The Open Library System — Re-invented, Implemented and Working

Jens Hofman Hansen

Special Consultant, IT, State and University Library, Victor Albecks Vej 1,
DK-8000 Århus C, Denmark,
jhh@statsbiblioteket.dk

Arne Sørensen

Head of IT, State and University Library, Victor Albecks Vej 1,
DK-8000 Århus C, Denmark

Abstract

A highly modularised and open system is described in this paper, and its ability to support the library's strategy. By focusing on a single problem, it is possible to solve it in a qualitatively or quantitatively different way that is dictated by business vision and user needs. This results in a highly innovative library system when seen as a whole.

Key Words: openness; library system; innovation; business strategy

In this paper we will present four essential elements of our library system that are completely modularised and replaceable. During the past few years, each part has undergone several adjustments and has taken a few huge developmental leaps. Together they support what we want to do as a library and what we want to deliver to our users. The individual elements and their interaction define the digital part of the State and University Library, Denmark (Figure 1).

The User Interface Reinvented

In 2005 the State and University Library accepted the fact that Google had become so dominant that the search engine set the standard for the way

Fig. 1: *This simplified model illustrates how we define our library system as a whole. It consists of separate replaceable modules. New modules can be added independently of the existing modules, and each module can be replaced with a new one.*
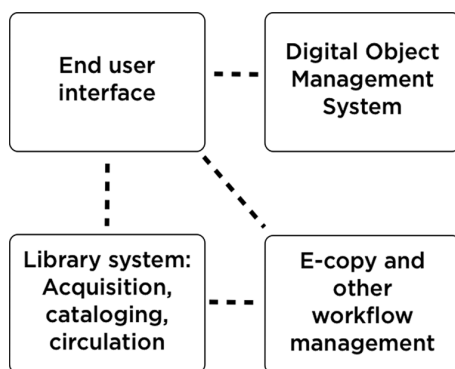
users find information on the web: they expect a single search field that gives access to all information irrespective of the collection it may originally come from or be in, whether it is physical or digital, whether they are articles, e-books, images or movies. Our existing search interface was anything but a Google-like interface: it had multiple search fields, radio buttons and drop-down menus, several search pages and no real relevancy algorithm at all.

When we accepted the users' search behaviour, we also decided to shift our focus from trying to teach users how to use *our* search interface into developing a search interface that was able to push information to our users on their premises. Certain information was never discovered by our users even though we knew that it was very valuable to them. Since we could not find a commercial system that was able to present our data in a suitable way for our focused users, the process of developing a simple but powerful search interface started.

Today, the new search technology behind the search interface is just one of many modules in an open library system that has been re-invented in most of its elements. By developing the new search interface as a separate module concentrating only on the search part, we got exactly what we wanted without spending development money on the elements that already worked fine. Underneath the new user interface, we kept the elements of the library system that handle physical material — the acquisition, the cataloguing, and the lending processes.

The main principles of our new search interface are:

Integrated Search and Sorting by Relevance

Users do not concern themselves with technicalities such as databases and data sources; first and foremost, they want some relevant resources. This is why we started to search across all the different data sources and to present results for the user in a single search result sorted by relevance. The integrated indexing of the different data sources into one index means that users in most cases are not even aware that the system is doing something completely new and different.

Post-filtering, not Pre-filtering

Much of what the users originally used to specify in the advanced search interface, without being quite able to predict the outcome, we now present within the search results as automatically generated clusters of interesting information. Clusters are also known as navigators or facets and have become quite conventional today — Google now presents navigators in the left-hand menu by default and so do several other widely used search-focused services.

Clusters can be used in searches where the user is trying to find something specific, or in searches where the user does not exactly know what the goal is, i.e., the user is looking for inspiration. Clustering is an example of a feature in a search system that takes the user by the hand and presents relevant suggestions during the search process, instead of imposing the burden of refining the search criteria on the user in the beginning of the search process (Figure 2).

Extra Information that Adds Value and Meaning

A library's greatest desire is to deliver relevant material of high quality to its users. Therefore, it is unfortunate if our users do not find all the quality material that is available on isolated library-driven websites.

