Petr Jirásek

3. ročník, obor Umělá inteligence a zpracování přirozeného jazyka

30. 11. 2012

esej na téma

Národní digitální knihovna

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

Oficiální stránky: http://www.ndk.cz/

Obsah

1. ÚVOD

V rámci této eseje bych se rád zabýval projektem Národní digitální knihovny, jehož snahou je opravdu rozsáhlá a konceptuální digitalizace kulturního bohatství naší země, která je zcela nezbytná v dnešní době informačních médií.

V rámci této práce je kladen důraz především na to, jaká je vlastně vize tohoto projektu, jaké výsledky a cíle jsou očekávány a také to, jakým způsobem jsou zajištěny některé vnitřní procesy v kombinaci s použitými technologiemi.

2. CHARAKTERISTIKA PROJEKTU

Národní digitální knihovna je souhrn aktivit Národní knihovny ČR spolu s partnerskými institucemi, mezi které kupříkladu patří Moravská zemská knihovna v Brně. Mezi tyto aktivity patří především snaha o digitalizaci knihovního bohatství, především významné části bohemikální produkce 19. až 21. století, dlouhodobé uložení dokumentů ve spolehlivém digitálním uložišti a zpřístupnění těchto dokumentů veřejnosti v souladu s autorskými právy daného díla.

3. PRŮBĚH

V únoru 2010, dva roky po přípravě projektové dokumentace a vize celého projektu, byl tento projekt započat podáním návrhu „Vytvoření národní digitální knihovny“ v rámci výzvy 07 Integrovaného operačního programu „Elektronizace služeb veřejné správy“, který byl poté v červnu téhož roku schválen.

Co se týče ekonomické stránky, projekt byl financován přibližně částkou 300 miliónů korun, přičemž 85 % této částky je čerpáno ze strukturálního fondu ERDF a 15 % z rozpočtu Ministerstva kultury ČR.

Projekt je rozdělen do těchto etap:

Etapa	Název etapy	Popis etapy	Od	Do
1.	Přípravná etapa	Příprava projektu, analýzy, studie proveditelnosti, vytvoření týmu projektu	01/2008	07/2010
2.	Investiční etapa 1	Výběr systémového integrátora, dodávky digitalizační technologie do MZK, instalace LTP, dodávky technologie pro digitalizaci do NK ČR.	08/2010	06/2012
3.	Investiční etapa 2	Integrace komponent digitalizace, LTP, testování procesů, školení pracovníků	07/2012	09/2012
4.	Poloprovozní etapa	Ověřovací provoz, vyhodnocení poloprovozu, digitalizace a tvorba metadat, konverze existujících dat	10/2012	12/2012
5.	Provozní etapa 1	Ověřovací provoz, vyhodnocení poloprovozu, digitalizace a tvorba metadat, konverze existujících dat	01/2013	12/2013
6.	Provozní etapa 2 a ukončení projektu	Plný provoz digitalizace, certifikace úložiště, harvesting do Europeany, ukončení administrace projektu a vyhodnocení	01/2014	12/2014

Dále, od roku 2015 – 2019 se předpokládá dlouhodobá udržitelnost projektu a od roku 2019 jeho pokračování.

4. CÍLE PROJEKTU

Hlavním cílem projektu je především digitalizovat bohemikální literaturu 19. – 21. století. Jedná se tedy o knihy vydané na území České republiky, napsané v češtině nebo pojednávající o Česku. Předpokládá se, že do konce roku 2019 bude zdigitalizováno více než 50 miliónů stran, což odpovídá zhruba 300 000 svazkům.

Významnou součástí projektu je také archivace dokumentů českého webu v rámci projektu NDK. Jedná se o dokumenty publikované v digitální podobě a v průběhu trvání projektu bude podchyceno cca 4 000 000 000 souborů, které tvoří webové stránky.

