AiP Beroun s.r.o., http://www.aipberoun.cz, tel.: +420 311 611 237, fax: +420 311 611 238
Talichova 807, Beroun, 266 01; IČO: 25778943, DIČ: 25778943

# ZDOKONALENÍ VIRTUÁLNÍHO BADATELSKÉHO PROSTŘEDÍ MANUSCRIPTORIA - VYUŽITÍ TEZAURŮ A DATABÁZÍ AUTORIT AGREGOVANÝCH CERL PRO VYHLEDÁVÁNÍ V MANUSCRIPTORIU

## Pilotní řešení

**Zpráva ke smlouvě Smlouva o spolupráci ve výzkumu a vývoji**

**verze 1.0**

AiP Beroun, autor Mgr. Olga Čiperová

# Obsah

# 1 Úvod o dokumentu

AiP Beroun uzavřela s Národní knihovnou České republiky dne 10.11.2011 Smlouvu o spolupráci ve výzkumu a vývoji: Zdokonalení virtuálního badatelského prostředí Manuscriptoria - Využití tezaurů a databází autorit agregovaných CERL pro vyhledávání v Manuscriptoriu.

## 1.1 Účel

Tento dokument tvoří zprávu k vytvoření a implementaci pilotní aplikace, která využije data poskytovaná CERL Thesaurus za účelem vylepšení možností vyhledávání informací o historických dokumentech v MnS. Popisuje a zdůvodňuje výsledné pilotní řešení a poskytuje návod k užívání vytvořené aplikace.

## 1.2 Předpokládaný čtenář

Tento dokument je určen především pro Zadavatele (NK ČR) a pro Řešitele úkolu (AiP Beroun) jako popis pilotního řešení. Dále je tento dokument určen všem, kteří se podílejí na rozvoji projektu Manuscriptorium jako uživatelé.

## 1.3 Termíny a konvence

Termíny a konvence použité v tomto dokumentu, pokud zde nejsou přímo vysvětleny, jsou popsány a definovány v dokumentu [1].

- M-Tez Pracovní název popisované Pilotní aplikace.

## 1.4 Reference

V dokumentu se odkazujeme na následující literaturu:

[1] Manuscriptorium v.2.0 – analýza systému, AiP Beroun 2004

[2] Manuscriptorium v.2.0 – Databáze historických názvů měst a její použití v Manuscriptoriu, Analýza projektu, AiP Beroun 2007

[3] Tvorba a využití dalších (autoritních) databází – zpráva k Dodatku č. 7, AiP Beroun 2009

[4] Autoritní databáze a řízené slovníky: pilotní řešení s využitím tezauru historických názvů měst - Zpráva k Dodatku č.8, AiP Beroun 2010

# 1 Úvod

Výzkumný úkol "Využití tezaurů a databází autorit agregovaných CERL pro vyhledávání v Manuscriptoriu" navazuje na výzkumný úkol „Autoritní databáze a řízené slovníky (Pilotní řešení s využitím tezauru historických názvů měst)", který byl realizován jako část dodatku č. 8 z r. 2010 ke „Smlouvě o spolupráci ve výzkumu a vývoji".

V letech 2009 a 2010 bylo úkolem implementovat do vyhledávacího systému Manuscriptoria externí autoritní databáze a tezaury. Výsledky těchto výzkumných úkolů byly publikovány v letech 2009 [3] a 2010 [4]. V těchto projektech byly externí databáze a tezaury zapojovány přímo do vyhledávacích formulářů katalogu Manuscriptoria a tak bylo uživatelům nabídnuto rozšíření a kontrola jejich dotazů pomocí jednotlivých autoritních bází.

V roce 2010 byla do Manuscriptoria zařazena možnost vyhledávání v geografické bázi „Stará města", která byla vytvořena v oddělení rukopisů a starých tisků NK ČR jako doplněk pro standardizaci zápisu geografických názvů při katalogizaci v bázi „STT - Staré tisky a mapy NK ČR 1501-1800". Pilotní řešení umožnilo uživatelům u vybraných položek ve formuláři Manuscriptoria ověřit zadaný geografický termín v externí bázi a případně seznam termínů pro dotaz rozšířit pomocí vyhledaných výsledků.

