VĚDECKÉ SPISY VYSOKÉHO UČENÍ TECHNICKÉHO V BRNĚ

Edice Habilitační a inaugurační spisy, sv. 597
ISSN 1213-418X

Otakar Čerba

IDENTICKÉ VAZBY PROPOJENÝCH PROSTOROVÝCH DAT

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta stavební
Ústav geodézie

Ing. et Mgr. Otakar Čerba, Ph.D.

IDENTICKÉ VAZBY PROPOJENÝCH PROSTOROVÝCH DAT

IDENTITY LINKS OF LINKED SPATIAL DATA

ZKRÁCENÁ VERZE HABILITAČNÍ PRÁCE

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
NAKLADATELSTVÍ VUTIUM

BRNO 2018

KLÍČOVÁ SLOVA

Propojená data, prostorová data, identické vazby, graf, zdroj propojených dat.

KEYWORDS

Linked data, spatial data, identity links, graph, Linked data resource.

MÍSTO ULOŽENÍ PRÁCE

Originál práce je uložen v archivu PVO FAST VUT v Brně

ISBN 978-80-214-5631-0
ISSN 1213-418X

Obsah

Představení autora

Jméno a příjmení: Otakar Čerba

Narozen: 17. června 1975 v Klatovech

Bydliště: Luční 898, Klatovy

Odborné zaměření – propojená prostorová data, sémantické aspekty prostorových dat, tematická kartografie, aplikace značkovacích jazyků a propojených dat v kartografii

Vzdělání

- Gymnázium Jaroslava Vrchlického Klatovy (1993)
- Fakulta aplikovaných věd Západočeské univerzity v Plzni – Geomatika (1999)
- Pedagogická fakulta Západočeské univerzity v Plzni – Učitelství zeměpisu pro střední školy, Učitelství výpočetní techniky pro střední školy (2000)
- Přírodovědecká fakulta Univerzity Karlovy v Praze – postgraduální studium oboru Kartografie, geoinformatika a dálkový průzkum Země (2012)

Přehled o praxi

- 2000-současnost Fakulta aplikovaných věd Západočeské univerzity v Plzni
- 2009-současnost České centrum pro vědu a společnost
- 2007-2013, 2014-2015 Help Service Remote Sensing

Stáže a studijní pobyty

- březen-červen 2014 – Food and Agriculture Organization (Itálie)
- listopad 2013 – Research Group Cartography, Department of Geodesy and Geoinformation, Vienna University of Technology (Rakousko)
- září 2012 – Department of Geography / Geology, University of Nebraska at Omaha (USA)

Řešené projekty (výběr)

- INTECOM – Výzkum a vývoj inteligentních komponent pokročilých technologií pro plzeňskou metropolitní oblast
- Peregrinus Silva Bohemica
- TB0500MV003 Metodika pro publikování prostorových informací ve formě otevřených dat (řešitel)
- SDI4Apps – Uptake of open geographic information through innovative services based on linked data

- SmartOpenData – Linked Open Data for environment protection in Smart Regions
- NeoCartoLink
- CentraLab – Central European Living Lab for Territorial Innovation
- Plan4business
- Plan4all
- EXLIZ – Excelence lidských zdrojů jako zdroj konkurenceschopnosti regionu
- A-Math-Net Síť pro transfer znalostí v aplikované matematice
- INSPIRE Thematic Working Group Natural Risk Zones
- Humboldt
- F0584/2011/F1d Geomatika multimediálně (řešitel)

Pedagogická činnost

- Výuka předmětů Tematická kartografie, GeoWeb, Geografická kartografie, Socioekonomická geografie pro geomatiku, Historie map a mapování, Úvod do geomatiky
- Vedení 18 bakalářských a 16 diplomových prací

Publikační činnost

- 16 publikací, 19 citací, h-index 2 (Web of Science)
- 23 publikací, 26 citací, h-index 3 (Scopus)
- 87 publikací, 154 citací, h-index 6 (Google Scholar)

Členství v odborných organizacích a skupinách

- Commission on Maps and the Internet, International Cartographic Association (zástupce České kartografické společnosti)
- Člen předsednictva České asociace pro geoinformace (CAGI)
- Člen České kartografické společnosti
- České centrum pro vědu a společnost (člen statutárního orgánu)

Úvod

V současnosti přestává být zásadním problémem většiny oborů pracujících s prostorovými daty a informacemi jejich množství a dostupnost. Klíčovým úkolem se stává především zajištění jejich kvality, včetně konzistence, aktuálnosti, přesnosti, podrobnosti, popisu pomocí metadat a sémantických anotací (prvky kvality prostorových dat jsou popsané například v publikacích [1] nebo [2]). S kvalitou velice úzce souvisí také možnosti sdílení prostorových dat a kombinace celých datových sad nebo pouze vybraných prvků pocházejících z různých zdrojů. Přesto v geomatice, geoinformatice a příbuzných oborech převažuje do jisté míry anachronický přístup, kdy se velké množství uživatelů nerozhoduje o datech na základě jednotlivých aspektů kvality a vhodnosti pro dané řešení, ale na základě pouhé dostupnosti (nikoli přístupnosti).

To však s sebou nese jistá rizika. Tím hlavním je složitá orientace v labyrintu prostorových dat a informací. Uživatelé tudíž často upřednostňují pořízení nebo nákup nových dat před (často marginální a jednoduchou) úpravou existujících datových sad nebo kombinací více existujících datových bází za účelem získání (odvození) potřebných informací. Množství nově sbíraných prostorových dat a informací celosvětově neustále prudce narůstá. Na tomto nárůstu se rapidně podílí především různá senzorová měření, produkty fotogrammetrie a dálkového průzkumu Země, geodetická měření, laser scan, ale i sekundární data vzniklá zpracováním dat primárních (originálně pořízených). Podle článku [3] zhruba 90% dat na celém světě vzniklo v posledních dvou letech[1]. Tento fakt víceméně potvrzuje i starší studie [4], která uvádí, že celosvětový objem dat se každé dva roky zdvojnásobí. Obě výše uvedené informace se sice nevztahují výhradně na prostorová data, ale je zcela jasné, že tato prudce rostoucí podmnožina obecných dat a informací bude celosvětové trendy spíše podporovat než vyvracet.

Rychlý nárůst objemu prostorových dat a informací je pochopitelný v případě periodicky automatizovaně sbíraných dat a informací, jako jsou například družicové snímky nebo senzorová měření, protože uživatelé potřebují a využívají aktuální informace například pro předcházení různým situacím souvisejícím s přírodními i dalšími riziky (například povodně nebo zemětřesení), případně jejich řešení, nebo pro úlohy zpracovávající aktuální dopravní situaci. Podobně náročné na množství dat jsou i komparativní studie ukazující změny nejen v prostoru, ale i čase. Přesto je legitimní otázka, zda není možné (a také levnější či efektivnější) nahradit sběr a přípravu nových datových sad lepším využíváním dat stávajících, včetně jejich propojování za účelem odvození nových informací a souvislostí. O nedostatečném využívání existujících dat vypovídá i další údaj – méně než 1% z celosvětových dat je analyzováno [5]. Jinými slovy, mnoho dat je pouze získáno (změřeno nebo spočítáno), ale již se s nimi neprovádí žádné další operace

[1] Vzhledem k datu publikování článku [3] je tato informace z hlediska absolutních čísel nesprávná, ale vhodně ilustruje prudký nárůst dat.

(kromě uložení), a často nejsou ani dále využívána.

Proto je možné v souvislosti se sběrem a dalším využíváním prostorových dat velmi často hovořit o plýtvání. To se týká nejen finančních prostředků určených na pořízení dat, ale také nákupu a provozu zařízení pro jejich získávání a ukládání, pracovní síly nebo času. Kromě výše uvedených úspor s potřebou efektivnějšího využívání a sběru prostorových dat souvisí například tvorba infrastruktur prostorových dat (SDI) na různých úrovních nebo potřeba mezinárodní spolupráce a vzájemného sdílení dat, jako je například směrnice INSPIRE – INfrastructure for SPatial InfoRmation in Europe [6].

Možným řešením, které může výše zmíněné plýtvání omezit, je přístup Linked Data (propojená data, [7]). Jeho hlavní výhodou je jednoduchý mechanismus umožňující typizované propojení dat z různých zdrojů [8]. Toto řešení zcela jistě nepředstavuje univerzální odpověď na všechny současné problémy prostorových dat. Dokáže však díky přístupu, který definuje různé typy provázání jednotlivých datových objektů s jinými prvky jiných datových sad, umožnit snadnější a korektnější propojení původně izolovaných datových sad, včetně možnosti odvozování nových informací. Navíc dodržování principů Linked Data umožňuje vyšší míru automatizovaného zpracování a propojování dat, včetně zlepšení srozumitelnosti, neboť propojená data mají úzkou vazbu na oblast sémantiky (například relace mezi pojmy v datech a slovníky, které tyto termíny definují). Mezi nejvýznamnější sady prostorových dat využívající princip Linked Data patří například GeoNames.org[2], LinkedGeoData.org[3] nebo datové sady produkované Ordnance Survey[4]. Důležitý je i podíl prostorových dat v největších Linked Data datových sadách Wikidata[5] a DBpedia[6] (více o podílu prostorových dat v DBpedia a Wikidata v textu [9]).

Účelem tohoto textu je přispět do diskuze týkající se aktuálních otázek spojených s prostorovými daty a informacemi ve formě propojených dat. Jak již bylo uvedeno výše, hlavním specifikem přístupu Linked Data, který ho odlišuje od tradičních tzv. „plochých dat“ (flat data), jsou především vazby na externí datové položky a slovníky [10]. Pro účely propojování původně izolovaných dat jsou nejdůležitější tzv. identické a podobnostní vazby (často souhrnně označované jako „identity links“; vazby na stejné nebo podobné objekty v jiných datových sadách), které

- mají významnou sociální funkci [11],
- jsou dobře standardizované,
- jsou navázané na sémantiku prvků,
- propojují data (různé datové reprezentace jednoho objektu),

[2] http://www.geonames.org
[3] http://www.linkedgeodata.org
[4] http://data.ordnancesurvey.co.uk/
[5] http://www.wikidata.org
[6] http://dbpedia.org

- umožňují získávat nové informace.

Chceme-li tedy hovořit o kvalitě prostorových dat publikovaných ve formě Linked Data, musíme se nutně zabývat i kvalitou identických a podobnostních vazeb. O tom, že je tento problém aktuální, svědčí i řada článků zaměřených na téma kvalita vazeb v Linked Data, jako jsou například [12], [13], [14] nebo [15]. Výzkum zaměřený na kvalitu prostorových Linked Data podporuje i fakt, že propojená data zatím nenašla výraznou odezvu v komerční sféře, jejich klíčovou doménou jsou především univerzity, výzkumná střediska a nekomerční projekty. Důvodem může být právě problematická kvalita, včetně kvality vazeb. O tom svědčí i výrok z článku [16], kde se uvádí, že samotné publikování dat v cloudu není důvodem pro jejich znovuvyužívání. V rozhodování, zda propojená data budou skutečně využívána, hraje důležitou roli také kvalita dat.

Hlavním cílem této publikace je přispět k diskuzi o využívání identických a podobnostních vazeb mezi objekty propojených dat, především v doméně dat prostorových. Přičemž získané výsledky jsou využity pro odvození pravidel pro využívání výše uvedeného typu vazeb v oblasti prostorových dat a také vytvoření návrhů pro zlepšení struktury a provázanosti prostorových propojených dat.

Použije-li se členění kvality, které ve svém článku nabízí [16], pak se tato práce zabývá především tzv. „trust in content“ (důvěrou v obsah). V jednotlivých fázích jsou ověřovány, testovány, srovnávány a aplikovány na testovací vzorek dat metody (metriky[7]) pocházející z různých vědeckých oborů (například teorie grafů nebo geografie dopravy).

Výzkum kvality identických a podobnostních vazeb prostorových propojených dat je realizován také z důvodů dosažení následujících druhotných cílů:

- Navrhnout nápravu a změny ve struktuře propojených prostorových dat za účelem zlepšení komunikace a eliminace případných chyb prostřednictvím doplnění explicitní sémantiky prostřednictvím vazeb Linked Data.
- Odhalit slabá místa zásadně snižující průchodnost Linked Data grafu (Data Network) pro vybrané prvky.
- Navrhnout prvky (tzv. bridging concepts) výrazně zlepšující průchodnost grafu (například spojující izolované podgrafy).
- Zpopularizovat Linked Data a jejich využívání především v oblasti geomatiky, geoinformatiky, geografie, věd o Zemi a dalších oborech využívajících prostorová data.
- Ukázat sociální funkce Linked Data v oblasti vybraných konceptů z geomatiky a příbuzných disciplín, což může souviset s propagací konkrétního oboru, ujasnění si postavení konkrétního oboru v systému věd, odlišení jednotlivých škol a lokálních zvyklostí, zohlednění pohledu laiků a expertů na jiné oblasti.

[7] Termín metriky pro tento typ metod byl převzatý z publikací [17], [18] a [19].

