

ZPRÁVA O ČINNOSTI NÁRODNÍ KNIHOVNY V PROJEKTU INTERPI ZA ROK 2014

INTERPI – Interoperabilita v paměťových institucích

Program aplikovaného výzkumu a vývoje národní a
kulturní identity (NAKI)
(DF11P010VV023)

Sestavila:
Marie Balíková



INTERPI

13. 11. 2014

Obsah

1	Analýza, ověřování, doplňování a úprava dat	3
1.1	Aplikace souboru národních autorit jako terminologického základu znalostní báze INTERPI	3
1.1.1	Soubor autorit v prostředí paměťových institucí	3
1.1.2	Dílčí autoritní soubory	4
1.2	Soubor tematických autorit	4
1.2.1	Struktura autoritního záznamu	5
1.2.2	Termín - lexikální jednotka	5
1.2.3	Definice termínu	5
1.2.4	Podoba termínu	6
1.2.5	Význam termínů, vztahy mezi tematickými termíny	7
1.3	Aplikace souboru národních autorit v procesu mapování/harmonizace terminologických zdrojů	8
1.4	Role Souboru věcných tematických autorit v procesu mapování/harmonizace terminologických zdrojů	8
2	Úprava autoritních termínů	10
2.1	Úprava souboru tematických autorit	10
2.1.1	Invertované tvary víceslovných termínů	10
2.1.2	Tvorba souboru kategorie osob	11
2.2	Úprava souboru geografických autorit	11
2.3	Aktualizace klasifikačního systému MDT	12
2.4	Příprava záznamů Národních autorit, export pro transformaci dat	14
3	Ověřování postupů mapování a harmonizace terminologických zdrojů	15
3.1	Vybrané tezaury	15
3.2	Proces porovnávání	16
3.2.1	Metody	16
3.2.2	Forma výstupu	16
3.3	Výsledky	17
3.4	Zhodnocení	18

3.4.1	Možnosti zlepšení	21
4	Problematika zámeckých knihoven	24

1 Analýza, ověřování, doplňování a úprava dat

V roce 2014 NKČR pokračovala v pracích spojených s analýzou, exportem a úpravu autoritních údajů jako základního předpokladu pro úspěšný vývoj znalostní databáze INTERPI. Byla vypracována studie o možnosti aplikace souboru národních autorit jako terminologického základu znalostní báze INTERPI. Zvláštní pozornost byla věnována analýze a úpravě tematických autorit a autoritních termínů označujících formu/žánr.

1.1 Aplikace souboru národních autorit jako terminologického základu znalostní báze INTERPI

Knihovnické obci slouží soubor národních autorit jako prostředek k unifikaci selekčních prvků bibliografických záznamů a jako takový má svou vnitřní strukturu a určité množství konkrétních informací odpovídající katalogizačním pravidlům AACR2, od roku 2015 katalogizačním pravidlům RDA.

1.1.1 Soubor autorit v prostředí paměťových institucí

Souborem národních autorit se tedy rozumí soubor autoritních hesel a s tím svázaný a propojený soubor odkazových hesel a poznámek. V kooperaci s paměťovými institucemi byla původní definice upravena na „Soubor národních autorit představuje soubor ověřených a unifikovaných jmenných a/nebo věcných selekčních údajů určených pro zpracování a vyhledávání dokumentů/informačních zdrojů s potřebným odkazovým a poznámkovým aparátem“. K této úpravě došlo v počáteční fázi kooperace mezi paměťovými institucemi, která měla spočívat především v důsledné aplikaci souboru národních autorit ve všech paměťových institucích. Stávající soubor národních autorit malým rozsahem doprovodných informací, tj. počtem variantních forem jména, rozsahem poznámek, tj. definic, biografických údajů však plně nepokrýval potřeby archivů, muzeí a galerií.

V další fázi kooperace, která se nyní odehrává na bázi konceptuálních modelů, se ukázalo, že pojem autorita, často navíc v odborné literatuře zužovaný pouze na pojem autoritní záhlaví, nezohledňuje všechny aspekty nutné pro dosažení interoperability v rámci paměťových institucí, proto se v současné době v prostředí paměťových institucí za základní pojem považuje entita a autoritní záhlaví slouží jako jeden z atributů entity, tj. její označení. Obsažný poznámkový aparát, o který by měly být záznamy souboru národních autorit rozšířeny, pak bude sloužit jako nástroj potřebné jednoznačné identifikace dané entity a podklad pro explicitní vyjádření sémantických vztahů a vazeb, do kterých daná entita vstupuje.

Explicitní vyjádření sémantických vztahů a vazeb potřebné pro strojové zpracování dat autoritní MARC záznamy plně nepodporují, protože vztahy a vazby nezbytné pro tvorbu ontologií/znalostníchází a také pro úspěšné mapování a efektivní harmonizaci řízených

slovníků používaných v paměťových institucích jsou zde vyjádřeny implicitně a ve většině případů není jejich význam pro internetové vyhledávače plně srozumitelný a uchopitelný. Nicméně soubory národních autorit, tím že pracují s jednoznačným označením entit a na toto jednoznačné označení navázanými početnými informačními zdroji (tištěnými a elektronickými dokumenty), představují pro tvorbu ontologickýchází dobré východisko. Nejinak je tomu i v případě souboru NAČR.

1.1.2 Dílčí autoritní soubory

Soubor národních autorit se skládá z dílčích autoritních souborů:

- Souboru personálních autorit, který slouží jako zdroj ověřených a unifikovaných záhlaví osobních jmen, jmen rodin a rodů a záhlaví typu autor/název;
- Souboru korporativních autorit, který slouží jako zdroj ověřených a unifikovaných záhlaví jmen korporací, institucí a akcí;
- Souboru názvových autorit, tj. autorit pro unifikované názvy a autorit typu autor/název; slouží k propojení záznamů pro dílo, které se vyskytuje v různých podobách a pod různými názvy nebo k rozlišení více děl s totožnými hlavními názvy;
- Souboru geografických autorit, který slouží jako zdroj ověřených a unifikovaných záhlaví geografických názvů;
- Souboru tematických autorit (dále také SVAT), který slouží jako zdroj ověřených a unifikovaných záhlaví tematických termínů vyjadřujících abstraktní entity (vlast, láska, radost), vědní a umělecké obory (matematika, malířství), objekty (hrady, požární technika), ale i specifické entity označující události nebo akce, které ovšem nejsou součástí korporativního záhlaví (staroměstská exekuce, 1621);
- Souboru chronologických autorit;
- Souboru formálních autorit, který se skládá z autoritních záznamů formálních deskriptorů a lze jej charakterizovat jako zdroj ověřených a unifikovaných záhlaví označujících žánry a formy dokumentů.

Z hlediska katalogizační praxe představují dílčí soubory autorit důležitý nástroj pro standardizaci údajů, z uživatelského hlediska zjednodušují a urychlují vyhledávání dokumentů a informací v nich obsažených.

