Budování vzájemně kompatibilních informačních systémů pro přístup k heterogenním informačním zdrojům a jejich zastřešení prostřednictvím Jednotné informační brány

Zpráva o výsledcích řešení výzkumného záměru v roce 2010

PhDr. Bohdana Stoklasová, hlavní řešitelka
Ing. Libor Coufal, Mgr. Jan Hutař,

Národní knihovna České republiky
Klementinum 190
110 00 Praha 1

14. prosince 2010

OBSAH

A Konstatační část

Úvodní poznámka vztahující se ke struktuře předkládané zprávy:

Předmětem výzkumné činnosti realizované ve výzkumném záměru *Budování vzájemně kompatibilních informačních systémů pro přístup k heterogenním informačním zdrojům a jejich zastřešení prostřednictvím Jednotné informační brány* je výzkum a vývoj směřující k vytvoření informačních systémů pro přístup k heterogenním informačním zdrojům, které budou navzájem kompatibilní do té míry, že bude možné je zastřešit tak, že se budou navenek (tj. pro koncového uživatele) prezentovat jako systém jediný. Jedná se o komplexní výzkumný záměr, který se v závěrečném roce řešení 2010 soustředí na oblast budování digitálních depozitních knihoven s ohledem na možnost jejich integrace v rámci Jednotné informační brány a nadnárodních portálů. Celkové výsledky a přínosy jsou shrnuty v kapitole B.2 Přínos řešitele a B.3 Posun znalostí. ***Výsledky řešení roku 2010, které budou uplatněny v rámci registrace výsledků (RIV) v roce 2011, jsou v textu uvedeny tučnou kurzívou a jsou shrnuty v kapitole C.1.***

A.1 Rešerše

Rešerše obsahuje publikační činnost řešitelů a dalších pracovníků NK ČR vztahující se k řešenému tématu za rok 2010. Je uspořádána tematicky podle jednotlivých oblastí, v jejich rámci abecedně podle jmen autorů. Publikace, které budou uplatněny v rámci registrace výsledků (RIV) v roce 2010, jsou uvedeny tučně.

Budování digitálních depozitních knihoven s ohledem na možnost jejich integrace v rámci Jednotné informační brány a nadnárodních portálů

COUFAL, Libor. Licence Creative Commons a jejich využívání ve školství a vzdělávání. In *LinuxExpo/Open Source Conference* [online]. Praha : Exponet, 20. 4. 2010 [cit. 2010-11-04]. Dostupný z WWW: <http://webarchiv.cz/files/dokumenty/konference/cc_ve_vzdelavani.pdf>.

COUFAL, Libor. Možnosti automatizované akvizice elektronických online časopisů pomocí nástrojů pro archivaci webu. In *20. seminář akvizičních pracovníků* [online]. Praha : Sdružení knihoven ČR, 17. 6. 2010 [cit. 2010-11-04]. Dostupný z WWW: <http://webarchiv.cz/files/dokumenty/konference/NTK_Akvizicni_seminar_2010.pdf>.

COUFAL, Libor. Preserving web archives : one size fits all? In *iPres 2010 : 7th international conference on preservation of digital objects : September 19-24, 2010, Vienna, Austria* [online]. Vídeň : iPres, 2010 [cit. 2010-11-04]. Diskuzní panel členů pracovní skupiny Preservation Working Group při IIPC (International Internet Preservation Consortium). Program dostupný z WWW: <http://www.ifs.tuwien.ac.at/dp/ipres2010/schedule.html>.

CUBR, Ladislav. Budování důvěryhodného systému trvalé identifikace digitálních dokumentů. *Knihovna*. 2010, roč. 21, č. 1, s. 23-31. ISSN 1801-3252.

CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů*. Praha : Národní knihovna ČR, 2010. 154 s. ISBN 978-80-7050-588-5 (brož).

CUBR, Ladislav; HUTAŘ, Jan; MELICHAR, Marek. Národní knihovna a předpoklady pro českou infrastrukturu trvalé identifikace digitálních dokumentů. In *Seminář ke krajské digitalizaci, NTK 1.3.2010 Praha*. 31 slidů. Dostupné z WWW: <http://pid.ndk.cz/dokumenty/prezentace-z-konferenci/narodni-knihovna-a-predpoklady-pro-ceskou-infrastrukturu-trvale-identifikace-digitalnich-dokumentu>.

GRUBER, Lukáš. *All rights reserved/Some rights reserved : beseda o free culture* [online]. Praha : Centrum studijních a informačních služeb FF UK, 2010 [cit. 2010-11-04]. Beseda o autorském právu, šíření informací a volném užívání děl. Dostupný z WWW: <http://iforum.cuni.cz/IFORUM-9871.html>.

GRUBER, Lukáš. Licence Creative Commons. In *Jak využít Open Access ve vaší publikační činnosti* [online]. Praha : VŠE, 2010 [cit. 2010-11-04]. Přednáška konaná dne 19. 10. 2010 v rámci Open Access Week. Popis přednášky dostupný z WWW: <http://www.vse.cz/zpravy/news.php?ID=11722>.

GRUBER, Lukáš. Licence Creative Commons. In *Open Access a Creative Commons : seminář konaný dne 28. 4. 2010* [online]. Brno : Centrum PARTSIP, 2010 [cit. 2010-11-04]. Popis semináře dostupný z WWW: <http://www.partsip.cz/akce/open-access-creative-commons-seminar>.

GRUBER, Lukáš. Manifest o volných dílech. *Ikaros* [online]. 2010, roč. 14, č. 6 [cit. 2010-11-04]. Dostupný z WWW: <http://ikaros.cz/node/6255>. ISSN 1212-5075.

HUTAŘ, Jan. Dlouhodobá ochrana dat. In *34. celostátní seminář knihovníků muzeí a galerií při AMG, 8.9.2010 Česká Lípa*. 25 slidů.

HUTAŘ, Jan. NDK. *Prezentace pro krajský úřad Jihlava 7.5.2010*. Slide 58-81. Dostupné z WWW: <http://www.ndk.cz/narodni-dk/prezentace-k-projektu-iop/prezentace-ndk-pro-kraje-jihlava/prezentace-pro-kraje-jihlava-7-5-2010/view>.

HUTAŘ, Jan. Dlouhodobá ochrana digitálních dat: co může vaše instituce udělat již dnes. *Archivy, knihovny, muzea v digitálním světě. Praha, NA ČR, 2.12.2010*. [online]. [cit. 2010-12-04]. Dostupné z WWW: <http://skip.nkp.cz/KeStazeni/Archivy10/den2/Hutar.pdf>.

FOJTŮ, A. Grey Literature at the Charles University in Prague. *Seminář ke zpřístupňování šedé literatury* [online]. Praha : Národní technická knihovna, 2010 [cit. 2010-12-08]. Dostupný z WWW: <http://nusl.techlib.cz/index.php/Seminare>. ISSN 1803-6015.

FOJTŮ, Andrea. Open source nástroje pro dlouhodobou ochranu digitálních dokumentů. *Archivy, knihovny, muzea v digitálním světě. Praha, NA ČR, 2.12.2010*. [online]. [cit. 2010-12-04]. Dostupné z WWW: <http://skip.nkp.cz/KeStazeni/Archivy10/den2/Fojtu.pdf>.

VYCHODIL, Bedřich. Úvod do problematiky (skrytých) nákladů na dlouhodobou archivaci. *Archivy, knihovny, muzea v digitálním světě. Praha, NA ČR, 2.12.2010*. [online]. [cit. 2010-12-04]. Dostupné z WWW: <http://skip.nkp.cz/KeStazeni/Archivy10/den2/Vychodil.pdf>.

A.2 Současný stav ve světě a v ČR

Budování digitálních depozitních knihoven s ohledem na možnost jejich integrace v rámci Jednotné informační brány a nadnárodních portálů

Problematika budování digitálních úložišť je jedním z klíčových témat řešených v paměťových institucích všech zemí, které již nashromáždily určité objemy digitálních dat. Česká republika je v této oblasti vysoce ceněna v rovině koncepční, koncepce Národní digitální knihovny (NDK) včetně centrálního digitálního úložiště se těší trvalému mezinárodnímu zájmu a částečně se uplatnily i v rámci zahraničních koncepcí.

Ministerstvo kultury a česká vláda přijaly Národní digitální knihovnu za strategickou prioritu, financování je schváleno a zajištěno v rámci Integrovaného Operačního Programu – IOP (Smart Administration).

NK ČR spolu s Moravskou zemskou knihovnu v Brně má v rámci projektu NDK tři hlavní cílové linie:

- urychlení digitalizace (dvě digitalizační centra v Praze a v Brně, nasazení masové digitalizace)
- dlouhodobá ochrana digitálních objektů (zdigitalizovaných i digital born dokumentů) - důvěryhodný digitální repozitář
- komfortní zpřístupnění a práce s dokumenty ze strany uživatele

Pro konkrétnější představu o obsahu i rozsahu zmíněného projektu NDK uvádíme několik čísel: jádro českého národního kulturního dědictví (dokumenty publikované na našem území od roku 1801 včetně + historické dokumenty do roku 1800 uložené v českých knihovnách) tvoří přibližně 1,2 milionu dokumentů, což představuje 350 milionů stránek. Digitalizace tohoto množství současným tempem by trvala zhruba 300 let. Projekt umožní digitalizovat těchto 350 milionů stránek během 20 let. Nejohroženější a nejvyužívanější dokumenty by měly být digitalizovány během prvních pěti let projektu v letech 2009-2014. V rámci projektu zároveň budou bezpečně uložena stávající data získaná archivací webu a během trvání projektu budou ukládána data nová. NK plánuje se účastnit výzkumů o způsobech dlouhodobého uložení dokumentů vzniklých archivací internetových stránek.

Takto ambiciózní projekt nás posune na přední místo v evropském i celosvětovém kontextu, vyžaduje však důkladnou přípravu. **Řada výsledků řešení tohoto výzkumného záměru je důležitým základem a vstupem pro řešení projektu NDK.**

NK ČR byla spoluřešitelem projektu DigitalPreservationEurope[1], který je jedním z hlavních evropských projektů v této oblasti. Pracovníci NK ČR dále využívají získané kontakty s významnými institucemi/knihovnami, které řeší podobné projekty masové digitalizace a dlouhodobé ochrany digitálních dat. NK ČR je mj. jednou z účastnických knihoven v rámci skupiny usilující o vytvoření univerzální sady požadavků (RFI) na systémy dlouhodobé ochrany dokumentů.

Problematikou, která je s velkými objemy digitálních dat nedílně spojena, je jejich jednoznačná identifikace - ideálně globálně jednoznačným a persistentním identifikátorem. V digitálním světě je z mnoha důvodů (vyhledávání, citace, manipulace s daty, sdružování, dlouhodobá ochrana aj.) velmi důležité být schopen jednoznačně označit a posléze identifikovat digitální objekt nebo logické kombinace objektů. Zvláště pokud instituce uchovává miliony digitálních dokumentů. *Životně důležité jsou identifikátory pro různé agregátory dat, nebo služby vyhledávání z více zdrojů dat/metadat apod.*

[1] http://www.digitalpreservationeurope.eu/

Typů takovýchto identifikátorů je více, v různých knihovnách a archivech ve světě jsou využívány např. ARC (NK Francie, California Digital Library); Handle – velmi rozšířený placený systém převážně na amerických univerzitách; DOI – identifikátor a komerční systém založený na systému Handle; URN:NBN – identifikátor a systém založený na čísle národní bibliografie NBN.

