



High North Research Documents – A Thematic Overlay Service of Open Access Documents

Open Repositories 11 July 2012

Obiajulu Odu and Leif Longva
University Library of Tromsø





High North Research Documents

Outline

- What is High North Research Documents?
- Why High North Research Documents?
- How did we build High North Research Documents?
 - Making use of BASE
 - Extracting relevant documents
 - A brief look from the technical side
- What we have achieved
- What we have learned



High North Research Documents – why?

- The Norwegian government has emphasized the importance of the northern areas
- The north is also of interest on an international level:
 - Politically and strategically
 - Environmental and climate change issues
 - Resource utilization
 - The northern sea route to the Pacific



High North Research Documents – why?

- The University of Tromsø (UiT) is the northernmost university in the world
- UiT want to profile itself as a key research institution on questions on the north
- A natural thing to do for the library at UiT to develop services around literature on the north



Research literature – the access issue

- The north is of interest to many parties
 - Politicians and governments – international, national and local
 - Business owners and executives
 - Interest groups and NGOs
 - Indigenous people
 - Interested laymen
- Access to research literature is often restricted by publishers' barriers



High North Research Documents – the idea

- Open Access research literature
 - Free access for anyone
 - The metadata are free to utilize



- What if we take advantage of this freedom, to extract all the OA research documents, relevant to the north?

High North Research Documents – the motivation

- The North is a cross-disciplinary theme
 - Traditional subject classification does not help us
- Library users have expressed the challenging task of finding research documents on the north, through existing search tools
- Can we extract relevant documents by analysing the freely available metadata?
 - Preferably through automatic algorithms
 - Minimizing the labour needed



Cooperation with BASE

- «The whole world» is available through aggregators like OAIster (OCLC) and BASE (Univ. of Bielefeld)
 - We do not need to do what they have done already
 - Could such services be data providers for us?
- BASE: Has harvested more than 36 mill records
 - From more than 2200 sources world wide
- We contacted BASE to cooperate in our project
- And was met with helpfullminded response



Cooperation with BASE

- All the metadata are free to utilize
- To best do what we intended to do, we were allowed to collect all the metadata records
 - With the help from BASE (<http://www.base-search.net/>)
- And then we could apply the method of extracting the records that are thematically interesting from a high north perspective



What do the metadata tell us?

- Our hypothesis:
- If selected keywords are present in the metadata, then the thematic scope of the document is of relevance to the high north

=>

- We need to carefully select the keywords



Finding the high north relevant documents

- A set of filtering words (keywords) applied on the metadata of the BASE records:
 - Geographically
 - Species' names
 - Language and folks (nations)
 - Other keywords
- So far, mainly English and Norwegian language keywords, plus Latin species' names
- Aiming to extract documents by automation, minimizing the need of manual selection



The quality of the keywords

- Are the meaning of the keywords unambiguous?
- Words may have different meanings in different languages
 - Examples:
 - The keyword 'labrador' is meant to extract documents on the area of Labrador in northeastern Canada. However, the word 'labrador' means farmer or peasant in Spanish.
 - The keyword 'sami' should refer to the sami people of the north. Sami is also name of a district in Burkina Faso, in Gambia, and in Greece



The quality of the keywords

- Keywords may be identical to person names
 - Example:
 - 'sami' (people in the north)
 - Sami is a common given name (in Turkey and in Finland). Sami Kama is a researcher at CERN
- Some keywords need to be defined more explicit
 - 'sami AND language'
 - 'sami AND people'
- The filter process is applied only to selected metadata elements: title, description, subject