1.2 Soubor tematických autorit

Soubor tematických věcných autorit je řízený a měnitelný abecedně uspořádaný soubor ověřených a unifikovaných věcných lexikálních jednotek selekčního jazyka (věcných selekčních údajů, indexačních termínů), mezi nimiž je možné definovat základní sémantické vztahy (ekvivalence, hierarchie, asociace). Na rozdíl od oborových tezurů je univerzální, pokrývá všechny vědní oblasti a obory, byť nerovnoměrně. Vzniká totiž induktivní metodou, tedy metodou zdola podle potřeb, které se vyskytnou při věcném zpřístupnění převážně tištěných dokumentů.

1.2.1 Struktura autoritního záznamu

Soubor se skládá z jednotlivých věcných autoritních záznamů, přičemž součástí každého takového záznamu jsou údaje:

- záhlaví (povinný údaj),
- variantní forma pro označení záhlaví (povinný údaj, existuje-li),
- nadřazený, podřazený, asociovaný pojem (povinný údaj, existuje-li),
- poznámkový aparát obsahující definici pojmu, pokyny o použití a jiné informace (v budoucnu povinný údaj, v současné době se údaje do autoritních záznamů doplňují),
- notační znak systematického selekčního jazyka související s autoritním záhlavím (povinný údaj),
- anglický ekvivalent (povinný údaj),
- údaj skupiny Konspektu (nepovinný údaj).

1.2.2 Termín - lexikální jednotka

Základním prvkem souboru je termín/lexikální jednotka, která bývá v odborné literatuře definována jako slovní vyjádření určitého pojmu, pokud možno ve formě substantiva nebo substantivního spojení. Termín, který označuje pojem, se v souboru autorit vyskytuje pouze jednou, což znamená, že označení pojmu je jedinečné. Termín, který není jednoznačný, být specifikován, rozšířen o upřesňující výraz, kterým může být přídavné jméno (vojenské operace) nebo výrazem - kvalifikátorem nebo relátorem, který se uvádí v závorce a je nedílnou součástí označení, např. intelligence (vrstva), intelligence (schopnost).

Specifikem SVAT (s ohledem na využívaný informační systém pro jeho tvorbu a správu) je, že požadavek na jednoznačnost platí pro preferovanou i variantní formu jména. Je-li třeba jeden termín odkázat na více preferovaných forem, k jednoznačnému rozlišení těchto variantních forem slouží ID záznamu, např. politické časopisy jsou odkázány na dva preferované termíny: časopisy politických stran a hnutí a společensko-politické časopisy. Tento princip nebude nutné aplikovat v znalostní databázi INTERPI.

V záznamu tematické autority existuje pouze jedna preferovaná forma jména, která se používá závazně při indexování k vyjádření určitého pojmu. Ekvivalent nebo kvaziekvivalent tohoto termínu se pak stává termínem nepreferovaným. Tento není dokumentu přiřazován, ale slouží jako uživatelský vstup do abecedního rejstříku, přičemž uživatel je odkázán na hlavní preferovaný termín (většinou je odkaz vytvořen přidáním zkratky viz.).

1.2.3 Definice termínu

Definice neboli vymezení rozsahu preferovaného termínu je důležitou součástí záznamu každé tematické autority univerzálního systému, neboť v takovém systému se často vyskytují výrazy,

kteřé se používají ve více oborech a odvětvích lidské činnosti současně; je proto třeba každý termín přesně specifikovat a definovat. Při tvorbě univerzálního souboru musí být zvláštní pozornost věnována obsahové stránce tematických lexikálních jednotek a integraci víceslovných spojení do souboru autoritních termínů.

1.2.4 Podoba termínu

Jednoslovné substantivum

Ideálním reprezentantem pojmu je jednoslovné substantivum v 1. pádě jednotného čísla, např. škola. S tímto ideálním stavem se však setkáváme zřídka, navíc v knihovnických systémech se tradičně používají deskriptory i nedeskriptory v množném čísle, tedy školy (viz níže).

Víceslovné termíny

Častěji se k označení pojmu používají víceslovné termíny, jejichž základ tvoří substantivum, a to buď adjektivní spojení (terminologické databáze), genitivní vazba (management kultury), nebo předložková vazba (teorie odpovědi na položku).

Komplexní termíny

V souboru tematických autorit se používají i komplexní termíny, tj. dvě souřadně spojená substantiva označující komplexní pojem, resp. vztah těchto pojmů, jejichž rozklad by vedl k vyhledání množství irelevantních dokumentů, např. rodiče a děti.

Pravopisná norma, psaní velkých a malých písmen, transliterace

Dodržuje se pravopisná norma, v případě existence pravopisných dublet se preferuje progresivní podoba. Při zápisu velkých a malých písmen se respektují pravidla českého pravopisu, obecné termíny, resp. vstupní část jejich záhlaví, se zapisují malými písmeny. Dodržuje se transliterace podle platných norem.

Cizojazyčné výrazy

Cizojazyčné výrazy je možné použít v případě, neexistuje-li adekvátní překladový výraz (benchmarking) nebo je-li termín běžně používán v daném oboru lidské činnosti nebo vědním oboru (diabetes mellitus).

Singulár a plurál

Počítatelná substantiva se uvádějí v plurálu (učitelé), abstrakta, počítatelná substantiva použitá jako abstrakta a názvy vědních oborů se zapisují v singuláru (vlast, pohyb (filozofie), matematika, kresba).

Slovosled

Dodržuje se přirozený slovosled, tedy pořadí adjektiva a substantiva, postponované přívlastky v odborném názvosloví se respektují (kyselina dusičná). Invertované tvary respektující prioritu substantiva jsou postupně do souboru doplňovány tak, aby byla usnadněna harmonizace rejstříků a heslářů používaných v paměťových institucích.

Zkratky, zkratková slova

U zkratk se preferuje plná rozepsaná podoba, zkrácená forma se odkazuje (DTP viz desktop publishing, DDC viz Deweyho desetinné třídění); běžně ustálené a všeobecně srozumitelné termíny se naopak preferují (syndrom získané imunodeficiency viz AIDS).

1.2.5 Význam termínů, vztahy mezi tematickými termíny

Významová nejednoznačnost označení pojmů, tedy termínů představuje obtížně řešitelný problém při tvorbě řízených slovníků všech typů, nejinak je tomu i v případě souboru tematických autorit. Synonymie se řeší vylučovacím odkazovým aparátem (abraziva viz brusiva), homonymie se řeší specifikací termínu pomocí vhodného přívlastku (morfologie rostlin, vysokoškolské koleje) nebo uvedením kvalifikátoru (relátoru) v závorce, přičemž závorkové doplnění je součástí deskriptoru (liška (houba), liška (savec)).

Významová stránka používaných termínů ovlivňuje také jejich sémantické vztahy. Mezi jednotlivými lexikálními jednotkami soboru tematických autorit lze definovat sémantické vztahy: vztah ekvivalence, vztah hierarchie a vztah asociace.

Vztah ekvivalence je jedním ze základních předpokladů fungování řízeného slovníku. Na jeho základě se rozlišuje preferovaná a variantní forma termínu, tedy deskriptor a nedeskriptor.

