Search Engine Optimization and implementation of Google Search Appliance in the Danish legal information system

By Søren Broberg Nielsen, Ministry of Justice, Denmark; Rasmus Lohals, NNIT A/S; Steffen Schalck, NNIT A/S and Nina Koch, Ministry of Justice, Denmark

Abstract

Part 1 of the paper will focus on Search engine optimization.

The paper will describe the demands and needs from the public to give general purpose search engines access to the data in the Danish legal information system, and the concerns such an access did present.

Our concerns were both of purely technical or operational character i.e. how would our system respond to a potential huge workload imposed by different crawlers, and concerns whether the different general purpose search engines page ranking systems would present a sound and true result list from a legal information point-of-view.

The paper will then describe the different methods we examined as preparation for opening the LIS to be indexed by other search engines, and as the last section in part 1 the final implementation of static meta data in the LIS, implementation of static and dynamic metadata on the documents in the LIS in accordance with the harvesting guidelines in the robots.txt.

Part 2 of the paper will report on the actual implementation of Google Search Appliance in retsinformation.dk

In the fall of 2009, two years after the new legal information system and the Official Journal online was launched, we made an online user survey, and the paper will report on the findings of the user survey. Especially the overall most demanded new feature — a Google-like search interface within the legal information system.

For the 25th anniversary of retsinformation.dk the Minister of Justice presented the implementation of Google Search Appliance to the end-users as a respond to the clear demand in the user survey. The paper will describe how we selected to use Google Search Appliance, what experiences we have heard of from others and what technical and operational demands Google imposes when you have a Google Search Appliance within you legal information system.

The technical implementation contains several different tasks, and the paper will discuss the different steps. First we made a proof of concept (PoC) in cooperation with a Google partner in Denmark to show the possibility to automatically load the GSA box with new and/or amended documents.

After the successful PoC we implemented a robust and operational method of loading the GSA within the production framework thus ensuring similarity of data in the different query engines. The last and most difficult task was how to tweak the PageRank™ result which came out of the Google Search Appliance, and the paper will describe the different approaches we took to overcome the default inadequate result delivered by the GSA box. Finally, the paper will state some of the shortcomings in the current version of Google Search Appliance.

Introduction

Retsinformation (translates directly into Legal Information) was established in 1985-1986 by the government. All primary and secondary legislation and every treaty that was in force on January 1, 1985, were incorporated, and nothing has been removed ever since. However, documents are marked as historical as legislation is repealed.

The first – and probably the most important – strategic decision made by the government was that data capturing was established as part of the process of issuing legislation. When a bill is passed by the Folketing (the Danish Parliament), the relevant minister/ministry is responsible for presenting it for the Royal Assent – thus becoming an act – for promulgating the act and since 1985 for publishing the act plus the metadata concerning the act in Retsinformation. Exactly the same goes for secondary legislation, so every civil servant in the central administration knows that if one issues delegated legislation or administrative orders, one's job is not done, until document and metadata are available in Retsinformation.

The second very important strategic decision concerning Retsinformation was made by the Folketing as an institution. It goes without saying that the Folketing was not bound by the governmental decision concerning legislation. However, the Folketing decided to take on a similar obligation concerning the legislative history behind the acts and has since 1985 uploaded the legislative material, the Hansard and its annexes.

From the very beginning the frontend was made for IBM 3270 terminals and later on terminal emulators, which could query the BRS database at the mainframe. But in August 1998 the government decided to open the new web based search interface on www.retsinfomation.dk and most importantly make it free of charge. However, the backend system was still running on the mainframe and the cost of operation varied with MIPS usage, which eventually led to a significant rise in the TCO of Retsinformation.

The last part of the mainframe was abandoned in September 2007, with the start of a completely new legal information system - Lex Dania production[1]. Lex Dania production is a wall-to-wall system, where drafting, proof-reading, introduction, passing, promulgation, publication, end-user access is supported by one single system, and today the frontend supports a standard multichannel strategy, i.e. web, web services for reuse and querying and mobile and tablet App's.

[1] Nina Koch: *Free Access to Legislation in Denmark: Advantages in Inter-institutional Cooperation - Design and Production, Law via the Internet Free Access, Quality of Information, Effectiveness of Rights, E.P.A.P. 2009*

Towards Search Engine Optimization for retsinformation.dk

Robots.txt and parliamentary questions

When retsinformation.dk was launched as the new frontend in September 2007 we had omitted to implement any controls and restrictions towards Web crawling in our database. This became very imminent shortly after as we received several complaints from Danish citizens who was concerned with their online privacy. Suddenly anybody could query their own name in general purpose search engines and find the act in Retsinformation which grants them Danish citizenship. We were not concerned with respect to the complaining citizens because in the preparatory works for the amendment of the act on the official journal introducing electronic promulgation this specific question had been thoroughly analysed[2]. However, we became aware of the unintended result in our frontend, and we had to urgently decide which solution we could implement without compromising the functionality on www.retsinformation.dk. We decided to implement the robots.txt which we had in our previous mainframe system. Thus, we completely excluded Web crawlers which respected robots.txt from all dynamic pages in our system, i.e. *all the legal information*. The reasons for this decision were:

- It was working in our old system.
- The workload generated by the Web crawlers was unknown and could potentially be very important as many documents in the database can be altered overnight. We have had such experiences on our mainframe based backend.
- Concerns about the search results created by the Google, Yahoo and others. Would they meet our standards as a provider of public legal information, as the ranking of the search result was outside our control?

Shortly after a discussion on a Danish tech forum[3] started regarding access to and reuse of public information, and suddenly someone posted their discontent with the robots.txt in the legal information system and claimed that we prevented public data from being liberated. Not only did the user post his discontent, he also opened a protest site[4] created to prove the possibility to index Retsinformation.