Cílem letošního výzkumného projektu, v návaznosti na projekt „Autoritní databáze a řízené slovníky (Pilotní řešení s využitím tezauru historických názvů měst)", bylo rozšířit pro uživatele možnosti vyhledávání a ověřování geografických termínů v dalších databázích. Pro pilotní řešení byla jako zdroj dat vybrána báze CERL Thesaurus – Places. Tento geografický tezaurus spravuje Consortium of European Research Libraries (CERL). Jedná se o jednu z nejrozsáhlejších světových bází geografických termínů, která obsahuje kromě autoritních záhlaví i informace o variantních a dokonce i fiktivních názvech. Její záznamy významně rozšiřují údaje z báze Starých měst.

Obsah CERL Thesaurus – Places je doplňován institucemi, které jsou sdruženy ve výše jmenovaném konsorciu CERL. Záznamy, které jsou k dispozici v internetovém rozhraní CERL Thesaurus, je možné vyhledávat zadáním geografického termínu a oproti bázi Starých měst jsou výsledky vyhledávání navázány díky souřadnicím na Google Maps. Kromě toho fungují odkazy na související autoritní záznamy z bází CERL (Printers, Authors, Corporate Bodies) a také na záznamy knih v katalogu Europeana. Je připraveno i propojení na záznamy CERL Reference Works (použitá literatura).

Pro usnadnění vyhledávání v několika geografických bázích jsme vytvořili zvláštní vyhledávací nástroj a začlenili jsme jej do systému Manuscriptoria jako samostatnou službu. Umožňuje v současné době prohledávání bází Stará města a CERL Thesaurus - Places. Při pokládání dotazu proběhne vyhledávání nejprve v bázi

Stará města a jeho výsledky jsou využity při tvorbě dotazu do báze CERL Thesaurus – Places.

Uživateli jsou tak „na jedno kliknutí" zprostředkovány informace, které by si jinak musel samostatně vyhledávat v několika databázových systémech a posléze složitě propojovat. Vyhledané výsledky jsou sestaveny pomocí logických operátorů do dotazu použitelného např. při vyhledávání v katalogu Manuscriptoria.

Toto řešení – samostatný vyhledávací nástroj - jsme zvolili vzhledem k rozdílné formě zápisu geografických jmen v tezaurech a v katalogu Manuscriptoria. Názvy v tezaurech jsou uváděny v prvním pádě jednotného (případně množného) čísla, ovšem geografické údaje v Manuscriptoriu jsou často vyskloňovány, takže neodpovídají svým tvarem termínům nalezeným v tezaurech. Uživatel má možnost vyhledané výsledky z geografických bází nejen přímo použít pro tvorbu dotazu do katalogu Manuscriptoria, ale může je podle svých potřeb upravovat přímo v prostředí M-Tez.

# 2 Struktura databáze „CERL Thesaurus - Places“

Struktura záznamu v databázi "CERL Thesaurus – Places" začíná autoritním tvarem názvu města či místa v 1. pádě. Toto záhlaví je v CERL Thesaurus uváděno často ve dvou variantách – v národním jazyce a jazyce tvůrce prvotního záznamu (viz hledání např. Praha, Brno, Olomouc). K záhlaví je připojena informace o lokaci města nebo místa a všechny zjištěné variantní podoby lokality - jak jazykové, tak grafické - opět v 1. pádě. Dalšími údaji jsou fiktivní názvy měst nebo míst, zeměpisné souřadnice, citace použité literatury a informace o tvůrcích záznamu.

## 2.1 Struktura záznamu

Data zpřístupněná v CERL Thesaurus Places využívají pro zápis interní formát typu MARC. Při vyhledávání, v návěšťovém zobrazení ani následném exportu dat nejsou obsažena všechna pole vyplněná v rámci interního formátu. V přehledu níže je uveden seznam těch položek, které jsou posléze předávány ve formě XML:

**001** - Kontrolní číslo - pole nemá žádné indikátory a podpole (neopakovatelné)

**215** – podpole a - Záhlaví (opakovatelné)

**356** – podpole a - Další informace k lokaci (opakovatelné)

**415, indikátory 01** - podpole a - Variantní názvy (opakovatelné)

**415, indikátory 11** - podpole a – Fiktivní názvy (opakovatelné)

Ukázku struktury do Manuscriptoria předávaného XML záznamu ukazuje následující Výpis 1.

```
<record xmlns='http://sru.cerl.org/ctas/dtd/1.1/' id='cnl00008136'
type='placeName'>
<info>
<display>Celle</display>
<geographicalNote>Deutschland, Niedersachsen, Lüneburg,
Celle</geographicalNote>
</info>
<nameForms>
<headingForm name='single'>Celle</headingForm>
<variantForm name='single'>Cesla</variantForm>
<variantForm name='single'>Shaesla</variantForm>
<variantForm name='single'>Zella</variantForm>
<variantForm name='single'>Cell</variantForm>
<variantForm name='single'>Cella Saxonum</variantForm>
<variantForm name='single'>Cella</variantForm>
```

```
<variantForm name='single'>Cellae</variantForm>
<variantForm name='single'>Cellis</variantForm>
<variantForm name='single'>Cellis Luneburg</variantForm>
<variantForm name='single'>Cellis Luneburgicis</variantForm>
<variantForm name='single'>Cellis Luneburgum</variantForm>
<variantForm name='single'>Zell</variantForm>
<variantForm name='single'>Zelle</variantForm>
<fictionalForm name='single'>Moskau</fictionalForm>
</nameForms>
<identifiers>
<identifier type='cerlthesaurus'>cnl00008136</identifier>
<identifier type='cerlthesaurus'>cnl00000150</identifier>
<identifier type='other' source='DE:GYMG'>SAUR-
2:00000591</identifier>
</identifiers>
</record>
```

**Výpis 1 - XML podoba jednoho záznamu autoritní databáze „CERL Thesaurus - Places"**

# 3 Implementace vyhledávání v databázích Stará města a CERL Thesaurus - Places

Pro vyhledávání v bázích Stará města a CERL Thesaurus – Places byla vytvořena samostatná pilotní služba v systému Manuscriptorium, díky které má uživatel možnost ověřit si nejen existenci hledaného geografického termínu, ale zobrazí se mu navíc informace o jeho variantních formách, případně fiktivních názvech, které byly zaznamenány jak na území České republiky, tak i na celoevropské úrovni.

Geografické termíny uložené v bázi Stará města obsahují primárně informace pouze o místech vydání prvotisků a starých tisků. Oproti tomu záznamy v databázi CERL Thesaurus – Places zahrnují širší spektrum. Jsou mezi nimi nejen geografické názvy míst vydání, ale i informace čerpané z provenienčních údajů nebo biografických dat autorů a korporací.

## 3.1 Popis způsobu vyhledávání v databázích Stará města a CERL Thesaurus - Places

Termín vepsaný uživatelem do vyhledávacího řádku je odeslán nejprve do báze Starých měst a následně jsou kompletní výsledky tohoto primárního dotazu automaticky pokládány jako další dotaz do báze CERL Thesaurus – Places. Díky tomu se významně rozšiřuje množství výsledných termínů, které odpovídají původně zadanému dotazu, i jejich vypovídací hodnota mimo jiné kvůli fiktivním názvům uváděným v CERL Thesaurus – Places.

V případě, že Stará města neobsahují uživatelem zadaný termín, je primární dotaz pokládán přímo do báze CERL. Vyhledané alternativní geografické termíny uživatel může podle potřeby označit a využít např. jako základ dotazu v katalogu Manuscriptoria.