Jedná se o synonymní jednotky, jejichž obsah je identický (označují stejný denotát²¹), liší se ovšem formou. První termín se závazně užívá při indexování, druhý není dokumentu přiřazován, ale slouží jako uživatelský vstup do abecedního rejstříku v podobě vylučovacího odkazu na termín preferovaný.

Hierarchický vztah nastává mezi lexikálními jednotkami téhož sémantického okruhu, kdy jeden pojem je podřazený druhému pojmu a to tehdy, jestliže k jeho identifikaci je třeba použít všechny znaky nutné k identifikaci pojmu nadřazeného, přičemž podřazený termín má minimálně o jeden znak, kterým se liší od nadřazeného termínu, více.

Vztah asociace vzniká mezi lexikálními jednotkami, které spolu významově souvisejí, avšak jejich vzájemný vztah není možno považovat za hierarchický.

1.3 Aplikace souboru národních autorit v procesu mapování/harmonizace terminologických zdrojů

Jak jsme již uvedli výše, představuje soubor národních autorit terminologický základ pro mapování a harmonizaci slovníků, které se mají podílet na znalostní bázi systému INTERPI.

Všechny složky souboru autorit, tedy jednotlivé dílčí soubory, se budou podílet na harmonizaci rejstříků, které budou obsahovat označení všech entit, tj. běžně používané typologicky nerozlišené indexy klíčových slov či jednotlivých segmentů řetězců předmětových hesel.

Filtrací těchto rejstříků obsahujících všechny termíny pomocí dílčích souborů NAČR získáme oddělené soubory potenciálních typologicky rozlišených deskriptorů: soubory personálních jmen, korporativních, soubor geografických termínů, obecných pojmů, atd.

Ještě výrazněji se uplatní soubor tematických autorit ať už při porovnávání řízených slovníků na všech jejich úrovních (lexikální, na bázi konceptu nebo na bázi instance) nebo při tvorbě řízeného slovníku/tezauru pro jednotlivé typy paměťových institucí na bázi rejstříku jimi používaných klíčových slov.

1.4 Role Souboru věcných tematických autorit v procesu mapování/harmonizace terminologických zdrojů

1. Soubor věcných tematických autorit je univerzální - jsou v něm zahrnuty termíny, případně terminologické soustavy ze všech oborů lidského poznání. Jednotlivé termíny jsou provázány s ekvivalentní notací MDT, která umožní expanzi do jednotlivých sémantických domén a zároveň usnadní tvorbu mikrotezaurů v jednotlivých vědních oborech;
2. Soubor věcných tematických autorit obsahuje aktuální terminologii; vytváří se induktivní metodou (metodou zdola), jednotlivé termíny jsou do něj vřazovány podle potřeb zpracovatele. Na jedné straně to představuje úskalí, protože SVAT neobsahuje všechny termíny jednotlivých terminologických soustav všech oborů. Na druhé straně to přináší velké výhody. Do souboru tematických autorit jsou neustále vřazovány jako preferované či nepreferované formy termínů současně používané termíny. Můžeme tedy konstatovat, že terminologie používaná v rámci SVAT je aktuální;
3. Soubor věcných tematických autorit používá víceslovné deskriptory - v praxi to znamená, že jsou používány termíny, které jsou specifické a detailní do té míry, že reprezentují daný (nově se objevivší) pojem adekvátním způsobem a zasazují ho tak do sémantického rámce daného vědního oboru;
4. Autoritní termíny v souboru tematických autorit jsou propojeny s ekvivalentní notací MDT, což v praxi umožní zařazení i specifických termínů používaných jednotlivými paměťovými institucemi, které tak mohou být federativní metodou navázány na obecnější termín, který tvoří součást univerzálního souboru;

5. Autoritní termíny jsou opatřeny anglickými ekvivalenty, což v praxi podpoří aplikaci tohoto souboru na mezinárodní úrovni.

2 Úprava autoritních termínů

2.1 Úprava souboru tematických autorit

V roce 2014 byly provedeny následující úpravy v souboru tematických autorit: do autoritních záznamů byly připojeny invertované tvary víceslovných termínů. Byl také vytvořen soubor obsahující označení kategorie osob.

2.1.1 Invertované tvary víceslovných termínů

Cílem této zásadní úpravy v souboru tematických autorit, tj. připojení invertovaných tvarů (odkaz typu viz) u víceslovných obecných termínů je napomoci orientaci uživatele při vstupu do systému INTERPI pomocí přístupových rejstříků a také usnadnit postup při harmonizaci/mapování řízených slovníků.

AUT - Úplné zobrazení záznamu

Zvolte formát: [Standardní](#) -- [MARC](#)

Záznam 1 z 1

Ident. číslo	ph303225
Záhlaví	 srovnávací metody
Odkaz. forma	vědecká komparace
	komparativní metody
	komparativní analýza
	srovnávací analýza
	metody srovnávací
	metody komparativní
	analýza komparativní
	analýza srovnávací
	komparace vědecká
Nadřazený termín	vědecká metodologie
MDT	 303.446
Angl. ekvivalent	comparative methods
Skupina Konspektu	 3 - Společenské vědy [18]

Bylo upraveno 15641 tematických termínů.

2.1.2 Tvorba souboru kategorie osob

Byl vytvořen soubor termínů obsahující kategorie/skupiny nepojmenovaných osob označených obecným názvem (ne jednotkových tříd vyjádřených vlastními jmény), které se v souboru autorit a v systému INTERPI uvádějí v plurálu. Forma v singuláru se uvádí jako variantní/alternativní forma a v systému INTERPI bude sloužit jako podklad pro tvorbu stručné charakteristiky konkrétní pojmenované osoby.

ph114921	knihaři	knihař
ph114923	knihkupci	knihkupec
ph114943	kojenci	kojenec
ph114944	zdravotně postižení umělci	zdravotně postižený umělec
ph114990	kosmonauti	kosmonaut
ph115059	íránisté	íránista
ph115081	křesťanští mystici	křesťanský mystik
ph115131	letečtí konstruktéři	letecký konstruktér
ph115154	lingvisté	lingvista
ph115219	manažerky	manažerka
ph115262	mentálně postižené děti	mentálně postižené dítě
ph115263	mentálně postižení	mentálně postižený
ph115333	mládež	mládež
ph115388	mystici	mystik
ph115403	měšťanstvo	měšťan
ph115418	nakladatelé a vydavatelé	nakladatel a vydavatel

Takto bylo upraveno 1800 autoritních termínů.

2.2 Úprava souboru geografických autorit

V souboru geografických autorit bylo v roce 2014 dořešeno propojení této databáze s Databází sídelních lokalit Čech, Moravy a Slezska cz_retro pomocí pole 052, kde se uvádí kód cz_retro. Byla vytvořena aplikace, která automaticky doplní kód cz_retro v poli 052 a propojí tak GEO termíny s bází cz_retro s pomocí hierarchické struktury uvedené v poli 951; takto lze postupovat v případech jednoznačných termínů. V záznamech geografických autorit, ve kterých nedošlo - z důvodů rozdílnosti údajů - k automatickému propojení na bázi údajů v poli 951, se postupně doplňuje pole 052 intelektuálně. Pro tyto případy byla vytvořena aplikace, která umožní zobrazit jednotlivé záznamy v bázi cz_retro přehledným způsobem, např.