However, in October 2008 Thomas R. Bruce made a presentation[5] at the 9th International Conference "Law Via The Internet", and we realized that we were not alone with our concerns e.g. the relevance of the search results generated by search engines such as Google. So we left Florence with confidence - we had made the correct choice by acting prudently. Moreover, it also made us wonder whether it would be possible to use some of the tools in SEO on a legal information system. Hence, by the end of 2009 we had the funding to start the project.

The project was just in time as the Minister of Justice in April 2010 in the legal affairs committee[6] was asked why lay people could not find the legislation they were looking for on the internet. This gave us the

---

[2] Bill nr. 106 of December 14 2005 general remarks pt. 3.5

[3] The discussion can still be found on http://www.version2.dk/blog/hvordan-starter-man-en-graesrodbevaegelse-7059

[4] The protest site is still running http://retsinformation.w0.dk/Forms/R0200.aspx!

[5] Thomas R. Bruce: *Foundlings on the Cathedral Steps, Law via the Internet Free Access, Quality of Information, Effectiveness of Rights, E.P.A.P. 2009*

[6] The question and answer can be found on http://www.ft.dk/samling/20091/almdel/reu/bilag/451/825798/index.htm and http://www.ft.dk/samling/20091/almdel/REU/spm/995/index.htm

opportunity to explain the course of our actions in public, and in October 2010 when the answer was given to the legal affairs committee the project was already implemented.

Initial suggestion for optimizing Retsinformation for search engine indexing

In order to allow search engines to crawl and index Retsinformation we needed to open up the directions in the robots.txt. We looked into the options we had for using the standard and extended robots' vocabulary and into inclusion and exclusion of different search engines and controlling the way crawlers would behave.

In our opinion, it is important to discriminate between a commercial business which optimizes their web site to attract customers by scoring high in search engine results and thereby directing users to their web site, and a legal content provider as Retsinformation, where the only mission is to provide users with the relevant content they are searching for. We are only interested in attracting users, if we can indeed provide the content they are searching for. This distinction defined the guideline for the way we should restructure and design the web content of Retsinformation.

In order to ensure the most optimal indexing of content by search engines, we sought to apply recommended SEO changes to the existing content on Retsinformation. This included optimizing meta tags such as title, keywords and description tags, as well as restructuring document body content with regard to the semantic structure. Additionally, we only wanted the search engines to index documents, which have the value “In force = True”, i.e. automatically omitting documents not in force.

Implementing search engine optimizations in Retsinformation

The overall strategy for the SEO change implemented on Retsinformation was that we strived to make it as simple as possible, minimize risks and make the change highly configurable. This was motivated by the fact that we could not test the changes before going live, and so we would never know beforehand how the result would come out. As a consequence, we would make the changes as simple as possible, and evaluate the result afterwards to see if we had made adequate optimizations.

We decided to keep the robots.txt directions to a bare minimum and not to use any filters. It was stripped down as to allow all user-agents and allowing indexing of all content. This meant that we would leave it to each served page to decide whether it should be indexed and/or followed by search engines.

The robots.txt file after SEO:

```
User-agent: *    # applies to all robots
Allow: /         # Let whoever understands Allow index retsinformation.dk
```

The textual content of all legal documents in Retsinformation is stored as static html which is created on the time of publication. Analysing and restructuring of this existing content as to optimize it for search engine indexing was considered a costly and potentially error prone endeavour, that could compromise the data quality and keeping to the overall strategy this was left out of the implementation.

We were then left with finding an optimal redesign of the meta content in Retsinformation which is database driven and generated at runtime, thus making it easy and relatively free of risks when making configurable changes with regard to SEO. Summing up, it was decided to keep the html changes, to the title, robots, keywords and description meta tags.

The html already contained all of these meta elements, however, the content of each meta element was not optimized for ensuring high page rank in search engine results. In order to get the indexing right it is generally recommended that central words describing the content should be positioned in the beginning of a given meta tag. To exemplify the changes to the <title> tag we made, we will show below the title meta tag of the act of social service before and after the change.

Before the change:

```
<title>retsinformation.dk – LBK nr 810 19/07/2012</title>
```

This only tells us that we are on the Retsinformation web site and gives the short term name for the act and does not really give a good description of the content being served.

And after the change was implemented:

```
<title>Serviceloven – Bekendtgørelse af lov om social service – retsinformation.dk</title>
```

Where "Bekendtgørelse af lov om social service" is the title of the act and "Serviceloven" is the popular title. This title tag is much more search engine friendly.

Instead we moved the short term name into the description meta tag together with the responsible ministry ("Social- og Integrationsministeriet"):

```
<meta name="description" content="LBK nr 810 af 19/07/2012 - Bekendtgørelse af lov om social service - Social- og Integrationsministeriet">
```

Furthermore, for the dynamic pages containing the legal documents we set the robots' directions depending on document status ("in force" versus "not in force") and document type, to either noindex or index. We only allow primary legislation, secondary legislation and international treaties in force to be indexed.

All the robots' directions, keywords and description meta tag for the static pages on Retsinformation are stored in a database and is loaded at runtime and cached. This way we have the flexibility of changing these after go-live, should we find it necessary. As an example of a static page in Retsinformation we will show below the resulting meta tags for the page showing the list of Danish primary legislation in force ("Gældende danske love")[7]:

```
<meta name="robots" content="index, follow"/>
<meta name="description" content="Oversigt over alle gældende love og lovbekendtgørelser." />
<meta name="keywords" content="hovedlov, hovedlove, lov, love, lovbekendtgørelse, lovbekendtgørelser" />
<title>Gældende danske love - retsinformation.dk</title>
```

In conclusion, we have only made simple, configurable changes to the existing Retsinformation content avoiding implementing changes that could potentially compromise data quality. All the directions and meta information optimized for search engine indexing are injected dynamically into each page at runtime. For static pages the information is configured via a database, and for the dynamic pages, i.e. the actual legal

[7] https://www.retsinformation.dk/Forms/R0210.aspx

information documents, search engine directions and meta data are calculated based on current document status and document type. We refrained from doing any advanced configuration of the robots.txt and until now it has been proven successful, with no noticeable performance hit incurred by letting all search crawlers index the site as they see fit.

