Zpráva ze služební cesty

Jméno a příjmení účastníka cesty	PhDr. Ladislav Cubr
Pracoviště – dle organizační struktury	Odbor digitálních fondů
Pracoviště – zařazení	výzkumný a vývojový pracovník Oddělení standardů
Důvod cesty	Návštěva Národní knihovny Francie (BNF)
Místo - město	Paříž
Místo – země	Francie
Datum (od – do)	11. -16. listopadu 2012
Podrobný časový harmonogram	11. 11. přílet, 12.-16.11 návštěva knihovny
Spolucestující z NK	Zuzana Kvašová, Rudolf Kreibich
Finanční zajištění	Projekt VAV 0136
Vztah k projektu	Systém trvalé identifikace v BNF a jeho vazba na digitální repozitář (LTP) a digitální archivaci
Cíle cesty	Získání nových zkušeností a znalostí v oblasti digitální archivace
Plnění cílů cesty	Zjištění specifického přístupu k řešení otázek trvalé identifikace a digitální archivace v rámci digitální archivace v BNF
Další podrobnější informace	navštívena hlavní pobočka BNF (Knihovna F. Mitterranda)
Podpora publicity projektu	Informace o českém systému trvalé identifikace

Související materiály	
Materiál	***Místo uložení***
Stránka BNF	http://www.bnf.fr/fr/la_bnf/sites/a.site_francois-mitterrand.html
Webová stránky vyvíjené aplikace	https://resolver.nkp.cz/
Zdrojové kódy vyvíjené aplikace	http://code.google.com/p/urnnbn-resolver-v2/

Datum předložení zprávy	27. listopadu 2012
Podpis předkladatele zprávy	Ladislav Cubr

	Datum	***Podpis***
Podpis nadřízeného		
Vloženo na intranet		
Přijato v mezinárodním oddělení		

Harmonogram návštěvy

Pondělí 12/11/2012
Základní představení práce webarchivu spadajícího pod oddělení povinného výtisku
Sklízení webových stránek z knihovnického hlediska

Úterý 13/11/2012
Digitální archivace webových stránek (příprava deponování sklizených webových stránek do digitálního repozitáře)
Sklízení webových stránek z technického hlediska

Středa 14/11/2012
Úvodní prezentace systému SPAR, digitálního repozitáře (LTP) vyvíjeného v BNF
Konceptuální návrh pro SPAR (metadata, ontologie, systém trvalé identifikace ARK)

Čtvrtek 15/11/2012
Konceptuální návrh systému SPAR (ingest, pre-ingest, archivace webu)
Prohlídka audiovizuálního oddělení v BNF (čtenářské hledisko, dočasné úložiště dat)

Pátek 16/11/2012
Technické řešení pro systém SPAR (virtualizace)
Závěrečné setkání – výměna poznatků a prezentace českého přístupu

Shrnutí základních poznatků z hlediska relevance pro NK ČR

BNF zhruba od poloviny minulého desetiletí vyvíjí ve spolupráci s externí firmou (ATHOS) systém pro digitální archivaci SPAR. Systém by měl být open-source. Tento systém je založen na standardu OAIS, široce užívaných metadatových schématech a specifické implementaci systému trvalé identifikace založené na schématu ARK (Archival Resource Key).

První funkční entitou systému, která byla v rámci systému SPAR vyvinuta, byla archivní entita. Archivní entita byla projektována na základě konceptu virtualizace uložení fyzických dat. Z následně vyvinutých částí OAIS je klíčovou funkční entitou systému SPAR depozitní entita (ingest). Plánovací entita (preservation planning) nebyla dosud v rámci vývoje systému SPAR vytvořena.

Datový model pro SPAR je založen na čtyřech stupních granularity: soubor – objekt – skupina – agregace. Soubor reprezentuje objekt (například 1 TIFF reprezentuje 1 stránku tištěné knihy). Skupina (tj. skupina souborů-objektů) je základní intelektuální jednotkou modelu a hlavním předmětem archivace (archivní balíček odpovídá úrovni skupiny). V některých případech je současně předmětem archivace též agregace (skupina skupin), například ARC (jako archivní balíček webarchivu) je na úrovni skupiny, celá sklizeň na úrovni agregaci. V tomto případě jsou metadata popisující agregaci deponována v samostatném informačním balíčku.

