VEŘEJNOSTÍ VYTVÁŘENÁ DATA – UNDERGROUND NEBO NOVÉ PŘÍLEŽITOSTI?

Jiří HORÁK[1]

[1] Institut geoinformatiky, Hornicko-geologická fakulta, Vysoká škola báňská – Technická univerzita Ostrava,

17. listopadu 15, 708 33 Ostrava, Česká republika
jiri.horak@vsb.cz

Abstrakt

S rozvojem internetu a sociálních sítí se objevují nové zdroje dat, vytvářené veřejností, které mohou být vhodně využity při realizaci různých průzkumů, marketingu, zlepšování navigace, organizace a regulace soukromé i veřejné dopravy. Zvláštní kapitolu tvoří přímo tzv. crowdsourcing, v případě geodat známější pod pojmem Volunteered Geographic Information (VGI). Patří zde zejména sdílení polohopisných geografických informací (např. tvorba digitálních map, sdílení lokalizačních záznamů), význam má ale i sdílení lokalizovaných pozorování (např. výskyt rostlinných či živočišných druhů, „lidová“ žurnalistika oznamující např. dopravní problémy), či vlastní provádění měření např. hluku, plynů, prachu atd. Vedle populárního OpenStreetMap lze najít ale i příklady vytváření pěší dopravní sítě z veřejností poskytnutých záznamů GPS a další dopravně či jinak orientované aplikace. Příspěvek poskytuje přehled možných aplikací a zamýšlí se nad problémy spojenými s využitím takových dat.

Abstract

The rapid progress of internet and social networks additionally generate a new data sources, created intentionally or non-intentionally by users. Such data can be utilised in various surveys, marketing, but also to improve navigation or improve regulation of both private and public transportation. The special segment of such data sources is called crowdsourcing, in the case of geographical data usually Volunteered Geographic Information. It covers sharing of geographical information (i.e. joined creating of digital maps), sharing of georeferenced observations like occurrences of special plants of animals, but also traffic problems, especially congestions, or measurements of noise, gas, dust. OpenStreetmap is not the only mapping project and the only way of utilisation of geographical data. Transport oriented applications may contain also projects to generate pedestrian networks from crowdsourced GPS records. The paper provides a review of possible applications and a reflection of issues linked with utilisation of such data sources.

Klíčová slova: crowdsourcing, Volunteered Geographic Information, internetové datové zdroje, GIS

Keywords: crowdsourcing, Volunteered Geographic Information, internet data source, GIS

1. ÚVOD

Cílem článku je upozornit na nové možnosti využití zdrojů dat, které vznikají na internetu. Současný web se někdy označuje jako web sémanticky propojených sociálních dat a to koresponduje i s vytvářením a využíváním nových zdrojů dat.

K základním typům datových zdrojů, zejména ve vazbě na možnost georeferencování dat, patří:

1. Datový sklad služby - poskytuje agregovaná a anonymizovaná data o chování uživatelů a využívání určitých internetových služeb. Kromě prostého zpřístupnění statistických dat o užívání služby může poskytovat i jistou aplikační nadstavbu a tak bychom mohli hovořit o Business Intelligence pro danou službu. V nejjednodušší podobě jsou uváděny četnosti návštěv webových stránek či počty využití služby. Takovou základní evidenci prohlížení jednotlivých stránek poskytuje např. Wikipedia, zatímco v případě vyhledávačů Google či Seznam jde již spíše o BI - Google Trends, statistiky Seznam.

Vybrané příklady použití datového skladu služby jsou uvedeny v článku Horák (2013).

2. Zdroje primárních dat - uživatelé vytváří primární data, sdílejí je v rámci určitých sociálních sítí s cílem poskytování vzájemných služeb, často označované jako crowdsoursing. Rozsah takových vytvářených dat je velmi široký – od informací obecného rázu (i jen připojované komentáře), přes měřená fyzikální data až společné vytváření produktů (cestopisy, digitální mapy apod.).
3. Zdroje dat vytvářené roboty – jsou to zdroje vytvářené cílenou automatizovanou činností, spočívající ve sběru, ukládání a případně agregaci dat dostupných na internetu. Může jít např. statistiku volných pracovních míst, sledování cen nemovitostí, účelové zpracování RSS zpráv nebo údajů o uživatelích sociální sítě. U těchto zdrojů dat je nutné řešit otázky etiky a autorských práv při využití takové služby. Vybrané příklady takových zdrojů dat jsou uvedeny v článku Horák (2013).

Vzhledem k rozsahu referátu se zaměřím pouze na zdroje primárních dat.

2. ZDROJE PRIMÁRNÍCH DAT

Je důležité tento zdroj vnímat dostatečně obecně. Nejde jen o tvorbu „amatérských" map. V nejjednodušší podobě to jsou projevy hodnocení či tagování, velmi populární v posledních letech na sociálních sítích. Některé studie ukazují, jak důležité mohou být tyto informace jako pomocné (metadata) při automatizovaném zpracování předmětných dat kolujících v sociálních sítích (např. pro klasifikaci a výběry snímků).