Results of SEO

We were of course curious about what the results would be on the internet. The go-live took place September 6, 2010 shortly before midnight, and we were looking for the Web crawlers to index the database. On September 21, 2010 queries in google.dk and other search engines were successful. We were very pleased with the results; queries about general terms such as “lov” [act] and “bekendtgørelse” [statutory order] made first in the result list, and queries about specific but very broad terms e.g. “social pension” did retrieve the act in force among the first 5 results. We only discovered one mishap, if you were looking for act in plural -“acts”, Retsinformation would only be placed in second or third spot. Acts in Danish is “love”, so the first spot was a dating site!

The next major test of the project was whether the SEO would direct users to the right document even if the document had changed in the legal information system e.g. would the consolidated act be replaced by the new consolidated act? Luckily we had the chance to test this immediately as a new consolidated version of the act on social service was published in Retsinformation on September 21. For the next 12 days the query "act social service" returned the repealed version of the act but then it returned the correct version of the act. This was very satisfying.

When the bill on electronic promulgation was passed in 2006[8] the Danish Parliament stated that it should be completely anonymous to use the official journal online. The Danish state should have no knowledge about whatever legal information citizens were searching for, hence many of the tools normally used in web analysis are not applicable to Retsinformation. We do not track the user interaction on our frontend, the origin of the users or their starting point, thus we have no specific statistics about the impact of the SEO. The numbers we do have, do not say anything specific, but daily average unique visitors was in February 2010 ca. 16.000 and in February 2011 ca. 20.500. Even though this is a significant increase in the use of Retsinformation we cannot conclude that is due to the SEO-project. This is mainly because we have seen a general increase in the usage of Retsinformation since September 2007, so the trend was only carrying on. But the trend has not been as strong in 2011 and 2012.

Google-like search interface

User survey

As mentioned in the first part of the paper the Danish Parliament decided in 2006 that the official journal should no longer be published on paper but on the web. To meet this demand we created a new production system and new frontends both for the official journal online (lovtidende.dk) and the legal information system (Retsinformation). The legal information system had go-live on September 24, 2007, and lovtidende.dk had go-live on January 1, 2008 although it was fully functional and running from September 24. We only needed the last quarter of 2007 to prove to ourselves, our minister and the parliament that it was working.

[8] See note 1

After one year of electronic promulgation the Ministry of Justice requested a survey of both end user systems, and in Nov-Dec 2009 we issued an online questionnaire[9]. During this time 1.284 respondents completed the questionnaire, and 97% of the users have used Retsinformation and 37% the official journal online.

The questionnaire gave the following information:

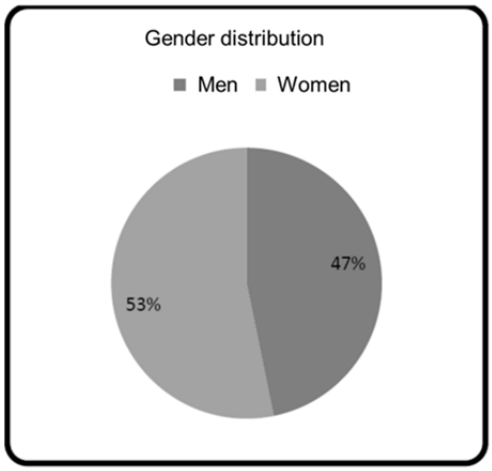


Figure 1 Gender distribution

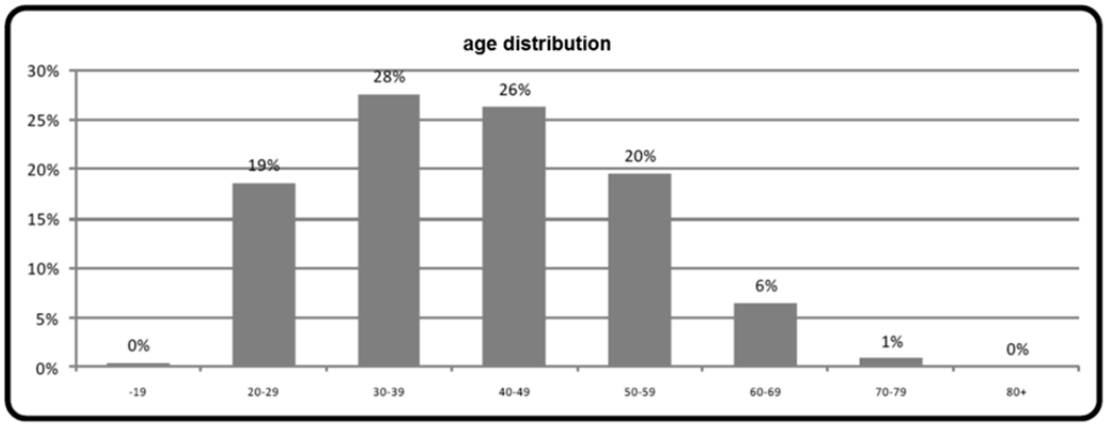


Figure 2 Age distribution

[9] Online questionnaire retsinformation.dk lovtidende.dk, Department of Civil Affairs December 21, 2009 by Userneeds A/S