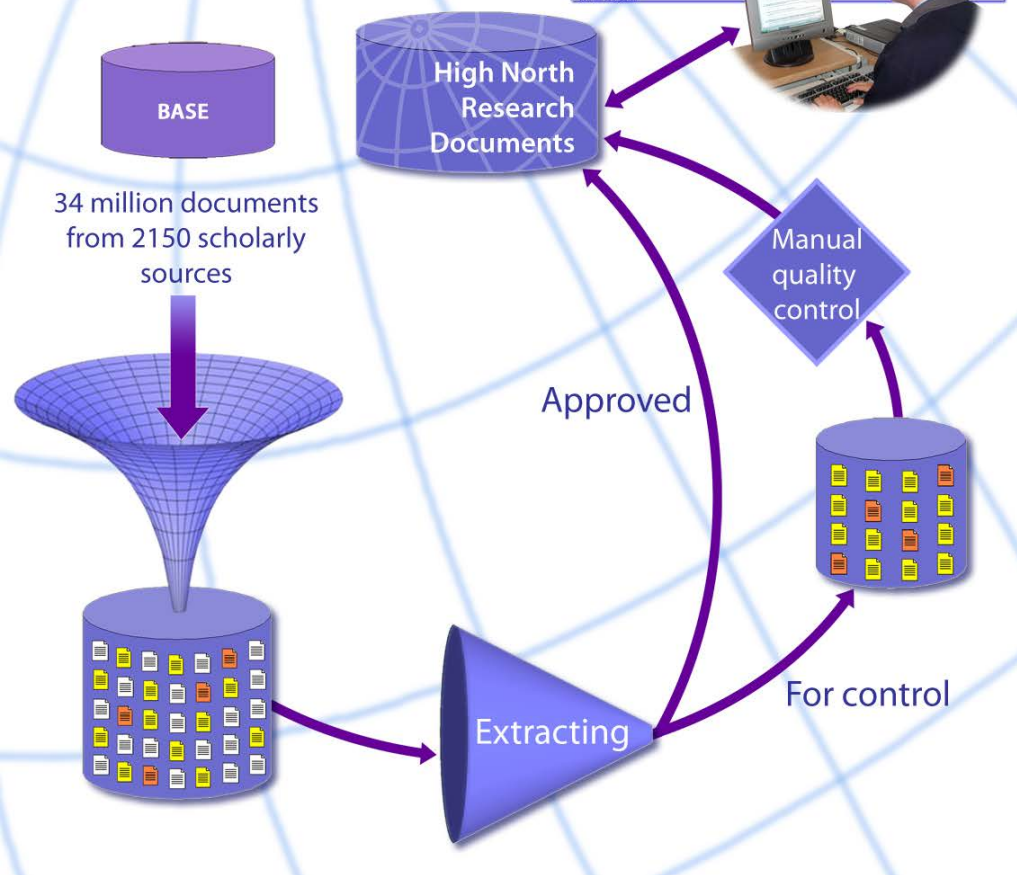


The quality of the keywords

- Geographical names:
- 'Newfoundland' – we do not want every document with University of Newfoundland as publisher
 - dc.publisher is not subject to filtering, but still:
- We need to do:
 - 'Newfoundland NOT University of Newfoundland'
 - Plus:
 - Records with 'University of Newfoundland' to be checked manually