Kvalita identických a podobnostních vazeb je testovaný na doméně prostorových (geografických, geoprostorových, geo-) propojených dat. Tyto oblasti jsou voleny jako pilotní nebo ilustrační ze čtyř hlavních důvodů:

- Vzhledem ke svému charakteru je výhodnější prostorová data a především vztahy mezi jednotlivými položkami modelovat pomocí grafových struktur (například založených na principu Linked Data) než prostřednictvím tradičních relačních databází, tzv. „flat data“ (například práce o landscape networks [20]).
- Publikace [11] uvádí, že „geografie je další faktor často propojující tematické domény“.
- Jak je patrné z Linking Open Data cloud diagram[8], geografická data a koncepty tvoří velice důležitou složku světa Linked Data.
- Autor se profesně a odborně zaměřuje na obory, jako jsou geomatika a kartografie. Proto budou jeho rozlišovací schopnosti při hodnocení shodnosti a podobnosti prvků vyšší než v jiných vědních oblastech.

Předložený text je rozdělený do osmi základních kapitol. V Úvodu je popsána současná situace na poli propojených prostorových dat a také jsou definovány cíle výzkumu popisovaného v tomto dokumentu. Následují kapitoly představující základní termíny výzkumu. Konkrétně se jedná o problematiku propojených dat, včetně identických vazeb, a grafových struktur pro popis vazeb propojených dat. Čtvrtá kapitola (Rešerše) uvádí výzkum publikovaný v této práci do celosvětového kontextu. V rešerši jsou zmíněny metriky pro kvantitativní analýzu sítí, jejichž výběr je implementován v metodice, která je jádrem následující kapitoly Metodika. Její struktura koresponduje s postupem pro posuzování kvality ekvivalentních vazeb, který se skládá ze tří hlavních procesů – vyhledávání, sběr a formalizace informací o identických vazbách; výběr metrik pro hodnocení vazeb, jejich kompozice a deklarace vhodných parametrů; konkrétní implementace metodiky. Šestá kapitola představuje krátký výčet experimentů, které byly uskutečněné v rámci ověřování metodiky. V následující sedmé kapitole (Výsledky) jsou výsledky experimentů diskutovány, komentovány a zobecňovány. Poslední pasáž textu (Závěr) shrnuje dosažené výsledky.

Propojená data

Propojená data (**Linked Data**), podrobnější informace jsou k dispozici v publikacích [21] a [8], představují jeden z moderních přístupů popisu dat a formalizace informací v oblasti informačních technologií. Jak vyplývá z názvu, hlavním principem je propojování dat. Toto propojování se odehrává na několika úrovních – jednotlivé objekty, sémantická úroveň, případně i v oblasti celých datových sad. Právě kvalitní (ověřené,

[8] http://lod-cloud.net/

standardizované, sémanticky popsané) vazby umožňují jednodušší sdílení a kombinování datových sad a jejich částí.

Linked Data jsou definována spíše na základě výčtu vlastností nebo principů (viz níže), ale v publikaci [10] se objevují dvě téměř totožné věty, které Linked Data označují za „sadu technik pro publikaci a propojování dat na webu s využitím standardních formátů a rozhraní."

Definice výčtem zahrnují často citované Linked Data principy a populární 5-star ranking schéma, obojí publikované v textu Linked Data [7].

Linked Data principy:

1. Používání URI (Uniform Resource Identifier) pro pojmenovávání jednotlivých prvků.
2. Používání HTTP (Hypertext Transfer Protocol) URI, aby názvy byly dohledatelné.
3. Poskytování informací pomocí standardů (RDF /Resource Description Framework/, SPARQL /SPARQL Protocol and RDF Query Language/) těm, kdo vyhledávají URI.
4. Zařazení vazby na jiná URI.

Pětihvězdičkové schéma pro Linked Open Data (5-star ranking scheme) je dalším výčtem vlastností nebo požadavků na propojená data.

1. Data jsou dostupná na webu pod otevřenou licencí.
2. Data jsou dostupná ve strojově čitelném formátu.
3. Data jsou dostupná ve strojově čitelném neproprietárním formátu.
4. Pro identifikaci jsou použité W3C standardy (RDF a SPARQL).
5. Data jsou propojená s ostatními daty především kvůli získání širšího kontextu.[9]

Mezi hlavní přednosti Linked Data přístupu patří podle [10] především kombinovatelnost s dalšími daty za účelem vytváření a získávání nových znalostí. Publikace [16] hovoří o tzv. „follow your nose" navigaci, která umožňuje prohledávání externích zdrojů a o kombinaci dat různého původu. Tento princip je použitý pro sběr informací o identických vazbách i ve výzkumu publikovaném v dalších částech tohoto textu.

Další předností Linked Data, která opět souvisí s vazbami na externí datové zdroje, je podle [10] „samodokumentovatelnost". Jinými slovy producent dat může propojit vlastní data s jinými objekty na webu, které uživateli poskytnou informace o těchto datech (například typ objektu, definice, popis, odkaz na originální zdroj apod.). Tato vlastnost je velice důležitá z hlediska komunikace, sdílení a kombinování jednotlivých

[9]Poznámka autora: Je třeba si uvědomit, že schéma je kumulativní. To znamená, že bez splnění předchozí podmínky není možné splnit následující.

datových sad a jejich prvků. Podobně i článek [22] také uvádí kombinovatelnost dat jako hlavní přednost Linked Data přístupu, ale zdůrazňuje ji především v souvislosti možností vývoje nových aplikací.

Jak vyplývá z přívlastku „linked“ („propojený“), klíčovou složku Linked Data přístupu tvoří vazby (linky, relace, vztahy), které uživateli umožňují provázání dat s daty a informacemi v jiných zdrojích. Podle [19] jsou vazby důležité z hlediska obohacení sémantiky, poukazování na nové zdroje informace a propojování datových sad.

Autoři publikace [11] rozlišují tři základní typy vazeb:

1. Relationship Links propojují příbuzné prvky v různých datových sadách (například knihu a jejího autora nebo město a významné rodáky). Tyto informace slouží především k začlenění vlastních dat širšího kontextu a k tomu získat další doplňující informace. Do této kategorie se řadí i topologické vazby, které jsou charakteristické pro prostorová data, a tzv. meronymické vazby (propojení části a celku).
2. Identity Links[10] umožňují identifikovat stejné nebo podobné objekty. Tento typ vlastností má podle [11] důležitou sociální funkci (Web of Data jako sociální systém), protože umožňuje zařadit do Web of Data různé pohledy na svět. Zdroj [23] charakterizuje tento typ vazeb jako linky, které definují, že dvě věci jsou identické nebo velmi podobné.
3. Vocabulary Links směřují od dat k slovníkovým položkám, které data popisují, charakterizují nebo definují. Právě tyto vazby doplňují sémantiku k datům a způsobují, že Linked Data jsou označována jako samoopisná (samodokumentovatelná), přičemž tento fakt umožňuje lidem i strojům data kvalitně zpracovávat a kombinovat.

Jak již bylo uvedeno výše, předmětem tohoto výzkumu jsou především vazby typu Identity Links, které umožňují pojímat Linked Data jako svébytný sociální systém. Hlavní důvody pro tuto paralelu popisuje článek [11]:

- Názorová různorodost: URI umožňují diferencovat popisy stejných fenoménů a tak vyjadřují různé pohledy poskytovatelů dat.
- Dohledatelnost: Používání různých URI umožňuje uživateli zjistit jednotlivé dílčí pohledy autorů dat.
- Neexistence jednoho bodu, jehož chyba by vedla ke kolapsu celého systému: Pokud by pro každý objekt na světě existovalo jediné URI, byl by celý systém velice

[10] V tomto textu budou označované jako identické vazby, ekvivalentní vazby, případně jako identické a podobnostní vazby. Používání předchozích výrazů jako synonym je dáno především faktem, že existující standardy identity links disponují různou přesností a striktností popisu vazby [23], takže v mnoha případech nepropojují pouze totožné prvky, ale i objekty s větší či menší mírou podobnosti. Hodnocení úrovně podobnosti geografických objektů a konceptů je však mimo rámec této práce. Zájemce o tuto problematiku se může seznámit s články [24], [25] nebo [26].

náchylný k celkovému kolapsu, nemluvě o vysokých nákladech na koordinaci, administrativu a byrokracii.

Grafové struktury pro popis vazeb propojených dat

Z důvodu přehlednějšího popisu, prezentace a možnosti vyhodnocení pomocí existujících metod jsou pro znázornění vazeb mezi jednotlivými instancemi geografických prvků použity grafy. Zmínku o grafech znázorňujících `owl:sameAs` vazbu má ve svém článku [22]. Tyto grafy jsou však neorientované a předpokládá se u nich symetričnost vazby, která však reálně neexistuje (viz [14]). Z těchto důvodů nejsou vhodné pro účely vyhodnocování a popisu kvality vazeb.

Publikace [19] uvádí tzv. **datovou síť** (Data Network). Ta je definovaná[11] jako orientovaný, označený graf $G = (V, E, L)$, kde V je množina uzlů, E množina hran a L množina popisků. Hrana $e_{ij} \in E$ propojuje uzly $v_i \in G$ a $v_j \in G$. K hraně $e_{ij} \in E$ je přiřazený popisek $l_{ij} \in L$. Hrany a popisky korespondují s predikáty RDF trojic, zatímco uzly reprezentují objekty a subjekty.

Článek [15] uvádí podobnou síť (založenou pouze na vazbách `owl:sameAs`) a nazývá ji **SameAs Network**, včetně definic[12]. Pomocí grafových struktur vyjadřují vazby mezi propojenými daty (nikoli pouze identické a podobnostní) také [27]. V tomto případě je graf označován jako Linked data graph. Další vyjádření propojení dat pomocí orientovaného grafu je k dispozici například v [28]. V tomto případě se však nejedná o propojená data ve smyslu Linked Data. Podle [15] zpracování identických vazeb v propojených datech ve formě grafových struktur poskytne odpovědi na otázky týkající se počtu, rozmístění a topologie těchto vazeb a zdrojů, které jsou jimi propojeny.

Přehled metrik

Rešerše je věnována výčtu dílčích metrik, které umožňují exaktně popisovat vlastnosti grafových struktur, které reprezentují identické vazby propojených dat. Metriky pro kvantitativní hodnocení prvků grafů nebo sítí nejsou pouze doménou matematiky, resp. teorie grafů. Provedená rešerše (celkově bylo nalezeno téměř 70 metrik, z nichž většina je

[11] Definice byla oproti původnímu článku autorem upravena a mírně rozšířena. To platí i pro další definice z tohoto zdroje.

[12] Definice víceméně představují lehce modifikovaná klasická tvrzení z teorie grafů a RDF, proto nejsou v této práci uváděny, ale pouze odkazovány.

uvedena v následujícím textu[13]) ukazuje řadu vědních disciplín a oborů lidské činnosti, kde tyto postupy nalézají své uplatnění. Jako příklady lze jmenovat sociologii [35], informační vědy [36], literaturu [37], biologii [32], medicínu [38], zdravotní péči [39], sport [40], [41] a především v současnosti velmi moderní sociální sítě [29], [34], [42]–[47].

Nejjednoduššími metrikami (tzv. strukturálními vlastnostmi sítí) jsou **počty základních komponent grafu**, tedy uzlů (size, network size) a hran (ties, edges). S těmito hodnotami pracují například analýzy [39], [46], [48], [49]. Počty se mohou vztahovat i na specifické typy uzlů – izolované, kořenové, listové a vnitřní vrcholy. Jedná se však o absolutní hodnoty, proto nejsou tak často používány, protože neumožňují efektivní srovnání více grafů.

Existují však metriky, které pracují s prvky grafu a přitom umožňují srovnání. Jedná se především o **hustotu** (density) [31], [36], [39], [41], [46], [48], [49], která je vyjádřena jako podíl počtu vazeb a maximálního počtu vazeb v grafu, který je vyjádřen zpravidla v procentech.

Další metrikou může být **dosažitelnost** (reachability). Ta je v [29] popisována jako průměrný počet propojení mezi dvěma prvky grafu.

Stupeň uzlu je základní metodou teorie grafů, která, jak vyplývá z názvu, popisuje především vrcholy grafu a jejich propojení. Tato metrika se především pro svoji jednoduchost používá v mnoha textech zaměřených na vyhodnocování grafových struktur [39], [42], [46], [49]–[52], včetně sítí v prostředí webu i propojených dat, jako například [19], [53].

Publikace [54] definuje stupeň uzlu v grafu G jako „počet hran grafu G, které obsahují vrchol v. Značí se $d_G(v)$.“ Stupeň uzlu tedy udává míru přímé propojenosti uzlu s dalšími vrcholy, což je v případě grafu znázorňujících identické a podobnostní vlastnosti klíčová informace při hodnocení provázanosti jednotlivých zdrojů dat a pojmů. Vzhledem k tomu, že počet uzlů je v každém grafu různý, je při srovnání více grafů důležité procentuální vyjádření a rozdíl skutečné hodnoty stupně uzlu od hodnoty maximálně možné, která je daná výrazem $n - 1$, kdy n je počet vrcholů grafu.