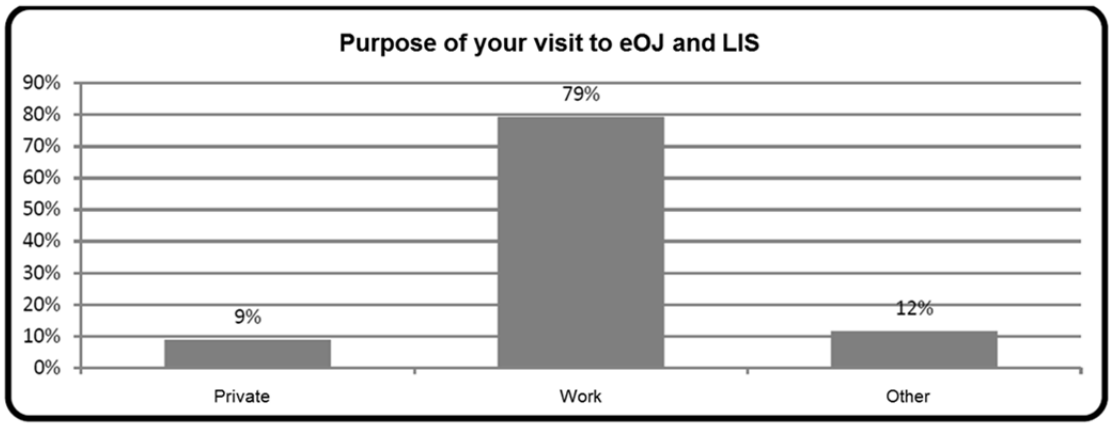


Figure 3 Purpose of usage

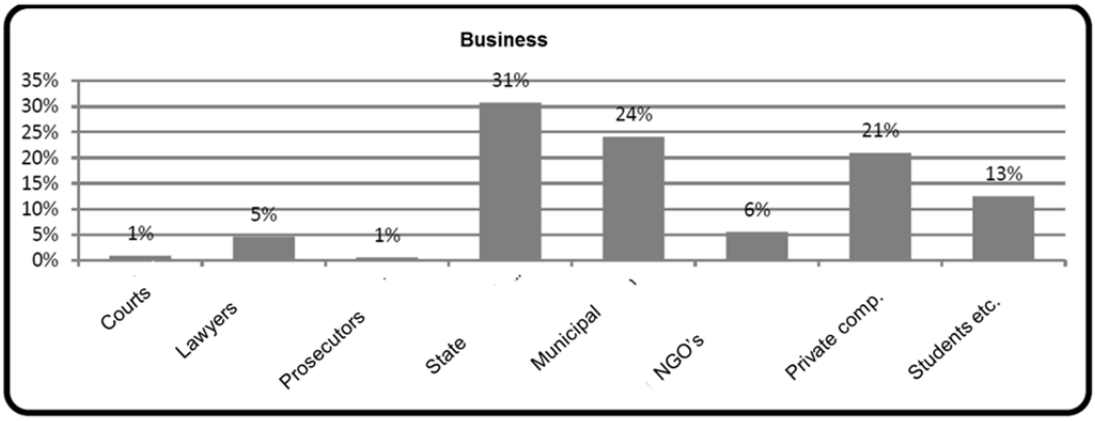
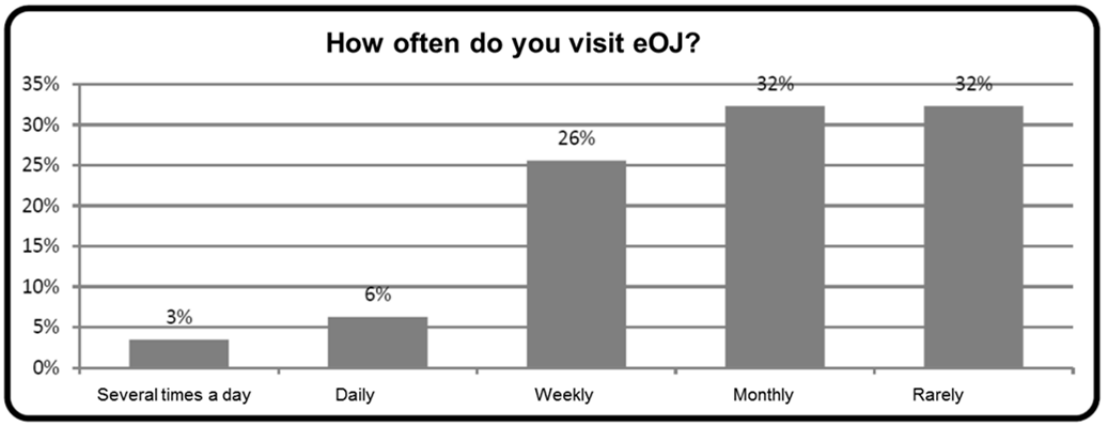


