Metodika je výsledkem řešení výzkumného projektu
TB02CUZK004 Vytvoření tezauru pro obor zeměměřictví a katastru nemovitostí,
poskytovatel Technologická agentura ČR, řešeném v letech 2014-2015.

TERMINOLOGICKÝ TEZAURUS PRO OBOR ZEMĚMĚŘICTVÍ A KATASTRU NEMOVITOSTÍ: CERTIFIKOVANÁ METODIKA

**RNDr. Adam Rambousek, Ph.D., Mgr. et Mgr. Vít Baisa, RNDr. Vít Suchomel,
doc. RNDr. Aleš Horák, Ph.D., Mgr. Lucia Kocincová**

Terminologický tezaurus pro obor zeměměřictví a katastru nemovitostí: Certifikovaná metodika

Autoři

RNDr. Adam Rambousek, Ph.D. (40 %)

Mgr. et Mgr. Vít Baisa (20 %)

RNDr. Vít Suchomel (20 %)

doc. RNDr. Aleš Horák, Ph.D. (10 %)

Mgr. Lucia Kocincová (10 %)

Dedikace

Metodika je výsledkem řešení výzkumného projektu *TB02CUZK004 Vytvoření tezauru pro obor zeměměřictví a katastru nemovitostí*, poskytovatel Technologická agentura ČR, řešeném v letech 2014-2015.

Cíle řešení projektu

Vývoj systému pro správu vícejazyčného tezauru, který umožní editaci a prohlížení tezauru, včetně importu dat z tematických databází (slovník VÚGTK, RUIAN a další) a exportu dat v konfigurovatelném formátu s možností nastavení kritérií pro předávání informací. Součástí systému bude také webová služba pro publikaci obsahu tezauru dalším aplikacím podle popsaného rozhraní (s využitím standardů WSDL a REST/JSON).

Oponenti

- prof. Ing. Václav Matoušek, CSc., ZČU Plzeň
- doc. Ing. Petr Rapant, CSc., Institut geoinformatiky, VŠB-TU Ostrava

Obsah

1 Cíl metodiky

Cílem předkládané certifikované metodiky je popsat nové metodické postupy při tvorbě, údržbě a rozšiřování terminologie oblasti zeměměřictví a katastru nemovitostí. Popisovaná metodologie se bude opírat o vlastnosti systému vyvinutému speciálně na podporu těchto činností - systém *Tezaurus pro obor zeměměřictví a katastru nemovitostí* (dále jen Tezaurus). Text metodiky podrobně popisuje využití a nastavení systému Tezauru za účelem systematického zpracování aktuální i budoucí terminologie z oblasti působnosti ČÚZK (Český úřad zeměměřický a katastrální), včetně technické dokumentace dodaného řešení a rozhraní služeb pro import, export a publikaci dat pomocí webové služby.

Tezaurus je definován jako *"řízený a měnitelný slovník deskriptorového selekčního jazyka uspořádaný tak, že explicitně zachycuje apriorní (paradigmatické) vztahy mezi lexikálními jednotkami" [1].* Podle obsahu a struktury je možné tezaurus dále dělit do různých druhů a kategorií, např. vícejazyčný nebo tematický. V případě tezauru ČÚZK jde o vícejazyčný tezaurus zaměřený na terminologii z oblasti působnosti ČÚZK, tj. pro obor zeměměřictví a katastr nemovitostí. Udržování, rozšiřování a aktualizace Tezauru umožní dlouhodobě konzistentní užívání termínů v informačních systémech a dokumentech ČÚZK.

Systém Tezauru podporuje předkládané metodické postupy pro tvorbu, údržbu a doplňování terminologické databáze v několika rovinách:

- základní slovníková aplikace (*DWS - dictionary writing system*) umožňuje klasické operace manuální práce se slovníkem v podobě terminologické databáze jako je editace, mazání a přidávání hesel nebo tvorby intraslovníkových propojení.
- konfigurovatelný import umožňuje dávkové propojení terminologické databáze s externími zdroji dat. Vzorové konfigurace popisují import z HTML dokumentů nebo CSV tabulek.
- automatické jazykové technologie založené na zpracování velkých doménových textových kolekcí (korpusů) ve více jazycích umožňují navrhovat kandidátské termíny, vztahy a překlady na základě:
    - analýzy statistických vlastností termínů v nových odborných textech
    - analýzy vztahových vzorů mezi termíny v českých odborných textech
    - srovnání kontextových vlastností kandidátských termínů v uvedených vícejazyčných doménových korpusech

  Tímto způsobem je možné průběžně zpracovávat návrhy nových termínů reálně užívaných v (současných i budoucích) odborných textech.
- prezentace dat pomocí slovníkové aplikace a webové služby doplňuje základní data o informace z korpusů jako jsou příklady užití termínů v reálných textech nebo termíny vyskytující se v podobných kontextech.

Předkládaná certifikovaná metodika popisuje metody pro vytvoření, rozšiřování a aktualizaci doménové terminologie pomocí systému Tezauru. Rovněž popisuje přidruženou webovou službu, která poskytuje nejen obsah Tezauru, ale umožňuje i manipulaci s daty Tezauru pomocí standardizovaného aplikačního rozhraní ve formátech WSDL a REST/JSON. Webová služba byla navržena tak, aby bylo zajištěno jednoduché propojení se stávajícími aplikacemi pomocí moderních standardů a aby bylo možné využít Tezaurus v nově vytvořených aplikacích. Webová služba je volně přístupná veřejnosti, tudíž kromě primárního cíle zajišťuje rovněž cíl vzdělávací: znalost terminologie a její jednotné užívání.

1.1 Metodické postupy

V této části uvádíme typické produkční situace a přehled jejich řešení formou odkazů do příslušných částí metodiky. Ve všech případech předpokládáme, že operace budou probíhat na dodané kompletní

instalaci (ve formě tzv.virtuálního stroje, viz 1.2 Výklad pojmů), která obsahuje funkční kompletní systém Tezauru včetně prvotního importu dat. Podrobný popis obsahu virtuálního stroje a instalačních softwarových balíků je součástí technické zprávy dodávané se systémem. Přehled provedených prvotních importů dat a jejich úprav je popsán v kapitole 2.3 Vstupní data.

1.1.1 Vytvoření nového tezauru

V aktuálně dodané aplikaci přímo tato operace není zapotřebí, protože Tezaurus je již naplněn všemi požadovanými daty, ale principiálně se zde aplikuje stejný postup, jako při *importu* většího množství termínů. Pro tyto operace je nutné *editorské oprávnění* pro přístup (viz 2.2.5.1 Správa uživatelů). Pro naplnění Tezauru daty jsou určeny následující postupy:

- Pokud jsou k dispozici **data termínů** ve formě tabulky (např. z programu Excel nebo LibreOffice), je možné tyto termíny hromadně přidat do Tezauru pomocí importu CSV dat, viz kapitola 2.2.6.1 Import uživatelských termínů. Při importu termínů je vždy možné (podle volby) přidávat buď jen nové termíny (v Tezauru neobsažené) nebo kompletně přepsat stávající termíny se stejnými názvy nebo identifikátory.
- Systém Tezauru obsahuje také konfigurace pro import dat z **externích zdrojů**, konkrétně publikovaného Terminologického slovníku VÚGTK a vybraných dat RÚIAN. V případě, že v budoucnu budou do těchto zdrojů přidány nové termíny, systém umožňuje jejich automatizované přidání do Tezauru. Podrobně je tento import termínů z externích zdrojů popsán také v kapitole 2.2.6.1 Import uživatelských termínů.
- Systém umožňuje také přidávání termínů na základě automatické analýzy dodaných odborných textů. Tento postup je určen např. pro nalezení potenciálních **neologismů** nebo termínů, které nejsou standardizovány, ale jsou přitom běžně **používané v odborných textech**. V případě, že budou k dispozici nové rozsáhlé elektronické kolekce odborných textů (zejména texty, ve kterých se vyskytují termíny dosud v Tezauru nezpracovávané), je možné tyto texty do systému nahrát a s využitím technik popsaných v kapitole 2.4.2 Automatická extrakce a návrh nových termínů vygenerovat *nové kandidátské termíny*. Konkrétní postup je popsán v kapitole 2.2.6.2 Import uživatelských dokumentů. Kandidátské termíny jsou do systému přidány odděleně od běžných termínů a musí před plným zařazením do hlavní hierarchie Tezauru projít expertní editací.

1.1.2 Úprava hesel tezauru

Upravovat hesla tezauru (pojmy, termíny, definice, odkazy, ...) může jen uživatel s *editorským oprávněním* pro přístup do systému (viz 2.2.5.1 Správa uživatelů). Úpravy hesel mohou být buď jednotlivé nebo hromadné a mohou probíhat pouze na základě expertních uživatelských znalostí a informací nebo s využitím kandidátských návrhů systému.

- Základní **úprava jednoho hesla** spočívá v jeho přímé změně pomocí editačního formuláře, viz souhrn v 2.2.3 Uživatelské rozhraní a podrobný popis v Příloze č. 3 Návod pro editory. V hesle je tak možné **změnit termíny nebo definice**, změnit **nadřazené pojmy** hesla, tj. zařadit heslo do konkrétního místa v hierarchii nebo celé heslo úplně smazat. Smazaná hesla jsou udržována ve zvláštním seznamu, ze kterého je možné je i vrátit zpět k editaci.
- Při editaci jednoho hesla je možné využít i **automatické návrhy** nadřazených pojmů nebo překladů (pro informaci, jak jsou tyto návrhy vytvářeny viz 2.4.3 Návrhy hyperonymických vazeb mezi termíny a 2.4.4 Návrhy překladových kandidátů z korpusů). Tyto informace slouží ovšem pouze jako podkladové, uvedené metody mají průměrnou úspěšnost cca 50%. V každém případě je tedy nutná expertní kontrola navrhovaných kandidátských dat.

Automatické návrhy jsou přímo součástí editačního formuláře (u hesel, kde jsou návrhy k dispozici), viz Příloha č. 3 Návod pro editory.

- V případě potřeby **hromadné úpravy** více hesel je možné použít funkci Export dat Tezauru ve formátu CSV, podrobněji viz 2.2.6.3 Export slovníkových dat Tezauru a z Příloha č. 3 Návod pro editory. Data v exportované tabulce je poté možné upravovat v klasických tabulkových editorech, jako je např. Excel nebo LibreOffice a po úpravách je zpět importovat do Tezauru. Při importu je nutné zvolit, zda se stávající data se stejným názvem nebo identifikátorem budou přepisovat nebo zda se importují jen nové termíny. Není proto možné současně upravovat termíny v systému i v exportované tabulce mimo systém.

1.1.3 Využití dat Tezauru v aplikacích třetích stran

Systém Tezauru obsahuje standardizované aplikační programové rozhraní (API) pomocí webové služby (*web service*), které umožní snadné **využití základních funkcí** Tezauru i jinými způsoby než umožňuje uživatelské rozhraní systému. Pomocí funkcí aplikačního rozhraní je možné v libovolné aplikaci (nutný je pouze přístup k Tezauru on-line přes Internet a uživatelský účet v systému s příslušným oprávněním) řešit následující situace:

- prohledávání a výpis dat Tezauru v libovolné formě (operace *search* a *getdoc*),
- změna, uložení nebo smazání vybraného hesla (operace *save* a *delete*),
- přidání nového hesla (operace *save*),
- vyhledání a vyznačení výskytů termínů ve vlastním textu (operace *highlight*). Tato metoda byla aplikována na definice stávajících hesel tezauru a poskytuje tak kontextové provázání definic.

Podrobný popis a příklady použití funkcí aplikačního rozhraní jsou uvedeny v Příloze č. 4.

1.2 Výklad pojmů a zkratek

Následující seznam uvádí stručné výklady nejčastějších pojmů a zkratek (převážně z oblasti jazykových technologií a slovníkových aplikací), které jsou používány dále v textu metodiky.

administrační rozhraní - část aplikace určena pouze správci, běžní uživatelé do této části aplikace nemají přístup
administrátorské oprávnění - speciální oprávnění ke vstupu do administrační části aplikace
aplikační rozhraní, **API** - (Application Programming Interface) standardizovaný způsob poskytování služeb Tezaurus jiným aplikacím, aplikace třetích stran mohou využívat funkce aplikace Tezaurus pomocí webové služby
automatická extrakce - viz extrakce
CSV - (Comma Separated Values) standardní textový formát pro uložení tabulek, kdy jeden řádek odpovídá jednomu řádku tabulky, obsah jednotlivých buněk je (typicky) oddělený čárkou
databáze - uspořádaná množina informací uložená v aplikaci
disambiguace - zjednoznačnění gramatických informací o slovu v daném kontextu, např. slovo *řeky* může být druhý pád jednotného čísla nebo první pád množného čísla, avšak ve slovním spojení *z řeky* je slovo *řeky* jednoznačně ve druhém pádě
doménový/specializovaný korpus - textový korpus sestavený z textů z určité oblasti (např. geologie)
DTD - (Document Type Definition) soubor pravidel určujících, jaké elementy a atributy můžeme použít v daném XML dokumentu
export dat - jednorázový způsob přenosu dat z aplikace Tezauru ke zpracování mimo tento systém (jinou aplikací nebo přímou expertní úpravou)

extrakce - automatický proces analýzy textů, jehož výsledkem jsou návrhy kandidátských termínů
heslo tezauru, pojem - základní jednotka slovníku, skládá se z více termínů, českých i ve více jazycích
hierarchie, stromová struktura - datová struktura obsahující jeden prvek nadřazený všem (kořenový prvek, kořen), nekoncové uzly a koncové uzly, každý uzel (kromě kořene) je spojen nejméně s jedním uzlem nadřazeným
HTML - (HyperText Markup Language) značkovací jazyk pro popis formátování webových stránek
hyperonymum - nadřazený pojem, nadtřída, obecnější pojem, například slovo *řeka* má hyperonymum *vodní tok*, opakem je hyponymum
hyponymum - podřazený pojem, podtřída, specifický pojem, například slovo *řeka* je hyponymem výrazu *vodní tok*, opakem je hyperonymum
identifikátor, ID - jednoznačný (obvykle číselný) kód hesla, přiřazený automaticky systémem
import dat - jednorázový způsob přenosu dat z jiných aplikací nebo zdrojů do aplikace Tezaurus
JSON - (JavaScript Object Notation) standardní textový formát pro výměnu datových struktur mezi různými programy přes počítačovou síť, používá se při volání webových služeb
kandidátský překlad - navržený překlad, jehož správnost musí potvrdit expert
kandidátská relace - vztah mezi termíny, jehož platnost musí potvrdit expert
kandidátský termín - slovo či slovní spojení, které by mohlo být termínem, rozhodnout o zařazení kandidátského termínu mezi termíny musí expert
klíčové slovo - viz kandidátský termín
klient - viz server/klient
konkordance - výpis všech výskytů konkrétního termínu v korpusu včetně omezeného kontextu (typicky ± 5 slov)
korpus - viz textový korpus
lemma - základní tvar slova (první pád jednotného čísla, u sloves neurčitek), např. slovo *mapou* má lemma *mapa*, slovo *jalovému* má lemma *jalový*
lemmatizace - určení lemmatu pro každé slovo v textu
morfologická analýza - určení lemmatu a gramatických informací o slovu (slovní druh, podle slovního druhu dále pád, číslo, rod, osoba, čas, způsob), ke každému slovu je přiřazeno lemma a seznam gramatických informací, např. pro slovo *řeky* je lemma *řeka* a gramatické informace *podstatné jméno, rod ženský, druhý pád jednotného čísla nebo první pád množného čísla*
nadřazený pojem - viz hyperonymum
paralelní/srovnatelný korpus - textový korpus obsahující stejné (nebo podobné) texty v různých jazycích (jeden text je překladem druhého nebo se zabývá stejným tématem) s vyznačením odpovídajících vět
podřízený termín - viz hyponymum
překladový ekvivalent - slovo nebo slovní spojení se stejným významem, ale v jiném jazyce (např. *řeka - river*)
překladový kandidát - viz kandidátský překlad
přihlašovací jméno a heslo - způsob autentizace uživatele, umožňuje přiřadit uživatelům systému odpovídající role a přístupová práva
server/klient - standardní architektura pokročilých výpočetních systémů, kdy jeden počítač (server) nabízí služby a klientské aplikace (často na jiných počítačích) se k serveru připojují a služby podle potřeby používají
stromová struktura - viz hierarchie
syntaktická analýza - (automatické) určení větných členů, tj. rozpoznání podmětu, přísudku, předmětu a dalších kategorií
termín - základní jednotka hesla, konkrétní instance pojmu, položka v hesle, slovo nebo slovní spojení v určitém jazyce

textový korpus - soubor textů v jednom jazyce, ve kterém lze vyhledávat výskyty konkrétních slov nebo základních tvarů v kontextech, korpus může být obecný nebo doménový, texty ve více jazycích mohou být uloženy v paralelních korpusech
tezaurus - speciální druh slovníku, který uživateli nabízí informace o synonymech (slovech se stejným významem), hyperonymech, hyponymech, někdy také o slovech opačného významu
token - základní jednotka textu: slovo, interpunkce, ostatní části textu (čísla, kódy), česky označována také jako *pozice*
tokenizace - rozdělení textu na tokeny/pozice (slova, interpunkci, ostatní, např. čísla nebo kódy), zpravidla na základě mezislovních mezer
uživatelské rozhraní - rozhraní aplikace Tezaurus určené uživatelům, obsahuje ovládací prvky (menu, vyhledávání) a prezentaci dat Tezauru
virtuální stroj - počítačový software, který izoluje aplikace na počítači tak, jako kdyby byly spuštěny na různých počítačích, na jednom fyzickém počítači může být spuštěno více virtuálních strojů, virtuální stroj je možné přenést na jiný fyzický počítač, který může být obsluhován i nekompatibilním operačním systémem
webová aplikace - aplikace, kterou může uživatel spustit z libovolného webového prohlížeče, není nutná instalace
webová služba - služba, kterou aplikace poskytuje jiným aplikacím, popis služby je dostupný pomocí standardu WSDL, data se přenášejí ve formátu JSON
WSDL - (Web Services Description Language) stardardní popis webové služby určený programům, které chtějí webovou službu využívat, popisuje, jaké funkce webová služba nabízí a jak je spustit
XML - (eXtended Markup Language) značkovací jazyk, kterým lze popsat data pomocí prvků (elementů) a jejich vlastností (atributů), data v XML tvoří hierarchii (např. *termín* [*překlady* [*překlad do angličtiny*, *překlad do ruštiny*], *definice*, *hyponyma* [*hyponymum 1*, *hyponymum 2*]])
XSLT - (eXtensible Stylesheet Language Transformations) formální způsob popisu převodu zdrojových dat ve formátu XML do libovolného jiného požadovaného formátu, nejčastěji HTML, jiného XML nebo libovolných jiných datových struktur.