Právě systém a kontext URN:NBN se na základě analýz z let 2008-2009 rozhodla používat a testovat NK ČR ve svém snažení o identifikaci digitálních objektů a případně i logických entit. URN:NBN je v současné době nejrozšířenějším systémem v Evropě na úrovni národních institucí (archiv, knihovna) a je využíván např. v Norsku, Finsku, Švédsku, Maďarsku, Německu, Rakousku, Švýcarsku atp.

V ČR se podobný systém zatím aktivně nevyužívá, některé instituce používají např. Handle a DOI, ale pouze pasivně jako příjemci těchto identifikátorů v metadatech dokumentů, nebo v rámci svých SW na repozitář, jako je např. DSpace[2], který je využíván na univerzitách pro uložení akademických prací (Ostrava).
Systém a vyvinutý SW pro správu, přidělování a resolvování identifikátorů by měl být základem národní služby pro URN:NBN, jak je tomu obvyklé v okolních státech (Německo).

Země různých částí světa, které mají určité zkušenosti s výzkumem a vývojem v oblasti archivace webu, spojují své síly a usilují o spolupráci, zejména na vývoji softwarových nástrojů a standardů. K tomuto účelu bylo v roce 2003 založeno konsorcium IIPC (International Internet Preservation Consortium), jehož členem se od května 2007 stala Národní knihovna České republiky. Naše mezinárodní aktivity a kontakty v této oblasti jsou velmi významné a naše výsledky snesou mezinárodní měřítka. NK ČR je zapojena i v evropském projektu LIWA – Living Web Archives[3].

Problematika vyhledávání v datech získaných sklízením webových stránek je velmi aktuální. Jen málo zemí nabízí prohledávání fulltextu, tj. kompletního indexu. NK takovýto index má a přes JIB jej lze prohledávat přes protokol SRU/SRW. Standard SRU/SRW je běžně používaný protokol pro vyhledávání v paměťových institucích. Jde o standardizovaný protokol pro vyhledávání realizovaný nad HTTP protokolem, výsledky dotazu jsou vráceny ve formě XML dokumentu odpovídajícímu standardu Dublin Core. Národní knihovna používá SRW/U (Search/Retrieve via the Web/URL) pro integraci fulltextu WebArchivu s Jednotnou informační bránou (JIB), Motivací pro nové řešení integrace fulltextu s Jednotnou informační bránou[4] byly následující nedostatky SRWLucene[5]:

- Neumožňuje shlukování výsledků podle domény
- Ve výsledcích vyhledávání chybí úryvky
- Rozlišuje mezi velkými a malými písmeny při vyhledávání, takže pro výraz Čapek nevrátí žádné výsledky, protože nutch všechna slova převádí na malá písmena.

Vzhledem k výše uvedeným nedostatkům SRWLucene jsme se rozhodli naimplementovat vlastní bránu mezi OpenSearch a SRU/SRW protokolem, která bude dostatečně obecná a bude možné jí integrovat se zdroji, které poskytují vyhledávání přes opensearch, jako je např. nutchwax.

[2] http://www.dspace.org/
[3] http://liwa-project.eu/
[4] http://www.jib.cz/V?RN=60521485
[5] http://code.google.com/p/oclcsrwlucene/

A.3 Vstupní data a cíl

Rekapitulace cílů uvedených v projektu pro jednotlivé oblasti a jejich zasazení do časového harmonogramu budou užitečnou pomůckou pro posouzení toho, které z vytčených cílů se podařilo/nepodařilo realizovat v roce 2010.

Budování digitálních depozitních knihoven s ohledem na možnost jejich integrace v rámci Jednotné informační brány a nadnárodních portálů

- Analýza existujících a nově vznikajících metadatových standardů (rovina bibliografická, administrativní, technická i ochranná) a návrh českých národních standardů (2006-2008).
- **Analýza digitálních depozitních knihoven v mezinárodním kontextu (2006-2010).**
- **Specifikace funkčních požadavků národního repozitáře s ohledem na jeho snadnou integraci v rámci portálů a Souborného katalogu ČR (2006-2009).**
    - Vzhledem k relevanci a posunům ve výzkumném záměru byla část problematiky u posledního cíle řešena i v roce 2010.

B Analytická část

B.1 Vlastní řešení

Pro snadnou orientaci jsou dosažené/plánované výsledky zvýrazněny v textu tučnou kurzívou.

3. Budování digitálních depozitních knihoven s ohledem na možnost jejich integrace v rámci Jednotné informační brány a nadnárodních portálů

- **Analýza digitálních depozitních knihoven v mezinárodním kontextu (2006-2010).**

V rámci příprav projektu Národní digitální knihovna proběhla v roce 2010, stejně jako předtím v letech 2008 a 2009, velká řada aktivit. Pokračovalo se v analýzách digitálních depozitních knihoven v mezinárodním kontextu. Spolupráce probíhala převážně s knihovnami, které jsou členy pracovní skupiny na vytvoření obecných požadavků (RFI) na LTP systém digitální depozitní knihovny, tj. národní knihovny Německa, Holandska, Velké Británie, Norska, Finska. Konzultace pokračují i s ostatními (např. Rakouská NK, Kongresová knihovna USA, NK Nového Zélandu, Španělska aj.) odkud má tým NK ČR a MZK nové informace o běhu obdobných projektů od tamních kolegů. Tým NK ČR prostudoval (ve spolupráci s týmem MZK, která je partnerem projektu) velké množství materiálů a navštívil stěžejní mezinárodní konference (IFLA, iPRES, Archiving aj.) věnované této problematice.

Vzhledem k blížícímu se výběrovému řízení na jednotlivé subsystémy a jejich systémovou integraci v rámci projektu NDK byly probíhající analýzy a konzultace velmi intenzivní. Jejich výstupem je lepší pochopení návazností a klíčových věcí pro budování digitální depozitní knihovny a pro dlouhodobou ochranu digitálních dat. Tyto poznatky se dají dále použít i mimo vlastní řešení projektu NDK, např. v jiných knihovnách a projektech. Budou také východiskem pro budoucí koncepci dlouhodobé ochrany dat pro ČR. Hmatatelným výstupem bude text zadávací dokumentace k projektu NDK (leden 2011), kde byly podstatně doplněny nároky na LTP systém, na digitalizační workflow a metadata. Jmenovitě metadata jsou určující pro další vývoj. Tak, jak budou využívána v projektu NDK (nové formáty = standardy využívané všude na světě), budou předepsána i v návazných projektech (např. VISK7 apod.). Ovlivní tedy interoperabilitu českých digitálních knihoven s ostatními systémy ve světě, protože bude odpovídat mezinárodním standardům.

V roce 2009 (prosinec) byl zahájen Proof of Concept (POC) na systém dlouhodobé ochrany. Jeho větší část se odehrála v roce 2010 (leden-červenec).
Cílem POC bylo porovnat funkcionalitu dvou kandidátů na zakoupení jako SW pro dlouhodobou ochranu. Chtěli jsme na speciálním vzorku našich stávajících dat zkusit následující – vložení dat do LTP systému; procesy dlouhodobé ochrany; celkovou použitelnost produktů; vhodnost našich dat na uložení v LTP systému a v neposlední řadě také připravenost pracovníků NK ČR na takovýto systém atd.
Zpráva z POC pro oba systémy byla dokončena v srpnu 2010. Samotné testování nám poskytlo obrovský vhled do problematiky, který by bylo nemožné získat jinak (např. návštěvou jiných knihoven, dodavatelských firem apod.). V rámci POC vznikly java aplikace na převod našich stávajících metadat do interních formátů obou systémů. To předpokládalo vytvoření mapování našich DTD do interních metadatových formátů obou LTP systémů. Základ těchto aplikací i mapování bude využit v projektu NDK při převodu stávajících dat.
Dále níže v textu je i krátký popis NDK projektu, řešení za rok 2010 a jeho vliv na výzkumný záměr a ostatní instituce v ČR.

- **Specifikace funkčních požadavků národního repozitáře s ohledem na jeho snadnou integraci v rámci portálů a Souborného katalogu ČR (2006-2009).**

V roce 2010 jsme se autorsky podíleli na vzniku publikace:

CUBR, Ladislav. Dlouhodobá ochrana digitálních dokumentů. Praha : Národní knihovna ČR, 2010. 154 s. ISBN 978-80-7050-588-5 (brož).
Tato publikace bude nahlášena jako jeden z výsledků řešení výzkumného záměru v roce 2010.

Jedná se o první průřezovou publikaci věnující se problematice dlouhodobé ochrany digitálních dat a návaznými tématy, která je dostupná v češtině. Kniha samotná je velmi informačně obsáhlá, vedle teoretické části popisuje veškerá dostupná řešení a příklady ze světa. Dá se očekávat, že publikace jako taková významně přispěje ke zvýšení povědomí o této problematice v ČR i na Slovensku v různých typech institucí, které mají za povinnost nebo chtějí dlouhodobě ochránit svá digitální data. Může také velmi dobře posloužit jako učební text na relevantních studijních oborech vysokých škol v ČR a na Slovensku.

V roce 2010 se řešení této části výzkumného záměru soustředilo vedle jiných na dvě oblasti

1) na jednoznačnou identifikaci digitálních objektů pomocí URN:NBN a
2) na zprovoznění SRU/SRW funkcionality nad daty z archivace webu tak, aby bylo možné prohledávat kompletní fulltext pomocí Jednotné informační brány

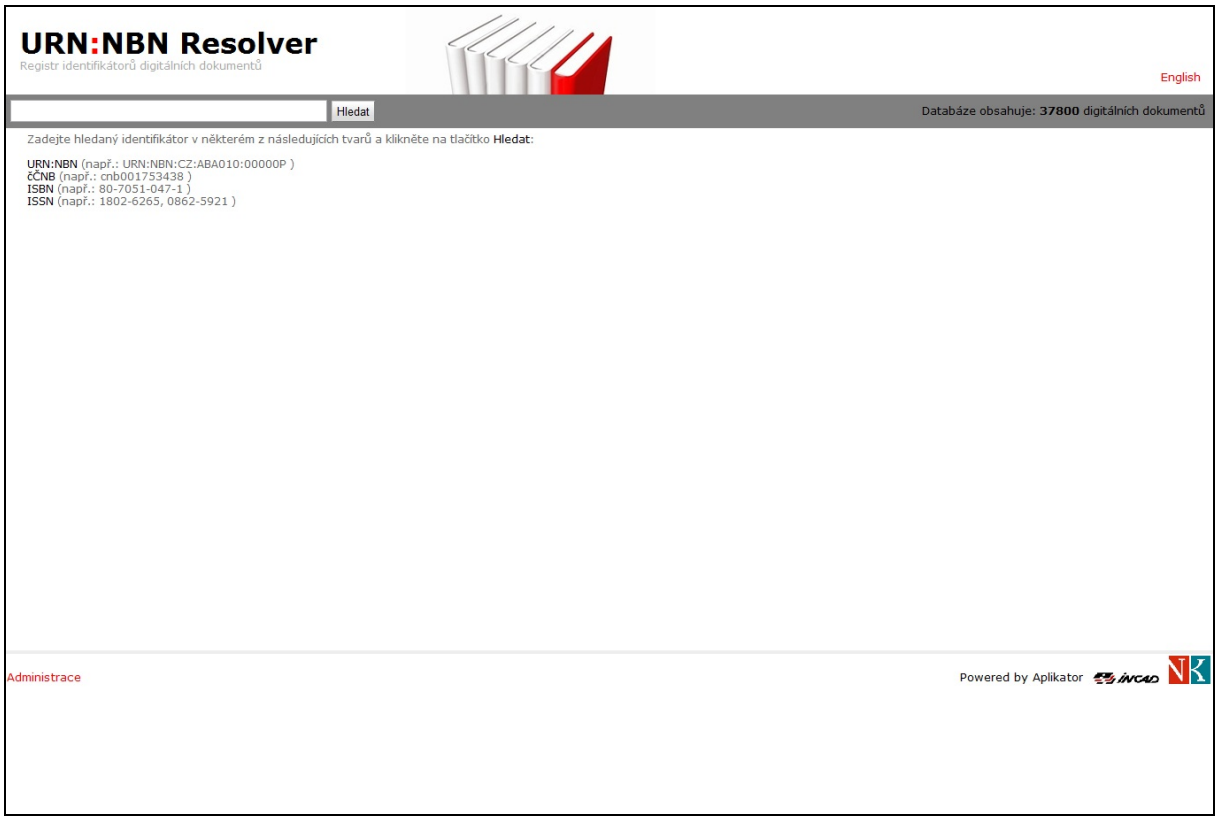
ad 1) Vývoj a pilotní test aplikace pro přidělování a správu URN:NBN

Cílem v roce 2010 bylo připravit prostředí a základní SW aplikaci pro pilotní test systému pro využití URN:NBN v NK ČR. Aplikace umožní přiřazování globálně jedinečného identifikátoru odpovídajícího pravidlům URN:NBN, dále jeho správu (administrátorský modul), vyhledávání dle identifikátoru a bude spolupracovat s již běžícími systémy NK, které jsou potřeba pro pilotní projekt.