Figure 4 Line of business



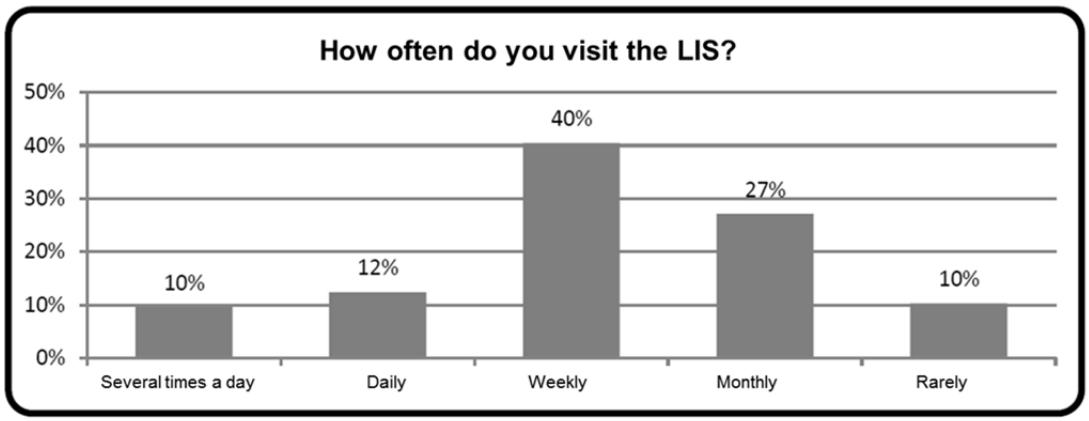


Figure 5 Usage of eOJ and Retsinformation

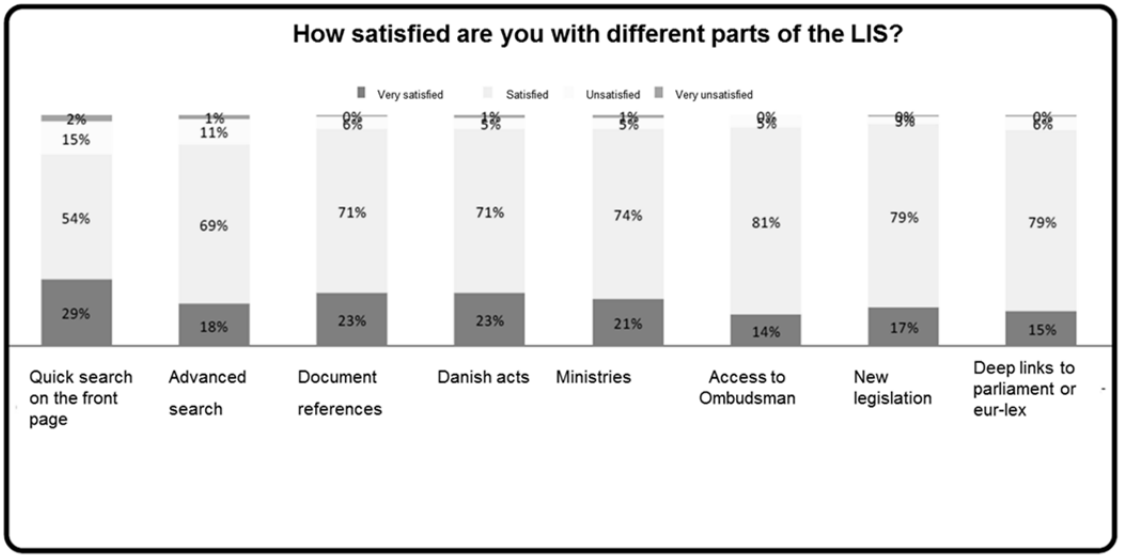


Figure 6 Quality of Retsinformation

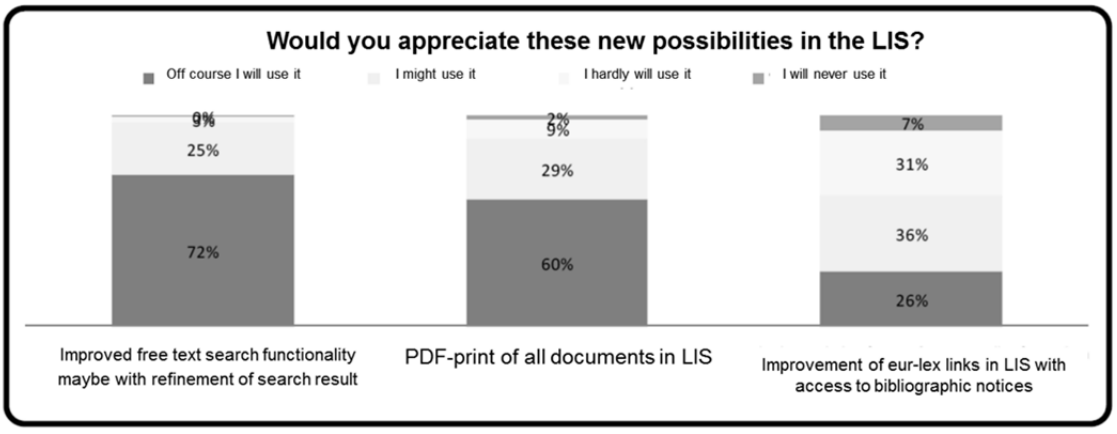


Figure 7 User demands

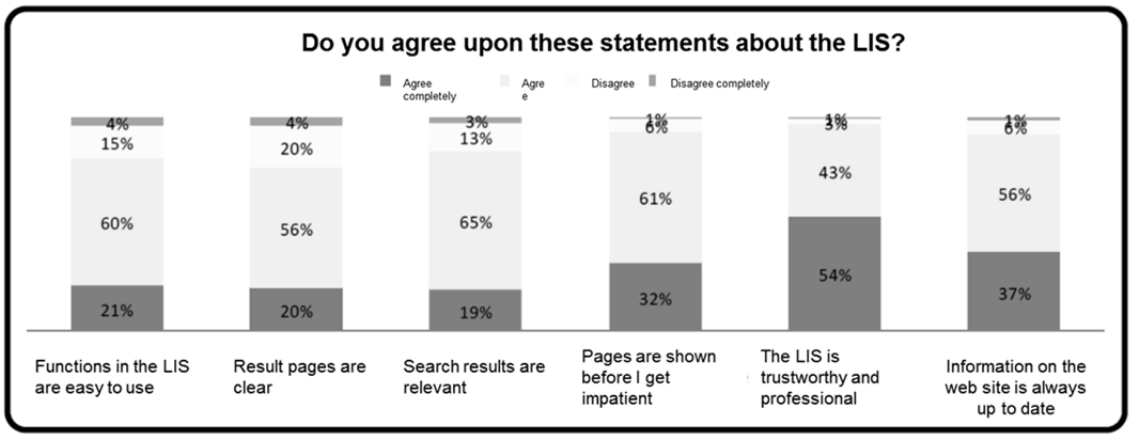


Figure 8 User evaluation

Generally speaking, the users rank our systems as very professional and trustworthy, and the usage is mainly related to work. This was not surprising as the decrease in subscribers of the paper official journal started years before the implementation of the electronic official journal, and the subscribers chose to use Retsinformation as a trustworthy alternative. From the statistics of the frontend system workload we also knew that usage is concentrated on weekdays from 08h00m to 17h00m.

More interesting were of course the concerns of our users, and most significantly was the dissatisfaction by all of the users regardless of professionalism with the search functionality. Search functionality is core business for a legal information system, and our users were all demanding a Google like search interface.

Specification for a new query method

With that result of the user survey we had to do something. Fortunately, Retsinformation was going to have its silver jubilee in February 2011, so the Minister of Justice decided to celebrate this event by responding to the user demand. Hence, the funding for a new search interface was established.

From the European Forum of Official Gazettes we knew that the UK had an early implementation of the Google Search Appliance (GSA) in legislation.gov.uk, and we could use some of the lessons learned by our UK colleagues:

- The Google Search Appliance was not satisfactory when used as search engine for structured search in metadata.
- The creation of a GSA index should be created as an integrated part of production flow untouched by human interaction.

Furthermore, we had learned from the SEO project that the quality of the search results and relevance were vital. So when the project started it was clear that a “Google like” search interface incorporated in Retsinformation should only provide results which could meet our standards for relevancy e.g. a consolidated act prevails guidelines, or a recent document prevails an older.

Proof of concept

Before doing a PoC Civilstyrelsen carried out an analysis of various relevant search engines together with NNIT, since establishing a “Google like” search interface did not necessarily include the use of Google

products. The aspects investigated were capacity, the technical platform, the interface model, usage on Retsinformation, ease of implementation, customization and not unimportant, the licensing model and price. Even though, there were several candidates among the renowned vendors of search engines; GSA was selected for several reasons.

- Installation of the GSA boxes was basically plug and play. They come as pre-installed boxes, ready to be inserted in a rack.
- The search algorithms are more or less the same as the ones on Google on the internet.
- The changes to Retsinformation were minor, as the search could be performed through web services. This ensured reuse of existing layout and functionality.
- Implementation of search suggestions were straight forward.
- Options for biasing the search result through an interface on the GSA box. Can be done run time for fine tuning the search results.

Civilstyrelsen together with NNIT and subcontractor Convergens, an experienced Google partner, decided to do a Proof of Concept before deciding upon how to go forth incorporating GSA technology in Retsinformation.

The goal of the PoC was to show that it was possible to index the legal documents in Retsinformation, and provide a simple Google-like search interface and a satisfying search experience bearing in mind the standards for relevancy laid out by Civilstyrelsen. A sub goal was to identify whichever limitations the technology may have that could potentially affect an eventual full scale implementation project.