Velmi významné jsou také zprávy od členů sociálních sítí, které dobrovolně reportují o zajímavých či mimořádných událostech kolem nich. V tomto případě můžeme mluvit o tzv. občanské žurnalistice (*civil journalism*). Veřejnost je prvním reportérem, přinášejícím zprávy z terénu o událostech. Goodchild (2008) dokonce mluví o lidech jako o inteligentních lidských sensorech, jejichž výhodou je schopnost provádět zpracování dat a jejich analýzu, tedy jisté předzpracování dat, což je odlišuje od HW senzorů.

Navíc veřejnost sama aktivně informace dále šíří prostřednictvím sociálních sítí. Nezbytnou podmínkou k tomu je mobilní internet a jeho všudypřítomnost.

Kromě zpráv o čemkoliv zajímavém se může veřejnost nasměrovat k získání cílených hlášení o výskytu požadovaných objektů (viz mapování výskytu vybraných živočichů či rostlin, výskyt tornád a jiných specifických meteorologických nebo astronomických jevů). Nejde tedy jen o příležitostná pozorování.

Další variantou může být získávání měřených dat (tedy nikoliv textových či obrazových zpráv) od amatérských pozorovatelů – např. meteorologické záznamy (teplota, výška sněhu, dohlednost), měření obsahu plynů, velikost hluku atd. V některých případech je občan využíván pouze jako nosič příslušného měřícího zařízení.

Souhrnně se mluví o amatérsky vytvářených datech, ovšem hranice mezi amatéry a profesionály je často problematická. Neutrálnější je proto označení crowdsourcing. V případě geodat se vžil pojem Volunteered Geographic Information (Goodchild, 2007).

V souvislosti s novými zdroji dat se musíme zabývat více sociálními aspekty, psychologií a sociologií, abychom dobře porozuměli způsobu vytváření dat, motivaci autorů a jejich chování. Využíváme data, která nevytvářejí zaměstnanci s jasnou motivací, a nevznikají v důsledku organizací garantovaných procesů. Tomu musíme přizpůsobit i postupy zpracování dat, zejména zvýšenou kontrolu a filtraci dat.

V první řadě je potřebné porozumět důvodům vzniku příslušného zdroje. Důvody motivace autorů mohou být (upraveno a doplněno z Goodchild 2008, Ding et al. 2009, Haklay, Weber, 2008):

- Snaha o zviditelnění, o uznání, exhibicionismus
- Altruismus
- Snaha pomoci, vyplnit existující mezeru
- Víra ve význam poskytování volných informací a zlepšení světa

- Odmítavý postoj k oficiálním datovým zdrojům (např. národním mapovacím agenturám)
- Potřeba být členem určitého sociálního prostředí a sociální skupiny
- Získat odezvu na své aktivity
- Uspokojení z vlastního procesu pořizování dat (rádi mapují, fotí, objevují nové možnosti atd.). Může tomu napomoci i samotná vhodná aplikace, která podněcuje a inspiruje k tvorbě popisů, doplňování informací. Např. Ames, Naaman (2007) tvrdí, že aplikace jako Flickr a ZoneTag stimulují lidi, aby fotky popisovali, zatímco lidé příliš nepopisují fotky na svém pevném či mobilním zařízení, přestože by toho mohli objektivně dobře využít, např. pro vyhledávání.

Je nutné se také zabývat ochranou takto vytvářených dat. Na ochraně soukromí pracuje např. projekt 7.rámcového programu EU First location bank (http://www.firstlocationbank.com/), který by měl mimo jiné umožňovat oddělit identitu osoby od identity záznamu, bezpečně ukládat datové záznamy v „bance“ a plně řídit nakládání s nimi, např. jejich veřejné využití, prodej apod.

Pokud se zaměříme na VGI, tj. na georeferencované informace, pak k hlavním oblastem využití patří:

- Sdílení polohopisných geografických informací – např. tvorba digitálních map, sdílení lokalizačních záznamů.
- Sdílení lokalizovaných pozorování – např. výskyt druhů, výskyt událostí, provádění měření
- Sdílení lokalizovaných fotografií.

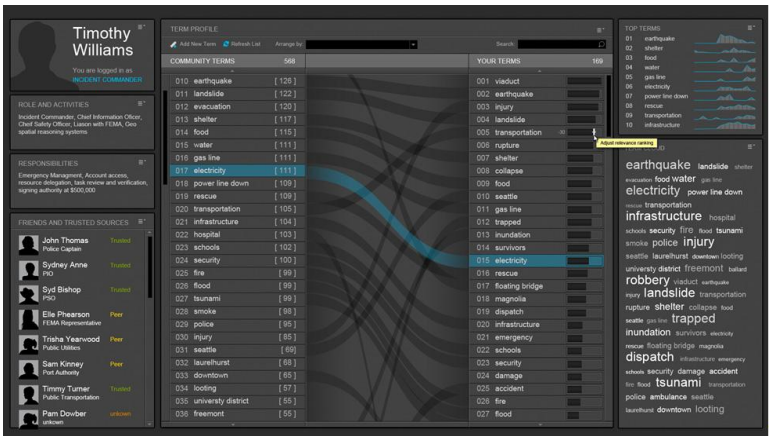
Dále uvedené příklady dokumentují občanskou žurnalistiku, Openstreetmap a Flickr.