In our new search system, valuable resources from several Danish library websites are shown in situations where these information fragments are most relevant. Since the introduction of the new search system, we have focused on adding value to the presentation of books and other materials. Today, books are presented not only with basic metadata, but often also the book cover, the

Fig. 2: Clusters, 'Did you mean?', availability, different types of ranking and sorting, a resource basket and other interface features are aspects we now control 100%. Being independent of the library system, we can accommodate what the users need and want with regardless of the update frequency of the traditional library system.
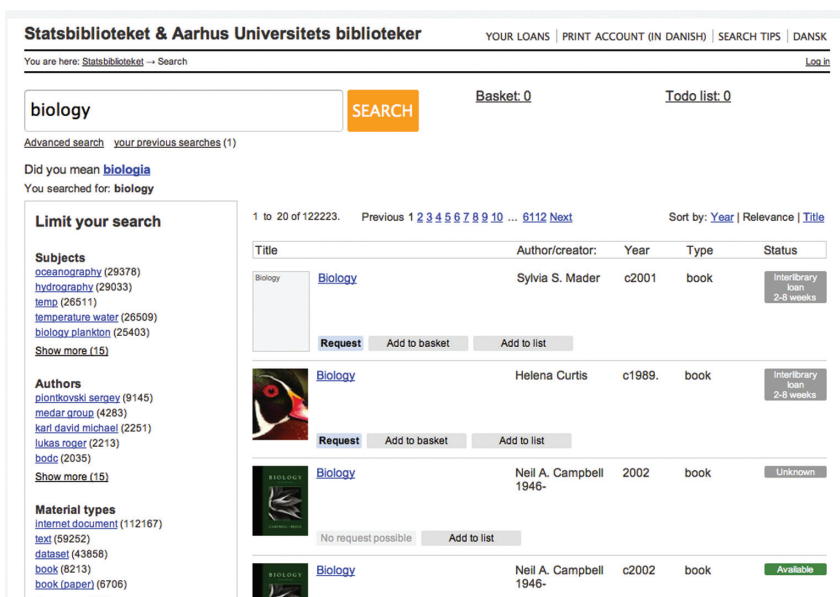

table of contents, an abstract and a link to an electronic version. This increases the user's ability to identify whether the book is the right one and gives value even before the user has the book in his hand.

With the new search system, we control the user interface completely and it is really just up to our own prioritisation what we want to do with it. On the basis of user studies and trends in user experience, we have been able to implement and experiment with features such as availability shown directly in the search results, 'Did you mean?' spelling suggestions, a resource basket, 'Other users who borrowed this, also borrowed …' and several other features (Figure 3).

We believe that our rapid development and radical new design have not only improved our own library business but also inspired vendors and other libraries to move their search user experience forward. We have noticed that

Fig. 3: Today, information about a book is more than just a simple index card: a link to the online edition, an abstract, the table of contents and similar items are all information that enriches the basic data about the book.

features we pioneered have now been incorporated into commercial systems as standard features. As another way of spreading our ideas and giving value to the rest of the world, we have also chosen to open source what we have developed. Today, our search technology is called Summa[1], and since it is a refinement and further development of the open source search engine library Lucene[2], we have found it suitable to open source Summa for everyone who wants to download it, use it and perhaps improve it.

The improvement in the user experience seems to be reflected in the number of materials borrowed by our users. As we have improved the interface and search experience in several iterations, we have also seen a steady increase in our lending and download statistics. It is always difficult to say what is actually the cause of the effect. However, we cannot really attribute the positive development to anything but our focused improvements. Anyway it is very motivating to see a rising tendency that stands in contrast to the falling lending figures of many libraries (Figure 4).

Fig. 4: Loans decreased during the period in which our system was down for a major update. But the number of loans clearly increases year by year.
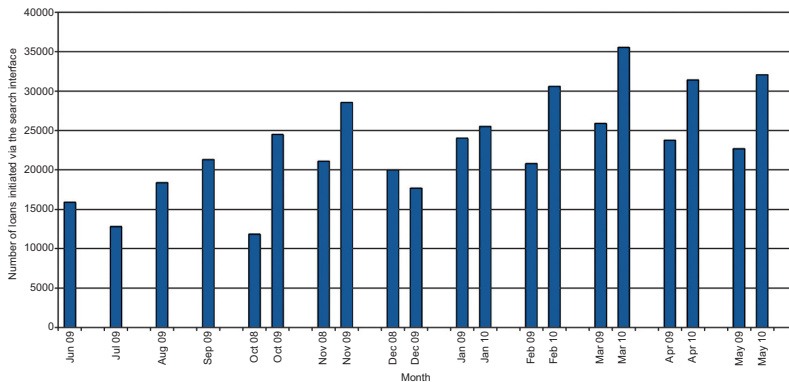
Replacing Important Elements of the Library System

