GEOKÓDOVÁNÍ OBJEKTŮ PODLE ADRESY

Jiří HORÁK[1], Jan CAHA[2],Tomáš INSPEKTOR[3], Pavel KUKULIAČ[4]

[1,2,3,4] Institut geoinformatiky, Hornicko-geologická fakulta, Vysoká škola báňská – Technická univerzita Ostrava, 17. listopadu 15, 708 33 Ostrava, Česká republika
[1]jiri.horak@vsb.cz; [2]jan.caha1@vsb.cz; [3]tomas.inspektor@vsb.cz; [4]pavel.kukuliac@vsb.cz

Abstrakt

Lokalizace institucí, podniků, škol a dalších organizací podle jejich adresy stále naráží na problém, že se při vytváření jejich seznamů nepoužívají vhodné standardy (dnes RÚIAN), ale používají se volně psané texty, které trpí řadou nedostatků a způsobují problémy při geokódování. U těchto seznamů je zpravidla k dispozici více adresních údajů, vč. administrativních jednotek, což nabízí možnosti zpřesněné lokalizace ve srovnání s použitím poštovní adresy, aplikované např. v mapových vyhledávačích. Na příkladu několika vzorků dat z Firemního monitoru (Albertina data), registru základních a středních škol a registru zdravotnických zařízení jsou porovnány výsledky geokódování s pomocí vyvinuté procedury využívající sady podpůrných administrativních údajů s výsledky textově flexibilního geokódování v mapových vyhledávačích.

Abstract

Adressed Based Geocoding of Objects: Location of institutions, companies, schools and other organisations according to their address still faces the problem that lists or registers of such objects are not created according to standards (especially code lists of address parts) and they contain free texts. Such texts have a number of shortcomings, and thus cause difficulty in geocoding. Usually these registers contain more than one piece of address data, including identification of territorial units, which enables better geocoding results than just usage of postal addresses. Using several data samples (companies drawn from the Albertina Company Monitor, register of primary and secondary schools, register of medical facilities) results of different types of geocoding are compared, mainly text based geocoding in map search engines and database procedures enabling administrative units supports.

Klíčová slova: geokódování, adresy, GIS, Google, Mapy.cz

Keywords: geocoding, address matching, GIS, Google, Mapy.cz

1. ÚVOD

Pojmy jako georefencování (georeferencing), geokódování (geocoding) či geolokace (geolocating) se často používají bez specifického rozlišení. I Longley et al. (2005) uvádí možnost jejich záměny. Pokud se přesto pokusíme o bližší vymezení, je zpravidla geokódováním chápáno nepřímé georeferencování, tj. určování polohy objektů a jevů prostřednictvím geokódů, a ne pomocí souřadnicových systémů.

Longley et al. (2005) upozorňují na požadavek stability systému geokódování v čase – pokud dochází k častým změnám pojmenování objektů či změně jejich polohy, je pochopitelně takový systém velmi náchylný k chybám.

Význam správného geokódování je důležitý jak pro oblast přírodních věd (viz např. Beaman, Conn, 2003), tak i společenských věd (např. Edwards et al. 2014).

Geokódování neboli adresování (address matching, geocoding) umožňuje připojit souřadnice k záznamům lokalizovaným adresou pomocí připojení k souboru, který obsahuje adresy i souřadnice. Geokódování využívá strukturované informace, kdy je adresa popsána v několika polích (Horák, 2013).

V ideálním případu lokalizujeme objekty, jejichž polohu lze dobře popsat pomocí adresy (adresního bodu). Pokud by byl k dispozici identifikátor adresního bodu, byla by úloha velmi jednoduchá. Zpravidla jsou však položky adresy zapsány volným textem nebo v případě seznamů poskytovaných různými institucemi pak

v podobě smíšené – část je zapsána pomocí identifikátorů (např. kód obce, kód okresu) a zbytek volným textem.

Cílem geokódování je lapidárně řečeno umístit bodové objekty do mapy. Těmito bodovými objekty mohou být například adresy, místa nehod, místa deliktů, adresy zákazníků, čísla parcel. Obzvláště komplikovanou situací je geokódování zastávek a dalších objektů systému veřejné dopravy (Zajíčková et al., 2014).

Pokud požadovaná poloha neodpovídá poštovní adrese, může se v textu popisu polohy objevit i referencování vůči názvu místa či objektu (např. před budovou radnice), někdy s využitím metrických informací (např. 180 m severně od adresy XY). V případě pokročilé analýzy volného textu se již hovoří o geoparsingu (Horák et al., 2011, Košková, Kafka 2009).

Za geokódování se dříve označovala pouze situace, kdy není k dispozici bodová vrstva, ale pouze liniová vrstva ulic s rozsahy domovních čísel. V ČR to znamená použití orientačních domovních čísel v adrese, které vytváří vhodný systém lokalizace v rámci ulice. Tak ji popisuje i např. Longley et al. (2005). Specializované funkce (k dispozici v desktopových programech pro GIS) rozlišují pravé a levé strany ulic, pracují s rozsahy domovních čísel v jednotlivých částech ulice a dovolují připojit i nepřesně nebo neúplně zadané adresy (překlepy v adresách, používání zkratek - ul., nám., tř. apod.). Poloha se pak určuje interpolací mezi krajními adresami (Horák, 2013).