2 Vlastní popis metodiky

2.1 Úvod

Systém Tezaurus pro obory zeměměřictví a katastr nemovitostí (dále jen Tezaurus) je uživatelsky dostupný pomocí webové aplikace Tezaurus. Tato aplikace zpřístupňuje data z oblasti působnosti ČÚZK uživatelům i aplikacím třetích stran. Přínosem aplikace je zobrazení odborných termínů a hierarchických vztahů mezi nimi (pomocí relace hyperonymie) včetně překladů do šesti jazyků (čeština, angličtina, francouzština, němčina, ruština, slovenština). Aplikaci lze použít pro poloautomatické vytváření terminologické databáze. Aplikace Tezaurus umožňuje import dat, díky kterému ji lze rozšiřovat, a export dat pro účely dávkového zpracování a použití dat Tezauru v aplikacích třetích stran. Součástí návrhu aplikace je i jednotný formát uložení dat ve formátu XML.

Aktuální data Tezauru byla v první fázi importována z několika dostupných datových zdrojů. Software pro import dat do Tezauru sestává ze sady nástrojů, které lze opakovaně požít pro strojový import z jednotlivých datových zdrojů. Výhodou tohoto řešení je snadná možnost aktualizace dat při změně původních zdrojů. Cílem importu termínů z různých zdrojů bylo vytvořit co nejširší terminologické podklady, které byly nejdříve automaticky a následně ručně aktualizovány tak, aby poskytly co nejkvalitnější prvotní naplnění databáze.

2.2 Popis systému

Systém Tezaurus slouží k zobrazování, editaci, údržbě a poskytování terminologických dat. Hlavní funkce systému umožňují:

- editaci termínů včetně údajů jako jsou definice, překlady nebo vazby na jiné termíny
- automatické návrhy kandidátských nových termínů na základě analýzy doménových textů
- automatické návrhy kandidátských překladů odvozených z analýzy kontextů v cizojazyčných doménových textech
- automatické návrhy kandidátských nadřazených termínů
- příklady užití termínů v reálných odborných textech
- návrhy kontextově podobných (slabě synonymních) jednoslovných termínů
- dávkové zpracování (tabulkový export/import)
- import z existujících datových zdrojů a export XML struktury
- standardizované aplikační rozhraní pro využití funkcí Tezauru v software třetích stran

Aplikace Tezaurus se skládá z modulů pro editaci hesla, správu systému, import a export dat, webové služby aplikačního rozhraní a korpusové funkce (automatické návrhy). Součástí aplikace jsou rovněž data, která byla agregována a poloautomaticky převedena z existujících zdrojů. Uvedené části aplikace a převod stávajících dat jsou popsány v této kapitole.

Aplikace Tezaurus vychází ze dvou vývojových projektů – část pro správu textových korpusových dat a extrakci termínů je rozšířením aplikace *Sketch Engine*, část pro správu slovníkových dat Tezauru je založená na vývojové platformě *Dictionary Editor&Browser* (DEB) [2].

Inovativní část aplikace spočívá zejména v propojení technik obou existujících projektů, které umožňuje jak práci se slovníkovými daty, tak práci s korpusy. Za naprosto nové lze považovat výstupy aplikace v oblasti automatické extrakce terminologie a automatické návrhy překladových

kandidátů při neexistenci vhodných dat (zarovnaných paralelních korpusů v doméně působnosti ČÚZK).

2.2.1 Výchozí aplikace

Vývoj systému Tezauru navazuje na dlouholeté zkušenosti Centra ZPJ, MU, Brno s aplikacemi pro výzkumné i komerční účely v oblasti jazykového zpracování dat a slovníkových aplikací. Implementovaný systém vychází z aplikací *Sketch Engine* a *platforma DEB*, které byly přizpůsobeny specifickým požadavkům a doplněny o nové funkce zaměřené na metodiku terminologické databáze. Veškerý dále zmíněný software je součástí dodané instalace systému Tezauru.
Sketch Engine [3] je korpusový manažer – webová služba umožňující vytváření, správu a jazykovou analýzu kolekcí textů (korpusů). Nástroj je využíván zejména pro lexikografickou práci, při tvorbě výkladových slovníků na základě velkých textových dat (textových korpusů) pro desítky jazyků. Uživatelé mohou v rámci *Sketch Engine* nahrávat textové dokumenty (.docx, .pdf, .txt, .html, .xml, .tmx atd.). Systém tyto dokumenty zpracuje pomocí jazykových nástrojů pro tokenizaci, morfologickou analýzu a disambiguaci, jednoduchou syntaktickou analýzu a všechna data indexuje ve speciální databázi. Korpusy je posléze možné velmi rychle plnotextově prohledávat na základě komplexních lingvisticky motivovaných dotazů, dále je možné extrahovat statistické analýzy, seznamy slov, trendy časových řad, seznamy kolokací, podobných slov atd. *Sketch Engine* je implementován v jazycích *C++* a *Python* (*Django*, *Cheetah*). V současnosti *Sketch Engine* systém zpracovává velké množství korpusů v mnoha jazycích, přičemž některé korpusy obsahují i desítky miliard slovních pozic. *Sketch Engine* je vyvíjen ve spolupráci Masarykovy univerzity a Lexical Computing Ltd., UK. Pro další informace o nástroji viz www.sketchengine.co.uk.

Platforma DEB je softwarová platforma vyvinutá v Centru ZPJ na Masarykově univerzitě používaná pro tvorbu lexikografických aplikací. Poskytuje databázové úložiště a moduly potřebné pro správu slovníkových dat, uživatelských účtů, vyhledávání v databázi nebo prezentace uložených údajů. Platforma využívá princip rozdělení na klientskou a serverovou část aplikace. Serverová část poskytuje funkce pomocí známého aplikačního rozhraní, klientská část aplikace zajišťuje prezentaci dat uživatelům.
Serverová část je implementována v programovacím jazyce *Ruby* [4], s využitím objektového modulárního návrhu architektury. Jako databázové úložiště se používá nativní XML databáze *Sedna* [5]. Pro vyhledávání dat v databázi se používá dotazovací jazyk *XQuery* [6].
Klientská část je implementována jako webová aplikace kombinující značkovací jazyk HTML s prezentačními styly CSS a skriptovacím jazykem *JavaScript*.
Klientská část aplikace komunikuje se serverem pomocí AJAX požadavků podle zveřejněného aplikačního rozhraní (API, viz 2.2.4), klientská aplikace načítá data ve formátu JSON [7].
Jednotlivá hesla Tezauru jsou v databázi uložena ve formátu XML a pro zobrazení uživatelům mohou být převedena do jiného formátu pomocí jazyka *XSLT* [8].

2.2.2 Jednotný formát dat

Součástí návrhu aplikace Tezaurus je návrh schématu XML, který umožňuje popsat data extrahovaná z různých zdrojů jednotným způsobem. Podrobný popis formátu XML používaného aplikací Tezaurus je v Příloze 1.
Formát byl navržen tak, aby umožňoval uchovat informace o názvu hesla, synonymech, překladech, pozici v hierarchii hesel a vazbách na další hesla.

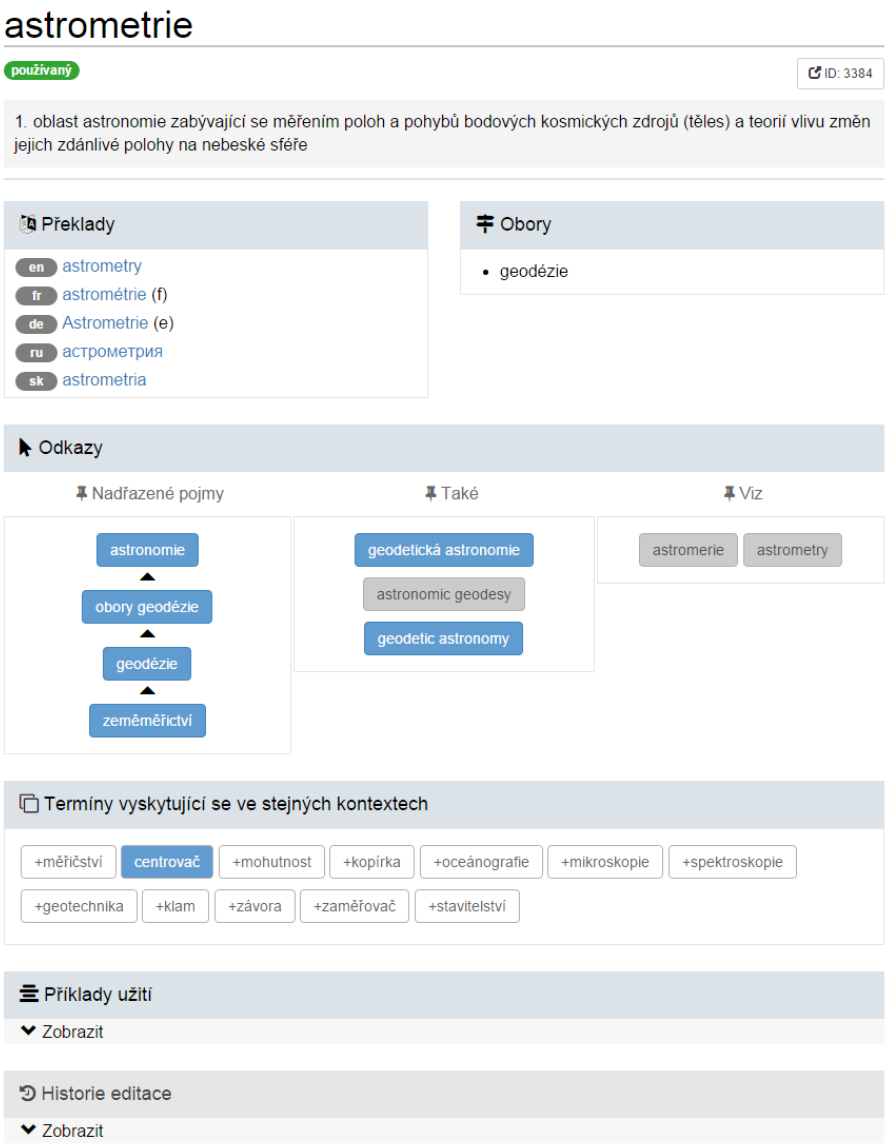
2.2.3 Uživatelské rozhraní

Uživatelské rozhraní aplikace se mírně liší v závislosti na roli uživatele (viz 2.2.5). Uživatele bez patřičného oprávnění nemají povolen import a editaci termínů. Uživatelé mohou procházet Tezaurus pomocí hierarchie hesel nebo vyhledávat konkrétní heslo ve vyhledávacím poli.

Heslo obsahuje název hesla, informaci o četnosti užití v korpusových datech, definici, překlady, zařazení do oboru a umístění v hierarchii hesel. Heslo může obsahovat reference na odbornou literaturu (viz obrázek 2.1). Množství informace je u každého hesla různé.

Po kliknutí na překlad nebo do části *''Příklady užití'* se zobrazí korpusová data obsahující konkordance (konkrétní výskyty daného termínu) v odborných textech.

Podrobnější informace o uživatelském rozhraní jsou obsaženy v návodu na používání aplikace Tezaurus (viz Přílohu 2).



Obrázek 2.1: Ukázka zobrazení hesla v Tezauru

2.2.4 Aplikační rozhraní (API)

Aplikační rozhraní aplikace Tezaurus (Application Programming Interface) umožňuje využití aplikace Tezaurus aplikacemi třetích stran. Ty mohou pomocí uvedeného rozhraní volat funkce vyhledávání, vyznačení termínů v zadaném textu a práce se zadaným heslem (vytvoření, zobrazení, úprava, odstranění).

Očekávané použití aplikačního rozhraní je v mobilních aplikacích využívajících data Tezauru, dále v aplikacích, které s daty Tezauru souvisejí (např. geografické informační systémy). Aplikace mohou využít libovolného ze dvou standardů pro komunikaci mezi aplikacemi (JSON a WSDL). Aplikace třetích stran by měly ošetřovat chybná volání webové služby. Technický popis API je obsahuje Příloha 4.

2.2.5 Správa aplikace

Pro správu systému slouží administrační rozhraní, které je přístupné jako webová služba na adrese `https://[jméno serveru]:8000/`. Uživatelé bez administrátorského oprávnění v tomto rozhraní mohou pouze změnit své přístupové heslo.
Uživatel s administrátorským oprávněním může spravovat uživatele a nastavení systému. Po instalaci je vytvořen administrátorský účet s následujícími údaji:

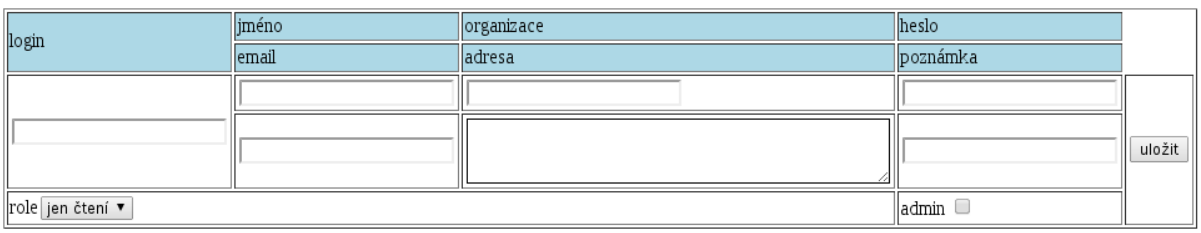
- přihlašovací jméno: *admin*,
- heslo: *tecu*

2.2.5.1 Správa uživatelů

Po kliknutí na odkaz "*správa uživatelů*" v administračním rozhraní se zobrazí seznam uživatelských účtů.

Vytvoření nového uživatelského účtu:

Pro vytvoření nového uživatelského účtu je potřeba vyplnit formulář "*nový uživatel*", povinné údaje: přihlašovací jméno a e-mail. Pokud není vyplněno pole heslo, vygeneruje se nové náhodné heslo. Přihlašovací jméno a heslo, spolu s dalšími informacemi je uživateli odesláno na zadaný e-mail. Dále je potřeba zvolit uživatelskou roli ze seznamu rolí, na výběr jsou možnosti "*pouze pro čtení*" (v rozhraní Tezauru může pouze číst zveřejněné údaje) a "*editor*" (může v rozhraní Tezauru upravovat údaje). Pokud má uživatel mít administrátorská oprávnění, je potřeba zaškrtnout pole "*admin*" (viz obrázek 2.2).



Obrázek 2.2: Vytvoření nového uživatelského účtu

Popis rolí:

- *jen pro čtení*: má možnost pouze prohlížet zveřejněné údaje ve webovém rozhraní Tezauru nebo pomocí API
- *editor*: má oprávnění v rozhraní Tezauru upravovat údaje
- *administrátor*: má oprávnění v rozhraní Tezauru upravovat údaje, spravovat uživatelské účty, provádět import a export databázových dat, provádět import nových dokumentů do korpusu

Úprava uživatelského účtu:

Po kliknutí na přihlašovací jméno v seznamu uživatelů se zobrazí formulář pro editaci údajů. Je možno upravit všechny údaje o uživateli, kromě přihlašovacího jména.
Po kliknutí na odkaz "*nové heslo*" se vygeneruje nové náhodné heslo, které se uživateli odešle e-mailem.
Po kliknutí na odkaz "*smazat účet*" a potvrzení bude uživatelský účet odebrán.
Uživatel je o všech změnách informován e-mailovou zprávou.

2.2.6 Uživatelská data

Aplikace podporuje kromě standardní editace hesel také nahrávání (import) uživatelských termínů nebo dokumentů, ze kterých jsou termíny extrahovány. Importované termíny, které ještě nejsou v databázi, jsou přidány do databáze termínů systému.

Součástí Tezauru je importní a exportní modul, pomocí kterého je možné přidávat hromadně dokumenty z jiných zdrojů. Importní modul pracuje s různými formáty, import dokumentů lze provést jednorázově či pomocí něj přidávat či aktualizovat data již v systému obsažená. Dokumenty lze ze systému exportovat ve formátu CSV nebo XML. Díky exportnímu modulu lze převést data Tezauru do jiného nástroje, XML editoru, případně XML databáze pro další použití.

2.2.6.1 Import uživatelských termínů

Funkce *Import* uživatelských termínů umožňuje nahrát seznam nových termínů do databáze termínů systému. Přidávané termíny musí být v souboru nahraném prostřednictvím uživatelova prohlížeče – uživatel nahraje soubor ze svého počítače a vybere formát vstupního souboru.
Všechny vstupní soubory musí být v kódování UTF-8. Import aktuálně podporuje následující předpřipravené formáty: VUTGK, CSV, RUIAN, TXT, HESLAR. Každý typ odpovídá jedné importní konfiguraci.