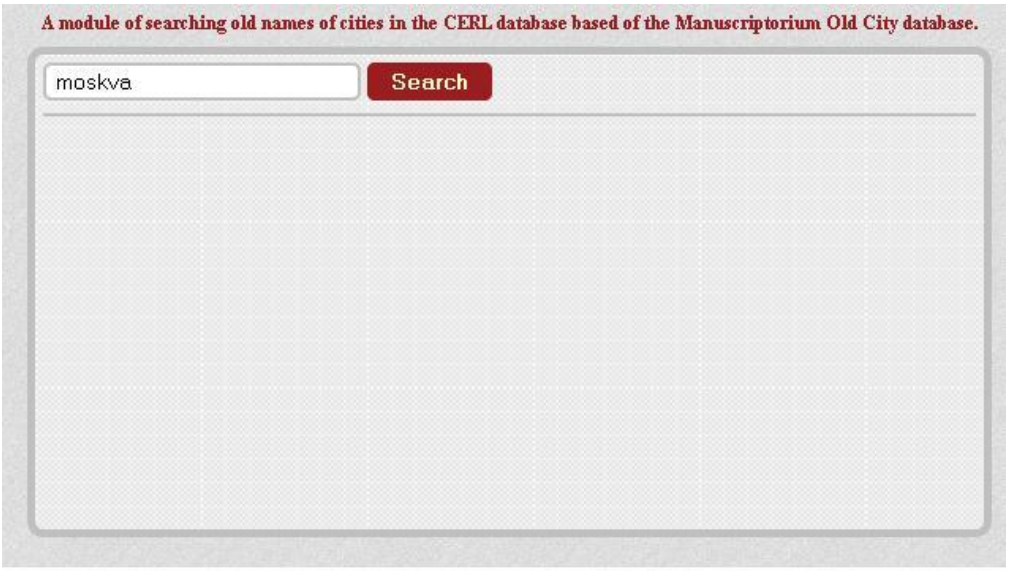
Princip vyhledávání v obou databázích je přehledně znázorněn na Obr. 1. Na příkladu termínu Moskva a jeho alternativních názvů v autoritní databázi je ukázán způsob prohledávání databází a postupného rozšiřování výsledků dotazu. Každá z bází je indexována tak, že by stačilo zadat jeden z alternativních názvů a byly by vyhledány kompletní záznamy s autoritní podobou názvu a všemi jeho variantami. Při dvoustupňovém prohledávání navíc uživatel získá informace i o autoritních záznamech z CERL Thesaurus - Places, v nichž figurují všechny termíny vyhledané v první bázi jako autoritní podoba, varianta, či fiktivní název.


**Obr. 1 Vyhledávání s využitím databází „Stará města“ a „CERL Thesaurus – Places“**

## 3.2 Popis uživatelského rozhraní

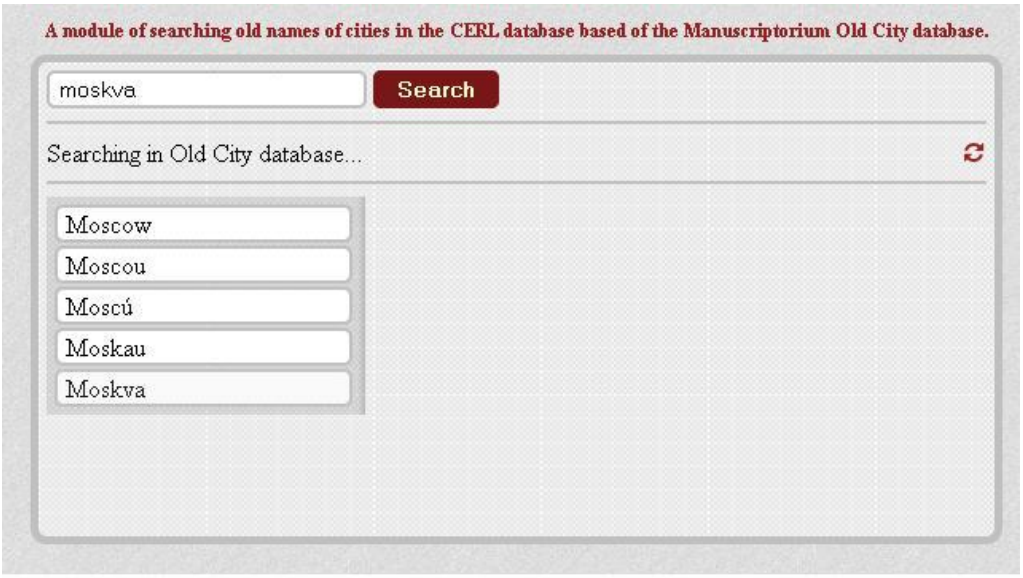
Pilotní řešení je k dispozici na serveru AiP Beroun v rámci testovací verze systému Manuscriptoria (192.168.10.12/cerl/).

Na následujících obrázcích je zobrazen postup vyhledávání. Uživatel nejprve vepíše do vyhledávacího řádku požadovaný termín - „Moskva“ a poté klikne na tlačítko „Search“ (viz Obr. 2).



