Knihovny.cz – building a discovery portal for Czech libraries

Petr Žabička
Division of Development, Research, Innovations and Digitisation, Moravian Library in Brno,
Czech Republic.
E-mail address: Petr.Zabicka@mzk.cz

Abstract:

Knihovny.cz has been built as a user-centred discovery portal integrating library catalogues, library services, digital libraries as well as other library resources. The portal aims to become the gateway to collections and services of Czech libraries. It is based on the VuFind open source discovery system with a number of modifications to improve user experience. It covers library collections and other information resources such as bibliographies, digital libraries and other online resources as well as library services. This paper provides insight on the decision-making process preceding the development of the portal that led to it being developed in-house by the Moravian Library. The paper then concentrates on the development process of the portal including user analyses, user experience studies and design decisions that led to the current version of the portal. To conclude, the paper discusses lessons learned during all phases of the design process, the difficulties encountered when promoting the portal and presents plans for further development of the portal, chiefly ongoing development of machine learning based improvements, the integration of interlibrary service component currently being developed by the National Technical Library a planned addition of a commercial index of non-Czech resources.

Keywords: Discovery Portal, VuFind, UX, Czech libraries

1 BRIEF HISTORY OF THE PORTAL

In 2010, Czech librarians started working on a new strategy – the Concept of Czech Libraries Development for the years 2011 – 2015 [Richter 2014]. In contrast to previous, rather formal, strategies the new strategy was created by a large body of librarians from many Czech libraries. The main vision of the strategy was to provide the user with library services anywhere any time and one of the main tools to achieve this was described as a "central portal of libraries.". In January 2012, the strategy was approved by the government.

To build the portal, a group of the largest Czech libraries came together and formed a Council of the Czech Library Portal. Early-on, it has been decided to run this portal on the

knihovny.cz (knihovny = libraries) domain which up until that time has been used to inform people about the various services provided by Czech libraries. Knihovny.cz later also became the self-describing name of the portal.

The Council then formed several working groups that defined the software requirements to procure the portal on a software as a service (SAAS) base. Over several following years several requests for information were issued and evaluated but it was becoming clear that because of the complexity of the planned portal and many special requirements the SAAS approach would be quite costly and had a high risk of not achieving the overall objective. Most of the libraries were not ready to join such a portal because their systems did not have the application programming interfaces (APIs) necessary to communicate with the planned portal and there was a serious risk that it would take years to connect even a few libraries into the portal which, in turn, would take several years to develop. Furthermore, once the SAAS contract came to an end, there was no guarantee that the original supplier would win the next request for proposals and the libraries would be forced to repeat the same process with a new vendor. It also became clear that much of the development work will need to be done even if the portal is based on an existing software.

This led to a decision to propose standards-based interfaces through which the portal will connect with the library systems [Rosecký 2013; Žabička 2013]. The proposal was chiefly based on DLF API recommendation [DLF 2008].

We have also decided to use OAI-PMH for bibliographic records enriched by the information on all items stored in a repeatable proprietary MARC field 996 – MARC-based holdings never took hold in Czech libraries. The 996 field has already been used for several projects of the National Library that needed a simple way to get the item-level information through OAI (the 996 field itself is being generated automatically by the given library system). As OAI-PMH has already been supported by all library systems available at the time, the modifications necessary to implement support for the 996 field were considered minor in comparison to any other method of harvesting the holdings data.

To provide the usual library catalogue services to the user, a subset of Z39.83 protocol also known as NCIP (NISO Circulation Interchange Protocol) has been specified for all library systems used by the Czech libraries – with one exception: Aleph library system already had a working, open and well-documented API available and we decided to use this API for Aleph. At the time of this decision, a group of German libraries set about to specify a DAIA/PAIA family of APIs (Document Availability Information API, Patrons Account Information API). The decision then has been to use the NCIP as mature, stable standard. However, if the decision was to be made today with what we already know, DAIA/PAIA may have been chosen instead.