Typické zpracování potom začíná porovnáním uliční adresy. Pokud se nemůže pro jméno ulice najít vhodný protějšek, použije pro cílovou adresu soubor zkratek. V dalším kroku se program zabývá čísly adres, pro které úspěšně porovnal jména. Z cílové adresy vezme číslo adresy a porovnává je s rozmezími každého úseku ve zdrojové tabulce, přičemž rozlišuje mezi levou a pravou stranou ulice. Výsledek geokódování se zapisuje pomocí kódu, který popisuje způsob nalezení shody a přesnost výsledku.

Často jsou při porovnání hledané a referenčních adres vybráni potenciální kandidáti a je jim přiřazeno skóre, které podává informaci o tom, nakolik si tyto dvě adresy odpovídají. Následně jsou uloženi podle dosaženého skóre do „souboru všech případů pro porovnání“. Uživatelé ArcGIS znají parametry jako je „Hlásková citlivost“, „Minimální skóre pro určení kandidátů“ či „Minimální porovnávací skóre pro výběr správného protějšku“, které se v tomto procesu používají.

2. DATA Z FIREMNÍHO MONITORU ALBERTINA

Albertina – Firemní Monitor (AFM) představuje komplex vzájemně provázaných a doplňujících se databází zaměřených na informace o firmách. Společnost byla založena v roce 1991 pod názvem Albertina Information Services, s.r.o., v roce 2011 se stala členem skandinávské sítě specializující se na monitoring a poskytování informací o podnicích a změnila svůj název na Soliditet, s.r.o.. Firma udává, že aktualizace dat se provádí na základě získávání informací zveřejňovaných v obchodním nebo živnostenském rejstříku (Kukuliač 2014).

Databáze rozlišuje Obchodní adresu a Registrovanou adresu. Registrovaná adresa je uváděna tak, jak je zapsána v příslušném registru. Obchodní adresa je upravena pro potřeby doručování, tj. respektuje požadavky pošty. Odpovídá registrované adrese, pokud není individuálně nebo z doplňkových zdrojů zjištěno, že firma je registrována na jiné adrese, než na které fyzicky sídlí. V tom případě je obchodní adresa upravena a registrovaná ponechána.

Při využití dat z AFM je nutné si uvědomit, že podnikatelské organizace nejsou v České republice evidovány podle místa podnikatelských aktivit, ale podle sídla organizace (podobně jako RES). V případě prostorových analýz ekonomických aktivit zpravidla potřebujeme provozovny. Použití sídel vede k tomu, že v území chybí významné podniky, jejichž sídlo je na jiném území (např. v Praze). Od roku 2003 databáze Albertina – firemní monitor poskytuje informaci o provozovnách většiny evidovaných ekonomických subjektů, avšak zahrnuje pouze informace týkající se lokalizace pobočky v daném území. U těchto provozoven již nejsou k dispozici ekonomické údaje ani údaje o počtech zaměstnanců, jako je tomu v případě sídel podniků.

Tab. 1 Popis struktury dat poskytovaných z Firemního monitoru (výběr polí s vazbou na adresu)

Název pole	Datový typ	Popis
ICO	Text	IČ firmy
DIC	Text	DIČ firmy
FIRMA40A	Text	Název firmy (1. řádek)
FIRMA40B	Text	Název firmy (2. řádek)
OBEC	Text	Obec (obchodní adresa)
ULICE	Text	Ulice (obchodní adresa)
PSC	Text	PSČ (obchodní adresa)
ULICE2	Text	Ulice 2 (pro pošt. účely)
POSTA	Text	Pošta
OKRES	Text	Okres
KRAJ	Text	Kraj
VELOBEC	Text	Velikost obce
ICZUJ	Text	Identifikační číslo ZUJ
POZUJ	Text	Obec s pověřeným obecním úřadem
RPZUJ	Text	Obec s rozšířenou pravomocí
REGFIRMA	Text	Registrovaný název firmy
REGOBEC	Text	Registrovaná obec
REGULICE	Text	Registrovaná ulice
REGPSC	Text	Registrované PSČ

Z hlediska lokalizace je významný atribut ICZUJ, který identifikuje základní územní jednotku (obec). Jeho naplněnost je vysoká (téměř 100% pro rok 2014), avšak asi 5% záznamů je špatně vyplněno, což způsobuje problémy. Proto se v praxi ukazuje jako vhodnější použít název okresu nebo PSČ pro vymezení lokalizace.

Dalším problémem je aktuálnost dat. Jsou evidovány i firmy, které již nepodnikají nebo podnikají pouze v jedné z mnoha pro sebe registrovaných činností (Horák et al., 2015).

2.1 Lokalizace podniků z AFM

K dispozici byly data pro období let 1999, 2004, 2009 a 2014. Celkově databáze obsahovala 1 776 285 záznamů pro Moravskoslezský a Jihomoravský kraj. Pro rok 1999 bylo evidováno 325 594 podniků, pro rok 2004 - 428 830 podniků, pro rok 2009 - 487 247 podniků a pro rok 2014 - 534 614 podniků. Lokalizace podnikatelských subjektů byla realizována v prostředí Microsoft SQL Server vůči referenční databázi stavebních objektů RSO. Propojením dat z databáze AFM s databází stavebních objektů byla provedena přímá lokalizace podnikatelských objektů. Vzhledem k nepřesnostem bylo nutné provádět harmonizaci dat a vícenásobné spojování s cílem zvýšit věrohodnost a úspěšnost lokalizace. Lokalizováno bylo celkem 1 753 499 podniků z celkového počtu 1 776 285 podniků. To představuje 99% úspěšnost lokalizace. Výsledky jsou obsaženy v disertační práci P.Kukuliače (Kukuliač, 2014), která se zaměřovala na hodnocení distribuce vybraných ekonomických aktivit v území a vývoj nové metody pro hodnocení prostorových vzorů ekonomických aktivit s využitím stochastické simulace.