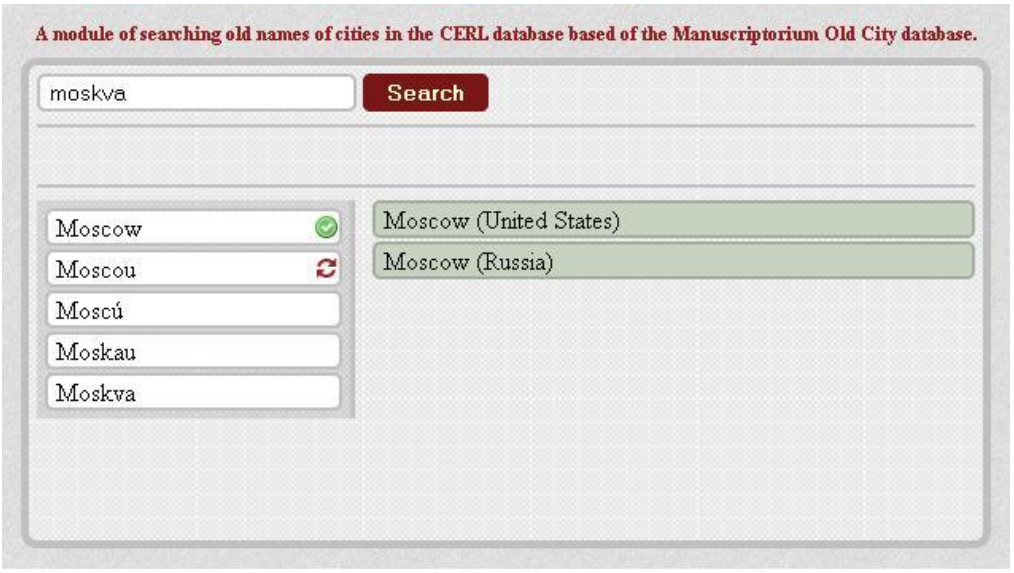

**Obr. 2 – Úvodní okno vyhledávacího nástroje**

Zadaný termín je vyhledáván nejprve v bázi Stará města a nalezené záznamy se po dokončení vyhledávání zobrazí v levé části obrazovky (Obr. 3).



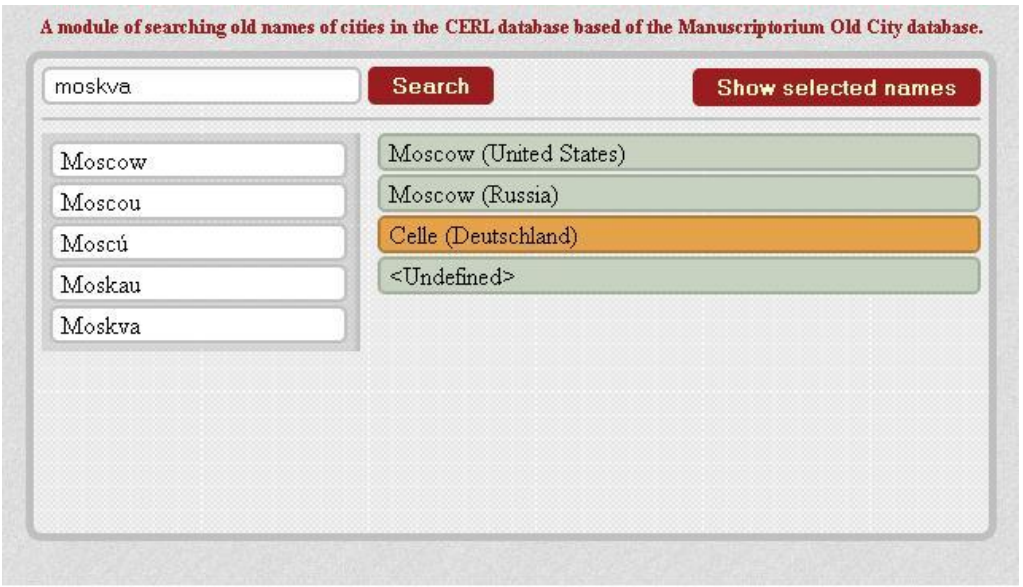

**Obr. 3 – Vyhledávání v bázi „Stará města"**

Po dokončení prohledávání Starých měst jsou postupně automaticky pokládány dotazy z každého nalezeného termínu do báze CERL Thesaurus – Places a výsledky se zobrazují v pravé části obrazovky (Obr. 4).



