Polar information sources – shining stars or black holes in the global Open Access network?

LEIF LONGVA *and* **STEIN HØYDALSVIK**
University Library,
UiT The Arctic University of Norway

Abstract

Open repositories holding scholarly documents are increasingly used for dissemination. These repositories commonly comply with the OAI-PMH standard, making it possible to automatically harvest the repositories, and build discovery services on top of scholarly repositories throughout the world. This is what Bielefeld Academic Search Engine (BASE) has done. In BASE, any record is easily searched and discovered, irrespective of how small or remotely located the repository is where the document is archived. High North Research Documents (HNRD) is an overlay service of BASE. The entire set of more than 100 million metadata records in BASE is subject to a filtering process, returning more than 700 000 records with relevance to the polar regions. We have analyzed the sources harvested by BASE, and the sources present in our HNRD service. This shows us from where the polar related scholarly outputs originates. Our analysis also reveals institution and regions that are more poorly represented in HNRD, compared to what should be expected. We believe there are several very interesting sources not following OAI-PMH and thus not present in BASE nor in HNRD. We invite PLC members to join us in an international cooperation to identify sources that are still not harvestable and thus not part of the global OA network. The next move would be to guide these sources and their mother institutions to migrate their sources to OAI-PMH enabled platforms. We further call on PLC members to cooperate in improving the dissemination, accessibility and discovery of polar related information through repositories.

High North Research Documents

High North Research Documents (HNRD) was launched in January 2012[1]. HNRD is a discovery service for open access scholarly literature and research data with relevance to the Arctic or the high north. The service is run by the library at UiT The Arctic University of Norway, in cooperation with Bielefeld University Library. The service is an overlay service of the Bielefeld Academic Search Engine (BASE).

All over the world, scholarly literature as well as research data are made openly available in repositories. Higher education institutions, and research institutes too, commonly have their Institutional Repository. And increasingly, scholarly documents are published as open access. However, openly available does not necessarily mean easily accessible.

The protocol The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[2] is commonly used to disseminate scholarly documents. This protocol enables services to harvest metadata from selected repositories and open access publishers' archives, and create a search service for scholarly documents and data, based on selected and possibly numerous repositories and sources. This is possible only as long as the repositories to be harvested satisfy the technical requirements of the OAI-PMH protocol.

BASE uses the OAI-PMH harvesting method and they harvest any harvestable source (they are aware of) with open access scholarly content, be it articles, reports, books or book chapters, conference objects, as well as research data, within any subject or research area. BASE is thus potentially a search service for all open access documents and research data, if all sources world wide were compliant to the OAI-PMH protocol.

Based on the entire pool of (per July 2018) more than 130 million records (documents and research data) in BASE, HNRD is doing a filtering process to select records that are relevant to the Arctic. As per July 2018, HNRD includes close to 1 000 000 records (780 000 documents and 190 000 data sets). And this is done through a rather simple algorithm filtering through the 130 million records in BASE, finding records that are relevant to the High North. The sources harvested by BASE do not need to adhere to some common metadata standard. The filtering process works more or less irrespective of which metadata schema is used, as long as the sources are OAI-PMH compliant. BASE does some normalizing of the metadata used by the various sources, and this is useful to the filtering process of HNRD.

Open ARI

HNRD has been operating for more than six years now. Our plan is now to do a pilot project to survey how the service may be revitalized.

UiT The Arctic University of Norway is devoted to research on and development of the Arctic. UiT's strategy (towards 2022) is named "Developing the High North"[3], and it lists several areas where UiT intends to play a major role in this respect:

- Energy, climate, society and environment: "Understanding what happens in the Arctic is key to understanding global climate change."
- Sami language, culture and quality of life
- Community development and democratisation: "The basis for collaboration and potential conflicts in the High North."
- Technology: "In a region characterized by long distances and a challenging climate, new technological solutions are needed to deliver welfare to the people living here"
- Sustainable use and management of resources

Developing and running a discovery service on scholarly literature and research data with relevance to the High North and the Arctic is thus falling nicely in line with this UiT strategy. And the UiT and its management has been backing our HNRD service, and is also backing our plan to now run a pilot project in order to revitalize the service.

The pilot project will be a cooperation between Norwegian Polar Institute and UiT The Arctic University of Norway. The revamped service may change name, tentatively to Open Arctic Research Index (Open ARI). We believe a service like HNRD is most useful for researchers and students who are working on Arctic related projects, but the potentials of the service has not been fully realized. The service needs to be developed further and managed closely, to become a vital service for the user community.

