RESOLVER – podklad pro další rozvoj v roce 2011

Informační podklad pro další rozvoj aplikace resolver, obsahuje:

1 Rozšíření resolveru URN-NBN 2011 – popis nově požadované funkcionality

Stručný úvod

U stávajícího resolveru URN:NBN (JAVA aplikace vyvinutá firmou INCAD, s databází na Oracle – http://resolver.nkp.cz, dokumentace a zdrojový kód na http://code.google.com/p/urnnbnresolver/) chceme rozšířit funkcionality.

V současnosti je jedinou instalovanou službou import dat z Registru digitalizace (http://www.registrdigitalizace.cz) a vracení dat do tohoto Registru.

Chceme, aby resolver dovedl získat metadata z univerzitních repozitářů a dalších zdrojů. Tato metadata by měl být schopen přijmout a uložit do své databáze. Tato metadata budou již obsahovat URN:NBN identifikátory přidělené mimo resolver, v systémech poskytovatelů dat.

Pro zajištění této funkcionality bude nutné:

1) Provést úpravy databáze resolveru – změny polí metadat a jejich nastavení, nové tabulky.
2) Upravit databázi tak, aby byl resolver schopen přijímat již hotová URN-NBN, resp. aktualizovat existující záznamy
3) Vytvořit rozhraní, službu (pravděpodobně OAI-PMH) pro sklízení, kontrolu (především kontrola úplnosti metadatového záznamu, kontrola duplicit, kontrola korektnosti /shoda se strukturou řetězce URN-NBN/, kontrola duplicit URN-NBN, kontrola funkčnosti odkazů URL vázaných k URN) a ukládání metadat do databáze resolveru.

Podrobnější popis

Spolupráci pro příjem metadat máme domluvenu s následujícími datovými repozitáři:

Repozitář Univerzity Karlovy

typ dat - vysokoškolské dokumenty (objem cca 30 tis. dokumentů)
tento repozitář obsahuje ještě jiné typy dat, naše zadání se však omezuje jen na jednu sbírku, a to VŠKP (bakalářské, magisterské, rigorózní, dizertační, habilitační práce)
systémy / zdroje dat:
repozitář – systém Digitool (http://repozitar.cuni.cz), katalog
metadata:
MARC21, DC, interní metadata repozitáře
Dostupné rozhraní: OAI-PMH

NUŠL (Národní technická knihovna)

typ dat - šedá literatura (objem max. 18 tis. dokumentů)
omezení - pouze to, co je majetkem NTK
systémy / zdroje dat:
repozitář – systém CDS Invenio (http://invenio.ntkcz.cz/), katalog
metadata:
MARC21, DC, NUŠL formát, interní metadata repozitáře
Dostupné rozhraní: OAI-PMH

Ikaros (online časopis) – v jednání

typ dat:
články časopisu – cca 8 tisíc článků
systémy / zdroje dat:
systém DRUPAL (http://ikaros.cz/)
metadata:

interní formát DB
Dostupné rozhraní: OAI-PMH modul pro Drupal

ÚKOL (ve spolupráci s ODO NK):
Ověřit, jak všechny tyto systémy fungují – technické řešení, možnosti vystavení a zasílání metadat, možnosti přijímání metadat, případně workflow ukládání dat a metadat – analýzy metadat máme již hotové

1.1 Pravidla pro přidělování identifikátoru (formát „urn:nbn:cz:B-C“) u nových registrátorů

ást B (identifikace podřízeného jmenného prostoru) je v gesci daného registrátora
návrh kódů:
Univerzita Karlova = uk / tedy: urn:nbn:cz:uk-C
NUŠL/NTK = aba013 / tedy: urn:nbn:cz:aba013-C
Ikaros = ik / tedy urn:nbn:cz:ik-C
Pro část C
Každý registrátor si v rámci této části bude sám přidělovat vlastní znakový řetězec
část C se musí skládat z 1-6 alfanumerických znaků
Registrátor sám ručí za korektnost URN (tj. že jeho jedinečnost a jeho soulad s pravidly)