Adresa podniku je popisována v exportu z databáze AFM dvěma atributy: Obec a Ulice. Atribut Obec obsahuje informaci o názvu obce a části obce, ve které se podniky nachází. Atribut Ulice pak zahrnuje definici názvu ulice a čísla popisného a orientačního (popř. evidenčního). Protože nelze přímo propojit jednotlivá pole adresy s databází RSO, používá se textového spojení do jednoho pole (v řadě variant), které

se následně porovnává se spojeným řetězcem adresních polí RSO. Tím se řeší formálně nesprávný zápis adres (např. přehození č.p. a č.o.) a případné chybějící části adresy.

Samozřejmě nalezení shody vyžaduje také obsahově správný zápis, tj. musí se eliminovat nadbytečné texty (např. název podniku v adrese typu „17.listopadu Fakult.nemocnice“), zkratky, překlepy a jiné problémy (p.o. box apod.). Aby bylo možné lokalizaci provést, bylo potřeba opravit názvy ulic, obcí a jejich částí, tak aby byly v souladu s databází stavebních objektů RSO. Např. adresa „1. československého armádního sboru“ musela nahradit všechny odlišné formy jejího zápisu jako „1.Čs. arm. sboru“, „1.čs.arm.sboru“, „Čs. armádního sboru“ či „Čs.arm.sboru“.

Pro zpřesnění lokalizace by bylo výhodné použít přímo kódu obce, který je v databázi AFM evidován. Problémem je, že ne vždy je tento kód správně uveden. Proto se raději využíval název okresu.

Aby bylo možné data z AFM lokalizovat, bylo provedeno vytvoření celkem 10 různých kombinací zápisu adres ze zdrojových databází (ČSÚ RSO 2002, ČSU RSO 2009, ČSU RSO 2012 a ČSÚ RSO 2014).

3. REGISTR ZDRAVOTNICKÝCH ZAŘÍZENÍ

Evidenci zdravotnických zařízení v příslušné části území vedou územní odbory ÚZIS (Ústav pro zdravotnické informace a statistiku). Registr obsahuje všechna zdravotnická zařízení s jejich základní kategorizací a kontaktními údaji. Adresní atributy jsou uvedeny v tab. 1.

Tab. 2 Struktura datového souboru (výběr adresních atributů)

Název atributu	Datový typ	Popis
ICO	text	identifikační číslo organizace
KRAJ	číslo	Vlastní číselník krajů
OKRES	číslo	Vlastní číselník okresů
ZUJEDN	text	kód obce ICOB (ČSÚ)
ULICE	text	Název ulice
CISDOM	text	Číslo domovní
CISORI	text	Číslo orientační
PSC	text	PSČ

4. DATA Z EVIDENCE ŠKOL

Adresář základních, středních a vyšších odborných škol je volně k dispozici na internetové stránce: http://stistko.uiv.cz/registr/vybskolrn.asp. Je pravidelně aktualizován a je možné si vyhledat všechny školy v ČR najednou nebo po jednotlivých krajích či okresech. Také lze data filtrovat podle jednotlivých typů škol. Seznam je možné exportovat do tabulky. Adresní atributy jsou v tab. 2.

Tab. 3 Struktura datového souboru (výběr adresních atributů)

Název atributu	Datový typ	Popis
RED_IZO	číslo	Označení identifikačního čísla ředitelství školy
IČO	text	identifikační číslo organizace
Území	text	kód okresu
ORP	číslo	kód obce s rozšířenou působností
Plný název	text	Plný název organizace
Zkrácený název	text	Zkrácený název organizace
Ulice	text	Název ulice (často včetně ČP)
č.p.	číslo	číslo
PSČ	číslo	poštovní směrovací číslo
Místo	text	název města
X	text	identifikátor školy/zařízení

Naplnění adresních atributů je různé – zatímco PSČ je vyplněno u všech záznamů, číslo popisné pouze u 11,9 %.

5. PROCES ZLEPŠENÉHO DATABÁZOVÉHO GEOKÓDOVÁNÍ PODLE ADRESY

Z evidence objektů či událostí se separuje primární klíč a adresní atributy, které se importují do prázdné tabulky UDALOST. Očekává se naplnění atributů OKRES, OBEC, CAST, ULICE, CISLOOR, CISLOPOP, UPRESNENI (doplňující poznámky k lokalizaci), PSC; případně další známé územní identifikátory jako např. identifikátor ORP, katastrálního území apod. Tyto atributy obsahují originální názvy a nesmí se v průběhu zpracování změnit.

Cílem zpracování je na základě jejich rozboru správně naplnit cílové atributy územní identifikace LAU1, NAZOB, KODOBCE, NAZCOB, KODCOB, NAZULICE, COR, CP. Kromě toho probíhá i částečně automatizované naplňování atributů určených pro evidenci objektů (TYPOBJ, NAZOBJ, PODOBJ). Na jejich základě pak jsou doplněny souřadnice identifikovaného místa.

