

Projekt Národní digitální knihovna – aktuální s tav projektu

RNDr. Tomáš Svoboda

Národní knihovna ČR,

tomas.svoboda@nkp.cz

INFORUM 2012: 18. konference o profesionálních informačních zdrojích



Abstrakt

Příspěvek popisuje aktuální stav projektu "Národní digitální knihovna" v květnu 2012, včetně informací o historii projektu a jeho technickém řešení.

POPIS A CÍLE PROJEKTU

Národní digitální knihovna je souhrn aktivit, které provozuje Národní knihovna ČR spolu s Moravskou zemskou knihovnou a dalšími partnerskými institucemi směřující k digitalizaci a zpřístupnění národního knihovního bohatství. Její nejvýraznější složkou je dotační projekt "Vytvoření Národní digitální knihovny" spolufinancovaný ze Strukturálních fondů EU (Evropského fondu pro regionální rozvoj) prostřednictvím Integrovaného operačního programu MV. V rámci tohoto dotačního projektu Národní knihovna České republiky a Moravská zemská knihovna v Brně zdigitalizují, dlouhodobě ochrání a zpřístupní významnou část svých fondů.

Projekt Národní digitální knihovna v sobě zahrnuje tři hlavní linie:

- 1. Digitalizace významné části bohemikální produkce 19.-21. století,** tj. knih vydaných na území České republiky, napsaných v češtině nebo pojednávající o Česku. Celkem do konce roku 2019 zdigitalizujeme více než 50 mil. stran, tedy přibližně 300 000 svazků.
- 2. Dlouhodobé uložení dokumentů ve spolehlivém digitálním úložišti.** Úložiště poskytne prostor pro bezpečné umístění dosud digitalizovaných dokumentů i digitálních dokumentů vytvořených či získaných v rámci dalších projektů.
- 3. Zpřístupnění digitálních dokumentů.** Pokud to autorské právo dovolí, bude zdigitalizovaný materiál zpřístupňován volně a zdarma.

Co budeme přednostně digitalizovat a zpřístupňovat?

- Dokumenty 19. a 20. stol., které jsou nejvíce ohroženy degradací kyselého papíru
- Uživatelsky nejzajímavější a nejžádanější publikace od r. 1801 do současnosti
- Dokumenty volně přístupné široké veřejnosti nebo prostřednictvím českých knihoven
- Další typy dokumentů podle kritérií významu, unikátnosti, fyzického stavu apod.

Projekt je financován částkou cca 300 mil. Kč v rámci Výzvy č. 07 Integrovaného operačního programu "Elektronizace služeb veřejné správy", prioritní osa 1: Modernizace veřejné správy, oblast intervence 1.1 Rozvoj informační společnosti ve veřejné správě.

Nositelem projektu je Národní knihovna ČR, partnerem je Moravská zemská knihovna v Brně.

ZÁKLADY

Projekt NDK navazuje na doposud realizované digitalizační projekty probíhající v České republice.

Digitální data jsou generována především v rámci tří velkých národních projektů.

Manuscriptorium (<http://www.manuscriptorium.com>) je systém pro vytváření sbírek a zpřístupnění informací o historických a vzácných dokumentech na Internetu, včetně virtuální digitální knihovny digitalizovaných dokumentů.

Kramerius (<http://kramerius.nkp.cz>) se zaměřoval na ochranu a zpřístupnění periodik, knih a ostatních dokumentů publikovaných od roku 1801. Velká část těchto dokumentů je silně ohrožena v důsledku tisku na kyselém papíře a/nebo častého používání.

WebArchiv (<http://www.webarchiv.cz>) je digitální archiv českých webových zdrojů, které jsou shromažďovány s cílem jejich dlouhodobé ochrany a zpřístupnění.

a dále jako další zdroj digitálních dat přichází smluvní spolupráce s firmou **Google** (v programu GoogleBooks), která digitalizuje části sbírek NK starší než 1800.

HISTORIE A POSTUP PRACÍ

PŘÍPRAVA

V únoru 2010 podala Národní knihovna ČR (NK ČR) společně s Moravskou zemskou knihovnou (MZK) jako partnerem Projekt „Vytvoření Národní digitální knihovny“ (NDK). Projekt byl podán v rámci Výzvy 07 Integrovaného operačního programu „Elektronizace služeb veřejné správy“. V červnu 2010 by projekt schválen. Jedná se o jeden ze základních stavebních kamenů konceptu eCulture, kterým sektor kultury významně přispívá k naplňování cílů Smart Administration.

