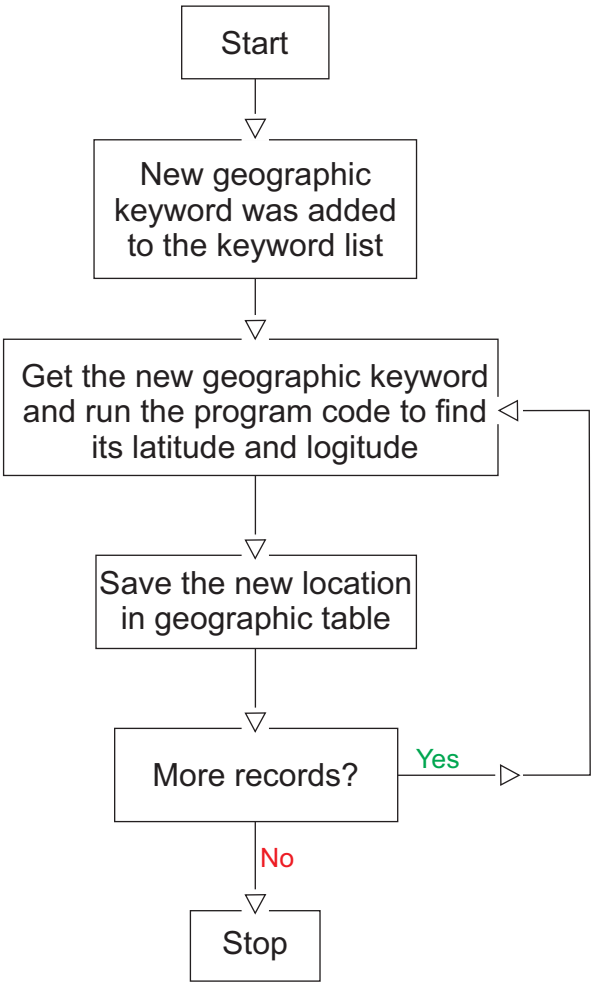


The geographic column will be used to extract the location and polt the record on map

8. Get the location of new place and add it to a table contains latitudes and longitudes.

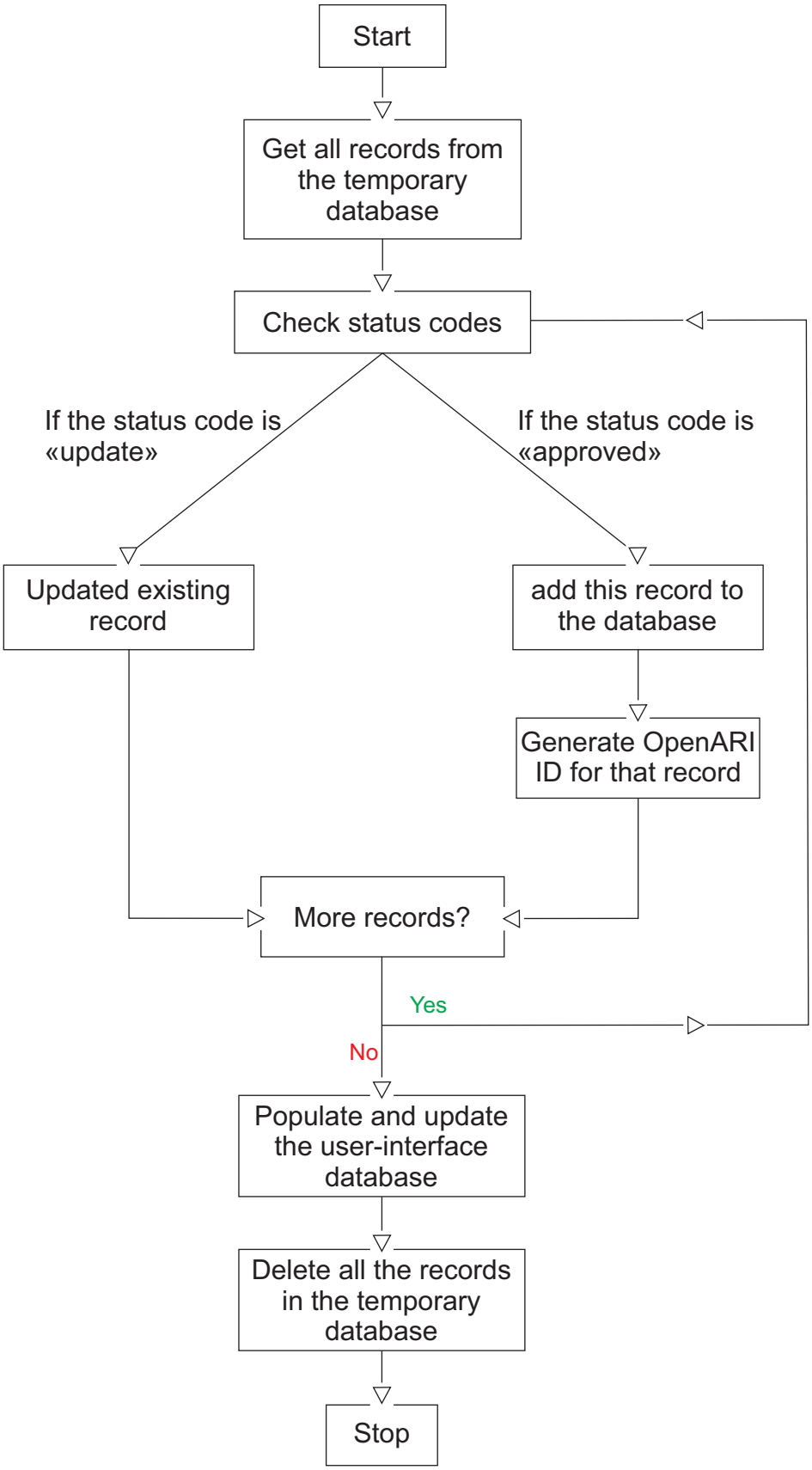
The following flow is to read a new geographic keywords and get the location of it.
We need this program only when a new geographic keyword is applied or a new record was added to the database with new location in the coverage item.