V případě propojených dat je potřeba pracovat s orientovanými grafy. Proto kromě celkového stupně uzlu uvažujeme také **vstupní a výstupní stupně**, které jsou dány počtem hran, které do vrcholu vstupují d_G^- a které z něj vycházejí d_G^+. Na základě hodnoty vstupních a výstupních stupňů uzlu je možné vrchol označit jako

[13]Do textu nakonec nebyly začleněny některé metriky, které nemohly najít uplatnění v popisovaném výzkumu. Jedná se například o metody pracující s ohodnocenými hranami (tie strength, intensity) publikované v [29]–[34] nebo metody využívající aktivní chování prvků sítě [29] nebo velmi specifické biologické metriky [32].

- izolovaný (isolated) – z něj ani do něj nevede žádná vazba (hodnoty obou stupňů jsou nula)[14],
- kořenový (source) – hodnota vstupního stupně uzlu je nula, zatímco hodnota výstupního stupně uzlu je vyšší než nula,
- listový (sink) – uzel není izolovaný, ale veškeré vazby vedou směrem do uzlu (hodnota vstupního stupně uzlu je vyšší než nula),
- vnitřní (internal) – hodnoty obou stupňů jsou vyšší než nula.

Podle [19] by cílem sledování stupňů uzlů a jejich následné modifikace (posilování, doplňování vazeb) mělo být přiblížení se k tzv. bezškálovým sítím. Tyto sítě představují ideální variantu, která je díky existenci hubů mnohem odolnější (robustnější) vůči vlivům náhodných chyb.

Dalším kritériem pro hodnocení ekvivalentních a podobnostních vazeb propojených dat je **souvislost grafu** (connectivity) [32]. Pokud bude graf souvislý, pak bude existovat propojení mezi všemi uzly, a tudíž bude možné propojit veškeré informace o každém konceptu.

Existují metriky, které jsou založené na kvantifikaci propojení jednotlivých uzlů. Základní metodou je výpočet **vzdáleností** (path length) $l(G)$ grafu G. Tato veličina je použita v síťových analýzách publikovaných například v [38], [46], [52], [56].

Inverzní hodnota vzdálenosti $l(G)$ se označuje jako **Global Efficiency** [38], [52], [57]. Global efficiency podává informaci o blízkosti uzlu k ostatním vrcholům sítě. Zdroje [38] a [52] popisují ještě tzv. **Local Efficiency**. Ta udává propojenost uzlů v dílčích částech grafu.

Se vzdáleností souvisí také další metrika označovaná jako **průměr grafu** (diameter). Ta je definovaná jako maximální vzdálenost mezi libovolnými uzly v grafu. Jako kritérium v síťových analýzách je použita například ve zdrojích [40] nebo [46].

Další důležitou metrikou je **centralita**. Podle Guereta [58], [19] „centralita indikuje kritickou pozici uzlu v topologii“. Proto tento typ metriky může být použitý pro identifikaci vhodného zdroje sémantických informací pro daný typ objektu. Zdroje [29]–[31], [33], [35], [38], [39], [56], [59]–[61] se zabývají centralitou obecně, včetně historie výzkumu centrality na úrovni grafů a sítí, nebo její aplikací na různých doménách. Využívání grafových struktur a příslušných metrik v oblasti propojených dat, konkrétně v rámci tzv. „recommender systems“ (poskytování informací uživateli, přičemž tyto informace jsou odvozeny na základě jeho předchozího chování), jsou popsány v článku [62]. Rolí centrality (jako jedné z metrik) na poli prostorových dat se zabývají například publikace [63] a [64].

V publikacích jsou nejčastěji zmiňovány tyto základní typy centrality [60], které budou

[14] Izolované uzly jsou spojené s dalším parametrem – **existencí sousedních uzlů** [55].

využity při tvorbě metodiky v rámci této práce:

- **centralita stupně** nebo centralita měřená stupněm uzlu (degree centrality),
- **centralita blízkosti** nebo centralita měřená blízkostí polohy ve středu sítě (closeness centrality),
- **centralita mezilehlosti** nebo centralita měřená středovou mezipolohou (betweenness centrality)

Kromě těchto centralit existují ještě další méně často využívané druhy, jako například Eigenvector centrality [34], [48], [49], [55], Barycenter Centrality [55], Information centrality [48], [49], Katz centrality [65], [66] nebo Reachability centrality [48], [49].

Centralita stupně C_d [31], [34], [35], [37], [44], [48], [60], [67]–[69] se určuje jako stupeň uzlu. V případě orientovaných grafů lze hovořit o tzv. outdegree C_{od} a indegree C_{id} centrality [39], [37].

Práce [70] uvádí, že „vrchol s vysokým počtem hran nebo více spojeními je ve struktuře grafu více centrální a má tak větší schopnost ovlivňovat ostatní. Vrchol, na který vede mnoho hran, lze označit za prominentní, přední či populární vrchol. Vrchol, ze kterého vede mnoho hran, lze naopak označit za vlivný vrchol – má vyšší šanci ovlivnit ostatní.“ Centralita stupně je využívána například v tzv. koeficientu efektivity (efficiency coefficient), který byl navržený Burtem [71] a použitý v analýzách publikovaných v [33] nebo [34].

Centralita blízkosti [31], [34], [35], [44], [47]–[49], [55], [60], [67] je definována jako průměrná nejkratší cesta mezi uzlem v a ostatními uzly grafu G. Nejvyššího hodnoty tohoto typu centrality signalizují, že uzel je dobře dostupný ze všech částí grafu. Z hlediska uzlů jako zdrojů sémantických propojených prostorových dat je vysoká hodnota centrality blízkosti důležitá z hlediska rychlého procházení datové sítě při získávání nových informací z reprezentací stejného geografického objektu. Podle [70] centralita blízkosti představuje „míru toho, jak dlouho bude trvat, než se informace rozšíří z daného vrcholu do všech ostatních vrcholů grafu.“

$$C_c(v) = \frac{1}{\sum_y d(y, v)}$$

kde $d(y, v)$ je délka nejkratší cesty mezi uzly y a v v grafu G.

Centralita mezilehlosti [19], [34], [35], [39], [44], [47], [49], [55], [58], [60], [67], [72] hodnotí uzly z hlediska „mezilehlosti“, tj. specifické polohy, kdy daným uzlem prochází velké množství cest mezi ostatními uzly [59]. Vysoké hodnoty centrality mezilehlosti indikují, že daný uzel tvoří „most“ uvnitř grafu, to znamená, že propojují do jisté míry samostatné nebo izolované podgrafy.

$$C_b(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

kde σ_{st} je celkový počet nejkratších cest v grafu G z uzlu s do uzlu t a $\sigma_{st}(v)$ je počet takových cest procházejících uzlem v.

Pro výpočet centrality mezilehlosti udává [19] zjednodušený vztah – poměr mezi počtem uzlů, které jsou přímými sousedy testovaného uzlu a jejich hrany jsou označeny vzhledem k testovanému uzlu jako vstupní, a počtem uzlů, které jsou také přímými sousedy testovaného uzlu a jejich hrany jsou označeny vzhledem k testovanému uzlu jako výstupní.

Publikace [58] a [19] považují za důležitou pro určování kvality vazeb především centralitu mezilehlosti, protože uzly s její nejvyšší hodnotou (tzv. bridges nebo brokers) bezprostředně ovlivňují tok informací v síti. Důležité je uvědomit si, že „most“ představují zároveň také „úzká hrdla“, tedy místa, jejichž chyba ovlivní prostupnost celé sítě. Změny sítě vyvolané výsledky vyhodnocení tohoto typu centrality mají za následek především snížení rozdílů v centralitě hubů sítě, čímž klesne náchylnost celé sítě k chybám cíleným na její klíčové prvky (huby).

Autority (authorities) a **středy** (hubs) [37], [44], [47], [55] představují dva specifické typy vrcholů grafu. Autorita je takový uzel, k němuž směřuje velké množství vazeb. Naopak střed je vrcholem, z něhož vychází velké množství propojení směrem ke zbytku sítě. Oba pojmy jsou úzce spojené s algoritmem HITS (Hyperlink-Induced Topic Search) [73]. Tento iterativní algoritmus umožňuje výpočet tzv. authority score $A(v)$ and hub score $H(v)$

PageRank je iterační algoritmus (navržený Larry Pagem a Sergeyem Brinem) počítající významnost jednotlivých vrcholů v síti. Základní princip algoritmu PageRank [74] spočívá v tom, že pro každý uzel se zjišťuje počet a také významnost (z pohledu PageRank) vrcholů, které jsou s ním přímo spojeny, kde je velikost kruhů vyjadřujících uzly přímo úměrná hodnotě koeficientu Page Rank). Pro síťové analýzy je tento algoritmus používán jako metrika například v publikacích [44], [45], [47], [55].

Se zjišťováním významu vrcholů souvisí i další méně rozšířené metriky publikované v [55]:

- **Important Neighbor Proportion** – procento významných uzlů mezi přímými sousedy.
- **Unknown Neighbour Proportion** – sousední vrcholy, u nichž nebyla zjištěna významnost (hodnota se vyjadřuje v procentech).
- **Shortest Distance to Known Important Classes** – nejkratší vzdálenost ke kterémukoli významnému vrcholu.

Shlukový koeficient (koeficient shlukování, Clustering coefficient) je publikován například v dokumentech [19], [31], [39], [40], [46], [52], [53], [75]–[77]. Tento lokální koeficient popisuje uzavřenost částí grafu a hustotu sítě kolem konkrétního uzlu. Podle publikací [75] i [19] se koeficient pro jednotlivé uzly grafu počítá jako poměr vazby mezi sousedy uzlu a maximálním možným počtem vazeb mezi těmito sousedy.

Celkový shlukový koeficient grafu se vypočítá jako průměrný koeficient všech uzlů grafu [46]. Podle [19] by případné změny v grafu (datové síti) měly zlepšit sdružování uzlů do lokálních skupin a zkrátit průměrnou cestu mezi takovými skupinami (podporovat tzv. small world network – viz publikace [78] nebo [79]. Ideální hodnota koeficientu je rovna 1. Této hodnoty je dosaženo, když je spojený každý uzel s každým.

Se shluky (klastry) souvisí také metrika nazývaná **Number of Connected Components** [46]. Ta představuje počet zřetelných klastrů obsažených v grafu.

Sítě jednotlivého aktéra (tzv. ego networks) [80]–[82] představují podgrafy vybraného uzlu, jeho přímých sousedů a jejich vzájemných propojení. Tento typ grafové struktury se používá především v oblasti sociálních sítí. Podle publikací [48] a [49] se pro analýzy takových sítí používají kromě tradičních metrik (jako například počet uzlů, počet hran nebo hustota sítě) následující metody (a jejich normalizované varianty):

- **EgoBetween** – procento nejkratších cest mezi uzly, které prochází egem (centrálním vrcholem).
- **TwoStepReach** – procento uzlů, které jsou z ega dosažitelné přes dva kroky.
- **WeakComp** – počet skupiny vzájemně propojených uzlů, které jsou mezi sebou propojeny pouze přes ego.
- **Brokerage** – počet uzlů, které mezi sebou nejsou přímo propojené, cesta mezi nimi vede přes ego. Podle publikace [31] je tento parametr úzce spojený s centralitou mezilehlosti.

Reciprocita vyjadřuje míru vzájemného propojení mezi dvojicemi uzlů. Vyjadřuje se zpravidla jako poměr mezi počtem vzájemně propojených uzlů a všech propojených uzlů [83]. Další možnosti výpočtu reciprocity jsou k dispozici na příklad v publikaci [84].

Jednou z dalších disciplín, kde je možné hledat systémy pro hodnocení kvality sítě (a tedy i vazeb v Linked Data), je geografie dopravy. Tato část geografie řeší síťové analýzy a dostupnost jednotlivých sídel (uzlů) v rámci komunikační sítě.

- **Beta index** – představuje nejjednodušší metodu hodnocení dopravní sítě. Jedná se o podíl počtu hran (e) a počtu uzlů (v).
- **Gama index** – jedná se o poměr mezi skutečným počtem spojení (hran) a maximálně možným počtem spojení. Hodnota se pohybuje mezi 0 a 1 a podle [85]

indikuje kompletnost sítě. Podle stejného zdroje je tento typ indexu vhodný pro sledování časových změn v síti.

- **Alfa index** – je podobný svojí konstrukcí Gama indexu, ale na rozdíl od počtu hran se v tomto případě pracuje s cykly grafu. Jedná se tedy o podíl skutečného počtu cyklů v grafu a maximálně možného počtu cyklů.
- **Eta index** – má význam pouze v případě ohodnoceného grafu, protože se vypočítá jako podíl celkové délky hran grafu a počtu hran.

Metodika

Studium kvality identických vazeb prostorových propojených dat lze nahlížet dvěma základními způsoby. V první řadě je možné hodnotit jednotlivé geografické koncepty, geografické objekty nebo jejich skupiny na základě toho, v jaké míře jsou zapojené do sítě propojených dat. Tedy, zda se vyskytují převážně izolovaně nebo jsou-li mezi sebou propojeny pomocí nějakého typu relace (v tomto případě identické vazby). Z tohoto hlediska je možné sledovat nejen kvantitativní údaje, ale vyhledávat i prostorové vzorce ukazující vztahy mezi určitými typy objektů nebo sémantických zdrojů. Takové prostorové vzorce mohou sloužit například ke zjišťování šíření informací (v tomto případě o tom, které zdroje propojených dat ze sebe navzájem čerpají informace) v prostoru propojených dat a identifikace nejvhodnějšího zdroje, který je vhodné využívat a případně také ovlivňovat jeho obsah za účelem zlepšení kvality dat v daném oboru.