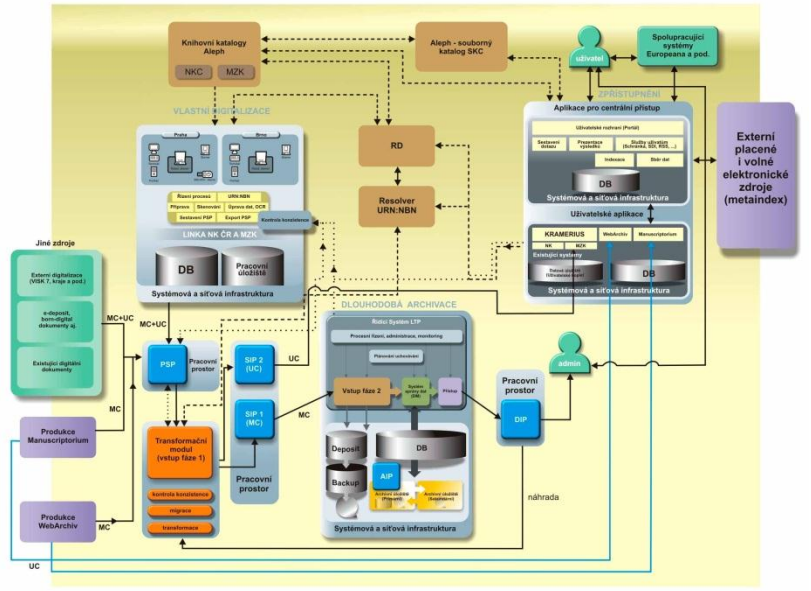
Dalším cílem je vybudovat systém pro zpřístupnění digitálního obsahu uživatelům a zprovoznění důvěryhodného a především dlouhodobě udržitelného repozitáře.

5. POPIS PROJEKTU

5.1. Repozitář

Systém repozitáře NDK vychází z požadavků na důvěryhodný digitální repozitář, tj. odpovídá konceptuálnímu modelu OAIS[1], přičemž základní složky systému dlouhodobé archivace v rámci NDK obsahují následující funkční submoduly:

- Subsystém digitalizace
- Subsystém dlouhodobého uložení dokumentů (LTP subsystém)
- Subsystém pro transformace a kontroly konzistence (Transformační modul)
- Subsystém zpřístupnění informací a dokumentů



Plné rozlišení: http://www.ndk.cz/obrazky/ndk_new/

V rámci dalších kapitol budou tyto základní submoduly zmíněny.

[1] Open Archival Information System

5.2. Digitalizace

Digitalizace je prováděna ve dvou lokalitách a to v Praze - Hostivaři a v Brně. Tam dochází ke skenování fyzických dokumentů, jako jsou papírové dokumenty, mikrofilmy apod. Tyto dokumenty v digitální podobě jsou poté doplněny o nezbytná metadata popisující technické, strukturální i administrativní informace. Tato metadata jsou buď tvořena manuálně nebo automaticky, přičemž v tomto případě jsou využity katalogy Národní knihovny a Moravské zemské knihovny a i jiné databáze, z kterých se metadata importují.

V případě obrazových materiálů může docházet také k jejím úpravám, jakou jsou ořezy, posuny apod.

V závěru jsou všechna tato data poslána jako kompletní balíček systému LTP[2].

Vybavení pro digitalizaci

Projekt disponuje tímto hardwarovým vybavením:

- Robotické skenery
- 2x 4DigitalBooks DL-3003
    - velké formáty
- 2x 4DigitalBooks DL-mini-I
- 4x Treventus
    - obtížnější svazky
- 2x Canon DR-X10C (na volné listy)
    - destruktivní digitalizace
- Ruční skenery

Konečný typ skeneru, který se použije pro digitalizaci konkrétního dokumentu, zcela závisí na charakteru tohoto dokumentu, přičemž obzvlášť u děl vysoké historické hodnoty, která jsou důsledkem svého stáři velice citlivá na práci s nimi, je třeba vybrat takové vybavení, které bude maximálně šetrné.

V rámci softwarového vybavení se využívá:

- SIRIUS
    - tvorba metadat a zpracování obsahových souborů
- Corel Xmetal
    - validace XML
- Adobe Photoshop
    - drobné úpravy obrazových dat
- Document Express DJVu Edition
    - konverze mezi JPEG a DJVu
- Abbyy FineReader
    - OCR[3]

Výstup digitalizace

Obrazy stránek jsou uchovávány v bezztrátovém formátu JPEG–2000 a rozpoznaný text systémem OCR je uložen ve formátu ALTO–XML.

[2] Long-Term Preservation
[3] Optical Character Recognition

5.3. Dokumenty z jiných zdrojů

Dokumenty, které budou uloženy v repozitáři, nemusí pocházet pouze z digitalizace provedené na pracovištích v Brně a Praze, ale mohou být importovány i z jiných zdrojů. Tyto dokumenty třetích stran, které již mají digitální podobu, pak prochází tzv. transformačním modulem (viz. Kapitola Systém pro transformaci a kontrolu konzistence).