The value of having a modular library system becomes more and more apparent: in 2010 we had to switch to a new library system because the old system (Horizon from SirsiDynix) was being discontinued. Having developed and tuned our own search interface, our demands for a replacement of this part

of the system were very clear: we simply needed a system that could handle information about our materials and users behind the scenes. The user search interface was not a parameter at all since we wanted to plug in the new system below our own search interface.

It was actually possible to change the system without our users being able to detect any changes after the implementation. However, since a replacement like this is really essential and almost unavoidably includes downtime, we carried out the change-over during the Christmas season in order to affect as few users as possible.

We have now replaced both the user interface and the back-end of our system, and the work has been done in completely independent processes. The strategic benefit and value have been clear both times: by having a modular library system it is possible to improve individual elements for the exact purpose that we have identified and prioritised — we have not been tied to one vendor that may have a different business and user focus from us.

Significantly Improving Workflows

Another benefit of the open library system is that we have been able to improve other elements of the system and dramatically reduce the need for manual work while at the same time enhancing the user experience and the feeling of receiving great service when using the library.

Since we do not have every journal from every year in electronic format, a significant number of orders for articles are still handled in a traditional way in our library: when users order a specific article, we basically have to find the physical journal in order to be able to deliver the article to the user. Today, this labour-intensive work is changing dramatically thanks to a module of our library system that represents our concept of digitisation-on-demand and reusing already digitised materials.

From Analogue to Digital

Formerly, users ordered an article through our web interface and we received an e-mail with the request. It was a manual process to keep track of the orders

and to find the actual articles. When an article was found, it was photocopied and sent to the user's home address if he did not want to come and get it at the library. The former process was actually quite a good service when seen from the user's perspective (except for the slowness and inflexibility of any analogue distribution system), but also a very expensive one when seen from the library's perspective. With this old workflow, there was no digitisation and no recycling of located articles. This also meant that the cost was constant and only rose if the article service got more popular.

Consequently, we looked at our workflow and identified where we could save manual work and better serve the users: by delivering the articles digitally, the users could get the articles by e-mail right after we had processed the order — a clear benefit when seen from a user perspective. But in building a reservoir of scanned articles with metadata, we have also started to reuse the scans and and no longer had to find the same articles again and again in our stacks. After all, some articles are more popular than others, which means that we process a lot of identical orders over time.

We have introduced the system (we call it 'e-copy') and our users now get scanned articles sent by e-mail. Shortly after the introduction, our recycle percentage was 12% of all ordered articles. This may seem low, but it is significant when taking into consideration that this represents more than 5,000 orders that would otherwise be processed manually.

When users enter all the data themselves, orders tend to include typos and at the same time we cannot expect users to type in the ISSN number of the journal they are looking for. The ISSN number is the key to matching users' orders to existing article scans, and since we wanted to increase our recycle percentage we integrated our system with the nationwide service called Bibliotek. dk.[3] Orders from Bibliotek.dk always include an ISSN number, since users can only order an article after they have found the journal with its metadata on the website. This means that we generally get orders that have a very high data quality.

Improving the Recycle Percentage

After we connected our processing system to the orders from Bibliotek.dk, our recycle percentage rose to 33%. With more than 19,000 deliveries handled automatically that would otherwise have to be processed manually, it is pretty

obvious that it makes sense to have isolated this part of our library system as a separate, improvable module. It may not be real innovation, but it certainly is an example of massive rationalisation combined with an improvement of the user experience (Figure 5).

Fig. 5: The recycle percentage seems to have stabilised at 33%, but it may slowly go up as the total number of scanned articles increases.

Year	Automatically delivered journal article requests	Journal articles scanned	Recycle percentage
2007	17	939	2%
2008	5,039	40,845	12%
2009	19,031	57,380	33%
2010	9,144	27,616	33%

Digitisation also presents problems: we are not allowed to distribute some articles digitally at all because of copyright issues. But we have developed a workflow that still allows us to save time by digitisation: we scan and keep all articles digitally, but the delivery is up to the user: he can come to the library and print the article himself by using his personal login or he can ask for the article to be mailed to him by snail mail for a small fee. We currently have a new agreement underway with the Danish organisation Copy-Dan[4] that distributes compensation to copyright holders. The new deal will cover even more journals and magazines that can be distributed digitally to all citizens.