The user community and the business model

So who are the user community? As any Higher Education institution, UiT The Arctic University of Norway is concerned with the needs of its own students and staff. But no university or research institution is an island. Research builds on previous research, and thus access to research results and research processes produced and performed anywhere in the global scholarly community, is to the benefit of research progress at large, as well as the progress of teaching and students' learning. This is one of the main motivations for open access to scholarly documents and research data. Easy access to well documented research yields the best and fastest progress in further research. And also any other use of the accessible documents and data, in business, public administration, in the work of NGOs and interest groups, or whoever may have use of the documents and data.

So the user community is basically anyone who may have interest in research on the Arctic. With open access documents and data, there is no competition involved between users. One user group is not blocking access for any others. On the contrary – any reuse of the research may produce further analysis and results, to the benefit again of others.

As mentioned above, UiT The Arctic University of Norway is, according to its formulated strategy, dedicated to focus on projects and topics with an Arctic scope. It was therefore never controversial for UiT to fund the development and the operation of the HNRD service. As for the refurbished service Open ARI, The UiT Library will fund the pilot project, that will lead to a recommendation of developing the full scale Open ARI or not (in addition to some in kind funding in terms of labour supply from Norwegian Polar Institute). Funding of the full scale service is a question to be investigated in the pilot project. We certainly do believe that if the pilot project recommends to go forth with the full scale project, the UiT as owner of the service will follow up with the necessary funding – possibly in cooperation with some co-owners and partner organizations.

The service Open ARI will be a discovery service on open access documents and data. It would be somewhat anachronistic to present this as a service requiring payments from its users. So the service itself will of course also be open access. The service will thus need funding from its owner and/or some sponsor(s), in order to be developed and launched. The bulk of the funding needed will be in the form of labor hours, hopefully also from a list of important partner institutions.

The pilot project

The pilot project will run for six months from this coming fall. The pilot will look into the needs of the user community (researchers, students and others) in order to make it an improved (compared to HNRD) service, and also describe how this may be achieved technically and organizationally. The pilot will survey possible cooperating partners world wide. If Open ARI shall succeed, active partners from centrally positioned institutions will be vital. The pilot shall end up with a prototype service based on the HNRD service and experiences. And finally the pilot shall describe a full scale project to develop an operating Open ARI service, and also describe a viable model for running, financing and managing the service.

The pilot will build on the experiences of HNRD and its cooperation with BASE. BASE is able to harvest the metadata of all and every repository with scholarly content, as long as these repositories are compliant to the OAI-PMH protocol. BASE is therefore a very useful partner in the project and service, and this partnership will be continued. The pilot will however go further, and look for important sources with Arctic related content, that are not captured by BASE, for the reason that they are not harvestable. The content of sources like these may be very interesting to include in Open ARI. The sources may have APIs that allows outsiders to harvest their content. Or we (Open ARI) may have to develop a tailor made API and get hold of the content.

One of the advantages of making use of the entire pool of harvested metadata in BASE, is that we are thus able to find and pull out interesting documents or data that resides in repositories that are not considered as important sources from an Arctic perspective. This "long tail" of repositories may hold one or a few documents or data sets that are of interest from an arctic perspective. Developing APIs on the other hand, we will only be capable to do for a limited number of repositories. So there may be interesting documents in some repositories that we do not discover. The best, for our Open ARI service, would therefore be that all repositories were OAI-PMH compliant, enabling a service like BASE and ours to harvest and extract the metadata.

We would like to stress that Open ARI does not require the many repositories to adhere to a standard metadata schema, beyond the OAI-PMH requirements. In order to find the interesting documents and data sets, Open ARI will search through the discovery metadata (e.g. title, description, abstract, subject and keywords) of each record, to find the records to include in Open ARI. These metadata, along with author, date and document type, and the (persistent) url back to the record in its repository, are what Open ARI will present following a performed search. We will do a basic mapping of these discovery metadata, so that all metadata presented appear uniformly. The user will further have access to the full record of metadata describing the individual document or data set, as well as the full text documents or data sets themselves, by clicking the url link.

