

METODIKA MAPOVÁNÍ A HARMONIZACE REJSTŘÍKŮ, ČÍSELNÍKŮ A ŘÍZENÝCH SLOVNÍKŮ APLIKOVANÝCH V PAMĚŤOVÝCH INSTITUCÍCH

CERTIFIKOVANÁ METODIKA

INTERPI – Interoperabilita v paměťových institucích

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

Zpracovali:

Marie Balíková (Národní knihovna)

Miroslav Kunt (Národní archiv)

Jana Šubová (Cosmotron Bohemia, s. r. o.)

Nadežda Andrejčíková (Cosmotron Bohemia, s. r. o.)



INTERPI

Předmluva

Jak již vyplývá z názvu, Metodika mapování a harmonizace rejstříků, číselníků a řízených slovníků aplikovaných v paměťových institucích je určena především pracovníkům paměťových institucí, kteří mají zájem podpořit myšlenku interoperability tím, že lokálně využívané specifické terminologické zdroje přizpůsobí jistým společným pravidlům. V této souvislosti je potřebné vyzvednout entuziasmus, ochotu a otevřenost odborníků z jednotlivých komunit, které provázeli a provázejí přípravu a realizaci této Metodiky i celého projektu INTERPI.

Rozdílnost a jedinečnost jednotlivých odborných komunit se projevuje především v přístupu, podle kterého se vybírají, reprezentují a uchovávají informace o jednotlivých obecných pojmech dané oblasti. Metodika proto nemůže být prostým a striktním návodem na výběr termínů a jeho uchování ve znalostní databázi INTERPI - takový přístup by předznamenal její neúspěch.

Předkládaná Metodika je především o pochopení a respektování rozmanitosti. Poskytuje přehled metod, které při mapování a harmonizaci terminologických zdrojů lze využít a rovněž doporučený postup pro realizaci konkrétních záměrů v této oblasti. V konečném důsledku je také doporučením, jak postupovat při integraci specifických terminologických zdrojů do znalostní databáze INTERPI, a to i v případě, že tyto terminologické zdroje doposud nebyly reprezentovány pomocí žádného technologického nástroje.

Všem odborníkům, kteří se na přípravě Metodiky podíleli, upřímně děkujeme a doufáme v pokračování spolupráce i na elektronické verzi Metodiky na stránkách projektu INTERPI (www.interpi.cz), kde předpokládáme zveřejňování doplňků a aktualizací Metodiky podle dalšího vývoje standardů v oblasti paměťových institucí.



Slovník

AACR2	Anglo-American Cataloguing Rules [Anglo-americká katalogizační pravidla] mezinárodně využívaná základní příručka pro bibliografický a katalogizační popis, druhé vydání
AAT	Art & Architecture Thesaurus [Tezaurus pro umění a architekturu] řízený slovník z oblasti architektury, umění a hmotné kultury, který obsahuje obecné pojmy zařazené do hierarchické struktury; je zpracován primárně v angličtině, ale obsahuje varianty termínů v dalších jazycích
AGROVOC	řízený slovník (tezaurus) z oblasti zemědělství, potravinářství, lesnictví, rybolovu, životního prostředí apod.
DDC	Dewey Decimal Classification [Deweyho desetinné třídění] univerzální klasifikační systém používaný především v zahraničí (anglicky mluvících zemích)
DEMUS	Dokumentace a Evidence MUzejních Sbírek počítačový systém pro evidenci a dokumentaci sbírek muzeí a galerií
Dublin Core	[Dublinské jádro] metadatový standard pro popis elektronických zdrojů
EuroVoc	vícejazyčný polytematický tezaurus Evropské unie
ISNI	International Standard Name Identifier [Mezinárodní standardní identifikátor jména] standard pro přidělování a správu mezinárodního identifikátor pro tvůrce
ISO	International Organization for Standardization [Mezinárodní organizace pro normalizaci] mezinárodní síť spolupracujících národních normalizačních organizací, zabývá se tvorbou mezinárodních norem
MARC	MAchine Readable Cataloging [strojem čitelná katalogizace] typ katalogizačního formátu
MDT	Mezinárodní desetinné třídění univerzální mezinárodně uznávaný klasifikační systém
MeSH	Medical Subject Headings [Předmětová hesla pro oblast lékařství] tezaurus z oblasti lékařství
MODS	Metadata Object Description Schema [Schéma pro popis metadat objektů] schéma umožňující prezentaci bibliografických dat v strojově čitelné podobě
MSAC	Multilingual Subject Access to Catalogues of National Libraries [Vícejazyčný věcný přístup do katalogů národních knihoven] mezinárodní projekt zaměřený na doplnění souborů národních věcných autorit o cizojazyčné varianty termínů
NAČR	Národní autority České republiky soubor národních autorit budovaný Národní knihovnou ČR
RDA	Resource Description and Access [Popis zdrojů a přístup k nim] aktualizace AACR2 - tedy revidovaná pravidla pro popis dokumentů
RDF	Resource Description Framework [Systém popisu zdrojů] obecný rámec dat (model metadat), která popisují zdrojový dokument ve strojově i lidsky čitelném formátu
SKOS	Simple Knowledge Organization System [Jednoduchý systém organizace poznání] formát pro reprezentaci tezaurů a řízených slovníků ve strojově čitelné podobě
SVAT	Soubor věcných autorit tematických část souboru Národních autorit ČR
VIAF	Virtual International Authority File [Virtuální mezinárodní soubor autorit] projekt a výsledek projektu, kterého cílem je spojit jednotlivé soubory národních autorit
VRA Core	Visual Resources Association Core datový standard vytvořený Asociací pro obrazové zdroje určený pro popis děl vizuální kultury



Obsah

0	Úvod.....	7
1	Teoretická východiska	9
1.1	Typologie terminologických zdrojů pro potřeby projektu INTERPI.....	9
1.1.1	Číselníky	9
1.1.2	Rejstříky a hesláře	10
1.1.3	Řízené slovníky a tezaury.....	12
1.1.4	Klasifikační systémy	14
1.1.5	Soubory autorit.....	16
1.1.6	Ontologie	17
1.2	Principy reprezentace pojmů v terminologických zdrojích.....	17
1.3	Harmonizace nebo mapování?	18
1.3.1	Postup při harmonizaci/mapování terminologických zdrojů	19
1.3.2	Metody porovnávání terminologických zdrojů	19
1.3.3	Metody mapování/harmonizaci terminologických zdrojů.....	26
1.3.4	Technologické nástroje pro reprezentaci terminologických zdrojů	26
2	Pozice a význam národních autorit.....	28
2.1	Charakteristika souboru tematických autorit	29
2.2	Podoba termínu	30
2.3	Význam termínů, vztahy mezi tematickými termíny	31
2.4	Aplikace souboru národních autorit v procesu mapování/harmonizace terminologických zdrojů	32
2.5	Role Souboru věcných tematických autorit v procesu mapování/harmonizace terminologických zdrojů	32
3	Aplikace metod mapování/harmonizaci slovníků v projektu INTERPI	34
3.1	Očekávané problémy specializovaných terminologických zdrojů	34
3.2	Postup mapování/harmonizace terminologických zdrojů na znalostní bázi INTERPI... ..	34
3.3	Hodnocení shody termínů a výsledky porovnávání	36
3.4	Výsledky mapování/harmonizace	36
3.5	Doporučení pro přejímání nových termínů do databáze INTERPI.....	37



3.5.1	Řešení homonymie, synonymie a polysémie	38
3.5.2	Sousloví - víceslovné termíny.....	38
3.5.3	Výskyt termínů obsahujících etnická adjektiva	39
3.5.4	Výskyt „situačních“, „lokálních termínů“ v terminologických zdrojích	39
3.5.5	Formální aspekty termínů používaných v terminologických zdrojích paměťových institucí	40
4	Seznam bibliografických odkazů.....	42

Seznam obrázků

Obr. 1: Ukázka rejstříku obsahujícího individuální i obecné pojmy.....	11
Obr. 2: Ukázka rejstříků, které jsou součástí popisu archivního fondu.....	12
Obr. 3: Příklad části informací o deskriptoru z tezauru AAT	14
Obr. 4: Ukázka záznamu systému MSAC	15
Obr. 5: Ukázka rejstříku MDT, ve kterém jsou notace MDT zpřístupněny s českými a anglickými ekvivalenty	15
Obr. 6: Ukázka podgrafů pro výraz „biologie“ ze dvou zdrojů (zeleně jsou označeny společné uzly)	23
Obr. 7: T-dimenzionální vektor reprezentující pojem z tezauru	24
Obr. 8: Kosinová vzdálenost dvou výrazů	25
Obr. 9: Příklad mapování na základě instancí.....	25



Seznam tabulek

Tab. 1: Přehled zastoupení jednotlivých typů terminologických zdrojů v paměťových institucích	9
Tab. 2: Příklad shodných termínů v třech zdrojích při porovnání na lexikální úrovni.....	20
Tab. 3: Příklad různých termínů ve třech zdrojích se stejným anglickým ekvivalentem (při porovnání na lexikální úrovni)	20
Tab. 4: Lexikální srovnání podobně znějících výrazů	21
Tab. 5: Porovnání termínů pomocí Levensteinovy vzdálenosti	22
Tab. 6: Porovnání termínů pomocí Jaro-Winklerovy vzdálenosti	22

0 Úvod

Projekt INTERPI si klade za cíl vytvořit infrastrukturu pro znalostní databázi paměťových institucí, tzn. poskytnout paměťovým institucím prostor pro uchovávání a výměnu dat o entitách, které jsou předmětem jejich zájmu. Otázky vytváření záznamů o entitách jsou řešeny v Metodice tvorby znalostního modelu INTERPI. Ze znalosti prostředí paměťových institucí a diskusí mezi odbornými komunitami však vyplynulo, že paměťové instituce potřebují rovněž sjednotit jejich existující rejstříky, hesláře, slovníky a soubory autorit tak, aby byl vytvořen snadnější přístup uživatelů k informacím o dokumentech, objektech a sbírkách a podpořena spolupráce mezi odbornými komunitami. Z tohoto požadavku formulovaného paměťovými institucemi vyplynula potřeba vytvořit předkládanou Metodiku mapování a harmonizace rejstříků, číselníků a řízených slovníků aplikovaných v paměťových institucích (dále jen Metodika).

Hlavním cílem Metodiky je poskytnout paměťovým institucím informace o metodách a postupech využívaných v projektu INTERPI v případech, že mají zájem slovníky, tezaury, hesláře sjednotit se znalostní databází INTERPI. Jde tedy především o vytvoření prostředí pro integraci oborových slovníků do znalostní databáze INTERPI, a to v co největší možné míře.

Metodika je primárně zaměřena na mapování a harmonizaci obecných pojmů, protože tato problematika je v oblasti paměťových institucí v Česku zatím řešena pouze okrajově, i když se specifické terminologické zdroje tematicky překrývají. Postupy uvedené v Metodice však lze s malými obměnami využít i při harmonizaci a mapování individuálních pojmů, což rozšiřuje její původní aplikační oblast.

Metodika zároveň může pomoci paměťovým institucím při tvorbě vlastních (oborově specializovaných) slovníků nebo tezaurů tím, že poskytuje informace o vhodných postupech výběru a zpracování termínů a vztahů mezi nimi. Znalostní databáze INTERPI zároveň poskytuje nástroj, který mohou využít instituce, které nemají technologické zázemí umožňující jim tvorbu vlastního slovníku a tezauru.

Při zpracování Metodiky jsme v prostředí paměťových institucí v Česku narazili na několik specifických rysů, které výrazným způsobem ovlivnily především samotný obsah Metodiky a v konečném důsledku i její název. Vzhledem k malému výskytu používaných řízených slovníků v archivech, muzeích a galeriích v Česku jsme původní záměr zaměřit Metodiku pouze na problematiku řízených slovníků rozšířili o další typy zdrojů termínů - tj. rejstříky, číselníky a hesláře.