Projekt NDK je financován z Integrovaného operačního programu EU částkou 255 milionů korun a spolufinancován z rozpočtu MK ČR částkou 45 milionů korun. Příspěvek 85% ze strukturálního fondu ERDF ve výši 254 946 300 Kč je doplněn 15% kofinancováním ze státního rozpočtu ve výši 44 990 700 Kč. Celkové způsobilé veřejné výdaje tedy činí 299 937 000 Kč.

NK ČR a MZK uchovávají ve svých fondech díky právu úplného povinného výtisku většinu monografií, periodik a dalších druhů dokumentů publikovaných na našem území (bohemika v užším slova smyslu), velké množství dokumentů vztahujících se k ČR publikovaných v zahraničí (bohemika v širším slova smyslu) a spravují bohaté historické fondy. Od roku 2000 spolupracují i na podchycení českého webu. Disponují tedy rozsáhlým a zároveň unikátním materiálem jedinečné kulturní, ale s ohledem na kontext Smart Administration především faktografické hodnoty.

PROJEKTOVÁ A INVESTIČNÍ ČÁST

V roce 2010 byl připraven obsahový základ projektu. V procesu bylo rozhodnuto o rolích externích dodavatelů: Projektový manažer a Systémový integrátor s tím, že tyto role převezmou komerční subjekty vzešlé z výběrových řízení.

Na začátku roku 2011 došlo ke změně hlavního koordinátora projektu a k urychlenému plnění harmonogramu projektu.

Prvním úkolem bylo uskutečnit dvě klíčová výběrová řízení na externí dodavatele služeb Projektového managementu a Systémového integrátora (neboli Generálního dodavatele). Výběrové řízení na dodavatele služeb Projektového managementu bylo uskutečněno jako první. Do soutěže přihlásilo deset uchazečů, všichni splnili kvalifikační požadavky zadavatele a žádný z nich nebyl vyloučen. Jako nejlepší byla vybrána nabídka firmy PragoData Consulting, s. r. o., s níž byla 30. 6. 2011 podepsána smlouva. Od 1. 7. 2011 PragoData převzala agendu projektu a připravila projektový tým na výběr Systémového integrátora.

Výběrové řízení na Systémového integrátora – generálního dodavatele technologií projektu – bylo podstatně složitější: Od doby vzniku projektu bylo potřeba zohlednit vývoj technologií, změny právních předpisů a rozpočet, který byl nižší oproti původním předpokladům z roku 2008. Výběrové řízení bylo vyhlášeno dne 4. 7. 2011, lhůta pro přijímání nabídek byla stanovena do 23. 8. 2011. Dva ze čtyř uchazečů byli vyloučeni, protože nesplnili zadání výběrového řízení. Sedmičlenná hodnotící komise jmenovaná ministrem kultury posuzovala nabídky jednotlivých uchazečů podle následujících hodnotících kritérií:

- výše nabídkové ceny (váha 50 %),
- kvalitativní, funkční a technické vlastnosti (váha 30 %),
- rozsah, kvalita a technické parametry služeb systémové integrace (váha 20 %).

Vítězem se stala firma Logica Czech Republic, s. r. o., čímž vznikl kompletní projektový tým, který začal pracovat na realizaci projektových cílů. Činnost firmy Logica byla zahájena kampaní intenzivních jednání mezi týmy zadavatele, Systémového integrátora a Projektového manažera, jejichž vyvrcholením bylo v závěru roku zpracování Prováděcího projektu, který se stal základem pro realizaci technologické části projektu pro následující období. Prováděcí projekt byl podroben velmi důkladnému připomínkovému řízení, v němž zadavatel uplatnil přes 700 připomínek, které byly postupně dodavatelem vypořádávány tak, že do nového roku projekt vstoupil připraven k akceptaci s jasnou perspektivou nadcházejících technologických a organizačních činností.

