Závěrečná zpráva z VISK 8b

## Cíle projektu

Cílem projektu bylo zahájit budování Centrálního portálu knihoven. Prezentační vrstva portálu měla být budována na bázi open source systému VuFind verze 2.x, který využívá k indexaci záznamů nástroj SOLR. Pro plnění portálu (sklízení metadat, jejich normalizace, deduplikace, obohacování) měl být použit modul Record Manager vyvinutý pro projekt Finna.

Výstupem pak měl být portál umožňující:

* vyhledávání a prezentaci výsledků jak v anonymním režimu, tak pro přihlášeného uživatele
* přihlášenému uživateli, jehož vstupní knihovna bude podporovat protokol NCIP nebo Aleph Restful-API, nabídne portál funkce nahrazující lokální OPAC knihoven

Předpokládanými zapojenými zdroji měly být katalogy a digitální knihovny zakládajících knihoven a vybrané specifické národní zdroje.

Cíle projektu byly naplněny, i když v několika případech došlo k realizaci cíle jinou cestou než projekt předpokládal. Je ale nutno zdůraznit, že cílem projektu nebylo systém dokončit, ale prokázat, že je zvolené řešení realizovatelné, což se podařilo, o čemž svědčí i to, že zakládající knihovny již na konci roku 2015 uzavřely vzájemné smlouvy o spolupráci na dalším budování portálu.

## Způsoby řešení

V roce 2015 v souladu s projektem MZK vytvořila tyto základní části portálu:

* Hardwarové zajištění provozu: zprovoznění velkokapacitního úložiště pro stažená data a rychlého úložiště na bázi SSD pro indexaci a zpřístupnění.
* Správce zdrojů pro sklízení, aktualizaci, deduplikaci, obohacování a indexaci dat do lokálního indexu v systému SOLR.
* Modifikovaný open source systém VuFind jako uživatelské rozhraní celého portálu.
* Napojení na knihovní systémy pomocí protokolu NCIP, resp. Aleph RESTful API.
* Přihlašování uživatelů do portálu a propojování uživatelských identit.
* Napojení na linkservery bylo řešeno indexací exportu relevantní části znalostní báze SFX do lokálního indexu CPK.

Pro vývoj a testovací provoz portálu bylo využito kapacity a výpočetního výkonu stávajících serverů MZK, které byly doplněny o velkokapacitní (SATA) a vysokorychlostní (SSD) datové úložiště NetApp.

Jádrem stávající vývojové infrastruktury jsou vývojový server cpk-front.mzk.cz a testovací server beta.knihovny.cz. Veškerý vývoj až na soubory s daty citlivého charakteru probíhá na serveru github.com a jeho zdrojový kód je volně dostupný. Na serveru GitHub na adrese <https://github.com/moravianlibrary/CPK/issues> je možné vznikající portál také připomínkovat, čehož využívají zejména pracovní skupiny řídící a připomínkující jeho vývoj.

Největším rozdílem ve vývoji byla realizace náhrady modulu Record Manager aplikací Správce zdrojů, která byla vyvinuta na míru potřebám portálu. Správce zdrojů je nyní klíčovou komponentou, odpovědnou za plnění portálu, tzn. pravidelné sklízení metadat, jejich normalizaci, deduplikaci a obohacování.a následnou indexaci do SOLRu. Nutnost vyvinout tento systém znovu od základu vyplynula z pečlivé analýzy jak původního Record Manageru, tak i zdrojů dat a provozních nároků na portál samotný. Původní systém nesplňoval nároky na dostatečnou robustnost a výkonnost a to nejen s ohledem na použitý programovací jazyk a databázový systém, ale také na jednotlivé procesy, především pak deduplikaci, která byla v původním Record Manageru řešena nevyhovujícím způsobem. Správce zdrojů umožňuje mimo jiné lepe kontrolovat jednotlivé procesy a je schopen ošetřit celou řadu výjimečných stavů, na kterých troskotal Record Manager a které vznikaly převážně z důvodu neúplného dodržování některých standardů původci dat. Do budoucna bude také Správce zdrojů mnohem lépe rozšiřitelný o modul plánování a spouštění úloh. Správce zdrojů nyní využívá databázi Postgres a je napsaný v jazyce Java.