- Konfigurace importu typu *VUGTK* byla připravena podle struktury hesel terminologického slovníku na webu VÚGTK (http://www.vugtk.cz/slovnik/, stav k říjnu 2015).
- *CSV* import pracuje s formátem, který lze exportovat v rozhraní Tezauru. Editoři si tak mohou exportovat vybraný podstrom hesel, pracovat s ním v externím programu (např. Excel, LibreOffice) a upravený CSV soubor opětovně importovat do Tezauru.
- Konfigurace *HESLAR* byla připravena pro import dat z Hesláře terminologického slovníku na webu VÚGTK (http://www.vugtk.cz/slovnik/heslar/heslar_rozbal.html, stav k říjnu 2015).
- *TXT* import pracuje s jednoduchým textovým souborem a předpokládá, že termíny v něm jsou uložené každý na samostatném řádku.
- *RUIAN* konfigurace dovoluje importovat vybrané číselníky (stát, kraje, regiony, vyšší územní samosprávní celky a okresy) z veřejného dálkového přístupu k datům registru územní

identifikace, adres a nemovitostí (RÚIAN) jako termíny s nadřazenými a pojmy podle hierarchie RÚIAN.

Data ze vzdálených externích zdrojů je pro účely importu nejprve nutné stáhnout na lokální úložiště. Pro tento účel je možné použít specializované nástroje pro hromadné stahování dat z Internetu, součástí dodané instalace systému jsou i nástroje na stažení uvedených zdrojů pomocí nástroje `wget`.

Konfigurace jsou ve formě skriptu jazyka Python a obsahují proměnnou `configuration` s klíči a hodnotami důležitými pro import. Každá konfigurace definuje, jestli se při importu mohou přepisovat již existující termíny v Tezauru (položka `overwrite`, viz níže). Chování lze změnit v importním formuláři (zaškrtnutím volby Přepsat existující termíny). Termíny jsou přepsány kompletně, existující záznamy se neslučují kromě poznámek.

Pro import CSV jsou důležité obsahy jednotlivých sloupců tabulky (klíč `columns`). Každý prvek seznamu `columns` odpovídá jedné položce XML struktury hesla, za dvojtečkou je číslo sloupce v CSV tabulce, ze kterého se zjistí hodnota položky (sloupce jsou číslovány od 0). Výčet všech povolených hodnot je součástí technické dokumentace dodávané se systémem. Následující příklad CSV konfigurace převádí informace z CSV tabulky níže do výsledného XML formátu použitého v Tezauru.

```
"delim": "    ",
"overwrite": True,
"columns": {
    "id": 0,
    "status": 1,
    "term": 2,
    "term_language": 3,
    "definition": 4,
    "hyperonym": 7
}
```

Příklad vstupu v CSV souboru (ve formě tabulky)

ID	ST	TERM	L	DEF	HYPER
4048	2	digitální mapa	cs	zmenšený generalizovaný konvenční obraz Země	mapa
				digitalizovaná mapa	digitální obraz
				digitální obraz Země	
2933		výšková kóta	cs	číslo vyjadřující požadovanou, popř. skutečnou výškovou polohu předmětu bez ohledu na měřítko, ve kterém je obraz předmětu nakreslený	

Výsledné XML vhodné pro přímé vložení do databáze Tezauru vypadá takto:

```
<entry id="4048" status="2">
  <terms>
    <term number="1" lang="cs">digitální mapa</term>
  </terms>
  <meta>
    <create_time>2015-09-09 12:34:20</create_time>
    <author>auto import</author>
  </meta>
  <defs>
    <def number="1">zmenšený generalizovaný konvenční obraz Země</def>
    <def number="2">digitalizovaná mapa</def>
```

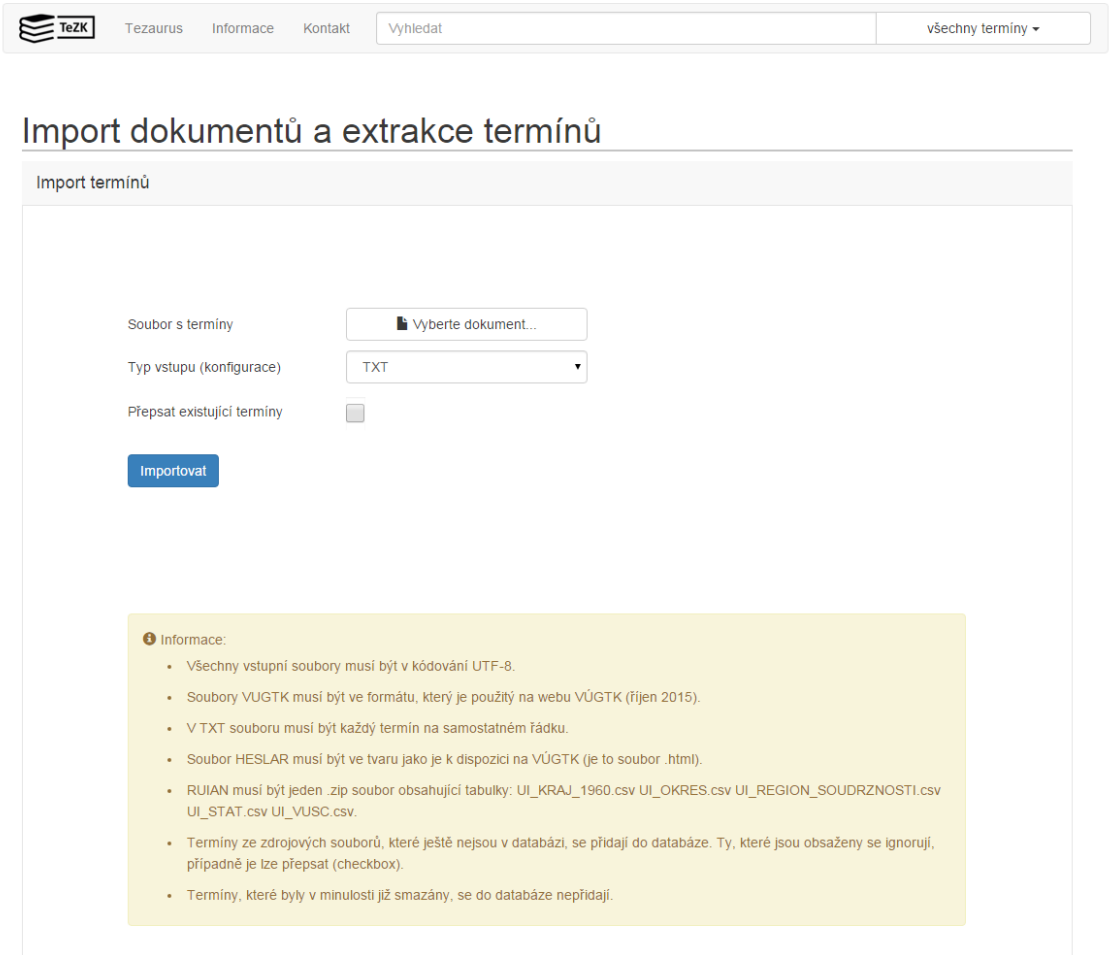
```
    <def number="3">digitální obraz Země</def>
  </defs>
  <hyper id="20388" />
  <notes>
    <note author="import" date="2015-09-09">import z CSV</note>
  </notes>
</entry>
<entry id="2933">
  <terms>
    <term number="1" lang="cz">výšková kóta</term>
  </terms>
  <meta>
    <create_time>2015-09-09 12:34:21</create_time>
    <author>auto import</author>
  </meta>
  <defs>
      <def number="1">číslo vyjadřující požadovanou, popř. skutečnou výškovou polohu předmětu bez ohledu na měřítko, ve kterém je obraz předmětu nakreslený</def>
  </defs>
  <notes>
    <note author="import" date="2015-09-09">import z CSV</note>
  </notes>
</entry>
```

Importní skript prochází postupně řádky CSV tabulky (předpokládá se formát, který vznikne exportem podstromu Tezauru v uživatelském rozhraní, tento formát je také popsán v CSV konfiguraci) a buduje na základě konfigurace XML strukturu hesel. V příkladu CSV vstupu výše je vidět, že formát dovoluje pracovat s vícenásobnými definicemi, případně vícenásobnými hyperonymy (zadanými slovně). Jednotlivá hesla jsou definována unikátními ID v prvním sloupci tabulky.

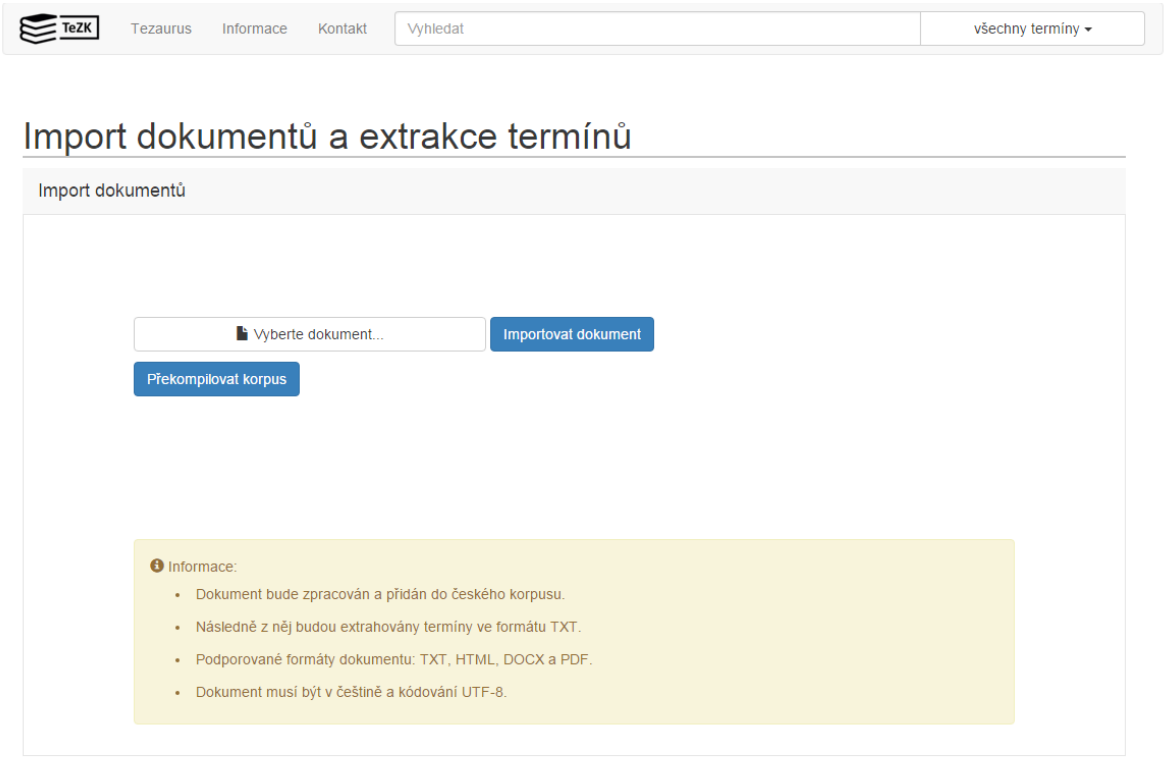
Konfigurace pro HTML (resp. slovník VÚGTK) má stejnou hlavní strukturu jako konfigurace CSV, důležitý je klíč `rules`, který obsahuje slovník nutný pro vytvoření XSLT šablony pro převod HTML souborů souborů ze stránek VÚGTK do podoby tabulky termínů, jejich překladových ekvivalentů, oborů atd. Tato tabulka je pak dále zpracována importním skriptem a převedena do XML souboru, který je přidán do databáze Tezauru. Podrobný popis konfigurace pro HTML soubory je součástí dokumentace dodané společně se systémem.

Importní část je dostupná z hlavní stránky Tezauru:

Obrázek 2.3: Odkazy pro import v rozhraní



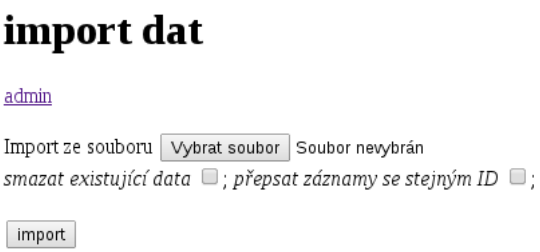
Obrázek 2.4: Formulář pro nahrání termínů



Obrázek 2.5: Formulář pro nahrání a následnou extrakci termínů z dokumentů ve formátu txt, html, docx a pdf

Data Tezauru, která odpovídají struktuře XML (viz 2.2.2), je možné importovat v administračním rozhraní aplikace, viz 2.2.5.

V rozhraní uživatel zvolí soubor z lokálního disku. Soubor se zkopíruje na server Tezauru, kde se pomocí XML schématu ověří, zda struktura souboru odpovídá XML struktuře schématu. Uživatel je upozorněn na případné chyby, v takovém případě k importu nedojde.
Uživatel také může zvolit, zda chce nahradit kompletní data z Tezauru novými daty (smazání původních dat a import nových), nebo aktualizovat a doplnit nové údaje (ponechají se původní data, doplní nová a přepíší aktualizovaná). Uživateli se zobrazuje průběh importu dat. Ukázka importního rozhraní je na obrázcích 2.4, 2.5 a 2.6.



Obrázek 2.6: Rozhraní pro import dat

2.2.6.2 Import uživatelských dokumentů

Funkce *Import uživatelských dokumentů* umožňuje nahrát dokument do textového korpusu českých odborných textů v systému. Přidávaný dokument je předán aplikaci prostřednictvím uživatelova prohlížeče – uživatel zvolí soubor ze svého počítače, aplikace soubor zpracuje a přidá do českého korpusu. Potom z rozšířeného korpusu extrahuje termíny a případné nové termíny zobrazí uživateli.
Požadavky na obsah souboru: Vstupní soubor musí být v kódování UTF-8 a v jednom z formátů TXT, HTML, DOC (*Microsoft Word binary file*), DOCX (*Office Open XML*), nebo PDF (*Portable Document Format*).
Proces zpracování nahraného souboru je podobný jako při vytváření původních korpusů dodaných v systému. Na rozdíl od těchto dat však celý proces proběhne online na počkání. Soubor je převeden na holý text, poté tokenizován, lemmatizován, morfologicky anotován, převeden do formátu vertikál (akceptovaného korpusovým manažerem) a přidán do adresáře se zdrojovými daty korpusu daného jazyka. Dále je korpus znovu zkompilován a jsou z něj extrahovány kandidátské termíny.
Nakonec jsou extrahované termíny zobrazeny a přidány do databáze termínů systému, kategorie automaticky navržených termínů.

2.2.6.3 Export slovníkových dat Tezauru

Data Tezauru je možné exportovat (tedy uložit pro následné hromadné zpracování, a to ruční nebo strojové) dvěma způsoby:

- Při prohlížení hesla uživatelem s *editorským oprávněním* je k dispozici funkce Export dat Tezauru ve formátu CSV. Podrobněji viz Příloha č. 3 Návod pro editory. Taková data je pak možné zpracovávat v tabulkových editorech, jako je např. Excel nebo LibreOffice.
- Kompletní výpis struktury tezauru ve výměnném formátu XML je možné získat v administračním rozhraní aplikace, viz kapitola 2.2.5.
  Uživateli se v prohlížeči zobrazí XML reprezentace kompletních dat Tezauru, odpovídající XML struktuře (viz 2.2.2). Tato data jsou vhodná např. pro automatické zálohování kompletní databáze, z uložené struktury je možné obnovit kompletní tezaurus včetně všech vazeb a metainformací, viz informace o importu XML struktury v administračním rozhraní.

2.3 Vstupní data

Základní soubor termínů byl do Tezauru převzat z několika zdrojů. Níže popisujeme použité zdroje, jejich specifika, zpracování a začlenění do datové struktury Tezauru. Zdroje dat byly převedeny automaticky do formátu XML zpracovatelného aplikací Tezaurus (viz 2.2.2). Pro konverzi byly vytvořeny speciální nástroje, které byly začleněny i do uživatelského rozhraní.

2.3.1 Popis slovníkových dat a jejich zpracování

2.3.1.1 Terminologický slovník VÚGTK

Hlavním zdrojem termínů byl *Terminologický slovník VÚGTK* (dále jen TS VÚGTK). Je dostupný na http://www.vugtk.cz/slovnik/ ve formě provázaných dokumentů ve formátu HTML.

Celý slovník (kolekce HTML souborů) byl stažen a převeden pomocí implementovaného nástroje do formátu XML se strukturou odpovídající Tezauru (viz 2.2.2). TS VÚGTK obsahuje kromě definic hesel také překladové ekvivalenty v několika jazycích a nadřazené obory. Tyto informace byly při konverzi zachovány a jsou součástí dat Tezauru.

Stažené soubory HTML byly hromadně validovány pomocí nástroje tidy (http://www.w3.org/People/Raggett/tidy/) . Výstupem byly validní XHTML soubory. Následně byly tyto soubory převedeny pomocí implementovaného nástroje v jazyce *Python* s využitím knihovny *xml.etree.ElementTree*. Ze souborů (jednotlivé soubory odpovídají termínům ve slovníku VÚGTK) byly extrahovány všechny dostupné informace:

1. termín a jeho synonyma
2. definice
3. obor, do kterého termín spadá
4. reference (literatura, jiné zdroje)
5. překlady do několika jazyků
6. nadřazený pojem

Výstupem převodního programu jsou data ve formátu XML se strukturou odpovídající Tezauru (viz 2.2.2). Nástroj `parse_vugtk_html.py` pro zpracování jednotlivých (X)HTML souborů je součástí systému Tezauru. Při aktualizaci dat TS VÚGTK je možné všechny pojmy znovu stáhnout, konvertovat a aktualizovat s nimi data v Tezauru.