2.1 Občanská žurnalistika

První možností jsou obecné zprávy o událostech v území.

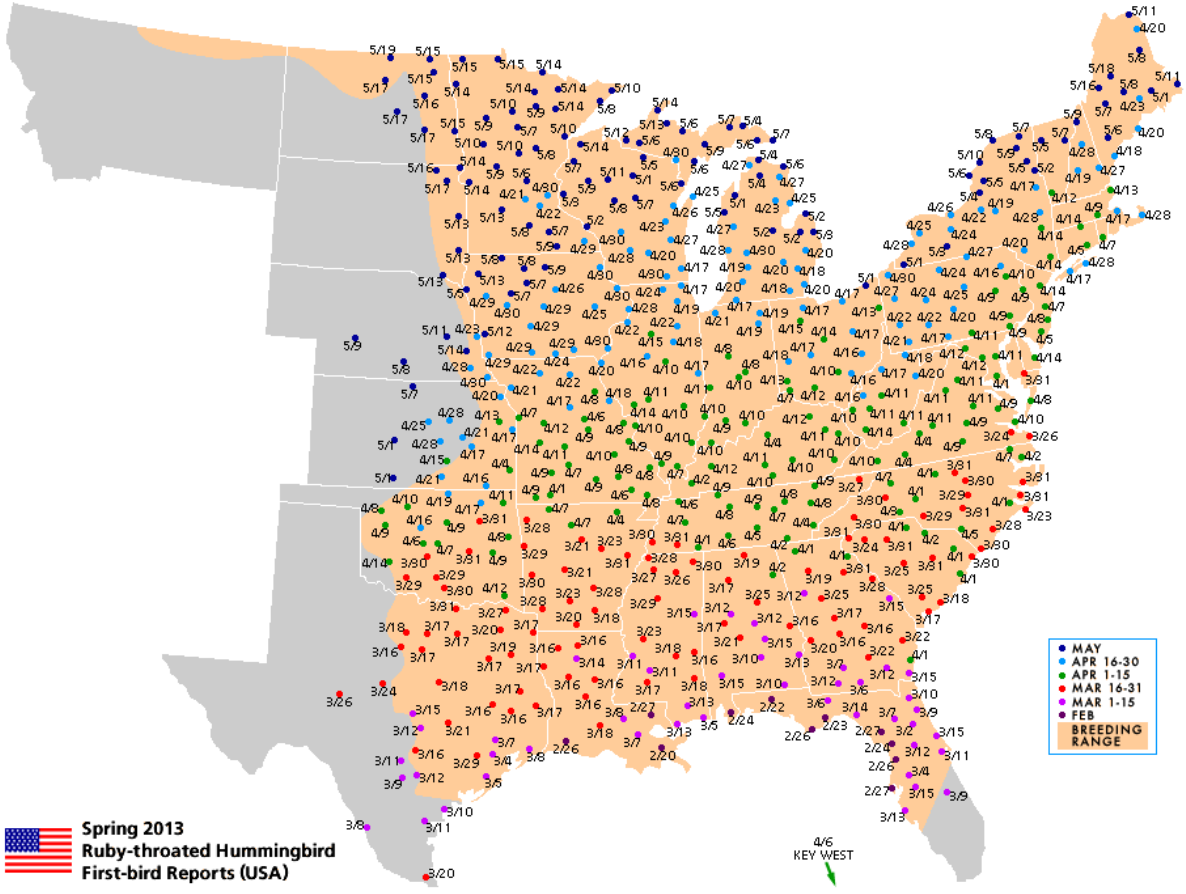
Je důležité, že pro podporu takových aplikací jsou k dispozici dostupné platformy např. GeoChat (http://instedd.org/technologies/geochat/) a Ushahidi (http://ushahidi.com/products/ushahidi-platform) a také standardy jako např. standardy Open Geospatial Consortium’s Sensor Web Enablement či Open GeoSMS, který umožňuje získat souřadnice, vložit je do SMS a tím zajistit interoperabilní komunikaci souřadnic místa a obsahu pro různá zařízení a aplikace. Open GeoSMS je implementován v Ushahidi a také v open source Sahana Disaster Management System. Jejich velkou výhodou je, že se dají použít na obyčejných telefonech, na chytrých telefonech ať již připojených na internetu (pak se mapuje událost na pozadí Google Maps, či jiného mapového serveru), nebo bez připojení (např. použití OpenStreetMap v offline režimu).

Jaký význam mohou mít takové obecné zprávy o událostech, ukazuje aplikace pro krizové řízení. V Pacific Northwest National Laboratory vyvinuli budoucí pracovní prostředí pro štáb krizového řízení, která umožňuje využívat velké objemy a proudy dat, včetně sociálních médií, zpracovat je a analyzovat tak, aby zodpovědné osoby dostávaly data šitá na míru (Boulous et al., 2011). Základem je profil uživatele, kde si definuje svoje zájmy a potřeby pomocí výběru výrazů a přidělení vah. Přicházející informace ze sociálních sítí (libovolné zprávy, které se šíří v sociálních sítích) se podle toho filtrují, posuzuje se shody výrazů nalezených v sociálních zprávách s výrazy, kterými specifikuje svůj profil uživatel (obr. 1), a jsou mu doručovány jen ty relevantní. Postupně se navíc přesnost systému zvyšuje, protože se systém dále učí na základě spokojenosti uživatele s doručenými zprávami.