To enable users to log in to the library services through the portal, SAML-based (Security Assertion Markup Language) authentication and membership in eduID.cz (Czech academic identity federation) run by CESNET (Czech research e-infrastructure) has been the logical choice. Some of the libraries with Aleph (including MZK) have already been members of the eduID.cz federation and used Shibboleth or SimpleSAMLphp to turn Aleph into an identity provider to give library users access to commercial electronic resources provided by the library.

As some libraries subscribe to electronic resources not available elsewhere, the portal might recommend to the user to register to a library where the given resource is available. In many cases, especially with remote access to the electronic documents, remote registration and remote access is the quickest way of obtaining the information needed. What remained was to

establish a secure way of checking the user identity. Therefore, further recommendation was to implement mojeID.cz as a source of reliable identities. Although mojeID.cz is not a library project, it provides infrastructure for identity validation and libraries can act both as an entity doing the identity validation as well as a service provider using that identity to provide some services – in this case a service of remote registration.

As most libraries collect a registration fee, online payments also became necessary. Even though the number of libraries offering online payments is still rather small, their number is growing as library system vendors add this functionality to their products. As a side effect, this allows users to pay online for other library services as well and saves a lot of time to both the staff and the users.

Since 2013, the libraries have been working with library system vendors to implement the above recommendations with the aim to have the interfaces available to all libraries that will be joining the portal. Currently, 8 library systems covering the vast majority of libraries have interfaces necessary to connect to the portal: Aleph from ExLibris (no Alma in the Czech Republic yet), Koha open source library system (no Evergreen library yet) and five local systems: Koniáš, developed in-house by the Municipal Library of Prague; DaWinci, from SVOP; Advanced Rapid Library, from Cosmotron; Verbis, from KP-SYS; Clavius, from Lanius. and Tritius, from Tritius Solutions.

The Moravian Library had been active in the field of open source library software development for a long time and by 2013 it had several years' worth of experience with the VuFind library system and developed several VuFind-based portals to bring together relevant bibliographic databases. While two of the portals were aimed at subsets of library collections (historickefondy.cz aggregates databases of old prints and manuscripts, narodnifonoteka.cz gathers data on sound recordings), the third one, CistBrno.cz has been set up to aggregate data from major libraries from Brno. In 2014, MZK started to test the interfaces of the individual library systems to see whether the systems will work with the central portal once the portal is operational. At the end of 2014, the board of directors of the participating libraries has decided that the portal will be based on an open source VuFind system and they tasked MZK with the development of the portal.

The work started immediately although it took several months to secure the funding necessary to find additional developers and acquire the storage necessary for the project. By the end of 2015, a beta version of the portal was online and the portal went to production on the 26th October 2016. MZK, in co-operation with the Brno University of Technology has also successfully applied for a R&D grant from the Ministry of Culture to enhance the portal functionality through the use of machine learning and associated technologies.

2 FUNCTIONAL REQUIREMENTS

As mentioned above, the aim of knihovny.cz was to provide a resource discovery service as well as a library service proxy. The user should also be able to link user accounts from several libraries and get the relevant information and services from those libraries in one place. As a discovery system, the portal indexes a wide range of information resources: library catalogues, bibliographic databases, tables of contents, full text and metadata content of digital libraries as well as bohemical resources outside libraries (Czech standards and patent databases, data on research and legal resources or newspaper article databases).

The portal itself does not provide interlibrary services but it will link to a document delivery system developed and run by the National Library of Technology; this is currently under development and should be finished by the end of 2017. This system should be able to

provide the user with interlibrary and delivery services in cases when it is not possible get the given document directly.

The portal also makes use of book covers and tables of content through ObalkyKnih.cz API. ObalkyKnih.cz is a collaborative project of Czech libraries, allowing libraries to scan book and magazine covers and tables of content and use them to enrich their catalogs. It was later enhanced to centralize user ratings and comments and became a de-facto standard part of a library catalogue.