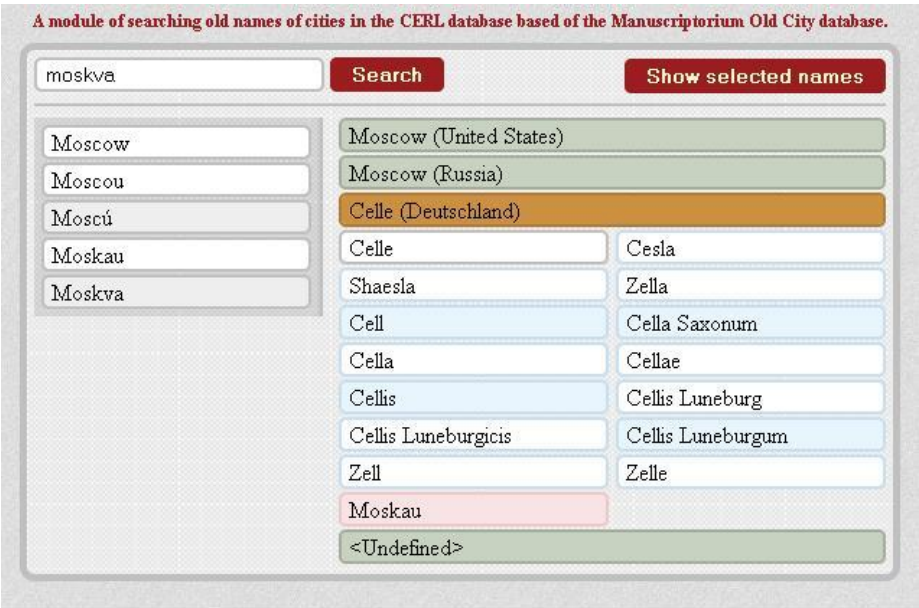

**Obr. 4 – Vyhledávání v bázi „CERL Thesaurus - Places"**

Seznam nalezených záznamů, z nichž je pro lepší přehled v prvním náhledu zobrazeno pouze autoritní záhlaví, je graficky rozlišen podle toho, zda se v jejich obsahu hledaný termín vyskytuje jako varitantní nebo fiktivní jméno. Záznamy obsahující termín ve variantních podobách nebo v záhlaví mají zelenou barvu, záznamy s termínem zapsaným jako fiktivní jméno jsou oranžové (Obr. 5).