V první pilotní fázi (do poloviny roku 2011) je aplikace určena pro využití v NK ČR, pro přidělování a zpřístupňování identifikátorů dokumentů projektu Kramerius, VISK7 a Norských fondů. Identifikátory budou přiřazeny všem digitalizovaným dokumentům z těchto projektů evidovaným v RD.CZ. V dalších fázích (mimo tento výzkumný záměr) bude služba poskytována i ostatním paměťovým institucím za předem určených podmínek.

Dokumenty, které v pilotní fázi dostanou URN:NBN musejí být v současné době uloženy v repozitáři NK nebo tam být uloženy později. Pilotní aplikace bude funkční a otestovaná do konce roku 2010, s případnými úpravami se počítá na rok 2011. Pilotní fáze se vědomě vyhýbá digital born dokumentům.

Velmi intenzivně probíhaly debaty s Odborem zpracování fondů, zejména kvůli číslu národní bibliografie a návaznostech na systém URN:NBN.



Obr. – výchozí rozhraní aplikace

Obecná funkcionalita

Návrh našeho pilotního řešení umožnil vytvoření nástroje, který dokáže přidělovat jedinečné identifikátory postavené na syntaxi URN:NBN pro digitální dokumenty (logické entity). Během implementace návrhu a vývoje nástroje resolver se potvrdila námi deklarovaná skutečnost[6], že je velmi komplikované a nelogické vytvářet aplikaci resolveru dříve, než bezpečně víme, jak bude vypadat životní cyklus digitálních dokumentů, kterým URN:NBN chceme přidělovat. Vzhledem k výše zmíněnému je logika pilotního řešení zatím omezena na dokumenty, které jsou registrovány v jednom balíčku jako zakázka v systému RD.CZ, a to pouze ty z nich, které jsou již reálně zdigitalizovány a budou trvale archivovány v digitálním repozitáři Národní knihovny ČR.

Nástroj resolver umožňuje rozeznávat na základě podřízených jmenných prostorů různé digitalizující subjekty (instituce) a přidělovat jim globálně jedinečné identifikátory, a zajistit, že žádný identifikátor nebude přidělen znovu. Tj. pokud dvě instituce zdigitalizují stejnou knihu, oba takto vzniklé digitální dokumenty budou mít různé URN:NBN, i kdyby byly naprosto shodné.

Nástroj umožňuje rovněž vyhledávat identifikované digitální dokumenty, a to nejen podle identifikátoru URN:NBN podle kterého vyhledává primárně, ale také podle dalších užívaných identifikátorů, a sice ISSN, ISBN a čČNB (číslo České národní bibliografie). Po vyhledání dostane uživatel relevantní metadata k vyhledávanému dokumentu, vidí zda a kým byl zdigitalizován a kde je zpřístupněn, včetně URL linku do konkrétní digitální knihovny.

[6] viz CUBR, Ladislav; HUTAŘ, Jan; MELICHAR, Marek. Kontrolní seznam pro strategii zajištění perzistence identifikátorů. *Knihovna*. 2009, roč. 20, č. 2. s. 54-62. Přístupné z WWW: <http://knihovna.nkp.cz/pdf/0902/090254.pdf>.

Obr. – výsledek hledání dle URN:NBN - výřez

Resolver je schopen přijmout a nadále udržovat URN:NBN přidělená jinými systémy, např. SW pro workflow digitalizace na konkrétním pracovišti (např. DocWorks nebo Sirius aj.). Druhou možností je, že workflow digitalizace bude volat resolver, ten přidělí URN:NBN.

Syntax identifikátoru

Syntax byla navržena takto - URN:NBN:CZ:XXX:12345A, kde:

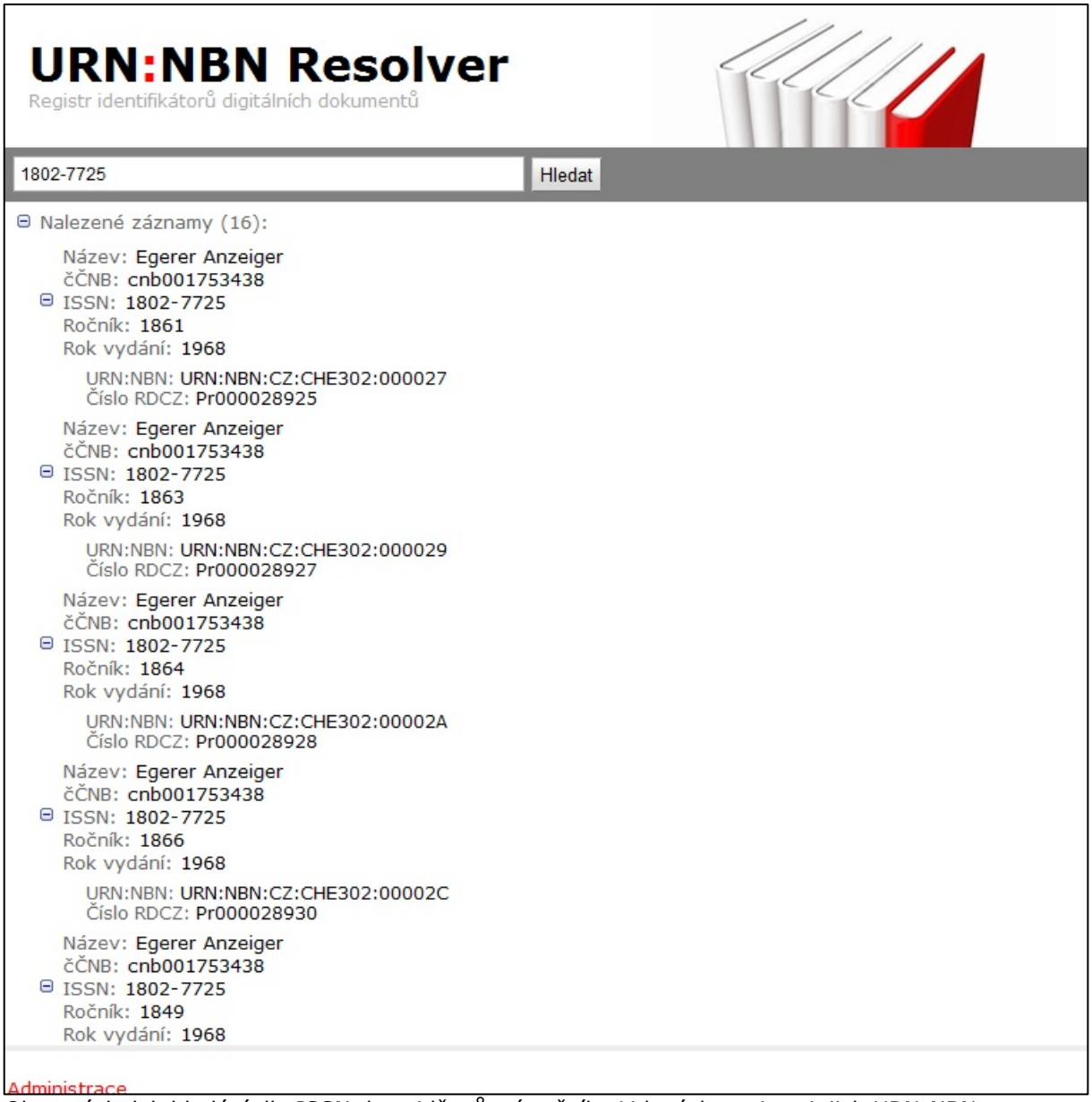
- XXX - kód vlastníka /původce / vydavatele (v pilotní fázi odpovídá údajům v RD.cz)
- 12345A - šest alfanumerických znaků - pro čísla a písmena (10 číslic + 26 znaků latinky) máme 36 na 6 kombinací, tedy cca 2,1 miliardy kombinací
- ve všech částech identifikátoru půjde o náhodná čísla, tj. nebudou označovat konkrétní číslo ročníku ani čísla (u periodik) apod.

Současné řešení

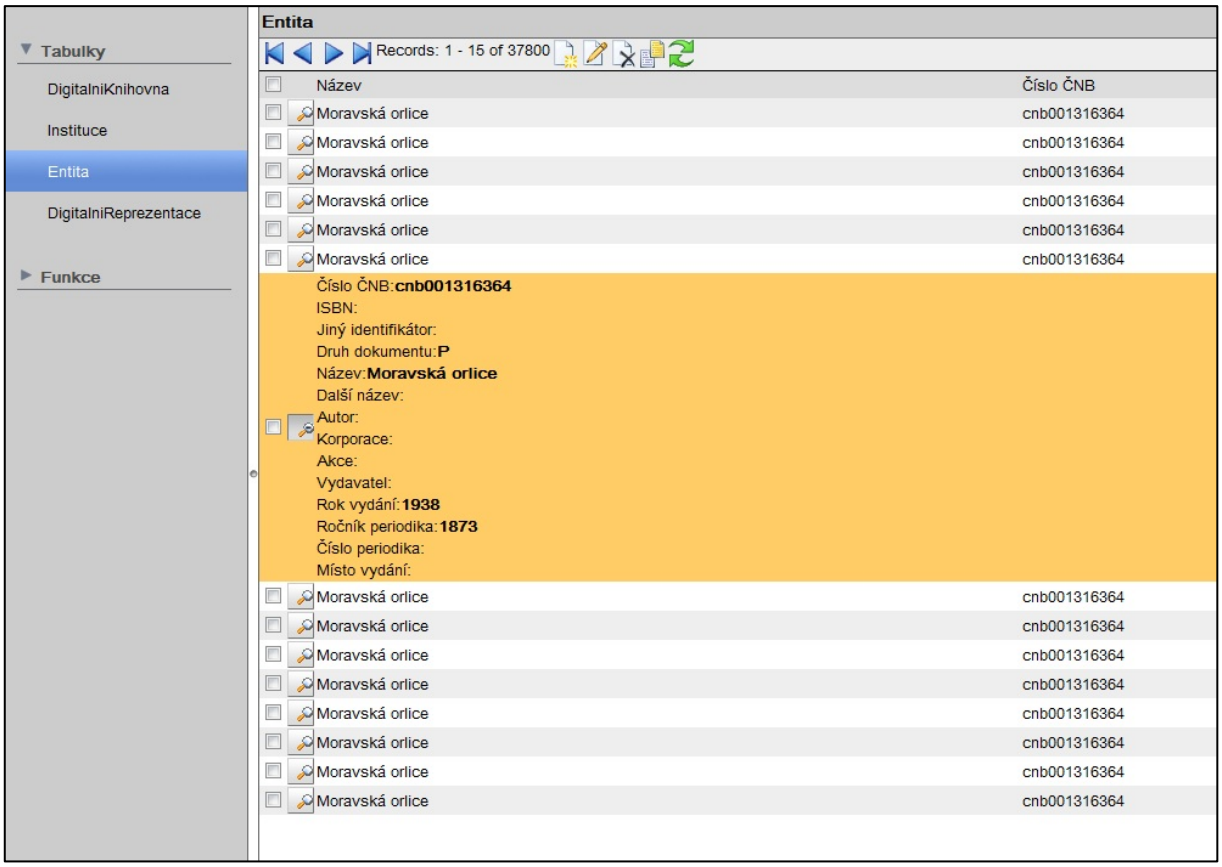
V pilotní fázi je resolver napojen na systém Registr Digitalizace (dále RD.CZ) a využívá jeho databáze. Do RD.CZ přicházejí data ve formě zakázek, které zpravidla reflektují způsob uložení dokumentů v depozitářích knihoven. Digitalizace probíhá obyčejně podle svazků monografií ovšem i periodik, které jsou svázány do svazků (dodatečně po akvizici) podle toho, jak se to vazačům hodilo, tj. bez obecně platných pravidel a jsou takto i digitalizovány. Jeden svazek tak může obsahovat různé množství čísel i ročníků, nebo naopak jen půl ročníku apod.). Ovšem metadata např. k číslu periodika v katalogu a ani v RD.cz neexistují. Do RD.CZ přichází takto zdigitalizované svazky jako zakázky a tedy jako celky (tj. digitální objekty), které v pilotní fázi dostanou přiděleno URN:NBN. Je jasné, že tato situace není konsistentní, např. pro uživatele, který hledá konkrétní číslo periodika a to konkrétní číslo by mělo mít své číslo URN:NBN, stejně jako např. celý ročník nebo i titul periodika. Takto má URN:NBN digitální objekt (celek) vzniklý digitalizací svazku, jehož obsah jako intelektuální entity je proměnlivý (velmi často celý ročník, dále půlročníky, dva nebo více ročníků apod.). To působí problémy hlavně u periodik, u monografií méně (často svazek=číslo zakázky).

