# Zpráva projektů “Příprava zapojení knihoven do centrálního portálu” a “Příprava zapojení knihoven do CPK – dodatečná žádost”

## Úvod

Projekt “Příprava zapojení knihoven do centrálního portálu” měl dva hlavní cíle:

1. Praktické ověření připravenosti knihoven (a jejich knihovních systémů) na zapojení do Centrálního portálu knihoven.

2. Praktické ověření náročnosti zapojení specifických (krajských/oborových) lokálních zdrojů do lokálního indexu.

Vzhledem k tomu, že projekt byl v prvním kole VISK8b krácen, podala MZK do 2. kola VISK8b dodatečnou žádost (název projektu “Příprava zapojení knihoven do CPK – dodatečná žádost” aby mohl být realizován celý zamýšlený projekt. Krácení na polovinu v prvním projektu se přitom týkalo pouze bodu 1, bod 2 byl realizován celý v rámci prvního projektu. Druhý projekt tak umožnil realizovat práce v rozsahu, zamýšleném prvním projektem.

Zjištěné skutečnosti a dosažené výsledky přispěly pak k tomu, že Rada CPK v listopadu rozhodla, že CPK nebude poptáván jako dodávka služby, ale že jej Moravská zemská knihovna začne budovat sama s využitím existujícího, volně dostupného systému VuFind, což později doporučila i Ústřední knihovnická rada.

V průběhu řešení projektů připravila MZK některá dílčí řešení, která budou využita při integraci knihovních systémů do portálu (např. gateway NCIP-Shibboleth) a získala řadu praktických zkušeností, které v příštím roce využije při budování portálu. Konkrétní poznatky týkající se řešení problematiky cíle 1 jsou uvedeny dále v textu.

V rámci řešení problematiky cíle 2 byli kontaktováni původci nejdůležitějších informačních zdrojů s cílem dosáhnout dohody u indexaci těchto zdrojů v lokálním indexu. Většina původců nemá proti indexaci svých zdrojů v CPK námitky, zejména pokud jde o zdroje volně dostupné, jiná situace je u zdrojů které volně dostupné nejsou nebo jsou, ale v omezené podobě. Zejména v těchto případech budou jednání s cílem podepsat smlouvu o poskytnutí zdrojů k indexaci pokračovat v příštím roce.

U řady volně dostupných zdrojů bylo jako nejschůdnější řešení vyhodnoceno postupovat formou indexace příslušného webu (řada původců o jiné možnosti ani neuvažovala). Pro tyto zdroje bylo vytvořeno testovací rozhraní s využitím Google Custom Search Engine.

Dále uvádíme podrobněji zjištění, která řešitelský tým učinil v průběhu řešení projektu:

## Knihovny

* je poměrně málo knihoven, které se aktivně a dobrovolně zapojily do testování. Další část knihoven by se do testování zapojila, ale nemá zatím jak (nesplňuje technické požadavky.
* pokud má CPK fungovat, tak bude nutné, aby byly knihovny aktivní nebo alespoň aby reagovaly v relativně krátkém čase.