První část Metodiky obsahuje teoretická východiska - především typologii zdrojů termínů, které se v paměťových institucích v Česku využívají, dále také informace o metodách vhodných pro porovnávání termínů a pro samotnou harmonizaci a mapování.

Druhá kapitola Metodiky je zaměřena na problematiku souboru národních autorit, a to především z hlediska jeho využití jako terminologického základu pro znalostní databázi INTERPI.

Třetí část představuje aplikaci metod mapování/harmonizace na konkrétní prostředí a podmínky projektu INTERPI. Podrobně uvádí postupy a podmínky přebírání termínů ze specializovaných slovníků do znalostní báze INTERPI. Navržené postupy vycházejí ze specifických rysů analyzovaných terminologických zdrojů a snaží se co v největší míře respektovat požadavky odborných komunit.

1 Teoretická východiska

Vypracování Metodiky mapování a harmonizace rejstříků, číselníků a řízených slovníků aplikovaných v paměťových institucích si v počátcích vyžadovalo sumarizovat teoretické i praktické poznatky z více oblastí - věcného zpracování v paměťových institucích, ze zpracování sbírek, objektů a dokumentů obecně, též z různých mezinárodních projektů zabývajících se harmonizací a mapováním slovníků a tezaurů, a v neposlední řadě také z oblasti informatiky a informačních technologií.

Tato část Metodiky představuje její teoretický základ především v oblasti terminologie a základních principů tvorby zdrojů termínů, které je potřebné mapovat/harmonizovat v jednotlivých typech paměťových institucí a rovněž v oblasti metod, které jsou v metodice aplikovány.

1.1 Typologie terminologických zdrojů pro potřeby projektu INTERPI

Pro potřeby Metodiky budeme považovat za terminologické zdroje všechny soubory termínů, které se používají v paměťových institucích při zpracování dokumentů, objektů a sbírek, bez ohledu na jejich obsah, formu a strukturu. V prostředí paměťových institucí v Česku se využívá široké spektrum terminologických zdrojů, jednotlivé typy si představíme blíže, přičemž jako kritérium pro jejich rozdělení využijeme míru organizace, strukturalizace, způsob a účel jejich vytvoření.

Tab. 1: Přehled zastoupení jednotlivých typů terminologických zdrojů v paměťových institucích

	Knihovny	Muzea	Galerie	Archivy
Rejstříky	ano	ano	ano	ano
Číselníky	ano	ano	ano	ano
Řízené slovníky	ano	ano	ne	ne
Tezaury	ano	ne	ne	ne
Autority	ano	ano	ano	ne
Ontologie	experimentálně	ne	ne	ne

1.1.1 Číselníky

Číselníkem se rozumí „seznam přípustných hodnot datového prvku obvykle ve formě dvojic, to znamená kódovaného údaje a hodnoty jeho kódu“¹. I další definice zdůrazňuje jako hlavní charakteristiku číselníku spojení hodnoty a kódu - číselník „je obvykle uspořádaný seznam entit (nejčastěji ve formě tabulky nebo katalogu), kde je každé konkrétní entitě přiřazen jednoznačný kód“². Číselníky obvykle nemají vyjádřeny vztahy mezi jednotlivými prvky, někdy obsahují odkazy z neplatných kódů nebo variantních vyjádření prvku (např. číselník zemí podle ISO 3166).

¹ Zákon č. 365/2000 Sb.

² Číselník. In *Wikipedie : otevřená encyklopedie* [online].

Číselníky se mohou vázat na konkrétní aplikační oblast (např. číselník typů objektů formátu Dublin Core, číselník studijních oborů), nebo mohou mít širší využití (např. číselník jazyků podle ISO 639); mohou mít platnost lokální (např. číselník druhu územní působnosti Českého statistického úřadu) nebo mezinárodní (např. kódy zemí podle ISO 3166).

V oblasti paměťových institucí se využívají obecné a mezinárodní číselníky (např. kódy zemí nebo jazyků) a také specifické číselníky, které se odvíjejí od používaného informačního systému a oblasti. Z důvodů možností aplikovaných informačních technologií se setkáváme s tím, že se i terminologické zdroje s vyšší mírou strukturalizace prezentují v informačním systému jako číselník (např. v systému DEMUS, který je využíván v muzeích).

Z hlediska Metodiky budou předmětem harmonizace a mapování především specifické číselníky. Obecné a mezinárodní číselníky není potřeba harmonizovat, mohou ale být využity při popisu entit (viz Metodika tvorby znalostního modelu).

1.1.2 Rejstříky a hesláře

Obecně termín **rejstřík** (někdy též **index**) představuje „sekundární dokument obsahující soupis významných pojmů nebo údajů objevujících se v dokumentu, uspořádaných podle jednotného principu (abecedně, systematicky, chronologicky) a doplněných odkazem na místo jejich výskytu“³. V informačních systémech se termín využívá obdobně a znamená soupis termínů vyskytujících se v záznamech a vybraných na základě určených kritérií. Rejstřík v informačních systémech představuje přístupový prvek k záznamům neboli pomůcku pro vyhledávání. Budování rejstříků nevyžaduje žádné speciální znalosti ani složité softwarové nástroje. Jednoduchost vytváření rejstříků je možné považovat za jejich největší nevýhodu, jelikož jejich kvalita značně závisí od zpracovatele.

V projektu INTERPI chápeme rejstřík jako jednoduchý seznam slov a slovních spojení (označení pojmů), bez jakéhokoliv vysvětlení jejich významu, tedy prostý seznam použitých termínů.

Nejčastěji se v paměťových institucích setkáváme s rejstříky abecedně řazených klíčových slov a s předmětovými hesláři, tj. abecedními seznamy předmětových hesel; v obou případech se jedná o pojmy nebo jejich řetězce používané při předmětové, neboli věcné klasifikaci kulturních objektů, které jsou předmětem popisu. Tyto rejstříky ve většině případů obsahují typologicky nerozlišené individuální i obecné pojmy např.:

1907–1918
3.–2.stol.př.n.l.
Domažlice-1431
Chmelnickij, Bogdan Michajlovič
Klopstock, Friedrich Gottlieb
Münzer, Tomáš, 1490–1525
oportunisté
projekční práce
Vítkov, bitva 1420
Zástřizly
Žešov

³ BALÍKOVÁ, Marie. Rejstřík. In *KTD* [online databáze].

Obr. 1: Ukázka rejstříku obsahujícího individuální i obecné pojmy

Sbírky Regionálního muzea v Litomyšli ONLINE

Vyhledávání
fond FOTOARCHIV

Filtr výpisu:

Předmět

Datace -

Autor rovná se

Popis obsahuje

Materiál obsahuje

Událost rovná se

Téma rovná se

Lokalita rovná se

Osoba rovná se

Inventární číslo rovná se

Výsledek seřadit podle:

Inventární číslo Vzestupně

TÉMA - slovník pro fond Fotoarchiv

- HRAČKY DĚTSKÉ
- interiér - DÍLNA ŘEMESLNICKÁ
- interiér - OBCHOD KNIHKUPECKÝ
- interiér - OBYTNÝ
- interiér - RESTAURACE
- KOČKA
- Komunální služby města Litomyšle
- KŮŇ
- MOTYČKA, František - jeho obrazy
- MŠE
- muzejní expozice, Litomyšl
- ODĚV DĚTSKÝ
- ODĚV MUŽSKÝ
- ODĚV ŽENSKÝ
- OSVĚTLENÍ VEŘEJNÉ
- PES
- REKLAMA
- skupina hudební H.G.H.
- skupina hudební HOMO LUDENS
- spolek OSTROSTŘELECKÝ, Litomyšl
- spolek ŘEZNÍKŮ A UZENÁŘŮ, Litomyšl
- spolek SOKOLSKÝ, Litomyšl
- spolek VLASTIMIL, Litomyšl
- sport - TENIS
- strana AGRÁRNÍ, Litomyšl

Přípravné práce spojené s potenciální harmonizací a mapováním těchto nerozlišených rejstříků klíčových slov jsou mnohem náročnější ve srovnání s harmonizací a mapováním rejstříků typologicky rozlišených (viz kapitola 3.2).

V archivech se význam a účel rejstříků více přibližuje původnímu významu rejstříku jako součástí informačního aparátu dokumentu. Rejstříky jsou tu vytvářeny jako součást archivních pomůcek a obvykle jsou rozděleny na tyto typy:

- rejstřík věcný - tj. předmětový,
- rejstřík místní - tj. zeměpisný nebo geografický,
- rejstřík osob a korporací - v některých případech je rejstřík korporací sloučen s věcným,
- rejstřík datací,
- rejstřík názvový,
- rejstřík čísel dokumentů.

První tři z výše uvedených se vyskytují nejčastěji, využívání ostatních závisí na možnostech informačního systému používaného při tvorbě archivních pomůcek.

Obr. 2: Ukázka rejstříků, které jsou součástí popisu archivního fondu

The screenshot shows the BADATELNA.EU website interface. At the top, there's a navigation bar with links: Základní informace, Informace o fondu, Inventář, and Rejstříky. A search bar is present with the text 'Vyhledávání:'. To the right, there's a dropdown menu for 'druh rejstříku:' with options: Vyberte druh rejstříku, Čísel dokumentů, Datací, Institucí, Místní, Věcný, Názevový, and Osobní. Below this, there's a table with the header 'Rejstříkové heslo - specifikace'. The table contains several rows of text, including 'Abdank-Abžoltovská, V., rozená Klichová - reemigrantka z Volyně', 'Abt, Alexandr - vojn československé zahraniční armády v SSSR', 'adresáře', 'Adventní nakladatelství v Brně', 'Afrika', 'akvarely', 'alba', 'Andruško, Josef - vojn československé zahraniční armády v SSSR', 'Antipovič, Dimitrij, Ing. - zemědělský odborník', 'Antonij - volyňský biskup', 'Argentina', 'armády viz též odboj, slavnosti vojenské, války, vojáci - 1. československý armádní sbor v Sovětském svazu', 'armády viz též odboj, slavnosti vojenské, války, vojáci - maďarská', and 'armády viz též odboj, slavnosti vojenské, války, vojáci - německá'.

Obdobným typem terminologického zdroje jako rejstřík je **heslář**, tj. „abecedně seřazený seznam hesel použitých při zpracování předem vymezené problematiky“⁴. S termínem heslář se obvykle spojuje soubor tematických, případně geografických hesel.

Abecedně uspořádané rejstříky předmětových hesel obsahují též typologicky nerozlišené označení entit, navíc upřesňující tematické, geografické a chronologické informace jsou uváděny jako podhesla, která jsou součástí řetězce předmětových hesel. Před potenciální harmonizací a mapováním je nutné tyto řetězce informací rozčlenit na jednotlivé segmenty; další postup je potom totožný jako u harmonizace a mapování nerozlišených rejstříků klíčových slov.

Rejstříky a hesláře jsou zastoupeny ve všech typech paměťových institucí a jsou tak důležitým terminologickým zdrojem, i když jejich harmonizace a mapování je z výše uvedených důvodů náročnější.

1.1.3 Řízené slovníky a tezaury

Vyšší úroveň organizace termínů představují řízené slovníky a tezaury.

Řízený slovník je „slovník lexikálních jednotek selekčního jazyka uspořádaný specifickým způsobem (např. zahrnuje vztahy ekvivalence, hierarchie a asociace), který slouží pro indexaci a vyhledávání dokumentů“⁵. Za řízený slovník je někdy považován i soubor autorit, ale souboru autorit budeme věnovat pozornost samostatně (viz kapitola 1.1.5).

⁴ JURČÁKOVÁ, Z. Heslář. In *Informačná výchova*, s. 111.

⁵ BALÍKOVÁ, Marie. Řízený slovník. In *KTD* [online databáze].