The GSA is a middleware search technology consisting of server hardware and software. It is highly configurable and can crawl websites and fileshares or can be fed documents for indexing. It is able to index a vast number of content types and will provide search results using the ranking algorithm found in Google on the internet. The default ranking can be biased using meta data biasing and source biasing and the indexed content can be structured into collections. Finally, the GSA provides a range of features such as search suggestions, stemming list configuration, self-learning score, search correction suggestions and search statistics.

Based on how we envisioned the GSA should be used in a production setup, we did a test setup using a test GSA server, feeding all existing html for all the legal documents in Retsinformation and did some initial testing on the ranking results.

Soon, it became apparent that the default unbiased search results produced by the GSA did not meet the standards for relevancy defined by Civilstyrelsen. We attempted to manipulate the ranking results using either meta data biasing, source biasing or a combination of both. Though we did manage to affect the ranking results to some extent, it did not come close to the standards we were looking for.

We then tried grouping the fed documents into collections as a solution. GSA provides a means to group specific URL patterns into collections, which then enables searches to target a specific collection and only return results within this collection. The documents were divided into three collections: “Primary legislation in force”, “Secondary legislation in force” and “Other” and now a composite search could be conceived with

results from Primary legislation in force collections preceding Secondary legislation in force which then again preceded the rest of the documents (the “Other” collection).

Using this approach, we managed to generate a satisfying ranking result with a Google-like one field search experience. The unified search result presented to the user is actually a composite search consisting of three searches, one for each of the collection, which is then post-processed and combined, and this is completely transparent to the end user.

During the PoC we did stumble upon some limitations using the GSA. The GSA imposes limitations on the size of each document it indexes. The total content of a document can only be 2.5 MB. The document will be indexed, but only the first 2.5MB of content will be included in the index. This limitation was not readily documented, but was something we discovered empirically. Afterwards Google have confirmed that there is indeed such a limitation in the GSA. This limitation was a real deal breaker as more than 60 documents in Retsinformation are much larger than 2.5 MB. We worked around this issue by stripping the documents used for the feed of all white space and html mark up. This way we succeeded in reducing the total content below 2.5 MB for the majority of documents currently in Retsinformation with only a handful exceeding this limit. Another limitation not well documented was that phrase searches (i.e. “this is a phrased search”) only works within the first 1.8 MB of a document. Searching for a phrase that occurs in the latter part of the document results in no hit for that document. This is a limitation that we did not found a solution for. One last peculiarity we found was that the GSA servers are not able to return the actual hit results beyond search result no. 1000. It will only tell you how many in total there are. However, it would seem unlikely that one would be interested in the search result ranked in 1001, and this is a limitation that we are willing to accept.