For international resources, API access to EBSCO Discovery Service has been acquired and will be available to the users in the second half of 2017. As the majority of portal users search for Czech language resources while EDS covers primarily English language resources we do not plan to mix the search results from the two indexes into one list as this might confuse users.

3 USER EXPERIENCE

Knihovny.cz portal was preceded by several librarian-oriented union catalogues and portals while the knihovny.cz target audience is library users. The change of focus from librarians to end-users means the portal has to be easy to use and easy to understand. To be able to provide local library catalogue services through the portal, the portal has to be consistent with the local library catalogues – it has to provide the same services over the same collections. This means the portal has to harvest the library databases in their entirety, including metadata records as is – with various errors or deliberate deviation from standards. However, such records must not negatively impact the overall user experience with the portal so there are rare cases when metadata records have been refused.

Although the portal itself is based on VuFind and would mostly work with the default VuFind user interface, we have decided to test our assumptions on user requirements. The Municipal Library of Prague, which has been in charge of this part of the project, contracted a web development company who met library users throughout the country to identify user needs and defined personas and scenarios that helped us to see the priorities and prepared a wireframe design of the portal. The four personas they came up with (reader of fiction, student, traditional researcher, digital scientist) could be split into two main groups – research library users and public library users.

The wireframe based on this analysis has not been influenced by the capabilities of VuFind or any other library system or the actual data available so in some cases it differed a lot from what was possible to achieve. For instance, one of the suggested functions was browsing through a list of authors with author pictures displayed with each name. This feature seemed to be attractive for users, however, in real life, we do not have pictures for the vast majority of the authors listed in the library databases. And at the same time, any alphabetical author list starts with a large number of abbreviations (anonymous authors of old prints, newspaper articles etc.). To work around this, a list of "recommended" authors has been compiled from which random five authors are displayed on the portal homepage.

Due to a limited budget, we have limited the development to one user interface. Therefore, it was necessary to find a compromise between the requirements of two principal personas: "the reader of fiction" and a "student/researcher". The standard VuFind interface was modified to match the wireframe as much as possible. As the next step, we hired user experience specialists to help us fine-tune the portal and put finishing touches to its graphic design.

Of the more notable UX-driven changes to the user interface, the most important was the asynchronous downloading of a search result set and a seamless switch from basic to advanced search and back without losing the search query and filtering (facet) settings.

Currently, the focus of the UX is centered on discovery index integration. To avoid confusing the user, the search interface will not change and the user will be presented with two result sets – one for the local (national) index and one for the international index. To avoid confusing the user, each of the result sets will be displayed in its own tab with the content of just one of the tabs visible.

4 TECHNOLOGY OF KNIHOVNY.CZ

The key to a successful search is an up-to-date index of correctly de-duplicated or interlinked documents. To achieve this, a new tool has been written to harvest the bibliographic records, de-duplicate, modify and enrich them before indexing them into a Solr index. At the time of this writing, the ~150 GB index resides on a NetApp SSD shelf and we are able to update the index in about 5 to 8 hours every night, while it would take about a week to harvest and 3 days to re-create the index from scratch on one machine. This index includes about 20 million records de-duplicated to below 13.5 million and includes also almost 44 million pages of OCR full texts from the National Digital Library project.

By default, the search results are ranked by relevancy using BM25 algorithm. The ranking system takes into account variables like the length of the document, how many times the search term appears in a given document and in the whole index, where was the search term found (in title, author, elsewhere in metadata, in table of contents or elsewhere in the document), how old is the document, and whether the document can be found in one of the libraries the user is registered in.

To improve the search experience, harvested records are currently enriched by data from the National Authority Database and MeSH (Medical Subject Headings) and converted or otherwise modified to fit one of the two main indexing profiles: MARC or Dublin Core. In relevant cases, full-text data is linked to the records – be it just a table of content or a whole document. Further metadata enhancement will be done for the bibliographic records of digitized books with full-text available but lacking subject headings. In such documents, we will attempt to generate Conspectus categories using machine learning techniques. This will keep the books visible when user filters the search results by subject.