**Databáze sídelních lokalit Čech, Moravy a Slezska
(CZ_RETRO)**

Základní vyhledávání

Lokalita Okres / Region Kraj

Rejstřík lokalit

© Ing. arch. Karel Kuřa (spolupráce Ing. Štěpán Mležíva)
 Společnost pro obnovu vesnice a malého města, 2013

Pro účely snadného vyhledávání a celkové uživatelsky přívětivější orientace je databáze cz-retro nyní zpřístupněna na adrese http://aleph.nkp.cz/web/cz_retro.htm, např.:

(CZ_RETRO)

Rejstřík lokalit

#	Lokalita / Region	Okres	Kraj
1	Loučka / Lipník nad Bečvou (ves)	Přerov	Olomoucký kraj
2	Loučka / Litovel (ves)	Olomouc	Olomoucký kraj
3	Loučka / Lovosice (býv. městská čtvrť)	Litoměřice	Ústecký kraj
4	Loučka / Nový Jičín (ves aglomerovaná)	Nový Jičín	Moravskoslezský kraj
5	Loučka / Třinec (vesnická čtvrť)	Frýdek-Místek	Moravskoslezský kraj
6	Loučka / Třinec (vesnická čtvrť)	Frýdek-Místek	Moravskoslezský kraj
7	Loučka / Valašské Klobouky (ves)	Zlín	Zlínský kraj
8	Loučka / Valašské Klobouky (hradiště)	Zlín	Zlínský kraj
9	Loučka / Valašské Meziříčí (ves)	Vsetín	Zlínský kraj
10	Loučka / Valašské Meziříčí (hradiště)	Vsetín	Zlínský kraj
11	Loučka / Židlochovice (ves)	Brno-venkov	Jihomoravský kraj
12	Loučka I / Valašské Klobouky (umělé kú)	Zlín	Zlínský kraj
13	Loučka II / Luhačovice (umělé kú)	Zlín	Zlínský kraj
14	Loučky / Benešov nad Ploučnicí (celek hist. osad)	Děčín	Ústecký kraj
15	Loučky / Bezručovice (víska)	Tachov	Plzeňský kraj
16	Loučky / Choceň (ves)	Ústí nad Orlicí	Pardubický kraj
17	Loučky / Jihlava (ves)	Jihlava	Kraj Vysočina

2.3 Aktualizace klasifikačního systému MDT

V tomto roce (2014) řešení projektu jsme pokračovali v teoretickém průzkumu o aplikaci klasifikačního systému MDT jako potenciálního propojovacího jazyka při řešení terminologické různorodosti termínů aplikovaných v jednotlivých paměťových institucích a jako nástroje pro tvorbu ontologií.

V roce 2014 pokračovaly práce na vývoji vícejazyčné překladové databáze MDT - UDC online.

České MDT Online obsahuje úplnou, standardní verzi MDT zahrnující 70 626 platných a přibližně 11 000 zrušených notací MDT. Více než 70 000 znaků je opatřeno českým překladem, zbytek textu je uveden v angličtině. Překlad bude do systému postupně doplněn. Tento nástroj má několik dalších funkcí umožňujících vyhledávání, prohlížení, analýzu, validaci a tvorbu notací MDT. Vyhledávání a zobrazení zrušených notací MDT pomáhá uživatelům dohledat zrušená čísla a jejich přeměrování, protože se mohou stále ještě vyskytovat při zpřístupnění knihovních fondů.

České MDT Online vzniká díky spolupráci mezi Konsorciem MDT a Národní knihovnou ČR.

Tento produkt nahrazuje předchozí MDT UDC Online (čtěte tiskovou zprávu).

Obr. UDC_1

V souvislosti s tím pokračují práce i na překladu a aktualizaci systému MDT, který nyní obsahuje 70626 znaků. Česká verze obsahuje nyní 64 899 přeložených znaků, zbývajících 5727 je přeloženo, po posouzení budou připojeny do vícejazyčné překladové databáze.

Universal Decimal Classification

MRF translator

STATISTICS EXPORTS ACTIVITY COMMENTS LOGOUT

Notation: 001.98 Search ADVANCED

Caption: Search

Browse: [v] Show unfinished only Show MRF2011 changes

001.98 Falsifikace ve vědě

001.99 Obstrukce a překážky pro šíření myšlenek

002 Dokumentace. Knihy. Písemnosti. Autorská činnost

002.1 Dokumenty (nosiče informací) jako předmět

002.2 Literární produkce jako celek

002.6 Dokumentační útvary a střediska

002.61 Mezinárodní dokumentační útvary a střediska

002.63 Národní dokumentační útvary a střediska

002.66 Dokumentační střediska firem, institucí, podniků, úřadů

003 Systémy psaní a písma

Source Language: English [en] Target Language: Czech [cs]

SEARCH: 001.98 Search Edit

CAPTION: Falsification of science

Falzifikace ve vědě

INCLUDING: False doctrines, fallacies (due to insufficient knowledge, bigotry, excessive zeal etc.)

Klamné závěry, omyly ve vědě (z důvodu nedostatečných znalostí, zaslepenosti, nadměrné horlivosti atd.)

Translator's Notes

Obr. UDC_2

Všechny práce spojené s překladem a aktualizací jednotlivých verzí klasifikačního systému MDT a s publikací klasifikačního systému MDT probíhají v úzké kooperaci s UDC Konsorciem v Haagu, které vytváří prostředí pro jejich realizaci, např. vývoj, hosting a údržbu multilingvární překladové databáze (pro různé verze vydání MDT - UDC Summary, UDC pocket edition, UDC MRF), porovnání a propojení různých verzí národního souboru klasifikačních znaků MDT - v našem případě porovnání verzí z let 2002 a 2003 s verzemi z let 2010 a 2012, publikaci klasifikačního systému jako linked open data (LOD) apod.

Česká verze MDT online bude zpřístupněna v prosinci 2014.

2.4 Příprava záznamů Národních autorit, export pro transformaci dat

Podle kritérií definovaných v roce 2013, na základě kterých byly exportovány záznamy Národních autorit pro testování transformace dat, byl i v této etapě vytvořen export stávajících autorit.

Bylo exportováno:

- 589764 personálních autorit označujících osoby,
- 14635 autorit označujících kombinace jména autora a název jeho díla,
- 94782 korporativních záhlaví,
- 3489 kombinací korporativních záhlaví a názvů díla,
- 26250 geografických autorit,
- 37132 autorit akcí,
- 2339 autorit unifikovaných názvů.