We have also widened the scope for the system to include orders of TV and radio broadcasts. Orders are processed in the same system as articles. TV and radio material is not yet delivered digitally due to copyright issues. At the moment, we distribute it on burned DVD's. The workflow is basically the same as for articles and we benefit from the workflow support built into the system already.

A Heavy Building Block: One Petabyte[5] of data

As a national library, it becomes really clear that we need an open library system when it comes to preservation and access to the enormous amounts of digital data representing our digital and digitised cultural heritage.

The diversity of our digital data is enormous: we record all Danish TV transmissions, all national radio transmissions, we harvest the entire Danish internet four times a year[6], we digitise LPs, old wax cylinders with recorded sound, radio news text manuscripts, newspapers and many other materials.

Digitising and preserving such a diversity of data is anything but simple, and it is crucial that the system we use for this task is replaceable with another system. In other words: when choosing or building a system for preserving digital data in these amounts, the first thing you need to get clear is your exit strategy. How do you get data out of the system if it no longer meets your needs? How do you migrate your data if they are no longer usable in their present format because of changes in the hardware or software available to the general public? These are questions that we need to be able to answer and incorporate into the system we use.

When you start looking into what system you need for preserving all your digital data at a library of this size, you soon find out that you need a highly generic system that can contain and preserve basically all types of digital information; also it must scale to enormous amounts of data. We have not been able to find a commercial system which met our requirements; when we were looking for potential systems, we learned that some systems were probably technically suited for the purpose, but that the way data were stored was black-boxed to a great degree, which also immediately disqualified the systems. Transparency is essential when we want to ensure that we can preserve data for the next 100 or 1,000 years.

Consequently, we have developed our own system called DOMS (Digital Object Management System) that preserves very different kinds of data. To give you an impression: it contains about half a million TV transmissions weighing in at nearly half a petabyte of data (500 terabytes). 300 million digitised radio news manuscripts have also been archived — a tremendous amount of valuable information although it does not take up many gigabytes. At the lower end of the data scale you find very small collections of e.g. 10 special tapes of sound that have been digitised for preservation or 500 wax cylinders that are very fragile and can only be played rarely in their physical form but endlessly in their present digitised form.

In total, we currently store about 1 petabyte of data when we combine the data from both our DOMS and our Internet archive. The data are duplicated

in three separate archives, which actually brings the amount of data to 3 petabytes.

One System for all Data

The whole concept of DOMS is precisely to have *one* system for all data. In the past, many small dedicated websites were developed for individual digitisation projects, websites that tended to die after a few years because they were programmed in programming languages that were not security updated any more or run on server software that had similar problems. Most of these websites were built to present the digitised information to the users, not to preserve the data.

By storing all data in the same system, it is possible for us to continuously monitor if data need to be migrated to other formats. We always know what formats our data are actually in and we can react if a format is becoming difficult to read.

This is also part of the reason why our Internet archive is kept separate from our DOMS: in contrast to data we generate by digitising materials ourselves, the Internet is not controlled by us, so other strategies for preserving and monitoring file formats are needed here. Furthermore, the Internet archive today consists of 4.5 billion separate files — this large amount does not really make it feasible to store it in a Fedora Commons[7] system.

Focus on Storage, not on Presentation

DOMS is a module in our library system that is intended to manage our digital objects. The focus is not on presenting them to our users. By keeping the data and our presentation of them in separate systems we make sure that both storage and preservation are in the best possible shape. At the same time we can always use the newest technologies and presentation techniques to disseminate the data (Figure 6).

Building a system like DOMS from scratch would be too much work and the risk of building a system which, due to its specific character, would not survive many years, would be too great. Therefore DOMS has been built on the basis of Fedora Commons, which is an architecture for storing, managing, and accessing digital content in the form of digital objects. Fedora defines

Fig. 6: Very different media types are digitised and stored in the same system called DOMS. This allows us to focus on long-term digital preservation and separate the presentation of the objects from storage.

a set of abstractions for expressing digital objects, asserting relationships between digital objects, and linking 'behaviours' (i.e., services) to digital objects. Fedora has a much more abstract perspective on data than other systems that define archived material (such as, e.g., Dublin Core). This means that we do not have only one data model, such as those which might originate from descriptions of a book or a CD. Having one single data model tends not to work that well when you start describing items that are really different in nature, such as moving images or various digital sound recordings.