Stav vývoje

* Obohacování záznamu o údaje Národních autorit je nyní realizováno pro jmenné autority. Jedná se o funkční předlohu pro obohacování o autority ostatních typů, které nebude příliš složité a bude v krátké době následovat.
* Ve stavu dokončování a testování je proces průběžných aktualizací dat, který je součástí modulu Správce zdrojů. Zde je kritickým okamžikem zejména nutnost zajistit aby procesy deduplikace a indexace denních aktualizací proběhly během jednoho dne a paralelně mohly probíhat i celkové importy dat z knihovních systémů nově zapojovaných knihoven.
* Zaindexování informací z linkovací serverů bylo ověřeno pro MZK a NLK. Rozhodnutí nezapojovat linkovací servery ale indexovat přímo jejich znalostní databáze vychází ze zjištění, že uživatelé nevědí, co znamená SFX, tuto službu proto často ani nenajdou a pokud ji najdou, tak často nedostanou relevantní informace. Přitom nejpodstatnější funkcí linkovacích serverů je přesměrování uživatele na plný text dokumentu, což je možné nejlépe realizovat právě zaindexovaáním znalostní báze SFX pro jednotlivé knihovny.
* Co se týče integrace knihovních katalogů, záměrně jsme nezapojovali ty knihovny, které nebyly připraveny plně, protože provazby mezi samotnou sklizní záznamu a propojením na knihovní systém jsou tak velké, že by v případě zapojování ne zcela připravené knihovny bylo nutné vynaložit nadměrné množství zbytečné práce na obejití jednotlivých nefunkčností. V případě všech knihovních systémů (Aleph, DaWinci, Koniáš, Koha, ARL, Clavius, Portaro) ale probíhalo průběžně postupné zprovozňování zapojení alespoň jedné instalace tak, aby po konečném odladění komunikačních rozhraní a formátů bylo možno rychle zapojit všechny knihovny s daným systémem. Do betaverze portálu je nyní zapojeno 12 knihoven, další jsou v různých stádiích rozpracovanosti.
* V několika případech je ještě otevřená otázka způsobu předávání informací o mapování některých prvků chování knihovního systému do portálu. Ukázalo se totiž, že mnohé činnosti, o nichž jsme předpokládali, že se v různých systémech chovají stejně, se ve skutečnosti chovají zásadně odlišně. Systém chceme směřovat k tomu, aby si tuto parametrizaci mohly dělat v administrátorském rozhraní CPK knihovny samy a tím se vyhnuly závislosti na správcích CPK. Zatím nedošlo k přenesení podpory funkcionality online plateb do portálu. Řešení bude pravděpodobně převzato z VuFindu MZK.

## Náročnější úkoly, které bylo nezbytné vyřešit

**Zadání**

Pro analýzy a samotný vývoj byly použity podklady dodané pracovními skupinami pro CPK. Jednalo se převážně o Projektový záměr verze 4 a seznam prioritních zdrojů k zapojení od Pracovní skupiny pro zdroje. V průběhu roku docházelo ke zpřesňování a doplňování těchto dokumentů o další, podrobnější podklady pomocí e-mailové korespondence, sdílených dokumentů a osobních jednání. Stanoviska jednotlivých osob zapojených do konzultací i stanoviska jednotlivých pracovních skupin se ale mnohdy významně lišila a příliš nedocházelo ke konsolidaci názorů. Tuto situaci pomohlo vyřešit rozhodnutí Rady CPK zredukovat počet zainteresovaných členů pracovních skupin na na nejaktivnější členy zřízením společného Expertního týmu a ponecháním ostatních členů v širší Pracovní skupině tak, aby mohli být dostatečně informování a podílet se nadále na vzniku portálu.

Při vývoji bylo snahou, aby se tam, kde je to možné, dodrželo zadání, což se až na několik odchylek podařilo splnit (zásadní koncepční rozdíl je v řešení implementace linkovacích serverů nebo ve způsobu využití systému Perun pro propojování identit).

Na základě praktických zkušeností ze zapojování prvních knihoven v průběhu roku byla dokumentace profilu protokolu NCIP upřesněna a doplněna. Tyto úpravy musely být následně zapracovány i do systémů zapojovaných knihoven.

**Správce zdrojů a národní index**

Sklizeň záznamů komplikovala pomalá odezva některých implementací OAI-PMH, což se především týkalo se prvotní sklizně.

V případě deduplikací jsme se pak potýkali převážně s otázkami nejednotně zpracovaných záznamů a nedodržování standardů (převážně u záznamů vzniklých v minulosti retrokonverzí).

Největší výzvou pak byl samotný fakt, že plánovaný portál je velký (množství zpracovávaných dat již nyní přesahuje 20 mil. záznamů) a manipulace s tak velkým indexem má svá specifika. Cílem dalšího snažení bude proto i optimalizace index SOLR tak, aby v tuto chvíli z vývojových důvodů prioritní rychlost zpracování dat byla v průběhu jara nahrazena optimalizací (zvýšením) rychlosti uživatelské odezvy, čemuž výrazně napomůže i přesunutí produkčního indexu na SSD úložiště.

**Identity**

Slíbená Shibboleth brána pro malé knihovny, jejichž zřízení bylo náročné hlavně z organizačních důvodů a domluvy CESNETem jako provozovatelem eduID, v tuto chvíli funguje pro Městskou knihovnu v České Třebové. Brány pro další knihovny budou následovat v roce 2016 po ostrém spuštění NCIPu v těchto knihovnách.

Propojování identit nebylo realizováno pomocí systému Perun dle původních plánů, ale vlastní implementací s využitím řešení vyvinutého pro portál Finna. V čase realizace totiž nebyl Perun dostatečně uživatelsky jednoduchý a použitelný a jeho API nenabízelo všechny potřebné funkce, ty byly dopracovány na náš podnět v posledních měsících roku 2015.