3 Ověřování postupů mapování a harmonizace terminologických zdrojů

Přístupy, které jsme uvedli v metodice pro automatizaci procesů harmonizace, jsme se snažili ověřit v praxi prostřednictvím desktopové aplikace s využitím relačního databázového systému. V této studii uvádíme prvotní výsledky průzkumu pro ověření základních algoritmů podobnosti. Ověření jsme provedli pomocí transformace získaných korpusů číselníků, slovníků a tezaurů používaných v našem prostředí pro zpracování objektů kulturního dědictví. Následně jsme vytvořili prototypovou aplikaci pro porovnání těchto zdrojů se souborem věcných autorit. Samotné srovnání bylo z časového hlediska realizováno na lexikální úrovni. Výsledek tohoto porovnání, což je patrné i v příloze, potvrzuje skutečnosti, které jsme uvedli v metodice a které se týkaly předzpracování vstupních dat za účelem zvýšení automatizovaného způsobu harmonizace.

3.1 Vybrané tezaury

Pro účely porovnání dat jsme zvolili data z těchto systémů:

1. **Autority** - byly brány jako základ pro srovnávání a s těmito daty jsme srovnávali data ze všech níže uvedených zdrojů. Ze souboru Národních autorit jsme vybrali, jak jsme již uvedli, jenom tematické termíny. Data autorit sestávala ze samotného unifikovaného termínu a všech termínů, které k němu mají nějaký vztah. Brali jsme termíny, které jsou v záznamech autorit uvedeny v polích MARC21 1XX, 4XX, 5XX.
2. **DEMUS** - tento systém je poměrně značně zastoupen v oblasti muzeí. Jednalo se o 2 tezaury a 4 číselníky:
 - Demus_VUM_materiály,
 - Demus_VUM_specifikace_materiálu,
 - Demus_VUM_specifikace_techiky
 - Demus_VUM_techiky.
 Nadřazený termín obsahovalo 1790 termínů. Všechny termíny obsahovaly také anglický a německý překlad.
3. **LANius, Clavius** - systémy zastoupeny především ve veřejných knihovnách, kde používají jednoduchá klíčová slova bez jakýchkoliv hierarchických vazeb.
6. **PSH - Polytematický strukturovaný heslář**, který spravuje Národní technická knihovna a je používán i jinými organizacemi v různých systémech. Data z tohoto hesláře jsme získali ve formě RDF SKOS a pro ověření byla použita tato data (každý termín obsahuje jednoznačný číselný aj URL identifikátor):

```
skos:prefLabel xml:lang="cs"
skos:prefLabel xml:lang="en"
skos:altLabel xml:lang="cs"
skos:altLabel xml:lang="en"
skos:narrower
skos:broader
skos:related
```

```
hlavní termín v českém jazyku
hlavní termín v anglickém jazyku
alternativní termín v českém jazyku (synonymum)
alternativní termín v anglickém jazyku (synonymum)
podřazený termín
nadřazený termín
příbuzný termín
```


skos:exactMatch

URL odkaz na českou wikipedii

Pro ilustraci uvádíme sumarizaci dat z jednotlivých systémů v následující tabulce.

Název tezauru	Počet termínů	Vstupní formát	Získané entity
Autority - věcné pojmy	36 265	řádkový MARC	Hlavní termín Asociativní termíny Nadřazené termíny Podřazené termíny Termín v EN jazyku
LANius, Clavius	73 071	CSV	Hlavní termín
Demus	2 882	CSV	Hlavní termín Nadřazený termín (ne u všech) Termín v EN jazyku Termín v DE jazyku
PSH	14 000	RDF/XML	Hlavní termín (CZ, EN) Alternativní termíny (CZ, EN) Nadřazené termíny Podřazené termíny Příbuzné termíny URL adresa na CZ wikipedii

3.2 Proces porovnávání

V procese porovnávání jsme se snažili pomocí více algoritmů uvedených v metodice najít shody, pravděpodobné shody a míru jejich pravděpodobnosti u termínů z výše uvedených zdrojů s termíny ze souboru věcných autorit. Výsledkem je výběr shodného, nebo s jakou pravděpodobností shodného termínu ze souboru věcných autorit s termínem uvedeným jako vstupní termín porovnávaných zdrojů, míra pravděpodobnosti je uvedena v procentech v rozsahu 0-100.

3.2.1 Metody

Porovnávání bylo realizováno na lexikální úrovni s využitím dvou základních metrik a to:

1. Vzdálenostní metrika Jaro: algoritmus, který je užíván pro srovnání řetězců, jako jsou jména, příjmení, ale také pojmenování jakýchkoliv objektů a podobně.
2. Vzdálenostní metrika Jaro Jaro-Winkler: modifikace předchozího algoritmu především s důrazem na prefixu předpony, které jsou typické i pro morfologii našeho jazyka. Díky tomu je možno lépe ošetřit srovnávání slov, které mají stejný slovní základ, můžou mít různé přípony či předpony, jako např jednotné a množné číslo ap. - škola - školy, nebo automobil - automobily, atd.).

3.2.2 Forma výstupu

Pro ilustraci uvádíme ukázkou výstupu porovnávání termínů z tezauru PSH s termíny věcných autorit. Porovnávání bylo realizováno za použití metriky Jaro-Winkler.



PSH - prefLabel	AUTHORITY	PSH ID	AUTHORITY ID	Podobnost (%)
harmonická analýza	harmonická analýza	PSH7554	ph328402	100
mírová hnutí	mírové hnutí	PSH9704	ph123128	97,3333
řešení problému	řešení problémů	PSH9325	ph137287	96,9474
reformovaná církev	reformované církve	PSH7823	ph116026	96,9474

3.3 Výsledky

Dále uvádíme výsledky porovnávání a mapování termínů z uvedených zdrojů.

Tabulka 1: Výsledky podobnosti termínů - metrika Jaro - početnost.

Tezaurus	Shoda - Vzdálenostní metrika Jaro					
	100%	(100%-97%>	(97%-95%>	(95%-93%>	(93%-90%>	(90%-0%>
	počet	počet	počet	počet	počet	počet
PSH - hl. termín	4592	48	143	246	627	8344
PSH - alt termín	887	17	35	67	211	10381
DEMUS	93	0	16	28	28	2717
LANius, Clavius	13266	275	868	1180	3126	54356

Tabulka 2: Výsledky podobnosti termínů - metrika Jaro - procentuální vyjádření

Tezaurus	Shoda - Vzdálenostní metrika Jaro					
	100%	(100%-97%>	(97%-95%>	(95%-93%>	(93%-90%>	(90%-0%>
	%	%	%	%	%	%
PSH - hl. termín	32,80	0,34	1,02	1,76	4,48	59,60
PSH - alt termín	7,65	0,15	0,30	0,58	1,82	89,51
DEMUS	3,23	0	0,56	0,97	0,97	94,27
LANius, Clavius	18,15	0,38	1,19	1,61	4,28	74,39

Tabulka 3: Výsledky podobnosti termínů - metrika Jaro-Winkler - početnost

Tezaurus	Shoda - Vzdálenostní metrika Jaro-Winkler					
	100%	(100%-97%>	(97%-95%>	(95%-93%>	(93%-90%>	(90%-0%>
	počet	počet	počet	počet	počet	počet
PSH - hl. termín	4592	168	355	787	2112	5986
PSH - alt termín	887	42	88	231	588	9762
DEMUS	93	16	27	39	110	2597
LANius, Clavius	13266	1006	1808	3750	10040	43201

Tabulka 4 Výsledky podobnosti termínů - metrika Jaro-Winkler - procentuální

Tezaurus	Shoda - Vzdálenostní metrika Jaro-Winkler					
	100%	(100%-97%>	(97%-95%>	(95%-93%>	(93%-90%>	(90%-0%>
	%	%	%	%	%	%
PSH - hl. termín	32,80	1,20	2,54	5,62	15,09	42,76
PSH - alt termín	7,65	0,36	0,76	1,99	5,07	84,17
DEMUS	3,23	0,56	0,94	1,35	3,82	90,11
LANius, Clavius	18,15	1,38	2,47	5,13	13,74	59,12

Tabulky 1 a 3 obsahují počty porovnávaných a namapovaných termínů z vybraných tezaurů.