Vyjmenované logické nekonzistence pilotního řešení nejsou problémem samotného resolveru, ten je již i v pilotní verzi schopen přidělit URN:NBN jakémukoliv digitálnímu objektu, který jako celek bude zaregistrován v RD.CZ nebo požádá o přidělení URN:NBN externě. Tj. pokud bude workflow digitalizace v NK nastaveno tak, že budou

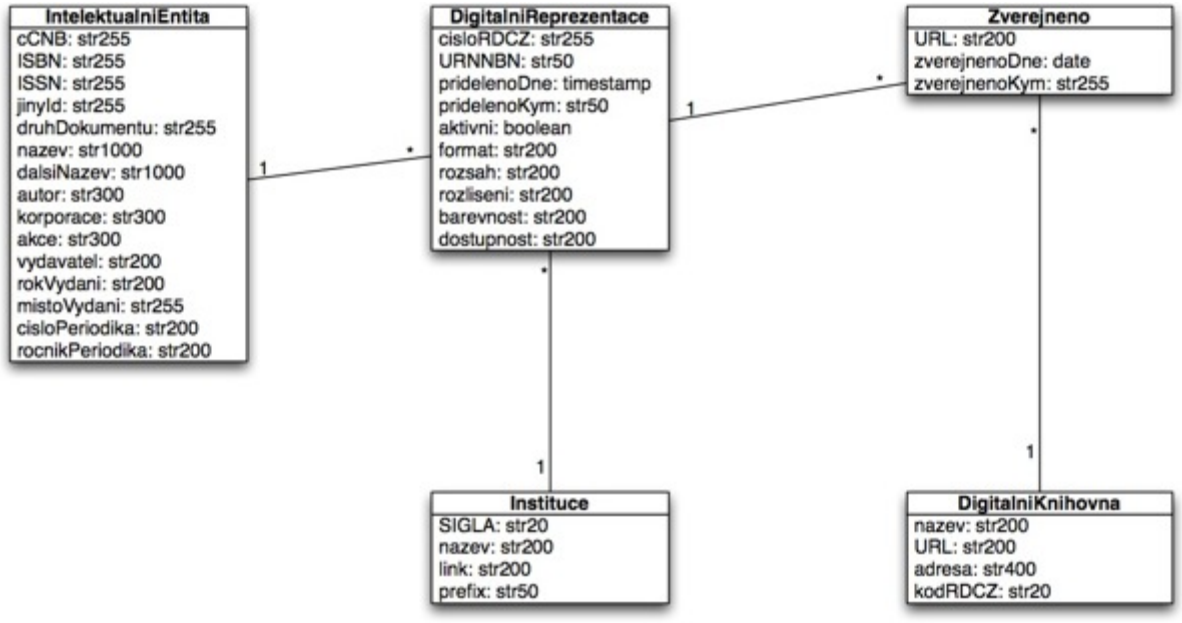
digitalizována periodika tzv. „na čísla", každé číslo bude mít svůj metadatový záznam, pak tato číslo mohou dostat URN:NBN. Stejně se to děje již nyní, pokud digitální objekt v RD.CZ je číslo, pak URN:NBN je přiděleno pro číslo. Tj. resolver je připraven na budoucí změny ve workflow digitalizace (opuštění čísel zakázek a zrušení fixace digitálního objektu na fyzický svazek dokumentů). To čemu se bude URN:NBN přidělovat, musí být jasné již během procesu digitalizace, jejíž workflow tak musí být nastaveno.



Obr.- výsledek hledání dle ISSN, lze vidět různé ročníky Lidových novin a jejich URN:NBN identifikátory (výřez)



Obr.- rozhranní administrace (výřez)



Obr. – definice struktury dat

Provázání workflow digitalizace a systému identifikace

Z předchozích odstavců vyplývá, že **je klíčové a nutné provázat systém trvalé identifikace s celým digitalizačním workflow a řízením digitálního životního cyklu dokumentu již od jeho rané fáze**[7]**.**
V přípravě projektu IOP/NDK se počítá s tím, že základní jednotkou granularity pro periodikum bude jedno číslo[8]. Tento postup bude aplikován v dalším vývoji systému trvalé identifikace URN:NBN. Je tedy potřeba stanovit standardy pro digitalizaci, které budou dodržovány institucemi, které dodávají data do RD.CZ, potažmo do projektů VISK7 apod.

Problematika aktuálních lokací digitálních dokumentů

Dalším tématem na budoucí řešení v dalších fázích výzkumu a reálného nasazení i pro ostatní instituce je „publikování" aktuálních lokací digitálního dokumentu. Tj. jde o to, aby instituce u dokumentů, kterým je URN:NBN přiděleno, buď zaslaly informace o aktuálním uložení (URL) nebo vystavily profil OAI-PMH pro resolver ze své digitální knihovny. Např. v současnosti jsou do RD.CZ dodána metadata o dokumentech (svazcích), které se budou digitalizovat nebo se již digitalizují. Není k nim ovšem logicky uvedeno URL, kde je zdigitalizovaný dokument uveden, protože zatím neexistuje. Podobně to bude ve workflow digitalizace projektu NDK. Jen velmi málo institucí do RD.CZ zpětně toto URL doplní. Tím jsme postaveni před otázku, odkud tato aktuální URL pro resolver brát, z jakého zdroje. Existují dvě možnosti – resolver bude muset využít protokol OAI-PMH ke sklízení metadat z jednotlivých digitálních knihoven spolupracujících institucí, tj. těch, kterým bylo povoleno si přidělovat nebo jsou jejich dokumentům přidělována URN:NBN. Zde opět narážíme na problém workflow a řízení dalších fází životního cyklu digitálního dokumentu. Druhou možnosti je ustanovení povinnosti, kdy by spolupracující knihovny musely dodávat resolveru informace o aktuálních lokacích (URL) dokumentu. Tato povinnost by se doplňovala s povinností udržovat aktuálnost URL dokumentů v digitální knihovně konkrétní instituce a s možností administrace své množiny dokumentů.

Technické řešení

- pilotní implementace systém Resolver URN:NBN je provozována na existující infrastruktuře Registru digitalizace v Národní knihovně
- Systém navazuje na ostatní řešení Národní knihovny - prostřednictvím Registru digitalizace přebírá bibliografické záznamy z Katalogu a informace o digitalizovaných předlohách z Krameria4
- URN:NBN resolver je webová databázová aplikace na platformě J2EE (Java 2 Enterprise Edition)
- může běžet v libovolném J2EE Servlet kontejneru, obvykle pod Apache Tomcat
- jako databáze může být použita většina obvyklých SQL databází
- systém využívá relační databáze Oracle 11g a Aplikačního serveru registru a napojení do vyhledávací služby FAST Registru
- vrstva GUI používá technologii Google Web Toolkit
- databázová vrstva je založena na frameworku Apache EmpireDb
- vyhledávací rozhraní je napsáno specificky pro potřeby Resolveru
- administrátorské rozhraní je postaveno na novém frameworku pro Relief 4
- pilotní implementace čte data z databáze RD.CZ, v dalších fázích bude doplněno API pro spolupráci s jinými systémy/zdroji dat

[7] viz (CUBR, Ladislav. Budování důvěryhodného systému trvalé identifikace digitálních dokumentů. Knihovna. 2010, roč. 21, č. 1, s. 23-31. ISSN 1801-3252.)
[8] Není možné, aby URN:NBN jednou identifikovalo celý ročník, jindy jen půlročník nebo podobné neúplné intelektuální entity.

Technický vývoj Resolveru zahrnoval

- definici databázové struktury aplikace do RDBMS Oracle
- vytvoření uživatelských formulářů pro administraci záznamů v Resolveru v systému R4
- vytvoření vyhledávací služby pro zadání identifikátoru
- sadu komponent pro komunikaci s ostatními systémy

Shrnutí

- jde o pilotní provoz, tj. plánovaná technická funkcionalita SW pro tuto fázi vývoje je vyřešena, základ technického řešení pro zapojení dalších institucí a projektu NDK je hotov
- netechnické věci související se strategií přidělování a odvíjející se od aktuálního workflow digitalizace je nutné ještě dořešit– očekáváme v další fázi v rámci projektu NDK
- resolver je maximálně flexibilní s ohledem na přidělování URN:NBN digitálním objektům – přidělí je čemukoliv, vliv na to co se bude přidělovat má workflow digitalizace a strategická rozhodnutí udělaná v ní (potažmo v projektu NDK a následně ve VISK7)
- v rámci NDK a řešení resolveru se ukázalo, že zcela jistě nastane situace, kdy URN:NBN budou přidělována i mimo resolver
- resolver bude umí zacházet i s URN:NBN, která budou podle určitých pravidel přidělována v jiných systémech (digitalizační workflow NDK nebo Sirius apod.)
- resolver by měl mít možnost sklízet OAI-PMH profily z digitálních knihoven zúčastněných institucí
- nelze vytvářet strategii resolveru, pokud neznáme finální podobu životního cyklu dokumentu (tj. tok dat v digitalizaci a dále)
- nelze vytvořit konzistentní systém identifikátorů, pokud ostatní zdroje dat, které by měl využívat, nejsou konzistentní
- je nutné stanovit závazné standardy pro data dodávaná do RD.CZ na základě workflow nebo URN:NBN oddělit od RD.CZ a udělat workflow tak, aby instituce, která URN:NBN potřebuje, s tím neměla příliš práce (tj. přidělit URN:NBN v procesu digitalizace a pak ho posléze sklízet z aplikace zpřístupnění)

Testovací verze je přístupná zatím jen na serveru vývojářské firmy http://sluzby.incad.cz/urnnbn/, do konce roku 2010 bude zprovozněna na následujícím URL http://resolver.nkp.cz/.

ad 2) Vývoj aplikace pro fulltextové vyhledávání ve WebArchivu včetně zpřístupnění fulltextových úryvků děl chráněných autorskými právy

Fulltextová indexace

Pro fulltextovou indexaci WebArchivu používáme nástroj nutchwax[9] vyvinutý organizací Internet Archive[10], který rozšiřuje funkcionalitu open source internetového vyhledávače nutch[11] o indexaci ARC souborů a ukládání specifických metadat pro WebArchiv.