2.3.1.2 Heslář VÚGTK

Heslář, taktéž dostupný online (http://www.vugtk.cz/slovnik/heslar/heslar_rozbal.html), byl zpracován automaticky a data z něj byla přidána do Tezauru. *Heslář* uvádí hierarchické členění termínů a obsahuje ekvivalenty termínů v anglickém jazyce. Nástroj konverze se snaží zachovat co nejvíce informací z *Hesláře* a jeho hierarchické struktury (v podobě hyperonymických vztahů v Tezauru). Vstupní data hesláře obsahují i některé nepravidelnosti, které je nutné po konverzi ručně opravit.

Data *Hesláře* jsou obsažena v jednom souboru, ve speciálním formátu umožňujícím zachytit hierarchii termínů. Pro účely extrakce informací z *Hesláře* byl vytvořen nástroj `parse_heslar.py` (taktéž součástí balíku), který čte data *Hesláře* a na výstup vypisuje seznam termínů s hyperonymickými vazbami, opět ve formátu XML pro Tezaurus (viz 2.2.2).

2.3.1.3 RÚIAN

Do hesláře byla integrována metadata a číselníky *Informačního systému územní identifikace* (ISÚI). Vzhledem k vhodnému rozsahu dat byly automaticky zpracovány číselníky vyšších územních prvků a územně evidenčních jednotek. Číselníky byly importovány ze souborů ve formátu CSV, které jsou zveřejněny na následující adrese:
http://www.cuzk.cz/Uvod/Produkty-a-sluzby/RUIAN/2-Poskytovani-udaju-RUIAN-ISUI-VDP/Ciselniky-ISUI/Vyssi-uzemni-prvky-a-uzemne-evidencni-jednotky.aspx

Údaje z číselníků byly převedeny z formátu CSV do formátu XML (viz 2.2.2) pomocí implementovaných nástrojů se zachováním hierarchických vztahů mezi položkami jednotlivých číselníků. Výsledná struktura v Tezauru tedy kopíruje strukturu územně evidenčních jednotek. Importní část (viz 2.3.2) obsahuje konfiguraci pro opětovné importování souborů CSV.

2.3.1.4 Další zdroje

Další termíny a metainformace lze získat např. zpracováním textových dat z české *Wikipedie* (portál Geografie, obory GIS, Geodézie, Kartografie, Geoinformatika atd.). Jinou možností je zpracování doménových dat získaných z internetových stránek, viz 2.3.3. Za zmínku také stojí terminologická báze IATE (http://iate.europa.eu), kterou spravuje *Translation Centre for the Bodies of the European Union*, zejména obor Geography. Výhodou IATE je překlad termínů do všech oficiálních jazyků Evropské unie.

2.3.2 Statistiky hesel

Tabulka 2.1 uvádí počty hesel, která byla automaticky extrahována ze slovníkových dat. Překlady s více ekvivalenty nejsou započítány.

Tabulka 2.1 Statistika automaticky extrahovaných hesel

	Všechna hesla
počet termínů	9 047
počet hyperonymických relací	10 448
definice (významy)	4 153
počet zařazení do domény	3 965
překladů celkem	22 989
překlady do angličtiny	8 540
překlady do němčiny	3 964
překlady do slovenštiny	3 536
překlady do ruštiny	2 785
překlady do francouzštiny	3 964

Aktuální statistiky hesel jsou dostupné v aplikaci Tezaurus v odkazu Informace.

2.3.3 Popis korpusových dat a jejich zpracování

Textová data pro korpus byla shromážděna dvěma metodami z veřejně dostupných internetových zdrojů. Všechny dále odkazované nástroje byly vyvinuty v *Centru zpracování přirozeného jazyka*, FI MU, Brno. Shromážděné dokumenty byly vyčištěny od netextového a nekvalitního obsahu nástrojem *JusText* [9] a zbaveny duplicit (podobných odstavců) nástrojem *Onion* [9]. Uvedené nástroje jsou

součástí dodaného systému Tezauru a v technické dokumentaci systému jsou uvedeny dodatečné informace o jejich případné instalaci.

Nejprve jsme získali obsah oborových webů pojednávajících o zeměměřictví a katastru nemovitostí. Jejich podrobná statistika s počty dokumentů a pozic je zahrnuta v Tabulce 2.2.

Tabulka 2.2 Statistika zpracovaného obsahu oborových webů

Zdroj	**Internetová doména**	**Dokumentů stáhnuto**	**Pozic stáhnuto**	**Dokumentů po deduplikaci**	**Pozic po deduplikaci**
Webové stránky resortu ČÚZK	www.cuzk.cz	16405	3137795	15322	2938663
Webové stránky VÚGTK	www.vugtk.cz	4659	6419950	3206	3639729
Webové stránky ČSGK	csgk.fce.vutbr.cz	241	77255	186	55911
Webové stránky KGK	www.kgk.cz	417	44814	414	26190
Webové stránky SFDP	www.sfdp.cz	192	35287	110	11081
Webové stránky kartografické společnosti	www.czechmaps.cz	94	108506	93	85659
Webové stránky a časopis Zeměměřič	www.zememeric.cz	8634	6100751	6247	2402238
Webové stránky a časopis Geodetický a kartografický obzor (2014)	www.egako.eu	31	298973	30	250175

Dále jsme shromáždili dokumenty z 1063 webových domén nástrojem *WebBootCaT* [10]. Klíčová slova, hlavní vstup pro tuto metodu, pochází ze seznamu hesel z vícejazyčného terminologického slovníku *VÚGTK, v.v.i.*. Výsledné dokumenty obsahují, kromě klíčových slov, podle kterých byly vyhledány, i kandidáty na další termíny v původním seznamu neobsažené. Tabulka 2.3 uvádí podrobnou statistiku těchto dokumentů.

Tabulka 2.3 Statistika shromážděných dokumentů použitých pro český korpus odbnorných textů

Tematická poddoména	**Dokumentů stáhnuto**	**Pozic stáhnuto**	**Dokumentů po deduplikaci**	**Pozic po deduplikaci**
globální navigační družicový systém	118	250833	117	221315
metrologie	144	867156	144	619482
fotogrammetrie a DPZ	42	244212	42	227731
geografická informace	55	805059	55	550681
mapování	213	858575	212	722080
kartografie	368	1358973	365	1124708
katastr nemovitostí	260	970951	259	776497
geodézie	190	575381	189	483679
teorie chyb	75	258345	75	218809
přístrojová technika	115	187106	113	173984
inženýrská geodézie	114	286846	113	242857

Dále byly do korpusu přidány texty právních předpisů a opatření související s zeměměřictvím a katastrem nemovitostí (stav zdrojů k 2. 7. 2015):

- Přehled právních předpisů souvisejících se zeměměřictvím a katastrem nemovitostí – 24 dokumentů podle přehledu:
  - http://www.cuzk.cz/Predpisy/Prehled-pravnich-predpisu-souvisejicich-se-zememer.aspx
- Právní předpisy v oboru zeměměřictví a katastru – 13 dokumentů podle přehledu:
  - http://www.cuzk.cz/Predpisy/Pravni-predpisy-v-oboru-zememerictvi-a-katastru.aspx
- Resortní předpisy a opatření – 43 dokumenty ze seznamů:
  - http://www.cuzk.cz/Predpisy/Resortni-predpisy-a-opatreni/Pokyny-CUZK-1-15.aspx
  - http://www.cuzk.cz/Predpisy/Resortni-predpisy-a-opatreni/Pokyny-CUZK-16-30.aspx
  - http://www.cuzk.cz/Predpisy/Resortni-predpisy-a-opatreni/Pokyny-CUZK-31-42.aspx
- Datové sady – 5 dokumentů z geoportálu:
  - http://geoportal.cuzk.cz/(S(qn4tqoqg02oega1vqydc1o4g))/Default.aspx?head_tab=sekce-02-gp&mode=TextMeta&text=dSady_uvod&menu=20&news=yes
- INSPIRE – 14 dokumentů z geoportálu:
  - http://geoportal.cuzk.cz/(S(lyfis0b5dkkvrox2b5b1q2h2))/Default.aspx?head_tab=sekce-04-gp&mode=TextMeta&text=inspire_uvod&menu=40&news=yes

Korpus byl sestaven a indexován pro rychlé vyhledávání v korpusovém manažeru *Manatee/Bonito* [11]. Celková velikost korpusu je nyní 12 689 548 pozic (9 757 005 slov, z toho 3 864 481 podstatných jmen) v 27 389 dokumentech.

Dále byly sestaveny nové jednojazyčné specializované korpusy z oblasti zeměměřictví a katastru nemovitostí v ostatních jazycích Tezauru, viz Tabulka 2.4. Data byla získána metodou *WebBootCaT* z internetových zdrojů, výchozí klíčová slova pochází z překladů slovníku *VÚGTK, v.v.i.* Korpusy byl sestaveny a indexovány pro rychlé vyhledávání v korpusovém manažeru *Manatee/Bonito*.

Tabulka 2.4 Statistika jednojazyčných korpusů pro ostatní jazyky Tezauru

Jazyk	**Dokumentů**	**Pozic**	**Zastoupeno webových domén**
angličtina	7 948	29 784 018	4 823
francouzština	5 265	15 357 934	3 259
němčina	3 335	8 709 6020	2 191
ruština	2 914	19 015 734	1 770
slovenština	2 943	10 252 449	1 528

Data v cizojazyčných korpusech jsou rozdělena dle použitých klíčových slov do subdomén, podrobnou statistiku viz Tabulka 2.5.

Korpusová data jsou využita pro automatickou extrakci a návrhy nových termínů (viz 2.4.2) a návrhy překladových kandidátů (viz 2.4.4).

Tabulka 2.5 Podrobná statistika cizojazyčných korpusů dle subdomén

Jazyk	**angličtina**		**francouzština**		**němčina**		**rušt ina**		**slovenština**	
Subdoména	dok.*	pozic	dok.*	pozic	dok.*	pozic	dok.*	pozic	dok.*	pozic
katastr nemovitostí	787	4 092 896	327	1 051 295	469	981677	182	654 650	449	1 432 914
kartografie	1 409	8 885 249	867	2 220 351	867	1 752 001	334	1 727 503	472	1 848 648
inženýrská geodézie	278	1 926 008	413	1 017 208	402	909 902	186	1 481 491	208	641 084
teorie chyb	424	1 032 637	229	456 889	121	859 374	219	2 001 121	44	141 331
geodézie	1 592	5 499 426	1335	3 772 204	55	362 269	534	5 151 819	619	2 059 502
geografická informace	1 058	2 298 700	300	1 964 178	260	732 265	496	2 702 370	176	781 635
globální navigační družicový systém	627	1 621 525	263	822 796	28	54 960	83	193 516	98	337 776
přístrojová technika	569	1 668 899	226	350 265	71	398 802	142	1 153 706	161	247 216
mapování	542	3 625 655	523	1 819 638	715	1 408 478	214	939 055	354	1 325 619
metrologie	494	2 016 289	217	848 343	212	662 707	370	1 715 243	75	372 530
fotogram-metrie a DPZ	240	858 757	566	1 024 089	135	583 196	182	1 268 450	287	994 806

*Pozn.: * zkr. “dokumentů”*

2.4 Přínosy aplikace pro tvorbu terminologické databáze

Přínosy aplikace Tezaurus lze vidět v možnostech tvorby terminologické databáze z oblasti působnosti ČÚZK. Jedním z výsledků je vytvoření hierarchie termínů v Tezauru, která popisuje hyperonymické vztahy mezi jednotlivými termíny. Užití algoritmů pro extrakci znalostí z textu přispívá k efektivnějšímu procesu vytváření terminologické databáze a překladového slovníku odborných termínů.

2.4.1 Hierarchie termínů v Tezauru

Stávající slovník TS VÚGTK (viz 2.3.1.1) byl doplněn o návrhy hyperonym, která byla v první fázi získána z *Hesláře* VÚGTK (viz 2.3.1.2) a v další fázi semiautomaticky kontrolována a doplněna. Pro každý termín se nejprve extrahoval jeden nebo více nadřazených pojmů podle stromové struktury a pro termíny, které se nenacházejí v hesláři, byl navržen pojem z domény. Slovník byl rovněž rozšířen o pojmy z hesláře, včetně překladů, odkazů a zbylých informací, a momentálně obsahuje přes 8 tisíc unikátních termínů vzájemně spojených do stromové struktury. Vrcholová hierarchie Tezauru byla kontrolována a upravena na základě informací v definicích i podle automatických návrhů vazeb tak, aby vytvořená hierarchie lépe popisovala hyperonymické vztahy mezi jednotlivými termíny. Všechna hesla byla rovněž kontrolována a převedena do nové struktury.

Níže uvádíme nejvyšší tři patra hierarchie Tezauru.

- zeměměřictví
  - geodézie
    - typy geodézie
      - nižší geodézie
      - vyšší geodézie
    - obory geodézie
      - kosmická geodézie
      - geodetická astronomie
      - fyzikální geodezie = fyzikální geodézie = geofyzika
      - geodynamika
      - inženýrská geodezie = inženýrská geodézie
    - teorie chyb
    - vyrovnávací počet
    - globální navigační družicový systém
      - technologie GNSS
      - určení třírozměrné polohy
      - parametry
    - geoinformatika
  - kartografie
    - obory kartografie
      - matematická kartografie
      - tematická kartografie
      - všeobecná kartografie
      - praktická kartografie
      - teoretická kartografie
      - počítačová kartografie
    - kartologie
      - dějiny kartografie
    - dokumentace
      - záznamy dokumentace
      - mapa

            - standardizace
            - symboly
            - projekce
            - kartometrie
        - technologie
            - tisk
            - formáty
            - postupy a metody
- katastr nemovitostí
    - katastr nemovitostí ČR
    - pozemkový katastr
    - nemovitost
    - veřejné knihy
- obecné termíny

2.4.2 Automatická extrakce a návrh nových termínů

Aplikace využívá seznam termínů extrahovaný systémem *Sketch Engine* běžícího na serveru pomocí dotazování korpusu zeměměřictví. Ke každému termínu ze seznamu hesel Tezauru aplikace nabízí funkce založené na korpusu:

- ukázky použití termínů ve větách, tzv. konkordanci,
- termíny vyskytující se ve stejných kontextech (slova sdílející kolokace s termínem, pouze jednoslovné termíny).

Seznam kandidátů na (nové) termíny je extrahován z korpusu zeměměřictví a katastru nemovitostí metodami srovnávání korpusů a extrakcí klíčových slov. [12, 13] Četnost slov a jmenných frází ve specializovaném doménovém korpusu srovnáváme s četností týchž slov a jmenných frází ve velkém obecném (nespecializovaném) korpusu *czTenTen12* [14]. Nejlepší kandidáti na termíny mají nejvyšší podíl četností. Navrhované termíny nejvíce charakteristické pro cílovou doménu jsou do aplikace začleněny taktéž pomocí *Sketch Engine*.

Prvních 100 kandidátů na víceslovné termíny, které byly automaticky extrahovány, je uvedeno v Tabulce 2.6.

Tabulka 2.6 Ukázka automaticky extrahovaných kandidátů na víceslovné termíny

státní správa zeměměřictví	podrobné měření	zvláštní předpis
správa zeměměřictví	lomový bod	výškový systém
katastrální mapa	český úřad	svaz geodetů
katastrální úřad	permanentní stanice	výměnný formát
zeměměřická činnost	věcné právo	nivelační přístroj
bodové pole	geodetický základ	vojenské mapování
katastrální operát	kartografická společnost	pozemková evidence
pozemková úprava	parcelní číslo	zemědělský půdní fond
souřadnicový systém	střední chyba	dálkový přístup
geometrický plán	list vlastnictví	soubor popisných informací
katastr nemovitostí	geografická informace	stabilní katastr
podrobný bod	topografická mapa	zeměměřická knihovna
mapový list	druh pozemku	geografický informační systém
prostorové datum	polohové bodové pole	Karlův vary
pozemkový úřad	dálkový průzkum	měření délek
katastrální zákon	hranice pozemků	kód kvality
podrobná informace	geodetická informace	český svaz geodetů
státní správa	bodová pole	digitální katastrální mapa
zeměměřický úřad	úřední oprávnění	svaz vědeckotechnických společností
referenční stanice	evidence nemovitostí	český svaz vědeckotechnických společností
model terénu	laserové skenování	digitální model terénu
určování polohy	výsledek zeměměřických činností	vědeckotechnická společnost
pozemková kniha	základní mapa	katastrální vyhláška
katastrální pracoviště	systematická chyba	zeměměřický inženýr
údaj katastru	nejistota měření	kartografická konference
obnova katastrálního operátu	tíhové pole	cestovní zpráva
totální stanice	obor geodézie	člen českého svazu
katastrální území	inženýrská geodézie	katastrální inspektorát
referenční systém	polygonový pořad	zjednodušená evidence
pozemkový katastr	seznam souřadnic	geodetická firma
digitální model	digitální mapa	právní vztah
mapové dílo	soubor geodetických informací	geodetické měření
geodetická práce	popisná informace	
identický bod	půdní fond	

2.4.3 Návrhy hyperonymických vazeb mezi termíny

Návrh nejpravděpodobnějších kandidátských hyperonymických vazeb pro daný (nový) termín je součástí editačního formuláře. Uživatel má možnost hyperonymum přidat buď dle svého výběru nebo si nechat aplikací navrhnout několik nejpravděpodobnějších kandidátů.

Samotná technika návrhu se přitom řídí dvěma zdroji údajů:

- Vazbami naučenými automaticky z textů korpusů.
- Vazbami ze stávajícího Tezauru podle podobnosti zařazovaného termínu s již zařazenými termíny.