Druhou možností je hodnotit kvalitu vazeb. Zda prvky vazby – objekt, subjekt a predikát – jsou ve vzájemném souladu (především z pohledu sémantiky) a zároveň odpovídají významu relace. V tomto případě hodnocení spočívá v klasifikaci vazeb na základě typologie (identické a podobnostní, případně podle jednotlivých standardů a především definic v nich uvedených, které popisují sílu a striktnost každé vazby). Další možností je hodnocení míry korespondence významu vazby a příslušných objektů a subjektů. Podobně jako v předchozích případech lze porovnávat jednotlivé typy vazeb mezi sebou. Tento způsob hodnocení není v případě identických vazeb vhodný. Měl by význam například pro porovnávání používání hierarchických a mereologických vlastností (vazeb typu „je speciálním případem“ a „je částí“).

Výzkum publikovaný v této práci se soustředí na oba způsoby hodnocení. Při sběru informací o identických vazbách jsou v první řadě identifikovány nekvalitní nebo chybné relace. Poté následuje nasazení metrik ukazující míru propojení geografických objektů nebo konceptů.

Sběr informací o identických vazbách probíhá v základních jednoduchých krocích, které spočívají v postupném procházení sítě propojených dat a porovnávání dvojic výskytů

reprezentací (instancí) stejného geografického objektu nebo jevu ve dvou různých sadách Linked Data. Výsledky jsou následně ukládány a analyzovány v grafech, které vyjadřují všechny dostupné výskyty reprezentací stejného geografického objektu nebo konceptu. Uzly grafu tvoří jednotlivé instance (resp. zdroje obsahující danou instanci) a hrany představují identické vazby.

V další fázi je možné takové hodnocení jednotlivých prvků propojených dat porovnávat (na základě statistického vyhodnocení různých typů chyb a výpočtů hodnot, které vyjadřují úroveň konektivity celé sítě – jednotlivé metody jsou uvedeny dále v této kapitole v částech věnovaných kvantitativnímu popisu sítě vytvořené identickými vztahy a výskyty reprezentace jednoho objektu v různých datových sadách) pro skupiny entit s podobnými vlastnostmi. Závěry této části výzkumu jsou pak zaměřeny na zobecnění výsledků evaluace, případně na identifikaci lokálních specifik, včetně odlišných národních nebo regionálních přístupů týkajících se terminologie, klasifikace nebo způsobu popisu jednotlivých prvků.

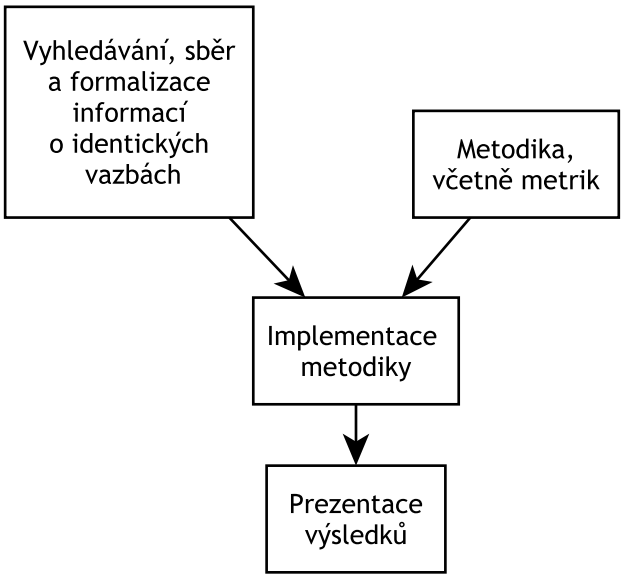
Jak již bylo uvedeno výše, podobně jako geografické koncepty nebo objekty lze testovat i zdroje poskytující jejich reprezentace v prostředí Linked Data. Na této úrovni je možné hodnotit především množství prvků z dané oblasti, které se vyskytují v jednotlivých datových a znalostních bázích, kvalitu poskytovaných informací (přičemž je nutné si uvědomit, že takové hodnocení by mělo být zajišťované především experty na dané domény, a tudíž je velmi obtížná automatizace) a vazby zdrojů propojených dat mezi sebou. Pokud budou srovnávány datové báze mezi sebou, pak výsledky mohou představovat především topologické (prostorové) vzory v datových sítích. Tyto informace lze následně využít pro zkvalitnění celé sítě propojených dat (vložení nových vazeb nebo uzlů).

Navržení způsobu hodnocení identických vazeb mezi objekty prostorových propojených dat je založené na čtyřech základních krocích (Obrázek 1), který ukazuje architekturu navrhovaného řešení, jež je částečně inspirovaná pojetím v publikaci [19]. V článku [27] je publikovaný také velmi podobný postup skládající se z následujících kroků – extrakce dat, tvorba grafu, výpočet míry zkreslení informací, kvantifikace a komparace výsledků.

Data a informace o identických vazbách propojených prostorových dat jsou sbírána především pro účely testování, ověřování správnosti a ilustrace způsobu použití i významu jednotlivých metrik a celé metodiky. Sběr dat může probíhat třemi základními způsoby:

1. Manuálně (procházením známých zdrojů sémantických dat, které jsou propojené v prostoru Linked Data)
2. Automaticky s využitím existujících služeb (v angličtině označovaných jako harvestor nebo crawler) – využívána byla především služba sameAs.org[15]

[15]Dále byly na základě informací ze zdroje [86] zkoumány služby a nástroje Sig.Ma, rkbexplorer nebo ObjectCoref. Dále byly bez většího úspěchu (funkčnost nebyla dostačující pro potřeby tohoto



Obrázek 1: Architektura procesu hodnocení identických vazeb prostorových propojených dat.

3. Automaticky pomocí skriptu vytvořeného v rámci této práce

Případné chyby, které ovlivňují vazbu mezi instancemi reprezentujícími geografické objekty, lze rozdělit na chyby

syntaktické Chyba má podobně jako v následujícím případu za následek kompletní nefunkčnost celé RDF trojice. Chyba se týká většinou zápisu predikátu nebo identifikátoru objektu, vznikla však, podobně jako v dalších typech chyb, na straně subjektu při zápisu identické vazby a jejího cíle. Může se jednat například o drobnou vadu (například záměnu písmen) v názvu vazby nebo identifikátoru objektu.

technické Chyba se může projevit nefunkčností (nedostupností) objektu. Jinými slovy se nepodařilo získat cíl vazby. Mezi technické chyby je možné řadit situace, kdy datová sada nebo konkrétní datový objekt není dostupný, funkční (ať už zcela nebo dočasně), byl zcela odstraněný (neexistuje) nebo došlo ke změně identifikátoru.

sémantické Chyba má za následek nedostatečné (nekorektní) fungování vazby. Nalezené výsledky mohou být například nepřesné nebo nejsou jednoznačné. Mezi sémantické nedostatky identických vazeb patří například následující případy:

- Vazba neodpovídá skutečnému vztahu objektu a subjektu.
- Objekt RDF trojice (cíl vazby) neodpovídá subjektu a vazbě.

výzkumu) testovány nástroje jako Silk, LDspider [87], Any23 nebo Falcons.

- Není dostatek informací a vazbě (predikátu), protože se jedná o nestandardizované (ani explicitně nepopsané, případně uzavřené) řešení.
- Absence vazby – dva objekty jsou sice identické nebo velmi podobné, ale nejsou spojeny příslušnou vazbou.

Vyhledávání, sběr a formalizace informací o identických vazbách probíhá podle následujícího postupu (Obrázek 2), který je zpracován ve formě série skriptů v jazycích Bash (Bourne again shell) s využitím prostředí UNIX shell (řídící proces celého postupu), XSLT (Extensible Stylesheet Language – Transformations) a programovacího jazyka R. Na obrázku 2 je znázorněna aktivita, používaný software nebo aplikace (v závorce pod každým krokem) a výstupní formát, resp. formát předávaných dat (v rámečku mezi jednotlivými kroky). Podobně jsou koncipována schémata i pro další části metodiky.

Metodika pro hodnocení identických vazeb mezi objekty prostorových propojených dat (Linked Data) (Obrázek 3) vzniká propojením dílčích metrik pro popis a hodnocení orientovaných grafů. Kromě metrik je při tvorbě a nasazení metodiky klíčové stanovení příslušných kvantitativních kritérií.

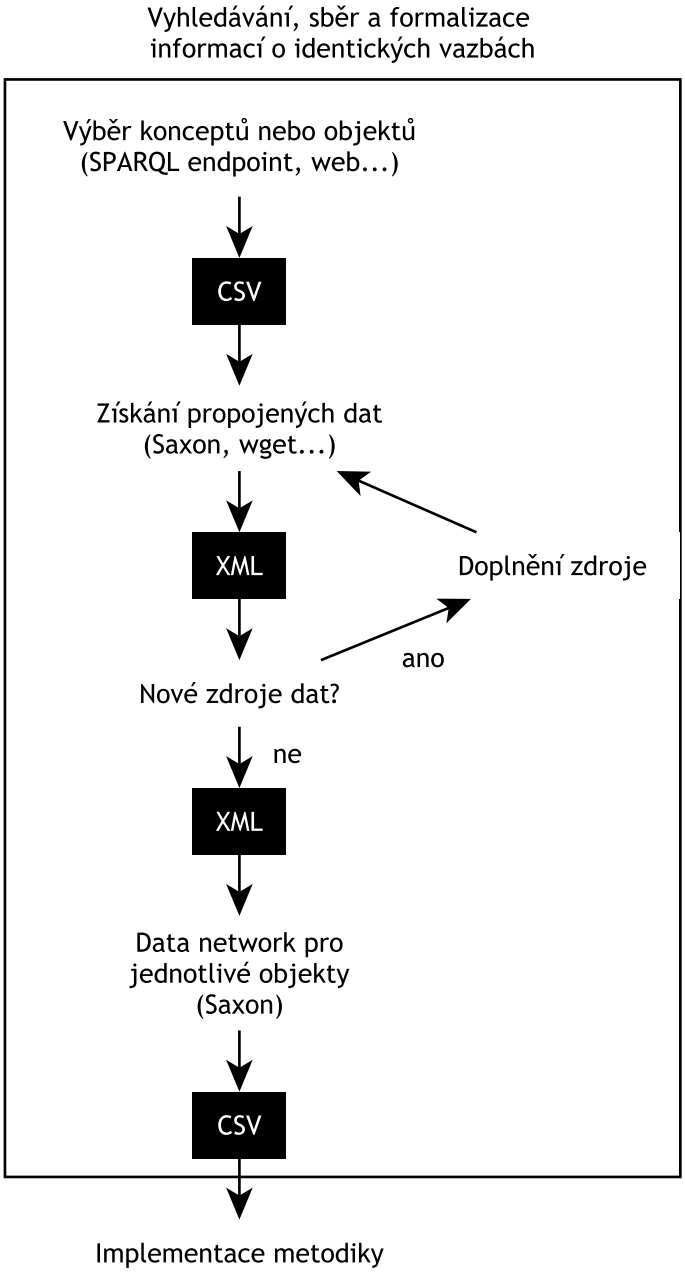
Metriky pro hodnocení identických vazeb pro prostorová propojená data představují postupy, kterými lze exaktně ohodnotit grafové struktury, které reprezentují relace mezi jednotlivými výskyty jednoho objektu v různých datových sadách. Metriky lze rozdělit do dvou skupin:

1. Metriky určené pro hodnocení uzlů v grafů, které jsou v publikovaných případových studiích určeny především pro určení kvality jednotlivých zdrojů sémantických dat.
2. Metriky pro hodnocení sítě, které udávají stav provázanosti konkrétního pojmu ve zdrojích Linked Data.

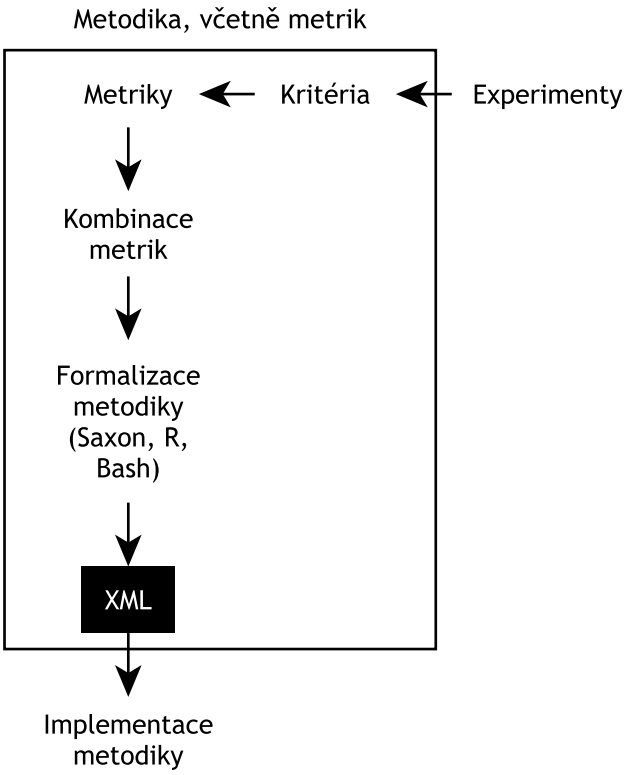
Hodnocení uzlu v rámci jednoho grafu poskytuje informace o zdroji propojených dat, který je uzlem reprezentovaný. Prvním požadavkem na uzel je samozřejmě nulový výskyt chyb – uzel nebude po eliminaci vazeb obsahujících sémantické nebo technické chyby izolovaný.