Obr. 1 Profil uživatele v systému PIE (Boulous et al., 2011)

Kromě obecných možností zpráv o různých událostech, kterými je reportér svědkem, je možné iniciativu veřejnosti zaměřit na vybrané zájmové objekty. Typickým příkladem je veřejností hlášený pozorovaný výskyt živočišných a rostlinných druhů, případně dalších přírodních jevů. Příkladem pro hlášení a mapování je web http://www.learner.org/jnorth/maps/Gallery.html pro oblast severní Ameriky. Tento web se soustřeďuje na hlášení výskytu motýlů druhu Monarcha stěhovavý, šedých velryb, orlů, jeřábu amerického, kolibříků, drozdů a délky slunečního svitu během dne. Existují i jiné specializované weby zaměřené na jednotlivé druhy včetně možnosti hlášení pozorování např. http://www.hummingbirds.net. Lokalizace se provádí plnou adresou, ZIP kódem nebo obecnějším popisem polohy, ale také přímým umístěním značky v Google Maps mash-up. Záznam se ukáže na mapě až po ověření. V roce 2013 hlásilo první výskyt kolibříků 5829 lidí. Výsledná registrace prvních pozorování ukazuje zřetelně závislost na zeměpisné šířce (obr. 2).



Obr. 2 První pozorování vybraného druhu kolibříků (ruby-throated hummingbird) v roce 2013 ve východní části USA (http://www.hummingbirds.net/map.html).

Jinou možností je sběr dat o hluku v jednotlivých místech. Kanhere (2011) popisuje využití lokalizovaných audio vzorků získaných na ulici, které veřejnost zasílá do centra prostřednictvím chytrých telefonů. V centru jsou data agregována a vytvářejí aktuální mapy hlukového znečištění ve městě každou hodinu.

Zajímavým příkladem využití crowdsourcingu s pasivní rolí uživatele je také aplikace Common Scents, kde se používá flotila kol, vybavených Sensaris City Senspody. Měří se CO, NOx, hluk, teplotu a relativní vlhkost (potřebná kvůli kalibraci dat) a měření se zaznamenávají společně s polohou a datem/časem (Boulos et al., 2011).

2.2 Digitální mapy

OpenStreetMap (OSM) představuje v současnosti asi nejrozsáhlejší crowdsource mapovací projekt. Jeho vznik se datuje do června 2004 na University College London (Haklay, Weber 2008). Počet registrovaných uživatelů se odhaduje na více než půl milionu (půl milionu dosaženo v únoru 2012 (Kasemsuppakorn, Karimi, 2013).

Vytvářet mapy mohou jen registrovaní uživatelé. Kromě tvorby vektorových topografických prvků lze také k nim připojovat fotky, audio záznamy či živě zaznamenávat externí GPS data (Haklay, Weber 2008). Atributy jsou ukládány jako páry „klíčové slovo; hodnota“ (např. „hospoda; Myslivna“).

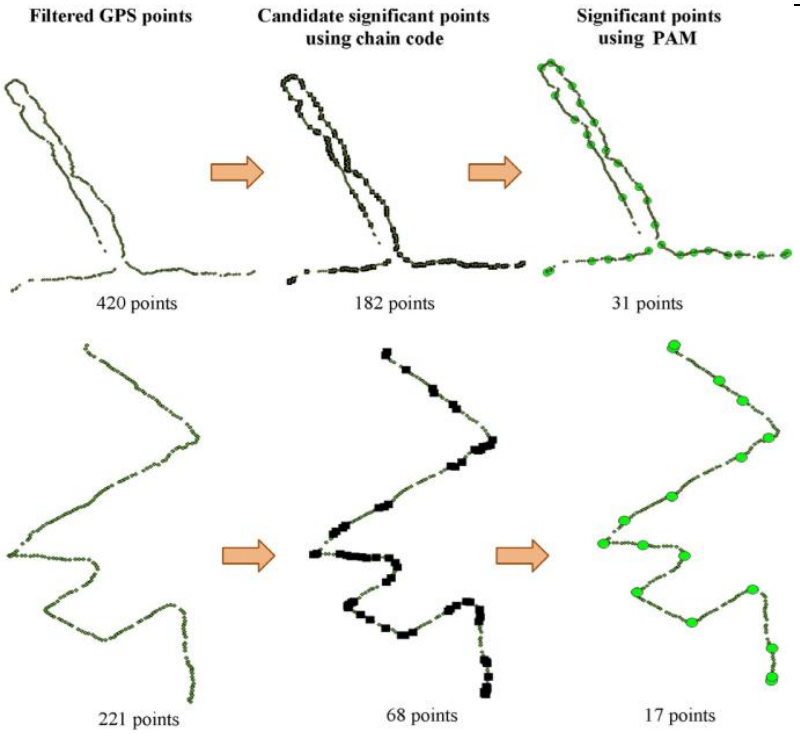
Pro někoho může být překvapením, že OSM využívá nejen amatérská data, ale také volně dostupné datové zdroje např. v USA data systému TIGER, nebo že využívá data veřejné správy. Např. město Rostock umožnilo v srpnu 2009 importovat všech 17000 městských budov do OSM (Over et al., 2010). Dokonce jsou známy i příklady spolupráce s komerčními subjekty. Díky poskytovateli navigačních dat Automotive Navigation Data (Haklay, Weber 2008) se první zemí s kompletním uličním pokrytím stalo Nizozemí. To jen ukazuje, jak tenká může být hranice mezi profesionálními a amatérskými systémy.