Zjišťování vazeb z textů korpusů je provedeno vyhledáváním vzorů ve specializovaném korpusu jako např.:

- *hyponymum* “je”/”jsou” *hyperonymum*
- *hyponymum* “a”/”nebo”/”či” “další”/”jiný”/”ostatní”/”podobný” *hyperonymum*

K vyhodnocení metody bylo extrahováno 50 nejvýznamnějších vazeb odpovídajících každému z těchto vzorů. 60 % kandidátů podle prvního vzoru, respektive 56 % podle druhého vzoru, bylo skutečně hypo/hyperonymickou vazbou.
Většina termínů je však v korpusu málo frekventovaná, tedy vazba nemusí být v korpusu vůbec přítomna. Proto technika návrhu kandidátských hyperonym využívá i vazby ze stávajícího Tezauru, na základě podobnosti zařazovaného termínu s již zařazenými termíny.

2.4.4 Návrhy překladových kandidátů z korpusů

Pro návrh překladových kandidátů je obvykle zapotřebí mít k dispozici zarovnaný korpus pro sledovaný jazykový pár, přičemž zarovnání je na úrovni vět nebo odstavců. Jelikož korpusy výše zmíněné byly vytvořeny odděleně, dokumenty v nich obsažené nejsou navzájem kompatibilní – nejedná se o vzájemné překlady. Korpusy se shodují pouze doménou (velmi pravděpodobně neobsahují ani jeden dokument, který by byl přeložen do jiného jazyka a byl obsažen v odpovídajícím korpusu) – takové korpusy se označují jako srovnatelné (comparable corpora). Běžně dostupné algoritmy pro vyhledávání překladových kandidátů termínů pracují s paralelními korpusy, nikoli se srovnatelnými, bylo tedy nutné vyvinout nový algoritmus, který překladové kandidáty navrhne pouze na základě extrahovaných termínů v jednotlivých jazycích a pomocí dostupných srovnatelných korpusů. Metody pro návrh překladových termínů s využitím skutečně paralelních korpusů (překladů) vykazují vyšší úspěšnost, vyžadují ovšem velké množství přeložených textů.

Algoritmus pro návrhy překladových ekvivalentů ze srovnatelných korpusů

Algoritmus využívá faktu, že ekvivalentní termíny (v tomto případě z domény geodézie, kartografie atd.) se často vyskytují se stejnými kolokacemi. Např. termín měření (anglicky measurement) se vyskytuje často se slovy chyba, přesnost, metoda. Ve srovnatelném anglickém korpuse se vyskytuje ekvivalent measurement často se slovy error, accuracy, method. Tyto kolokace jsou překlady kolokací v českém korpusu. Pomocí překladového slovníku lze tedy zjistit jaké dva termíny sdílí nejvíce kolokací a na základě této statistiky vytvořit seznam nejvhodnějších překladových kandidátů pro další ruční kontrolu a začlenění do Tezauru.

Algoritmus prochází postupně pro daný jazykový pár všechny možné páry termínů v prvním a druhém jazyce a zjistí velikost množiny společných kolokací (společná kolokace je taková, kdy kolokace v prvním jazyce lze přeložit pomocí dostupného slovníku na nějakou kolokaci v druhém jazyce). Výstupem algoritmu je seznam všech možných překladových párů utříděný podle velikosti množin společných kolokací.

Tento algoritmus nebude nikdy v principu dosahovat takové kvality extrakce kandidátských překladů jako kdyby se pracovalo se zarovnanými korpusy. Tabulka 2.7 obsahuje příklad seznamu překladových kandidátů pro jazykový pár angličtina-čeština. První sloupec obsahuje počet společných kolokací, tučně jsou vyznačené správné překlady.

Tabulka 2.7 Ukázka automaticky vytvořených kandidátských překladů pro jazykový pár angličtina-čeština

24	**land**	**nemovitost**
24	**earth**	**zem**
24	data	datum
24	**calculation**	**výpočet**
23	**map**	**mapa**
23	input	výpočet
22	distance	měření
21	reference	obr
21	parameter	výpočet
21	**measurement**	**měření**
20	**terrain**	**terén**
20	analysis	datum
20	**accuracy**	**přesnost**
19	estimation	výpočet
19	estimate	výpočet
19	**computation**	**výpočet**
19	**cadastre**	**katastr**

Při vyhodnocení úspěšnosti algoritmu na termínech, které se vyskytují v seznamech kandidátských termínů pro češtinu i pro daný cílový jazyk, dosahuje metoda pro nejlepších 20 shod následující přesnost: angličtina 34 %, francouzština 21 %, němčina 40 %, ruština 24 % a slovenština 47 %. Správnost překladu se kontrolovala vůči překladům, které jsou obsaženy v Tezauru.

3 Srovnání novosti

Předkládaná metodika je výsledkem unikátní kombinace uplatnění nejnovějších technik oblasti jazykového inženýrství a návrhu aplikace ověřených lexikografických postupů při tvorbě, udržování a využití dat systému *Tezauru pro obor zeměměřictví a katastru nemovitostí*.

Metodika popisuje využití systému Tezauru jednak pro prezentaci dat technologických termínů a vztahů mezi nimi pro účely zpracování odborných textů i pro informování široké veřejnosti, a jednak možnosti a techniky udržování obsahu Tezauru do budoucna s možnostmi automatických návrhů nových termínů v publikovaných textech i podpory jejich kandidátských překladů do dalších jazyků.
Novost metodiky tak spočívá v systematickém návrhu práce s Tezaurem v moderním softwarovém systému využívajícím nejnovější výsledky v oblasti zpracování jazykových a lexikografických dat. Tato kombinace technik je jedinečná nejen v rámci vývoje podobných systémů v ČR, ale i ve světě. Předchozí výsledky (Terminologický slovník a Heslář VÚGTK) neumožňovaly práci s odbornými termíny v takovém rozsahu. Oproti těmto výsledkům je nová zejména hierarchie hesel, překlady hesel, zařazení hesel do domén, zobrazení příkladů užití pro heslo i jeho překlady. Z hlediska vytváření terminologických databází jsou nové zejména funkce automatické extrakce klíčových slov, návrhy hyperonymických vazeb a návrhy překladových kandidátů.

Tvůrce metodiky i systému Tezauru, tedy tým *Centra zpracování přirozeného jazyka na Masarykově univerzitě v Brně*, zde zúročil mnohaleté zkušenosti se základním i aplikovaným výzkumem v oblasti zpracování jazykových dat i tvorby slovníkových aplikací (podobné nástroje vytvořené *Centrem ZPJ MU* využívají přední světoví vydavatelé slovníků, jako je např. *Oxford University Press* nebo *MacMillan*).

Automatické techniky vyvinuté pro účely Tezauru jsou založené na výsledcích vlastního i mezinárodního výzkumu z posledních deseti let a zejména jejich aplikace na český jazyk a na danou terminologickou oblast tak reprezentuje zcela unikátní technologie.

4 Uplatnění metodiky

Metodika je určena pro široké spektrum uživatelů, od členů hlavního terminologického orgánu v dané oblasti, tedy *Terminologické komise ČÚZK*, přes odborníky z oblasti zeměměřictví až po širokou veřejnost při styku s informacemi zejména ve vztahu ke katastru nemovitostí, registru územní identifikace, adres a nemovitostí nebo obecně terminologie geodézie a kartografie.

Hlavní část metodiky je věnována popisu metodických postupů při tvorbě, údržbě a rozšiřování terminologické databáze oboru zeměměřictví a katastru nemovitostí. Postupy jsou prezentovány pomocí primárních funkcí systému Tezauru zaměřených jednak na obecnou správu systému, pravidla pro různé úrovně práce se systémem, technický popis konfigurací, a jednak na údržbu a doplňování dat Tezauru, a to buď přímou editací termínů, jejich překladů, definicí a vazeb, nebo na základě automatických metod navrhujících nové kandidátské termíny z textů, jejich překlady do pěti jazyků, heuristické určení provázání nových termínů na stávající databázi Tezauru apod.

Část metodiky popisující využití dat Tezauru odbornou i širokou veřejností představuje jednak možnosti uživatelského prohlížení a vyhledávání v Tezauru, a jednak popisuje parametry standardizovaného aplikačního rozhraní pro strojové zpracování obsahu Tezauru v libovolných navazujících aplikacích.

5 Seznam literatury

[1] BALÍKOVÁ, Marie. Tezaurus. In: KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV) [online]. Praha : Národní knihovna ČR, 2003- [cit. 2015-10-08]. Dostupné z: http://aleph.nkp.cz/F/?func=direct&doc_number=000001649&local_base=KTD.

[2] RAMBOUSEK, Adam. Creation and Management of Structured Language Resources [online]. Brno, 2015 [cit. 2015-09-25]. Disertační práce. Masarykova univerzita, Fakulta informatiky. Vedoucí práce Aleš Horák. Dostupné z: http://is.muni.cz/th/60380/fi_d/.

[3] Kilgarriff, Adam, et al. The Sketch Engine: Ten Years On. In Lexicography (2014): 1-30. (2001): 1-37.

[4] David Flanagan and Yukihiro Matsumoto. The Ruby Programming Language. O'Reilly, first edition, 2008.

[5] A. Fomichev, M. Grinev, and S. Kuznetsov. Sedna: A Native XML DBMS. Lecture Notes in Computer Science, 3831:272, 2006.

[6] Scott Boag, Don Chamberlin, Mary F. Fernández, Daniela Florescu, Jonathan Robie., and Jérôme Siméon. XQuery 1.0: An XML Query Language (Second Edition), 2010. Dostupné z: http://www.w3.org/TR/xquery.

[7] Douglas Crockford. JSON, The Fat-Free Alternative to XML. In Proceedings of XML 2006, Boston, USA, 2006. Dostupné z: http://www.json.org/xml.html.

[8] James Clark. XSL Transformations (XSLT) Version 1.0, 1999. Dostupné z:http://w3.org/TR/xslt.

[9] Pomikálek, Jan. Removing boilerplate and duplicate content from web corpora. Disertační práce, Masarykova univerzita, Fakulta informatiky (2011).

[10] Baroni, Marco, Adam Kilgarriff, Jan Pomikálek, and Pavel Rychlý. WebBootCaT: instant domain-specific corpora to support human translators. In *Proceedings of EAMT*, pp. 247-252. 2006.

[11] Rychlý, Pavel. Manatee/Bonito - A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masarykova univerzita, 2007. p. 65-70. ISBN 978-80-210-4471-5.

[12] Kilgarriff, Adam. Comparing Corpora. In *International Journal of Corpus Linguistics* 6 (1)

[13] Kilgarriff, Adam. Simple maths for keywords. In *Proceedings of Corpus Linguistics*. 2009.

[14] Suchomel, Vít. Recent Czech Web Corpora. In *Proc. 6th Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Tribun EU, 2012. s. 77-83.

6 Seznam publikací

Aleš Horák, Adam Rambousek, Vít Suchomel a Lucia Kocincová.
Semiautomatic Building and Extension of Terminological Thesaurus for Land Surveying Domain.
In Aleš Horák, Pavel Rychlý. *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Tribun EU, 2014. s. 129-137, 9 s. ISSN 2336-4289.
Výsledek vznikl v rámci specifického výzkumu na vysoké škole.

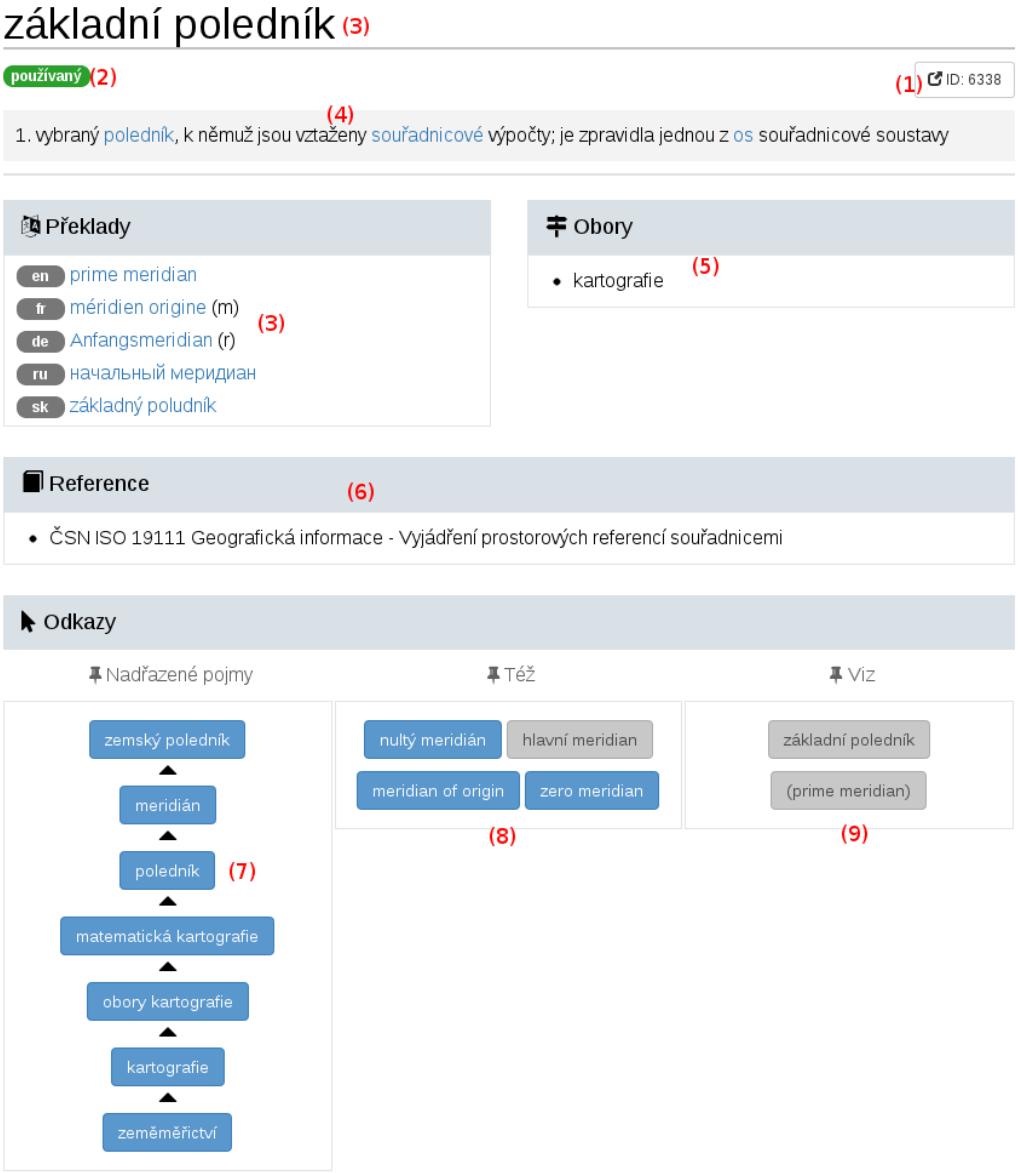
Příloha č. 1 Struktura hesla

Hesla tezauru se ukládají do databáze v jednotném formátu XML. V této kapitole jsou popsány všechny prvky, které lze použít ve struktuře jednoho hesla. Formálně je struktura popsána pomocí DTD schématu. Díky návrhu formátu XML je struktura hesla v případě budoucí potřeby snadno rozšiřitelná o další položky.
Pro efektivní a rychlou práci nejen s heslem, ale i s celým systémem, každé heslo obsahuje nejen základní atributy a vazby, ale i odkazy na zdroje dat a možnost přidávaní (systémových i uživatelských) komentářů.

Základní struktura položek jednoho hesla má stromovou strukturu a obsahuje následující položky (u každé položky uvádíme i odkaz na příslušnou odpovídající část v následující obrázek hesla):

- Kořenový element *entry* pro uložení záznamu jednoho hesla. Obsahuje atributy: *id* - unikátní identifikační číslo hesla, používá se pro vyhledávání hesel a pro odkazy mezi hesly (na obrázku **(1)**), *status* - označení stavu hesla (povolené hodnoty: 1=terminologický termín, 2=používaný termín, 3=používaný nezpracovaný termín, 4=zamítnutý termín, 5=automaticky navržený termín) (na obrázku **(2)**).
  Dále obsahuje následující položky:
  - element *terms* pro jednotlivé termíny
  - element *defs* pro definice
  - element *domains* pro obory
  - element *references* pro odkazované zdroje
  - element *hyper* pro odkazy na nadřazené pojmy
  - element *alsos* pro odkazy na příbuzná hesla
  - element *sees* pro odkazy na podobná hesla
  - element *notes* pro poznámky o historii editace.
- Element *terms* obsahuje alespoň jeden element *term*. (na obrázku **(3)**)
  - Element *term* slouží pro uložení jednoho termínu. Termín se zadává jako textová hodnota elementu. Dále může obsahovat atributy: *lang* - jazyk termínu, *number* - pořadové číslo termínu, *refno* - odkaz na jiný termín (např. propojení zkratky a plného rozepsání zkratky), *abbr* - informace, zda je termín zkratka, *morf* - doplňující morfologická informace (např. slovní druh).
- Element *defs* může obsahovat libovolný počet elementů *def*. (na obrázku **(4)**)
  - Element *def* slouží pro uložení vysvětlení jednoho významu hesla, jedné definice. Zadává se jako textová hodnota elementu. Dále může obsahovat číselný atribut *num* - pořadové číslo významu.
- Element *domains* může obsahovat libovolný počet elementů *dom*. (na obrázku **(5)**)
  - Element *dom* slouží k uložení jednoho oboru, pod který daný pojem spadá. Zadává se jako textová hodnota elementu.
- Element *references* může obsahovat libovolný počet elementů *ref*. (na obrázku **(6)**)
  - Element *ref* slouží k uložení jednoho odkazu na zdroj daného pojmu. Zadává se jako textová hodnota elementu.
- Element *entry* může obsahovat libovolný počet elementů *hyper*. (na obrázku **(7)**).
  - Element *hyper* slouží k uložení odkazu na jeden nadřazený pojem. Jako povinný atribut *id* se zadává identifikátor nadřazeného pojmu.
- Element *alsos* může obsahovat libovolný počet elementů *also*. (na obrázku **(8)**)
  - Element *also* slouží k uložení jednoho odkazu na příbuzné heslo, zadává se jako textová hodnota elementu. Může obsahovat atribut *lang* pro upřesnění jazyka.

