ZPRÁVA O ČINNOSTI NÁRODNÍHO ARCHIVU V PROJEKTU INTERPI ZA ROK 2012

INTERPI – Interoperabilita v paměťových institucích

Program aplikovaného výzkumu a vývoje národní a kulturní identity (NAKI) (DF11P01OVV023)

Zpracoval:
Miroslav Kunt

INTERPI

7. 11. 2012 | verze 1.0

Obsah

1 Úprava, doplňování a analýza entit pro potřeby archivů

1.1 Tvorba geografických entit typu sídlo

Stejně jako v roce 2011 pokračovala práce na pořizování geografických entit typu sídlo ze dvou hlavních zdrojů.[1] V případě jmen sídel v Čechách, resp. jejich historických jmen od středověku do 19. století se ukázala náročnost úkolu, která spočívá v identifikaci jména uvedeného v citacích dobových pramenů (často latinsky a německy), kdy citace obsahuje celé nebo zkrácené věty, často s několika jmény, za ní následuje citace pramene atd. To vše znesnadňuje identifikaci jména dotčené obce z dané doby. Obdobně se také některá jména od doby, kdy byl zdroj (Profous, Antonín: Místní jména v Čechách) publikován, změnila.

Zpracování jmen sídel Slovenska podle vlastivědného slovníku z druhé poloviny 70. let[2] bylo v roce 2011 dokončeno a připraveno pro databázi INTERPI. Další území bývalého Československa, Podkarpatská Rus, bylo zpracováno podle dobových lexikonů, takže zachycuje nejnověji stav z roku 1930. Snaha získat informace od kolegů z Ukrajiny nebyla úspěšná, avšak podle údajů dalšího spolupracovníka projektu, ing. arch. Kuči, část novějších dat k dispozici má.Proto mu byla data zpracované Národním archivem počátkem roku 2012 také předána.

1.2 Entita liniová stavba - přehled úseků a vzniku železničních tratí Česka

S ohledem na potenciál a nedostupnost zdroje byl prováděn popis železničních tratí podle rukopisu uloženého v tzv. Archivu Českých drah v Praze-Libni.[3] Jedná se o unikátní soubor dat o vzniku nejen jednotlivých železničních tratí, ale i stanic a zastávek. Tak lze v celé šíři demonstrovat problematiku liniových i bodových staveb, které mohou navíc měnit trasu/polohu.

1.3 Korporativní entita – spolky

Pro účely budoucího zpracování entit korporace typu spolek je upravován soupis spolků (jména spolků), který vznikal v Národním archivu v minulých 20 letech. Data ve stávající podobě nebyla použitelná, proto se přikročilo k jejich čištění. V této souvislosti byl doplněn i funkční model v případě korporací o data působnosti. U spolků totiž fakticky nelze stanovit ani jejich vznik, nehledě na nemožnost definování tohoto okamžiku v různých historických obdobích (je jím ustavující schůze, schválení stanov, ustavení přípravného výboru apod.?).

---

[1] Profous, Antonín: Místní jména v Čechách. Jejich vznik, původní význam a změny. Nakladatelství Československé akademie věd, 1954-1960, díl I-IV + dodatky (V); Hosák, Ladislav – Šrámek, Rudolf: Místní jména na Moravě a ve Slezsku I, II. Academia Praha, 1970, 1980

[2] Kol.: Vlastivedný slovník obcí na Slovensku I-III. VEDA, vydavateľstvo Slovenskej akadémie vied, Bratislava 1977-1978

[3] Josef Panáček: Historická příručka ČSD 1838 - 1938. České kraje.

1.4 Záznamy o původcích typu korporace

V průběhu roku probíhaly práce na pořizování záznamů o původcích – částečně na základě prací na Základních pravidlech pro zpracování archiválií, částečně za účelem praktického odzkoušení datové struktury budoucí databáze INTERPI a jejího funkčního modelu. Tak byly zpracovány podrobné záznamy o regionálních korporacích různých typů z oblasti východních Čech, současných korporacích v oblasti státní správy. Současně byly shromážděny podklady pro další práci (zejména korporace 1848-1918), ovšem jejich zpracování je závislé na stabilizaci funkčního konceptu a datové struktury, protože pozdější úpravy by byly neefektivní.