Na základě procentuální lexikální podobnosti pojmů jsme celý interval rozdělili na 6 významnějších podintervalů. První interval, resp. hodnota 100% představuje úplnou lexikální shodu porovnávaných termínů (úplná shoda textových řetězců). Každý následující podinterval zaznamenává počty porovnávaných termínů s jejich nejvyšší nalezenou podobností s některým termínem z národních autorit.

V tabulkách 2 a 4 jsou uvedeny procentuální hodnoty vypočtené z počtů termínů v tabulkách 1 a 3. Jednotlivé tezaury neobsahují stejný počet termínů, proto by bylo uvedení jenom počtu porovnaných a namapovaných termínů málo informativní.

V příloze č. 1 uvádíme grafy, které znázorňují získané hodnoty.

3.4 Zhodnocení

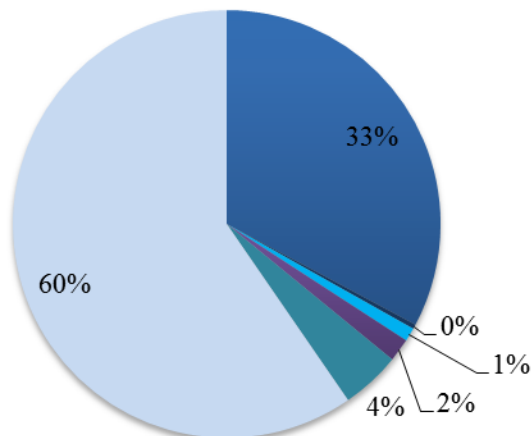
Z uvedených experimentů vyplývá, že metrika JaroWinkler dosahuje přesnější výsledky a dokáže objevit větší množství lexikálně podobných slov, které se liší především sufixem, tedy příponou slova. Největší lexikální podobnost termínů s termíny záznamů věcných autorit dosahuje tezaurus PSH. Naopak nejmenší lexikální podobnost byla dosažena u tezaury DEMUS, kde je možné pomocí lexikální podobnosti automaticky namapovat jen asi 4% záznamů. Nutno však podotknout, že celkově jde jenom o malý počet záznamů a 4% shodu bude možno zvýšit už uvedeným předzpracováním dat. Taky je důležité, kromě intervalu záznamů se 100% shodou, vzít v úvahu ještě další intervaly, jak je možno vidět v příloze. Data v těchto intervalech se liší především formou zápisu termínu, např. jednotné, nebo množné číslo. Jde tedy opět o předzpracování dat, které může tyto výsledky značně upravit ve prospěch automatizace celého procesu. Na porovnávání nemají vliv velká a malá písmena, protože obě metriky s tímto počítají a nevykazují v těchto případech rozdíly. Následující grafy přesně ukazují, jaká je úspěšnost automatizace procesu harmonizace jednotlivých zdrojů i podle daných metrik.



Procentuální podobnost termínů PSH s autoritami pomocí metriky Jaro

Interval podobnosti pojmů:

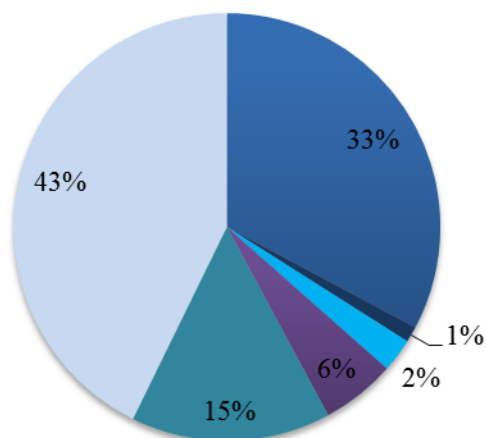
- 100%
- (100% - 97%>
- (97%-95%>
- (95%-93%>
- (93%-90%>
- (90% -0%>



Procentuální podobnost termínů PSH s autoritami pomocí metriky JaroWinkler

Interval podobnosti pojmů:

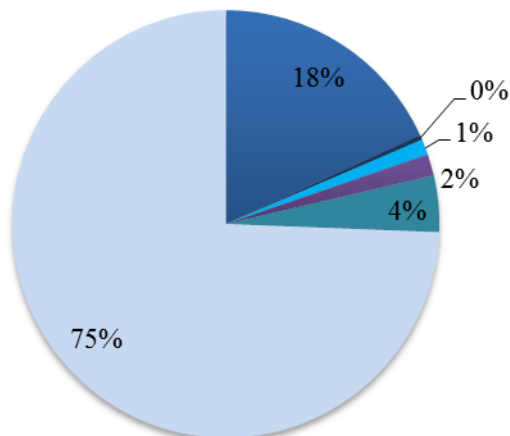
- 100%.
- (100% - 97%>
- (97%-95%>
- (95%-93%>
- (93%-90%>
- (90% -0%>



Procentuální podobnost termínů Lanius s autoritami pomocí metriky Jaro

Interval podobnosti pojmů:

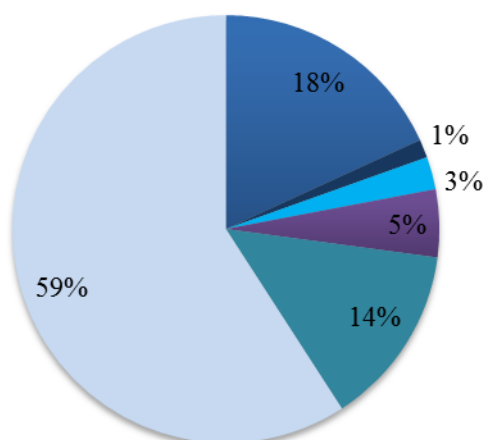
- 100%
- (100% - 97%>
- (97%-95%>
- (95%-93%>
- (93%-90%>
- (90% -0%>



Procentuální podobnost termínů Lanius s autoritami pomocí metriky JaroWinkler

Interval podobnosti pojmů:

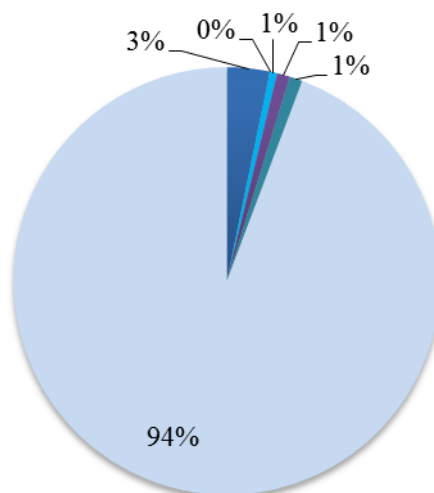
- 100%
- (100% - 97%>
- (97%-95%>
- (95%-93%>
- (93%-90%>
- (90% -0%>