- Element *sees* může obsahovat libovolný počet elementů *see*. (na obrázku **(9)**)
    - Element *see* slouží k uložení jednoho odkazu na podobné heslo, zadává se jako textová hodnota elementu. Může obsahovat atribut *lang* pro upřesnění jazyka.
- Element *notes* může obsahovat libovolný počet elementů *note*. (ve výchozím nastavení se uživateli nezobrazuje)
    - Element *note* slouží pro uložení informace o jedné změně hesla. Komentář ke změně se ukládá jako textová hodnota. Dále obsahuje atributy: *number* - pořadové číslo změny, *xpath* - určuje část hesla, ke které se poznámka vztahuje, *refno* - přesnější určení upravené části hesla, *author* - editor, který provedl změnu, *time* - datum a čas úpravy.



Zobrazení závislostí prvků ve stromové struktuře

entry, @id, @status

- terms
  - term, @lang, @number, @refno, @abbr, @morf
- defs (definice)
  - def, @num
- domains (obory)
  - dom
- references (zdroje)
  - ref
- hyper, @id (id nadřazeného termínu)
- alsos
  - also, @lang
- sees
  - see, @lang
- notes (komentáře, poznámky, historie změn)
  - note, @number, @author, @time, @xpath, @refno
  - xpath odkazuje na část hesla, ke kterému se poznámka vztahuje, proto jsou všechny části jednoznačně číslované @number
  - @refno je číslo zdroje v references.ref v případě, že se poznámka týká importu

Ukázka konkrétního hesla v XML formátu (interní formát Tezauru)

```
<entry id="3203">
  <terms>
    <term lang="cz" number="1">měřický bod</term>
    <term lang="en" number="2" refno="1">survey point</term>
    <term lang="fr" number="3" refno="1">point m géodésique</term>
    <term lang="de" number="4" refno="1">Vermessungspunkt r</term>
    <term lang="sk" number="5" refno="1">meračský bod</term>
  </terms>
  <meta>
    <create_time>2014-05-16 12:01</create_time>
    <author>auto import</author>
  </meta>
  <defs>
    <def lang="cz" number="1">bod kteréhokoliv z bodových polí,
     který tvoří podklad pro další měření; pokud splňuje podmínky
     stanovené ČSN 73 0415, nazývá se geodetický bod
    </def>
  </defs>
  <domains>
    <dom>geodézie</dom>
  </domains>
  <references>
    <ref>ČSN 73 0415 Geodetické body</ref>
  </references>
  <hyper id="1024"/>
  <notes>
```

```
    <note number="1" author="import" time="2014-05-16 12:01:13"
          xpath="/entry">automatický import hesla
    </note>
    <note number="2" author="novak" time="2014-05-22 14:10:27"
          xpath="/entry/terms/term" refno="5">doplněn slovenský
                překlad
    </note>
  </notes>
</entry>
```

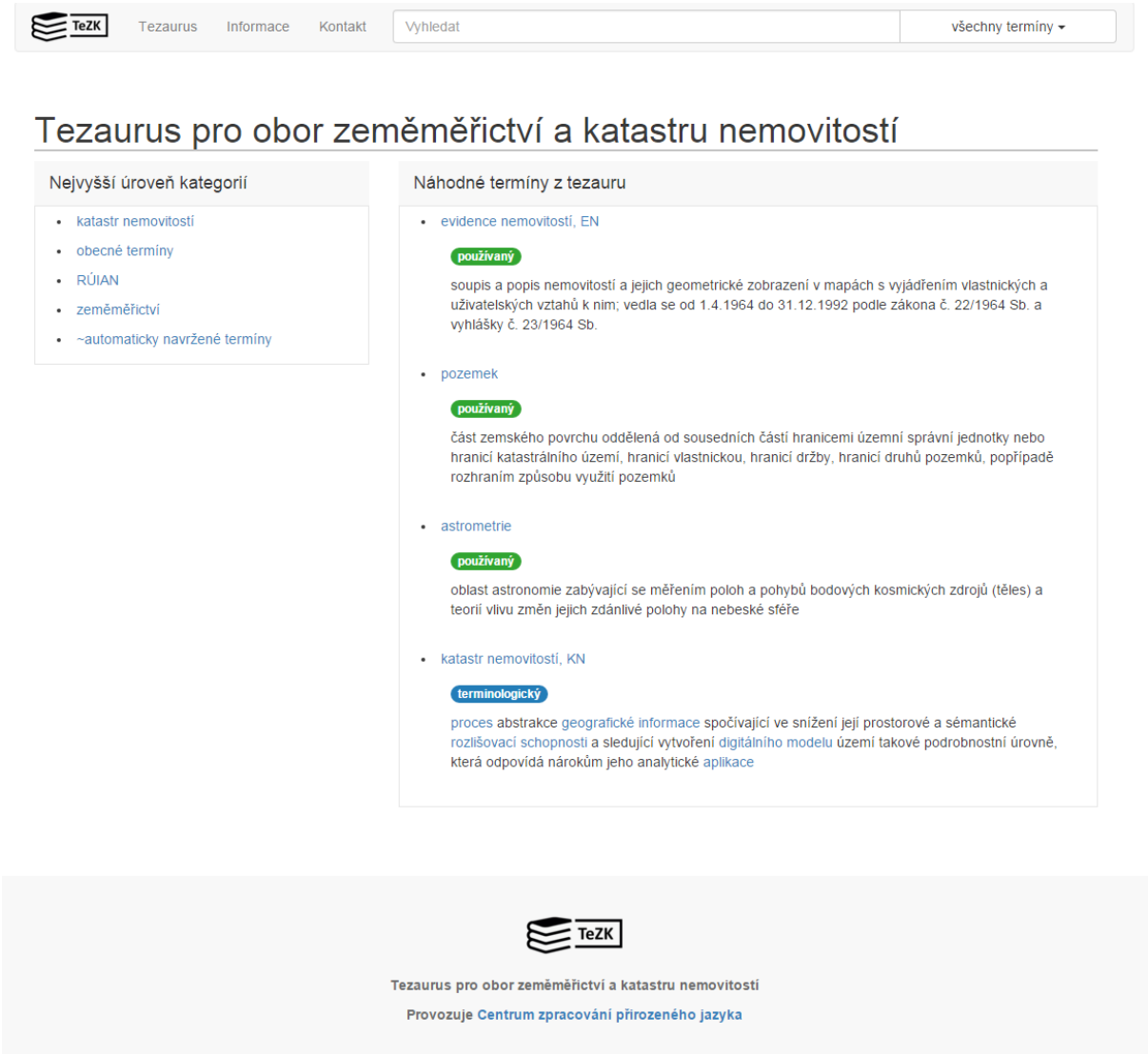
DTD schéma

```
<!ELEMENT entry (terms, meta, defs, domains, references, alsos, sees,
hyper+, notes)>
<!ATTLIST entry id CDATA #REQUIRED>
<!ELEMENT terms term+>
<!ELEMENT term PCDATA>
<!ATTLIST term
  lang   CDATA #REQUIRED
  number CDATA #REQUIRED
  refno CDATA #IMPLIED
>
<!ELEMENT meta (create_time, author+)>
<!ELEMENT author PCDATA>
<!ELEMENT create_time PCDATA>
<!ELEMENT defs (def+)>
<!ELEMENT def PCDATA>
<!ATTLIST def
  number CDATA #REQUIRED
  lang CDATA #REQUIRED
>
<!ELEMENT domains (dom+)>
<!ELEMENT dom PCDATA>
<!ATTLIST dom
  number CDATA #REQUIRED
  lang CDATA #REQUIRED
>
<!ELEMENT references (ref+)>
<!ELEMENT ref PCDATA>
<!ATTLIST ref
  number CDATA #REQUIRED
>
<!ELEMENT alsos (also+)>
<!ELEMENT also PCDATA>
<!ATTRLIST also
  lang CDATA #REQUIRED
>
<!ELEMENT sees (see+)>
<!ELEMENT see PCDATA>
<!ATTRLIST see
  lang CDATA #REQUIRED
>
<!ELEMENT hyper>
<!ATTRLIST hyper
  id CDATA #REQUIRED
```

```
>
<!ELEMENT notes (note+)>
<!ELEMENT note PCDATA>
<!ATTLIST note
  number CDATA #REQUIRED
  author CDATA #REQUIRED
  time CDATA #REQUIRED
  xpath CDATA #REQUIRED
  refno CDATA #IMPLIED
>
```

Příloha č. 2 Návod pro uživatele

Po otevření aplikace ve webovém prohlížeči na adrese `http://[jméno serveru]/tecu` se zobrazí úvodní stránka aplikace s možností prohledávání a hierarchického procházení Tezauru, stránka se základními informacemi o Tezauru a kontaktní informace.



Obrázek P2.1: Úvodní stránka aplikace

Záložka *Tezaurus* obsahuje slovníkové údaje z Tezauru, podrobnější popis níže.

Na záložce *Informace* se zobrazují podrobnější informace o Tezauru včetně aktuálních detailních statistik o datech a referenčních odkazech.

Záložka *Kontakt* obsahuje kontaktní údaje autorů a správců Tezauru.

Vyhledávání

Na všech záložkách je možno do pole *Vyhledat* vložit hledaný termín, který chcete najít v Tezauru. Již během psaní hledaného termínu do pole se nabízejí možnosti vyhledaných termínů. Pokud kliknete na některou z nabízených možností, zobrazí se úplný slovníkový záznam pro zvolené heslo.

Filtrování podle statutu hesla

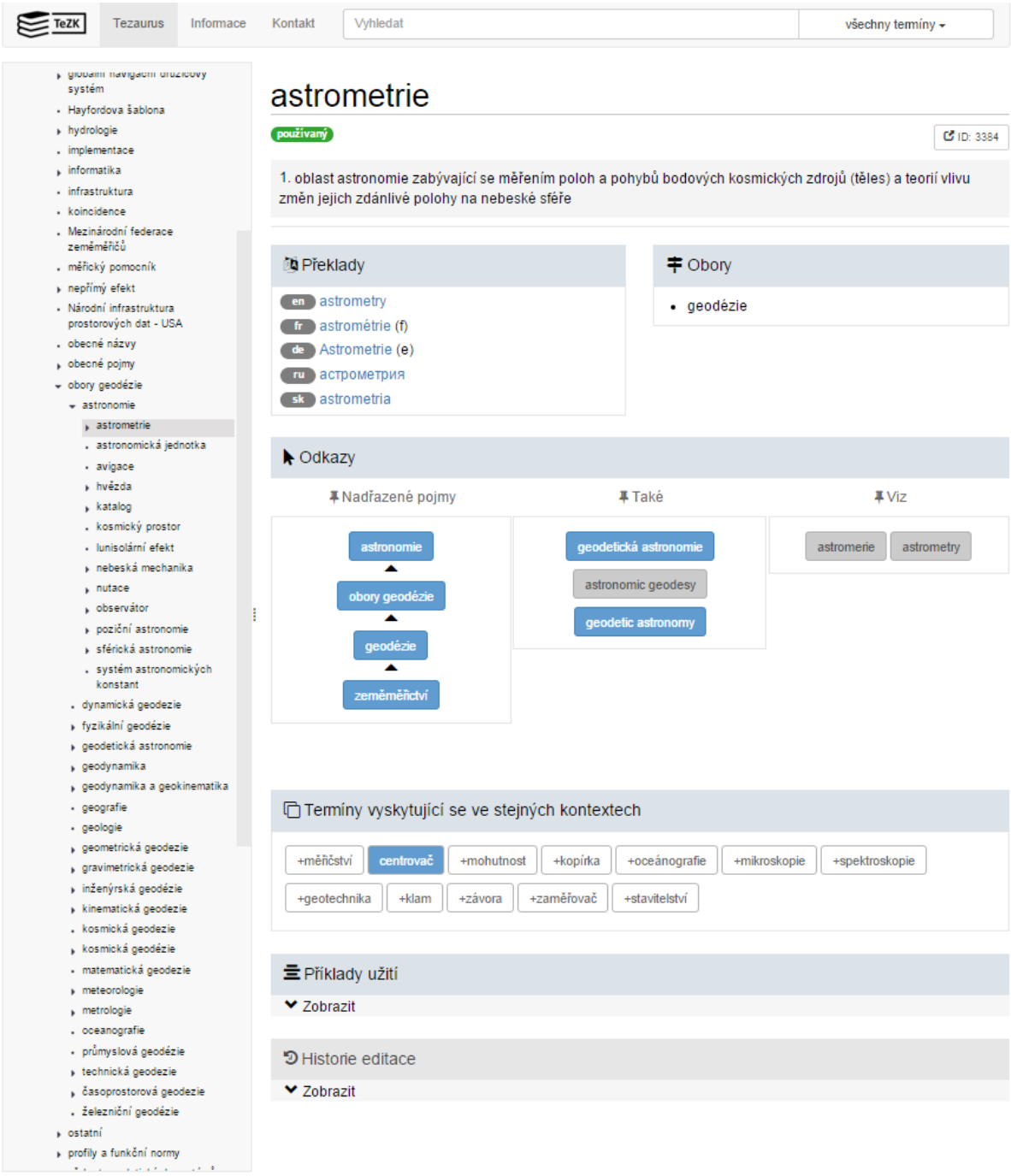
Každé heslo v Tezauru má určen statut podle důvěryhodnosti informací a termínu. Jsou evidovány čtyři statuty hesla:

- terminologické (schválené *Terminologickou komisí*)
- používané (používané, ale nezařazené mezi terminologické termíny)
- používané nezpracované (zatím nezpracované *Terminologickou komisí*)
- odmítnuté (zamítnuté/smazané kandidátní termíny z automaticky navržených)

Při práci s Tezaurem můžete pomocí výběru z nabídky (vpravo od pole *Vyhledat*) zvolit druh hesel, s nimiž chcete pracovat. Např. můžete prohlížet pouze terminologická hesla, pokud chcete v psaném textu používat pouze schválené termíny.

Strom Tezauru

Na záložce *Tezaurus* se v levé části zobrazuje strom hyperonymických vztahů mezi hesly Tezauru. Ve výchozím stavu vidíte pouze nejvyšší úroveň stromu. Pokud má termín podřízené termíny, zobrazuje se před názvem termínu šipka. Po kliknutí na šipku se otevře další úroveň stromu. Po kliknutí na název termínu se v pravé části okna zobrazí podrobné údaje o termínu.
Pokud do Tezauru vstoupíte kliknutím na výsledek vyhledávání nebo pomocí odkazu na určitý termín, zobrazí se strom termínů již otevřený na zvoleném hesle.



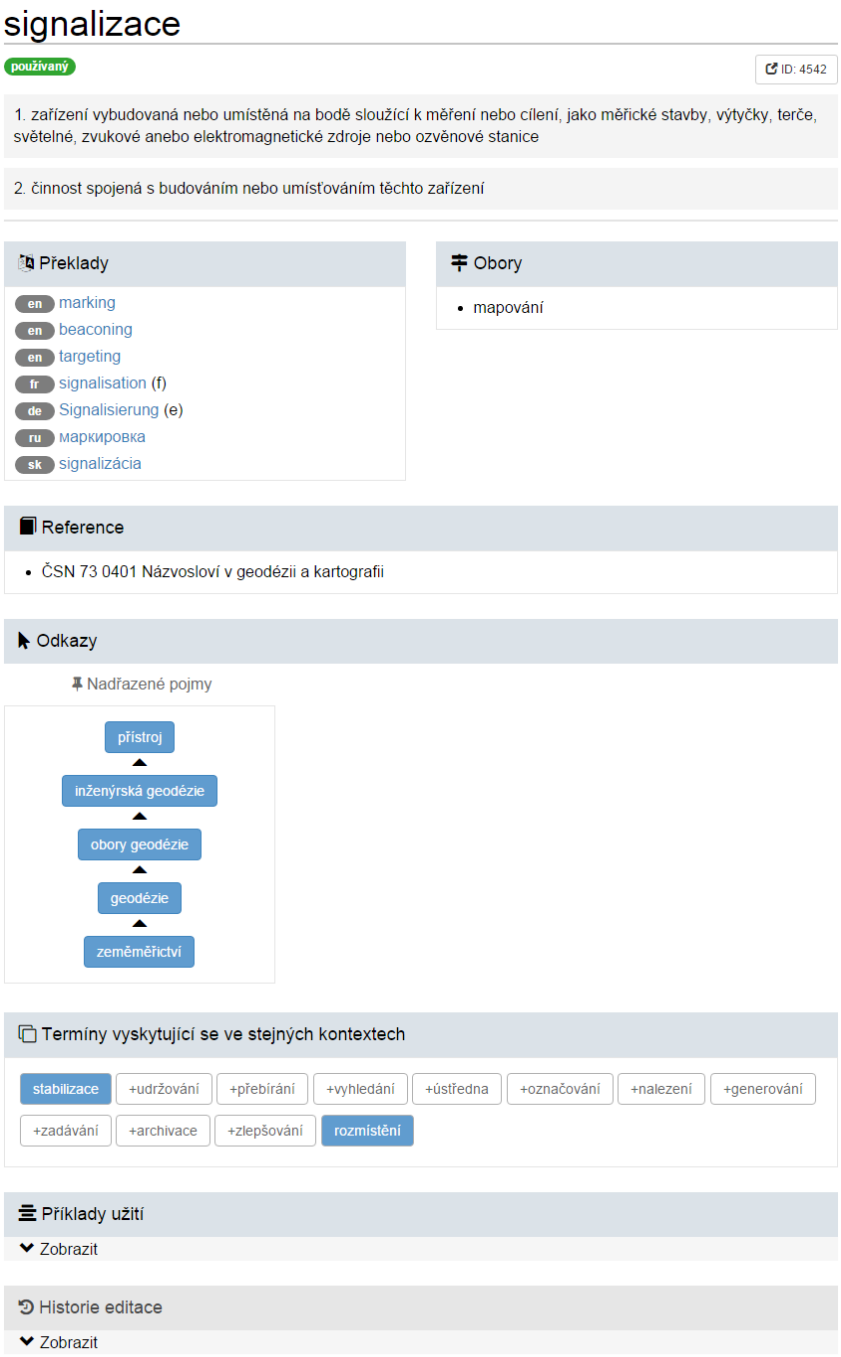
Obrázek P2.2: Strom Tezauru a zobrazení hesla

Podrobné informace o termínu

V záhlaví se nachází termín a jeho varianty. Pod záhlavím je informace o statutu termínu (vlevo) a ikona (vpravo), která slouží jako přímý odkaz na daný termín (tento odkaz můžete sdílet a příjemci se zobrazí zvolený termín bez nutnosti vyhledávání).