2 Komparace rejstříků archivních pomůcek Národního archivu

V souvislosti s přípravou na přechod k sdílené databázi INTERPI bylo provedeno předběžný test kompatibility rejstříkových hesel archivních pomůcek Národního archivu s dosavadní databází národních autorit ČR. Zadání předpokládalo následující postup:

- strojově porovnávaná vstupní data jsou v MS Excell nebo databázi MS SQL 2007. Obsahují identifikaci a heslo – slovo nebo skupinu slov (textový, resp. alfanumerický řetězec). Pro účely snazšího porovnání v následných krocích (uživatelem) zadavatel vstupní data rozšíří o další informace (popis archiválií, ke kterým se data váží);
- referenční data jsou dostupná pomocí protokolu Z 39.50 z databáze Národních autorit ČR (viz http://autority.nkp.cz/zakladni-informace/pristup-do-baze-autorit-pres-z39.50);
- z referenčních dat se zpracovávají všechny podoby jména (preferovaná i nepreferovaná podoba);
- cílem je porovnat zmíněná hesla ze vstupních dat s hesly dat referenčních, najít úplnou shodu, částečnou shodu nebo podobnost (např. na základě začátku jednotlivých slov) a nabídnout párování vstupních a referenčních dat k odsouhlasení zadavateli;
- po odsouhlasení jednotlivých propojení vstupních a referenčních dat pracovníkem zadavatele dojde k uložení informace o referenčních datech do dat vstupních tak, aby vazba mezi porovnávanými a referenčními daty byla jednoznačná a trvalá (identifikátor i řetězec-jméno);
- nástroj nebo skripty, které vzniknou, budou (pokud vyhovují definici autorského díla) licencovány jako Open-source a zpřístupněny zadavateli, který je oprávněn s nimi v rámci licence Open-source nakládat.

Práce byly zadány formou veřejné zakázky malého rozsahu výzvou třem dodavatelům, přičemž kvalifikačním kritériem byla znalost komunikačního protokolu Z 39.50 a hodnotícím kritériem se stala nejnižší cena a nejkratší doba realizace. Ze soutěže vyšla jako vítěz firma Cosmotron Bohemia, s.r.o., které provedla dle zadání porovnání dat.

Termíny z rejstříků byly rozděleny do těchto skupin:

1. tematická hesla
2. jména osob
3. geografická jména
4. korporace

Termíny v každé skupině byly analyzovány a pro každou skupinu termínů byl stanoven postup pro hledání odpovídajících autorit v bázi národních autorit. Očekávaným výsledkem byl soubor terminů doplněný o informace o autoritě nebo autoritách, které termínu potenciálně odpovídají. Byly doplněny informace: záhlaví, kód záznamu v bázi národních autorit a případně poznámka (podle toho, o kterou skupinu terminů se jednalo). Součástí byl také výraz v selekčním jazyku PQF, který se použil pro vyhledání.

Jako hlavní problém byl při vyhledávání prostřednictví protokolu Z39.50 hned v úvodě stanoven obsah indexů pro vyhledávání. V bázi národních autorit není dostupný index, který by obsahoval

pouze záhlaví. Současné indexy jak pro jmenné, tak pro věcné autority obsahují všechny odkazové formy jména. Při této struktuře indexu je velmi obtížné hledat termíny podle jednotlivých slov, protože mezi výsledky jsou zařazeny také autority, které jedno z hledaných slov obsahují v záhlaví a druhé v odkazové formě. Hledání podle jednotlivých slov by bylo výhodné především v případě tematických hesel z rejstříků Národního archivu, kde se používá invertovaný zápis hesel.