Optimální uzel v rámci grafu by měl mít následující vlastnosti, které souvisí s významnou pozicí vrcholu v rámci grafu. Bude-li mít uzel takové postavení v rámci grafu, pak je možné o zdroji s velkou pravděpodobností prohlásit, že je populární, často odkazovaný a/nebo poskytující odkazy, což do jisté míry může svědčit o jeho kvalitě ve smyslu poskytování dat a informací. Optimální pozici lze popsat následujícími větami:

1. Uzel je propojený na velké množství ostatních uzlů.
2. Uzel je dosažitelný z mnoha uzlů.
3. Uzel leží blízko ostatních uzlů (z hlediska délky nejkratší cesty).
4. Uzel propojuje nezávislé podgrafy sítě.



Obrázek 2: Architektura – vyhledávání, sběr a formalizace informací o identických vazbách.



Obrázek 3: Architektura – metodika, včetně metrik.

Předchozí seznam lze snadno vyjádřit pomocí vybraných metrik:

- Centralita stupně[16] – zdroj je pomocí identické vazby propojený na mnoho dalších datových sad.
- Centralita blízkosti[17] – zdroj leží blízko ostatních uzlů (z hlediska délky nejkratší cesty).
- Centralita mezilehlosti[18] – zdroj propojuje nezávislé části Linked Data prostoru.
- Autorita[19] – zdroj je pomocí identických relací dosažitelný z mnoha uzlů (dalších zdrojů propojených dat).
- Hub[20] – zdroj poskytuje propojení na další zdroje.
- Page Rank – zdroj je spojený s velkým množstvím kvalitních uzlů, kde tato kvalita je daná především mírou propojenosti těchto vrcholů.

Hodnocení postavení prvku (konceptu nebo datové položky) v síti propojených dat je možné zjišťovat pomocí metrik, které detekují parametry celého grafu (a nikoli jednotlivých uzlů jako v předchozím případě). Implementovaný graf (typu sameAs Network) jako celek ukazuje zdroje, které obsahují reprezentaci prvku, a identické vazby mezi těmito reprezentacemi. Ideální objekt z hlediska identických vazeb propojených dat se dá popsat následujícími výrazy:

1. Reprezentace objektu v jednotlivých zdrojích Linked Data by měly obsahovat minimální množství chyb.
2. Reprezentace objektu by měly být součástí vysokého počtu zdrojů.
3. Reprezentace objektu by měly být hustě propojeny pomocí identických vazeb.
4. Propojení jednotlivých reprezentací by mělo být maximálně homogenní.

Podobně jako v části věnované hodnocení uzlů grafu, i v tomto případě je možné k předchozím větám, které vyjadřují vlastnosti ideálních prvků Linked Data sítě, přiřadit konkrétní metriky, které umožňují kvantifikaci, a tudíž i možnost srovnání.

1. Chybové prvky jsou eliminovány ve fázi sběru informací o identických vazbách.
2. Počet uzlů v grafu.
3. V tomto případě jsou vhodné dvě metriky
   (a) hustota sítě (míra propojení jednotlivých reprezentací[21]),

[16] Tato metrika má stejný význam jako stupeň uzlu.

[17] Tato metrika zastupuje celkovou vzdálenost a metodu Global Efficiency, které mají podobný význam.

[18] Centralita mezilehlosti má podobný význam jako metoda EgoBetween.

[19] Pro zjednodušení může být nahrazena vstupním stupněm uzlu, případně odpovídajícím typem centrality stupně.

[20] Pro zjednodušení může být nahrazena výstupním stupněm uzlu, případně odpovídajícím typem centrality stupně.

[21] Tato veličina může být nahrazena hodnotou podobné metriky Gama index.

   (b) reciprocita (do jaké míry existují vzájemné vazby mezi reprezentacemi objektu),
4. Shlukový koeficient – koeficient deklarující existenci nezávislých komponent grafu, které snižují jeho homogenitu.

Implementace (Obrázek 4) zahrnuje výběr a aplikaci (vzájemné propojení a stanovení potřebných kritérií) metrik uvedených v předchozím kroku na data o identických vazbách získaná v prvním kroku metodiky. Do implementační fáze data o vazbách vstupují ve formě grafových struktur, kdy jeden graf odpovídá jednomu datovému objektu.

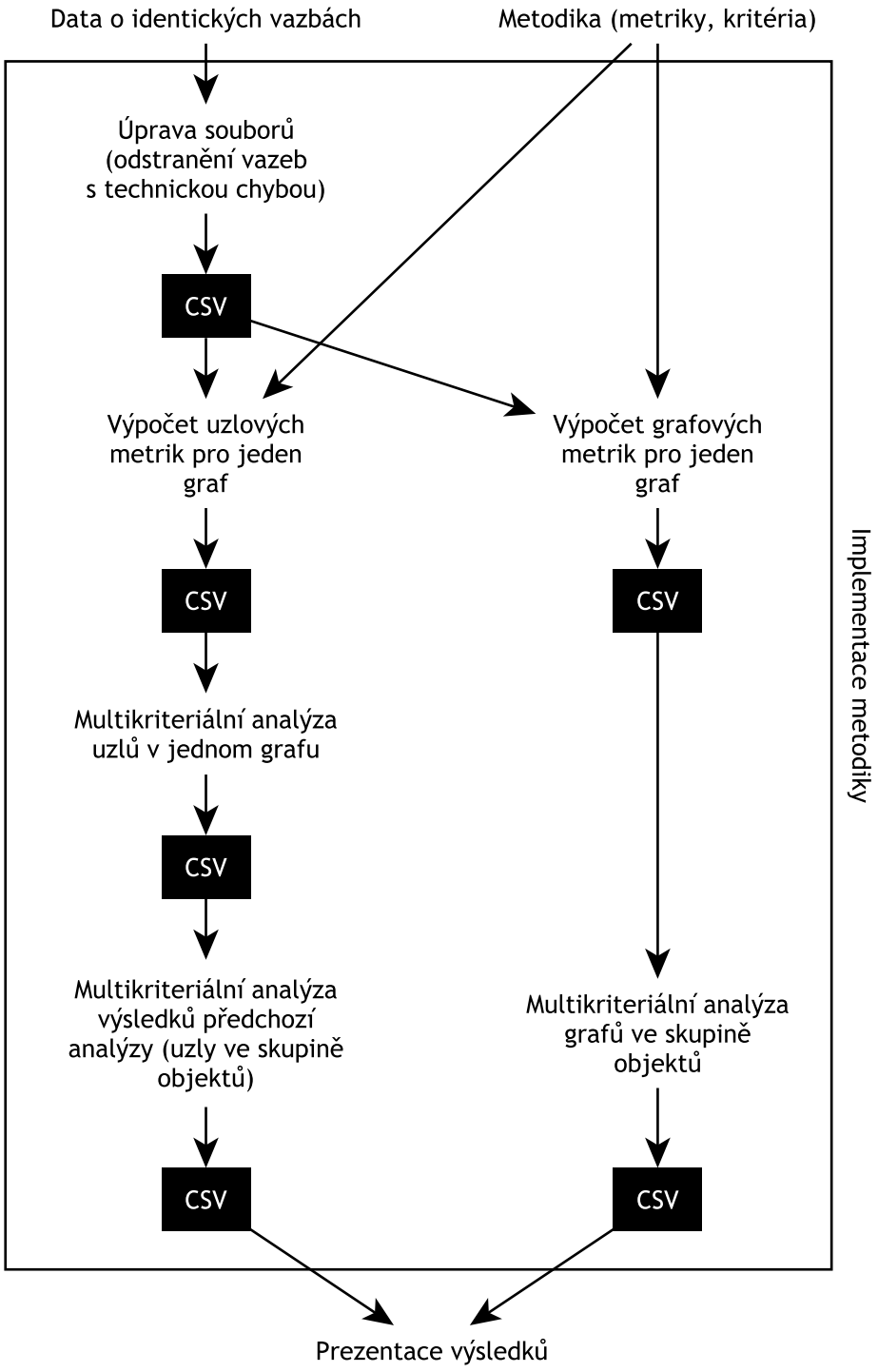
Před aplikací metrik jsou vstupní data nejprve upravena. Modifikace dat spočívá v odstranění vazeb, ve kterých je identifikována technická chyba (například data neexistují nebo nejsou dostupná) a dále odstranění takto vzniklých izolovaných uzlů, které nevstupují do následující analýzy.

Souhrnné hodnoty, v případě uzlových metrik pro celý graf, v případě grafových metrik pro skupinu objektů, jsou zjišťovány pomocí multikriteriální (vícekriteriální) analýzy [88]. Ta spočívá ve výběru nejlepší alternativy pomocí definice několika kritérií (v případě této práce se jedná o metriky), jejich vah (na počátku byla pro všechna kritéria stanovená hodnota 1) a způsobů výpočtu. Vzhledem k tomu, že jsou k dispozici nominální data, je možné zvolit přístup multikriteriální analýzy, který pracuje přímo s kvantitativními hodnotami.

Experimenty

Jednotlivé experimenty (případové studie) analyzují uzly (reprezentující zdroje propojených prostorových dat) a grafy (objekty propojených prostorových dat) na základě kvantitativních údajů získaných aplikací metrik. Testování probíhá na úrovni dílčích metrik, ale také pomocí souhrnných hodnot pro oba typy metrik, které jsou získány prostřednictvím metody vážených součtů jako jedné z technik multikriteriální analýzy. Vyzdvihovány jsou především extrémní hodnoty (u většiny metrik se jedná o maxima), kdy by takové zdroje dat měly představovat vhodné datové báze využívající princip propojených dat pro daný objekt, typ objektu, případně prostorová data jako celek. Grafy, které dosáhly vhodné extrémní hodnoty, představují objekty (typy objektů), které by mohly sloužit jako příklad dobré praxe (z hlediska využívání identických a podobnostních vazeb) pro transformaci prostorových dat do podoby Linked Data. V některých případech je u jednotlivých veličin ověřována také jejich variabilita, která vyjadřuje stabilitu hodnot jednotlivých metrik, což do jisté míry může být chápáno jako spolehlivost zdroje nebo homogenita způsobu popisu datových objektů.

Následující seznam obsahuje výčet realizovaných případových studií.



Obrázek 4: Architektura – implementace metodiky.

- Hlavní města ve střední Evropě
- Evropské mezinárodní silnice
- Hraniční řeky
- Uzlová letiště
- Republiky
- Srovnávací studie – globální a lokální data (hory v České republice a stratovulkány na celém světě), data s odlišnou geografickou lokalizací (hlavní města v Evropě a Africe)

Poslední experiment je realizovaný na všech získaných datech. Jeho výsledky by měly tvořit základ pro zobecnění poznatků týkajících se identických vazeb mezi objekty propojených prostorových dat. Celkově byly v rámci různých skupin dat získány údaje o 4 502 objektech.

Výsledky

Z důvodu nedostatku prostoru nejsou v tomto textu publikovány konkrétní výsledky jednotlivých experimentů, včetně dat, grafického vyjádření a statistického vyhodnocení. Tyto informace jsou dostupné v habilitační práci autora[22].

Z výsledků získaných v dílčích případových studiích a také během zpracování celého vzorku dat lze získat informace o vhodných zdrojích propojených dat, které obsahují prostorová data, a také o způsobu, jakým jsou v oblasti prostorových dat zavedeny identické vazby. V první řadě je možné z provedených experimentů odvodit význam jednotlivých metrik pro hodnocení uzlů a celých grafů.

Centralita stupně popisuje pozici uzlu z pohledu přímých propojení na ostatní vrcholy grafu. Z hlediska propojených prostorových dat se tedy jedná o počet zdrojů, které jsou pomocí identických vazeb spojeny s datovou sadou, jejíž centralita stupně je zkoumána. Jinými slovy jde o pozici kritickou z hlediska zapojení zdroje pomocí identických vazeb do systému Linked Data. Zdroje s vysokou hodnotou centrality stupně mají velké množství přímých sousedů, a tudíž i potenciál mnoha nepřímých vazeb na další zdroje nebo z dalších zdrojů. Proto lze říci, že zdroje s vysokou centralitou mají v síti propojených dat prominentní postavení, jak z hlediska ovlivňování ostatních zdrojů, tak z hlediska robustnosti a odolnosti vůči chybám celé sítě.

Centralita blízkosti vyjadřuje, jak dobře je uzel dostupný ze všech částí grafu. Z hlediska uzlů jako zdrojů sémantických propojených prostorových dat je vysoká hodnota centrality blízkosti důležitá z hlediska rychlého procházení datové sítě při získání nových informací z reprezentací stejného geografického objektu. Tato metrika nemá pro

[22] http://gis.zcu.cz/projekty/Identity_links/

hodnocení uzlů jako zdrojů propojených dat velký význam, z tohoto důvodu jí byla při volbě vah přiřazena poměrně nízká hodnota.