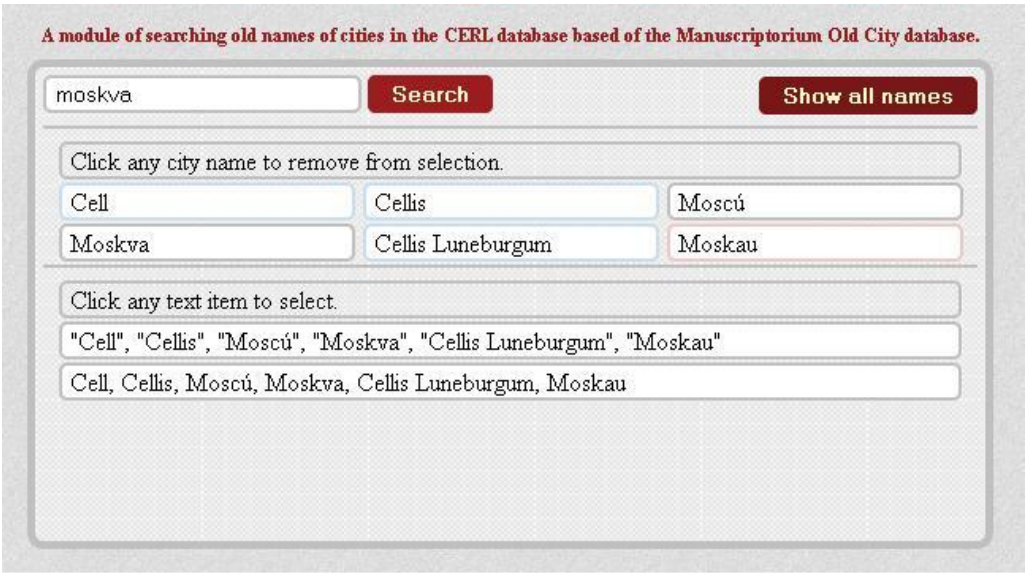

**Obr. 4 – Zobrazení vyhledaných záznamů z obou bází, odlišení záznamů s termínem zapsaným jako fiktivní název.**

Po dokončení vyhledávání si uživatel může rozkliknout jednotlivé položky v seznamu vyhledaných termínů, aby viděl celé záznamy. V otevřeném záznamu jsou graficky odlišeny termíny uvedené jako záhlaví, variantní a fiktivní názvy. Ze všech těchto alternativ (včetně termínů vyhledaných z báze Starých měst) si uživatel může kliknutím vybrat požadovaný termín (Obr. 5).




**Obr. 5 – Zobrazení celého záznamu z báze „CERL Thesaurus – Places" a označení požadovaných termínů**

Vybrané termíny je možné si zobrazit pomocí tlačítka „Show selected names" v další obrazovce. Všechny vybrané termíny jsou vypsány v horní části obrazovky. V dolní části jsou termíny připraveny ve dvou variantách pro vložení do clipboardu. První nabízí termíny ve formě frází oddělené čárkami (logický operátor OR), v druhé variantě jsou termíny ve slovním režimu opět odděleny čárkami. Uživatel má možnost si jednu z variant zakliknout a po zkopírování využít např. při tvorbě dotazu do katalogu Manuscriptoria (Obr. 6).




**Obr. 6 – Zobrazení vybraných termínů (možnost zkopírování do dotazu pro Manuscriptorium)**

# 4 Zhodnocení / Závěr

Pilotní řešení je v současné době dokončeno. Na straně zpracovatele byla ověřena funkčnost všech částí připravené služby a také obsahový překryv dat z báze Stará města se záznamy z CERL Thesaurus – Places. Možnost výběru z vyhledaných alternativních názvů a vložení do clipboardu nabízí uživatelům základní komfort při práci s výsledkem vyhledávání.

Způsob implementace respektuje známá omezení (předchozí zpráva [4]), jímž je především v datech chybějící informace o jazyce, ve kterém jsou zapsány variantní názvy. Toto se týká i CERL Thesaurus – Places. V něm navíc mohou nastat problémy při identifikaci autoritní podoby názvu, protože po redakci databáze jsou u některých záznamů k dispozici dvě záhlaví (např. Prag / Praha, Olomouc / Olmütz).

Na základě zprávy z roku 2010 jsme u nové vyhledávací služby přistoupili k zobrazování vyhledaných výsledků z CERL Thesaurus – Places ve dvou krocích. Nejprve je uživateli k dispozici seznam vyhledaných záhlaví a až po výběru požadovaného záhlaví si uživatel prohlédne celý obsah záznamu. Při prohlížení si může vybrat preferované termíny pro další práci (např. přípravu dotazu do Manuscriptoria). Jako praktická vizuální pomůcka při práci slouží uživateli barevné rozlišení jednotlivých typů geografických názvů (záhlaví, variantní název, fiktivní název).

Jako další krok bychom badatelům chtěli dát k dispozici možnost položit dotaz „na jedno kliknutí" sestavený z vybraných termínů přímo do Manuscriptoria.

Pilotní řešení potvrdilo předpoklad, který byl formulován ve zprávě z roku 2010 [4], že je možné efektivně propojit vyhledávání geografických termínů v bázi „Stará města" a následně dotaz rozšiřovat pomocí údajů z báze CERL Thesaurus – Places. Potvrdilo se nám, že pro systém Manuscriptoria není technickým problémem připojovat i další externí databáze, pokud budou vhodným způsobem poskytovat svá data ve zdrojové podobě ke stažení a následnému zpracování v dalších systémech nebo budou disponovat nějakou formou aplikačního rozhraní tak, aby mohly být dalšími systémy využívány.

Další fází ve využívání externích zdrojů by mohlo být připojení ostatních bází spravovaných CERLem (Consortium of European Research Libraries) do systému Manuscriptoria (jména autorů, názvy institucí, jména tiskařů). V tezauru osobních jmen je obsažena např. valná většina středověkých autorů. Záznamy jsou navíc propojeny se souvisejícími autoritními záznamy v ostatních bázích CERLu a některé dokonce s katalogizačními záznamy dokumentů v Europeaně.