Open ARI is thus not holding any full text documents or data sets, but merely a limited set of metadata for each record. Open ARI is therefore helping the repositories to disseminate their content to new readers and users. The authors will thus enjoy enhanced visibility and a wider circle of students, researchers and others who may find their documents and data sets interesting, which possibly may lead to new citations any kind of new reuse of their output. So Open ARI is not competing with anyone, but rather creating a win-win situation for all parties involved.

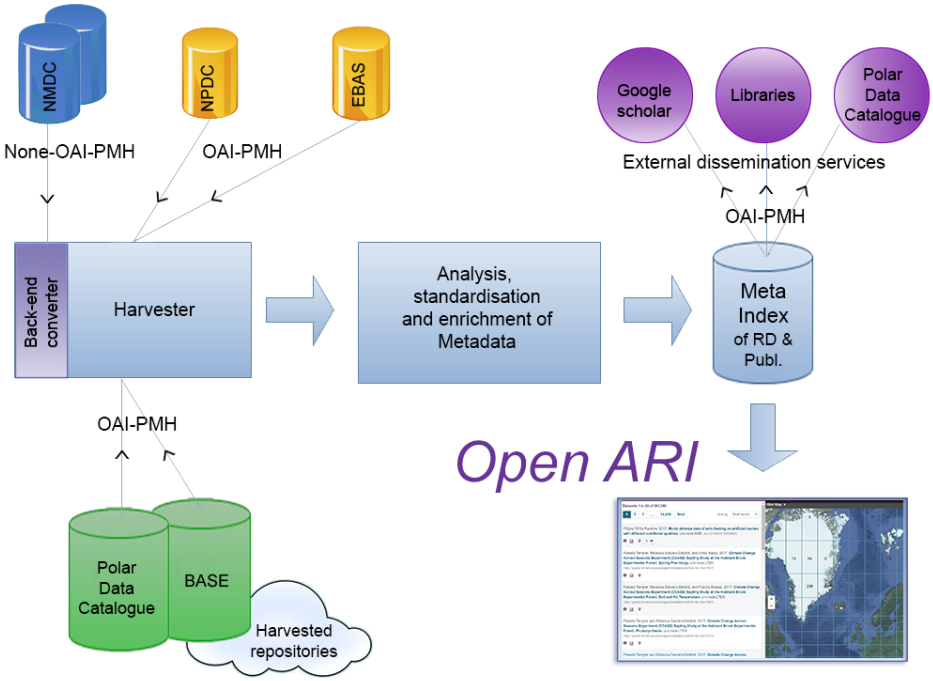

Figure 1. Illustration of the principles behind Open ARI

Another important question is all the documents and data that are behind paywalls. Should Open ARI include these in some way? Records of these type of documents and data are much more difficult to include, because of the paywalls. Their metadata is namely also normally behind paywalls, and not easily accessible. The pilot project will look into this, and see if there are ways to include also non open access documents and data sets, in a comprehensive (Open) ARI.

Coverage of the current service

Currently, per July 1. 2018, the HNRD service holds close to 1 000 000 records, from close to 3400 content providers. This encompasses more than 50% of all content providers in BASE, which is rather amazing.

Looking at the various contributing sources, we have made an effort to analyse from which country the content of HNRD originates. Some sources are international in their scope, and these fall in the category "International sources" in the table below. The remaining sources we have attached to their native country, and summed up the number of records from each country. The analysis is limited to a summing of sources contributing 300 records or more. We get the following distribution of the most important nations:

Country	Number of records*	% of total records
International sources	242 573	27,36 %
USA	161 430	18,21 %
Canada	146 502	16,52 %
Germany	126 691	14,29 %
France	35 124	3,96 %
United Kingdom	27 598	3,11 %

Australia	21 356	2,41 %
Russia	19 221	2,17 %
Denmark	16 214	1,83 %
Norway	15 843	1,79 %
Sweden	11 087	1,25 %
Finland	9 749	1,10 %
Iceland	9 381	1,06 %
Belgium	8 356	0,94 %
Japan	7 389	0,83 %
Other countries	28 106	3,17 %

* Counting only sources contributing 300 records or more.

We see that the list is dominated by western countries. Important countries like Russia and Japan contributes approximately 2 and 1 percent respectively. While China, an important country with interests in the Arctic, is outside this top list all together, contributing merely 0,3 % of the records. This is even less than China's contribution to BASE (which is approximately 0,6%). This may indicate that the filtering process is not good enough towards the Chinese records. But we need to look closer into this in order conclude.