Centralita mezilehlosti indikuje, že daný uzel tvoří „most“ uvnitř grafu mezi jeho více či méně izolovanými součástmi (podgrafy). Zdroje s vysokou centralitou blízkosti jsou atraktivní především proto, že při prohledávání síťového grafu pro daný objekt nebo koncept budou s velkou pravděpodobností objeveny (protože přes ně vede většina cest mezi ostatními uzly), a tudíž budou zohledněny i informace, které takové zdroje obsahují. Je však nutné si uvědomit, že takový zdroj představuje riziko ve smyslu tzv. „úzkého hrdla“. To znamená, že pokud dojde k chybě (například ke změně persistentních URI nebo výpadku serveru), může tento problém a jakékoli jeho narušení vést až ke kolapsu celého systému identických vazeb pro daný objekt. Je nutné zmínit i důraz na kvalitu obsahu takového datového zdroje. Vzhledem k jeho poloze v síti ho budou uživatelé často zpracovávat, a proto budou případné chyby v obsahu ve velké míře přebírány i do uživatelských řešení. Tento fakt má i druhou stránku, která ukazuje přednosti crowdsourcingových řešení – časté využívání zdroje s velkou pravděpodobností povede i k tomu, že případné chyby budou rychle odhaleny a odstraněny.

Autorita představuje pravděpodobně jedinou metriku, která je, alespoň nepřímo, spojená s kvalitou zdroje. Jak vyplývá z názvu, jedná se o kvantitativní vyjádření stavu, že na uzel grafu odkazuje mnoho dalších vrcholů. Podle textu [70] je možné označit vrchol grafu (v přeneseném významu tedy i zdroj propojených dat reprezentovaný tímto vrcholem) s vysokým skóre autority za „prominentní, přední, či populární“. Vzhledem k tomu, že i při používání dat platí tržní principy, budou takové zdroje s vysokou mírou pravděpodobnosti i kvalitnější než ostatní datové sady. Popularita zdroje vyvolává kromě jeho častého využívání také zpětnou vazbu v podobě oprav možných chyb a přidávání nových informací. Takový zdroj je pomocí identických relací dosažitelný z mnoha dalších datových sad publikovaných podle zásad Linked Data přístupu.

Střed (Hub) je vrcholem, z něhož vychází velké množství propojení směrem ke zbytku sítě. Autor publikace [70] označuje vrchol s vysokým skóre středu jako vrchol „vlivný“. Lze tedy říct, že takový zdroj bude hojně využívaný pro vyhledávání dalších reprezentací objektů. Pokud bude uživatel potřebovat získat široké portfolio informací a dat, které nejsou součástí jedné datové sady, pravděpodobně začne prohledávat Linked Data prostor právě od některého z důležitých středů. To samozřejmě znamená, že obsah takového datového zdroje bude zřejmě častěji publikovaný (a analogicky k textu týkajícího se centrality mezilehlosti i verifikovaný), než tomu tak bude v případě zdrojů, které tvoří listové uzly grafu.

Page Rank je iterační algoritmus, který zjišťuje významnost jednotlivých vrcholů v síti na základě důležitosti uzlů, které jsou s takovým vrcholem přímo propojené. Zdroj propojených dat s vysokou hodnotou Page Rank je spojený s velkým množstvím kvalitních uzlů, kde tato kvalita je daná především mírou propojenosti těchto zdrojů.

Velikost grafu udává počet uzlů v grafu. V případě grafů publikovaných v této práci (tzv. datových sítí) tato metrika ukazuje, v kolika vzájemně propojených zdrojích Linked Data je objekt zmíněný. Hodnota velikosti grafu tedy souvisí s popularitou ve sféře propojených dat a především s potřebou publikovat taková data jako Linked Data, která je vyšší u takových datových sad a objektů, které mohou být nahlíženy z různých kontextů, a tudíž se mezi uživateli objevují odůvodněné požadavky na kombinaci informací o objektu z různých datových sad.

Hustota grafu je definovaná jako poměr počtu relací v grafu k hodnotě maximálního možného počtu vazeb. Tato metrika poskytuje informaci o tom, „jak moc jsou jednotlivé reprezentace objektu ve zdrojích propojených prostorových dat informovány o existenci jiných instancí objektu“. Je však nutné si uvědomit, že hustota nijak neřeší spojitost grafu.

Reciprocita grafu vyjadřuje míru vzájemného propojení mezi dvojicemi uzlů. Vysoká hodnota reciprocity znamená, že existuje velké množství obousměrných spojení v orientovaném grafu, což zvýší jeho průchodnost, která se promítne do lepší možnosti získávání informací o jednom objektu z různých zdrojů.

Shlukový koeficient odhaluje více či méně izolované skupiny v grafu. Tyto skupiny snižují robustnost a odolnost sítě vůči vnějším chybám (například technické poruchy u zdrojů, které propojují jednotlivé shluky) a zároveň mohou představovat skupiny zdrojů, které si jsou nějakým způsobem podobné (například z hlediska původu a zdroje informací, vzniku datové sady, obsahu, klasifikačních systémů a podobně).

Cílem testování a hodnocení pomocí uzlových metrik bylo nalezení „ideálního“ zdroje pro propojená prostorová data. Výraz „ideální“ je zapsaný v uvozovkách, protože je zcela jasné, že se nejedná o jeden dokonalý zdroj, ale o více datových sad, které vyhovují konkrétním účelům. V předchozích odstavcích jsou popsané jednotlivé metriky, jejichž úkolem je právě kvantifikovat vhodnost zdrojů dat pro jednotlivé účely. Je však nutné mít stále na paměti, že výsledky byly získány z limitovaného vzorku dat. Omezení tohoto vzorku, které zároveň představují možnosti budoucího směru dalšího výzkumu, se vztahují ke

- sběru dat – nejsou uvažovány zdroje, které nejsou svázané identickými vazbami s datovou sadou DBpedia (například GEMET, EuroVoc, NAL Thesaurus nebo AGROVOC, který obsahuje značné množství prostorových dat [89]);
- zpracování dat – pro datové sady, které nebyly v době získání validní z pohledu RDF (to znamená, že nebyly strojově zpracovatelné pomocí běžných technologií bez nutnosti programování), nebylo možné nalézt další vazby (i když v mnoha případech existují), a proto takové zdroje tvoří v grafu listové uzly;
- množství dat ve vzorku, které by bylo možné téměř neomezeně (limitujícím faktorem by zde bylo pouze technické řešení) zvětšovat za účelem získání přesnějších

informací.

Za zdroje vhodné pro propojení prostorových dat pro konkrétní účely můžeme označit takové, které jsou

1. propojené na velké množství dalších zdrojů (mají vysokou hodnotu centrality stupně a také skóre středu a autority), a tudíž jsou schopné zpřístupnit nebo naopak poskytnout další informace o objektu prostorových dat – podle případových studií je ve všech skupinách testovaných dat nejdůležitější DBpedia, v případě hlavních měst Wikidata (to platí pro hodnoty centrality stupně i skóre středu, skóre je popsáno níže);
2. blízko ostatním uzlům (mají vysokou centralitu blízkosti), a tudíž existuje menší riziko, že při narušení některé vazby nebudou odkazované informace dostupné – podle experimentů realizovaných v této práci disponují nejvyššími hodnotami centrality blízkosti DBpedia a Wikidata (v případech hlavních měst a republik);
3. díky své klíčové poloze, kdy propojují různé do jisté míry nezávislé části Linked Data prostoru (mají vysokou centralitu mezilehlosti), často procházené během získávání informací z různých prezentací jednoho objektu nebo konceptu – také v tomto případě jednotlivé experimenty vygenerovaly nejlepší skóre pro databázi DBpedia a pro množiny dat týkajících se hlavních měst i pro Wikidata;
4. často odkazované z jiných datových sad (mají vysokou hodnotu skóre autority), a tudíž představují na poli hodnoceného tématu dat populární zdroj a autoritu z pohledu poskytovaných informací – z hlediska jednotlivých experimentů je hodnocení autority nejvíce různorodé, přičemž jako nejdůležitější zdroje z pohledu této metriky se jeví GeoNames.org (většinou pro tradiční geografické prvky jako sídla, hory nebo státy), Wikidata a Yago (obojí spíše pro netypické datové sady jako závodní okruhy, ale také pro řeky nebo stratovulkány).
5. co nejdokonaleji integrované do prostoru propojených dat (což mimo jiné znamená i přímá propojení s klíčovými datovými sadami, která jsou kvantifikována hodnotou Page Rank) – jednotlivé případové studie z tohoto hlediska vyzdvihují především datové sady DBpedia (ve všech případech s výjimkou evropských mezinárodních silnic má tento zdroj nejvyšší hodnocení Page Rank), Wikidata a Yago (pro okruhy F1, IndyCar a také pro vzorek složený ze všech získaných dat).

V celkovém zkoumaném vzorku dat představuje DBpedia nejvhodnější zdroj z hlediska všech centralit a také skóre středu. Jinými slovy lze říci, že DBpedia obsahuje mnoho odkazů na další datové zdroje, díky nimž propojuje izolované části grafů. Wikidata a částečně také Yago, jsou nejvhodnější sadou z hlediska autority (odkazů vycházejících z jiných zdrojů a mířících na Wikidata nebo Yago). Z hlediska Page Rank, které lze chápat jako komplexnější metriku než ostatní, se jako nejlepší jeví DBpedia, těsně následovaná databázemi Wikidata a Yago.

Výsledky multikriteriální analýzy pomocí uzlových metrik jako kritérií přináší podobné

výsledky. Nejlépe hodnoceným zdrojem je DBpedia. Výjimkou jsou středoevropská hlavní města, kde vychází mírně lepší hodnocení pro Wikidata (jedná se zřejmě o anomálii ve vzorku způsobenou aktivním přístupem editorů produktu Wikidata ve střední Evropě). Jinak jsou v téměř všech případech Wikidata hodnocena jako druhá nejvhodnější. Posledním zdrojem, který má vysoké hodnoty váženého součtu, jsou GeoNames.org, která představují druhý nejvhodnější zdroj pro kategorie dat hory v České republice, uzlová letiště a stratovulkány.

Na základě výsledků multikriteriální analýzy lze datové zdroje (zkratky pro jednotlivé zdroje jsou používány z úsporných důvodů, jejich vysvětlení je k dispozici v příloze) rozdělit na skupiny, které mají podobné hodnoty váženého součtu:

- DBpedia – Tento vyniká svojí hodnotou váženého součtu vysoko nad ostatní zdroje. Je to dáno jednak kvalitou samotné databáze a také způsobem získávání dat pro případové studie, kdy DBpedia byla používána jako počátek vyhledávacího procesu.
- Wikidata, Yago, GeoNames.org – Skupina, která představuje dominantní datové zdroje nejen z hlediska vícekriteriální analýzy, ale také z pohledu jednotlivých metrik.
- VIAF – Zdroj dat VIAF (společný projekt národních knihoven a katalogizačních systémů) se projevil především v některých dílčích metrikách u konkrétních skupin dat (například centralita blízkosti a autorita v případě středoevropských hlavních měst).
- LinkedGeoData, Deutschen Nationalbibliothek, Library of Congress Name Authority File – Jedná se o skupinu nestejnorodou z hlediska původu (dva zdroje poskytované významnými národními knihovnami a Linked Data verze OpenStreetMap), ale obecně se jedná o kvalitní zdroje z hlediska vazeb i z pohledu poskytování dostatečného množství prostorových dat.
- Bibliothèque nationale de France, Identifiants et Référentiels, FAST Authority File, ISNI, WebNDL, MusicBrainz, GADM – Další skupina zdrojů s podobnými hodnotami se již výrazně odlišuje od předchozích. Je tvořena především databázemi, které jsou poskytované menšími národními knihovnami, oborovými knihovnami nebo specifickými projekty. Zdroj GA (Database of Global Administrative Areas, GADM) představuje jeden z dalších zdrojů čistě prostorových dat (hranice administrativních území).
- National Library of Israel, Transparency International, Eurostat Linked Statistics, World Bank Linked Data, OpenEI, Linked Web APIs a další – V této skupině jsou velice důležité zdroje z hlediska statistických informace (například Transparency International, Eurostat, World Bank), které jsou dostupné jen pro státy. Jejich

skóre je nízké, protože se nevyskytují ve všech vzorcích dat.

- Ostatní zdroje již jsou nesourodé z hlediska výsledků multikriteriální analýzy a marginální z pohledu prostorových dat.

Z hlediska tvorby identických vazeb v nových datových sadách propojených prostorových dat lze využít následujících doporučení, která byla odvozená z experimentů a dalších poznatků získaných v rámci tohoto výzkumu.