Tezaurus můžeme chápat jako řízený a měnitelný slovník lexikálních jednotek opírající se o slovní zásobu jednoho či více přirozených jazyků, který vyjadřuje sémantické vztahy mezi lexikálními jednotkami. Jde tedy o deskriptorový selekční jazyk založený na deskriptorech, jejichž forma a vztahy jsou standardizovány slovníkem se speciální strukturou. Norma ISO 5964 definuje tezaurus jako slovník řízeného selekčního jazyka uspořádaný tak, že explicitně zachycuje apriorní vztahy mezi pojmy. Jde o vztahy nadřazenosti, podřazenosti, shody či podobnosti významu, přičemž jednoznačně určuje preferovaný termín. Z hlediska úrovně hierarchického uspořádání můžeme hovořit o jedno či víceúrovňových tezaurech a z hlediska jazykového zase o jedno či vícejazyčných tezaurech. Většinou jsou budovány odborníky pro danou oblast - obor (téma), kde se snaží zachytit vše, co ji vystihuje a tvoří tak základní terminologii oboru.

Tezaury mohou být vytvořeny pro mezinárodní použití (např. AAT, MeSH, EuroVoc) nebo pro lokální aplikaci v jedné zemi nebo instituci.

Z hlediska mapování a harmonizace představují řízené slovníky a tezaury specifický problém. Jejich vysoká míra organizace a strukturalizace je výhodná především proto, že umožňuje lépe identifikovat význam jednotlivých termínů (deskriptorů, jak se v případě tezaurů označují preferované termíny). Při mapování tezaurů a řízených slovníků je možné nejednoznačné termíny identifikovat pomocí nadřazených nebo podřazených termínů. Na druhé straně, v případě tezaurů a řízených slovníků můžeme stěží hovořit o harmonizaci (pokud nejde pouze o sladění formy termínů), protože zařazení termínů do struktury vychází ze specifických potřeb daného odboru, pro který je tezaurus nebo řízený slovník vytvářen.

Obr. 3: Příklad části informací o deskriptoru z tezauru AAT

ID: 300054144

Record Type: concept

 **dance (discipline)** (performing arts, arts (broad discipline), ... Disciplines (facet))

Note: Performing art form realized in rhythmic movement, creating visual designs by a series of poses and tracing of patterns through space in the course of measured units of time.

Terms:

dance (discipline) (**preferred**,C,U,LC,English-P,D,U,U)
dancing (discipline) (C,U,English,UF,U,U)
舞蹈 (C,U,Chinese (traditional)-P,D,U,U)
wǔ dào (C,U,Chinese (transliterated Hanyu Pinyin)-P,UF,U,U)
wu dao (C,U,Chinese (transliterated Pinyin without tones)-P,UF,U,U)
wu tao (C,U,Chinese (transliterated Wade-Giles)-P,UF,U,U)
dans (C,U,Dutch-P,D,U,U)
danza (C,U,Spanish-P,D,U,U)

Facet/Hierarchy Code: K.KD

Hierarchical Position:

 Activities Facet
 Disciplines (facet) (G)
 disciplines (G)
 humanities (G)
 <arts and related disciplines> (G)
 arts (broad discipline) (G)
 performing arts (G)
 dance (discipline) (G)

Additional Notes:

Chinese (traditional) 指一種表演藝術，表演者於一段特定的時間內，在某空間中展現富節奏的肢體動作，並經由一連串的姿勢和形態模擬來創造視覺元素。

Dutch Uitvoerende kunstvorm uitgevoerd in de vorm van ritmische beweging, waarbij visuele ontwerpen worden gecreëerd door een reeks poses en van bepaalde patronen in de ruimte gedurende afgemeten tijdeenheden.

Spanish Forma de arte interpretativa realizada en movimiento rítmico, creando diseños visuales por medio de una serie de poses y un trazado de modelos a través del espacio en a lo largo de unidades de tiempo medidos.

Related concepts:

involves **choreographers**

..... (people in the performing arts, <people in the arts>, ... People) [300025649]

1.1.4 Klasifikační systémy

Speciálním terminologickým zdrojem jsou **klasifikační systémy**, které představují „logické uspořádání množiny klasifikovaných pojmů na základě vztahů a závislostí mezi charakteristickými znaky těchto pojmů. Struktura klasifikačního systému je zakotvena v klasifikačním schématu. Seznam klasifikovaných pojmů vyjádřených notacemi a opatřených jazykovými ekvivalenty je obsažen v klasifikačních tabulkách. Vyjadřovacím prostředkem klasifikačního systému je systematický selekční jazyk“⁶.

Klasifikační systémy mohou být rovněž vytvářeny na mezinárodní úrovni (např. MDT) a tehdy jsou mnohdy řešeny jako vícejazyčné, nebo jsou vytvářeny na lokální úrovni, pro potřeby jedné instituce.

⁶ BALÍKOVÁ, Marie. Klasifikační systém. In *KTD* [online databáze].

Notace mezinárodních klasifikačních systémů se využívají také v řízených slovnících, tezaurech nebo autoritách jako způsob vyjádření termínu, který není závislý na jazyce. Představují tak důležitý prvek pro mapování nebo harmonizaci terminologických zdrojů. Notace klasifikačního systému může poskytnout informaci o významu termínu v případě nejednoznačných termínů.

Příklad projektu MSAC (Obr. 4) ukazuje možnost připojení ekvivalentů termínů ve slovenštině, slovinštině, makedonštině, litevštině a lotyšštině k původním českým termínům. Připojení bylo realizováno na základě notace MDT a anglického ekvivalentu uloženého v záznamu českého termínu. Projekt MSAC byl realizován v doménách sociologie a právo.

Obr. 4: Ukázka záznamu systému MSAC

072 7	a 316.8 x Manželství a rodina. Sexuální vztahy 2 Konspekt 9 18
	a 343 x Trestní právo 2 Konspekt 9 16
	a 616.89 x Psychiatrie 2 Konspekt 9 14
080	a 316.835.2 c sociologie d sociology 2 MRF_2003
	a 343.542.5 c trestní právo d criminal law 2 MRF_2003
	a 616.89-008.442.38 c psychiatrie d psychiatry 2 MRF_2003
150	a incest 7 ph121057
450	a krvesmilstvo 0 o
550	a psychopatologie 7 ph124819
	a sexuální morálka 7 ph125509
	w g a psychosexuální poruchy 7 ph241652
	w g a sexuální trestné činy 7 ph125513
	w g a trestné činy proti lidské důstojnosti 7 ph436052
75007	a incest 2 eczenas
75037	a kraujomaiša b 343.542.5 c baudžiamoji teisė 2 lit
	a kraujomaiša b 616.89-008.442.38 c psichiatrija 2 lit
	a kraujomaiša b 316.835.2 c sociologija 2 lit
75047	a Incest b 316.835.2 c sociologija 2 mac
	a Incest b 343.542.5 c krivično pravo 2 mac
	a Incest b 616.89-008.442.38 c psihijatrija 2 mac
75057	a incest b 316.835.2 c sociológia 2 slo
	a incest b 343.542.5 c trestné právo 2 slo
	a incest b 616.89-008.442.38 c psichiatria 2 slo
75067	a Incest b 343.542.5 c criminal law 2 slv
	a Incest b 616.89-008.442.38 c psychiatry 2 slv
	a Incest b 316.835.2 c sociology 2 slv

V některých informačních systémech využívaných v paměťových institucích jsou klasifikační systémy - resp. jejich notace a jazykové ekvivalenty - prezentovány v podobě jednoduchého číselníku, např. vybrané znaky MDT (Obr. 5).

Obr. 5: Ukázka rejstříku MDT, ve kterém jsou notace MDT zpřístupněny s českými a anglickými ekvivalenty

<u>608.34:663/664</u> -- geneticky modifikované potraviny / genetically modified foods
<u>61</u> -- lékařství / medicine
<u>61</u> -- zdravotní / health education
<u>61:001.4</u> -- lékařská terminologie / medical terminology
<u>61:001.891</u> -- lékařský výzkum / medical research
<u>61:001.894</u> -- lékařské objevy / medical discoveries
<u>61:004.7</u> -- elektronické zdravotnictví / eHealth
<u>61:007</u> -- lékařská informatika / medical informatics
<u>61:007</u> -- lékařská kybernetika / medical cybernetics
<u>61:061.23</u> -- lékařské společnosti / medical societies

1.1.5 Soubory autorit

Specifické postavení při zpracování a zpřístupňování informací (zejména v knihovnách) sehrávají soubory autorit.

Autority definují knihovny jako „soubor ověřených a unifikovaných jmenných a/nebo věcných selekčních údajů určených pro zpracování a vyhledávání dokumentů s nezbytným odkazovým a poznámkovým aparátem“⁷. Informace o autoritě jsou zaznamenány v podobě záznamu autority (autoritním záznamu), který je definován jako „autorizovaná forma jména kombinovaná s dalšími informačními elementy, které identifikují a popisují jmenovanou entitu a mohou také odkazovat na příbuzné autoritní záznam“⁸. **Soubor autorit** představuje potom soubor autoritních záznamů.

Soubory autorit můžeme považovat za řízené slovníky, jejichž cílem je umožnit vyhledávání dokumentů a jiných informačních objektů zpracovávaných v knihovnách. Ve většině případů obsahují pouze ty selekční prvky, které byly použity při zpřístupnění knihovnických fondů; rovněž standardy a pravidla, podle kterých se utvářejí, odpovídají potřebám knihoven a pokud mají sloužit i potřebám jiných komunit, musejí být těmito požadavkům přizpůsobeny. Ukazuje se též, že při zpracování těchto termínů knihovníci neuvádějí potřebné množství informací a nezkoumají všechna fakta související s daným termínem, ale věnují se pouze těm, které jim umožní daný termín jednoznačně vymezit vůči ostatním termínům.

Soubory autority jsou budovány většinou na národní úrovni a rozlišují se podle typu autoritních záznamů (viz kapitola 2).

Problematika souboru autorit je specifická v tom, že jsou v oblasti knihoven poměrně rozšířené, což vede k vytváření projektů zaměřených na mezinárodní spolupráci při jejich vytváření, harmonizaci a propojování - např. personální autority a autority korporací jsou na mezinárodní úrovni mapovány v rámci projektu VIAF⁹.

Mezinárodní projekty a spolupráce v oblasti souborů autorit se zaměřuje rovněž na aplikaci společného identifikátoru pro vybrané entity - např. v roce 2012 vstoupila v platnost ISO norma ISO 27729:2012 - International Standard Name Identifier, tj. Mezinárodní standardní identifikátor jména. Pole této normy identifikátor ISNI trvale identifikuje tzv. „veřejné identity“. Veřejnou identitou se podle této normy rozumí jméno, pod kterým daná entita v současnosti je, nebo v minulosti byla, veřejně známa. Většina reálných osob má jednu veřejnou identitu (jediné jméno), některé jich ale mohou mít více, podobně jako tomu je při zpracování personálních autorit, jestliže tatáž osoba publikovala různé žánry pod různými jmény. Cílem je, aby „veřejné identity“ byly identifikovány pomocí identifikátorů ISNI napříč různými obory tvorby obsahů. Norma také definuje způsob tvorby těchto identifikátorů, stejně jako hlavní principy jejich přidělování, samotný registrační systém ISNI, registrační metadata a referenční databáze ISNI.

⁷ VODIČKOVÁ, Hana. Řízený slovník. In *KTD* [online databáze].

⁸ *ISAAR (CPF) : Mezinárodní standard pro archivní autoritní záznamy korporací, osob a rodů*.

⁹ Pro více informací viz: <http://www.viaf.org>

Příklady mezinárodních projektů v oblasti souboru autorit podtrhují význam aktivit v oblasti harmonizace a mapování různých zdrojů informací o entitách.