OSM je také zajímavé z pohledu organizace sociální sítě, zejména snaze o vzdělávání tvůrců map a jejich koordinaci. Pro začínající editory se organizují tzv. mapping party. Koordinace probíhá v oblasti pokrytí chybějícího území nebo doplnění chybějících atributů. Tvůrci OSM jsou např. požádáni, aby sbírali specifický typ potřebné informace (např. typ střechy a počet pater, potřebný ke generování 3D krabicových modelů budov) (Over et al., 2010). Nejvyšší formou koordinace je příprava vlastních mapovacích kampaní – jedna z prvních proběhla v roce 2006 na ostrově Wight, kde se cca 30 účastníků podílelo na kompletním zmapování ostrova (Haklay, Weber 2008).

OSM neslouží jen pro vytváření snadno dostupných map území. Sbíraná data mohou být využita i jiným způsobem.

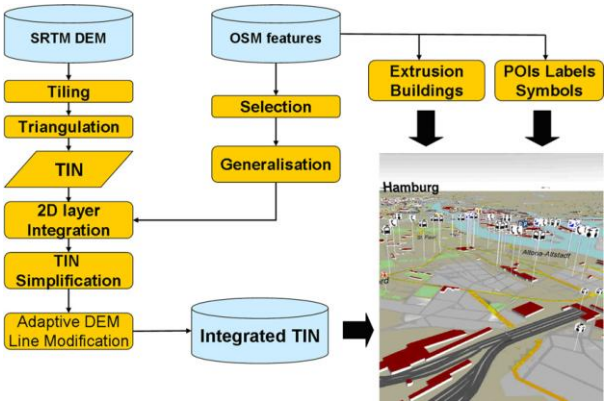
Kasemsuppakorn a Karimi (2013)] použili záznamy GPS dostupné na OSM k automatickému vygenerování pěší dopravní sítě, která je jinak jen velmi obtížně a nákladně získatelná. Vhodné GPS trasy vybrali pomocí klíčového slova „pedestrian“ v tagech tras. Nový algoritmus, navržený pro automatizované zpracování tras, zahrnuje:

1. Předzpracování – zajišťuje zejména odstranění chyb jako je problém GPS Time-To-First-Fix a nahodilé chyby (výsledky obskurních signálů GPS). Filtrace se provádí na základě podmínek, že počet družic musí být větší než 3 a HDOP menší než stanovený limit.
2. Výběr významných bodů – nejdříve se počítá směr jednotlivých kroků, kóduje se do jednoduché růžice směrů, následně se vyberou kandidáti na významné body a použije se místní shlukování k nalezení významných bodů – metoda Partitioning Around Medoids (PAM) (Kaufman, Rousseeuw, 1987) (obr. 3).
3. Konstrukce pěší sítě - postup je schopen nejen vytvářet nové sítě, ale také aktualizovat stávající sítě a propojovat jejich segmenty.



Obr. 3 Postupná redukce GPS bodů (Kasemsuppakorn, Karimi, 2013)

Over et al. (2010) demonstruje použití dat OSM pro generování 3D modelu města, v kombinaci s výškovými daty ze Shuttle RADAR Topography Mission (SRTM) (obr. 4). Část dat z OSM (klasifikace využití území, silniční a železniční síť) používají jako fixní při tvorbě digitálního modelu reliéfu (společně s daty SRTM), ostatní data z OSM se považují za živá a aktualizují se denně nebo týdně. Automatizovaně generují 3 nové kategorie objektů (popisy, budovy a body zájmu), které umísťují do 3D scény. Při tvorbě 3D modelů budov se využívá počtu pater, které tvůrci OSM zapisují k budovám. Další atributy (typ střechy) byly navrženy k doplnění.



Obr. 4 Postup tvorby městské 3D scény v OpenStreetMap (Over et al. 2010)

Samozřejmě je potřebné také otevřeně mluvit o problémech OSM. Podle Haklay, Weber (2008) k nim patří:

- není jasná vhodnost dat pro různé účely,
- vliv geografie a participace v projektech,
- průběh aktualizace dat,

- licencování – původně Creative Common Framework, která je ale pro některé úkoly nevhodná, protože vytvořená mapa musí být šířena pod stejnou (tedy volnou) licencí, což ale nelze použít např. při vizualizaci citlivých údajů,
- nerovnoměrné pokrytí (heterogenita). Haklay (2009) provedl analýzu OSM v Anglii a zjistil, že za 4 roky existence OSM je pokryto 29% Anglie, z toho asi 24% jsou ale digitalizované linie bez atributů. V jednotlivých místech může být pokrytí velmi dobré, např. uliční síť v Hamburgu byla již v roce 2009 hotova z 99,8% (http://www.openstreetmap.de/presse/2008-10-24-hamburg-stat.html).
- nerovnoměrná aktivita lidí – 99,8 registrovaných uživatelů neprovádí žádná vylepšení mapy a skutečnými tvůrci je jen několik osob.