On the other hand, if the stored data can be used in practice, we cannot introduce a new description scheme every time we archive a new object. Consequently, the State and University Library has developed selected highly structured data models — e.g. we have one for sound, one for moving images, etc. This work has also made us one of the twelve core committers to Fedora Commons. Everything we develop is open source so we do not just benefit from the work ourselves but we also spread the results to the rest of the world and anyone who wants to benefit from the developed system.

DOMS is growing bigger but it is built to scale. We have tested it for durability of 200 million objects, which should be enough to cover the next five years

of digitisation and preservation in one single instance of our DOMS. Once we have reached this limit, parallelisation could be a possible strategy for scaling up the collection even further.

Another aspect we are working on in order to be able to present information from DOMS smoothly to our users is the handling of rights. Individual data collections have different clearances and copyright issues. Some data are available to all, some are only available for researchers, and again some are available only if you pay a fee.

An Automatically Generated User Interface for DOMS

In order to be able to handle many kinds of data and data models, we designed a generic user interface for DOMS to be used by our staff. This makes it easier for staff who digitize collections to input and enrich data. The interface is generated on the basis of the specific data definition, which enters DOMS. This provides easy data input and data enrichment for the staff who work every day with digitising our collections.

Fig. 7:
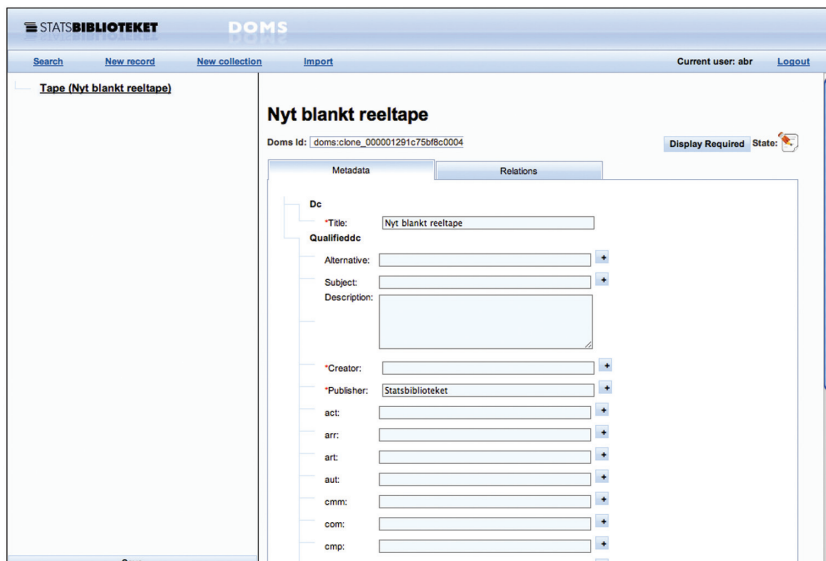
Open is the Only Way to Go

All the steps of our journey towards a library system which supports our business needs and our users' information needs, tell us that open systems are nearly always the only way to go. We are constantly working on modularising different elements of our great technological eco-system in order to be able to optimise each part to meet the purpose for which they were built. Some modules are self-developed, some are open sourced, and some are commercial products, so openness comes in different shapes and sizes. But all modules exchange data and functionality so they can be combined and intertwined in various ways. It is also very clear to us that optimising building blocks often is a very innovative process — by focusing on a single problem, it is possible to solve it in a qualitatively or quantitatively different way that is not dictated by anything but our business vision and user needs. Furthermore, simply by combining building blocks in the right way, we bring tremendous value to our users — and that is exactly what we believe the library business is all about.

Notes

[1] http://www.statsbiblioteket.dk/summa/

[2] http://lucene.apache.org/java/docs/

[3] Means "library.dk". Can be seen at www.bibliotek.dk

[4] www.copydan.dk

[5] A petabyte equals 1,000 terabytes.

[6] http://netarkivet.dk/

[7] http://www.fedora-commons.org/