1.1.6 Ontologie

Ontologie představuje popis vztahů mezi kategoriemi jisté části reálného světa, je obecnější než řízený slovník a obecnější než tezaurus tím, že obsahuje více významových vztahů (nejen vztahy asociace a hierarchie). Nenahrazuje tezaurus, protože neobsahuje doporučené termíny. Řeší rozklad doménové znalosti na základní znalostní prvky - elementy a specifikuje sémantické vztahy mezi nimi.

Ontologie představují nový směr ve tvorbě terminologických zdrojů, zdrojů informací o entitách a také nový směr pro prezentaci těchto dat v podobě vhodné pro sémantický web (např. finský projekt ONKI¹⁰).

1.2 Principy reprezentace pojmů v terminologických zdrojích

Z hlediska mapování a harmonizace terminologických zdrojů je důležitá též problematika reprezentace pojmů v jednotlivých typech terminologických zdrojů používaných v paměťových institucích. Na principech, které se při tvorbě terminologických zdrojů využívají, závisí především struktura, rozsah i formální stránka zvoleného označení pojmu - tedy termínu.

Pro úplnost je potřeba objasnit vztah mezi pojmem a termínem. **Pojem** je definován jako „forma myšlení odrážející podstatné vlastnosti předmětů, jevů objektivní skutečnosti. Pojmy mohou být všeobecné a jedinečné, univerzální a neutrální, konkrétní a abstraktní, pozitivní a negativní. Tvorba nového pojmu nebo jeho rozšíření, modifikace známého pojmu je často spojená s dokazováním“¹¹. Reprezentantem pojmu je **termín**, tj. „slovo nebo slovní spojení použité k označení pojmu“¹².

V souvislosti s tvorbou různých typů terminologických zdrojů se v literatuře z oblasti informační vědy vyskytuje i termín **lexikální jednotka**, který představuje označení pojmu v systému selekčních jazyků. Lexikální jednotka je „sekvence písmen, číslic a dalších specifických znaků používaná pro označení určitého pojmu. V selekčních jazycích založených na bázi přirozeného jazyka je lexikální jednotka reprezentována převážně slovním vyjádřením, obvykle ve formě substantiva nebo substantivního spojení. Termín se někdy používá i v oblasti systematických selekčních jazyků pro označení klasifikačního znaku, resp. obecně pro označení základního prvku jakéhokoliv selekčního jazyka. Z hlediska struktury selekčního jazyka představuje

¹⁰ LAPPALAINEN, Mikko - FROSTERUS, Matias - NYKIRI, Susanna. Reuse of library thesaurus data as ontologies for the public sector. In *IFLA 2014* [online].

¹¹ JONÁK, Zdeněk. Pojem. In *KTD* [online databáze].

¹² JONÁK, Zdeněk. Termín. In *KTD* [online databáze].

lexikální jednotka jednoduchý pořádací znak. Soubor použitých lexikálních jednotek tvoří lexikum selekčního jazyka.“¹³

Pro mapování a harmonizaci terminologických zdrojů jsou důležité principy, které se využívají pro výběr termínů.

Z procesního hlediska lze řízené terminologické zdroje vytvářet induktivní metodou (tj. „zdola“ s dokumentem v ruce) výběrem vhodných termínů ze slov obsažených ve zpracovávaných dokumentech, nebo deduktivní metodou (metodou shora) na základě analýzy terminologické oblasti oboru, pomocí oborových slovníků apod. Výhodou prvního postupu je to, že ke konkrétnímu termínu je navázán nějaký dokument, ve kterém lze význam termínu ověřit a takto vytvořený terminologický zdroj odráží aktuální stav oblasti reprezentovaný v dokumentech. V druhém případě je výhodou komplexnost terminologického zdroje a případně jeho propracovanější struktura.

Z hlediska formální stránky zvolených termínů lze v terminologických zdrojích rozlišit rovněž dva principy: při prvním se používají jednotlivá slova, která se kombinují až při jejich využití při zpracování dokumentů nebo objektů, při druhém se používají přesnější víceslovné termíny nebo kombinace slov.

1.3 Harmonizace nebo mapování?

V souvislosti s problematikou sjednocování tezaurů, řízených slovníků, terminologických systémů a soustav nebo různých standardů se používají dva termíny: harmonizace a mapování. Někdy ve stejném významu, někdy v rozličných souvislostech.

Termín **harmonizace** znamená uvedení v soulad, sladění, odstranění rozdílů¹⁴. Termín harmonizace se objevuje v souvislosti se systémy norem (např. harmonizace norem státu s mezinárodními normami), v oblasti legislativy (např. harmonizace norem státu s legislativou EU) apod. V oblasti paměťových institucí se používá především v souvislosti se soubory národních a lokálních autorit. Harmonizace lokálních souborů autorit se soubory národních autorit v těchto případech znamená úpravu obsahu a formy lokálních autorit tak, aby odpovídaly obsahu a formě národních autorit. Důležité je, že při harmonizaci autorit je možné zachovat i specifické rysy lokálních autorit (například specifické poznámky, biografická data apod.), které soubor národních autorit neobsahuje. Z praxe harmonizace autorit budeme vycházet i při vymezení tohoto termínu pro potřeby Metodiky. Pod pojmem harmonizace (slovníků, heslářů, číselníků) budeme chápat proces, jehož cílem je odstranit rozdíly, jimiž se harmonizovaný systém odlišuje od jiného systému, se kterým má být uveden v soulad. Odstranění rozdílů může spočívat ve změně formy termínů, ve změně vztahů mezi termíny apod.

Termín **mapování** se v oboru informační věda používá ve dvou významech. Jednak ve smyslu vytváření map reprezentujících související objekty, entity (např. mapování vědy) a potom ve

¹³ BALÍKOVÁ, Marie. Lexikální jednotka. In *KTD* [online databáze].

¹⁴ *Slovník spisovné češtiny* [online databáze]. *ABZ : slovník cizích slov* [online databáze].

smyslu „propojení každého prvku dané množiny s jedním nebo více prvky druhé množiny“¹⁵. Ve druhém významu se termín mapování spojuje například s mapováním standardů a formátů (např. MARC - MODS, MARC - RDA, Dublin Core - VRA Core). V aplikaci na oblast řízených slovníků a rejstříků se mapování týká především propojení tezaurů, klasifikačních systémů apod. (např. DDC - MDT), přičemž se mapované systémy (termíny v mapovaných systémech) nikterak neupravují. Pro potřeby projektu INTERPI bude mapování znamenat propojení mezi entitami ve znalostní databázi INTERPI a termíny (pojmy) ze slovníku, rejstříku, číselníku používaném v konkrétní paměťové instituci.

Z výše uvedené definice obou termínů vyplývá, že cílem projektu INTERPI i této Metodiky je harmonizace oborových terminologických zdrojů (včetně jejich integrace do znalostní databáze INTERPI nebo alespoň jejich propojení s ní), která vytváří předpoklad pro sémantickou interoperabilitu - termíny jsou sjednocené, mají stejnou formu a je zajištěn jejich stejný význam a tím i použití. Jsme si však vědomi, že při heterogenitě prostředí v paměťových institucích bude úspěchem aplikace Metodiky i „pouhé“ mapování slovníků, heslářů, číselníků na znalostní databázi INTERPI, které umožní dále se získanými výsledky pracovat. Také je potřebné zdůraznit, že v praxi se harmonizace a mapování často prolínají, nebo na sebe navazují - harmonizace (především formální stránky termínů) je předpokladem pro automatické mapování a naopak mapování může být předpokladem harmonizace (především z hlediska obsahové stránky termínů). Jiným způsobem lze říci, že harmonizace bez mapování nemá smysl a stejně tak mapování bez harmonizace. V Metodice proto oba termíny nebudeme důsledně rozlišovat, pouze ve specifických případech, kdy se popisované metody a postupy budou týkat pouze jednoho z nich.

1.3.1 Postup při harmonizaci/mapování terminologických zdrojů

Při plánování projektů harmonizace a mapování se vždy vychází z potřeb uživatelů, tj. laické i odborné veřejnosti. Uživatelské požadavky na sjednocení nebo propojení heterogenních terminologických zdrojů jsou zpravidla podmíněny náročností procesu vyhledávání v těchto heterogenních zdrojích. Uživatelé potřebují najít dokumenty a objekty související s daným tématem, osobou, místem atd., které je zajímá. Rozličnost terminologických zdrojů používaných při popisu dokumentů a objektů uživatelům vyhledávání značně komplikuje. Cílem mapování a harmonizace je poskytnout standardizovaný jednotný přístupový bod k požadovaným informacím.

Proces harmonizace nebo mapování je zahájen výběrem terminologických zdrojů používaných v paměťových institucích. Následuje analýza zdrojů z hlediska struktury zdroje, formy vybraných termínů, poté výběr metod porovnávání terminologických zdrojů (resp. jednotlivých termínů v nich zastoupených).

1.3.2 Metody porovnávání terminologických zdrojů

Při mapování/harmonizaci terminologických zdrojů lze využít tyto základní způsoby porovnání termínů v nich obsažených:

¹⁵ Oxford dictionaries [online].

1. automatické porovnávání - s využitím softwarových nástrojů,
2. intelektuální porovnávání - založené především na vyhodnocování shody termínů na základě individuálního posouzení termínů zpracovatelem, informačním specialistou nebo doménovým expertem,
3. poloautomatické porovnávání - systém generuje potenciální kandidáty určené k mapování/harmonizaci, které jsou poté intelektuálně posouzeny a vyhodnoceny.

V současné praxi se nejčastěji používá třetí varianta. Lze konstatovat, že podíl intelektuálního porovnávání závisí na specifičnosti oborového slovníku.

Z hlediska rozsahu porovnávaných objektů rozlišujeme 3 základní skupiny metod porovnávání terminologických zdrojů:

1. metody na lexikální úrovni - do porovnávání jsou zahrnuty pouze samotné termíny,
2. metody na konceptuální úrovni - do porovnávání jsou zahrnuty termíny a jejich okolí/vztahy (nadřazené, podřazené termíny atd.),
3. metody založené na instancích - do porovnávání jsou zahrnuty termíny a dokumenty/objekty, v popisu kterých se vyskytují.

Metody porovnávání na lexikální úrovni

Lexikální metody jsou zaměřeny na určování míry podobnosti textových řetězců, tj. termínů z terminologických zdrojů. V ideálním případě najdeme stejná slova, či slovní spojení (termíny, řetězce znaků) použité v obou porovnávaných zdrojích.

Tab. 2: Příklad shodných termínů v třech zdrojích při porovnání na lexikální úrovni

Pojem	Identifikátor		
	Národní autority	EuroVoc	AGROVOC
politika	PSH8309	131558	6062
informatika	PSH6548	100223	3864
přírodní vědy	PSH11969	141614	4318
trh práce	PSH1264	142776	28713

V jiných případech jde zase o použití synonym, kdy v různých zdrojích jsou významově stejné pojmy vyjádřeny jinými výrazy. Ve vícejazyčných slovnících, či tezaurech, může být pojítkem těchto výrazů právě jejich stejný cizojazyčný ekvivalent (nejčastěji anglický termín). V jiných případech může být určující např. notace klasifikačního systému.

Tab. 3: Příklad různých termínů ve třech zdrojích se stejným anglickým ekvivalentem (při porovnání na lexikální úrovni)

Společný anglický pojem	Národní autority	EuroVoc	AGROVOC
water pollution	znečištění vody	znečišťování vody	znečištění vody
public health	veřejné zdraví	zdraví veřejnosti	zdravotnictví
animal husbandry	chov hospodářských zvířat	chov hospodářských zvířat	chov zvířat

K porovnávání textových řetězců (termínů), které nejsou zcela identické lze využít metrik pro hledání podobnosti, jako jsou Normalizovaná Levensteinova vzdálenost (Normalized Levenstein

Distance), Jaro-Winklerova vzdálenost (Jaro-Winkler Distance) nebo níže uvedená Graph Edit Distance.