Pod záhlavím jsou vysvětlující definice termínu. Následují překlady termínu do dalších jazyků. Po kliknutí na přeložený termín se zobrazí ukázky použití termínu v textech získaných ze specializovaného korpusu z oboru zeměměřictví v konkrétním jazyce.

Pokud se termín vyskytuje v odborné literatuře nebo zdrojích, zobrazí se citační informace v části Reference.



Obrázek P2.3: Zobrazení termínu

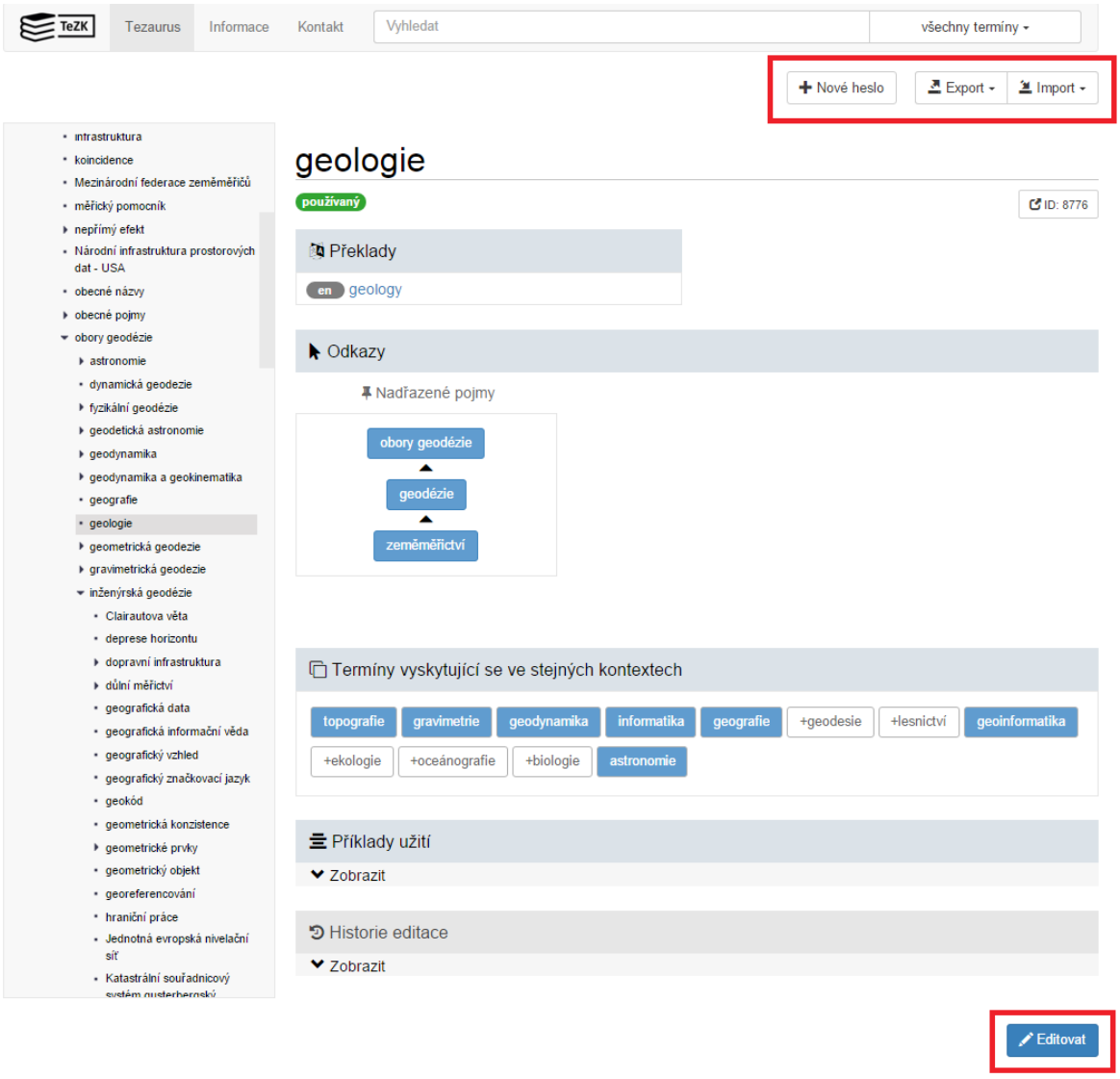
V části “*Odkazy*” se zobrazují nadřazené termíny pro daný termín a případně další odkazy na termíny. V části “*Příklady užití*” se zobrazí ukázky použití termínu v českém korpusu z oboru zeměměřictví (obrázek P2.4).
Pro jednoslovné termíny je možné pomocí korpusu zjistit podobná slova, resp. slova, která se vyskytují v podobných kontextech. Jazykově se tedy nejedná o plná synonyma, ale o slova, která se v určitých kontextech chovají podobně jako hlavní termín. Pokud je k dispozici dostatečný vzorek dat pro daný termín, zobrazí se v části “*Termíny vyskytující se ve stejných kontextech*” navržené podobné termíny. Termíny, které existují v Tezauru, se zobrazují modrou barvou a po kliknutí na termín se zobrazí podrobné údaje o tomto termínu. Termíny, které v Tezauru neexistují, se zobrazují šedou barvou se symbolem “+”. Kliknutím na tento odkaz se přidá odpovídající nové heslo (otevře se editační formulář s možnosti upřesnění dalších položek) do Tezauru. Seznam automatických synonym je uspořádán sestupně podle míry podobnosti s hlavním termínem.

Obrázek P2.4: Ukázka použití termínu v českém korpusu

Práce s údaji pomocí další aplikace / strojově zpracovatelná data
Pokud potřebujete údaje z Tezauru využít v další aplikaci nebo potřebujete získat data ve strojově zpracovatelné podobě, můžete použít aplikační rozhraní Tezauru (viz sekci 2.2.4 metodiky a Přílohu 4).

Příloha č. 3 Návod pro editory

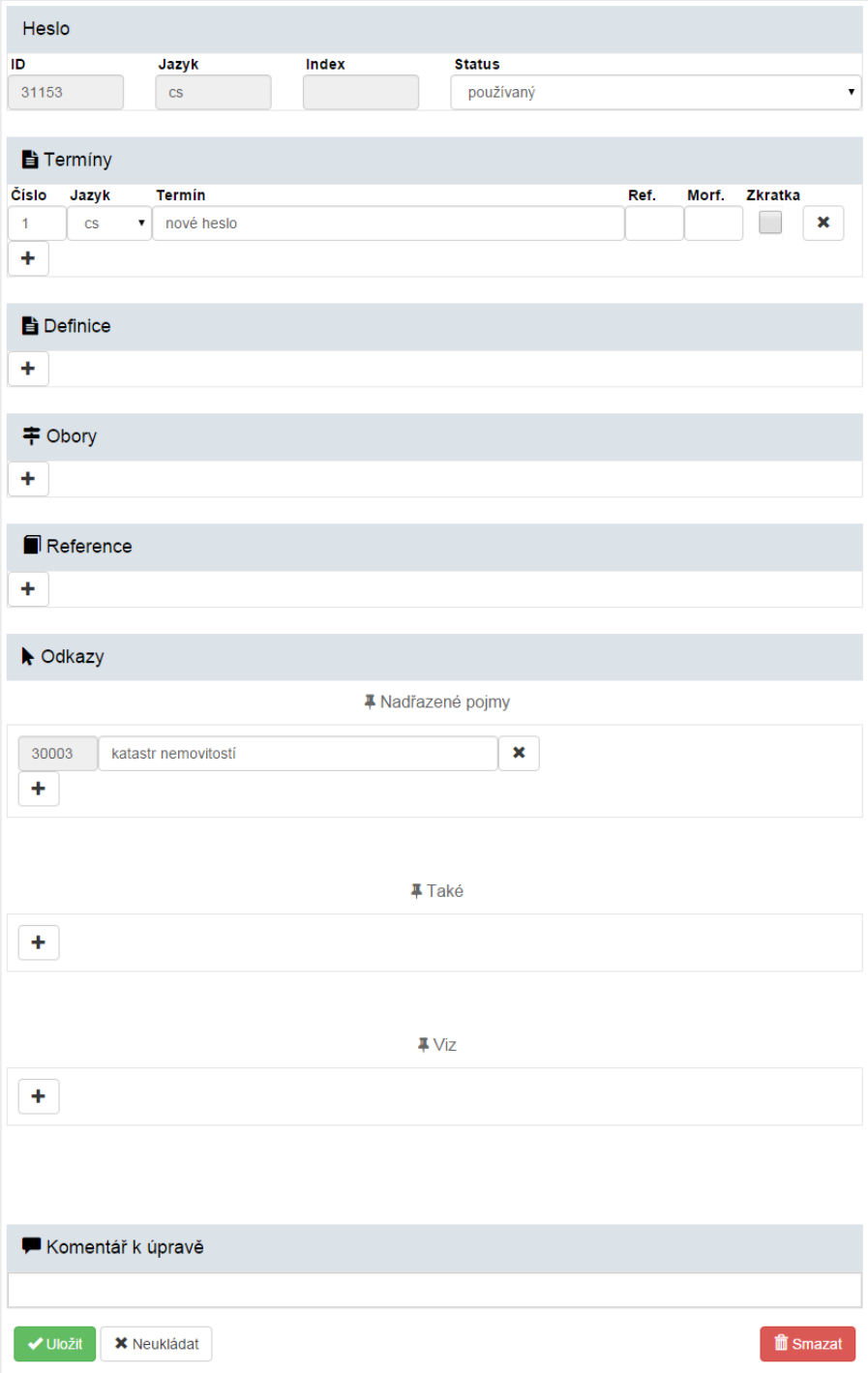
Uživatelé s oprávněním pro editaci hesel se přihlašují do zabezpečené části webové aplikace na adrese `http://[jméno serveru]/tecu`.
Po přihlášení se zobrazí webové rozhraní aplikace Tezaurus stejné jako pro nepřihlášené uživatele, s několika odkazy pouze pro editory, viz obrázek P3.1.



Obrázek P3.1: Rozhraní pro editory

Vytvoření nového hesla v Tezauru

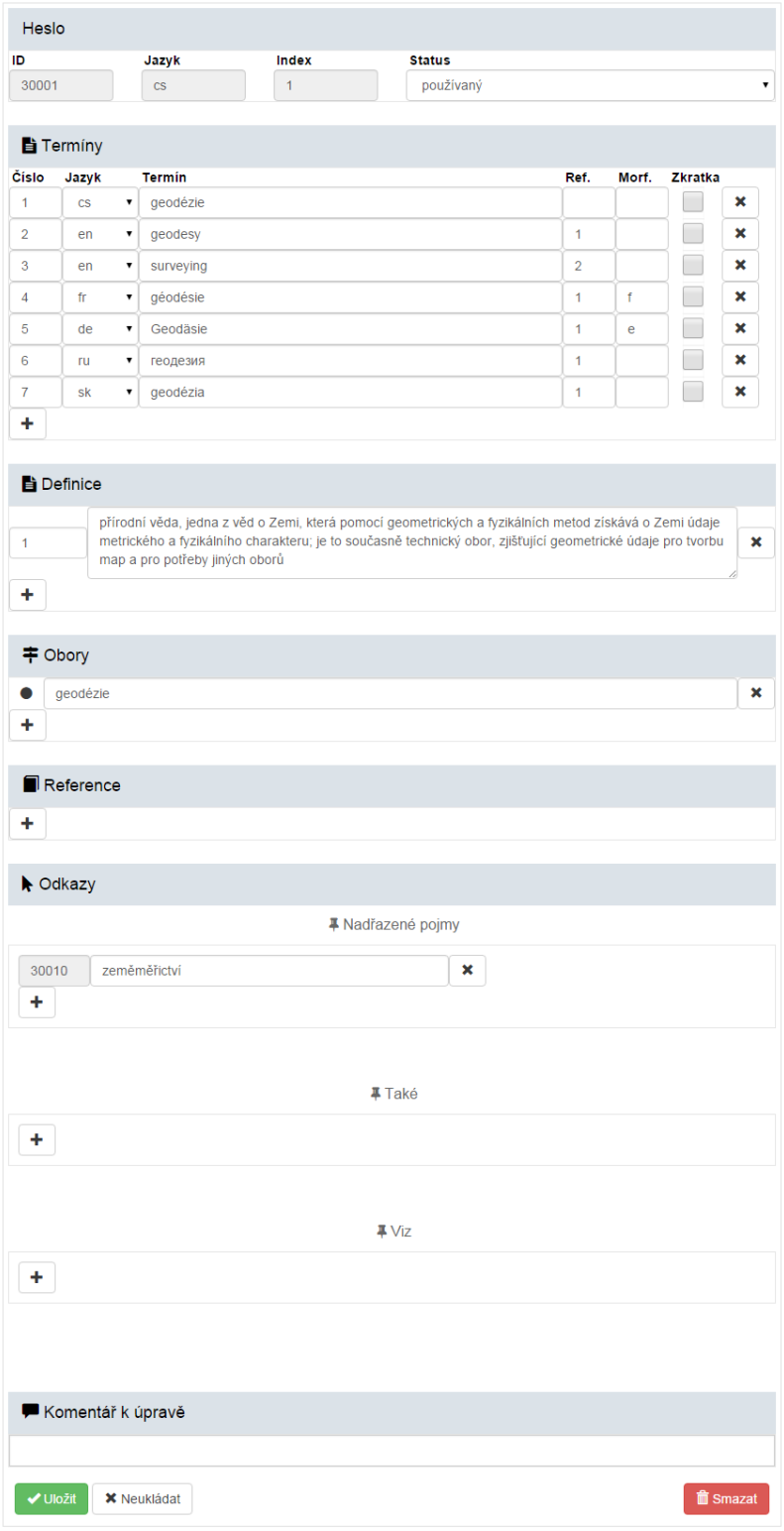
Po kliknutí na odkaz "*Nové heslo*" dojde k vytvoření nového prázdného hesla v Tezauru. Pokud je zobrazeno jiné heslo, nově vytvořené heslo bude mít přiřazené stejný nadřazený pojem. Poté klikněte na odkaz "*Editovat*" pro úpravu informací o hesle.



Obrázek P3.2: Editační formulář pro nově vytvořené heslo

Editace hesla v Tezauru

Po vyhledání hesla klikněte na odkaz "*Editovat*", který se editorům zobrazuje u podrobných informací o hesle, pod heslem. Zobrazí se formulář pro úpravu údajů.



Obrázek P3.3: Editační formulář pro dříve vyplněné heslo

U všech položek hesla, které se mohou v hesle vyskytnout vícenásobně, je možno přidat novou položku tlačítkem “**+**”. Libovolnou z položek můžete odebrat tlačítkem “**x**”. Všechna pole formuláře jsou doplněna nápovědou, která uživatele informuje, jaké informace může do pole zadávat. Formulář také obsahuje základní kontrolu správnosti vyplněných dat. Pokud je některé z polí nesprávně vyplněno, nedovolí formulář celé heslo uložit a chybně vyplněné pole se zvýrazní.

V části “*Heslo*” se zobrazuje needitovatelný jednoznačný identifikátor hesla. Také zde lze nastavit statut hesla (podrobněji viz Přílohu 2, filtrování podle statutu) výběrem z možností: *terminologický, používaný, používaný nezpracovaný, odmítnutý*. Jako výchozí volba pro nová hesla se nabízí “*používaný*”.

V části “*Termíny*” se zadává samotný termín a jeho překlady do různých jazyků. Pro každý termín či překlad uživatel zadá:

- *pořadové číslo termínu* - použije se pro seřazení v případě více variant termínu v jednom jazyce,
- *jazyk termínu* - je možno přímo vyplnit dvoupísmennou zkratku jazyka (*cs* pro češtinu, *en* pro angličtinu, *fr* pro francouzštinu, *de* pro němčinu, *sk* pro slovenštinu a *ru* pro ruštinu) nebo vybrat ze seznamu nejčastějších jazyků,
- *termín.*

Pokud zadáváte překlad termínu, můžete využít automatické návrhy překladu (na základě textových korpusů, viz sekci 2.4.4 metodiky). Po vyplnění českého termínu přidejte tlačítkem “**+**” nové pole, zvolte požadovaný jazyk a klikněte na tlačítko “*Návrh*”. Pokud jsou k dispozici dostatečná data pro výpočet, zobrazí se 10 nejpravděpodobnějších překladových ekvivalentů pro zvolený jazyk. Po výběru se návrh přenese do pole pro termín a lze ho upravit.

V části “*Definice*” se vyplňují vysvětlení termínu. Je možno zadat více definic, pro každou definici je potřeba vyplnit:

- *pořadové číslo definice* (určuje pořadí a preferenci definic) a
- *text definice.*

V části “*Obory*” lze zadat oborové domény, do kterých daný termín spadá.

V části “*Reference*” lze zadat odkazy na normy či odbornou literaturu, které se vztahují k danému termínu. Odkazy by měly být dostatečně jednoznačné pro identifikaci zdroje či publikace.