Procentuální podobnost termínů DEMUS s autoritami pomocí metriky Jaro

Interval podobnosti pojmů:

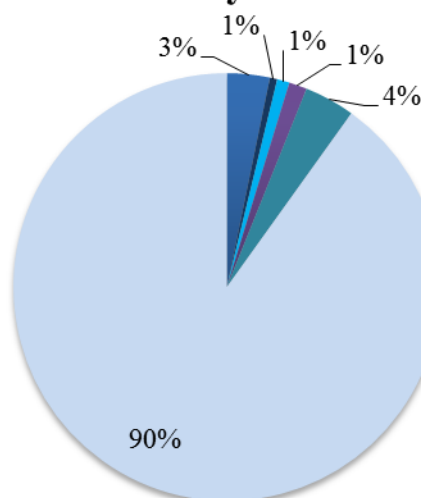
- 100%
- (100% - 97%>
- (97%-95%>
- (95%-93%>
- (93%-90%>
- (90% -0%>



Procentuální podobnost termínů DEMUS s autoritami pomocí metriky JaroWinkler

Interval podobnosti pojmů:

- 100%
- (100% - 97%>
- (97%-95%>
- (95%-93%>
- (93%-90%>
- (90% -0%>



3.4.1 Možnosti zlepšení

Jak již bylo uvedeno výše, v procesu transformace zdrojů do jednotného datového úložiště jsme nevykonávali žádné předzpracování těchto dat, jak jsme uvedli v metodice. Tu je poměrně široký prostor pro vylepšení výsledků automatizace tohoto procesu, což dokazují i výstupy v příloze. Je jasné vidět, jak je to s mírou shody při srovnávání termínů v jednotném a množném čísle, či termínů, které jsou v jiném pádu, nebo které jsou uvedeny v jednom ze zdrojů v invertované podobě. Při samotné realizaci projektu harmonizace by toto

předzpracování do značné míry zvýšilo míru automatizace daného procesu. Dále jsme nevytvářeli grafy a podgrafy, především pro porovnávání tezaurů jako je PSH, EUROVOC, či AGROVOC, protože jde o časově mnohem náročnější procesy; můžeme však uvést, že i v těchto případech je možné s využitím metod uvedených v metodice do značné míry harmonizaci různých zdrojů zautomatizovat. Na základě našich zkušeností, ke kterým jsme při řešení našeho projektu dospěli, můžeme říct, že nejvhodnějším výchozím zdrojem, na který je možno úspěšně mapovat všechny užívané zdroje, jsou Národní tematické autority.

Množné vs. jednotné číslo

Některé tezaury uvádí termíny v jednotném čísle a jiné zase v množném čísle. Tento problém částečně řeší metrika JaroWinkler, která dokáže pracovat se slovním základem a rozdílnou koncovkou zohledňuje v míře podobnosti, jak je patrné v uvedených výstupech. Problematickou je však stanovení prahové hranice podobnosti. V našich pokusech se jako optimální prahová hranice ukázala hodnota 97% v metrice JaroWinkler a XXX v metrice Jaro.

Řešením tohoto problému může být

- přechod na jednotnou formu srovnávaných termínů (dohoda o používání jednotného nebo množného čísla, neinvertovaná podoba zápisu termínu, ap.),
- lemantizace a stemování s použitím korpusu českého jazyka na nalezení základního tvaru termínů.

Závorky s vysvětlujícím kontextem

Některé termíny používají umístění upřesňujícího textu v závorkách za základním termínem. Např.:

- | | |
|--------------------|--------------------------|
| ▪ Lanius a Clavius | české pověsti (beletrie) |
| ▪ AUTHORITY | české pověsti |

Textová podobnost této dvojice s použitím metriky JaroWinkler je pouze 91,85%.

Nebo termín:

- | | |
|-------------|----------------------------|
| ▪ PSH | chicagská škola ekonomie |
| ▪ AUTHORITY | chicagská škola (ekonomie) |

s podobností 98,57%.

Spojovníky

Problémem je také nejednotné užívání spojovníků a s tím spojené zpracování slov

- | | |
|-------------------|-------------------|
| ▪ Lanius, Clavius | zenový buddhismus |
| ▪ AUTHORITY | zen-buddhismus |

s podobností slov 91,85%.

nebo termín

- | | |
|-------------|-------------|
| ▪ PSH | zeta-funkce |
| ▪ AUTHORITY | zeta funkce |

s podobností 96,36%.

Používání velkých písmen

Používání velkých a malých písmen není jednotné a stejné ve všech zdrojích Např.:

- | | |
|-------------|------------|
| ▪ PSH | sms zprávy |
| ▪ AUTHORITY | SMS zprávy |