One important issue is of course the languages of the documents and their metadata. The filtering process of HNRD strives to include documents and data sets of any language. However, we realize that there is still a way to go, until we have a good coverage of the various languages used. Here is a table showing the distribution of the various languages appearing most frequently in HNRD:

Language	**Number of records**	**% of total records**
English	630 261	65,08 %
Unknown	268 495	27,72 %
French	14 441	1,49 %
Spanish	13 971	1,44 %
Portuguese	12 184	1,26 %
Norwegian	11 895	1,23 %
Icelandic	6 660	0,69 %
Russian	5 569	0,58 %
German	5 123	0,53 %
Swedish	4 357	0,45 %
Finnish	3 959	0,41 %
Japanese	2 522	0,26 %
Danish	2 413	0,25 %
Chinese	1 639	0,17 %
Polish	1 289	0,13 %

Not surprisingly, English is the predominant language. This can be explained from the fact that English is the dominating language in scholarly output. But to some extent, we know that it is also due to the fact that the filtering process of HNRD is not capturing languages others than English and Norwegian good enough.

Another issue to discuss is where to draw the geographic borderline. Which areas belongs to the Arctic, and thus which geographically located topics should be included. In HNRD, the definition drawn by Arctic Monitoring and Assessment Programme is used, with some minor adjustments:

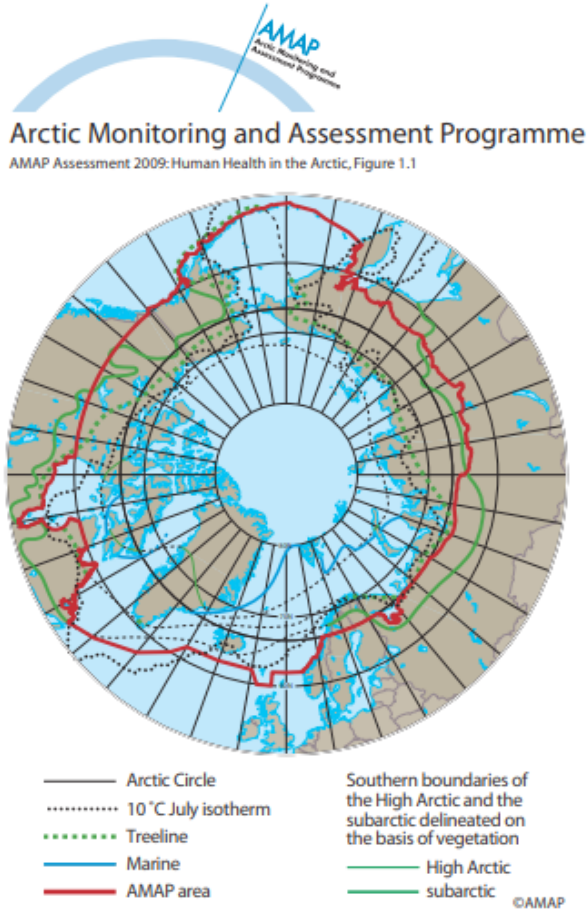

Figure 2. Illustration of the various definitions of the Arctic[4]

In HNRD, also documents and data sets about the Antarctic region is included, since many issues will be relevant to the Arctic (e.g. ice and cold waters). In the Open ARI pilot project the geographic boundaries of what to include and not from an Open ARI service, including the question of Antarctica, will be discussed.

Invitation to cooperation

Our pilot project will survey possible cooperating partners world wide. Interested institutions, who work within the thematic scope of the Arctic, are hereby invited to contact us, so we can start discussing how we may cooperate in order to develop a best possible Open ARI service, to the benefit of all scholars as well as others who have interests in the Arctic.

We need to develop a service that covers all languages used in scholarly documents and data sets, that covers all geographic areas of the circumpolar Arctic region, and covers all subjects areas and themes as long as the content is relevant to the Arctic. If we can achieve that, the Open ARI service will become a very useful service to the user community.

References

1) LONGVA, L.; HØYDALSVIK, S. High North Research Documents : your source for research documents on the North. Polar libraries bulletin (2012) no. 68 p. 7-9 http://hdl.handle.net/10037/4733
2) See https://www.openarchives.org/pmh/
3) Developing the High North. UiT's strategy towards 2022. https://uit.no/Content/572401/cache=20181304152812/Developing%20the%20High%20North_web.pdf
4) From https://www.amap.no/