Tab. 4: Lexikální srovnání podobně znějících výrazů

Výraz A	Výraz B	Normalized Levenshtein Distance	Jaro-Winkler Distance
politik	politika	0,875	0,983
politik	poliklinika	0,546	0,847
kantón	kanón	0,833	0,961
hrady	hady	0,800	0,940

Přesnost těchto metrik lze zvýšit rovněž určitým předzpracováním textových řetězců. Výběru způsobu předzpracování předchází analýza použitých termínů a způsobu jejich tvorby. Je nutno přihlížet ke specifickým rysům a morfologii přirozeného jazyka, který je v daném zdroji použit. Důležité je si uvědomit, že je veliký rozdíl při lematizaci anglických a českých termínů, jelikož v angličtině se vyskytuje pouze malé množství přípon, naopak v českém jazyce jako jazyce flektivním jsou přípony běžné. Lematizace (stemování) je způsob, kdy se při hledání shody využívá základ slova nebo slov, např..

akci*ové* společn*ost* → akci společnost
 akci*ová* společnost → akci společnost

V jiných případech může být nápomocna eliminace speciálních znaků, nebo slov (diakritika, stop slova¹⁶, interpunkční znaménka ap.) nebo též převod velkých počátečních písmen na malá.

V případě, že víceslovné termíny jednoho zdroje jsou tvořeny v přímém pořadí (společenské vztahy) a ve druhém zdroji se využívá inverze (vztahy společenské), využije se pro předzpracování termínů tokenizace.

Normalizovaná Levensteinova vzdálenost (Normalized Levenstein Distance): $dist_{lev}(s_i, s_j)$ je minimální počet operací (vkládání N_{ins} , mazání N_{del} nebo nahrazování N_{sub} jednoho znaku), které jsou nezbytné pro transformaci jednoho textového řetězce na jiný s ohledem na maximální délku obou těchto řetězců.

Formální zápis funkce vypadá následovně:

$$dist_{lev}(s_i, s_j) = \frac{N_{ins} + N_{del} + N_{sub}}{\max(|s_i|, |s_j|)}, dist_{lev} \in [0, 1]$$

Jestliže použijeme Levensteinovu vzdálenost, tak můžeme podobnost dvou výrazů definovat takto:

$$sim_{lev}(s_i, s_j) = 1 - dist_{lev}(s_i, s_j), \quad sim_{lev} \in [0, 1]$$

Čím vyšší je shoda termínů, tím více se výslední číslo blíží hodnotě 1.

¹⁶ Slova, která se v daném jazyce vyskytují často, ale nenesou žádnou významovou informaci, mají zpravidla pouze syntaktický význam. Typicky se jedná o spojky, předložky.

Tab. 5: Porovnání termínů pomocí Levensteinovy vzdálenosti

Výraz 1 (Authority)	Výraz 2 (EuroVoc, AGROVOC)	Podobnost
případová studie	případové studie	0,9375
znečišťování vody	znečištění vody	0,7647
chov hospodářských zvířat	chov zvířat	0,4400
veřejné zdraví	zdraví veřejnosti	0,1176

Jaro-Winklerova vzdálenost (Jaro-Winkler distance) je metrika, jejíž výpočet se realizuje pomocí koeficientu Dice distance¹⁷, přičemž se bere v úvahu poměr počtu společných znaků a délka řetězců. Metrika přihlíží rovněž na poměr společných znaků a transpozicí (výskyt opakujících se společných znaků) jakož i délku společného prefixu. Metrika je vhodná na měření vzdálenosti slov ve flektivních jazycích.

Podobně jako u předchozí metriky, i v tomto případě hodnota 1 znamená maximální shodu, hodnota 0 znamená žádnou shodu.

Tab. 6: Porovnání termínů pomocí Jaro-Winklerovy vzdálenosti

Výraz 1 (Authority)	Výraz 2 (EuroVoc, AGROVOC)	Podobnost
případová studie	případové studie	0,98330
znečišťování vody	znečištění vody	0,93034
chov hospodářských zvířat	chov zvířat	0,79758
veřejné zdraví	zdraví veřejnosti	0,00000

Metody porovnávání na konceptuální úrovni

Porovnání pomocí vztahů je většinou nezbytné v případě specializovaných tezaurů a řízených slovníků, ve kterých se vyskytují nespecifikované termíny, jejichž význam je dán kontextem celého tezauru nebo řízeného slovníku. Metody na konceptuální úrovni jsou důležité též při ověřování správnosti nalezené shody na lexikální úrovni.

Metody porovnávání na konceptuální úrovni vycházejí ze vztahů termínů - tj. nadřízený termín (broader term - BT), podřízený termín (narrower term - NT), související termín (related term - RT), nepreferovaný termín (use for - UF), preferovaný termín (use - USE).

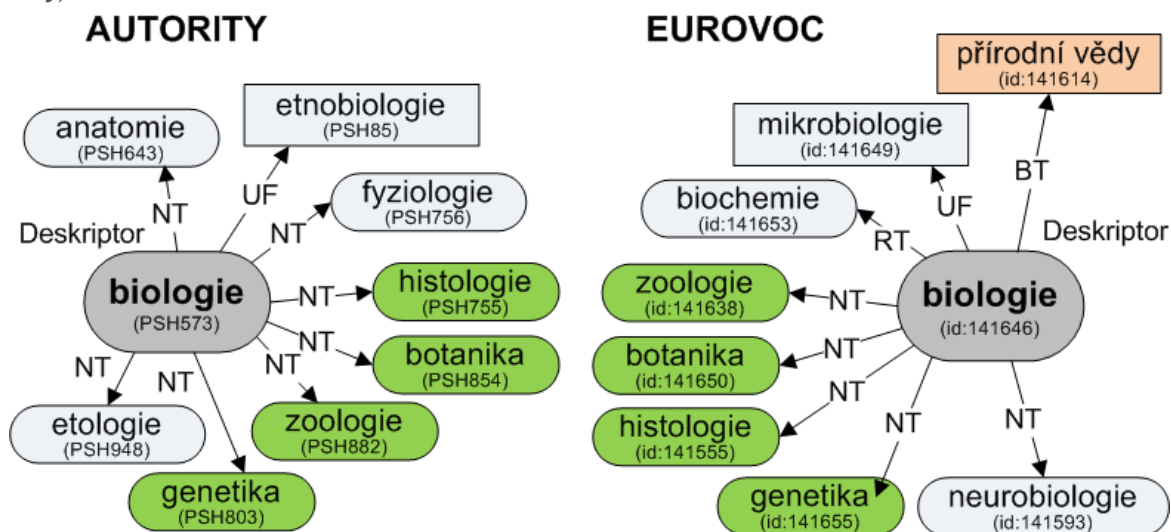
Podobnost se může hledat minimálně dvěma způsoby:

- metodou společných grafů,
- metodou vektorové podobnosti.

Metoda společných grafů hledá stejné nebo velmi podobné podgrafy, kde uzly představují pojmy tezauru a hrany představují výše uvedené vztahy.

¹⁷ Dice koeficient je metrika, která porovnává řetězce podle množství společných dvojic po sobě následujících znaků.

Obr. 6: Ukázka podgrafů pro výraz „biologie“ ze dvou zdrojů (zeleně jsou označeny společné uzly)



Na měření podobnosti pojmů v této (grafově) reprezentaci termínů se využívá metrika Graph Edit Distance (GED). Metrika určuje nejmenší počet uzlů a vztahů, které je nutno přiřadit, vymazat nebo nahradit, aby se graf g_1 rovnal grafu g_2 . Výpočetní složitost takto definované metriky je velmi vysoká, proto byly navrženy tři varianty GED:

- pojmová podobnost (Conceptual similarity),
- vztahová podobnost (Relational similarity),
- grafová podobnost (Graph similarity).

Jestliže máme dva grafy g_1 a g_2 . Tak nech $g_c = mcs(g_1, g_2)$, je maximální společný podgraf grafů g_1 a g_2 . GED definuje podobnost mezi grafy jako kombinaci hodnot jejich pojmové podobnosti s_c a vztahové podobnosti s_r .

Pojmová podobnost s_c určuje, kolik pojmů mají dva grafy g_1 a g_2 společných:

$$s_c = \frac{2n(g_c)}{n(g_1) + n(g_2)}$$

kde funkce $n(g)$ určuje počet společných pojmů (uzlů) v grafu g .

Vztahová podobnost s_r udává, jak podobné jsou vztahy mezi stejnými pojmy v obou grafech:

$$s_r = \frac{2m(g_c)}{m_{g_1}(g_1) + m_{g_2}(g_2)}$$

kde $m(g_c)$ je počet hran (vztahů) v grafu g_c a $m_{g_i}(g_i)$ je počet hran v bezprostředním sousedství s grafem g_c . (Bezprostřední sousedství $g_c \subseteq g_i$, $v g_i$ pozůstává z hran g_i kde alespoň jeden konec hrany přináleží g_c).

Grafová podobnost sim_g

Jestliže máme graf g , který chceme mapovat (propojit) s grafem g_T , pak můžeme metriku grafové podobnosti vyjádřit jako kombinaci předchozích dvou metrik:

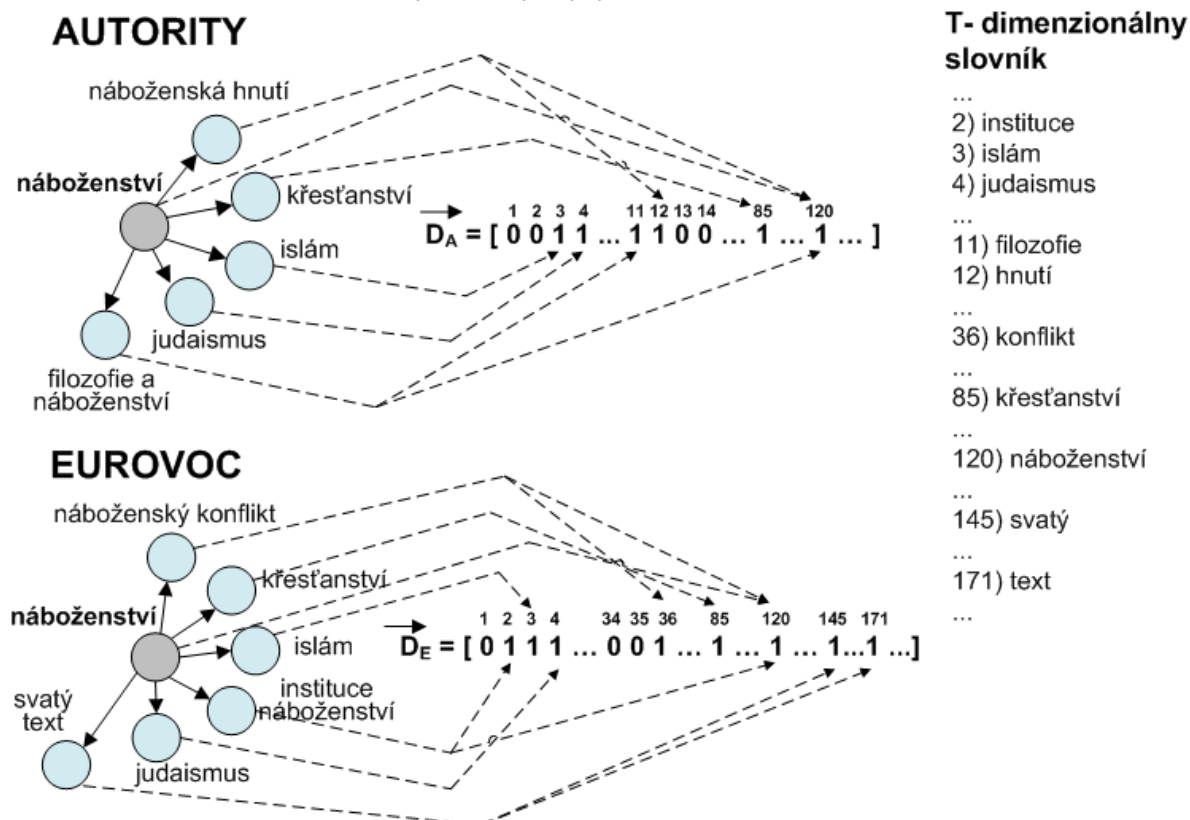
$$sim_g = \frac{N_c(g, g_T) + E_c(g, g_T)}{N(g_T) + E(g_T)}$$

kde $N_c(g, g_T)$ je počet společných uzlů v grafu g a g_T ; $E_c(g, g_T)$ je počet společných hran mezi g a g_T ; $N(g_T)$ počet uzlů v grafu g_T ; $E(g_T)$ počet hran v grafu g_T .