The de-duplication process then tries to find records describing the same items in different libraries or databases. The de-duplication is a rather complex algorithm optimized for speed and quality of de-duplication. It takes into account not just the individual bibliographic records but also metadata, showing which bibliographic records have been manually merged by the administrators of the national union catalogue in the National Library of the Czech republicto make sure that their valuable work is re-used. At the same time, digital library records are linked to metadata records from library catalogues.

One of the data sources the portal gets from individual libraries may also be a MARC export of a SFX knowledgebase. Using this data, the portal can point the user to an electronic resource subscribed by the library. As some of the libraries import this data into their own catalogues, there is another challenge involved – to ensure that nothing has been imported twice (once through the library catalogue and once directly by the portal).

Similar problem arises with e-books lending: many Czech libraries subscribe to one or more e-book lending providers but often they do not provide all the e-books available from a given provider to their users: some libraries choose just the e-books they have in print, while other

libraries provide just the books they do not have in print; some libraries import the e-books into their catalogues and other libraries just merge the links to the existing print book records.

To improve the overall user satisfaction, the portal provides autocomplete service for entering the search term. In Knihovny.cz, the autocomplete function provides suggestions splits into three types (title, author and subject) so that user can select the best match. This and other functions of the portal will be subject to further research thanks to a five-year research project started in 2016 in cooperation with a Brno University of Technology research team experienced in knowledge technologies and semantic enrichment of content.

5 FUTURE WORK

After integration of the EDS index into the portal, the only thing missing will be integration of interlibrary services – the "Get It!" service which is currently under development by the National Library of Technology. Without it, we cannot promote it to users outside Prague – users from Prague usually find what they need in one of the libraries located there. Also, even though the target audience of the portal is not librarians but library users, it will be librarians who will be "selling" the portal to the users. Therefore, we have to "sell" the portal to librarians which is never an easy task because they have much different usage patterns from the general public. Currently, librarians still form a rather large group of portal users. This we infer from the usage statistics where advanced search still makes about 20% of the total searches while on systems where the majority of the users are not librarians the usage of advanced search is much lower.

Apart from the EDS and "Get It!" integration, the portal will undergo just minor changes and bug fixes – the main changes will be under the hood. A new server will make the portal more robust and allow for better response times. Some parts of the portal will be refactored and new libraries and databases will be added to the index. Hopefully, these changes will sell the portal both to the librarians as well as to the end users.

References

ILS Discovery Interfaces. [online]. Washington: Digital Library Federation, 2008 [cit. 2013-04-09]. Available from: http://old.diglib.org/architectures/ilsdi/.

RICHTER, Vít and MILLEROVÁ, Kateřina. The Development of Czech Libraries, 1990–2013. LIBRARY TRENDS, Vol. 63, No. 2, 2014 (Libraries in a Postcommunist World: A Quarter of a Century of Development in Central and Eastern Europe and Russia: Part 1, edited by Hermina G.B. Anghelescu), pp. 161–182. DOI 10.1353/lib.2014.0041

ROSECKÝ, Václav and ŽABIČKA, Petr. Požadavky na výměnu dat mezi knihovním a discovery systémem na příkladu implementace systému VuFind v Moravské zemské knihovně. *Knihovna* [online]. 2013, roč. 24, č. 1, s. 79-88 [cit. 2016-02-14]. Dostupný z WWW: <http://knihovna.nkp.cz/knihovna131/13179.htm>. ISSN 1801-3252.

ŽABIČKA, Petr. Knihovní systémy a jejich integrace do CPK. Duha [online]. 2013, roč. 27, č. 3 [cit. 2016-02-14]. Dostupný z WWW: <http://duha.mzk.cz/clanky/knihovni-systemy-jejich-integrace-do-cpk>. ISSN 1804-4255.