## Data

* není novinkou že bibliografická data knihoven nejsou zdaleka tak konzistentní, jak by bylo ideální a je potřeba s tím počítat. Na rozdíl od Souborného katalogu, který nedostatečně zpracované a chybové záznamy odmítá, CPK bude nucen, pokud má sloužit jako náhrada za lokální OPACy, sbírat všechny záznamy. Existující nekonzistence a chybovost zápisu má největší negativní vliv na určení typu dokumentu, což se projeví na kvalitě deduplikací a zobrazení faset.
* Pro deduplikace je klíčové, aby identifikátory použité pro párování byly v dobré kvalitě.
  + Doplnění deduplikačního klíče o další řetězce mimo ISBN je samozřejmé. Ale zbytečně velké množství záznamů se nespáruje kvůli duplicitně použitým ISBN. Přestože jsou duplicitně obsazená ISBN agenturou ISBN označeny a převedeny do podpole z, do velkého množství knihoven tato informace už nedoputuje a duplicita je objevena až ve chvíli spojení s daty jiných knihoven.
  + Při zpracovávání záznamů jsme vyvinuli kontrolní mechanismus, který porovnává ISBN oproti bázi NKC01 a vytváří návrh oprav pro jednotlivé knihovny.
* Identifikátory jiné než ISBN jsou na tom ještě hůře, protože knihovní systémy kontrolní mechanismus na duplicitnu ISBN v rámci svých databází mají, ale ne vždy jej aplikovaly například na nakladatelské číslo, číslo ČNB, OCLC číslo.
* Vzhledem k tomu, že proces zpracování dat v CPK bude čistě automatický a nebude možné do něj zasahovat manuálně, testujeme možnost využití výsledků deduplikací Souborného katalogu ČR, kde je proces deduplikace doplněn o manuální dočišťování.
* Různé knihovny mají různou míru veřejné dostupnosti různých údajů, ať již proto, že některé údaje vúbec neevidují nebo proto, že je považují pouze za služební. Proto je důležité, aby knihovny posílali pouze ty údaje, které mají být uživatelům CPK přístupné. Měla by to být otázka nastavení na straně OAI provideru jednotlivých knihoven, ne vždy je to ale možné. V takových případech je potřeba, aby
  + byla knihovna schopna předat CPK jednoznačnou informaci o viditelnosti bibliografického záznamu
  + byla schopna předat informaci o viditelnosti exemplářů (podpole q obsahující 0 v poli 996)
* Problematika vícesvazkových děl, které jsou některými knihovnami popisovány zdola a jinými shora je problematická, přestože bychom uměli takové záznamy automaticky sloučit. Prezentace takto sloučených záznamů není uživatelsky dostatečně srozumitelná, hlavně pak v návaznosti na zobrazení konkrétních exemplářů a objednávky/rezervace. Uživatelsky mnohem jednodušší se jeví zpracovat a sloučit bibliografické záznamy vícesvazkových děl tak jak přijdou (tj. neslučovat záznamy popisu zdola a shora) a následně doporučit další svazky v seznamu “podobných” záznamů.
* Je nutné aby id záznamu bylo jedinečné v rámci jedné celé přispívající knihovny. Problematické je pokud knihovna buduje více databází a jako id používá například sekvenčně strojově přidělené číslo. V takovém případě je důležité, aby id bylo doplněno i nějaký prefix tak, aby byla zachována jedinečnost id v rámci instituce.
* Mnohé instituce mají interně ve svých datech duplicitní záznamy. Obvykle to vzniklo sloučením databází více institucí (např fakult) do jedné. V takovém případě je potřeba rozhodnout zda tyto záznamy spojit do jednoho i přes to, že takové spojení lokální katalog nemá.

## Fasety

* Pro použitelnost faset (filtrů) je klíčovým předpokladem očekávají určitá kvalita záznamů a to zejména v oblasti kódovaných údajů, jejichž využití je pro fasety z důvodu jednoznačnosti zápisu nejjednodušší. Některé záznamy, převážně ty staršího data vytvoření, neobsahují jiný údaj pro určení druhu dokumentu než rozsahy a prefixy signatur, jindy je nutné heuristické určování z nejednoznačně zapsaného fyzického popisu či všeobecné poznámky. Tyto nestandardní způsoby určování jsou použitelné pro zpracování menšího množství dat nebo omezeného množství zdrojů. Ve velkém množství je potřeba počítat s tím, že takovéto záznamy spadnou do položky Ostatní (Neurčeno) a dost možná také nedojde k jejich spárování v rámci procesu deduplikace. Z funkčního hlediska nedojde k omezení, pouze selekční údaje v záznamu nebudou moci být obohaceny o údaje z případných shodných záznamů z jiných knihoven.
* V tuto chvíli ještě není možné otestovat dopad zavedení RDA. Ve chvíli kdy by všechny záznamy byly v souladu s RDA, výrazně by to zjednodušilo rozpoznávání formátů. Reálně ale musíme počítat s tím, že bude nutné do procesu zpracování vsadit rozpoznání použitých pravidel a podle toho zvolit správné mapovací postupy.
* Na základě analýzy implementace národních autorit se ukázalo, že fasety nebude možno přímočaře obohacovat o údaje z autoritních údajů. Toto obohacování by totiž u faset způsobila multiplicitní zobrazení ekvivalentních hesel pod sebou, což by bylo pro uživatele kontraproduktivní. Práce s nadřazenými a podřazenými pojmy bude naopak předmětem dalších analýz. Obohacování o údaje z autoritní báze bude ale používáno pro indexaci a tedy vyhledávání a zobrazení.