Často se také diskutuje otázka přesnosti, nejenom pokrytí, a objevuje se řada článků, srovnávající výsledky OSM s oficiálními datovými zdroji v dané oblasti. Např. Over et al. (2010) oceňovali přesnost v Německu a zjistili horizontální přesnost kolem 10 m, výškovou přesnost asi 25 m. Absolutní přesnost pěších cest generovaných z bodů je asi 5 m. Haklay (2009) na základě srovnání přesnosti v Londýně se záznamy Ordenance Survey určil, že v průměru se liší záznamy asi do 6 m.

OpenStreetMap není jediným projektem, který umožňuje vytvářet a využívat sdílené digitální mapy. Wikimapia (http://www.wikimapia.org) vznikla v roce 2006 a představuje prostředí pro vytváření mapy, připojení fotek, komentářů a vazeb na Wikipedia. V červenci 2012 měla síť úpajně přes 1 mil. uživatelů, kteří společně vytvořili 19 mil. zákresů.

Jinou, úzce specializovanou sítí, určenou pro milovníky outdoorových aktivit, je WikiLoc (http://www.wikiloc.com). Umožňuje sdílení doporučených tras, fotek, popisů a různých klasifikací. V srpnu 2013 měli asi 790 tis. uživatelů a 1,5 mil. outdoorových tras.

2.3 Systémy pro sdílení fotografií

Jedním z nejvýznamnějších systémů pro sdílení fotografií je Flickr (http://www.flickr.com), a to díky počtu uživatelů, objemu dat i snadnému přístupu pomocí Java API. Jeho vlastníkem je Yahoo!, který jej v roce 2005 koupil od tvůrců. V březnu 2013 měl asi 87 registrovaných uživatelů a celkově obsahuje více než 3 miliard fotek, z toho více než 100 mil. geotagovaných (Hong et al., 2013).

Flickr dělí sbírky fotek (datové sady) na 2 typy - individuální fotosady a veřejné skupiny. Kolekce fotografií je možné klasifikovat do 8 předdefinovaných kategorií (např. aktivita, místo, věc, osoba) a lze si také vytvářet vlastní třídy.

Více než 50 % fotosad a rovněž fotek nemělo v roce 2009 vůbec žádný popis (Stvilia, Jörgensen, 2009). Vedle pojmenování je možné k fotce připojit tagy, tedy klíčová slova. Bohužel jsou tagy často nejednoznačné, neúplné a příliš personalizované. Dají se rozdělit na objektivní a subjektivní. Je to mimo jiné i důsledkem toho, že uživatelé často nerozumí významu a správnému způsobu tagování. Oddělovač tagů je mezera, což vede při špatném použití k tomu, že se mezi tagy objevují spojky, a na druhou stranu i k používání skládaných slov bez mezer, které však nejsou jazykově korektní (např. blackandwhite či projectmanagement) (Ding et al., 2009), což ovšem může komplikovat zpracování.

Význam tagů u fotografií je velký, protože nám významně pomáhá ve výběru fotek, jejich vyhledávání, a také umožňuje další analýzy a aplikace, např. odhady rozlišení mezi profesionálními a amatérskými fotografy (Ding et al., 2009), doručování personalizovaných informací, analýzu pohybu turistů (De Choudhury et al., 2010), (Asakura, Iryo 2007), (Lewa, Mckerchera, 2006), (Zheng et al., 2012), rozeznávání krajinných prvků (Li et al., 2009), vytváření agregované znalosti o geografické oblasti (Ahern et al., 2010), (Zheng et al., 2012), měření podobnosti snímků, zlepšení tagování obrázků, doporučení vhodných tagů pro neoznačené obrázky, zlepšení úspěšnosti nalezení podobných obrázků na základě tagů, zavedení systému a zlepšení sémantiky tagů (Ding et al., 2009), predikce umístění fotek z vizuálních, textových a časových charakteristik či jejich série (Crandall et al., 2009), vztahy mezi textovými koncepty, geografickými místy a událostmi (Rattenbury et al., 2007), vytváření nových aplikací (Ahern et al., 2010), (Hong et al, 2013).

Např. Li et al. (2009) navrhují novou metodu určování relevance (objektivnosti) tagů. Na základě obrazové podobnosti hledají podobné obrázky, přečtou jejich tagy, a podle překryvu a frekvence určit ty tagy, které jsou objektivní, protože se nacházejí i u podobných obrázků. Rovněž podle tagů vyhledávali podobné obrázky (metoda Tag-Based Image Retrieval) a dosáhli pozoruhodně vysoké přesnosti. Zkoušeli také doplnit tagy u neoznačených obrázků, ale navržený postup zatím nedosáhl uspokojivých výsledků (obr. 4).