1.2 Nové případy užití („use cases) pro uvedené registrátory

1.2.1 Přidělení URN-NBN současným datům

přidělování většímu objemu dat, jednorázově a zpětně (tzn., že data jsou již vyprodukována a uložena, nemají URN:NBN)
Předpokládané workflow:
1) registrátor si sám dodatečně vygeneruje URN a přidělí ho svým stávajícím dokumentům (do metadat, tj. v databázi / do katalogu apod.)
2) takto přidělené URN však nesmí být vidět navenek – nesmí se zatím zobrazovat v katalogu, ani na úvodní html stránce dokumentu apod.
3) registrátor zašle dávkově do resolveru celý balík takto přidělených URN, včetně povinných metadat (včetně URL) vážících se k těmto URN, která vyžaduje resolver
4) resolver zkontroluje korektnost URN a pokud:
5a) URN je korektní, pak:
5.1a) resolver zašle registrátorovi potvrzení, že URN jsou v pořádku
5.2b) odteď může registrátor tato URN uvádět i navenek (v katalogu, na webu apod.)
5b) URN není korektní
5.1b) resolver vrátí chybné URN (včetně záznamu?) zpět registrátorovi, ten věc dá do pořádku, a zašle do resolveru opravené URN
5.2b) vracíme se do bodu 4)
6) resolver si uloží URN s přidruženými metadaty do své DB
7) odteď již resolver může přesměrovávat z URN na aktuální URL
anebo na metadatový záznam vygenerovaný z DB (tato funkcionalita funguje již nyní)

Možné problémy

Z nějakého důvodu se může stát, že registrátor nebude schopen si sám generovat část C (nemá na to nástroje, nebo ztratí informace o tom, co již přidělil apod.), proto:
je třeba zvážit reálné riziko této situace

pokud riziko opravdu existuje, pak:
navrhnout řešení, které by předpokládalo, že si registrátor nainstaluje speciálně vyvinutý generátor čísel, který bude mít u sebe lokálně a který se bude schopen dotazovat centrálního resolveru, odkud má pokračovat atd.

ÚKOL:
vytvořit technický návrh s rozpisem kompletního nejvhodnějšího postupu pro výše uvedený use case, včetně rozpisu všech potřebných systémů (co jaký systém bude dělat, jak a v jaké posloupnosti, jaké jsou varianty pro různé situace, jaké jsou nejvhodnější protokoly apod.)
předpokládáme, že by měl být užit protokol OAI-PMH (resp. OAI-PMH harvester)
nutno zvážit, zda je, či není třeba vytvořit generátor k lokální instalaci jako službu navíc, ačkoliv my nepředpokládáme, že by toto byl reálný problém (každá DB by to měla umět sama?)
vyřešit, jak zajistit, aby do doby, dokud nebude URN-NBN zkontrolováno resolverem, nebylo nikde vidět navenek (v katalogu, na html stránce s dokumentem apod.)
pokud uživatel klikne na URN (zapsané ve tvaru URL /tj. HTTP adresa resolveru + URN/) v katalogu daného registrátora, resolver musí poznat tuto situaci (tj. dotaz směřuje z katalogu), a přesměrovat přímo na html stránku digitální knihovny daného registrátora
pokud uživatel zapíše URN daného registrátora přímo do resolveru (tj. dotaz byl přímo zadán do okna resolveru), pak resolver nejprve zobrazí metadatovou stránku (příklad: http://resolver.nkp.cz/urnnbn/#urn:nbn:cz:aba001-0001qc), na které je uveden seznam aktuálních URL, a teprve kliknutím na dané URL je uživatel přesměrován na konkrétní adresu
důvod: v prvním případě uživatel již viděl v katalogu metadata, nepotřebuje je vidět
tato funkcionalita musí fungovat, ale je možné, že nebude vždy využívána

1.2.2 Aktualizace metadat v resolveru

Pokud registrátor změní URL (nebo další metadata) k URN – nastanou varianty řešení:
1) registrátor okamžitě pošle nový údaj do resolveru, aby aktualizoval svou DB
2) resolver naopak v pravidelných režimech bude sklízet z repozitářů aktuální metadata
3) registrátor pouze pošle resolveru informaci „teď sklízej, aktualizovali jsme metadata“

ÚKOL:
rozhodnout, jaká z těchto variant je nejlepší
navrhnout postup

1.2.3 Přidělování pro nově vznikající data

Na rozdíl od 1.2.1 nebude toto přidělování tak masivní
Registrátor bude generovat URN hned při vzniku metadat / záznamu k dokumentu ve své digitální knihovně
jinak stejný postup jako 1.2.1

