High North Research Documents – a new thematic and global service reusing all open sources

Leif Longva, Obiajulu Odu, University of Tromsø

Abstract

High North Research Documents is a thematically tailored search service, based on all the open and freely available research documents in the world. The relevant documents are identified through their freely available metadata records. A set of specialized keywords are applied to this end, in a filtering algorithm. An automatic filter process is emphasized, minimizing the need for manual labor in the process. The filtered records relevant to high north are ingested into a customized DSpace repository which serves as end user search and discovery interface. The metadata records used as input are made available through a co-operation with Bielefeld University Library and their BASE service, the harvester of open repositories worldwide. High North Research Documents is available at http://highnorth.uit.no

Keywords

Open Repositories, Open Access, Metadata Reuse, OAI Protocol, Dublin Core, Harvester, Thematic Search Service, High North Area, DSpace, PHP, Mysql

Introduction

The world-wide proliferation of freely available scholarly documents has further intensified the need for discovering, sharing, exchanging and reuse of scholarly information across research communities. Generalist search engines as Google Scholar, metadata harvesters as BASE, indexes as DOAJ are access providers for large masses of open scholarly contents. However, these search engines still require a highly significant effort for research stakeholders to search and retrieve thematic documents for targeted needs.

Lately, access to the thematic documents relevant to the circumpolar high north has gained much attention. Also, there are growing interests among politicians, public administrators, business executives, NGO members, as well as researchers and students in accessing these documents. A crucial need exists for interface that will improve accessibility and search for documents relevant to high north area.

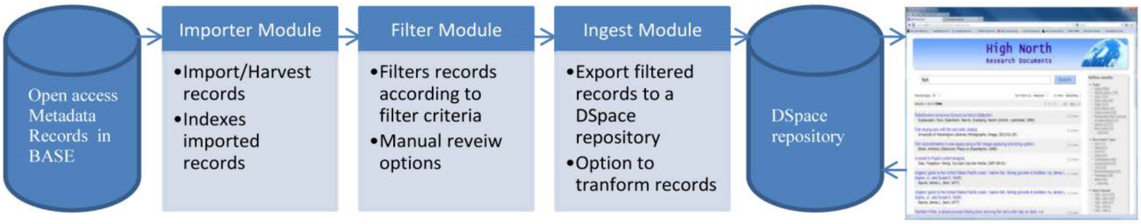
This proposal describes the development of the High North Research Documents (HNRD) system at the University of Tromsø. The system improves the propagation of thematic documents relevant to high north area by providing *advanced services* that import metadata records from open access sources, filter out those metadata records relevant to the high north topic into a single point of access by means of a search and discovery approach. Metadata figures prominently in HNRD development of the advanced services.

Collection of Metadata Records for High North - how?

For open resources, metadata is a valuable asset that should be shared with external systems. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is widely used standard for disseminating metadata records. One of the world's largest collections of OAI metadata is BASE at the Bielefeld University Library. BASE has, so far, harvested more than 34 million documents from more than 2000 resources. BASE's collection policy is to harvest *all* of the OAI repositories in the world. BASE has, perhaps, the widest subject variety of any digital contents. Therefore, creating consistent enhanced and thematic metadata is one of the biggest challenges using metadata records from such a harvester as BASE. HNRD project is cooperating with Bielefeld University Library in utilizing the metadata records stored in BASE in creating the HNRD service with the high north relevant documents.

Overview of HNRD Services

In order to develop a new thematic and global service reusing all metadata records from BASE, we employ three separate modules in HNRD. These modules are open-source applications written in PHP.


Open access Metadata Records in BASE
Importer Module
•Import/Harvest records
•Indexes imported records
Filter Module
•Filters records according to filter criteria
•Manual reveiw options
Ingest Module
•Export filtered records to a DSpace repository
•Option to tranform records
DSpace repository
High North
Research Documents

Firstly, the **Importer Module** enables the collection of metadata records deposited in BASE. Currently, the collection of metadata records from BASE is done using SSH based protocol, *rsync*. This is used because OAI-PMH as a transfer tool is ill-matched to the task of large one-time data transfer, compared to an *rsync* transfer of records. We intend to use the module in future to reap the benefit of the OAI-PMH for new additions and updates of metadata records. The module stores all DC attributes in the metadata as separate fields in a database. In order to improve system efficiency, the fields are indexed using full-text properties of the database, MySQL .

The second, a **Filter Modul**e**,** offers both a manual and an automatic filter option. This module leverages Importer Module's use of indexing and the Mysql search specification. The service defines filter criteria and on how to extract metadata records relevant to the high north.