Celý proces má 3 etapy:

1. Příprava
2. Harmonizace
3. Geokódování

5.1 Příprava

Příprava zahrnuje importování a nastavení aktuálních územních číselníků a generování aktuálních seznamů pro jednotlivá cílová pole (názvy obcí, částí obcí, ulic atd.). Územní číselníky se přebírají z RSO ČSÚ, je ale možné využít RUAIN. Následně se generují těžiště jednotlivých kombinací adresy - např. souřadnice pro kombinaci COB-NAZULICE-CP-COR, souřadnice pro kombinaci COB-NAZULICE-CP, souřadnice pro střed ulice v části obce COB-NAZULICE, souřadnice pro střed ulice v obci KODOBCE-ULICE, souřadnice pro část obce, souřadnice pro obec atd. Protože těžiště může být určeno ze skupiny bodů, které se v dané kombinaci adresních atributů shodují, eviduje se kromě průměrného středu také RMSE a počet adres, ze kterých bylo těžiště vypočteno.

Při procesu harmonizace je možné využít více těchto seznamů těžišť – pokud se nenalezne daná kombinace adresy v aktuálním roce (např. z důvodu přejmenování ulice), prohledají se následně starší seznamy.

Současně se provádí generování jedinečných názvů objektů, které lze použit pro identifikaci (unikátní seznam názvů obcí, seznam jedinečných názvů částí obcí atd.). Jeho smyslem je možnost přiřadit jedinečný kód obce či kód části obce nalezeným jednoznačným a nezaměnitelným názvům.

Další přípravnou operací je generování variantních názvů (zejména zkrácené z původních, např. Ústí n.L.).

5.2 Harmonizace

Cílem je správně naplnit atributy KODOBCE, KODCOB, NAZULICE, COR, CP; nebo identifikátor referenčního objektu.

Referenčním objektem rozumíme jakýkoliv objekt v území, jehož název je jednoznačný alespoň v určitém prostorovém kontextu (Pražský hrad, nádrž Lipno, obchodní centrum Tesco – nezaměnitelné v dané části obce). Je zřejmá jistá analogie ke Geonames, POI, placenames. Upřednostněny jsou objekty, které mají stálou polohu v území (a tedy i souřadnice resp. souřadnicový obdélník), lze ale i využít i mobilní, pokud neexistuje spolehlivější identifikace místa (vražda v Orient Expresu, linka č. 8 v Ostravě apod.).

Identifikace správných územních jednotek pro hledaný objekt se provádí shora dolů, protože to umožňuje omezit vyhledávání v dalším kroku jen na relevantní územní rámec.

Proto se nejdříve určují správné vyšší administrativní jednotky (pokud jsou známy). Klíčovým krokem je určení správné obce (tedy správně identifikovat NAZOB a KODOBCE), následuje určení správné části obce, pak správné ulice a nakonec se vyplní CP a COR (již bez kontroly na existenci v oficiálním seznamu).

V zásadě se proces snaží obsah pole ztotožnit s některou z položek v oficiálním seznamu. Pokud se to nedaří, aplikuje se „očista“ (smažou se znaky, které se v daném seznamu nemají vyskytovat), aplikuje se

segmentace textového pole a hledá se název v některém ze segmentů, používají se opravy (seznam minulých oprav), aplikují se vygenerované variantní názvy (např. seznamy zkratek).

Vyhledává se vždy jen v již shora omezeném území (tj. název části obce jen v rámci již známé obce, název ulice v rámci ČOB atd.).

Proces končí co nejvyšším vyplněním všech cílových polí.

Tento postup není samozřejmě vždy ideální – stačí např. přehodit název obce a název části obce, nebo určení části obce v jiné obci, což se zatím řeší přes opravy (tab. známých oprav).

Speciální postupy jsou potřebné zejména v případě, pokud je v adrese zapsán objekt, často křižovatka nebo liniový objekt (např. silnice, linka MHD).

5.3 Geokódování

Jde o relativně jednoduchý proces, kdy se v jasně definovaném pořadí zkouší ztotožnit kombinace harmonizovaného zápisu adresy hledaného objektu se známými kombinacemi jednotlivých adresních polí s těžišti (jejich příprava viz kap. 5.1).

Pořadí ztotožňování vypadá přibližně následovně: GPS souřadnice, Objekt, přesná shoda v adrese (část obce + název ulice + CP + COR), horší shoda v adrese (např. obec + název ulice + CP) až po aproximativní určení ulice, nebo pouze části obce či jen obce.

Tím se zajistí, že hledanému objektu (resp. adrese) budou přiděleny souřadnice s nejvyšší úrovní přesnosti, které v daném případě bylo možné dosáhnout. Pokud byla adresa špatně rozpoznána, poskytuje systém alespoň přibližnou polohu.

6. GEOKÓDOVÁNÍ POMOCÍ API MAPY.CZ

Mapy.cz společnosti Seznam.cz poskytují přístup ke geokódování pomocí API, které je k dispozici na adrese: http:/api4.mapy.cz/geocode.

Vhodně formátovaný řetězec adresy je posílán na tuto adresu, služba vrací XML, jehož součástí jsou i zeměpisná šířka a délka.

API rozhraní Mapy.cz vrací nejen souřadnice zadané adresy, ale také seznam firem, které se na uvedené adrese nacházejí (pravděpodobně placená reklama).