Hlavním datovým formátem pro dlouhodobou archivaci digitalizovaných dokumentů je v současnosti v BNF formát TIFF. Hlavním metadatových schématem pro SPAR je METS.

Pro jednotlivé dílčí předměty metadatového popisu jsou užita tato schémata: PREMIS, DC (bibliografická metadata), MIX (technická metadata pro TIFF), MPEG-7 (technická metadata pro zvuková data a video), CONTAINER MD (technická metadata pro balíčky ARC z webarchivu), TEXT MD (pro textová data). V současnosti je hlavním předmětem zájmu archivace probíhající digitalizace, starší data z digitalizace nebo data jiné povahy budou do systému SPAR deponována později. BNF

plánuje do konce roku 2012 začít proces deponování balíčků webarchivu, zpočátku pouze aktuální sklizně.

Vzhledem k tomu, že digitalizace v BNF stále užívá původní metadatový formát (RefNum), vyvinutý v 90. letech 20.století, proces deponování digitalizovaných dat do systému SPAR je založen na úvodní úpravě deponovaných dat (modul pre-ingest) do metadatového schématu pro systém SPAR.

Z hlediska organizačního řízení je základním prvkem systému SPAR systém dohod mezi oddělením BNF, starajícím se o provoz systému SPAR (IT oddělení), a ostatními odděleními BNF (například oddělení digitalizace nebo webarchivu, dodávající nebo mající v bezprostřední budoucnosti úmysl dodávat do systému SPAR svá data), nebo externími subjekty. SPAR v těchto dohodách deklaruje úroveň archivačních služeb. BNF definuje čtyři úrovně archivace podle možností operací s archivovanými datovými formáty – od bitové ochrany po plnohodnotnou dlouhodobou archivaci. Podmínkou nejvyššího stupně archivace je to, že BNF vlastní kompletní dokumentaci k danému formátu (to je případ formátu TIFF).

Úroveň archivačních služeb se odvíjí podle aktuálních finančních a znalostních možností BNF a podle jejích zákonných povinností, a dále podle typu jednotlivých dat (např. digitalizovaná kniha x zvuková data). Tyto dohody s producenty dat jsou také archivovány jako informační balíčky, a sice jako specifický typ informačního balíčku, tzv. referenční balíček.

Schéma pro METS je pro systém SPAR strukturován tak, aby jej bylo možné automaticky přemapovat do formátu RDF (Resource Description Framework), s kterým pracuje systém SPAR v rámci datové entity OAIS (správa databáze) pro potřeby vyhledávání apod. BNF navrhla vlastní implementaci formátu RDF, založenou primárně na převzetí některých částí několika zavedených ontologií (např. DC) pro popis prvků a jejich vztahů. Každý termín je jednoznačně identifikován schématem INFO:URI.

Schéma pro RDF v rámci systému SPAR obsahuje zhuštěné informace o informačních balíčcích (tj. nepřebírá opakující se informace o formátech jednotlivých souborů) – jeho výhodou je nižší velikost metadat. Při ztrátě databázových údajů je možné RDF vždy znovu automaticky vygenerovat z metadat uložených v balíčcích (jednotlivé archivní balíčky obsahují vždy datové formáty a příslušná metadata v METS a dalších vnořených metadatových formátech).

Pro identifikaci dat slouží v BNF specifická implementace systému identifikace založeného na schématu ARK (Archival Resource Key). Systém identifikace v BNF zohledňuje primárně technické hledisko předmětu identifikace (identifikace archivních balíčků a jejich verzí). BNF je z hlediska užití systému ARK výjimkou mezi význačnými národními knihovnami světa. Systém ARK je totiž primárně rozšířen zejména v americkém prostředí univerzitních knihoven, nikoliv v prostředí národních knihoven. Prostřednictvím systému ARK je v BNF možné identifikovat pouze vyšší úrovně datového modelu (tedy skupinu a agregaci).