Implementation of the GSA search in Retsinformation

Based on the result from the POC it was decided to begin implementation of GSA in Retsinformation.

The software system comprising the backend document handling system (Lex Dania), the publication engine and the end user publicly available web frontend (including Retsinformation) is a custom build .NET application complex. Everything has been designed with modular style architecture, with well-defined interfaces and modules with separation of concern, configurability and change in mind. Introducing a new system component such as GSA can be done relatively easy and with minimal risk.

The implementation consisted of 4 parts: 1. Setup the GSA servers and configure them to our needs, 2. Build an automated feed job that could feed documents to the GSA as they are publicised 3. Integrate this job into the existing publication engine, and 4. Design and build a new web search interface into the existing retsinformation.dk web site.

In production we use two GSA servers (GB-7007) setup in a mirrored replication setup; the index is replicated and always in sync. This provides a load balanced setup, with manual fail-over possibilities. Based on the PoC we configured three collections and made only one meta data biasing configuration for a certain document type. We also set up stemming lists for the most popular search terms (the main laws and acts) so searches for “social service” also would return hits on “social services”.

The publication engine was the main worry because it is mission critical that publications do not fail, and this engine is basically the heart of the application complex. It consists of a collection of independent

applications each being responsible for a single task during the publication process. The tasks are configured, scheduled, bundled and communicate with each other through a database. So we basically just had to build a new discrete GSA feed application, plug it in to this modular framework, and let the existing publication tasks communicate which documents should be fed to the GSA and which documents should have their meta data updated. Figure 9 illustrates this setup.

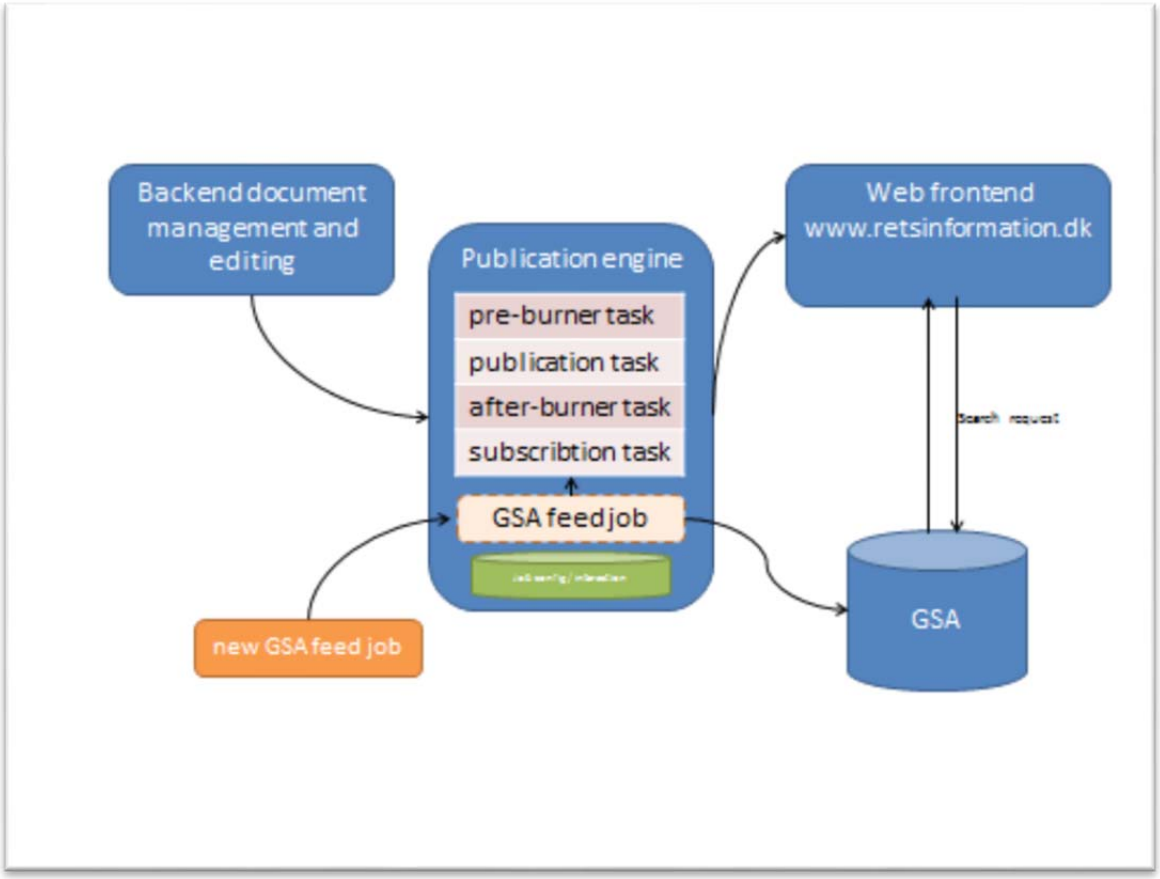


Figure 9 High level overview of Lex Dania production

Having built and configured a feed job, we were left with designing a search interface. The design was really an extension of the existing search web interface. Now the user have two tabs in the main search window ”Simple search” and “Field search” (figure 10), and using a cookie we remember which tab was last used should the user return to the web site. We use the search suggestion feature, and the search correction feature (ie. “Did you mean?”) provided by the GSA as illustrated in figure 11.