Cícha (2013) popisuje, že geokódování s využitím API mapy.cz na rozdíl od Google Maps Geocoding API nemá téměř žádnou schopnost aproximace a proto pokud je v zadání chyba (stačí jedno špatné písmeno), proces geokódování neproběhne. Proto vyžaduje čistotu vstupních dat. V současnosti díky pokroku vývoje uvedeného API se výsledky nejeví tak náchylné na čistotu dat. Cícha implementoval využití API pomocí skriptu v R. Této funkce jsme pro geokódování vzorků dat využili.

V případě rozpoznání pouze ulice je bod lokalizován do její poloviny, či do jinak stanoveného referenčního bodu (seznam.cz, 2013, in Cícha, 2013).

Omezení služby objemem není známo, proces však nesmí být vyhodnocen jak DoS útok a proto se požaduje, aby požadavky byly omezeny na max. 1 za sekundu (seznam.cz, 2012, in Cícha, 2013). Využití služby je bezplatné, a to i pro komerční účely.

7. GEOKÓDOVÁNÍ POMOCÍ GOOGLE MAPS GEOCODING API

Služba je dostupná na adrese https://maps.googleapis.com/maps/api/geocode/output?parameters, kde output je specifikace formátu (json nebo XML).

Z parametrů je vyžadována adresa nebo její komponenty. Pro zápis adresy se má používat způsob dle národní poštovní služby. Dodatečné prvky v adrese, jako obchodní názvy, jednotky, označení bytů či pater se má z adresy předem odstranit (Google, 2014).

Služba vrací zeměpisné souřadnice ve WGS84, ale také řadu dalších informací, jako jsou identifikované segmenty adresy (ulice, lokalita (politická), administrativní jednotka různých úrovní, země (politická), ZIP kód, správně (= americky) formátovaná adresa, geometrie (location – LAT a LON, typ lokalizace, souřadnice SV a JZ okraje obdélníku které ohraničují objekt, typ adresy (např. uliční adresa) a status. Pokud není status OK, vrací kromě kódu vysvětlení i zprávu s popisem důvodu.

Vyhledávání se dá omezit regionem, boxem, a různými komponentami adresy, jenže pokud nenajde odpovídající protějšek, vrací souřadnice i mimo dané omezení (Google, 2014).

Obecně služba vrací jen 1 výskyt, i když je adresa nejednoznačná.

Pro mobilní zařízení či uživatelské vstupy na web doporučuje Google geokódování na straně klienta.

Využití Google Maps Geocoding API implementoval v prostředí R ve své bakalářské práci Cícha (Cícha, 2013). Popisuje, že nejjednodušší způsob geokódování v R nabízí balík ggmap skrze funkci geocode(), kde stačí zadat jako argument adresu. Volitelným argumentem se nastavuje požadovaný výstup, zda se požadují pouze souřadnice x a y, nebo také adresa či další údaje. Jeho funkce jsme využili pro geokódování adres.

Podrobný popis je k dispozici na adrese https://developers.google.com/maps/documentation/geocoding/.

7.1 Omezení použití:

Volně dostupné API omezuje použití na 2500 žádostí za 24 hodin a nanejvýš 5 žádostí za sekundu.

V případě registrovaných uživatelů Google Maps API for Work je to 100000 žádostí za 24 hodin a nanejvýš 10 žádostí za sekundu.

V případě vyhodnocení užití jako porušující pravidla či narušení bude vaše služba dočasně přerušována a v případě pokračování bude váš přístup blokován.

Co je ale zásadnější, je omezení, které říká, že Geocoding API je možné používat jen ve spojení s Google map a že geokódování bez zobrazení výsledků na mapě je zakázáno.

8. ZPRACOVÁNÍ A VÝSLEDKY

Pro potřeby lokalizace různých cílů dojíždění v území byly lokalizovány adresy jednotlivých firem a institucí s využitím různých zdrojů – zde uvádíme Albertina Firemní monitor, databázi zdravotnických zařízení a databázi školských zařízení v ČR. Pro testování výsledku geokódování byly vybrány malé vzorky dat z jednotlivých databází.

Ke geokódování bylo využito – funkce pro volání Google Maps Geocoding API, funkce pro geokódování v rámci API mapy.cz (obě funkce byly implementovány v prostředí R studentem Cíchou v rámci jeho BP), výsledek databázového geokódování s pomocí RSO (řetězcová varianta 1 pro zpracování AFM dle kapitoly 2.1 a pokročilá varianta 2 pro zpracování zdravotnických zařízení a školských zařízení dle kapitoly 5).

Každá z funkcí vrátila v případě úspěchu při geokódování souřadnice ztotožněné adresy. Ty pak byly pro každý vzorek dat samostatně porovnány. Pro každou skupinu bodů geokódovaných pro tutéž adresu byly vypočteny průměrný střed a směrodatná vzdálenost (standard distance) z důvodu snazší detekce problémových lokalizací. Směrodatná vzdálenost přes 1 km byla použita jako limit pro detekování problémové adresy, kde se výsledky geokódování jednotlivými nástroji výrazně liší. U takové skupiny byly kontrolovány jednotlivé adresy a nalezená špatná umístění označena jako hrubá chyba. Některé příklady jsou v tab. 4.