Visual Search		**Suggested tags by different methods**		
Image	*Accuracy*	*baseline* [7]	*baseline* [8]	*tagRelevance*
	0.77	*flower* *red* macro nature garden	*flower* *red* macro *rose* garden	*flower* *red* macro *rose* garden
	0.00	2006 family japan beach vacation	2006 cat family campcourtney august12006	icehockety hockey family hurricane cat
	0.10	2006 wedding japan park vacation	2006 pepperell wedding japan park	japan ***bike*** hiking park texas

Obr. 5 Srovnání přesnosti tagování 3 neoznačených obrázků pomocí metody tag relevance a jiných (Li et al., 2009).

Je zřejmé, že Flickr neslouží jen jako zdroj fotek a fotografického porovnávání míst, která nás v mapě zajímají. Jednou z pozoruhodných možností je identifikace turistických oblastí a tras z posloupnosti fotografií, která provedl Zheng et al. (2012) ve 4 světových městech (Paříž, Londýn, New York a San Francisco). Celkově bylo zpracováno cca 800 tisíc fotek od 23 tisíc uživatelů Flickr, kde ke každé fotce jsou k dispozici ID, GPS souřadnice, datum a čas, identifikátor fotografa, resp. toho, kdo je umístil na web, a připojený text. Fotky každého fotografa seskupili po dnech. Následně provedli rozlišení turistů a neturistů (metoda filtrace na základě výpočtu entropie pro počet fotek v pravidelné mřížce) s cílem vyloučit z dalšího zpracování data, pocházející od neturistů. Poté identifikovali oblasti zájmu v jednotlivých městech pomocí shlukovacího algoritmu DBSCAN s využitím hustoty lokalizovaných snímků. Dále studovali pravděpodobnost přesunu od jedné oblasti zájmu k druhému pomocí Markovových řetězců. Následovala topologická analýza tras, kdy analyzovali počet oblastí zájmu navštívených během 1 trasy a popsali nejpopulárnější trasy (obr. 5). Analýza pak pokračovala klasifikací tras na náročné a odpočinkové, popisem jejich vlastností. Tyto informace využili v další klasifikaci tras, jejich seskupování (hierarchické shlukování) a pokročilé analýze.

Obr. 6 Hlavní oblasti zájmy a trasy přesunu mezi nimi v San Francisco (Zheng et al., 2012)

2.4 Doporučený postup při využití primárních internetových zdrojů dat

Ke sběru a zpracování dat je možné doporučit následující obecný postup:

- Příprava – Je potřebné dobře zvážit výběr informačních zdrojů, ověřit jejich vhodnost pro daný účel a oblast, a prověřit dodržení legislativních a etických pravidel. Následně je potřebné se zabývat volbou vhodných parametrů sběru a způsobu tvorby nové kolekce, zda se mají využít všechna data či jejich část pořízená náhodným či definovaným systematickým výběrem, případně jejich kombinací. Samozřejmě pokud je to potřebné, musí se vyvinout a otestovat vlastní programový nástroj pro takový sběr dat.
- Sběr dat – Zahrnuje jednorázové či systematické opakování stahování či jiného těžení dat. Využívá se vhodné formy uložení dat, včetně všech potřebných metadat.
- Zpracování dat – Zahajuje se především kontrolou dat na jejich úplnost, rozsah, dodržení stanovených integritních omezení, explorační analýzou dat s cílem především detekce anomálií. Následují již podle účelu různé varianty filtrace (selekce) dat, jejich validace vůči realitě nebo srovnání s jinými zdroji, analýza dat, aplikace vhodných statistických nástrojů, a samozřejmě i volba vhodné formy výstupů.
- Interpretace dat – při interpretaci výsledků je nutné vzít v úvahu zájmy tvůrců dat (proč vznikal tento primární zdroj dat), možné vychýlení vůči sledované výzkumné otázce či hypotéze, aspekty ovlivňující výsledek, rozsah dat, možné ovlivnění vnějšími událostmi při sběru dat atd.

3. ZÁVĚR

V současnosti se na internetu vytváří řada nových datových zdrojů, zejména v důsledku rozvoje sociálních sítí a vzniku nových služeb.

K hlavním výhodám těchto zdrojů dat patří dostupnost (data jsou veřejně a stále dostupná), následně i snadná verifikace výsledků, bezplatnost, možnost kontinuálního sledování v čase, velký objem dat, možnost integrace více zdrojů dat pro jednu úlohu, aktuálnost (neustále živý zdroj dat s možností sledování změn) a v některých případech i snadnost interpretace (většinou snadná interpretovatelnost).

Ke správnému a efektivnímu využití však uživatel však musí nejdříve zvážit řadu aspektů, které komplikují jejich využití. Jde zejména o důvěryhodnost (jsou to netradiční zdroje dat, nejsou řízeny s jasnou autoritou a

garancí), ochrana osobních údajů (analýzy osobních záznamů, fotek apod.), nutnost řízené agregace dat, problém jednostrannosti a závislosti dat (data vázaná na jednoho poskytovatele, jistý profil původních uživatelů), problém harmonizace při spojování více zdrojů (řešení nekonzistencí metodických, časových, sémantických), případně i chudost dat (malá výbava atributy).