NDK kupříkladu čerpá ze zdrojů jako je WebArchiv a Manuscriptorium.

WebArchiv

WebArchiv je digitální archiv „českých“ webových zdrojů a v rámci aktivit projektu NDK je také zahrnut.

Data z tohoto archivu budou uložena ve formátu ARC nebo WARC souborů o velikosti přibližně 100 MB. Předpokladem jsou dvě sklizně „českého webu“ ročně, přičemž po ukončení sběru se budou data ukládat do LTP systému a do systému zpřístupnění.

Během 5ti let bude zpracováno přibližně 1 572 souborů denně. Integrace souborů, které vzejdou z činnosti WebArchivu, nebude probíhat kontinuálně, ale dávkově.

Manuscriptorium

Projekt Manuscriptorium disponuje digitálními dokumenty z oblasti historických fondů, jako jsou rukopisy, mapy, listiny apod. I v tomto případě bude snaha o to, tyto díla ukládat do LTP systému.

5.4. Systém pro transformaci a kontrolu konzistence

Tento systém zajišťuje, aby všechna data vstupující do LTP systému byla ve stejném formátu v závislosti na vstupním charakteru informací (dokumenty, obrázky apod.). Data, která jsou již digitalizovaná, pocházejí z jiných zdrojů apod., musí touto transformací projít tak, aby se docílilo logické ochrany LTP systému a udržela jeho konzistence.

5.5. Dlouhodobá ochrana digitálních dat

Dlouhodobá ochrana digitálních dat probíhá na několika úrovních tak, aby nemohlo dojít k nenávratné ztrátě těchto informací.

V rámci fyzické bitové ochrany se ukládají data na páskách ve třech duplikátech v oddělených lokalitách, přičemž se pravidelně provádí jejich kontrola a údržba podle automatizovaného schématu v systému LTP.

Dále se provádí i tzv. logická (formátová) ochrana, kde je cílem zachovat přístup k těmto digitálním datům uživatelům, v závislosti na technickém prostředí té doby a umožnit jak vyhledávání, tak porozumění obsahu a smyslu. Systém LTP si tak udržuje přehled o všech formátech uložených v uložišti, sleduje jejich udržitelnost a navrhuje provádění prezervačních úkonů.

5.6. Zpřístupnění

Zpřístupnění digitalizovaných dat probíhá ve dvou liniích. První z nich spočívá v umístění PDF výstupu na trvalou adresu tak, aby bylo umožněno vyhledávačům tyto soubory indexovat a zobrazit ve výsledcích svého vyhledávání. Druhou linií je zpřístupnění prostřednictvím aplikace Kramerius.

Dále je v současné době vytvořena aplikace pro centrální přístup, která zajistí pohodlný přístup k dokumentům z různých aplikací jejím prostřednictvím. Díky tomu je možné vyhledávat ve více zdrojích najednou, aniž by uživatel tušil, ve kterém z nich by měl daný dokument hledat. Navíc, pokud se informace o dokumentu budou nacházet na více místech, pak tyto související informace budou u dokumentu také uvedeny.

6. VÝSLEDKY

Projektovým manažerem se na základě výběrového řízení stala společnost PragoData Consulting s r.o. V rámci výběrového řízení na systémového integrátora a dodavatele koncem roku 2011 zvítězila Logica Czech Republic s r.o. ve spolupráci s Albertina icome Praha (AiP), přičemž dalšími zájemci byly společnosti jako IBM, Techniserv IT nebo Ness.

Subdodavatelem vybavení pro digitalizaci (viz. Vybavení pro digitalizaci v kapitole Digitalizace) byla společnost NUPSESO CZ. V oblasti pro dlouhodobou ochranu dat byla zvolena AiP Safe a jejich řešení.

Během připomínkového řízení, které probíhalo koncem roku 2011 k prováděcímu projektu, který byl připravován systémovým integrátorem, bylo vypořádáno přes 700 připomínek.

V současné chvíli se celý projekt nachází v pilotní fázi, přičemž začátkem roku dojde k zintenzivnění práce. Pilotní provoz slouží primárně k vyladění pracovních procesů a odladění softwaru.

Také již existuje neveřejná nová verze aplikace Kramerius, kde je možné sledovat již digitalizovaná data. Všechny submoduly repozitáře jsou již implementovány, ač některé jeho součásti nejsou stále veřejné a prochází procesem dolaďování.