Fulltextová indexace se skládá z následujících fází:

1. Import obsahu dokumentů z ARC souborů - z každého textového dokumentu jsou extrahována metadata, text a v případě HTML stránek ještě navíc odkazy. Výsledky jsou ukládány do tzv. segmentů.

[9] http://archive-access.sourceforge.net/projects/nutch/
[10] http://www.archive.org/
[11] http://nutch.apache.org/

2. Aktualizace databáze crawleru - tato část je sice z našeho pohledu zbytečná, neboť používáme pro sklízení Heritirix a ne crawler nutche, ale z jistých technických důvodů ji nelze vynechat.
3. Invertování odkazů - každému dokumentu je přiřazen seznam stránek, které na něj odkazují.
4. Vygenerování pageranku pro hodnocení relavance stránek - je vygenerován textový soubor obsahující na každém řádku URL dokumentu, podle kterého je lexikograficky setříděn a počet externích odkazů (tzn. odkazů z jiných domén), které na daný dokument odkazují.
5. Indexace - ze segmentů, které byly vytvořeny v první fázi, se generuje invertovaný soubor a ke každému dokumentu se navíc ukládá hodnota pageranku. Seznam metadat ukládaných do indexu je v následující tabulce:

pole	**popis**	**příklad**
segment	20100326225912	segment, má význam pouze pro nuchwax
title	Národní knihovna	titulek stránky (z obsahu elementu title)
content		textový obsah dokumentu pro generování úryvků
url	http://narodni-knihovna.cz/	URL dokumentu
digest	sha1:NO2WDXITSO6MDWUBNK3BXZAPZCSLQGE6	otisk (hash) z obsahu dokumentu
collection		jméno kolekce (nepovinné, nepoužíváme)
date	20081018190624	čas sklizně dokumentu
type	text/html	MIME typ dokumentu
length	28138	velikost dokumentu v bytech
boost	5.0	relevance dokumentu pro řazení výsledků, hodnota je rovna $\log_{10}$N, kde N je počet externích odkazů na tento dokument

Fulltextová indexace probíhá po částech, výsledné indexy je třeba sloučit do jednoho a následně odstranit z indexu duplicitní dokumenty, které rozlišujeme podle MD5 haše.

Úskalí při fulltextové indexaci:

1. Špatné či chybějící deklarované kodování dokumentu - nástroj nutchwax byl modifikován tak, že identifikuje kódování dokumentu stejným způsobem jako wayback, u kterého je již odladěné. Pokud deklarace kódování chybí, použije se heurisitika. Úprava spočívala v ukládání HTTP hlavičky s kodováním při importu segmentů, kterou sám o sobě nutchwax zahazuje, a využitím této informace při indexaci. Toto řešení lze využít i v distribuované verzi.
2. Problém s občasnými pády byl vyřešen přechodem na Javu od IBM.
3. Extrakce textu z některých nekorektně vytvořených PDF dokumentů je stále problémem, protože z některých dokumentů po extrakci vypadne "čínský čaj" a získat původní text není schopen ani Acrobat Reader.
4. Spam

Při detekci znakové sady se postupuje následovně (při prvním pozitivním výsledku se nepokračuje):

1. Deklarace znakové sady v HTTP hlavičce odpovědi serveru

2. Deklarace znakové sady v prologu HTML dokumentu
3. Jednoduchá heuristika, pro každou českou znakovou sadu (UTF-8, ISO-8859-2, CP-1250) se spočítá celkový počet českých znaků s diakritikou ze začátku dokumentu a jako výsledek se bere znaková sada s nejvyšším dosaženým počtem.

Možnosti omezení vyhledávání dokumentů na volně dostupné:

1. Mít dva indexy, jeden úplný a druhý jen s volně dostupnými dokumenty vygenerovány jednou za čas z úplného indexu stejným způsobem, jakým odstraňujeme duplikáty.
2. Odlišit ve výsledcích hledání volné a nedostupné dokumenty, např. ikonou či poznámkou.
3. Modifikace indexu je problematická, změnit metadata dokumentu lze jen tak, že ho odstraníme a přídáme znovu do indexu.

Odstraňování dokumentů z indexu je poněkud komplikované, protože index v Lucene nelze upravit na místě ("in place") a je tudíž náročnější na místo na disku (potřebujeme až M dodatečného volného místa, kde M je velikost původního indexu):

1. Nejprve je vygenerována bitmapa, kde každý bit reprezentuje jeden dokument a jeho hodnota indikuje, zda má či nemá být ponechán.
2. Při slučování indexů se přečte bitmapa každého indexu a do výsledného indexu jsou přidány jen dokumenty, které nebyly označeny jako smazané. Sloučit lze jeden či více dokumentů.

Statistika fulltextu

Rychlost indexace se pohybovala v rozmezí 1000 až 1500 ARC souborů za den, statistiky za jednotlivé roky lze najít v následující tabulce:

java	IBM JRE 1.6.0
procesor	4 x Intel(R) Xeon(R) CPU E5420@2.50GHz
operační paměť	8 GB
diskové pole	10 TB

rok	**zaindexované ARC soubory**	**počet dokumentů v tisících**
do roku 2005	915	2499
2006	1446	2143
2007	2395	3880
2008	8856	18450
2009	25363	15612
2010	7342	8365
celkem	46317	50940

Vyhledávání ve fulltextu

Vyhledávání ve fulltextu je dostupné na http://war.webarchiv.cz/nutch/search, zvýrazňování výsledků ve waybacku zajišťuje javascript, který hledaná slova získá z redirectu a funkčnost není garantována u všech stránek. Aktuálně (k 31.10.) jsou zaindexované nasmlouvané zdroje od roku 2003 do března 2010. Postupně se bude index rozšiřovat o nové sklizně.

Formát dotazu:

- Výsledek obsahuje pouze stránky, které obsahují všechna slova v dotazu.
- Lze vyhledávat i fráze, hledanou frázi je třeba uzavřít do dvojitých uvozovek, např. "Národní knihovna".
- U dotazů nezáleží na velikosti písmen.
- Určitý výraz můžete z vyhledávání vyřadit vložením znaménka mínus před něj, např. vyhledávání football -NFL najde všechny stránky týkající se fotbalu, ale neobsahující slovo "NFL".
- Dokumenty lze vyhledavát i podle času, např. dotaz "Národní knihovna" date:2005. Čas je porovnávan prefixově, tzn. pro dotaz "date:2005" najde všechny dokumenty z roku 2005, pro dotaz "date:200512" všechny dokumenty z prosince 2005. Lze vyhledávat i podle intervalu, např. "Národní knihovna" date:200511-200601.
- Pokud nás zajímají pouze výsledky z domény www.nkp.cz, použijeme pole site, např. site:www.nkp.cz "Národní knihovna".
- Vyhledávat lze i podle mime typu, např. dotaz "Národní knihovna" type:application/pdf najde všechny PDF dokumenty, které obsahují frázi "Národní knihovna".

Poznámky k hledání:

- Vyhledávat lze i dlouhé fráze skládající se z několika vět, takže není problém vzít pár vět z odstavce nějakého článku a hledat, kde všude byl článek převzat. Šlo by to využít i pro detekci podobných dokumentů.
- Nutch neumí detekovat podobné dokumenty při zobrazování výsledků, proto se často stává, že podobné dokumenty mají shodnou relevanci a tudíž jsou ve výsledcích u sebe.
- Detekce znakové sady Nutchem u některých dokumentů není optimální, na řešení se pracuje.
- U některých PDF vypadne při převodu na text "čínský" čaj.
- Původně chyběla možnost stránkování výsledků jako u Googlu, řešením je použít XSL šablonu distribuovanou s nutchwaxem, která transformuje výsledky z Open Search do HTML a podporuje stránkování.

Oficiální dokumentace k vyhledávání je na http://wiki.apache.org/nutch/FAQ#Searching.

Vyhledávání ve fulltextu je dostupné na:

- Webové rozhraní na http://war.webarchiv.cz/nutch/search, které využívá open search a výsledky z XML transformuje do HTML za pomocí XSL šablony.
- Open search rozhraní na http://war.webarchiv.cz:8080/WebarchivSearcher/opensearch.html
- Jednotná informační brána, název zdroje je WebArchiv – vyhledávání v plných textech.
- Metalib Masarykova univerzity pod názvem Webarchiv.

Obr. Úplné zobrazení nalezeného výsledku fulltextového vyhledávání v JIB

SRU/SRW protokol, zapojení do JIB

Jednotná informační brána podporuje protokol SRU/SRW díky skriptům napsaných v Perlu vyvinutých finskou národní knihovnou, které transformují dotazy a odpovědi do (z) protokolu SRW/U a zpřístupňují je tak jádru systému Metalib, který je srdcem JIB. Veškeré parametry dotazu jsou v případě SRU součástí URL, v případě SRW se pro volání a vracení výsledku používají webové služby. Dotazy jsou vyjádřené v jazyce CQL (Contextual/Common Query Language), což je poměrně silný a lidsky čitelný jazyk pro dotazování nad vyhledávacími stroji.

CQL podporuje:

- booleovské operátory (AND, NOT, OR). Příkladem dotazu je "'auto' OR 'motocykl'".
- relační operátory (=, >, <, ...). Příklad dotazu je "year > 2000".
- dotazy na příslušný klíč, např. "dc.autor='Božena Němcová'".
- podrobnější popis a příklady dotazů lze najít v A Gentle Introduction to CQL[12].