Tab. 4 Příklady adres a jejích chyb

Vzorek	ULICE	CISDOM	PSC	NAZOB	Správné X	Správné Y	Chybné X	Chybné Y	Výsledek geokódován
Zdravotnická zařízení	ZS-UNEX a.s.		78391	Uničov			-7729961	-6993905	Pouze G-ul-PSC, jenže špatně
Zdravotnická zařízení	Zdrav. středisko	900	73571	Dětmarovice	-457361	-1096124	-565677	-1195293	Správě db, chybně G-ul-PSC
Zdravotnická zařízení	NHO a.s. ambulance Energetiky		70702	Ostrava Kunčice					Pouze G-ul-PSC
Zdravotnická zařízení		1	78326	Bílá Lhota	-565590	-1106672	-7764152	-6916454	Špatně G-ul-PSC, správně G-ul-obec
Školy	Kmochova 1823		43401	Most			-781250,1	-986398	Pouze G-ul-obec, jenže špatně
Školy	Zámek 1	1	56401	Žamberk	-598928	-1060176	-202382	-649231	Správně db, G-ul-obec, špatně G-ul-PSC
AFM	106		74719	Závada	-477724	-1087894	-496398	-1232600	Správně db, špatně G-ul-obec
AFM	205 - areá		74767	Hrabyně					pouze G-ul-PSC a M-ul-PSC, ale oboje špat.

9. POROVNÁNÍ VÝSLEDKŮ GEOKÓDOVÁNÍ

Souhrnné výsledky různých postupů geokódování jsou uvedeny v tabulce 5, po odstranění hrubých chyb pak v tabulce 6.

Tab. 5 Počet vrácených adres u jednotlivých nástrojů geokódování

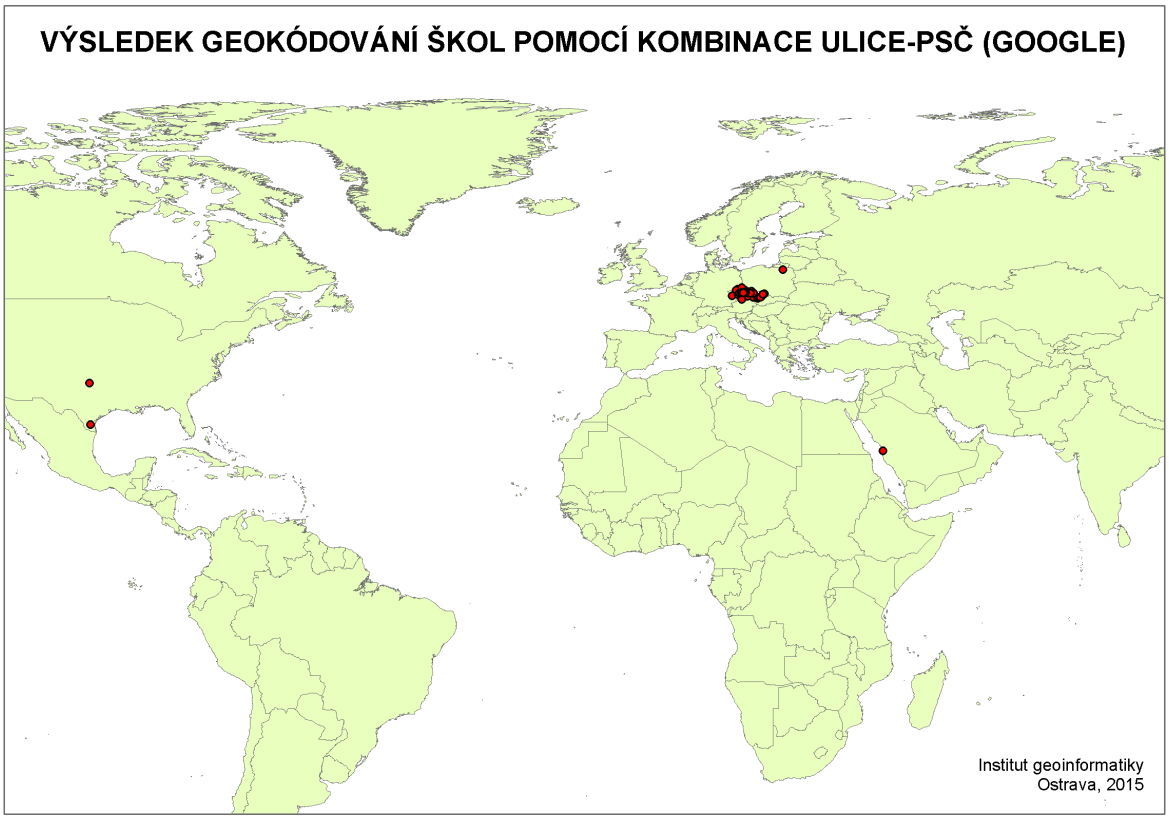
Vzorek	Počet	Databázové1	Databázové2	G-ul-obec	G-ul-PSC	M-ul-obec	M-ul-PSC
AFM	82	74		81	73	75	56
Zdravotnická zařízení	151		147	149	148	131	125
Školy	61		52	61	60	53	47

Tab. 6 Počet vrácených adres u jednotlivých nástrojů geokódování po odstranění hrubých chyb