Aplikace pro centrální přístup je řešena prostřednictvím systému Vufind, přičemž dodavatel dosud aplikaci nedal k dispozici veřejně. NDK ale nebude využívat přímo jejich verzi, jelikož dodavatel pracuje s verzí 1.3 a v současnosti je připravován přechod na verzi 2.0. Proto tato vyšší verze aplikace bude pouze přebírat relevantní data z aplikace dodavatele.

Dosud bylo zdigitalizováno přes 200 000 stran dokumentů, přičemž podle interního plánu se předpokládá, že by mělo být do konce tohoto roku zdigitalizováno 2 %, tedy 520 000 z celkových 26 miliónů, které by měly být hotovy do konce roku 2014. Každé z pracovišť (Praha a Brno) by tak mělo zdigitalizovat 13 miliónů stran, přičemž dvě směny na jednom pracovišti by měly denně zpracovat 26 400 stran.

Odhaduje se, že cíl pro tento rok bude splněn. Jako největší překážka v plnění těchto cílů se zatím jeví nedostatek personálu.

7. VLASTNÍ ZHODNOCENÍ

Doufám, že výsledná úroveň centrální aplikace, která je v současné době neveřejná, svou propracovanosti umožní uživatelům využít maximální potenciál tohoto projektu. A to ve všech směrech. Věřím, že nebude kladen důraz pouze na funkcionální část, ale také na část uživatelského rozhraní. Jestliže bude rozhraní špatně navržené a pro uživatele těžko přístupné a to především z pohledu jeho ovládání, byla by to skutečně velká škoda, jelikož by zbytečně vznikla překážka pro širokou veřejnost v užívání tak zajímavé služby, jakou NDK může být.

Dále oceňuji snahu o digitalizaci národního kulturního bohatství a považuji ji i z hlediska moderní informační společnosti za zcela nezbytnou. Umožňuje nejen uživatelům zpřístupnit zajímavá kulturní díla, ale také díla, která jsou často zchátralá svým stářím, přičemž budou dlouhodobě uchovány pro další generace a budou k dispozici k dalšímu studiu a analýze.

Také oceňuji kroky NDK ve směru projektu WebArchiv, jelikož v dnešní technologicky pokročilé době kulturní dědictví nemusí mít pouze fyzickou formu, ale také formu elektronickou a jsem skutečně rád, že zde existují instituce, které si to uvědomují.

8. PODĚKOVÁNÍ

Děkuji panu Ing. Petrovi Žabičkovi za poskytnutí informací ohledně aktuálního stavu projektu a současných výsledků.

9. ZDROJE

http://www.ndk.cz/narodni-dk
http://knihovna.nkp.cz/knihovna91/humesto.htm
http://www.digitalpreservationeurope.eu/publications/presentations/KoncepceStav.pdf
http://www.inforum.cz/pdf/2012/svoboda-tomas.pdf
http://www.ndk.cz/narodni-dk/dotacni-projekt-vytvoreni-narodni-digitalni-knihovny/studie-proveditelnosti
http://www.ndk.cz/narodni-dk/prezentace-k-projektu-iop/inforum-2012/narodni-digitalni-knihovna
http://www.skipcr.cz/akce-a-projekty/dokumenty/akm-2011/Svoboda.pdf
http://www.manuscriptorium.com/index.php?q=cs
http://www.ndk.cz/narodni-dk/prezentace-k-projektu-iop/digitalna-kniznica-29-3-2011-jasna-pod-chopkom/tomas-foltyn-nove-digitalizacni-pracoviste-narodni-knihovny-cr-a-jeho-hardwarove-a-softwarove-vybaveni/view

10. METADATA V DC

<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<meta name="DC.Title" content="Národní digitální knihovna" />
<meta name="DC.Creator" content="Petr Jirásek" />
<meta name="DC.Subject" content="Národní digitální knihovna" />
<meta name="DC.Subject" content="digitální knihovna" />
<meta name="DC.Date" content="2012-12-30" />
<meta name="DC.Description" content=" Tato esej pojednává o tom, jaká je vlastně vize projektu Národní digitální knihovny, jaké výsledky a cíle jsou očekávány a také to, jakým způsobem jsou zajištěny některé vnitřní procesy v kombinaci s použitými technologiemi." />
<meta name="DC.Format" scheme="IMT" content="application/pdf" />
<meta name="DC.Source" scheme="URL" content="http://www.ndk.cz/>
<meta name="DC.Language" scheme="RFC3066" content="cze" />