Filter criteria are carefully constructed using among others specific key-words, phrases or combination with Boolean operands. These key-words include geographical names, species names, names of people and languages, as well as other categories of key words. So far, English and Norwegian language versions of the key-words are dominant in our list, plus species names in Latin.

Extracting those metadata records relevant to the high north is achieved by applying the filter criteria on the metadata records. We have three classes in our classification of filtered metadata records. First, there is the class of "Automatically approved", which represents the high north relevant (positive) instances. The second is “Manual control”, which represents the instances where we need to check the high north relevancy manually. To validate these records, we use specific tools which highlight key-words found in the records. The tools allow us to approve or reject a record. The last class is “Discard”, which are instances we regard as not high north relevant at all. Discarded records are entered into a blacklist of records, to make sure those records for ever are excluded from HNRD. Based on filter results, we may choose to update the filter criteria terms.

The third, an **Ingest Module** which provides the functionality to transform metadata records relevant to the high north into DSpace XML format and ingested them into a DSpace repository. The Ingest service includes other transformation options to upload and add custom information into metadata records. For example, a simple transformation could be created to insert key-words into metadata record to facilitate building of facets.

End User Search and Discovery Interface

One of the requirements of the HNRD system was to provide end user with both a regular search interface and faceted search. Faceted search provides among others intuitive way to refine search results by category (facet). A second requirement was the provision of feedback mechanism, where end users can enter suggestions on how to improve the HNRD system, for example by informing us about an irrelevant metadata record in the HNRD system.

HNRD project adopted DSpace as platform for end user search and discovery interface. This is because 1) we have local expertise that made DSpace an attractive choice and 2) DSpace employs two modules in its design to address our requirement issues. The modules are called “Manakin” and “DSpace Discovery”. Manakin creates an abstract framework that provides for the creation of individual, customized repository interfaces. "DSpace Discovery", a faceted search feature gives us the ability to create ‘Views/Topics/Limit by/Refine Search Results’ over the data, based on properties (metadata) of records. Out of the box, DSpace Discovery comes with facets for Author, Subject and Date issued. We employed the module to create more high north specialized facets for additional metadata values. HNRD extended the functionality of Manakin to include an integrated feedback system where end users can report among others lack of relevance of a document, if a metadata record is without the pointer to the digital full text material.

The *HNRD* applications include a scheduling facility, making it possible to set-up schedules to automatically and routinely collect, filter and ingest new records.

The lessons learned in collecting high north metadata from BASE

The issue of metadata quality is an important factor in the HNRD system. Even if all other aspects of the system worked perfectly, poor quality metadata would degrade the quality of the resulting contents relevant to high north.

Lack of full text material

Some of the metadata that passed through the HNRD filter criteria are purely bibliographic, i.e. metadata without the pointer to the full text material. Since our goal is to collect only metadata records with full text materials, we strive to remove this kind of records from HNRD. Due to the volume of records collected, we cannot check each URL to see if it is pointing to digital full text material.

Managing duplicates

Some of the records that passed through the NHRD filter criteria have duplicates. Too many duplicates in a result list can affect the end user's comfort. A phenomenon that can generate duplicates in HNRD is for example, if a publication is written by authors from several collaborating institutions. If so, this publication may be archived on the server of each institution and BASE harvests these institutions.

Managing DC fields and values

We found out that some repositories exposed their DC field information by using non-standardized text strings. For example, in document Type field, *text* and *article* were used to describe the same article. They were different variation in Date value: *1988-04-00, 1981, ca. -91, n.d*. Other fields such as Language pose the same problem as Type and Date.

The HNRD project is working on introducing some standardization on these fields in the next release of HNRD.

Discussion and future work

The ease to reuse metadata as well as the documents, is one of the great advantages of freely available documents. We have presented HNRD as a system that provides a set of advanced services that collect and reuse metadata records stored in BASE, and then filters out documents relevant to high north. The results are exposed to end user through a platform for search and discovery.

The quality of the metadata is a key factor to the efficiency of the reuse. Our work shows that the quality of the filtering service was affected by the quality of the metadata. We are working to find other strategies and methods to improve the filtering of metadata fields, in order to improve the quality of High North Research Documents. Defining new words or phrases, and translating them into other languages, is a priority in the work. Analysis of non-relevant records is one of the means by which the filter criteria may be improved.

The end user interface includes a feedback function by which the end users may help us in the development of the filter criteria.

HNRD was officially launched in January 2012. The filter service allows us to extract over 128.000 relevant records found within more than 34 million records in BASE. The *HNRD* implementation provides services covering all subjects and many languages, and it is free and open for all at the URL, http://highnorth.uit.no. HNRD is an example we believe may be followed by others.