Dalším problémem byla skutečnost, že indexy pro osobní jména a korporace odkazují do společného indexu jmenných autorit a indexy pro geografické, věcné autority a autority pro formu a žánr odkazují do společného indexu věcných autorit, přesto že všechny mají přidělené samostatné atributy protokolu Z39.50. Výsledkem na dotaz na osobní jméno může být ve výjimečných případech autorita korporace podobně jako na dotaz na geografické jméno autorita věcná.

Odhalení těchto problému je důležité pro vymezení postupů pro spolupráci jiných paměťových institucí.

Navržené propojení na národní autority se bude postupně posuzovat pro každou skupinu terminů individuálně. Na základě výsledků se bude přistupovat k dalším úpravám dat. Z výsledku v následujících podkapitolách plyne především již dříve předpokládaná skutečnost, že největší aktuální význam z hlediska interoperability mají geografické entity (zeměpisný rejstřík), kde lze docílit automatizovaného propojení. Dosavadní nedostatek, nedostatečná identifikace geografické entity pouze záhlavím bude ještě v databázi NAČR odstraněna uváděním hierarchického začlenění entity do okresu, kraje atd. (týká se sídel).

Naopak u entity typu osoba je propojení obtížné zejména z důvodu potenciální chybné identifikace osoby. Ukazuje se tak nezbytné uvádět bližší charakteristiku důsledně jak v archivních pomůckách, tak v databázi a používat tyto údaje k porovnání. Překvapením bylo nízké procento propojených jmen korporací. To lze vysvětlit tím, že zejména historické korporativní entity nejsou v databázi národních autorit dostatečně zastoupeny a také stejnými problémy jako u věcného rejstříku: formalizace je jiná než v databázi národních autorit, zpravidla je heslo doplněno podheslem (např. „OSN - Evropská hospodářská komise“).

2.1 Tematické hesla

Hlavním problémem skupiny tematických hesel při harmonizaci jsou rozdíly ve vytváření termínů – využívá se jednotné číslo a inverzní pořadí ve slovním spojení. Doplňky k terminu se píší za pomlčku. Také se využívají příliš úzké a specializované termíny.

Pro získání propojení alespoň na nejširší termín byl zvolen pro vyhledávání pouze hlavní termín (tj. termín po pomlčku nebo po závorku).

```
kroj – slovácký => kroj
portrét – J. Munzar => portrét
pes – Fráček => pes
```

Vyhledávaných bylo 13016 termínů – propojení se záznamem v národních autoritách bylo navrženo u 4079 termínů (potenciální úspěšnost 31%).

2.2 Jména osob

Ve skupině jmen osob byl hlavním problémem způsob zápisu různých údajů přímo do termínu pro jméno. Součástí tak byli doplňky jako „dr.“, variantní formy jména (viz). Současně byl zvolen různý způsob oddělení jména a příjmení – s čárkou nebo bez čárky. Tyto nejednoznačnosti ve zápisu si vyžádali rozsáhlejší pravidla pro stanovení termínu pro vyhledávání. Termíny neobsahovali žádné životní data, v případě frekventovaných jmen nebylo možné spolehlivě určit, o kterou autoritu jde.

Pravidla pro výběr termínů pro vyhledávání:

- v případě, že termín obsahuje slovo “viz”, vyhledává se podle části před “viz” a podle části za “viz”; část za “viz” může obsahovat dvě jména, oddělené pomlčkou – potom se vyhledává osobitně podle prvního a druhého jména – všechny termíny se upraví podle dalších pokynů,
- z termínu se vezme část po první čárku,
- v případe, že je termín před čárkou jen jedno slovo, vezme se i část po další čárku,
- v případe, že termín končí tečkou, tečku odstranit,
- v případe, že termín začíná textem “dr.” text “dr.” se odstraní.

```
Jung Josef => Jung Josef
Sadat, Anvar => Sadat Anvar
Munzar Jan, vnuk Jana Munzara staršího => Munzar Jan
Trtíková-Růžková Miluška, roz. Jánská => Trtíková-Růžková Miluška
Doležal, J. => Doležal J (použije se zkrácení)
Goll Jaroslav, dr. viz Bachmann Adolf, dr. -Braf Albin, dr. => Goll Jaroslav +
Bachmann, Adolf + Braf Albin (vyhledávají se postupně 3 termíny, výsledek se
spojí do jednoho)
Guckler Josef viz dr.Kryštůfek František => Guckler Josef + Kryštůfek František
(vyhledávají se postupně 2 termíny, výsledek se spojí do jednoho)
Hellmer Karel viz dr.Felgel Robert - Habermann Josef => Hellmer Karel + Felgel
Robert + Habermann Josef (vyhledávají se postupně 3 termíny, výsledek se spojí
do jednoho)
```

Vyhledávaných bylo 12663 termínů – propojení se záznamem v národních autoritách bylo navrženo u 5033 termínů (potenciální úspěšnost 40%).

2.3 Geografické jména

Problémem geografických termínů z rejstříků je také jejich nejednotnost zápisu a používání různých forem doplňků. Zároveň nebylo možné předpokládat zda autorita v bázi národních autorit obsahuje v záhlaví rozlišovací prvek (doplněk) v podobě okresu, obce nebo kraje. Termíny ze skupiny geografických jmen se proto vyhledávali ve více fázích od nejpřesnějšího po nejobecnější termín.

Pravidla pro výběr termínu byly stanoveny:

- v případě, že termín obsahuje text „viz“ vyhledává se termín za textem „viz“ – upravuje se podle dalších pravidel
- z termínu se vybere část po závorku
- z termínu se vybere část po první čárku – toto bude 1. varianta termínu

- v případě, že část za čárkou obsahuje text „ob.“ nebo „obec“ vybere se část za tímto textem – toto bude doplněk
- v případě, že část za čárkou obsahuje text „okr.“ nebo „o.“ vybere se část za tímto textem – toto bude doplněk, v případě, že je táto část dlouhá jen 2 znaky, nebere se do úvahy
- v případě, že část za čárkou obsahuje text „o.“ ale neexistuje žádná další část za tímto textem (čárka se jako text nepočítá) doplní se k 1. variantě text „okres“
- 1. varianta termínu se spojí s doplňkem (podle prvého nebo podle druhého pravidla) v tvare do podoby „1. varianta doplněk“ – vznikne 2. varianta resp. 3. varianta

```
Běšiny, o. Klatovy => Běšiny + Běšiny Klatovy
Bílá Voda, ob. Červená Voda, o. Ústí nad Orlicí => Bílá Voda + Bílá Voda Červená
Voda + Bílá Voda Ústí nad Orlicí
Ostrava, o., => Ostrava okres (vyhľadáva sa bez atribútu pre frázu)
kraj severomoravský => kraj severomoravský (vyhľadáva sa bez atribútu pre frázu)
Západní Evropa viz Evropa západní => Evropa západní
Zalitavsko (Translajtánie) => Zalitavsko
Brtníky, okr. DC => Brtníky
```

Vyhledávaných bylo 24331 termínů – propojení se záznamem v národních autoritách bylo navrženo u 17149 termínů (potenciální úspěšnost 70%).

2.4 Korporace

Skupina s termíny korporací byla tvořena dvěma soubory a obsahovala tak různorodé termíny, že nebylo možné stanovit obecnější pravidla pro vyhledávání – z termínu se pouze odstranili doplňky za čárkou a v závorce.

```
TISKAŘSKÉ A VYDAVATELSKÉ DRUŽSTVO ROLNICKÉ V PRAZE, ZAPSANÉ SPOLEČENSTVO S
RUČENÍM OBMEZENÝM => TISKAŘSKÉ A VYDAVATELSKÉ DRUŽSTVO ROLNICKÉ V PRAZE
Spolek českých mediků (Praha) => Spolek českých mediků
Omladina (Čechy) => Omladina
```

Vyhledávaných bylo v prvním souboru 10900 termínů – propojení se záznamem v národních autoritách bylo navrženo u 171 termínů (potenciální úspěšnost 1.5%), ve druhém souboru 1885 termínů – propojení se záznamem v národních autoritách bylo navrženo u 268 termínů (potenciální úspěšnost 14%).