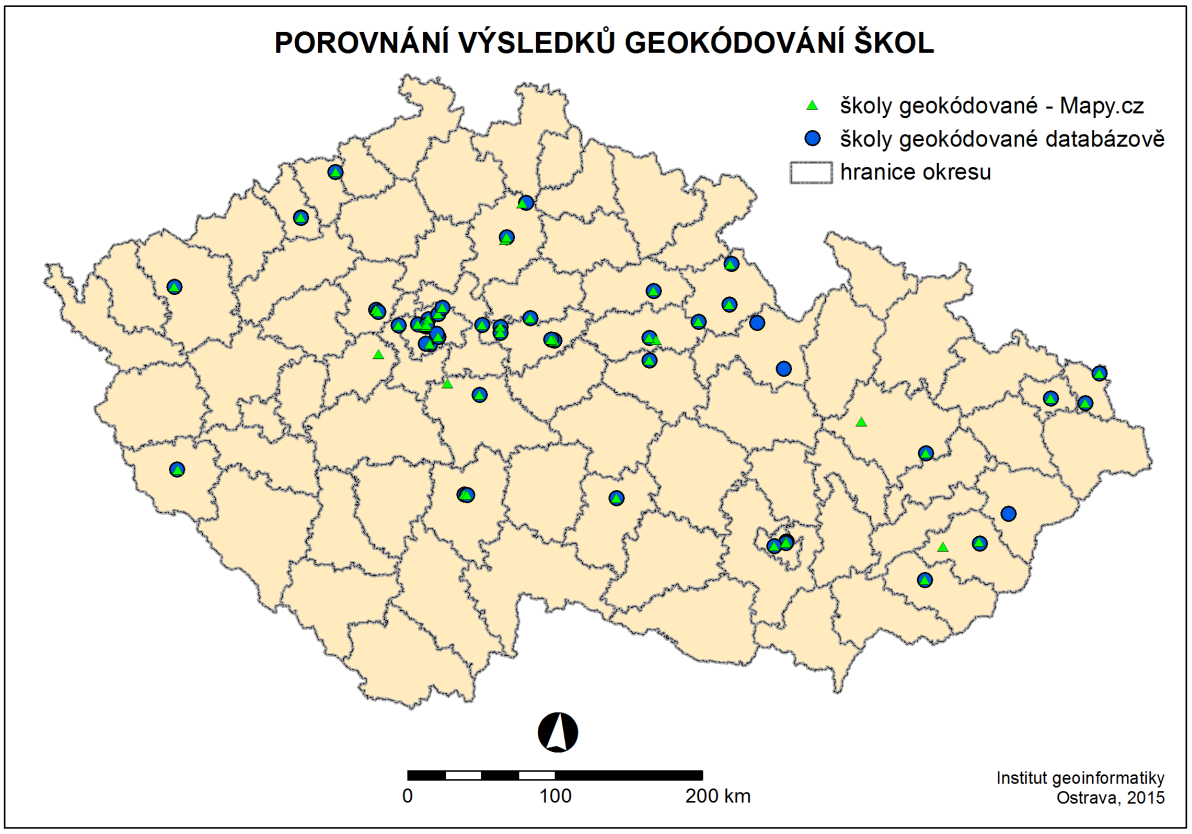
Vzorek	Počet	Databázové1	Databázové2	G-ul-obec	G-ul-PSC	M-ul-obec	M-ul-PSC
AFM	82	74		80	62	75	55
Zdravotnická zařízení	151		145	149	122	131	117
Školy	61		52	60	52	52	41

Z toho vyplývá, že absolutně nejlepší výsledky má Google, varianta zadávání ulice-obec. Celkově jako nejhorší se jeví zadávání kombinace ulice-PSČ jak pro Google, tak pro Mapy.cz. Poměrně často dochází k tomu, že je nabízena poloha mimo ČR. Na obr. 1 jsou 3 případy českých škol, z nichž 2 byly lokalizovány v USA a jedna v Saudské Arábii, přitom důvod není zřejmý.