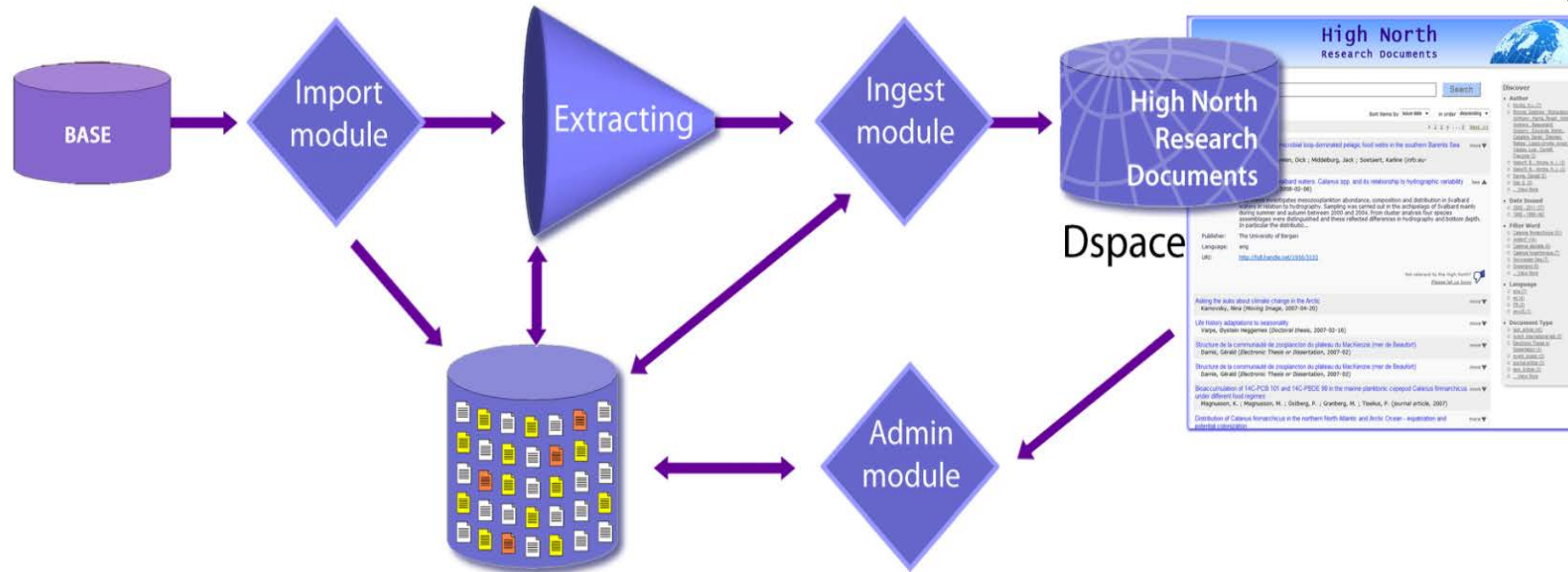
High North Research Documents



High North System Model

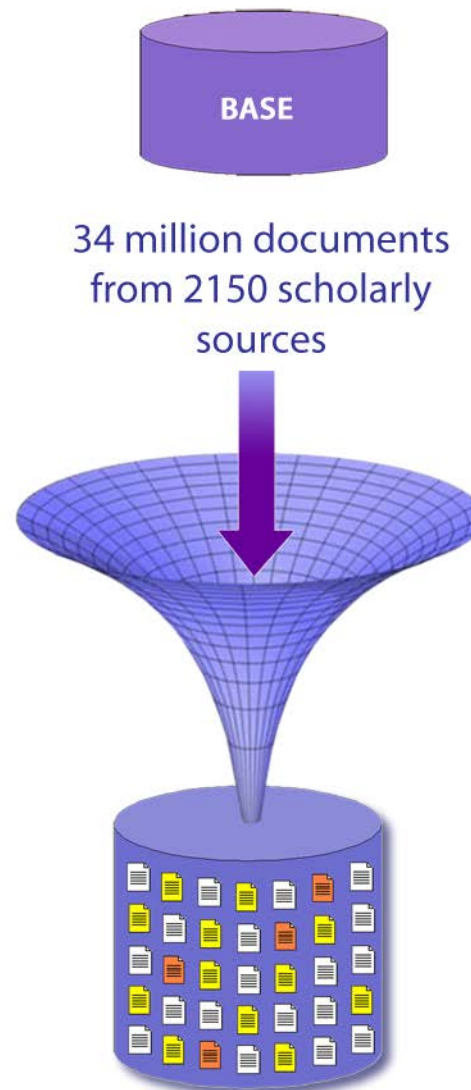


User search & discovery



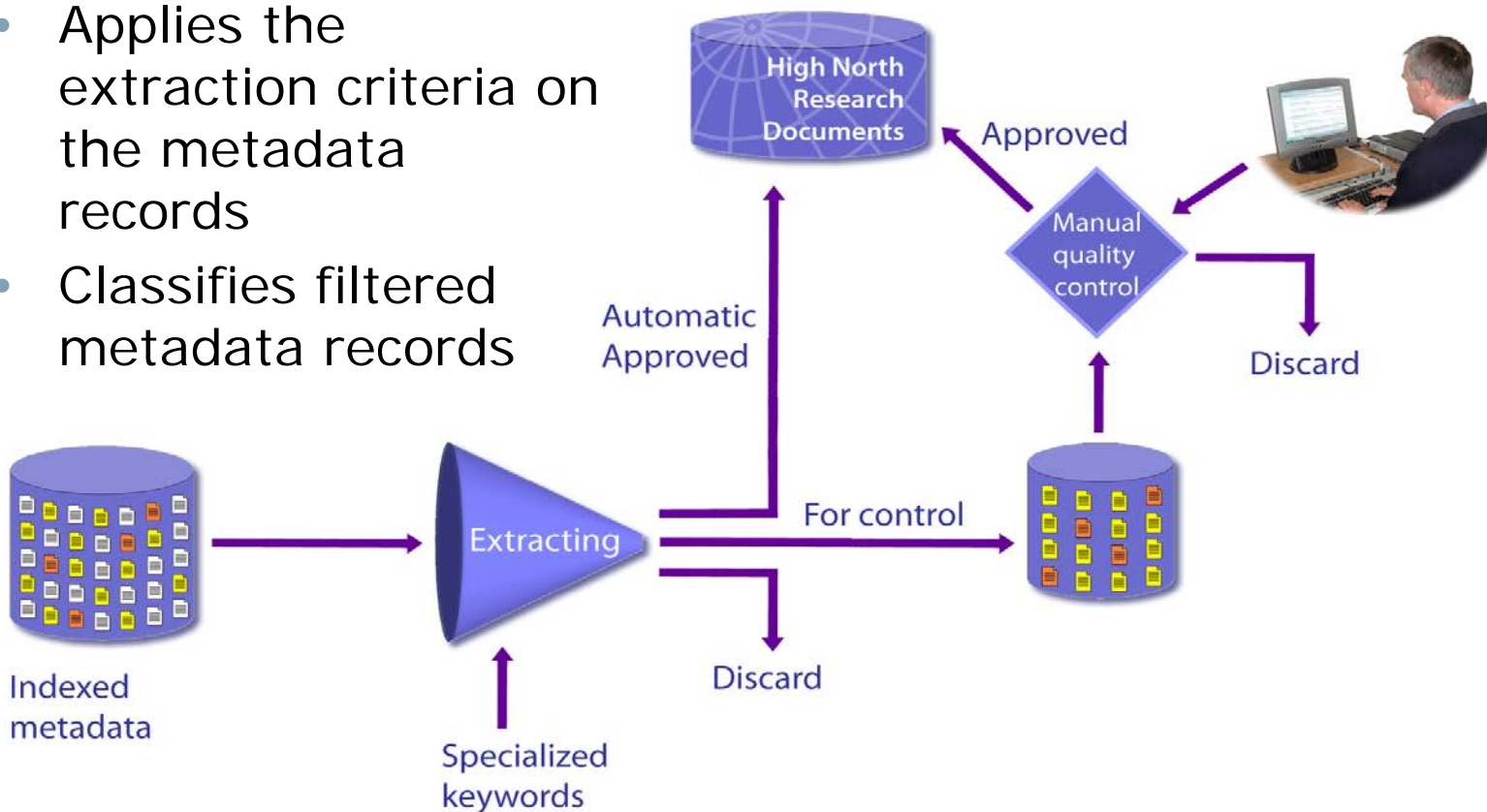
Import Module

- Gets metadata records from BASE using SSH based protocol, *rsync* for large one-time data transfer
- Stores and indexes all records in database using full-text properties of the database, MySQL