3 Příprava metodických a metodologických postupů, jejich zavedení v archivech

Rok 2012 byl rokem novelizace právních předpisů v oblasti archivnictví a spisové služby. Hlavním důvodem změn (kromě opravy chyb) bylo nastavení pravidel pro elektronickou archivaci v návaznosti na projekt národního digitálního archivu. Postup projektu INTERPI se přímo promítal do následujících činností, které INTERPI zpětně ovlivňovaly.

3.1 Zadávací dokumentace technologie národního digitálního archivu

Jedním z úkolů celostátního archivního portálu je správa informací o původcích (entita typu osoba nebo korporace). Proto bylo již při tvorbě zadávací dokumentace zajištěno, aby výsledky projektu INTERPI byly v oblasti archivů aplikovány (že tomu tak není v případě projektu národní digitální knihovny, lze litovat):

*„Archivní portál umožní přijetí (prostřednictvím webového rozhraní nebo webovou službou) dotaz obsahující a) jméno nebo část jména, b) IČO, c) ID autoritního záznamu. Dotaz portál postoupí prostřednictvím webové služby znalostní databázi paměťových institucí a jí vrácený výsledek znázorní v definované podobě nebo zpřístupní prostřednictvím webové služby IS příslušného archivu. Struktura dat a webové služby znalostní databáze paměťových institucí jsou předmětem projektu výzkumu NAKI č. DF11P01OVV023 (*http://autority.nkp.cz/interpi*).*

Archivní portál umožní zpracovávat (vytvářet, modifikovat, znázorňovat) geografický, jmenný nebo věcný autoritní záznam ve znalostní databázi paměťových institucí a jeho zpřístupnění webovým rozhraním nebo webovou službou dle A.1.25 až A.1.28 obdobně.“[4]

3.2 Novela zákona 499/2004 Sb., o archivnictví a spisové službě ve znění pozdějších předpisů

Do zákona byla při přípravě novely včleněna ustanovení o evidenci původců, z nichž nejvýznamnější je § 18c odst. 3, který kodifikuje povinný popis původců v elektronické podobě jednotlivými archivy. „Popis a evidence původců“ je postavena na roveň (paralelně vedle ní) dosavadní evidenci Národního archivního dědictví. Při aplikaci tohoto ustanovení je zamýšleno užívání databáze INTERPI, pro niž komunikaci bude zprostředkovávat celostátní archivní portál, jak je zřejmé z citovaných ustanovení zadávací dokumentace národního digitálního archivu.

[4] Požadavek A.1.30 a A.1.31 Zadávací dokumentace k zakázce „NDA, zadávací řízení na dodavatele technologií ICT a implementaci a vývoj SW“. Národní archiv, cit. [30.10.2012]. Dostupné z WWW <https://web.nacr.cz/zakazky/NDA_projekt_ISNDA/dokumenty/NDA-VZ-na-dodavatele-IS-NDA_v36.pdf>

3.3 Novela vyhlášky 645/2004 Sb., kterou se provádí zákon o archivnictví a spisové službě

Novelizovaná vyhláška uvádí podrobnosti k zákonem stanoveným povinnostem. Příprava těchto ustanovení byla již plně navázána na práce v rámci projektu INTERPI. V souladu s legislativními pravidly, zvyklostmi a postupy nebylo však možné některé pojmy vhodně upravit. Jedná se zejména o následující problémy:

1. U osob při popisu původců vypuštěn na zásah ve vnitrorezortním připomínkovém řízení „životopis“, který ovšem zůstal jako povinná součást úvodu inventáře
2. nebylo možné použít pojem „rod“ (nežádoucí vazba na občanský zákoník a zákon o rodině) ani „korporace“
3. při legislativně-technické úpravě došlo k dodefinování „sídla“ na rozsah celé adresy, i když to nebylo původním úmyslem