Práce na projektu, na kterých se podílí zadavatel i externí dodavatelé, jsou organizovány v tematických Pracovních skupinách pro jednotlivé obory:

- pro Digitalizaci – zabývá se problematikou skenování a navazujících činností
- pro Dlouhodobou ochranu – řeší otázku jak zachovat digitální dokumenty čitelné a srozumitelné pro budoucnost
- pro Zpřístupnění – zajišťuje zveřejňování digitálních dokumentů uživatelům
- pro infrastrukturu ICT – stará se o technickou stránku projektu
- Projektová kancelář – zajišťuje administraci projektu

FÁZE DODÁVEK:

Právě probíhá a její součástí jsou následující činnosti:

- Instalace technologií - do 30.4. dodáno investiční vybavení a začíná integrace.

- Výkonnostní testy komponent - probíhají testy, které ověřují reálné výkonnosti skenerů a dalších hardwarových a softwarových komponent.
- Integrační testy
- Výkonové testy výrobních linek - proběhnou 08/2012 a jejich cílem je v režimu plného provozu vyzkoušet výkonnost linky jako celku.

OD 2013: PLNÝ PROVOZ.

Už dříve v průběhu roku 2012 během testů je simulován reálný provoz a jeho výsledky jsou zahrnuty do výstupů projektu.

TECHNICKÉ ŘEŠENÍ

DIGITALIZACE A POSTPROCESSING

Dokument vybraný pro digitalizaci prochází zpracováním v subsystému digitalizace s použitím nástrojů tzv. „digitalizačního workflow“; první metadata se načtou z knihovního systému Aleph pomocí načtení čárového kódu a předávají se do digitalizačního workflow a do Registru digitalizace (dále RD); digitalizační workflow také přiděluje dokumentům identifikátory (URN:NBN), které dále spravuje aplikace Resolver URN:NBN.

Skenery (převážně robotické) produkují sadu obrázků stránek představující daný dokument. Zvláštním případem je destruktivní digitalizace postradatelných dokumentů prováděná dokumentovým skenerem.

Skenery: DL-3003, DL-mini-I, Treventus, Canon a další

V postprocessingu jsou obrázky (stránky) ořezány, zorganizovány a procházejí procesem OCR.

Výstupem je balíček obsahující ořezané stránky v JPEG2000 a soubory ALTO-XML (výstup OCR) pro další zpracování.

TRANSFORMACE

V pracovním prostoru balíčky PSP dále zpracovává transformační modul.

Jsou zkontrolována metadata a vytvořeny:

- balíček SIP1 pro LTP systém a
- data pro zpřístupnění – PDF soubory a vstupní balíčky pro aplikaci Kramerius

Transformační modul přiděluje URN:NBN externím dokumentům, které nepřicházejí z digitalizačního workflow.

Transformační modul sleduje tok dokumentů z externích zdrojů a kontroluje konzistenci UC a MC mezi LTP a aplikacemi zpřístupnění.

V případě pozdější změny dat nebo metadat dojde k úpravě archivní kopie a opakované vygenerování uživatelských kopií.

Data z projektů Manuscriptorium a WebArchiv nebudou v transformačním modulu upravována pro zpřístupňující aplikace. Zde vložení do zpřístupňující aplikace probíhá „nezávisle na projektu NDK“ mimo transformační modul, před či po vložení dat do LTP systému.

Data z jiných zdrojů určená k archivaci a zpřístupnění v systému NDK jsou vkládána do pracovního prostoru transformačního modulu, který zajišťuje jejich převod do balíčků SIP1 a SIP2 a jejich odeslání do LTP a aplikací zpřístupnění.

LTP

Podle OAIS.

Dodavatelem je AipSafe. Výhodou české firmy je značná flexibilita v organizaci implementačních prací a v akceptaci požadavků zákazníka.

Fyzická bezpečnost:

Archivní data uchovávána na páskách ve 3 oddělených lokalitách, periodické kontroly a údržba probíhají podle automatizovaného schématu v systému LTP.

Formátová údržba:

Systém LTP si udržuje přehled o všech formátech uložených v úložišti, periodicky sleduje jejich udržitelnost a navrhuje provést prezervační úkony.

ZPŘÍSTUPNĚNÍ

- Soubor PDF vygenerovaný transformací z archivní kopie je zpřístupněn na webu na trvalé adrese.
- Aplikace Kramerius (v NK ČR i MZK) zpracuje balíček SIP2 a zajistí zpřístupnění uživatelských kopií v prostředí webové aplikace.
- Zpřístupňovány jsou uživatelské kopie vyprodukované oběma pracovišti digitalizace (Praha, Brno) i uživatelské kopie externích dat.