Opensearch[13] je jednoduchý protokol pro vyhledávání na webu, základem je jednoduchý XML dokument, který popisuje syntaxi URL pro volání a seznam podporovaných formátů (JSON, RSS, HTML). Příklad takového XML dokumentu je:

```
<?xml version="1.0" encoding="UTF-8"?>
<OpenSearchDescription xmlns="http://a9.com/-/spec/opensearch/1.1/">
    <ShortName>Web Search</ShortName>
    <Description>Use Example.com to search the Web.</Description>
    <Tags>example web</Tags>
    <Contact>admin@example.com</Contact>
    <Url type="application/rss+xml"

 template="http://example.com/q={searchTerms}&amp;format=rss"/>
</OpenSearchDescription>
```

Každý opensearch deskriptor obsahuje jeden či více elementů Url, které obsahují v atributu type vrácený formát (RSS, JSON, HTML) a v atributu template syntaxi URL pro volání, která obsahuje následující parametry, které jsou při dotazu nahrazeny za příslušnou hodnotu parametru:

[12] http://zing.z3950.org/cql/intro.html
[13] http://www.opensearch.org/Home

parametr	povinný	význam
{searchTerms}	ano	dotaz, hledaný výraz
{startPage?}	ne	číslo požadované stránky (umožňuje stránkování výsledků)
{count?}	ne	počet vrácených výsledků
{startIndex?}	ne	zobraz výsledky od (umožňuje stránkování výsledků)

Příklad volání, které vrátí prvních pět výsledků pro dotaz MZK ve formátu RSS je: http://war.webarchiv.cz:8080/nutch/opensearchquery=MZK&start=0&hitsPerPage=5&format=rss

Markantní rozdíl mezi opensearch a SRU/SRW je, že opensearch nepředepisuje na dotaz žádná omezení a může jím být prakticky cokoliv. V SRU/SRW formu dotazu předepisuje jazyk CQL. Příklad dotazu v CQL, a jeho ekvivalentu pro nutch ilustrujeme v následující tabulce, nutch podporuje "googlovské" výrazy:

dotaz v CQL	ekvivalent dotazu pro nutch
dc.title="hello" AND dc.date=2008	title:"ahoj" date:2008
"příklad dlouhé fráze"	"příklad dlouhé fráze"
"první fráze" AND "druhá fráze"	"první fráze" "druhá fráze"
"obsahuje" NOT "neobsahuje"	"obsahuje" -"neobsahuje"
cql.serverChoice="studená válka"	"studená válka"
dc.author="novak"	- (nelze přeložit)
"auto" OR "motocykl"	"auto" "motocykl" (není ekvivalentní)

Předposlední dotaz nelze přeložit, neboť pole autor z Dublin Core nelze namapovat na odpovídající pole v nutch a v takových případech vrátíme chybu. Poslední dotaz není ekvivalentní dotazu v CQL, protože nutch nepodporuje operátor OR a výsledný dotaz najde všechny dokumenty, ve kterých se vyskytují zároveň slova auto a motocykl, zatímco dotaz v CQL najde dokumenty, které obsahují slova auto nebo motocykl. Při implementaci rozhraní mezi opensearch a SRU/SRW musíme počítat s tím, že ne veškeré dotazy v CQL půjdou přeložit do opensearch. Mapování polí z Dulin core, na které se dotazuje metalib, na pole nutche je přitom následující:

formulář v metalibu	pole v CQL	nutch
všechna pole	cql.serverChoice	implicitní pole
název	dc.title	title
rok	dc.date	date
předmět	dc.subject	chyba, ekvivalent neexistuje
autor	dc.author	chyba, ekvivalent neexistuje
ISSN	dc.identifier	chyba, ekvivalent neexistuje
ISBN	dc.identifier	chyba, ekvivalent neexistuje

Příklad výsledku dotazu SRU/SRW ve formátu Dublin Core:

```
<?xml version="1.0" encoding="UTF-8"?>
<srw xmlns="info:srw/schema/1/dc-v1.1">
    <title>MZK Brno</title>
    <publisher>www.mzk.cz</publisher>
    <date>2008-10-18</date>
```

```
    <description>uryvky</description>
    <identifier>odkaz na živou verzi/</identifier>
    <identifier>odkaz do waybacku</identifier>
    <identifier>všechny časové verze</identifier>
    <format>text/html</format>
</srw>
```

Vyvinuli jsme tedy překladač z jazyka CQL na ekvivalentní dotaz v OpenSearch pro nutchwax, který lze ovšem využít i pro jiné zdroje s podobnou syntaxí (otestovali jsme ho i na google books) a který bere v potaz výše uvedená úskalí překladu. Dále jsme vyvinuli konfigurovatelnou bránu mezi SRU/SRW a OpenSearch, do které jsme začlenili tento překladač a XSL šablonu, která konvertuje RSS výsledky nutchwaxe do formátu Dublin Core. Zdrojové kódy rozhraní mezi opensearch a SRU/SRW jsou dostupné na adrese http://code.google.com/p/opensearch-to-srw-gate/.

Konfigurace brány

Brána se konfiguruje v textovém souboru ve formátu java properties, který definuje URL s opensearch deskriptorem, XSL šablonu pro transformaci výsledku, definici jmenných prostorů, XPath výraz pro separaci jednotlivých výsledků a XPath výraz, který vrátí počet celkových výsledků. V XPath výrazech se můžeme odvolávat na definované jmenné prostory. Pro tranformaci dotazu musíme definovat podporované logické operátory (AND, OR, NOT) a mapování polí z CQL na odpovídající opensearch ekvivalent. Většina vyhledáváčů totiž podporuje výrazy ve tvaru *klíč:hodnota*.

```
opensearch.url=http://war.webarchiv.cz/warcs/googlebooks.xml
opensearch.xsl_template=/home/app/opensearch/googlebooks.xsl
opensearch.namespace.opensearch=http://a9.com/-/spec/opensearchrss/1.0/
opensearch.namespace.atom=http://www.w3.org/2005/Atom
opensearch.records_xpath=//atom:entry
opensearch.total_records_xpath=//opensearch:totalResults/text()
operator.and=AND
operator.or=OR
operator.NOT=-
key.srw.serverChoice=
key.cql.serverChoice=
key.cql.any=
key.title=intitle
#dublin core elements
key.dc.creator=inauthor
key.dc.title=intitle
key.dc.publisher=inpublisher
key.dc.identifier=isbn
```

Příklad konfigurace pro Google books

Při psaní XSL šablony musíme dávat pozor na definici jmenných prostorů, google books vrací Dublin Core ve jmenném prostoru s URI *http://purl.org/dc/terms*, Metalib je ovšem vyžaduje ve jmenném prostoru *info:srw/schema/1/dc-v1.1*, google books vrací výsledky ve formátu DC, stačí u všech elementů DB změnit jmenný prostor.

```
<?xml version="1.0" encoding="utf-8"?>
<xsl:stylesheet version="1.0"
    xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
    xmlns:dc_google="http://purl.org/dc/terms"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
```

```
    xmlns:srw_dc="info:srw/schema/1/dc-v1.1"
    xmlns:atom="http://www.w3.org/2005/Atom"
    exclude-result-prefixes="atom dc_google">
    <xsl:output method="xml" encoding="utf-8" indent="yes" />

    <xsl:template match="/">
        <srw_dc:dc>
            <xsl:apply-templates/>
        </srw_dc:dc>
    </xsl:template>

    <xsl:template match="//dc_google:*">
        <xsl:variable name="element" select="name()"/>
        <xsl:element name="{$element}">
            <xsl:value-of select="text()"/>
        </xsl:element>
    </xsl:template>

    <xsl:template match="text()"/>
</xsl:stylesheet>
```

XSL šablona pro google books

Loňským výsledkem tohoto záměru byl poloprovoz SRU/SRW rozhraní pro vyhledávání nad fulltextovým indexem a jeho integraci s metavyhledávacími portály. V letošním roce jsme se v této oblasti posunuli na úroveň provozu, který bude nahlášen jako jeden z výsledků řešení za rok 2010.

Krátký popis projektu NDK a jeho vlivu na výzkumný záměr

V roce 2010 byl schválen projekt Národní digitální knihovna. V jeho rámci vznikne v roce 2011 pilotní provoz a po testování v roce 2012 ostrý provoz kompletní digitální knihovny s moduly digitalizace, dlouhodobá ochrana a zpřístupnění.

Vzhledem k tomu, že jak masová digitalizace, tak dlouhodobá ochrana a zpřístupnění digitálních dokumentů jsou aktuální, ale velmi nová témata dneška, kde získání aktuálních informací a praktických zkušeností rozhodně není snadnou ani levnou záležitostí, bude se jednat o významný příklad pro další české knihovny a paměťové instituce. V následujícím textu jsou popsány ty části projektu NDK, které se týkají problematiky řešené v tomto výzkumném záměru (resolver URN:NBN a SRU/SRW).

Digitalizace

Pro zvládnutí procesu přípravy a samotné digitalizace, archivace digitálních dokumentů a pro automatické propojování všech částí digitálního systému s katalogy NK ČR a MZK jsou podstatné dva typy identifikátorů: číslo národní bibliografie (čČNB zavedeno v roce 2009) pro jednotlivé publikace (manifestace intelektuálních entit) a URN:NBN, které bude sloužit jako unikátní persistentní identifikátor digitálních dokumentů (nebo jejich logických částí).

Počítá se s tím, že URN:NBN bude přidělováno v resolveru (viz vyvíjený poloprovoz v NK) a to pro vrchní úrovně intelektuálních entit (titul periodika, svazek monografie). Resolver bude tyto identifikátory spravovat, vyhledávat a vracet relevantní lokace, kde je digitální dokument dosažitelný atd. Naproti tomu ve workflow digitalizace budou také přímo přidělovány URN:NBN identifikátory, ale logickým entitám typu ročník, číslo, vnitřní část. Tyto v digitalizaci přidělené identifikátory budou shromažďovány v resolveru, který se o ně bude starat stejným způsobem jako o identifikátory, které přidělil sám, včetně návazných funkcí.

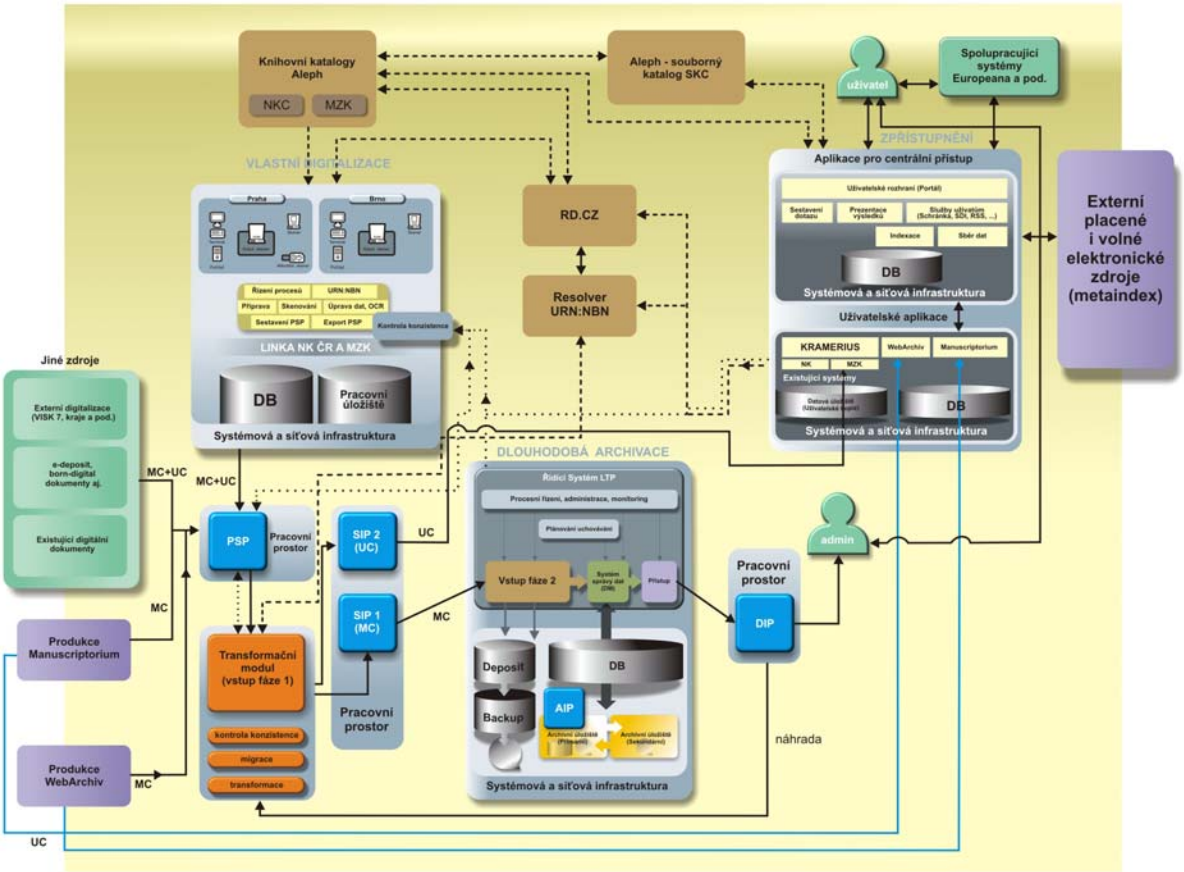
Resolver také bude přidělovat, a také je spravovat, identifikátory logickým entitám/dokumentům, které neprojdou digitalizačním workflow NDK a budou určené k archivaci v LTP systému.