New geographic keyword



9. Add new records / Updated the records that existing in our final database.

All the approved records should be added to our final database. All records marked by “update” should be used to updated the existing records. During the update, we will keep the record ID, the record title and the geographic keyword and replace all the other fields.



Appendix F

Abstracts and outreach activities of the OpenARI's team during the pilot project to distribute the message and the idea of the project.

Archiving and collecting Arctic datasets: Open Arctic Research Index*

Abu-Alam, T.S., Karl Magnus Nilsen, Obiajulu Odu, Stein Høydalsvik

Universitetsbiblioteket, UiT Norges arktiske universitet

tamer.abu-alam@uit.no

The number of digital repositories containing publications and datasets on the Arctic region are increasing enormously. Users want relevant information according to their query with minimum interval of time. Scholars are compelled to search the individual repositories to get their desired documents.

Open Arctic Research Index (Open ARI), a planned service at the UiT - The Arctic University of Norway, aims to collect and index all the openly available Arctic-related publications and datasets in a single open-access metadata index. By providing a simple search dialog box to the index, users can search all these repositories and archives in a single operation.

The project investigates how such a service can support researchers in their research by making results from Arctic research more visible and better retrievable based on a standardized, interdisciplinary metadata set. The project started by clarifying the need of new technical solution to collect all the published material using algorithms that allow the best way of filtering relevant records. We have defined 115 possible national and international collaborators who can feed the Open ARI with content. The team will analyze the success opportunities and the challenges in order of planning a full-scale management model.

Keywords: Open ARI, Dataset repositories, Arctic Research, Polar Sciences

* Submitted to the Open Repositories 2019 conference - Hamburg, June 10-13, 2019

Open Arctic Research Index (Open ARI): A new horizon in Archiving and collecting Arctic datasets*

Abu-Alam, T.S.
(Universitetsbiblioteket, UiT Norges arktiske universitet; tamer.abu-alam@uit.no; ORCID: 0000-0001-6020-365X)

Open ARI is a planned service at the UiT - The Arctic University of Norway in order to collect, sort and archive all the openly available publications and datasets that were published on the Arctic region. This new service will be available as an open-access database to the users through-out an interactive searchable front-end. The pilot project will investigate how such a service can support researchers in their research by making results from Arctic research more visible and better retrievable through a common search index based on a standardized, interdisciplinary metadata set. Moreover and for a better overview for the polar sciences, the new Arctic database will include, as well, examples from the Antarctic region. As a pilot project, we started by clarifying the need of new technical solution by which the Open ARI will be able to collect all the published material using algorithms that allow the best way of filtering processes. Also we are now in a stage to define all the possible national and international collaborators who can support and feed the Open ARI with content from their internal databases. A group of scientist and researchers will be formed as a reference group who will show us the needs of the scientific community to be sure that our final product will meet the interest of the users. By the end of the pilot project, the team will analysis the success opportunities and the challenges in order of planning a full scale management model.

Keywords: Open ARI, Dataset repositories, Arctic Research, Polar Sciences

* Presented in the Munin Conference on Scholarly Publishing conference - Tromsø, November, 2018