1. Identických a podobnostních vazeb by mělo být realizováno co největší množství, přičemž je potřeba uvážit vhodný standard pro popis vazby a také shodu mezi obsahem (nikoli pouze názvem) propojovaných prvků.
2. Není nutné vytvořit vazby na všechny dostupné zdroje propojených dat (například kvůli pracnosti budování vazeb a nárůstu velikosti souborů s daty, viz následující příklad). Upřednostňovány by měly být ty zdroje, které
   - získaly příznivé hodnocení pomocí metriky Page Rank a multikriteriální analýzy,
   - jsou mezi sebou v prostorou propojených dat vzdálené (z hlediska počtu hran nutných k jejich propojení),
   - tvoří v prostoru Linked Data do velké míry izolovaný shluk (nová datová sada by pak měla úlohu prvku sítě, který propojuje nezávislé podgrafy),
   - jsou příbuzné z hlediska obsahu,
   - tvoří v síti neuzavřené trojúhelníky, jež by se po přidání nového zdroje změnily v regulérní shluky,
3. Je vhodné požádat odkazované zdroje o zpětné vazby na novou datovou sadu a případně se domluvit na předání částí kódů s vazbami nebo jiném způsobu vytěžení nových dat.
4. Je nutné pravidelně kontrolovat fungování identických vazeb, aby se předešlo problémům ve využívání dat a aby se nový zdroj nestal v očích uživatelů nespolehlivým.

Závěr

Prezentovaný výzkum se zabývá problematikou propojených dat a jejich vazby na data prostorová. Propojená data představují v současnosti velice aktuální téma, které je akcentováno jak ve vědeckém výzkumu, tak v aplikační sféře. Propojená data jsou podporována Evropskou Unií, která spolufinancuje a spolufinancovala řadu projektů zaměřených na toto téma (například MELODIES, SDI4Apps, SmartOpenData, EUCLID – EdUcational Curriculum for the usage of LInked Data nebo Linked Data for Libraries). Problematikou Linked Data se zabývá řada respektovaných institucí (národní knihovny, statistické úřady, univerzity) nebo komerčních společností od malých firem jako je

OpenLink Software až po giganty v oblasti informačních technologií jako je Google.

Z hlediska propojení Linked Data a prostorových dat již existují některé slovníky nebo datové produkty publikované jako propojená data. Ze skupiny slovníků přímo zaměřených na prostorová data a informace je nutné vyjmenovat GeoSPARQL, který kromě jiného obsahuje exaktně definované topologické vazby, ISA Programme Location Core Vocabulary (obsahuje například slovník pro zápis adres), Basic Geo (WGS84 lat/long) Vocabulary (souřadnice) nebo registry INSPIRE transformované do podoby Linked Data. Také národní mapovací agentury poskytují ve formě propojených dat nejen databáze prostorových objektů, ale také slovníky nebo ontologie typů objektů. Mezi hlavní propagátory Linked Data na úrovni mapových agentur patří například Ordnance Survey ve Velké Británii nebo americká USGS (United States Geological Survey) a její produkt U.S. National Map.

V České republice začínají vznikat pokusy o tvorbu propojených prostorových dat a slovníků publikovaných jako Linked Data a zaměřených na prostorová data a informace na Českém úřadu zeměměřičském a katastrálním, Institutu plánování a rozvoje hlavního města Prahy, Českém vysokém učení technickém (Fakulta elektrotechnická) nebo Západočeské univerzitě v Plzni (Fakulta aplikovaných věd). Poslední jmenované pracoviště, na němž působí autor této práce, se zaměřuje na tvorbu ontologie pro výměnný formát digitální technické mapy, která aspiruje na to stát se univerzálním katalogem typů objektů prostorových dat pro veřejnou správu. Druhým počinem Západočeské univerzity na poli propojených prostorových dat je tvorba a správa databáze Smart Points of Interest, která obsahuje zhruba 30 miliónů bodů rozmístěných po celém světě a publikovaných jako Linked Data, včetně využívání výše jmenovaných slovníků a identických i topologických vazeb na další datové sady, jako jsou GeoNames.org, LinkedGeoData, DBpedia nebo Wikidata.

Problematika propojených prostorových dat není důležitá jen z pohledu akademického výzkumu, ale může být zajímavá pro praxi z několika hledisek:

- Úspora (nejen finanční) při pořizování nových dat a také při správě vlastních dat – zvláště v případě rozsáhlých databází prostorových dat je důležité, aby se zbytečně neopakovaly často banální atributy jednotlivých datových objektů. Linked Data přístup je v tomto případě v souladu s principy, na nichž je postavena evropská směrnice INSPIRE [6] a které přisuzují klíčovou roli faktu, že prostorová data jsou dostupná přímo od jejich majitele, pořizovatele nebo správce. Propojená data v tomto případě představují technologickou platformu, která zajišťuje tuto přístupnost přímo na úrovni dat a nikoli prostřednictvím externích nástrojů, po jejichž použití je ještě ve většině případů nutná harmonizace takových integrovaných dat.
- Nové informace a souvislosti – různé datové sady nemusí nutně obsahovat pouze

redundantní data, ale především data, která se vzájemně doplňují. Linked Data umožňují velmi jednoduše (díky pravidlům pro propojená data) takové datové sady a především objekty v nich uložené propojit, a tak získat nové informace a souvislosti. Tato vlastnost propojených dat je důležitá především v těch oborech, kde se dá na jeden prvek reprezentující prostorovou entitu nahlížet několika způsoby.

- Komunikace – díky propojení datových položek na prvky ontologií, tezaurů nebo kontrolovaných slovníků má uživatel takových dat velice jednoduchou možnost zjistit význam nebo definici položky v datech, přičemž obojí bývá často kontextově závislé. Tento fakt usnadňuje komunikace mezi uživateli, kteří nejsou napojeni na stejné terminologické základy, využívají různou legislativu, pocházejí z odlišných vědních oborů nebo absolvovali různě zaměřené vzdělávání.

Tato práce se nezabývá propojenými prostorovými daty jako celkem, ale pouze jedním aspektem propojených prostorových dat. Tímto aspektem jsou identické vazby (někdy označované jako ekvivalentní nebo identické a podobnostní). Cílem práce je popsat využívání identických vazeb mezi reprezentacemi objektů propojených dat v doméně dat prostorových.

Výstupy práce jsou určené především pro ty uživatele, kteří by chtěli využít existující sady propojených prostorových dat, ale neorientují se v nich. Druhou skupinou uživatelů jsou zájemci o tvorbu databází propojených prostorových dat, kteří chtějí, aby jejich data splňovala podmínky pětihvězdičkového klasifikačního systému, a tudíž potřebují vytvořit identické vazby mezi svými daty a objekty v externích datových sadách. Výsledky výzkumu lze rozdělit do dvou skupin:

1. Doporučení vhodných zdrojů propojených prostorových dat, které jsou klasifikovány podle vlastností těchto zdrojů. Tyto vlastnosti byly kvantitativně vyjádřeny pomocí metrik. Například vlastnost, že zdroj je často odkazovaný z jiných datových sad, a tudíž s velkou pravděpodobností půjde o datový zdroj často využívaný a zřejmě také kvalitní (alespoň z hlediska potřeb uživatelů), je vyjádřena pomocí skóre autority. Celkem bylo navrženo šest metrik pro uzly grafu (centralita stupně, centralita blízkosti, centralita mezilehlosti, skóre autority, skóre středu a Page Rank) řešících různé parametry zdrojů z pohledu identických vazeb. Tyto metriky byly poté sumarizovány pomocí vícekriteriální analýzy. Jako nejvíce doporučované zdroje se ukázaly tyto datové sady: DBpedia a Wikidata; pro některé konkrétní účely nebo skupiny dat mohou mít značné uplatnění GeoNames.org, VIAF nebo Yago. Specifické je postavení LinkedGeoData – tato datová sada sice obsahuje velké množství prostorových objektů (jedná se o kopii OpenStreetMap), ale disponuje pouze malým počtem vazeb na externí data a ani jiné datové sady zatím nevyužívají potenciál LinkedGeoData a neposkytují na ni velké množství odkazů prostřednictvím identických vazeb. Z hlediska prostorových dat jsou důle-

žité ještě datové sady obsahující především ekonomická nebo politicko-geografická data, jako například Eurostat, Transparency International nebo World Bank.

2. Doporučení pro doplňování identických vazeb do existujících nebo nově tvořených datových sad. Tato doporučení se odvíjejí především od grafových metrik (hustota grafu, velikost grafu, reciprocita grafu a shlukový koeficient), které umožňují kvantifikovat dílčí parametry grafu (datové sítě obsahující zdroje propojených prostorových dat jako uzly a identické vazby jako hrany). Tyto parametry jsou podobně jako v předchozím případě shrnuty pomocí multikriteriální analýzy. Doporučení se netýkají pouze vhodných zdrojů (viz předchozí bod), ale doplnění datové sítě do stavu, kdy by se jednalo o systém, který bude robustní a odolný vůči lokálním chybám a zároveň aby realizace identických vazeb byla efektivní z hlediska pracnosti.

Výsledky výzkumu publikovaného v habilitační práci mohou přispět k lepšímu výběru vhodných datových zdrojů (tezaurů, znalostních bází, ontologií, kontrolovaných slovníků) pro jednotlivé oblasti prostorových dat a také k nalezení vhodného způsobu popisu prostorových objektů a konceptů ve formě propojených dat i jejich propojení z různých externích zdrojů. Je třeba si uvědomit, že Linked Data přístup, včetně implementace identických vazeb, znamená zcela nový pohled na prostorová data, neboť po jeho zavedení by využívání prostorových dat nemuselo být omezené žádnými technickými ani legislativními bariérami (viz pětihvězdičkový klasifikační systém). Taková data by byla publikována v univerzálním, otevřeném a nezávislém formátu, využívala by identifikátory objektů v podobě URI (systém, který je prověřený v oblasti internetu) a především by byla navázána na jiné datové sady. Uživatel by se tedy mohl soustředit na obsah dat a jeho využívání pro vlastní potřeby, přičemž by nemusel řešit technologické otázky týkající se harmonizace dat. Jednou z překážek v nastolení tohoto ideálního stavu je to, že běžní uživatelé a vlastníci prostorových dat zatím nemají informace o tom, jak najít vhodné datové zdroje pro své účely a jak vlastní data publikovat, aby splňovala standardy Linked Data a zároveň, aby takové publikování dat neúměrně nezatěžovalo jejich poskytovatele. Právě tuto mezeru v dostupných informacích se snaží zacelit tato práce.

Předložená práce vychází z autorových aktivit v oblasti propojených prostorových dat na Západočeské univerzitě, v Office of Knowledge Exchange, Research and Extension (Food and Agriculture Organization of the United Nations) a v projektech SDI4Apps, SmartOpenData a Metodika pro publikování prostorových informací ve formě otevřených dat[23]. Identické vazby jako hlavní téma byly zvoleny proto, že jsou považovány za klíčový prvek propojených dat. Jak již bylo několikrát v textu práce uvedeno, umožňují připojit k datům nové informace z jiných datových bází, které jsou nezávislé z hlediska pohledu a správy, nebo přidat sémantickou informaci pomocí vazby na slovník, onto-

[23]Veřejná zakázka Technologické agentury České republiky TB0500MV003.

logii nebo tezaurus. Právě kombinace dat z různých zdrojů bez nutnosti taková data přímo vlastnit a spravovat je rozhodně a nezpochybnitelně hlavní výhodou Linked Data přístupu.

Seznam literatury

[1] R. Devillers, A. Stein, Y. Bédard, N. Chrisman, P. Fisher, and W. Shi, "Thirty years of research on spatial data quality: Achievements, failures, and opportunities," *Transactions in GIS*, vol. 14, no. 4, pp. 387–400, 2010.

[2] P. A. van Oort, *Spatial data quality: From description to application*. Wageningen Universiteit, 2006.

[3] A. Dragland, "Big data–for better or worse," *SINTEF, retrieved on July*, vol. 22, 2013.

[4] J. Gantz and D. Reinsel, "Extracting value from chaos," *IDC iview*, vol. 1142, pp. 1–12, 2011.

[5] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," *IDC iView: IDC Analyze the future*, vol. 2007, pp. 1–16, 2012.

[6] I. Directive, "Directive 2007/2/ec of the european parliament and of the council of 14 march 2007 establishing an infrastructure for spatial information in the european community (inspire)," *Published in the official Journal on the 25th April*, 2007.

[7] T. Berners-Lee, "Linked data: Design issues." 2006.

[8] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pp. 205–227, 2009.

[9] J. Macura, "Porovnání projektů wikidata a dbpedia jako zdrojů prostorových dat," Bachelor Thesis, University of West Bohemia, 2016.

[10] D. Wood, M. Zaidman, L. Ruth, and M. Hausenblas, *Linked data*. Manning Publications Co., 2014.

[11] T. Heath and C. Bizer, "Linked data: Evolving the web into a global data space," *Synthesis lectures on the semantic web: theory and technology*, vol. 1, no. 1, pp. 1–136, 2011.

[12] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality

assessment for linked data: A survey," *Semantic Web*, vol. 7, no. 1, pp. 63–93, 2015.

[13] M. Acosta, A. Zaveri, E. Simperl, D. Kontokostas, S. Auer, and J. Lehmann, "Crowdsourcing linked data quality assessment," in *International semantic web conference*, 2013, pp. 260–276.

[14] L. Ding, J. Shinavier, T. Finin, and D. L. McGuinness, "Owl: SameAs and linked data: An empirical study," 2010.

[15] L. Ding, J. Shinavier, Z. Shangguan, and D. L. McGuinness, "SameAs networks and beyond: Analyzing deployment status and implications of owl: SameAs in linked data," in *International semantic web conference*, 2010, pp. 145–160.