Nové formáty dat a metadat
V rámci příprav na digitalizaci v NDK bylo rozhodnuto o přistoupení k běžným standardům metadat tak, jak jsou využívána v podobných projektech ve světě. To znamená:

- MODS, MARCXML a DC pro bibliografická metadata
- METS – jako kontejner pro celý balík dat a metadat + vyjádření logické i fyzické struktury dokumentů
- PREMIS – technická a administrativní metadata
- MIX – technická metadata

Proběhla také analýza formátu JPEG2000 a jeho nasazení pro archivní a uživatelskou kopii.

LTP systém



Úkolem LTP (Long Term Preservation) systému je zajistit trvalé (po dobu, dokud budou mít uchovávané digitální objekty význam pro uživatele) uchování digitálních nebo digitalizovaných dokumentů v takovém stavu, ve kterém je budou moci uživatelé použít. Uchovávané zdroje musí být vyhledatelné, uživatel je musí mít možnost v pro něj běžném technickém prostředí zobrazit tak, jak zamýšlel jejich tvůrce, a uživatel musí být schopen jim porozumět, pochopit jejich obsah a smysl. Dosažení těchto cílů předpokládá uchování nejen bit streamu reprezentujícího daný digitální objekt, ale také uchování dalších informací (kontextu), které umožní objekt vyhledat, adekvátně technicky zobrazit a uživateli objektu i porozumět.

Příprava dokumentů pro vstup do LTP systému bude začínat již na pracovišti digitalizace nebo při harvestingu ve WebArchivu. Odtud budou data přicházet již se základními metadaty (popisnými, strukturálními, administrativními, po kontrole kvality obrazu a zpracování OCR). Před vstupem do LTP systému musí data i metadata projít Transformačním modulem.

Transformační modul přeměňující veškerá příchozí data a metadata do vnitřních metadatových formátů LTP systému (tvorba a předávání SIP balíčku do LTP) bude předmětem vývoje. S ohledem na vysokou úroveň standardizace v Česku bude tento modul využitelný i v dalších českých knihovnách a paměťových institucích. V tomto modulu se promítnou znalosti nabité během provádění analýz ostatních projektů a knihoven, jakož i během POC.

Trvalé identifikátory
Trvalé identifikátory tvoří základní infrastrukturu pro dlouhodobou ochranu digitálních dat. Úspěch jakéhokoli systému identifikace je založen především na kvalitní analýze a dobře nastavené administrativě. V digitálním světě poskytují identifikátory jiné možnosti než ve světě analogovém: mohou zdroje nejen identifikovat, ale i přímo zpřístupňovat (tzv. resolution services). Zajištění trvalé dostupnosti zdrojů zprostředkované nějakým systémem identifikace pak komplexnost celého problému značně zvyšuje.

Vlastní technologická implementace systému pro trvalou identifikaci je založena na URN:NBN standardu. V první fázi v polovině roku 2010 vznikla analýza funkčních a nefunkčních požadavků na SW, která by se pak měla odrazit ve vývoji a následně poloprovozu aplikace pro správu a provoz identifikátorů. Ve fázi poloprovozu se omezujeme pouze na zdroje zdigitalizované v NK a v projektu VISK7. Rozšíření funkčnosti pro ostatní knihovny a další profily a funkce pro born-digital dokumenty (především WebArchiv) je plánováno pro rok 2011, po analýze a vyhodnocení poloprovozní fáze.

Zpřístupňování

Zpřístupnění digitálních dokumentů spadajících do Národní digitální knihovny můžeme rozdělit do několika skupin:

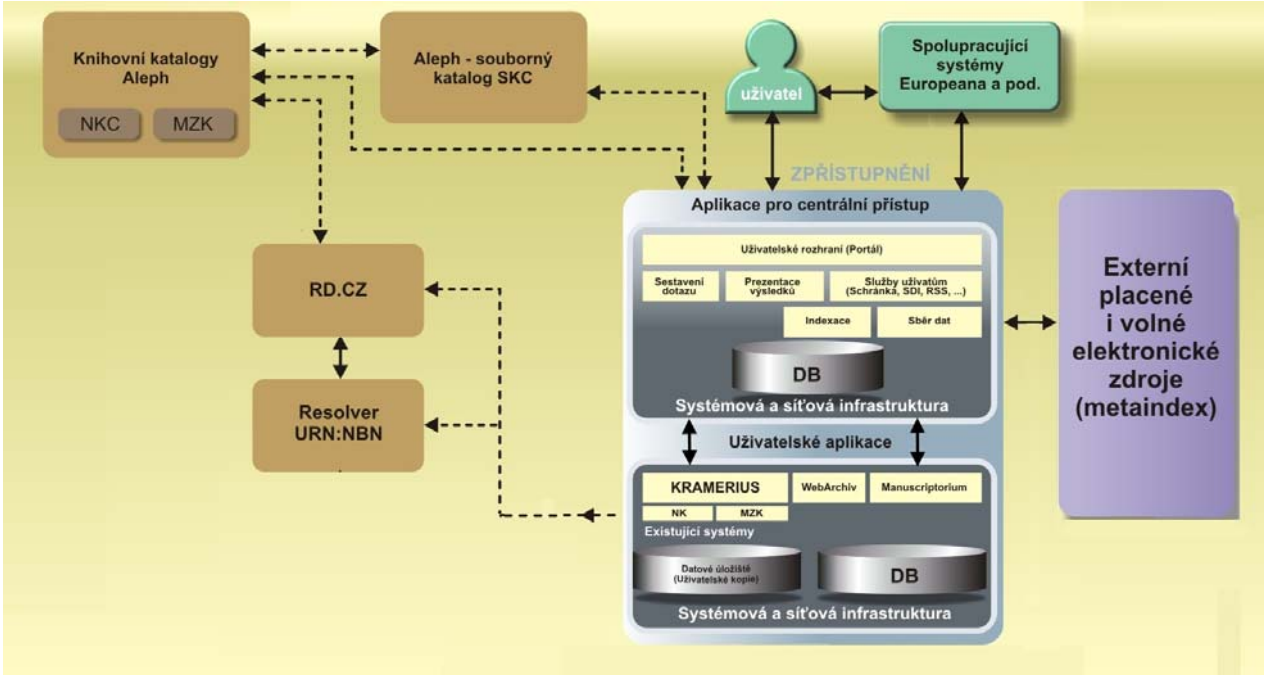
- Existující rozhraní: zpřístupnění pomocí stávajících aplikací Kramerius, WebArchiv, Manuscriptorium
- Aplikace pro centrální přístup: zpřístupnění pomocí zastřešujících systémů a portálů orientovaných na uživatele knihoven nebo zájemce o kulturní dědictví
- Zpřístupnění pomocí externích portálů, vyhledávačů a služeb (mimo oblast kultury)

- **Existující systémy: zpřístupnění pomocí stávajících aplikací Kramerius, WebArchiv, Manuscriptorium**

Rozvoj těchto aplikací má dlouhou tradici a je předmětem samostatných výzkumných záměrů a domácích i zahraničních projektů.

- Aplikace pro centrální přístup**: zpřístupnění pomocí zastřešujících systémů a portálů orientovaných na uživatele knihoven nebo zájemce o kulturní dědictví**

Bude využívat externích indexů např. z placených databází; z vyhledávačů (Google) a bude tyto indexy doplňovat interním indexem připojených aplikací. V našem případě budou těmito připojenými aplikacemi Kramerius NK a MZK, Manuscriptorium, WebArchiv a případné další (katalogy NK a MZK, souborný katalog).



- **Zpřístupnění pomocí externích portálů a vyhledávačů (mimo oblast kultury)**

V kontextu IOP: Elektronizace služeb veřejné správy a Smart Administration vstupujeme do pro nás zcela nové oblasti, kde půjde o zpřístupnění našich digitálních dokumentů nikoli našim více či méně proškoleným uživatelům, ale běžným občanům např. v rámci Portálu veřejné správy. Nejdůležitějším výstupem projektu Národní digitální knihovna v tomto kontextu bude co nejjednodušší zpřístupnění faktografie obsažené v plných textech digitálních dokumentů.

V této pro nás nové oblasti spolupráce bude jistě velký prostor pro výzkum a vývoj.

B.2 Přínos řešitele

Přínos řešitele vyplývá z popisu vlastního řešení v kapitole B1. Nejcennější je přínos řešitele v oblastech, kde řešení přineslo výrazný posun znalostí viz kapitola B3 – Posun znalostí.

B.3 Posun znalostí

K nejvýznamnějšímu posunu znalostí došlo v roce 2010 v dále uvedených oblastech. ***Nejvýznamnější výsledky zásadního významu, které budou uplatněny v rámci hlášení do RIV, jsou vytištěny tučnou kurzívou.***

3. Budování digitálních depozitních knihoven s ohledem na možnost jejich integrace v rámci Jednotné informační brány a nadnárodních portálů

V roce 2010 jsme se autorsky podíleli na vzniku publikace shrnující problematiku dlouhodobé ochrany dat, včetně přístupů a možných řešení.

CUBR, Ladislav. Dlouhodobá ochrana digitálních dokumentů. Praha : Národní knihovna ČR, 2010. 154 s. ISBN 978-80-7050-588-5 (brož).

Dále jsme publikovali příspěvek v recenzovaném periodiku Knihovna:

CUBR, Ladislav. Budování důvěryhodného systému trvalé identifikace digitálních dokumentů. *Knihovna*. 2010, roč. 21, č. 1, s. 23-31. ISSN 1801-3252.

Nejhodnotnějších výsledků jsme dosáhli v těchto oblastech:

- perzistentní identifikátory
  - analýza a návrh využití URN:NBN na datech NK a VISK7;
  - **vývoj aplikace pro NK ČR (poloprovoz)**
  - koordinace užívání identifikátorů, mezinárodní spolupráce
- **vývoj a testování brány SRU/SRW – OpenSearch rozhraní pro vyhledávání nad fulltextovým indexem a jeho integrací s metavyhledávacími portály: výsledek na úrovni R- software**

Velkého pokroku jsme dosáhli v testování, vývoje a finálního nasazení SRU/SRW rozhraní pro vyhledávání nad fulltextovým indexem a jeho integrací s metavyhledávacími portály. Tento počin významně obohacuje možnosti vyhledávání v datech vzniklých archivací internetových stránek, která započala v roce 2000.

Velmi důležité je i nasazení, zatím pilotního, systému pro přiřazování, správu a identifikaci digitálních dokumentů – resolveru URN:NBN a to i přesto, že jde zatím o systém fungující pouze pro data, která jsou uložena v jedné instituci (NK). Je to východisko pro další rozvoj, jehož cílem je národní systém URN:NBN pro různé instituce.

Významného pokroku jsme dosáhli v oblasti specifikace funkčních požadavků národního digitálního repozitáře a systému na dlouhodobou ochranu digitálních dat. Specifikace funkčních požadavků byla jádrem Studie proveditelnosti a v podstatně rozšířené verzi i základem pro odborné podklady pro výběrové řízení na systémového integrátora a dodavatele jednotlivých subsystémů pro projekt NDK. Text obsahuje specifikaci funkčních i nefunkčních požadavků včetně kvalifikačních předpokladů nejen pro systém dlouhodobé ochrany (digitální repozitář) a jeho integraci v rámci různých portálů a katalogů, ale i o funkční specifikaci linek a pracovních postupů (včetně personálních nároků) pro masovou digitalizaci.