Metoda vektorové podobnosti

Metoda srovnává vektory, které reprezentují pojmy z harmonizovaných tezaurů. Jestliže máme vektor d binárních vstupů skládající se ze samotného termínu, příbuzných a propojených termínů, tak nejdříve z cílového tezauru sestojíme slovník 'normalizovaných' termínů. Když T je dimenze daného slovníku, pak jsou pojmy ze zdrojového (q) i cílového (d) tezauru reprezentované jako T -dimenzionální vektor ($d = [x_1, x_2, \dots, x_T]$). Prvek vektoru x_i představuje přítomnost/nepřítomnost daného i -tého termínu mezi termíny, které charakterizují pojem d (viz Obr. 7).

Obr. 7: T -dimenzionální vektor reprezentující pojem z tezauru



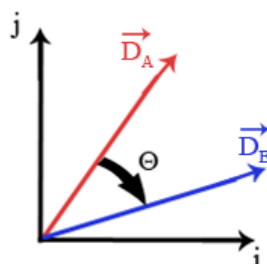
Podobnost pojmů (vyjádřených v podobě binárních vektorů), může být měřena pomocí kosinové vzdálenosti. Formálně to vyjádříme takto:

$\vec{D}_A$ a $\vec{D}_E$ jsou dva binární vektory $[w_1, w_2, \dots, w_{|T|}]$ výrazů ze zdrojového a cílového tezauru. Podobnost na základě kosinové vzdálenosti mezi $\vec{D}_E$ a $\vec{D}_A$ je

$$\text{sim}(\vec{D}_A, \vec{D}_E) = \frac{|\vec{D}_A \times \vec{D}_E|}{|\vec{D}_A| \cdot |\vec{D}_E|}$$

kde $|\vec{D}_A|$ a $|\vec{D}_E|$ jsou normy vektorů reprezentujících výrazů.

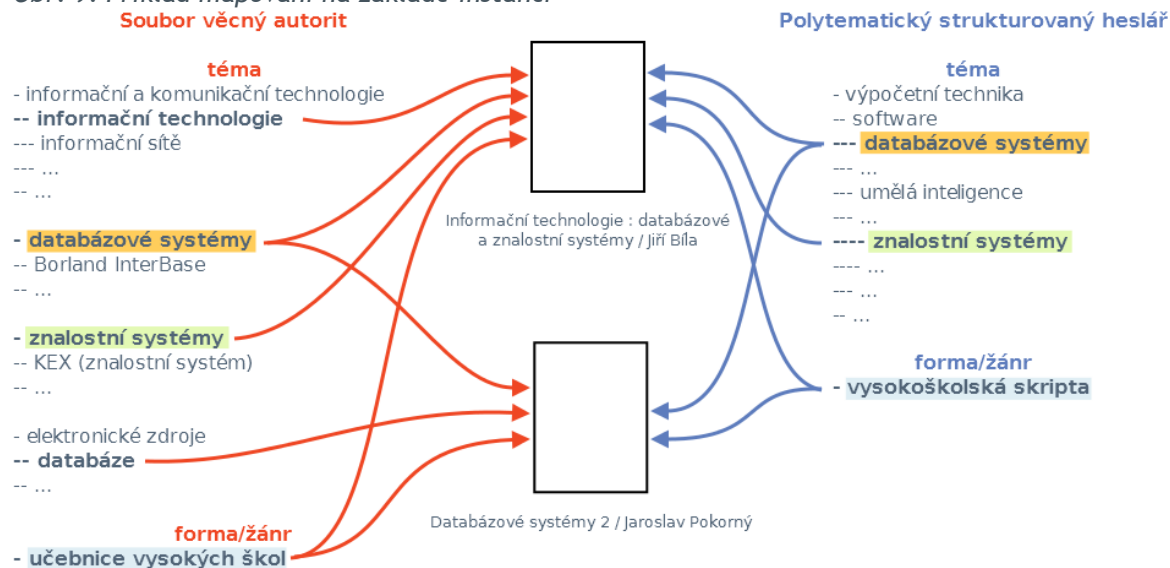
Obr. 8: Kosinová vzdálenost dvou výrazů



Metody porovnávání na základě instancí

Metody z této skupiny porovnávají společnou množinu objektů ze dvou nebo více digitálních sbírek, přičemž objekty každé sbírky jsou označeny, klasifikovány pojmy jednoho tezauru, slovníku. Tímto způsobem se na základě lexikálních a sémantických podobností hledají ekvivalentní, nadřazené, podřazené a související pojmy mezi zdrojovým a cílovým tezaurem, slovníkem.

Obr. 9: Příklad mapování na základě instancí



Výsledky porovnání slovníků (termínů ze slovníků) jsou důležité při rozhodování, zda jsou termíny shodné, či nikoli a zda je možné uvedené termíny propojit.

Je třeba zdůraznit, že uvedené metody porovnávání se využívají při automatickém porovnávání a jsou součástí specializovaných programů.

1.3.3 Metody mapování/harmonizaci terminologických zdrojů

Pokud na základě analýzy a porovnání terminologických zdrojů přistoupíme k mapování/harmonizaci, máme na výběr dvě základní metody:

- metoda korelace,
- metoda federace.

Při porovnávání konkrétních termínů je obě metody možné kombinovat a využívat současně.

Metoda korelace

Tato metoda je založena na usouvztažnění termínů pocházejících z jednotlivých zdrojů na základě jejich porovnání. Spočívá v přiřazení jednoho termínu k ekvivalentnímu termínu pocházejícímu z jiného zdroje. Pokud je nezbytné, chybějící termín z jednoho zdroje se do druhého doplní. Tento postup je vhodný a možný pro mapování/harmonizaci zdrojů, které nejsou tematicky příliš vzdálené a jsou na stejné úrovni specifikace.

Metoda federace

Metoda federace znamená uvádění odborných termínů pod jedním obecnějším univerzálním termínem. V případě INTERPI se využití této metody předpokládá u vysoce specializovaných termínů (např. regionálních variantách označení pro stavby lidové architektury apod.), které se propojí s obecným termínem.

Při aplikaci obou metod je výhodou, pokud zdroje obsahují například notaci MDT, která je jazykově nezávislá a termín obvykle zařazuje do širší kategorie. Mapování/harmonizace na základě MDT je tak obvykle dostatečně efektivní i při automatickém zpracování.

1.3.4 Technologické nástroje pro reprezentaci terminologických zdrojů

V procesu mapování/harmonizace jsou důležité taktéž otázky reprezentace terminologických zdrojů ve formátech, které umožní jejich strojové zpracování. Pro reprezentaci tezaurů a řízených slovníků se nejčastěji využívají formáty SKOS¹⁸. Představují aplikaci RDF¹⁹ specializovanou na reprezentaci systémů organizace poznání takovým způsobem, který je vhodný pro sémantický web.

V projektu INTERPI se počítá s využitím formátu SKOS pro prezentaci entit ze třídy obecný pojem. Prezentace jiných terminologických zdrojů určených k mapování/harmonizaci by proto rovněž měla být realizována ve formátu SKOS.

V případě ostatních tříd entit bude záležet na požadavcích pro rozsah dat reprezentovaných ve vybraném formátu. Pokud by bylo nutné zachovat maximum informací o entitě, bude nutné

¹⁸ Pro více informací viz: <http://www.w3.org/2004/02/skos/intro>.

¹⁹ Pro více informací viz: <http://www.w3.org/RDF/>.

využít proprietární formát, který umožní prezentovat specifické typy vztahů a událostí. Informace o struktuře a principech proprietárního formátu INTERPI budou zveřejněny jako součást modelu kooperace nebo jako součást dokumentace poloprovozu.

2 Pozice a význam národních autorit

Soubor národních autorit tvoří terminologický základ znalostní databáze 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 RDA²⁰.

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.

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;

²⁰ Pro více informací viz stránka Národní knihovny s přehledem interpretací pravidel RDA: <http://www.nkp.cz/o-knihovne/odborne-cinnosti/zpracovani-fondu/katalogizacni-politika/rda>.

- 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 označující 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.

2.1 Charakteristika souboru 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 tezaurů 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. 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).

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, musí

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.).

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

2.2 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 deskripty i nedeskripty 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

Počitatelná substantiva se uvádějí v plurálu (*učitelé*), abstrakta, počitatelná 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*).

2.3 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 varianční 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.

2.4 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.

2.5 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 vědných oborech;

²¹ Denotát je charakterizován jako „vztah mezi výrazem a tím co označuje. Denotátem jedinečného jména je individuální předmět, denotátem obecného jména třída předmětů“ JONÁK, Zdeněk. Denotát. In *KTD* [online databáze].

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.

3 Aplikace metod mapování/harmonizaci slovníků v projektu INTERPI

Základem pro mapování/harmonizaci slovníků pro projekt INTERPI byl zvolen soubor národních autorit a to z toho důvodu, že je oborově nezávislý, má strukturu tezauru a obsahuje notace MDT, což umožňuje identifikovat kontext nejednoznačných termínů (viz kapitola 2.1).

3.1 Očekávané problémy specializovaných terminologických zdrojů

Očekávané problémy, které mohou vyplynout ze specifických rysů terminologických zdrojů, lze obecně shrnout v následujících bodech:

1. problémy s významem termínů - předpokládá se, že terminologické zdroje budou obsahovat:
 - příliš specifické termíny, které se v databáze INTERPI nenacházejí,
 - jednoslovné termíny obecného významu (*operace*),
 - významově málo specifické termíny (*život a dílo*),
 - příliš obecné, nerozlišené termíny používané namísto užších termínů - je potřeba poznamenat, že tento nedostatek lze odhalit pouze porovnáním na základě instancí (viz kapitola 1.3.2);
2. problémy s formou termínů - terminologické zdroje mohou využívat při tvorbě termínů:
 - jednotné/množné číslo,
 - inverzi/přímé pořadí slov,
 - velká/malá písmena,
 - jinou pravopisnou normu,
 - řetězce předmětových hesel (to znamená spojení více obecných hesel pro vyjádření specifického tématu).

Uvedené problémy lze eliminovat vhodným výběrem metod porovnání, předzpracování termínů a vhodných metod harmonizace/mapování.

3.2 Postup mapování/harmonizace terminologických zdrojů na znalostní bázi INTERPI

Cílem INTERPI není skutečné namapování všech oborových slovníků z oblasti paměťových institucí na jeden společný slovník, ani vytvoření jednoho univerzálního slovníku. Toto je především v kompetenci a v zájmu jednotlivých odborných komunit. Cílem projektu INTERPI je poskytnout nástroje, pomocí kterých je možné i vysoce specializované slovníky harmonizovat (mapovat) s obecnějšími tak, aby bylo možné propojit navzájem informační zdroje a informační objekty zpracovávané, uchovávané a zpřístupňované v jednotlivých sbírkách paměťových institucí.

Proces mapování slovníků zahrnuje tyto kroky:

1. výběr terminologického základu - tj. výběr slovníku nebo souboru termínů, které budou tvořit základnu pro mapování,
2. výběr a realizace formální reprezentace terminologického základu - např. SKOS,
3. zpracování charakteristiky lexikálních jednotek v terminologickém základu - např. používání jednotného/množného čísla, invertované/přímé pořadí...,
4. výběr a analýza zdrojových slovníků, jejichž termíny se budou mapovat na terminologický základ,
5. výběr vzorku termínů ze zdrojových slovníků a jejich reprezentace ve vybrané struktuře,
6. výběr metod porovnávání, proces porovnávání a vyhodnocení výsledků,
7. harmonizace/mapování terminologický zdrojů.

