

# A tail with no end

The Norwegian National Library pilot project "The High North" – an evaluation after one year

12 June 2008

---

*"... older content tends to score higher because it's had longer to accumulate incoming links. In other words, search inverts the usual priority of content; older is often the better."*

[Chris Anderson – "The Long Tail", 2006]

## 1. Background for the pilot project

Norwegian writers, book publishers, publishers of periodicals and other right holders and the National Library have the common goal of making literature and other copyrighted material accessible, and of giving the public insight into and knowledge of the diversity of the Norwegian cultural heritage.

In order to achieve this goal, the National Library and representatives of Norwegian right holders<sup>1</sup> have been cooperating for about a year in a pilot project in which Norwegian works are made accessible via the Internet within a specific topic – **The High North**. The aims of the pilot project include the gathering of experience with technological solutions, linking questions, making agreements, user behaviour etc., experience that can form the basis for possible future agreements on digital mediation of copyrighted material.

The terms of the cooperation are set out in a separate agreement, see [http://www.nb.no/content/download/1949/16107/version/2/file/avtale\\_GBR.pdf](http://www.nb.no/content/download/1949/16107/version/2/file/avtale_GBR.pdf)

The pilot was launched on April 24, 2007. After one year the participants wish to evaluate how users view the service. This short report sums up the results of a quantitative analysis of user behaviour as of April 24, 2008. A more detailed evaluation will be provided after the end of the pilot period in Oct / Nov 2008.

## 2. Functionality

### *Definitions*

In this report the following terms are defined:

A **user** is a person with a unique IP address<sup>2</sup>.

A **visit** occurs when a user opens a digital document for viewing.

A **viewing** of one or several book pages or periodical pages occurs when a user "visits" a work and studies the page(s) on a computer terminal. A "page" is displayed.

---

<sup>1</sup> Right holders are represented by the Norwegian Publishers' Association, the Norwegian Authors' Union, the Norwegian Non-Fiction Writers and Translators Association, the Norwegian Critics' Association and the rights organization LINO. By separate agreement with BOBO, GRAFILL, Forbundet Frie Fotografer and Norges Fotografforbund, LINO has been authorized to enter agreements on behalf of these organizations' members in the pilot project.

<sup>2</sup> Behind each IP address there may be several persons, since a company, organization etc. may operate using a single IP address.

### *Search and retrieval in NBdigital*

The main entrance to content in NBdigital is by using the National Library's search engine. From the search window the user can search among 3 million entries: web pages, catalogue records and more than 1 million digital objects.

The National Library's search is an integrated search which enables lookups in many different source databases, independent of type of material and subject category. The user searches the whole collection simultaneously, and so the lists of search results display content and information in a new way. At the same time, dynamic navigators are generated based on the metadata attached to the objects, enabling the user to refine the results using structured data.

### *Full text display of books and periodicals*

The books in NB digital are subjected to OCR and structure treatment, and metadata are added. The text resulting from the OCR, and its metadata, are indexed by the NB search engine.

So it is possible to search in the traditional way for title and author (metadata), and at the same time search for words and phrases in the content itself (free text). When following search results in the text, the user is shown the pages in the book where the phrase was found and can browse on from there. An automatic structural analysis is also performed, in which the list of contents, if there is one, is annotated, and page numbers in the book are verified so that the display of digital pages conforms to the actual pagination.

The user can turn pages in the work, go to a specific page and select between different qualities and display modes. The pages are displayed as jpg images of actual book pages. They appear as single pictures. Downloading of complete books or the full OCR text has not been enabled.

Users of the library system BIBSYS can move directly from the display of the book to placing an order through BIBSYS. Periodicals are scanned in full, but access is restricted so that users can read only articles that have been cleared through agreements.

### *High North material as a separate collection*

High North materials are tagged in the NB catalogue, enabling them to appear as a separate collection within the digital library. It is possible to restrict the search to this subset of the index.

### High North portal

When a search query gets results in material tagged as "High North" in the catalogue, or if the search term is found on a pre-defined list of High North related terms, a "super hit list" is activated, which is the entrance to the High North portal page <http://www.nb.no/highnorth/>. This page places the High North term in a wider context, offering news feeds, articles, presentations and a large collection of links to relevant High North resources all over the world.

### *Visibility*

The High North material lies open to external search engines, allowing the robots to find and index the content. This makes the material searchable from outside NBdigital.

Except for being covered by the press, broadcasters and trade publications, the service has not been advertised in TV commercials or any other form of paid marketing.

## 2. Numbers after one year of operation

The usage survey was based on 394 book titles and 230 articles from periodicals (a total of 634 titles).

Unique pages: 81,468 (71,425 book pages, 10,048 article pages).

Unique IP addresses: 2,482

332 of the users (13,4%) used both books and periodicals.

Visits: 3,798 (books) and 1,463 (articles).

Works visited: 308 (books) and 192 (articles), 78% and 83%

Total unique pages displayed: 23,014 (books) and 4,739 (articles)

Unique pages displayed: 10,535 (books) and 1,622 (articles)

Pages displayed per visit (average): 5.23 (books) and 2.02 (articles)

In 56% of the titles 1 - 20% of the content was displayed.