C Návrhová část

C.1 Výsledky řešení

Dosažené a dosud neuplatněné výsledky

Následují výsledky řešení dosažené v roce 2010, které budou zavedeny do evidence RIV v roce 2011.

B – odborná monografie

CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů*. Praha : Národní knihovna ČR, 2010. 154 s. ISBN 978-80-7050-588-5 (brož).

J_{rec}- článek v odborném periodiku (časopise)

CUBR, Ladislav. Budování důvěryhodného systému trvalé identifikace digitálních dokumentů. *Knihovna*. 2010, roč. 21, č. 1, s. 23-31. ISSN 1801-3252.

Z_{polop}– poloprovoz

Název výsledku: **SRU/SRW rozhraní pro vyhledávání nad fulltextovým indexem**
Popis výsledku: SRU/SRW rozhraní pro vyhledávání nad fulltextovým indexem a jeho integraci s metavyhledávacími portály a s prohlížecím rozhraním Wayback
Tvůrci výsledku: Libor Coufal, Vlastimil Krejčíř, Lukáš Kopáč, Petr Žabička
Garant výsledku: Libor Coufal

Z_{polop}– poloprovoz

Název výsledku: **Resolver URN:NBN**
Popis výsledku: Vytvořené prostředí a základní SW aplikace pilotního testu pro využití URN:NBN v NK ČR umožní přiřazování identifikátoru, jeho správu, vyhledávání dle identifikátoru a zpřístupnění odkazu na digitální dokument. V pilotní fázi budou přiřazeny identifikátory všem digitalizovaným dokumentům evidovaným v RD.CZ, které vznikly v rámci programu Kramerius, VISK7 a Norských fondů.
Tvůrci výsledku: Jan Hutař, Ladislav Cubr, Incad
Garant výsledku: Jan Hutař

C.2 Závěr

V rámci řešení komplexního výzkumného záměru *Budování vzájemně kompatibilních informačních fondů …* se v NK ČR podařilo dosáhnout výsledků, které mají velký význam pro vědu a výzkum ve všech profilových oborech NK, v celém oboru knihovnictví a informační věda a v neposlední řadě v NK ČR a ostatních knihovnách i jiných paměťových institucí.

C.3 Návrhy opatření

Pro zajištění dalšího rozvoje ve všech oblastech zastřešených výzkumným záměrem *Budování vzájemně kompatibilních informačních fondů*

1. Budování digitálních depozitních knihoven s ohledem na možnost jejich integrace v rámci Jednotné informační brány a nadnárodních portálů v mezinárodním kontextu

2. Optimalizace využití heterogenních informačních zdrojů prostřednictvím jejich integrace v rámci Jednotné informační brány

3. Budování digitálních depozitních knihoven s ohledem na možnost jejich integrace v rámci Jednotné informační brány a nadnárodních portálů

bude, vhledem k ukončení záměru v roce 2010, nutné zajistit jejich nepřetržité a včasné financování z těchto zdrojů:

- výzkum a vývoj: vlastní rozpočet NK, výzkumný záměr (MK ČR), evropské projekty
- provoz a využití výsledků: vlastní rozpočet NK, VISK (MK ČR), evropské projekty

Některé aktivity, řešené v rámci záměru v roce 2010, budou pokračovat dále v normálním ostrém provozu a bude potřeba jen částečný výzkum na přidávání nových funkcionalit (problematika SRU/SRW). Tento rozvoj a provoz musí být nepřetržitě financován:

- ideálně z rozpočtu NK
- z výzkumných záměrů (MK ČR)
- projektů NAKI (MK ČR)

V případě rozvoje a výzkumu systému jednoznačné identifikace za použití URN:NBN bude zcela jistě nutné financování dalšího výzkumu tak, aby vznikla služba identifikace digitálních objektů na národní úrovni, využitelná nejen knihovnami, ale i archivy, muzei apod. Financován musí také být provoz služby, který si bude žádat pracovní síly, technologie apod. Financování výzkumu a vývoje by mělo být poskytnuto z*:*

- projekty NAKI (MK ČR)
- výzkumné záměry MK ČR
- evropské projekty

Financování provozu ideálně z:

- rozpočet NK
- program VISK (MK ČR)

V rámci výzkumného záměru byly zahrnuty oblasti řešení, kde by jakékoli přerušení kontinuity znamenalo nejen pozastavení vývoje, ale těžký, mnohdy nemožný, návrat do výchozího stavu a nenávratné ztráty (budování spolehlivého digitálního úložiště, archivace webu, provoz a rozvoj složitých systémů).

S ohledem na význam záměru, který přesahuje rámec NK a oboru knihovnictví i oboru informační věda, by jakákoli diskontinuita v řešení znamenala citelnou ztrátu pro výzkum a vývoj v mnoha oborech i pro běžné služby českých i zahraničních knihoven.

D Použití finančních prostředků

D.1 Komentář

Poznámka: Podrobný rozpočet obsahuje příloha F1. Žádost o úpravu rozpočtu a schválení úpravy rozpočtu obsahuje příloha F5.

Institucionální podpora:

Investice
Z položky byly zakoupeny v souladu s určením položky – další technické vybavení - především komponenty pro rozšíření serverového prostředí v CDH (server pro WebArchiv aj.) v celkové hodnotě 733 278 Kč. Částka nad limit 210 000 Kč je vkladem NK ČR.

Služby
Z prostředků na služby byly hrazeny především externí služby související s rozvojem WebArchivu včetně propojení s portálem JIB, vývoj softwarové aplikace Resolver URN:NBN od firmy Incad, vydání monografie Dlouhodobá ochrana digitálních dokumentů v celkové výši 366 546 Kč. Částka nad limit 366 000 Kč je vkladem NK ČR.

Cestovné
Z položky cestovné byly hrazeny částečně zahraniční cesty řešitelů v celkové výši 194 932 Kč. Částka nad limit 194 000 Kč je vkladem NK ČR.
Podrobný rozpis je uveden v příloze F3. V příloze F4 jsou podrobně uvedeny i další zdroje financování zahraničních cest řešitelů záměru. Všechny cestovní zprávy jsou obsaženy v příloze F7.

Mzdy
Mzdy v celkové výši 390 000 Kč byly využity na mimořádné odměny řešitelů a dalších pracovníků podílejících se na řešení výzkumného záměru.

Pojištění
Pojištění tvoří povinných 34% zákonných odvodů k položce mzdy. Nedočerpaná částka byla převedena na položku služby.

FKSP
FKSP tvoří povinná 2% zákonných odvodů k položce mzdy. Nedočerpaná částka byla převedena na položku služby.

Odpisy
Odpisy ve výši 80 000 Kč se vztahují k digitálnímu úložišti.

Přesuny mezi jednotlivými položkami byly povoleny MK ČR (viz Příloha F6)

Přečerpané částky na všech položkách byly uhrazeny z rozpočtu NK.

Vklad NK ČR:

V předloženém záměru se NK ČR zavázala vložit v roce 2010 do aktivit souvisejících s realizací záměru celkem 18 446 000 Kč (21 096 – 2 650 000 Kč). I když institucionální podpora byla snížena z požadovaných 2 650 000 Kč na reálných 1 375 000 Kč, vklad NK byl s ohledem na podporu nových souvisejících aktivit zejména v souvislosti s budováním digitální knihovny podstatně vyšší než plánovaná částka (celkem 19 776 545 Kč). Vklad NK ČR tvoří kromě výše uvedených drobných vkladů (přečerpané částky) v celkové výši 6878 Kč osobní náklady zaměstnanců a k nim vztažené režijní náklady. Vkladem NK ČR je i úhrada dalších cest řešitelů z rozpočtu NK ČR. Tato položka je v roce 2010 relativně

nízká. Důvodem je úhrada řady cest řešitelů výzkumného záměru ve 2. pololetí 2010 z cestovného projektu NDK.

Řešení záměru vyžaduje spolehlivé fungování základu digitálního úložiště, několika serverů a velkého množství koncových stanic, které je nutné udržovat a postupně obnovovat, nezbytnou podmínkou je kvalitní a nákladné síťové připojení. Do provozních nákladů je započtena i poměrná část nákladů na externí subjekty podílející se na provozu základního i aplikačního SW zejména pro digitální knihovnu včetně nezbytného programování specifických aplikací. NK ČR dokrývá ze svého rozpočtu náklady na cestovné řešitelů záměru. Podrobnosti uvádí přílohy F4 a F6.

E Resumé a klíčová slova

E.1 Resumé a klíčová slova v češtině

Resumé:
Předmětem výzkumné činnosti realizované ve výzkumném záměru *Budování vzájemně kompatibilních informačních systémů pro přístup k heterogenním informačním zdrojům a jejich zastřešení prostřednictvím Jednotné informační brány* je výzkum a vývoj směřující k vytvoření informačních systémů pro přístup k heterogenním informačním zdrojům, které budou navzájem kompatibilní do té míry, že bude možné je zastřešit tak, že se budou navenek (tj. pro koncového uživatele) prezentovat jako systém jediný. Jedná se o velmi komplexní výzkumný záměr, který v sobě integruje výzkumnou činnost v několika vzájemně provázaných oblastech: optimalizace věcného zpřístupnění dokumentů s ohledem na integraci v mezinárodním kontextu (kombinace vyhledávání v plných textech a řízených slovnících, konkordance klasifikací, aplikace metody Konspektu); optimalizace využití heterogenních informačních zdrojů prostřednictvím jejich integrace v rámci Jednotné informační brány (jednotné prostředí, jednotné kladení dotazů, jednotné výstupy, vlastní prostředí, přidané služby); budování digitálních depozitních knihoven s ohledem na možnost jejich integrace v rámci Jednotné informační brány a nadnárodních portálů. Výsledky dosažené v průběhu řešení záměru ve všech uvedených oblastech jsou srovnatelné s výsledky nejvyspělejších zemí v dané oblasti.

Klíčová slova:
Informační systémy * portály * jmenné zpracování * věcné zpřístupnění * integrace informačních zdrojů * digitální úložiště

E.2 Abstract and key words in English

Abstract:
The aim of the research plan **"Building of Mutually Compatible Information Systems for Access to Heterogeneous Information Resources under the Umbrella of the Uniform Information Gateway"** is research into, and development of, information systems for access to heterogeneous information resources that will be mutually compatible to such an extent that it will be possible to put them under one umbrella in such a way that for the external environment (i.e. for the final user) they will work as a single system. It is a very comprehensive project that integrates research activities in a number of related subjects: optimisation of subject-based access to documents with an emphasis on the international context (a combination of searches in full texts and controlled vocabularies, concordance of classifications, application of Conspectus principles); optimisation of the use of heterogeneous information resources by their integration into the Uniform information gateway (uniform environment, uniform queries, uniform outputs, user's own environment, extended services); building of digital repositories to be integrated under the umbrella of the Uniform information gateway and other portals. Research results achieved during the NL involvement in the research plan are comparable with those achieved in countries known as most advanced in this area.

Key words:
Information systems * portals * bibliographic description * subject access * integration of information resources * digital repositories

F Přílohy

F1 Monografie Dlouhodobá ochrana digitálních dokumentů
F2 Periodikum Knihovna 2010, roč. 21, č. 1.
F3 Rozpis rozpočtu
F4 Čerpání finančních prostředků – podrobné tabulky a další doklady
F5 Financování služebních cest řešitelů z rozpočtu NK ČR
F6 Žádosti o úpravu rozpočtu, schválení úpravy rozpočtu
F7 Vklad řešitele – struktura
F8 Prohlášení o dostupnosti účetních dokladů
F9 Cestovní zprávy z cest hrazených z výzkumného záměru