Obsahují-li zdrojové slovníky předmětové řetězce (např. města - Česko), předzpracování spočívá v oddělení jednotlivých částí řetězce. Porovnávat se budou prioritně první části řetězců, protože se předpokládá, že jsou obsahově nejvýznamnější. Ze souborů termínů, které tvořily další části řetězců je potřeba odstranit nevýznamná slova a spojení (např. život a dílo, použití). Vzniklý soubor termínů může být rovněž předmětem mapování/harmonizace, ale je pravděpodobné, že bude vykazovat znaky typově nespécifikovaného rejstříku - tj. bude obsahovat obecné i individuální pojmy.

V případě, že se analýzou zdrojového slovníku (především v případě heslářů a rejstříků) zjistí, že obsahují kromě obecných pojmů také individuální pojmy, bude potřeba přistoupit k filtrování obecných pojmů. Při filtrování lze postupovat intelektuálně - instituce, která rejstřík poskytl, jej zkontroluje, každý termín posoudí a vyčlení ty, které reprezentují obecné pojmy. Automatické filtrování lze provést na základě porovnání termínů se znalostní databází INTERPI. Předpokladem automatického filtrování je zpřístupnění rejstříku ve strojově čitelné proprietární struktuře. Jednotlivé termíny budou porovnány s databází INTERPI a označeny třídou entity (pokud se v databázi vyskytnou).

Následuje porovnání, které je založené na vybraných metodách popsanych v kapitole 1.3.2. Metody, které vyžadují zpřístupnění dalších objektů kromě termínů, je možné aplikovat při dodržení těchto podmínek:

- metoda porovnávání na konceptuální úrovni - zdrojový slovník musí mít dostupné informace o vztazích termínů;
- metoda porovnávání na základě instancí - pro zdrojový slovník musí být k dispozici databáze dokumentů, které využívají zdrojový slovník při popisu. Databáze dokumentů musí být přístupná pomocí standardizovaných nástrojů (např. protokol Z39.50, popsané webové služby, atp.). Současně se předpokládá, že databáze dokumentů obsahuje publikované dokumenty, u kterých je pravděpodobné, že se nacházejí v databázích spravovaných Národní knihovnou.

3.3 Hodnocení shody termínů a výsledky porovnávání

Je třeba si uvědomit, že výsledkem aplikací metod porovnávání termínů není automaticky mapovaný/harmonizovaný slovník daného terminologického zdroje. Výsledek obvykle obsahuje informaci o předpokládaném spojení mezi terminologickými zdroji.

Jako první se používá metoda porovnání na lexikální úrovni. Shodu termínů lze z hlediska jeho preference/nepreference v terminologickém základu nebo ve slovníku daného terminologického zdroje hodnotit v následujících stupních:

1. preferovaná (nebo jediná) forma termínu ze slovníku daného terminologického zdroje se shoduje s preferovaným termínem z terminologického základu,
 - výsledkem procesu je: termín ze slovníku daného terminologického zdroje, shodný preferovaný termín z terminologického základu, identifikační číslo záznamu pojmu z terminologického základu;
2. preferovaná (nebo jediná) forma termínu ze slovníku daného terminologického zdroje se shoduje s variantním termínem z terminologického základu,
 - v tomto případě je výsledkem procesu: termín ze slovníku daného terminologického zdroje, shodný variantní termín z terminologického základu, identifikační číslo záznam pojmu z terminologického základu, preferovaný termín pojmu z terminologického základu;
3. nepreferovaná forma termínu ze slovníku daného terminologického zdroje se shoduje s preferovaným termínem z terminologického základu,
 - výsledkem procesu je: nepreferovaný termín ze slovníku daného terminologického zdroje, preferovaný termín ze zdrojového slovníku, shodný preferovaný termín z terminologického základu, identifikační číslo záznamu pojmu z terminologického základu;
4. nepreferovaná forma termínu ze slovníku daného terminologického zdroje se shoduje s nepreferovaným termínem z terminologického základu,
 - výsledkem procesu je: nepreferovaný termín ze slovníku daného terminologického zdroje, preferovaný termín ze zdrojového slovníku, shodný preferovaný termín z terminologického základu, identifikační číslo záznamu pojmu z terminologického základu, preferovaný termín pojmu z terminologického základu;
5. bez shody.

V případě, že slovníkem daného terminologického zdroje je řízený slovník nebo thesaurus, který obsahuje vymezené vztahy mezi termíny, je vhodné především na termíny s nejvyšším stupněm shody (1) uplatnit metodu založenou na porovnávání konceptů, aby se vyvrátila náhodná shoda řetězců a dokázala také obsahová shoda. Samozřejmě použití metod založených na porovnání konceptů není vyloučeno ani při dalších stupních shody (2-4), je pouze vhodné zvážit náročnost hodnocení a očekávanou přesnost výsledku a tím efektivitu takového hodnocení.

3.4 Výsledky mapování/harmonizace

Na základě výsledků porovnání termínů lze přistoupit k harmonizaci/mapování terminologických zdrojů. V případě, že termíny s nejvyšším stupněm shody (1) byly také ověřeny metodou založenou na porovnávání konceptů, je možné termíny harmonizovat/mapovat automaticky. V jiných případech se při mapování/harmonizaci postupuje manuálně. Mapování/harmonizaci by

měli vykonávat pracovníci ze specializovaného pracoviště, které poskytlo slovník daného terminologického zdroje.

Mapování spočívá především ve vložení *identifikačního čísla pojmu/termínu* ze slovníku daného terminologického zdroje do záznamu entity v znalostní databázi INTERPI. Identifikační číslo se vkládá spolu s označením zdroje (viz Metodika tvorby znalostního modelu). Další kroky se odvíjejí od stupně shody.

V případě, že *preferovaný termín* ze slovníku daného terminologického zdroje je shodný s *preferovaným termínem* ze znalostní databáze INTERPI, je vhodné k preferovanému termínu doplnit zkratku specializovaného slovníku a označit tím, že jeho forma vyhovuje pravidlům INTERPI i specializovaného slovníku.

V případě shody na stupni 2 (tj. *preferovaný termín* ze slovníku daného terminologického zdroje je shodný s *nepreferovaným termínem* ze znalostní databáze INTERPI) je na rozhodnutí zpracovatele, zda bude akceptovat preferovaný termín INTERPI jako preferovaný termín specializovaného slovníku (původní preferovaný termín zdrojového slovníku je možné uchovat jako variantní termín v záznamu INTERPI, pokud se tam již nevyskytuje). Nebude-li preferovaný termín akceptovat, může vložit další preferovaný termín s označením specializovaného slovníku - zůstane tak zachováno jeho specifikum. Podobně se postupuje i při stupni shody 3 (*nepreferovaný termín* ze slovníku daného terminologického zdroje je shodný s *preferovaným termínem* ze znalostní databáze INTERPI).

V případě, že se *nepreferovaný termín* ze slovníku daného terminologického zdroje shoduje s *nepreferovaným termínem* ze znalostní databáze INTERPI (stupeň shody 4) je pravděpodobné, že jsou termíny využívány v jiném kontextu - pokud to tak není, je možné podobně jako v předešlých případech akceptovat preferovaný termín INTERPI nebo doplnit nový preferovaný termín podle specializovaného slovníku.

V případě, že záznam termínů ve slovníku daného terminologického zdroje obsahuje také variantní/nepreferované termíny, lze je doplnit do záznamu entity. Je na rozhodnutí zpracovatele, uzná-li variantní termíny jako obecně platné, nebo je vymezí pouze pro specializovaný slovník.

3.5 Doporučení pro přejímání nových termínů do databáze INTERPI

Přejímání nových termínů do databáze INTERPI není vyloučeno a to ani v případě, že se jedná o vysoce specializované termíny. Pouze je doporučeno využít metodu federace - tj. vytvořit propojení mezi vysoce specializovaným termínem a nejbližším obecným termínem, který se dá pro jeho označení využít.

Při přejímání nových termínů do databáze INTERPI je třeba se zaměřit na následující problematiku okruhy.

3.5.1 Řešení homonymie, synonymie a polysémie

Ideální stav, podle kterého by každý pojem měl jedno označení/pojmenování, se v reálné situaci nevyskytuje. Často se v procesu mapování/harmonizace budeme setkávat se situací, kdy jeden pojem může být označen více termíny, hovoříme o synonymii; naopak jeden termín může označovat více pojmů, v tomto případě hovoříme o homonymii nebo polysémii.

Synonyma bývají označována jako slova souznačná, tj. slova nebo slovní spojení se vzájemně stejným nebo podobným významem, např. *terapie pohádkou* a *pohádkoterapie*. Problém synonymie se v terminologických zdrojích řeší uvedením variantní formy jako odkazu viz, např. preferovaný termín zvolíme *terapie pohádkou* a termín *pohádkoterapie* odkážeme jako nepreferovaný.

Homonymum je slovo nebo jiná jazyková jednotka, popř. její tvar (morfém, gramatický tvar, slovo, slovní spojení, věta), znějící nebo psaná stejně, ale různého významu. Ve většině případů jde o slova souzvučná mající rozdílný význam, po významové stránce je souvislost mezi nimi čistě náhodná, např. *kolej* (jako ubytování pro studenty) a *kolej* (jako součást drážního svršku). Od homonymních slov se odlišují **polysemní slova**, tj. slova stejně znějící, mající různý význam, ale mezi jejich významy existuje nějaká souvislost, např. termín *operace* označuje ve všech kontextech nějakou akci.

Různé významy homonymních a polysémních slov se v terminologických zdrojích musí rozlišit specifikací termínu pomocí adjektiva, např. pro homonymum *operace* se použije: *binární operace*, *bankovní operace*, *bojové operace*, *operace srdce*; případně uvedením kvalifikátoru (relátoru) v závorce, který je součástí deskriptoru, např. *morfologie (biologie)*, *morfologie (lingvistika)*.

Doporučení pro harmonizaci/mapování v projektu INTERPI

Toto zjednotnění a upřesnění významu významově nejednoznačných slov potřebnou specifikací daného termínu bude efektivně využito při harmonizaci a mapování terminologických zdrojů.

3.5.2 Sousedství - víceslovné termíny

Ideálním reprezentantem termínu je *jednoslovné substantivum*. V jednotlivých vědních oborech však narůstají požadavky na pojmenování/označení nových pojmů. Protože počet slov v jazyce je limitován, často pojmenování nových pojmů vznikají kombinováním stávající slovní zásoby. Vznikají tak *sousloví*, ustálená slovní pojmenování, víceslovné termíny. Nejčastějším typem sousloví je spojení podstatného jména s přívlastkem shodným (*český jazyk*, *kočka domácí*, ...) nebo s přívlastkem neshodným (*ministerstvo zdravotnictví*, *postavení mimo hru*, *skok do dálky*, ...).

Po významové stránce představuje sousloví ustálené několikaslovné pojmenování, které má význam jako celek a označuje jeden pojem. Jednotlivá slova mají často pozměněný nebo

obrazný význam a není je možné nahrazovat synonymy. Sousloví se vyznačují se ustáleným slovosledem, např. školní budova, otakárek fenýklový.

Víceslovné termíny představují sémantickou jednotku s jediným jí vlastním významem. Nelze je plnohodnotně nahradit formálním rozkladem na jednotlivé komponenty. Proto se nyní upouští od tohoto formálního rozkladu víceslovných termínů (sousloví) dříve doporučovaného normou ČSN 01 0188 a striktně v případě předmětových hesel dodržovaného, takže např. „psychologie osobnosti“ se zachovává a nenahrazuje se dvěma významově širšími termíny („osobnost“ „psychologie“).