Usage by year of publication – based on 280 most displayed titles (sorted by IP address) for books and periodicals together:

<b>Year of publication *</b>	<b>0</b>	* defined as 2007. No works in the service
<b>1 - 2 years</b>	<b>8 %</b>	
<b>3 - 6 years</b>	<b>12 %</b>	
<b>7 - 11 years</b>	<b>17 %</b>	
<b>&gt; 11 years</b>	<b>62 %</b>	

Usage by topic and genre - based on 100 most displayed titles (sorted by IP address) for books and periodicals together :

<b>Fiction</b>	<b>9 %</b>	
<b>Nonfiction</b>	<b>91 %</b>	Broken down:
<b>(Cultural) history, biography</b>	<b>70 %</b>	
<b>Social studies, strategic studies</b>	<b>7 %</b>	
<b>Nature, outdoors, travel</b>	<b>18 %</b>	
<b>Language, literature</b>	<b>9 %</b>	

Number of works : number of pages displayed - based on the 100 most displayed titles (sorted by IP address) - for books and periodicals together :

<b>2 works</b>	<b>5 %</b>
<b>20 works</b>	<b>25 %</b>
<b>100 works</b>	<b>66 %</b>

The service was linked to several search engines in January 2008. Subsequently there was a measurable increase in the number of visits and pages displayed.

### 3. An evaluation

The usage survey presented above constitutes a purely numerical description. At the end of the pilot project we plan to perform a comprehensive evaluation which will include a qualitative analysis.

Comments to some of the main findings:

- The fact that 78% of the books and 83% of the articles were visited, documents that the service has been of interest to the users. The project group has not been able to find comparable surveys anywhere else that could serve as a relevant reference. There was one example concerning usage frequency: "Nearly three-quarters of sessions saw content viewed ..." <sup>3</sup> The survey quoted analyzes a project at University College London, and describes students' use of literature on the syllabus.
- The fact that 62% of the titles in the service that were visited, were first published 11 or more years ago, may signal that the internet has become a viable distribution channel for literature, especially for those titles that are hard to access in other ways. One can also see that there is a good distribution of interest among subject areas. The fact that such titles are made available, increases demand. And it is the niches that are credited. This phenomenon is often referred to as "the long tail", a term first invented by US writer Chris Anderson. The point is that the majority of products (e.g. books) available has approximately the same value as the titles most in demand, simply because there are so many of them and they are sold over a long period. Anderson says, "This is not just a quantitative change, but a qualitative one, too. Bringing niches within reach reveals latent demand for non-commercial content. Then, demand shifts towards the niches [...], creating a positive feedback loop that will transform entire industries – and the culture – for decades to come." <sup>4</sup> The numbers in the analysis at hand clearly seem to display a "tail effect".
- The number of users and the number of page displays increased considerably after the works were made searchable through the great search engines in January 2008. This conforms with surveys in other countries. One example: "Those people accessing via search engine were most likely to record more views in a session and were more likely to view textpages." <sup>5</sup>
- The web site for the service contains links to more than 200 sites / documents with content that is crucial to the High North theme. Since these works were born digital, they are not included in the index of digitized High North objects, and as such do not influence the numbers in the survey.
- The supply of books and periodicals in the service has not been constant, but growing. The most recent list of additions (approx. 100 titles) began to be added to the service in April 2008. Accordingly, they constitute a considerable portion of the material, while also being so newly added that their number of visits are negligible. Over time this will influence the average numbers for visits and usage.
- The other surveys mentioned above (notes 3 and 5), analyze two digitization projects at English and US academic institutions. The project team has studied these analyses, but found that only to a certain degree can they be used as a frame of reference; the difference in target groups, length of the project period and criteria for analysis all point this way. On the other hand, the trends and patterns in user behaviour documented there, support the probability of the results in the present analysis.

---

<sup>3</sup> David Nicholas et al.: "SuperBook", in the programme for the conference "Online Information 2007", London December 4 – 6, 2007, p. 50.

<sup>4</sup> Cf Chris Andersson: *The Long Tail. How Endless Choice is Creating Unlimited Demand*, Random House 2006, p. 26. The quote at the beginning of this report is from Anderson's article "Google and the Long Tail of Time", [http://thelongtail.com/the-long-tail/2006/04/google\\_and\\_the\\_.html](http://thelongtail.com/the-long-tail/2006/04/google_and_the_.html), p. 1.

<sup>5</sup> Nichols, op.cit., p. 54. See also David Nichols et al.: "What does usage data tell us about users?", in the programme for the conference "Online Information 2007", London December 4 – 6, 2007, p. 84

#### **4. The road ahead**

One of the goals of the pilot projects has been to gather experience which can give impulses to further cooperation on digital mediation of copyrighted material.

The survey here presented shows good user traffic in the service, and the material requested shows considerable distribution among subject areas, and an interesting distribution among years of publication / age of material. Since a portion of the content was only recently incorporated in the service, the project team has decided to extend the project period by 15 months. This will also allow for improved quality assurance of the results.

In parallel with the wish to refine the pilot project, the project team has initiated work on considering a broadening of the cooperation. Such a broadening will be detailed when the pilot project is evaluated at the end of October 2008.