Přes zmíněné nedostatky se podařilo prosadit úpravu, která zaručuje řádné pořízení informací o původcích a je kompatibilní s INTERPI. Zcela novým způsobem byly definovány částečně korporace, přičemž byl opuštěn v dosavadní archivní legislativě přítomný koncept pouze právnických osob: „u archiválií, jejichž původcem je skupina osob vystupujících pod stejným názvem a sdružených k dosažení shodného účelu“. Důvodová zpráva k předmětnému § 12b tyto skutečnosti reflektuje, stejně tak databázi INTERPI:

„Koncepce evidence původců je zvolena tak, aby se tato evidence stala informační databází o původcích archiválií, a to v rozsahu umožňujícím jednak vyhledávání s využitím odkazů (rozšířené vyhledávání na základě poskytnuté dílčí nebo velmi omezené jedinečné informace), jednak umožňující poskytování údajů o těchto původcích. Původci jsou pro účely stanovení údajů a skutečností vedených v evidenci rozčleněni podle dlouhodobých zkušeností archivů tak, aby nebyla opomenuta žádná z existujících možností, a aby tedy byly evidenčně postižitelné záznamy o veškerých existujících původcích archiválií v digitální podobě. Pro původce jsou tudíž v jednotlivých kategoriích stanoveny údaje pro „fyzické osoby“, právnické osoby (včetně zájmových sdružení osob ve smyslu občanského zákoníku) a skupiny osob vystupujících pod stejným názvem a sdružených k dosažení shodného účelu. Poslední uvedená kategorie je historicky poměrně četným a z hlediska významu archiválií důležitým původcem, i když nemá postavení právnické osoby ve smyslu občanského zákoníku. Jedná se o skupiny fyzických, právnických nebo fyzických a právnických osob, které se na základě dobrovolnosti sdružily k dosažení určitého cíle a které přitom nejsou založeny zákonem stanoveným způsobem, resp. pro svou činnost nesplňují zpravidla žádné vnější formální znaky [zejména stolní společnosti – např. Dobročinná stolní společnost „Žebrácká elita“ Písek 1929-1941 (Státní okresní archiv Písek, NAD 1721), Pionýrská skupina 1. sjezdu SSM Chrudim při ZŠ Olbrachtova Chrudim 1983-1990 (Státní okresní archiv Chrudim, NAD 2096), sjezdy – např. Sjezd abiturientek bývalého dívčího reálného gymnasia Brno 1916-1970 (Archiv města Brna, NAD 790)].“

3.4 Návrh nových Základních pravidel pro zpracování archivního materiálu

Nová Základní pravidla pro zpracování archiválií fakticky znamenají po 50 letech převrat v metodice základních archivních činností s výjimkou předarchivní péče. Jejich koncepce vychází, jak bylo naznačeno v loňské průběžné zprávě, z mezinárodních archivních standardů, především ISAD(G) a ISAAR(CPF). Právě druhý jmenovaný standard se stal základem pro popis původců archivních fondů, který se promítl i do legislativy (viz výše). Zároveň musela být stanovena pravidla pro tvorbu názvů – a to nejen v případě základních entit.

Práce na Základních pravidlech pro zpracování archiválií byly v případě kapitol „popis původců“ a „tvorba názvů a jmen“ koordinovány a konzultovány i v rámci projektu INTERPI. Tyto konzultace významně přispěly k definici entit (nejen základních) a také k zamyšlení nad udržitelností či neudržitelností některých metodických postulátů ať již v oblasti archivnictví nebo knihovnictví.

Protože se jedná o základní metodiku pro archivy, byla její přípravě věnována velká pozornost. Pravděpodobně v prosinci bude celý materiál zveřejněn na webu Ministerstva vnitra k odborné diskusi.