[16] S. Bechhofer *et al.*, "Why linked data is not enough for scientists," *Future Generation Computer Systems*, vol. 29, no. 2, pp. 599–611, 2013.

[17] S. Tartir, I. B. Arpinar, M. Moore, A. P. Sheth, and B. Aleman-Meza, "OntoQA: Metric-based ontology quality analysis," 2005.

[18] C. Guéret, S. Wang, and S. Schlobach, "The web of data is a complex system—first insight into its multi-scale network properties," in *Proceedings of the eccs*, 2010, vol. 10.

[19] C. Guéret, P. Groth, C. Stadler, and J. Lehmann, "Assessing linked data mappings using network measures," in *Extended semantic web conference*, 2012, pp. 87–102.

[20] E. Estrada and Ö. Bodin, "Using network centrality measures to manage landscape connectivity," *Ecological Applications*, vol. 18, no. 7, pp. 1810–1825, 2008.

[21] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, "Linked data on the web (ldow2008)," in *Proceedings of the 17th international conference on world wide web*, 2008, pp. 1265–1266.

[22] G. De Melo, "Not quite the same: Identity constraints for the web of linked data." in *AAAI*, 2013.

[23] H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuinness, and H. S. Thompson, "When owl: Sameas isn't the same: An analysis of identity in linked data," in *International semantic web conference*, 2010, pp. 305–320.

[24] B. Bennett, "Application of supervaluation semantics to vaguely defined spatial concepts," 2001, pp. 108–123.

[25] B. Bennett, "Physical objects, identity and vagueness," in *KR*, 2002, pp. 395–408.

[26] B. Bennett and P. Agarwal, "Semantic categories underlying the meaning of 'place'," in *International conference on spatial information theory*, 2007, pp. 78–95.

[27] I. Tiddi, M. d'Aquin, and E. Motta, "Quantifying the bias in data links," in *In-*

ternational conference on knowledge engineering and knowledge management, 2014, pp. 531–546.

[28] Q. Lu and L. Getoor, "Link-based classification," in *ICML*, 2003, vol. 3, pp. 496–503.

[29] N. M. Tichy, M. L. Tushman, and C. Fombrun, "Social network analysis for organizations," *Academy of management review*, vol. 4, no. 4, pp. 507–519, 1979.

[30] M. Emirbayer and J. Goodwin, "Network analysis, culture, and the problem of agency," *American journal of sociology*, vol. 99, no. 6, pp. 1411–1454, 1994.

[31] C. Haythornthwaite, "Social network analysis: An approach and technique for the study of information exchange," *Library & information science research*, vol. 18, no. 4, pp. 323–342, 1996.

[32] N. Blüthgen, J. Fründ, D. P. Vázquez, and F. Menzel, "What do interaction network metrics tell us about specialization and biological traits," *Ecology*, vol. 89, no. 12, pp. 3387–3399, 2008.

[33] A. Abbasi, J. Altmann, and L. Hossain, "Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures," *Journal of Informetrics*, vol. 5, no. 4, pp. 594–607, 2011.

[34] O. Cimenler, K. A. Reeves, and J. Skvoretz, "A regression analysis of researchers' social network metrics on their citation performance in a college of engineering," *Journal of Informetrics*, vol. 8, no. 3, pp. 667–682, 2014.

[35] L. Freeman, "The development of social network analysis," *A Study in the Sociology of Science*, 2004.

[36] E. Otte and R. Rousseau, "Social network analysis: A powerful strategy, also for the information sciences," *Journal of information Science*, vol. 28, no. 6, pp. 441–453, 2002.

[37] A. Agarwal, A. Corvalan, J. Jensen, and O. Rambow, "Social network analysis of alice in wonderland," in *Workshop on computational linguistics for literature*, 2012, pp. 88–96.

[38] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *Neuroimage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[39] A. G. Dunn and J. I. Westbrook, "Interpreting social network metrics in healthcare organisations: A review and guide to validating small networks," *Social Science &*

Medicine, vol. 72, no. 7, pp. 1064–1068, 2011.

[40] P. O. Vaz de Melo, V. A. Almeida, and A. A. Loureiro, "Can complex network metrics predict the behavior of nba teams?" in *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining*, 2008, pp. 695–703.

[41] F. M. Clemente, M. S. Couceiro, F. M. L. Martins, and R. S. Mendes, "Using network metrics in soccer: A macro-analysis," *Journal of human kinetics*, vol. 45, no. 1, pp. 123–134, 2015.

[42] G. Kossinets and D. J. Watts, "Empirical analysis of an evolving social network," *science*, vol. 311, no. 5757, pp. 88–90, 2006.

[43] A. Bajaj and R. Russell, "AWSM: Allocation of workflows utilizing social network metrics," *Decision Support Systems*, vol. 50, no. 1, pp. 191–202, 2010.

[44] I. Varlamis, M. Eirinaki, and M. Louta, "A study on social network metrics and their application in trust networks," in *Advances in social networks analysis and mining (asonam), 2010 international conference on*, 2010, pp. 168–175.

[45] B. Hajian and T. White, "Modelling influence in a social network: Metrics and evaluation," in *Privacy, security, risk and trust (passat) and 2011 ieee third inernational conference on social computing (socialcom), 2011 ieee third international conference on*, 2011, pp. 497–500.

[46] T. Spiliotopoulos and I. Oakley, "Understanding motivations for facebook use: Usage metrics, network structure, and privacy," in *Proceedings of the sigchi conference on human factors in computing systems*, 2013, pp. 3287–3296.

[47] I. Varlamis, M. Eirinaki, and M. Louta, "Application of social network metrics to a trust-aware collaborative model for generating personalized user recommendations," in *The influence of technology on social network analysis and mining*, Springer, 2013, pp. 49–74.

[48] T. Zimmermann and N. Nagappan, "Predicting defects using network analysis on dependency graphs," in *Software engineering, 2008. icse'08. acm/ieee 30th international conference on*, 2009, pp. 531–540.

[49] R. Premraj and K. Herzig, "Network versus code metrics to predict defects: A replication study," in *Empirical software engineering and measurement (esem), 2011 international symposium on*, 2011, pp. 215–224.

[50] L. Ding, T. Finin, and A. Joshi, "Analyzing social networks on the semantic web," *IEEE Intelligent Systems (Trends & Controversies)*, vol. 8, no. 6, pp. 815–820, 2004.

[51] Y. Theoharis, Y. Tzitzikas, D. Kotzinos, and V. Christophides, "On graph features of semantic web schemas," *IEEE Transactions on Knowledge and Data Engineering*,

vol. 20, no. 5, pp. 692–702, 2008.

[52] Q. K. Telesford *et al.*, "Reproducibility of graph metrics in fMRI networks," *Frontiers in neuroinformatics*, vol. 4, p. 117, 2010.

[53] R. Gil, R. García, and J. Delgado, "Measuring the semantic web," *AIS SIGSEMIS Bulletin*, vol. 1, no. 2, pp. 69–72, 2004.

[54] R. Čada, Z. Ryjáček, and T. Kaiser, *Diskrétní matematika*. Západočeská univerzita, 2004.

[55] F. Thung, D. Lo, M. H. Osman, and M. R. Chaudron, "Condensing class diagrams by analyzing design and network metrics using optimistic classification," in *Proceedings of the 22Nd international conference on program comprehension*, 2014, pp. 110–121.

[56] M. Bode, K. Burrage, and H. P. Possingham, "Using complex network metrics to predict the persistence of metapopulations with asymmetric connectivity patterns," *ecological modelling*, vol. 214, no. 2, pp. 201–209, 2008.

[57] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," *Physical review letters*, vol. 87, no. 19, p. 198701, 2001.

[58] C. Guéret, P. Groth, F. Van Harmelen, and S. Schlobach, "Finding the achilles heel of the web of data: Using network analysis for link-recommendation," in *International semantic web conference*, 2010, pp. 289–304.

[59] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, pp. 35–41, 1977.

[60] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1978.

[61] B. Wellman, "Network analysis: Some basic principles," *Sociological theory*, pp. 155–200, 1983.

[62] P. Ristoski, M. Schuhmacher, and H. Paulheim, "Using graph metrics for linked open data enabled recommender systems," in *International conference on electronic commerce and web technologies*, 2015, pp. 30–41.

[63] R. Sinha and R. Mihalcea, "Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity," in *Semantic computing, 2007. icsc 2007. international conference on*, 2007, pp. 363–369.

[64] K. Coursey and R. Mihalcea, "Topic identification using wikipedia graph centrality," in *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics, companion volume:*

Short papers, 2009, pp. 117–120.

[65] P. Grindrod, M. C. Parsons, D. J. Higham, and E. Estrada, "Communicability across evolving networks," *Physical Review E*, vol. 83, no. 4, p. 046120, 2011.

[66] J. Zhao, T.-H. Yang, Y. Huang, and P. Holme, "Ranking candidate disease genes from gene expression and protein interaction: A katz-centrality based approach," *PloS one*, vol. 6, no. 9, p. e24306, 2011.

[67] S. P. Borgatti and M. G. Everett, "A graph-theoretic perspective on centrality," *Social networks*, vol. 28, no. 4, pp. 466–484, 2006.

[68] E. Yan and Y. Ding, "Applying centrality measures to impact analysis: A coauthorship network analysis," *Journal of the Association for Information Science and Technology*, vol. 60, no. 10, pp. 2107–2118, 2009.

[69] I. Varlamis, M. Eirinaki, and M. Louta, "Application of social network metrics to a trust-aware collaborative model for generating personalized user recommendations," in *The influence of technology on social network analysis and mining*, Springer, 2013, pp. 49–74.

[70] M. Nykl, "Určování vyznamnosti vrchol grafu: PageRank a jeho modifikace," Technical report No. DCSE/TR-2013-09, University of West Bohemia, 2013.

[71] R. S. Burt, *Structural holes: The social structure of competition*. Harvard university press, 2009.

[72] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, "Network analysis in the social sciences," *science*, vol. 323, no. 5916, pp. 892–895, 2009.

[73] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[74] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, 1999.

[75] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[76] S. R. Sundaresan, I. R. Fischhoff, J. Dushoff, and D. I. Rubenstein, "Network metrics reveal differences in social organization between two fission–fusion species, grevy's zebra and onager," *Oecologia*, vol. 151, no. 1, pp. 140–149, 2007.

[77] Y. Qu, W. Ge, G. Cheng, and Z. Gao, "Class association structure derived from linked objects," 2009.

[78] S. Milgram, "The small world problem," *Psychology today*, vol. 2, no. 1, pp. 60–67,

1967.

[79] L. A. Adamic, "The small world web," in *International conference on theory and practice of digital libraries*, 1999, pp. 443–452.

[80] S. Borgatti, "Structural holes," *analytictech. com*, vol. 20, no. 1, pp. 35–38, 1997.

[81] M. Everett and S. P. Borgatti, "Ego network betweenness," *Social networks*, vol. 27, no. 1, pp. 31–38, 2005.

[82] V. Arnaboldi, M. Conti, A. Passarella, and F. Pezzoni, "Analysis of ego network structure in online social networks," in *Privacy, security, risk and trust (passat), 2012 international conference on and 2012 international confernece on social computing (socialcom)*, 2012, pp. 31–40.

[83] M. E. Newman, S. Forrest, and J. Balthrop, "Email networks and the spread of computer viruses," *Physical Review E*, vol. 66, no. 3, p. 035101, 2002.

[84] S. S. Wasserman, "A stochastic model for directed graphs with transition rates determined by reciprocity," *Sociological methodology*, vol. 11, pp. 392–412, 1980.

[85] J.-P. Rodrigue, C. Comtois, and B. Slack, *The geography of transport systems*. Routledge, 2013.

[86] A. Hogan, A. Polleres, J. Umbrich, and A. Zimmermann, "Some entities are more equal than others: Statistical methods to consolidate linked data," in *4th international workshop on new forms of reasoning for the semantic web: Scalable and dynamic (nefors2010)*, 2010.

[87] R. Isele, J. Umbrich, C. Bizer, and A. Harth, "LDspider: An open-source crawling framework for the web of linked data," in *Proceedings of the 2010 international conference on posters & demonstrations track-volume 658*, 2010, pp. 29–32.

[88] S. Greco, J. Figueira, and M. Ehrgott, "Multiple criteria decision analysis," *Springer's International series*, 2005.

[89] R. Palma, T. Reznik, M. Esbrí, K. Charvat, and C. Mazurek, "An inspire-based vocabulary for the publication of agricultural linked data," in *International experiences and directions workshop on owl*, 2015, pp. 124–133.

Abstract

The goal of this habilitation thesis is a development and testing of the methodology for the evaluation of identity links of spatial Linked Data. The methodology is composed of metrics focused on a quantification of properties of graph nodes (for example degree centrality or Page Rank) and properties of graphs (for example reciprocity of graph or clustering coefficient). The graph represents a geographical object in the Linked Data space, where graph nodes are Linked Data resources and edges mean identity relations. Final values of metrics are summarized by multiple-criteria decision analysis with particular metrics as criteria and alternatives stand for either Linked Data resources or data networks of spatial data objects. Research results are composed of description and comparison of resources containing spatial data interlinked by identity relations. There is also published a set of recommendation for spatial Linked Data development with focus on identity links. The thesis includes several case studies illustrating and optimizing metrics as well as the complete methodology.