**Uživatelské rozhraní**

Od počátku bylo jasné, že jako uživatelské rozhraní portálu bude použit VuFind. Požadavky na něj ale od počátku jasně ukazovaly, že jej bude nutné do velké míry modifikovat.

V průběhu roku se několikrát ukázalo, že jednotné rozhraní pro všechny knihovny a uživatele není ideální. Protože zapojované knihovny se liší jak ve své velikosti, tak ve svých funkcích, mají jejich uživatelé značně různorodé potřeby. Do jisté míry se to projevilo ve výsledcích uživatelského testování, které bylo provedeno paralelně jako projekt MKP. Výsledky testování se výrazně přiklonily k potřebám malých veřejných knihoven. Částečně pak byly ale korigovány expertní skupinou a zástupci větších i menších knihoven.

Vzhledem k rozhodnutí vyvinout prozatím jen jedno rozhraní pro všechny uživatele je nezbytné dojít u uživatelského rozhraní ke kompromisu mezi oběma směry, resp. zajistit takovou uživatelskou podobu portálu, která by pokročilejší funkce nabídla v podobě, která nebude komplikovat použitelnost základní podoby portálu.

**Logistika vývoje, návaznosti a financování**

Snaha vyvinout poměrně rozsáhlý systém v krátkém časovém období na zkoušku nás nutila k tomu dělat v mnohých případech rozhodnutí vedoucí k rychlému, ale ne vždy nejlepšímu možnému řešení. Zajištění souvislého financování na období delší než 1 rok by usnadnilo dlouhodobé plánování a by navíc mnohem lépe řešilo otázku vybudování funkčního a stabilního týmu.

**Další externí limity**

MZK se v průběhu loňského roku snažila získat do správy doménu knihovny.cz, což by v mnohém usnadnilo jak vývojové práce, tak následnou komunikaci s externími systémy jako je EduID a CESNET. V době psaní tohoto textu je již připravena dohoda s Národní knihovnou o převzetí správy domény Moravskou zemskou knihovnou.

Agregátor digitálních knihoven - Česká digitální knihovna - byla zprovozněna až na samém konci roku 2015, takže s fulltexty jsme pracovali na vzorcích sklizených z jednotlivých knihoven. Problematiku indexace fulltextů proto máme realizovanou a testovanou na vzorcích dat z Krameria Národní digitální knihovny, provozované MZK.

Pomalá okamžitá odezva od některých knihovních systémů by mohla být limitujícím faktorem používání CPK, pokud nedojde k jejímu brzkému vyřešení. Na základě těchto zjištění jsme veškerá uživatelská rozhraní pracující s externími daty přepsali do asynchronní podoby, aby pomalá odezva z jedné knihovny nezpůsobila kolaps celého systému.

Komplikace v komunikaci pak nastávaly u knihoven, kde technický kontakt pro zapojení do knihovny do CPK nebyl zároveň členem některé z pracovní skupiny pro CPK. Toto by měla řešit nová struktura pracovních skupin

## Využití finančních prostředků

Poskytnutá dotace byla využita na mzdové prostředky (včetně odvodů) týmu, který portál vybudoval a na pořízení datových úložišť.

Celkem 2.499.000 Kč bylo použito na úhradu mezd pracovníků MZK.

Z financí projektu došlo zároveň k posílení datového úložiště MZK tak, aby mělo dostatečnou kapacitu i rychlost pro provozování portálu v pilotním období.

Nejprve bylo stávající datové úložiště posíleno o diskové pole Netapp FAS2554 s osazením 24x 2TB disky a následně na konci roku díky navýšení rozpočtu projektu ještě o polici s SSD o kapacitě 12x400 GB. Oba nákupy poskytly dostatečný prostor na vývoj portálu, multiplicitní vývojové kopie a zálohy dat jak pro běžný provoz, tak pro vysoké výpočetní nároky, které přináší deduplikace a indexace. SSD pak poskytnou navíc komfort v podobě rychlé odezvy uživateli po spuštění samotného portálu.

499.000 police pro datové uložiště Netapp FAS2554 s osazením 24x 2TB disky

760.000 police SSD pro datové uložiště Netapp o kapacitě 12x 400 GB

## Závěr

Výsledkem prací v roce 2015 je funkční jádro portálu, který je v režimu testování přichystaný pro první uživatele na adrese beta.knihovny.cz. Zapojování knihoven postupuje průběžně podle toho, jak se daří ve spolupráci s knihovnami a dodavateli knihovních systémů odlaďovat jejich komunikační rozhraní. MZK přistupuje k zapojování knihoven tak, že po odladění zapojení na jedné knihovně jsou teprve postupně zapojovány další knihovny se stejným knihovním systémem. Právě nedostatečný počet plně do portálu zapojených knihoven vedl k rozhodnutí Rady CPK ze dne 20. října 2015 k posunutí veřejného spuštění portálu na léto 2016 (respektive na 23. 6. 2016) tzn. do doby, než bude možné zapojení všech nebo alespoň významné většiny zakládajících knihoven.