Nicméně se tyto prvky v rejstřících stále vyskytují, musí jim být proto při harmonizaci věnována patřičná pozornost.

Sousloví může být v hovorovém jazyce nebo ve slangu staženo do jednoslovného pojmenování, např. obývací (obývací pokoj), nákladák (nákladní auto), liduška (lidová škola umění). Tyto prvky jsou z procesu harmonizace či mapování vyloučeny.

Doporučení pro harmonizaci/mapování v projektu INTERPI

Sousloví, víceslovné termíny vyjadřují adekvátním způsobem daný pojem, jsou významově přesnější, jednoznačnější. Problém působí jejich jednoznačná identifikace, jejich rozpoznání je někdy velmi složité. K identifikaci sousloví v terminologických zdrojích v projektu INTERPI poslouží porovnání s terminologickým základem, tedy se souborem tematických termínů (SVAT).

3.5.3 Výskyt termínů obsahujících etnická adjektiva

V terminologických zdrojích paměťových institucí se vyskytují *víceslovné termíny*, v nichž substantivní základ je zpřesněn *etnickým adjektivem*, např. „čeští prezidenti“, čeští legionáři“ apod. V souboru věcných tematických autorit se tento typ termínů vyskytuje pouze omezeně, a to tehdy je-li tento termín všeobecně uznáván, např. „české baroko“ a dále v případech, kdy je, byť implicitně, vázán na národní jazyk, případně se vztahuje k oblasti umění, např. čeští spisovatelé, české výtvarné umění...

Pro potřeby paměťových institucí bude možnost integrovat víceslovné termíny obsahující etnická adjektiva zachována.

Doporučení pro harmonizaci/mapování v projektu INTERPI

Víceslovné termíny obsahující etnická adjektiva nebudou při harmonizaci/mapování vyloučeny, budou zachovány a intelektuálně posouzeny.

3.5.4 Výskyt „situačních“, „lokálních termínů“ v terminologických zdrojích

Za situační či lokálně se vyskytující termíny považujeme *termíny*, které jsou *příliš detailní*, specifické pro daný kontext, dané prostředí, danou instituci, daný dokument. Charakterizují specifický fenomén, souvisejí s jedním informačním zdrojem a s lokálním prostředím, např.

„literární tvorba olomoucká“, „významné osobnosti města Jaroměřic“. Tato slovní spojení nepovažujeme za ustálená, obsahují dva odlišné sémantické vrcholy „literární tvorba“ + Olomouc; „významné osobnosti“ + Jaroměřice.

Doporučení pro harmonizaci/mapování v projektu INTERPI

Při harmonizaci/mapování budou tyto prvky identifikovány porovnáním s terminologickým základem (SVAT) a budou jako situační termíny z první fáze procesu mapování vyloučeny.

Výjimku tvoří termíny, které by z pohledu univerzální databáze mohly být považovány za situační, pro odbornou komunitu v dané sémantické doméně však představují specifický odborný termín, který by měl být jako takový do univerzální databáze zařazen. Jde o termíny typu „olomoucké baroko“, „Sèvreský porcelán“. Podobné případy budou řešeny individuálně a v diskusi s odborníky.

3.5.5 Formální aspekty termínů používaných v terminologických zdrojích paměťových institucí

Jednotné / množné číslo

V knihovnických terminologických zdrojích se termíny vyskytují v množném čísle, zatímco termíny uvedené v číselnících a rejstřících muzeí, galerií a v některých případech i archivů se vyskytují v jednotném čísle. Tato rozdílná praxe je daná anglosaskou tradicí.

Bylo dohodnuto, že v integrované znalostní bázi INTERPI se budou, pokud to bude možné, respektovat racionální požadavky všech paměťových institucí; bylo tedy přijato kompromisní řešení, že do souboru SVAT, který tvoří základ pro harmonizaci a mapování tematických rejstříků, budou doplněny tvary termínů i v jednotném čísle tak, aby tento soubor vytvářel adekvátní nástroj pro kooperativní tvorbu, aplikaci i sdílení entit mezi všemi paměťovými institucemi.

Doporučení pro harmonizaci/mapování v projektu INTERPI

V souboru obsahujícím entity ze třídy entit obecný pojem budou doplněny tvary označení/pojmenování v jednotném čísle.

Nepravopisná forma zápisu

V rejstřících se vyskytují i termíny, jejichž zápis neodpovídá pravopisné normě, v některých případech je na ně navázáno i větší množství dokumentů, např. „archeologický výzkumy“. Tyto i další pravopisné prohřešky (1.světová válka), budou v rejstřících opraveny před započítáním procesu harmonizace.

Invertovaný slovosled

Víceslovné termíny používané v heslářích předmětových hesel a jejichž zápis se řídil normou ČSN 01 0188, budou před započítáním vlastní harmonizace/mapování tzv. normalizovány; jejich

tvár bude redukován na kořen/základ slova, bude uplatněna podmínka nerespektování pořadí slov.

Doporučení pro harmonizaci/mapování v projektu INTERPI

Nepravopisné termíny budou opraveny, invertovaná podoba víceslovných termínů bude normalizována. Do souboru tematických autorit budou doplněny invertované tvary víceslovných termínů, např. „vysoké školy“ (přirozený slovosled), školy vysoké (invertovaný slovosled) může být uveden jako variantní označení.

4 Seznam bibliografických odkazů

ABZ : *slovník cizích slov* [online databáze]. [cit. 2014-09-09]. Dostupné z WWW: <http://slovník-cizich-slov.abz.cz>.

BALÍKOVÁ, Marie. Klasifikační systém. In *KTD : Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online databáze]. Praha : Národní knihovna České republiky, 2003- [cit. 2014-09-09]. Dostupné z WWW: <http://aleph.nkp.cz/cze/ktd>.

BALÍKOVÁ, Marie. Lexikální jednotka. In *KTD : Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online databáze]. Praha : Národní knihovna České republiky, 2003- [cit. 2014-09-09]. Dostupné z WWW: <http://aleph.nkp.cz/cze/ktd>.

BALÍKOVÁ, Marie. Rejstřík. In *KTD : Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online databáze]. Praha : Národní knihovna České republiky, 2003- [cit. 2014-09-09]. Dostupné z WWW: <http://aleph.nkp.cz/cze/ktd>.

BALÍKOVÁ, Marie. Řízený slovník. In *KTD : Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online databáze]. Praha : Národní knihovna České republiky, 2003- [cit. 2014-09-09]. Dostupné z WWW: <http://aleph.nkp.cz/cze/ktd>.

BARTOLONI, Giacomo - FRANCESCONI, Enrico. Sharing knowledge by conceptual mapping: the case of EU thesaural interoperability. In *JURIX 2010 : legal knowledge and information systems*. December 2010.

BIASIOTTI, M.A. - FARO, S. - FRANCESCONI, E. Thesaurus mapping for promoting semantic interoperability of european public services. In *eChallenges e-2011 : conference proceedings* [online]. 2011 [cit. 2014-09-09]. Dostupné z WWW: <http://www.ittig.cnr.it/Ricerca/Testi/biasiotti-faro-francesconi2011.pdf>

Číselník. In *Wikipedie : otevřená encyklopedie* [online]. Aktualizováno 2014-07-25 [cit. 2014-09-09]. Dostupné z WWW: <http://cs.wikipedia.org/wiki/%C4%8C%C3%ADseln%C3%ADk>.

ČSN 01 0172 ISO 5964. *Pokyny pro vypracování a rozvíjení vícejazyčných tezaurů*. 1992.

ČSN 01 0193. *Jednoязыčný tezaurus. Složení, uspořádání a formální úprava*. 1988.

ČSN ISO 5127-1 (01 0167). *Dokumentace a informace. Slovník. Část 1, Základní pojmy*. 1993.

ČSN ISO 5127-2 (01 0183). *Dokumentace a informace. Slovník. Část 2, Tradiční dokumenty*. 1994.

ČSN ISO 5127-6 (01 0163). *Dokumentace a informace. Slovník. Část 6, Selekční jazyky*. 1994.

DOERR, Martin. Semantic problems of thesaurus mapping. In *Journal of digital information* [online]. 2001 [cit. 2014-09-09], vol. 1, no. 8. Dostupné z WWW: <https://journals.tdl.org/jodi/index.php/jodi/article/view/31/32>.

EUROPEANA. *Europeana Semantic Elements : specification and guidelines* [online]. 2013-07-14 [cit. 2014-09-09]. Dostupné z WWW: <http://pro.europeana.eu/documents/900548/2eee7beb-b9d8-4532-a089-8e8d6df38ce7>.

IFLA. *Functional requirements for subject authority data (FRSAD) : a conceptual model*. München : De Gruyter Saur, 2011. ISBN 978-3-11-025323-8. 74 s.

ISAAR (CPF) : *Mezinárodní standard pro archivní autoritní záznamy korporací, osob a rodů*. 2. vyd. Praha, 2009. Dostupné z WWW: <http://www.ica.org/download.php?id=1646>.

JONÁK, Zdeněk. Denotát. In *KTD : Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online databáze]. Praha : Národní knihovna České republiky, 2003- [cit. 2014-09-09]. Dostupné z WWW: <http://aleph.nkp.cz/cze/ktd>.

JONÁK, Zdeněk. Pojem. In *KTD : Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online databáze]. Praha : Národní knihovna České republiky, 2003- [cit. 2014-09-09]. Dostupné z WWW: <http://aleph.nkp.cz/cze/ktd>.

JONÁK, Zdeněk. Termín. In *KTD : Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online databáze]. Praha : Národní knihovna České republiky, 2003- [cit. 2014-09-09]. Dostupné z WWW: <http://aleph.nkp.cz/cze/ktd>.

JURČÁKOVÁ, Z. Heslár. In KATUŠČÁK, Dušan - MATTHAEIDESOVÁ, Marta - NOVÁKOVÁ, Marta. *Informačná výchova : terminologický a výkladový slovník*. Bratislava : Slovenské pedagogické nakladateľstvo, 1998. ISBN 80-08-02818-1.

KRAMER, Ralf - NIKOLAI, Ralf - HABECK, Corinna. Thesaurus federations : loosely integrated thesauri for document retrieval in networks based on internet technologies. In *International journal on digital libraries*. September 1997, vol. 1, no. 2, p. 132-152.

LAPPALAINEN, Mikko - FROSTERUS, Matias - NYKIRI, Susanna. Reuse of library thesaurus data as ontologies for the public sector. In *IFLA 2014* [online]. [cit. 2014-09-09]. Dostupné z WWW: <http://library.ifla.org/819/1/086-lappalainen-en.pdf>.

MALTESE, Vincenzo - HOSSAIN, Bayzid Ashik. *SAM : a tool for the semiautomatic mapping and enrichment of ontologies* [online]. Trento : University of Trento, 2012 [cit. 2014-09-09]. Dostupné z WWW: <http://eprints.biblio.unitn.it/3965/1/techRep007.pdf>.

Oxford dictionaries [online]. Dostupné z WWW: <http://www.oxforddictionaries.com/>.

RITZE, Dominique, - ECKERT, Kai. Thesaurus mapping : a challenge for ontology alignment? In *CEUR workshop proceedings* [online]. 2012 [cit. 2014-09-09], vol. 946 (Ontology matching). Dostupné z WWW: http://ceur-ws.org/Vol-946/om2012_poster8.pdf.

Slovník spisovné češtiny [online databáze]. Ústav pro jazyk český, 2011 [cit. 2014-09-09]. Dostupné na WWW: <http://ssjc.ujc.cas.cz/>.

VODIČKOVÁ, Hana. Řízený slovník. In *KTD : Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online databáze]. Praha : Národní knihovna České republiky, 2003- [cit. 2014-09-09]. Dostupné z WWW: <http://aleph.nkp.cz/cze/ktd>.

Zákon č. 365/2000 Sb., o informačních systémech veřejné správy a o změně některých dalších zákonů, ve znění pozdějších předpisů.