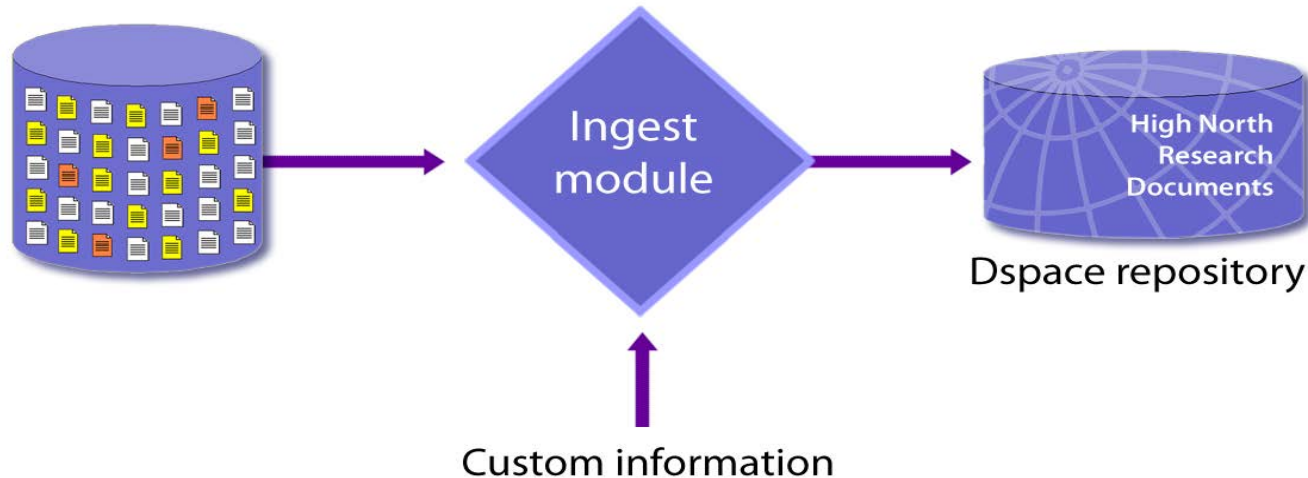
Extract Module

- Applies the extraction criteria on the metadata records
- Classifies filtered metadata records



Ingest Module

- Transforms metadata records relevant to the high north into DSpace XML format
- Imports them into a DSpace repository
- Options to add custom information to records:
 - Information that facilitates building of facets, etc



Admin Module

- To add, edit, or display all filtering words (keywords)
- To edit the status of a record or records:
 - Blacklist / Reject
 - Approved
 - Manual control
- By clicking on a keyword (searching), we can get all the records that this keyword has a match on
- Manages feedback reports



Search and Discovery Interface

- Why Dspace?

- We have local expertise in the house
- Provides end user with both a regular search interface and faceted search/browsing
- Provides for the creation of individual, customized repository interfaces - Manakin



<http://dspace.org>

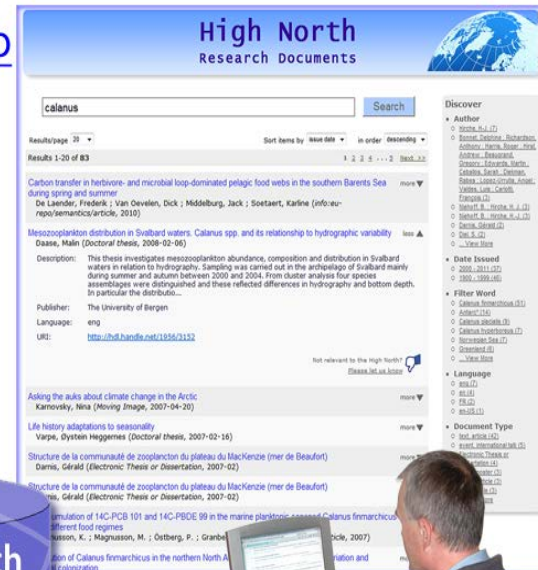
Search and Discovery Interface

highnorth.uit.no

- Extended DSpace to provide feedback mechanism such that end users can:
 - Enter suggestions on how to improve the service
 - Inform us about any irrelevant metadata record
 - Inform us about any record with non-existing URL



DSpace





High North RD v 1.1

- 151 000 documents extracted
 - From more than 50% of the sources appearing in BASE – and from all over the world
 - Many different languages
 - Even if we apply mainly English and Norwegian and Latin in the filtering process
 - Any subject, but weight on the hard sciences
- Developing the list of keywords is a priority
 - More and better keywords (and phrases)
 - Translating the list to more languages



Am 26.01.2012 20:30, schrieb Amanda Graham:

> This looks like it's going to be a superb resource. As a northern and circumpolar studies instructor, and on behalf of my students, I'm thrilled.

>

> Amanda Graham

> Yukon College

Comments:

Wonderful site! Thank you for putting it together. One problem: The link to the "Handbook for Alaska Legislators" from 1957 is broken. URL:

<http://hdl.loc.gov/loc.gdc/mtfgc.1016>

Date: 1/26/12 6:38 PM

Email: swoodham@alaskadispatch.com

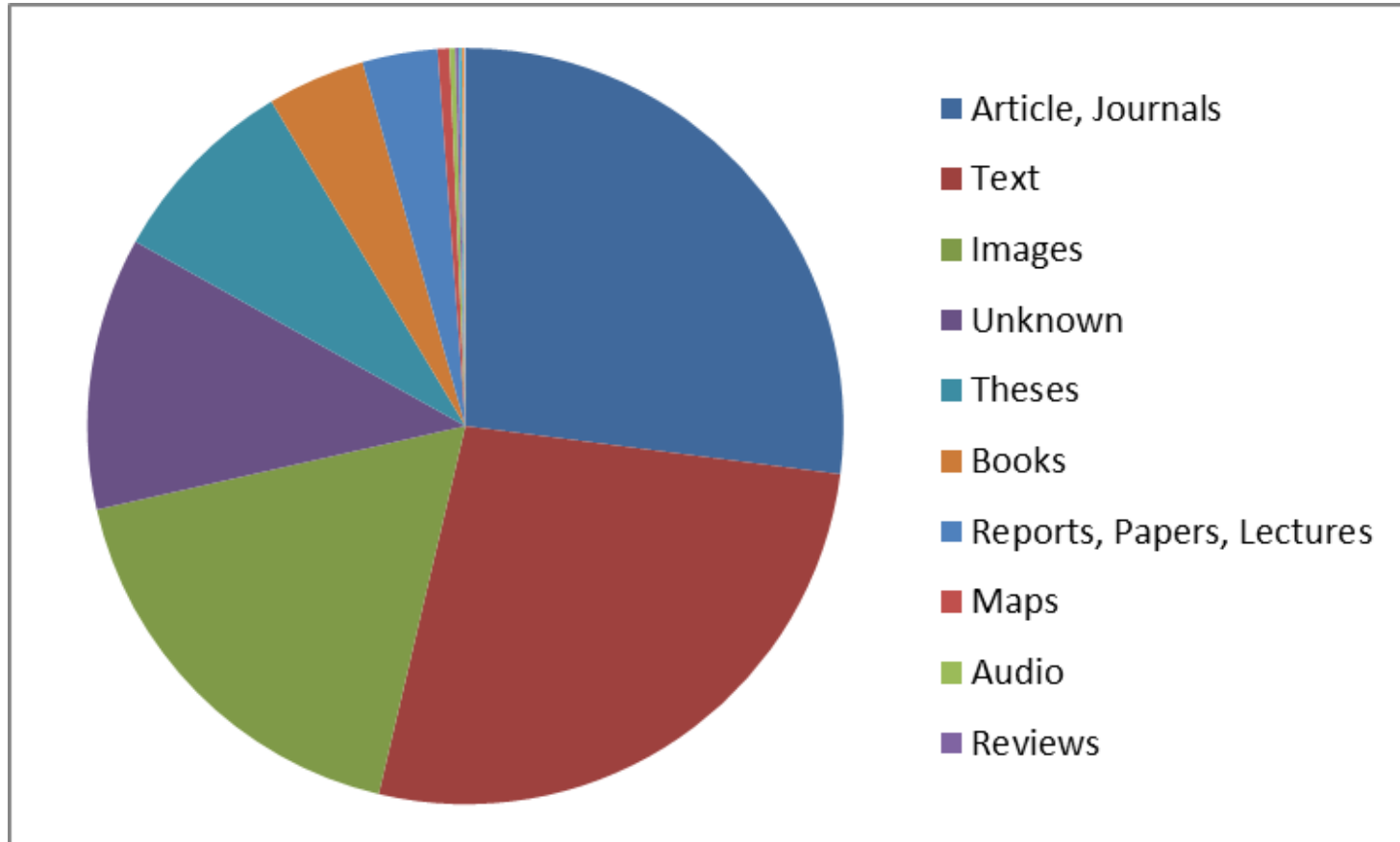


The problem of non-OA records

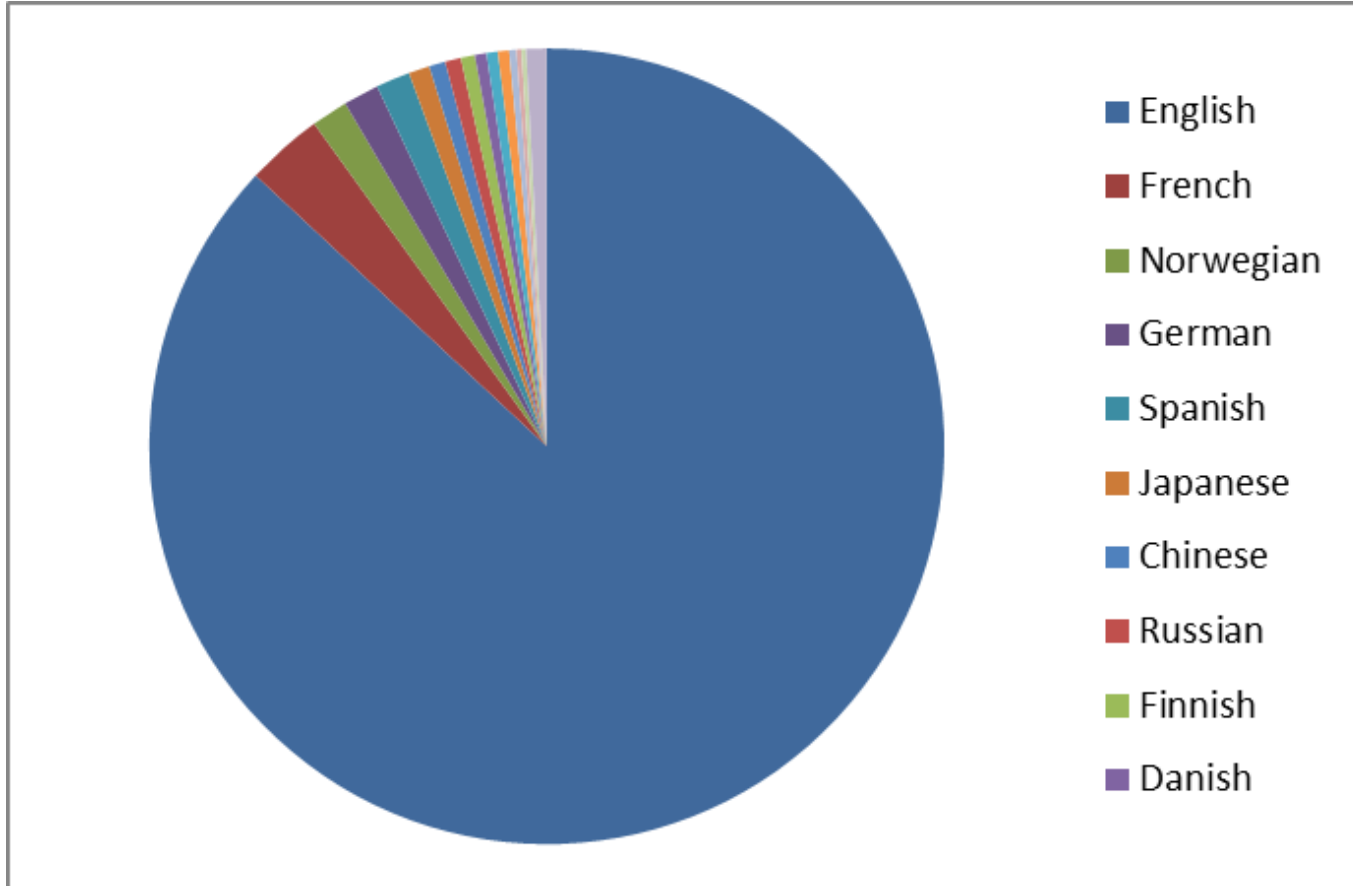
- Many data providers supply BASE with
 - records with metadata only
 - records where the documents have access restrictions
- This would be OK if these records could easily be identified and omitted from High North (and BASE)
 - dc:rights – should be used to indicate restrictions on access
 - How to identify records without any full text documents?
- We need to weed out many non-OA records from High North RD



Document types in High North RD



Languages in High North RD




High North: Home - Mozilla Firefox

File Edit View History Bookmarks Tools Help

HN RD High North: Home +

High North Research Documents



Send Feedback About

Explore the High North

High North Research Documents includes all freely available research documents thematically relevant to the high north. The service covers all subjects and many languages, and is free and open for all.

High North Research Documents ver. 1.1
Powered by [DSpace 1.7.2](#)
Data is provided by [BASE - Bielefeld Academic Search Engine](#)
© 2011-2012 [University of Tromsø Library](#)

Contains 151085 records
Last Updated: 3. July 2012

[Contact Us](#) | [Send Feedback](#)



<http://highnorth.uit.no>