Článek se zaměřil na některé zajímavé možnosti využití těchto datových zdrojů. Vhodnými příklady jsou např. analýzy chování veřejnosti, jejich profilování, zjišťování názorů, aplikace dat pořízených uživateli (fotky či GPS trasy či zákresy objektů na OpenStreetmap, pozorování kolibříků), získávání nových kontaktů, ale i podpora krizového řízení (sběr primárních informací, early warning), návrhy nových aplikací atd.

Veřejností vytvářená data nemusí být v protikladu s oficiálně vytvářenými zdroji dat. Poskytují výhodné možnosti vzájemného doplňování a koexistence. Tvoří neoddělitelnou součást geoinformační infrastruktury a podporují její udržitelnost.

LITERATURA

1. Ahern, S., Naaman, M., Nair, R., Yang, J.: World Explorer: Visualizing Aggregate Data from Unstructured Text in Geo-Referenced Collections. In: *Proc. of the 7th ACM/IEEE Joint Conference On Digital Libraries*. (2007) 1-10.
2. Ames, M., Naaman, M.: Why We Tag: Motivations for Annotation in Mobile and Online Media. In: *Proc. of Conf. On Human Factors In Computing Systems*, Vols 1 and 2. (2007) 971-980.
3. Asakura, Y., Iryo, T.: Analysis of tourist behaviour based on the tracking data collected using a mobile communication instrument. *Transport. Res. Part A: Policy and Pract.* 41, 7 (2007) 684–690.
4. Boulos M.N.K., Resch B., Crowley D.N., Breslin J.G., Sohn G., Burtner R., Pike W.A., Jezierski E., Chuang K.-Y.S.: Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. *International Journal of Health Geographics*, vol. 10, 67 (2011).
5. Crandall, D. J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: *Proc. of the 18th International Conference on World Wide Web*. ACM, New York (2009) 761–770.
6. De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., Yu, C.: Constructing travel itineraries from tagged geo-temporal breadcrumbs. In: *Proc. of the 19th International Conference on World Wide Web* (WWW'10). ACM, New York, (2010) 1083–1084.
7. Ding, Y., Jacob, E.K., Zhang, Z., Foo, S., Yan, E., George, N.L., Guo, L.: Perspectives on Social Tagging. *Journal of the American Society for Information Science and Technology*, Vol. 60 (12) (2009) 2388–2401.
8. Goodchild, M.: Assertion and Authority: User-Generated Geographic Content. In: *Proc. of AGIT 2008*, Salzburg, 1-4.7.2008.
9. Goodchild M.F.: Citizens as sensors: web 2.0 and the volunteering of geographic information. *GeoFocus* (Editorial), N. 7, (2007), 8-10. ISSN: 1578-5157
10. Haklay, M., Weber, P.: OpenStreetMap: User-Generated Street Maps. *Pervasive computing*, vol. 4 (2008) 12-18.
11. Haklay M.: How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, vol.37(4) (2009) 682-703.
12. Hong, R., Zha, Z.-J., Gao, Y., Chua, T.-S., Wu, X.: Multimedia encyclopedia construction by mining web knowledge. Signal Processing 93 (2013) 2361–2368.
13. Horák J.: Nové internetové datové zdroje. In Pokorný J., Šaloun P. (eds.): Datakon 2013, 13.-15.10. 2013, VŠB-TU Ostrava, pp.65-95. ISBN 978-80-248-3189-3

14. Kanhere S.S.: Participatory Sensing: Crowdsourcing Data from Mobile Smartphones in Urban Spaces. In: Proc. of *2011 12th IEEE International Conference on Mobile Data Management (MDM)*. (2011) 3-6

15. Kaufman, L., Rousseeuw, P.J.: Clustering by means of medoids. In: Dodge, Y. (Ed.), *Statistical Data Analysis based on the L1 Norm*. Elsevier, Amsterdam. (1987) 405–416.

16. Lewa A., Mckerchera, B.: Modeling tourist movements: A local destination analysis. Annals Tourism Res. 33, 2 (2006) 403–423.

17. Kasemsuppakorn, P., Karimi, H.A.: A pedestrian network construction algorithm based on multiple GPS traces. Transportation Research Part C 26 (2013) 285–300.

18. Li, Y., Crandall, D. J., Huttenlocher, D. P.: Landmark classification in large-scale image collections. In: *Proc. of the International Conference on Computer Vision*. (2009) 1957–1964.

19. Over, M., Schilling, A., Neubauer, S., Zipf A.: Generating web-based 3D City Models from OpenStreetMap: The current situation in Germany. *Computers, Environment and Urban Systems 34* (2010) 496–507

20. Stvilia, B., Jörgensen, C.: User-generated collection-level metadata in an online photo-sharing system. *Library & Information Science Research* 31 (2009) 54–65.

21. Zheng, Y.-T., Zha, Z-J., Chua, T-S.: Mining travel patterns from geotagged photos. *ACM Trans.Intell. Syst. Technol.* 3, 3, Article 56 (May 2012), 18 stran.

22. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from flickr tags. In: *Proc. of the ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York (2007) 103–110.