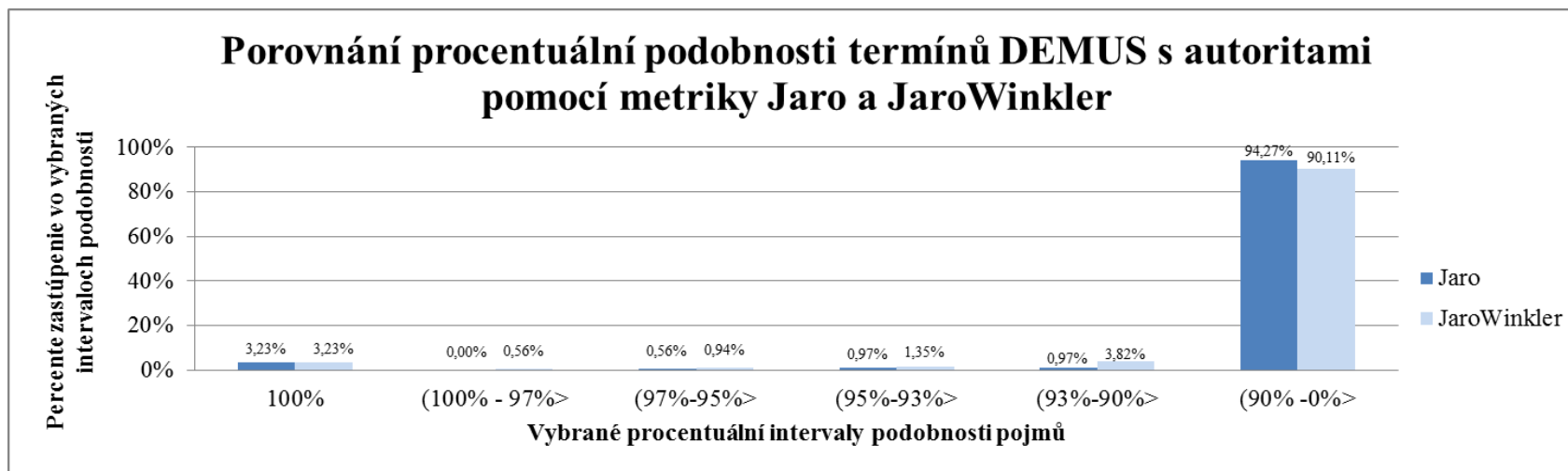
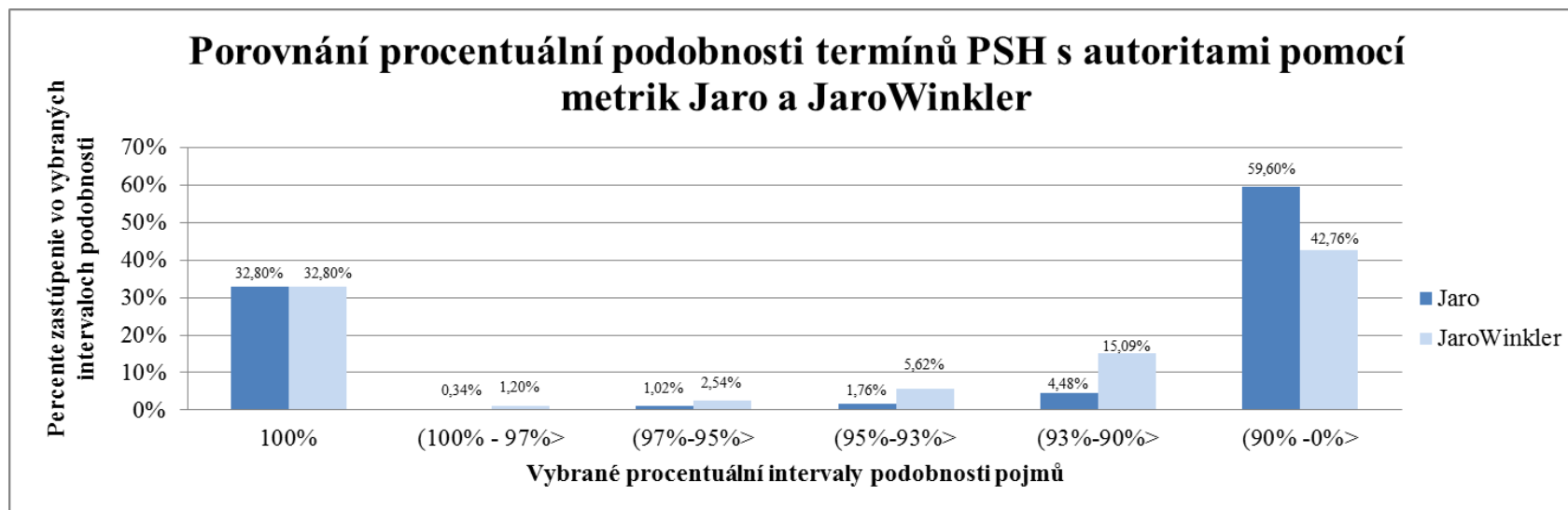
Implementované metody JaroWinkler aj Jaro mají funkce, které zohledňují porovnávání velkých a malých písmen.

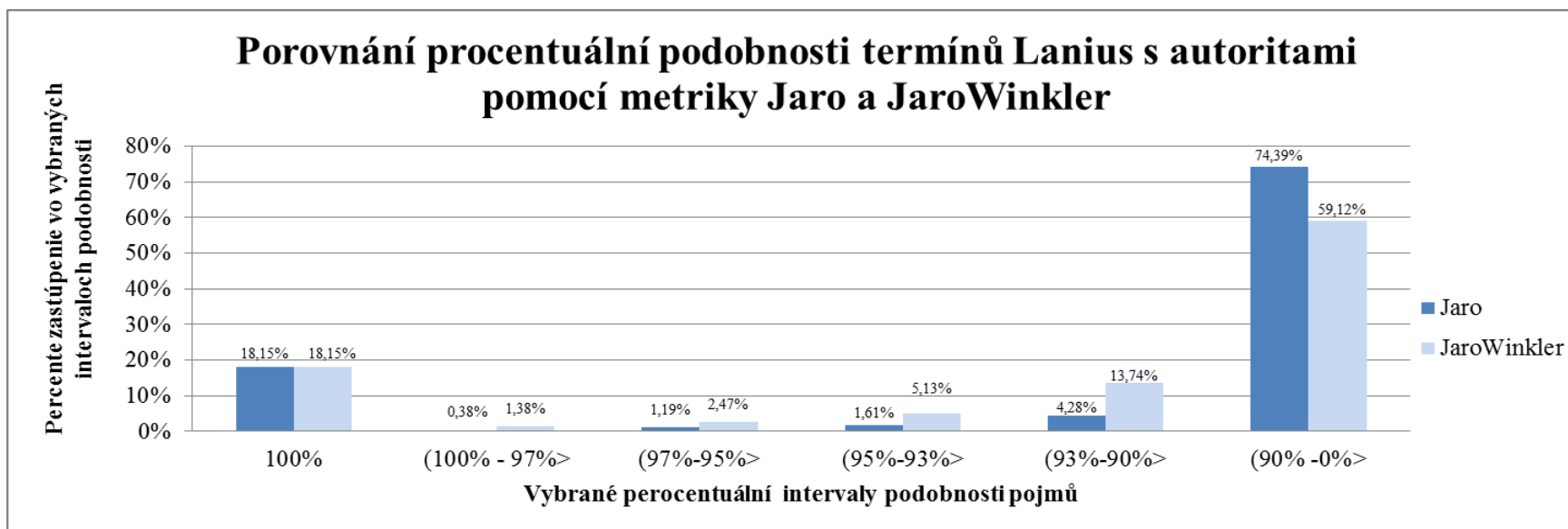
4 Problematika zámeckých knihoven

Zámecké knihovny a jejich fondy vynikají svou pohnutou novodobou historií a úsilím, které bylo vynaloženo na uchování tohoto typu kulturního dědictví. Fondy zámeckých knihoven, jejich zmapování, digitalizace a zpřístupnění jsou také předmětem různých projektů, budoucích potenciálních partnerů projektu INTERPI. Proto byla vypracována úvodní studie, jejímž hlavním cílem bylo na příkladu čtyř zámeckých knihoven poukázat na celou šíři problémů souvisejících se začleněním tohoto typu entit do znalostní databáze INTERPI. Ukazuje se např., že již označení zámeckých knihoven, které je součástí souboru Národních jmenných autorit, např. musí být podrobena důkladné analýze. V rámci znalostní databáze INTERI bude jistě vhodnější používat „přímé“ označení tohoto typu knihoven, tedy „Zámecká Lokšanská knihovna“ a současnou preferovanou formu v Národních autoritách „Březnice (zámek). Lokšanská knihovna“ uvést jako formu variantní.

Studie tvoří přílohu č. 2.

Příloha 1: Grafy s výsledky ověřování metod porovnání termínů





Příloha 2: Studie: Zámecké knihovny v Čechách, na Moravě a ve Slezsku, pod odbornou správou Knihovny Národního muzea