V části “*Nadřazené pojmy*” se zadávají hyperonyma (nadřazené termíny) daného termínu. Nadřazený pojem již musí být uložen v databázi. Do textového pole začněte psát termín a podobně jako při vyhledávání termínu se průběžně při psaní nabízí možnosti odpovídající zadanému textu. Kliknutím na některou z možností termín vyberete a do pole před termínem se automaticky doplní jednoznačný identifikátor hesla. Za termínem se nachází seznam automaticky vytvořených návrhů, ze kterých je možné vybrat nadřazený pojem.

V částech “*Také*” a “*Viz*” je možno zadat příbuzné termíny či termíny, kde mohou uživatelé dohledat další informace. Pro každý termín se zadává:

- *dvoupísmenný kód jazyka* (lze přímo zadat nebo vybrat z nabízených možností) a
- *vlastní termín* (může jít o termín obsažený v tezauru nebo libovolný text mimo obsah tezauru, v případě, že jde o termín z tezauru, se při zobrazení automaticky identifikuje a zobrazí jako odkazs).

Po provedení úprav v hesle zapište v části “*Komentář k úpravě*” vysvětlující poznámku k úpravám, které jste provedli. Poznámka bude viditelná dalším editorům v historii úprav hesla, spolu s časem úpravy a jménem editora.

Provedené úpravy se uloží do databáze po kliknutí na tlačítko “*Uložit*”.

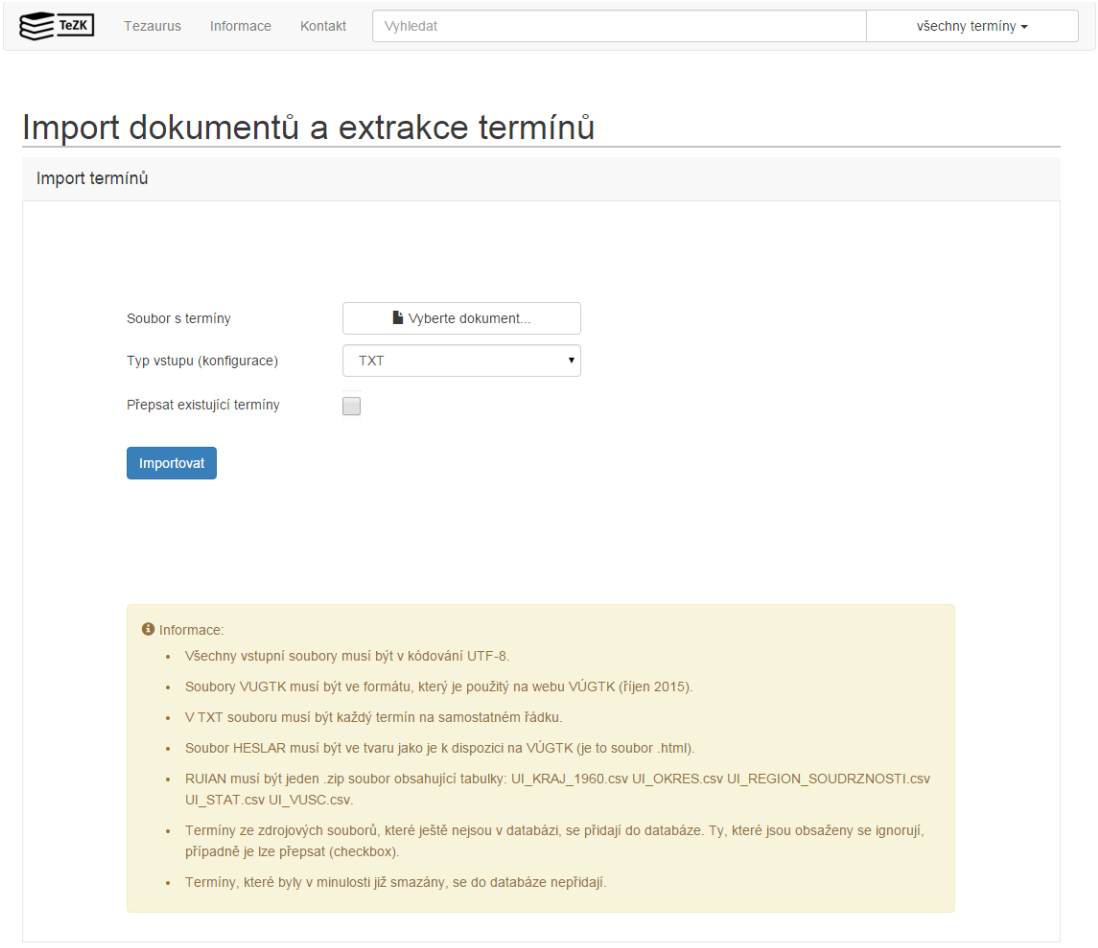
Smazání hesla z Tezauru
Vyhledejte heslo, které chcete odstranit. Kliknutím na odkaz “*Editovat*” zobrazte formulář pro úpravy. Ve formuláři klikněte na tlačítko “*Smazat*”. Po potvrzení bude heslo odstraněno z Tezauru a přesunuto do oddělené databáze. Heslo se již nezobrazuje v Tezauru ani při vyhledávání, ale v případě potřeby je možno heslo obnovit a vrátit zpět.

Export stromové struktury Tezauru
Po kliknutí na odkaz “*Export stromu*” v horní liště se uživateli zobrazí textová reprezentace stromové struktury Tezauru. Tuto reprezentaci je možno uložit a použít pro další práci. Informace v této reprezentaci nelze zpětně importovat do tezauru. V případě, že uživatel chce upravit hyperonymické vztahy hesel, je potřeba použít přímo editaci hesla nebo export/import údajů ve formátu CSV.

Export dat Tezauru ve formátu CSV
Podrobné informace o heslech Tezauru lze exportovat ve formátu CSV, s nímž je možné pracovat např. v tabulkovém procesoru. Po úpravě lze hesla importovat zpět do Tezauru (viz dále Import termínů do Tezauru).
Pro export vyhledejte v Tezauru požadované heslo. V administrační liště na odkaz “Export” a vyberte export “Aktuálního podstromu”. Poté se nabídne k uložení soubor ve formátu CSV, který obsahuje podrobné informace o zvoleném hesle a všech heslech jemu podřízených v hyperonymické struktuře.
Pro každé heslo se uloží informace ve sloupcích v následujícím pořadí: ID, status, termín, jazyk termínu, definice, obor, reference, hyperonyma, odkaz viz, jazyk odkazu, odkaz též, jazyk odkazu. Nové heslo vždy začíná na novém řádku s ID. Pokud je v řádku prázdná buňka pro ID, pokračují informace o daném hesle. Jednotlivé položky (termíny, definice, hyperonyma, odkazy, reference, obory) je potřeba psát vždy do samostatné buňky na novém řádku.

Import termínů do Tezauru
Termíny lze importovat i z existujících zdrojů. Na stránce importu termínů (odkaz vpravo nahoře, viz obrázky P3.4 a P3.5) lze nahrát soubor a zároveň vybrat typ importovaných dat. Systém obsahuje několik konfigurací pro zdroje použité v Tezauru. Lze importovat seznam termínů uložený v jednoduchém textovém souboru, jeden termín na řádek. Můžete rovněž importovat CSV soubory (které lze připravit v libovolném tabulkovém procesoru) s předdefinovanou strukturou. Formulář obsahuje stručnou nápovědu. Jakmile se soubor přenese na server, je zpracován pomocí zvolené importní konfigurace a termíny ze souboru jsou automaticky přidány do databáze. Aplikace ignoruje všechny termíny v databázi již obsažené a dříve smazané. Je možné zaškrtnout volbu přepsání existujících termínů (zapnutá implicitně pouze pro import z CSV souborů).

Obrázek P3.4: Odkazy pro import v rozhraní



Obrázek P3.5: Formulář pro nahrání termínů

Přidání nové konfigurace

Každý typ vstupu má odpovídající konfiguraci (modul v jazyce Python) ve složce `/var/www/import/confs` pojmenované `xxx_conf.py`. Novou konfiguraci lze přidat nahráním nového modulu do této složky. K tomu lze použít například nástroj scp (přístup k nahrávání konfigurací má primárně pouze správce systému).

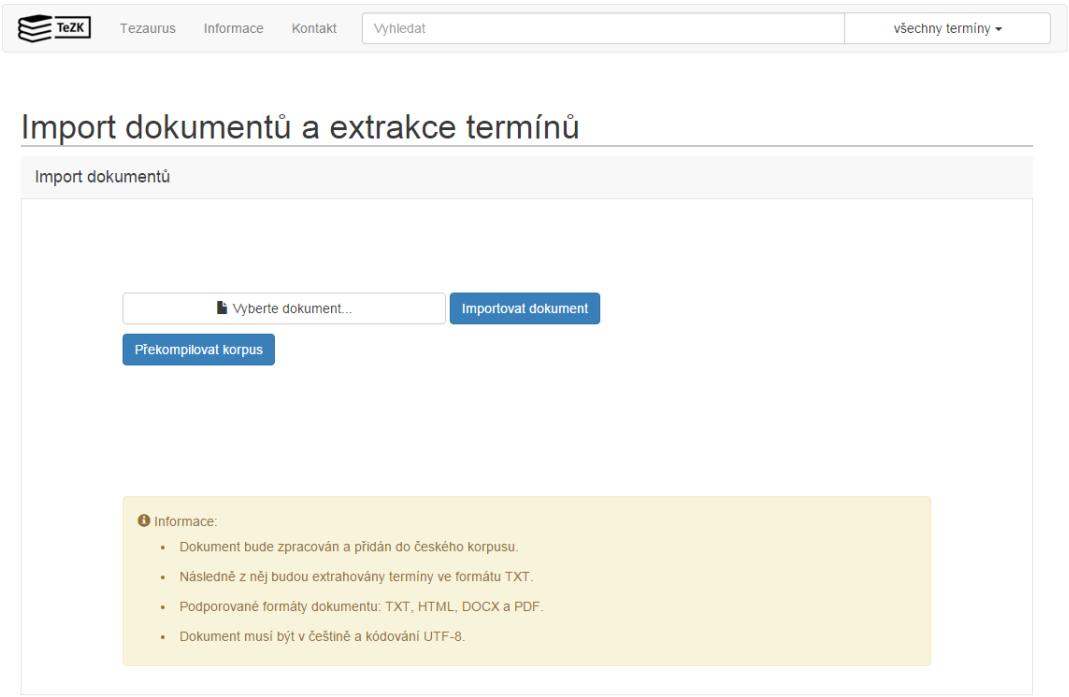
```
scp nova_konfigurace_conf.py [jméno serveru]:/var/www/import/confs
```

Struktura CSV konfigurace je popsaná v sekci 2.2.6.1. V rozhraní (formuláři) se automaticky načte tato přidaná konfigurace a zobrazí se popis konfigurace podle klíče `title` definovaného v modulu.

Import dokumentů

V druhém importním formuláři, viz obrázek P3.6, lze nahrát dokumenty obsahující texty ze souvisejících domén (geografie, geodézie, katastr nemovitostí,…). Import podporuje několik formátů (.txt, .html, .docx a .pdf). Jakmile je soubor přenesen na server, extrahují se z něj doménové termíny a ty jsou automaticky přidány do databáze Tezauru (pokud v ní ještě nejsou) do kategorie “automaticky navržené termíny”.

Při extrakci termínů se porovnávají statistické charakteristiky slovních spojení v obecném a ve specializovaném korpusu. Jako kandidátské termíny jsou navržena slovní spojení, která jsou (statisticky) charakteristická pro nahrávaný text, nikoli obecné fráze. Systém vybere nanejvýš 50 termínů s nejvyšším poměrem relativních četností výskytů v nahrávaném dokumentu a v obecném korpusu. Více o metodě extrakce termínů v sekci 2.4.2 Automatická extrakce a návrh nových termínů.



Obrázek P3.6: Formulář pro nahrání a následnou extrakci termínů z dokumentů ve formátu txt, html, docx a pdf

Příloha č. 4 Návod na používání aplikačního rozhraní webové služby

Aplikační rozhraní aplikace Tezaurus umožňuje využití aplikace Tezaurus aplikacemi třetích stran. Ty mohou pomocí uvedeného rozhraní volat funkce práce z Tezaurem (vyhledávání, zobrazení hesel, práce s heslem).

Pro aplikační rozhraní platí stejné nastavení přístupu jako k uživatelskému rozhraní. Bez autentizace je možno použít funkce rozhraní pro vyhledávání a zobrazení hesel. Pro použití aktivních operací (úprava hesel, odstranění hesel) je potřeba autentizace pomocí uživatelského účtu s editorským oprávněním.

U všech metod je možné přepínat mezi zobrazením návratové hodnoty ve formátech JSON a WSDL parametrem `format` (hodnota `json` nebo `wsdl`). Aktuální adresa rozhraní `https://[jméno serveru]/tecu`.
V případě chybného volání metody vrátí API chybové hlášení:

- pro formát JSON objekt s parametry `error=true` a `msg` s popisem chyby,
- pro formát WSDL se vrací XML dokument s elementy `result=error` a `message` s popisem chyby.

Aplikační rozhraní lze volat přímo prostřednictvím požadavku HTTP (HTTP request). Příklad volání:
`http://[jméno serveru]/tecu?action=search&search=vodopis&format=json`
Návratová hodnota:

```
[
 {
  "highlight":"hydrografická mapa",
  "head":"hydrografická mapa","id":"4906"
 },
 {
  "highlight":"<span class=\"search-highlight\">vodopis</span>",
  "head":"vodopis","id":"9710"
 }
]
```

Jiný příklad volání:
`http://[jméno serveru]/tecu?action=getdoc&id=9710&format=json`
Návratová hodnota (výpis je zkrácený):

```
{
   "entry": {
      "hyper": {
            "@term": "druhy mapy",
            "@id": "11899"
      },
      "@id": "4906",
      "terms": {
            "term": [
                  {
                  "@lang": "cz",
                  "$": "hydrografická mapa"
                  },
                  {
                  "@lang": "ru",
```

```
                "$": "гидрографическая карта"
                }
            ]
      },
      "defs": {
         "def": {
            "$": "mapa zobrazující jako hlavní téma tekoucí a stojaté vody"
         }
      }
   }
}
```

Níže následuje popis jednotlivých funkcí aplikačního rozhraní včetně parametrů a výsledků.

Vyhledávání

Parametry	

`action`
`search`
`format`
	search
	hledaný výraz
	json/wsdl

Výsledek	
JSON	pole objektů s nalezenými výsledky (každý objekt obsahuje ID hesla, termín a termín se zvýrazněním hledané části)
WSDL	XML dokument, ve kterém objekty `result-entry` obsahují nalezené výsledky (každý objekt obsahuje ID hesla, termín a termín se zvýrazněním hledané části)

Vyznačení termínů v textu

Parametry	

`action`
`text`
`format`
	highlight
	text k vyznačení
	json/wsdl

Výsledek	
JSON	text s označenými částmi, které odpovídají termínů v Tezauru, ke každému termínu je doplněn identifikátor pro odkaz do Tezauru
WSDL	XML dokument, který obsahuje původní text s označenými částmi, které odpovídají termínů v Tezauru (jako objekt term), ke každému termínu je doplněn identifikátor pro odkaz do Tezauru

Zobrazení hesla

Parametry	
`action` `id`	getdoc ID záznamu

`format`	json/wsdl
Výsledek	
JSON	objektová reprezentace XML podoby hesla
WSDL	XML dokument hesla

Vyhledání hlavní úrovně tezauru

Parametry	
`action` `format`	get_top json/wsdl
Výsledek	
JSON	pole objektů s termíny (každý objekt obsahuje ID hesla, termín a počet podřízených termínů)
WSDL	XML dokument, ve kterém objekty `result-entry` obsahují nalezené výsledky (každý objekt obsahuje ID hesla, termín a počet podřízených termínů)

Vyhledání podřízených hesel

Parametry	

`action`
`id`
`format`
	get_subtree
	ID hesla, pro které se hledají podřízené termíny
	json/wsdl

Výsledek	
JSON	pole objektů s termíny (každý objekt obsahuje ID hesla, termín a počet podřízených termínů)
WSDL	XML dokument, ve kterém objekty `result-entry` obsahují nalezené výsledky (každý objekt obsahuje ID hesla, termín a počet podřízených termínů)

Vyhledání cesty ve stromu

Parametry	

`action`
`id`
`format`
	get_path
	ID hesla, pro které se hledá cesta
	json/wsdl

Výsledek	
JSON	pole, které obsahuje ID hesel postupně od zadaného hesla ke kořenu stromu
WSDL	XML dokument, ve kterém objekty `result-entry` obsahují ID hesel postupně od zadaného hesla ke kořenu stromu

Uložení upraveného hesla

Parametry	

`action`
`id`
`data`
`format`
	save
	ID hesla
	objektová (JSON) reprezentace hesla
	json/wsdl

Výsledek	
JSON	při úspěšném uložení objekt s parametry `ok=true`, `msg=saved` a `id` s ID hesla
WSDL	XML dokument s elementy `result=saved` a `id` s ID hesla

Vytvoření nového hesla

Parametry	

`action`
`data`
`format`
	save
	objektová (JSON) reprezentace hesla
	json/wsdl

Výsledek	
JSON	při úspěšném uložení objekt s parametry `ok=true`, `msg=saved` a `id` s přiděleným ID hesla
WSDL	XML dokument s elementy `result=saved` a `id` s přiděleným ID hesla

Odstranění hesla

Parametry	

`action`
`id`
`format`
	delete
	ID hesla
	json/wsdl

Výsledek	

JSON	Při úspěšném odstranění objekt s parametry
	`ok  true`
	`msg  deleted`
	`id  ID odstraněného hesla`

WSDL	XML dokument s elementy
	`result  deleted`
	`id  ID odstraněného hesla`