1.2.4 Vystavení metadat z resolveru jiným systémům

Rozšířit resolver o možnost, aby z něj mohly naopak data sklízet jiné subjekty

ÚKOL
K resolveru nainstalovat OAI-PMH repository nebo něco podobného

2 Úprava databáze

ÚKOLY:
Rozšíření stávajících polí/tabulek dle našeho zadání
Úprava datového modelu dle našeho zadání
Spojení některých polí na základě parametru „typ dokumentu“ dle našeho zadání
Příklad – pokud chceme zadávat metadata k identifikátoru článku, nepotřebujeme vidět pole ISBN, pokud k monografii, nepotřebujeme vidět ISSN, pokud k vysokoškolské kvalifikační práci, pak se ukáží speciální podpole (např. „diplomová práce (Mgr.)“)

3 Rozšíření účtů

Možnost samostatné správy pro každý podřízený prostor (UK, NTK, IKAROS)
V současnosti pouze jeden účet
Jako login například kód registrátora
Jednotliví registrátoři si budou moci - manuálně - upravovat údaje, ale pouze ty, které jsou v rámci jejich podřízeného prostoru (část „B“ v URN)

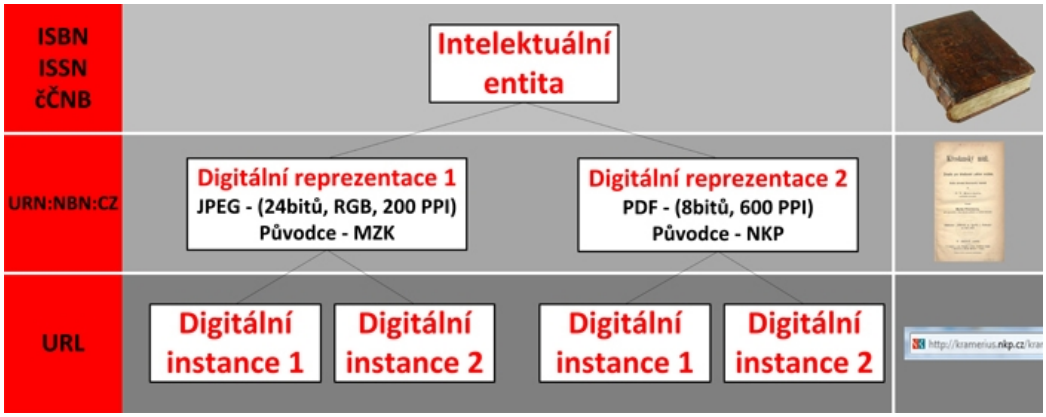
4 Základní informace o resolveru

NK ČR využívá systém trvalé identifikace podle de facto standardu URN-NBN (dále též jen jako „URN“)
Centrální registrační autoritou je NK ČR
NK ČR provozuje centrální službu (resolver) na této adrese http://resolver.nkp.cz/
Dokumentace je na http://code.google.com/p/urnnbnresolver/
Struktura URN pro ČR má tuto podobu - urn:nbn:cz:B-C

Část „urn:nbn:cz:“ je závazná a stejná pro všechny identifikátory URN v ČR
B = kód podřízeného registrátora (2-6 alfanumerických znaků)
6 znaků je vyhrazeno siglám, registrátor bez sigly musí mít méně jak 6 znaků
C = alfanumerický znakový řetězec jedinečný v rámci „B“ (1-6 AN znaků)
Poznámka: u URN nezáleží na velikosti písmen

Ke každému URN v DB resolveru vyžadujeme skupinu metadat (bibliografická, technická, administrativní) – viz datový model, schéma Schéma

4.1 Současný datový model (pro digitalizované dokumenty)



- Digitální dokument označovaný identifikátorem URN-NBN je jednou digitální reprezentací jedné intelektuální entity

Intelektuální entita

- je intelektuálním obsahem dokumentu
- je to, co má jako celek bibliografický záznam
- digitálním vyjádřením intelektuální entity je digitální reprezentace
- táž intelektuální entita může být vyjádřena několika digitálními reprezentacemi
- může mít identifikátor CNBN (čČNB) (Marc21 – pole 015), ISBN, ISSN