## OAI-PMH

* OAI-PMH se ukázalo jako optimální pro pravidelné aktualizace údajů v portálu (jak pro nově zpracovávané údaje, opravené záznamy nebo záznamy mazané). Jako problematické se ale OAI-PMH bohužel ukázalo pro počáteční sklizeň záznamů. Přenos většího počtu záznamů především u menších knihovních systémů je poměrně pomalý a často nevysvětlitelně padá. Ne vždy se podaří v takových případech znovunavázat spojení.
* Také se ukázalo, že pro sklizně tak velkých objemů dat, je Record Manager, který vyvinula finská národní knihovna pro sklizeň dat do portálu Finna a který MZK v letošním roce používala pro testování OAI-PMH ostatních knihovních systémů nedostatečný. Record Manager je proto nyní upravován tak, aby zvládal lépe výpadky během přenosu dat, aby pracoval efektivněji s ohledem na plánované objemy dat, které budou přenášeny a případně i aby byl schopen počáteční sklizeň rozdělit do více částí.
* Během testování jsme pomáhali na vyžádání ladit konfiguraci OAI provideru jednotlivým knihovnám. Vyzkoušeli jsme, že maximální množství záznamů posílané v rámci jedné dávky tak, aby přenos byl optimální, je 50. Nejideálnější je, aby se množství odesílaných záznamů v rámci jedné dávky řídilo délkou záznamů a tedy objemem přenášených dat spíše než počtem záznamů.

## NCIP

* Posun ve vývoji protokolu NCIP pro účely CPK nebyl zdaleka tak velký, jak jsme očekávali. Abychom mohli otestovat funčnost pracovní skupinou navrženého NCIP rozhraní, vyvinuli jsme vlastní geteway mezi xcnip2toolkitem a API Alephu. Toto řešení bylo funkční natolik, abychom pronikli do hloubky NCIP protokolu a v návaznosti na tom mohli vyvinout NCIP konektor pro vufind, který dokáže toto rozhraní využívat a dokáže tak v reálném provozu otestovat implementace NCIP v jednotlivých knihovních systémech.
* Přestože je tedy nyní možné připojit přes NCIP do centrálního portálu ostatní alephovské knihovny, neplánujeme toto rozhraní pro ně použít, protože by šlo jen o vložení nadbytečné vrstvy, která je vhodná pro testování, ale v ostrém provozu by mohla vnést do systému zbytečná rizika. Alephovské knihovny budou proto nadále připojovány přímo přes Aleph Restful API, aby toky dat co nejkratší.
* Nabyté zkušeností nám umožní snadněji připojit do portálu ostatní knihovní systémy, jakmile budou mít jednotlivé knihovny NCIP zprovozněný.

## Další zkušenosti

* Obecně jsme získali mnoho zkušeností jak se zpracováním větších objemů dat, tak s různými způsoby jejich prezentace. Za klíčovou považujeme především srozumitelnou, resp. pochopitelnou prezentaci sloučených údajů a zároveň optimalizaci indexů tak, aby provoz celého portálu byl efektní.

## Závěr

Řešení obou projektů pomohlo ověřit reálný stav připravenosti knihoven na vstup do Centrálního portálu. Zjištěný stav a získané zkušenosti řešitelského týmu tohoto projektu přispěly k přehodnocení způsobu zajištění portálu, jehož vývoj by měla Moravská zemská knihovna zahájit v roce 2015.