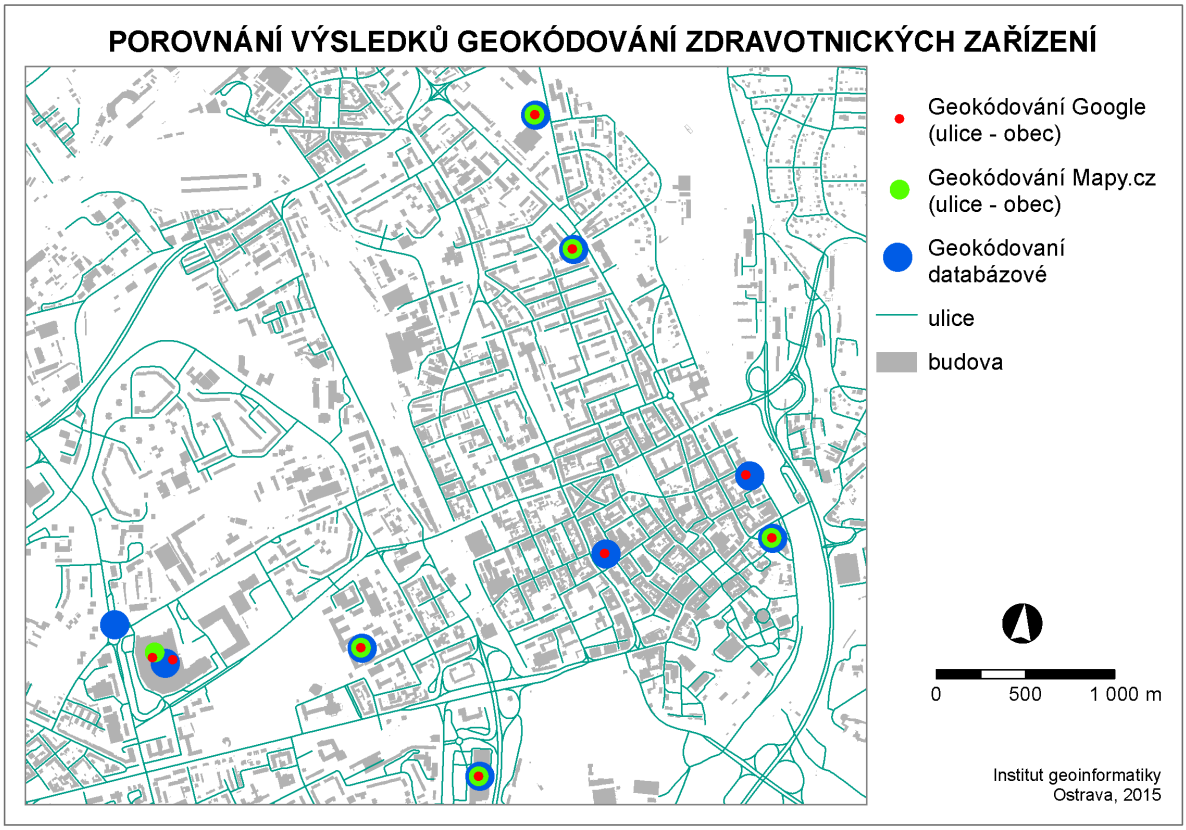
Ani databázový proces zpracování není pochopitelně zcela bez chyb. Nicméně příčiny chyb v něm lze snadno analyzovat a systém vylepšovat učením. Jistou výhodou je i možnost detekovat chyby mimo systém geokódování. Např. v případě zdravotnických zařízení byly zjištěny 2 hrubé chyby ve zdrojích – špatně zadaný kód obce (Karvinská hornická nemocnice s kódem obce Havířov) a špatné souřadnice 1 adresy v Prostějově dle RSO ČSÚ. Na obr. 3 je rovněž vidět, že jeden ze 2 bodů ZZ nebyl správně umístěn do OC Futurum, ale na ulici před ním (nedostatek v rozpoznání čísla).



Obr. 1 Mapa světa s ukázkou lokalizace Google pomocí kombinace ulice-PSČ



Obr. 2 Výsledek geokódování škol dle Mapy.cz a databázové zpracování



Obr. 3 Centrum Ostravy s ukázkou rozdílů v lokalizaci geokódovaných adres.

10. ZÁVĚR

Databázové geokódování podle adresy je výrazně pracné. Nicméně má větší potenciál pro kvalitní automatizované zpracování – používání různých známých zkratek a náhrad, využití různých číselníků, seznamů referenčních objektů, využití starších seznamů, které umožní lokalizaci neaktuálních, ale v zápisu adres se stále vyskytujících územních názvů. Významnou výhodou je možnost náhodného simulování polohy v rámci určené územní jednotky (pokud není známa přesná poloha), což je pro analýzy mnohem vhodnější řešení než koncentrace všech bodů např. do středu ulice a vytvoření falešné anomálie.

Existují i další možnosti, např. geokódovací aplikace Nominatim projektu OpenStreetMap, který však v případě více kandidátů neposkytuje vhodný výběr nejlepší z nich (Cícha, 2013).

Aplikace pro pokročilé databázové geokódování se dále vyvíjí.

11. PODĚKOVÁNÍ

Příspěvek byl podpořen grantem GAČR 14-26831S Prostorové simulační modelování dostupnosti. Děkujeme Ing. Antonínu Orlíkovi za transformaci souřadnic.

LITERATURA

Beaman, R. S., Conn, B. J. 2003: Automated geoparsing and georeferencing of Malesian collection locality data. In: Telopea, vol. 10(1), pp. 43–52. National Herbarium of New South Wales, Sydney.

Cícha, V. (2013): Správa, analýza a prezentace zdravotnických prostorových dat pomocí R. Olomouc, 2013. Bakalářská práce. Univerzita Palackého.

Edwards, S. E.; Strauss, B.; Miranda, M.L. 2014: Geocoding Large Population-level Administrative Datasets at Highly Resolved Spatial Scales. Transactions in Gis, 18 (4), pp. 586-603.

Google, 2014: The Google Geocoding API. On-line. https://developers.google.com/maps/documentation/geocoding/

Horák J.: Zpracování dat v GIS. 242 stran. VŠB-TU Ostrava 2013. 3.vydání

Horák J., Tesla J., Ivan I. 2015: Hodnocení dojížďky do zaměstnání v moravskoslezském kraji. In Sborník GIS Ostrava 2015.

Horák J., Belaj, P., Ivan I., Nemec P., Ardielli J., Růžička J.: Geoparsing of RSS News. Studies in Computational Intelligence, Volume 381, 2011, Pages 353-367. ISSN: 1860949X.

Košková, I., Kafka, Š. 2009: Geoparser – automatické vyhledávání geografických lokalizací v textu. In: Proceedings of Geoinformační infrastruktury pro praxi. 100 pp. MSD, Brno.

Zajíčková, L., Voženílek, V., Burian, J., Tuček, P., 2014, Demand specifications for geodata within a public transport system. Conference Proceedings SGEM 2014, 14th International Multidisciplinary Scientific GeoConference, STEF92 Technology Ltd., 8s.