Digitální reprezentace

- je tvořena množinou počítačových souborů
- tyto soubory reprezentují jak vlastní obsah, tak metadata
- její reprodukcí (softwarové / hardwarové zprostředkování) dochází k zpřístupnění intelektuální entity uživateli
- je popsána technickými metadaty - informace o formátu, digitalizaci apod.
- jedna intelektuální entita může mít více různých digitálních reprezentací
- zásadní podmínka - jiný vlastník dokumentu = (vždy) jiná digitální reprezentace
- jednotlivé digitální reprezentace se liší na základě stanovených signifikantních vlastností
    - pokud nejsou stanoveny, platí pouze jediná signifikantní vlastnost - vlastník
- po migraci do nových formátů zůstává digitální reprezentace stejná, nemění se URN
- digitální reprezentace má identifikátor URN (pokud je přidělen v našem systému)

Digitální instance

- je identickou kopií digitální reprezentace umístěnou na konkrétní internetové adrese
- identita na bitové úrovni (shodnost kontrolních součtů)
- jedna digitální reprezentace může existovat ve více digitálních instancích
- má identifikátor (lokátor) URL

5 Současný stav resolveru URN-NBN

5.1 Popis stavu

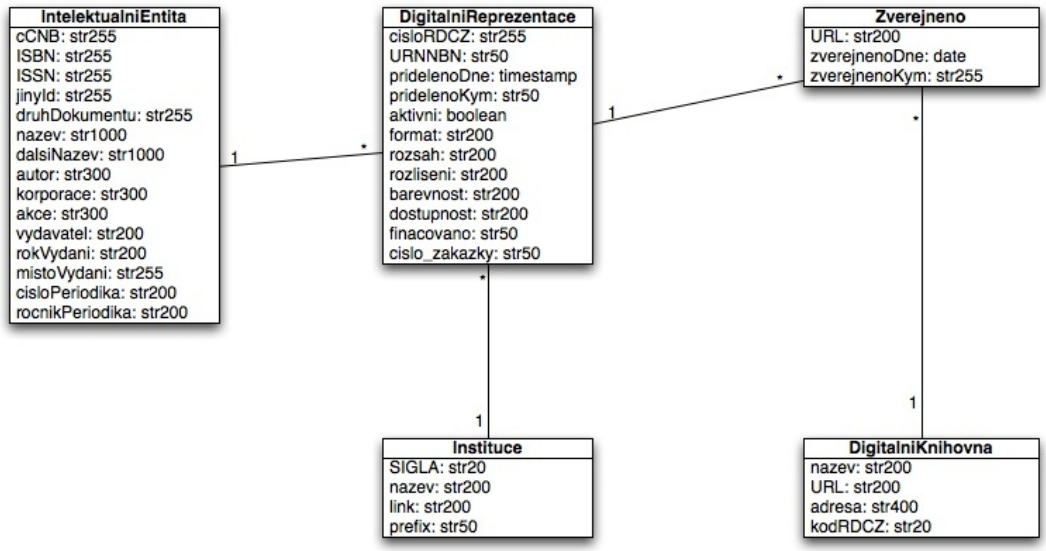
- Zdrojem dat jsou výhradně metadata z RD-CZ – je to jediné komunikační rozhraní, které resolver v tuto chvíli má
- Identifikace pouze některých digitalizovaných dokumentů
- vývoj firma Incad v roce 2010 – technická specifikace viz http://code.google.com/p/urnnbnresolver/wiki/Uvod?tm=6
- **resolver je ve správě NK, neprobíhají žádné další úpravy**
- na resolver není žádná servisní smlouva – vše je v rukách NK

5.2 Popis funkcí

- Přidělení jednoznačného identifikátoru URN:NBN v rámci českého národního prostoru (URN:NBN:CZ)
- Správa přidružených metadat (metadata k identifikátoru) v databázi
- Možnost manuální administrace (vkládání a úpravy dat ručně)

5.3 Datový model

5.3.1 Schéma



- Model je určen pouze pro digitalizované monografie a seriály, v případě seriálu se jedná o provizorní model.

5.3.2 Typologie dokumentu

- Ve stávající aplikaci jsou pouze tyto dva typy dokumentu:
- BK = monografie v současném RD-CZ
- SE = část seriálu v současném RD-CZ

5.4 Sestavení, instalace, konfigurace a správa uživatelských účtů

- viz http://code.google.com/p/urnnbnresolver/wiki/Instalace

5.5 Administrace

- viz http://code.google.com/p/urnnbnresolver/wiki/Administrace