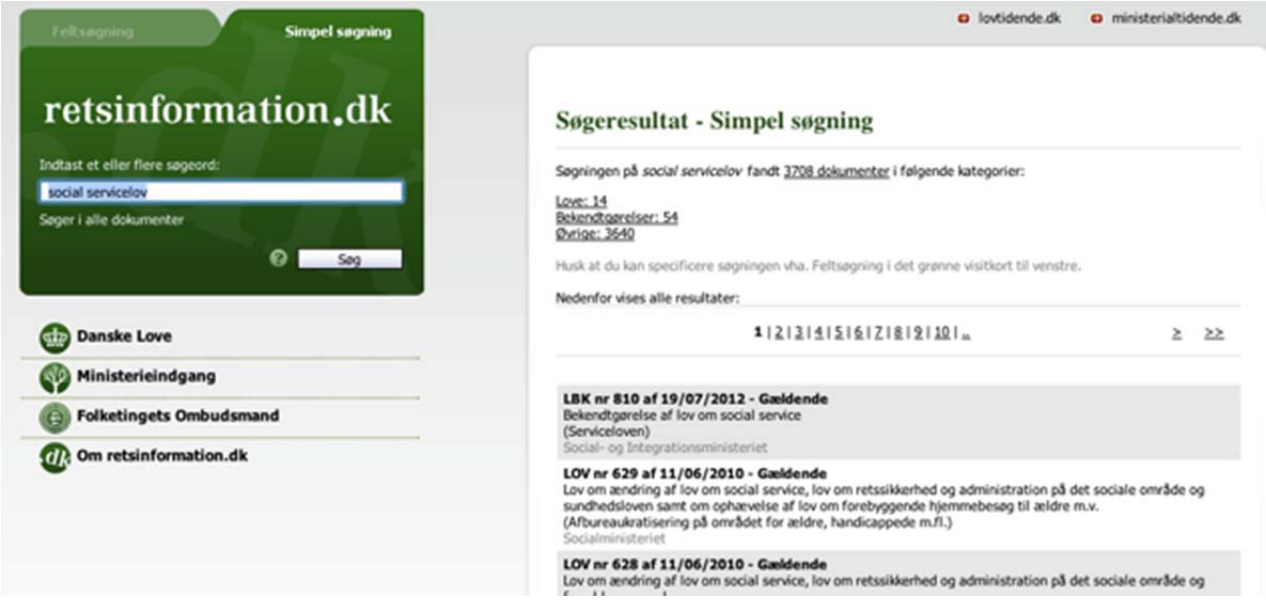
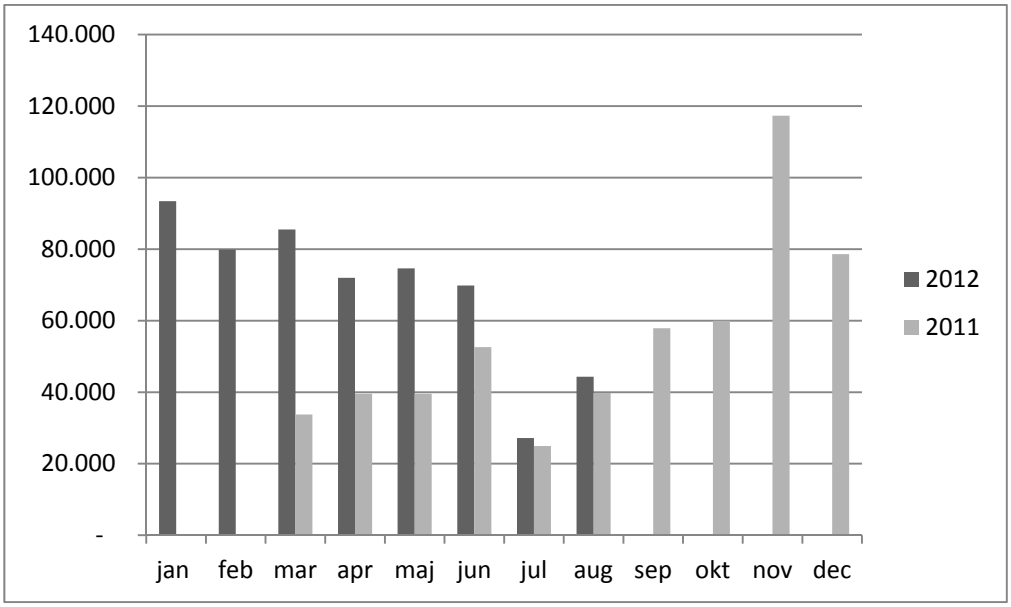


Figure 10 The web search interface

Figure 11 Illustrating search suggestion and search correction provided by the GSA

Outcome of Google Search Appliance

After the silver jubilee go-live we have been tracking the usage of the GSA, and as the chart below shows the usage has been stabilized around 75.000 monthly queries, with smaller usage during the holiday season. The abnormality in November 2011 is due to the usability study made at the University of Copenhagen.



Several mails in our help desk have shown that our end users have been very pleased with the new search interface, but it has also revealed some surprising information. Users that we would normally rank as expert users e.g. legal drafters and even law school professors, praise the simple search.

From the data used to create the suggestion list we can also see some very peculiar queries. Users query in simple search with typical metadata e.g. Nr. Year, Date, Doctype, or they use standard phrases e.g. "act on", "amending act of".

In November 2011 a class of Associated Professor Ralf Molich at the University of Copenhagen made a usability study of Retsinformation. The purpose of the study was to teach students about usability and the means to study it, and Retsinformation was only the object not the subject. Nevertheless it was interesting and fun.

The main findings of the students were on the positive side; speed and GSA implementation - on the negative side; very difficult to understand a legal information system and poor online help.