Identické vazby propojených prostorových dat

Otakar Čerba

Habilitační práce

Vysoké učení technické v Brně
2017

Abstrakt

Cílem habilitační práce je vytvoření a testování metodiky pro hodnocení identických vazeb mezi objekty prostorových propojených dat. Metodika se skládá z dílčích metrik zaměřených na kvantifikaci vlastností uzlů v grafech (například centralita stupně nebo Page Rank) a vlastností celých grafů (například reciprocita grafu nebo shlukový koeficient). Graf představuje geografický objekt v prostoru propojených dat, kde uzly grafu jsou zdroje propojených dat a hrany reprezentují identické vazby. Výsledné hodnoty metrik jsou dále zpracovány pomocí multikriteriální analýzy, v níž jednotlivé metriky tvoří kritéria analýzy a alternativy představují buď zdroje propojených dat, nebo datové sítě objektů prostorových dat. Výstupy výzkumu se skládají z popisu a srovnání datových zdrojů obsahujících prostorová data a propojených pomocí identických vazeb. Dále je publikována sada doporučení pro tvorbu propojených prostorových dat s ohledem na identickém vazby. Práce obsahuje několik případových studií, které ukazují a pomáhají optimalizovat metriky i celou metodiku.

Klíčová slova: propojená data, prostorová data, identické vazby, graf, zdroj propojených dat.

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). A 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 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.

Keywords: Linked Data, spatial data, identity links, graph, Linked Data resource.

Poděkování

Za pomoc, mnohdy i nevědomou nebo nepřímou, při vytvoření této práce jmenovitě děkuji následujícím osobám (v abecedním pořadí, bez akademických, vědeckých a pedagogických titulů): Phil Archer, Caterina Caraciollo, Jiří Cajthaml, Václav Čada, Aleš Čepek, Diviš Čerba, Otakar Čerba, Prokop Čerba, Eva Čerbová, Kristýna Čerbová, Ján Feranec, Josef Fryml, Hana Hladíková, Karel Charvát, Karel Janečka, Karel Jedlička, František Ježek, Jan Ježek, Štěpán Kafka, Johannes Keizer, Jáchym Kellar, Michal Kepka, Martina Kepka Vichrová, Dzmitry Kozhukh, Zbyněk Křivánek, Miroslav Lávička, Radovan Machotka, Tomáš Mildorf, Barbora Musilová, Pavel Novák, Ján Pravda, Jiří Pyšek, Petr Rapant, Tomáš Řezník, Jiří Šíma, Pavel Vlach; dále projektům CentraLab, Exliz, plan4business, SDI4Apps, SmartOpenData a všem osobám v nich zapojeným.

Obsah

Seznam obrázků

Seznam tabulek

Seznam zkratek

API Application Programming Interface
Bash Bourne-again shell
CSV Comma-separated values
DAML The DARPA Agent Markup Language Homepage
DPZ Dálkový průzkum Země
DTD Document Type Definition
FAO Food and Agriculture Organization of United Nations
FAST Faceted Application of Subject Terminology
FTP File Transfer Protocol
GEMET GEneral Multilingual Environmental Thesaurus
GRDDL Gleaning Resource Descriptions from Dialects of Languages
HITS Hypertext Induced Topic Search
HTTP Hypertext Transfer Protocol
HTTPS Hypertext Transfer Protocol Secure
INSPIRE INfrastructure for SPatial InfoRmation in Europe
KLDR Korejská lidově demokratická republika
LAN Local Area Network
LD Linked Data
LOD Linked Open Data
MCDA Multiple-criteria decision analysis
N3 Notation3
NAL National Agricultural Library
NTIS Nové technologie pro informační společnost
OIL Ontology Inference Layer
OWL Web Ontology Language

PDF Portable Document Format
PUNTIS Podpora udržitelnosti centra NTIS
RDF Resource Description Framework
RDFS RDF Schema
RELAX NG REgular LAnguage for XML Next Generation
RIF Rule Interchange Format
SAARC South Asian Association for Regional Cooperation
SDI Spatial Data Infrastructure, Infrastruktura prostorových dat
SKOS Simple Knowledge Organization System
SPARQL SPARQL Protocol and RDF Query Language[1]
SPOI Smart Points of Interest
UNECE United Nations Economic Commission for Europe
URI Uniform Resource Identifier
USA United States of America; Spojené státy americké
USGS United States Geological Survey
VIAF Virtual International Authority File
W3C World Wide Web Consortium
WGS World Geodetic System
XML Extensible Markup Language
Yago (YAGO) Yet Another Great Ontology
ZČU Západočeská univerzita

[1] Rekurzivní zkratka.

Kapitola 1

Úvod

Aktuální problémy prostorových dat a propojených dat

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 Big Data, for better or worse: 90% of world's data generated over last two years [3] zhruba 90% dat na celém světě vzniklo v posledních dvou letech[2]. Tento fakt víceméně potvrzuje i starší studie Extracting Value from Chaos [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 (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

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

tvorba infrastruktur prostorových dat (SDI) na různých úrovních nebo potřeba mezinárodní spolupráce a vzájemného sdílení dat, jako je například směrnice INSPIRE [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[3], LinkedGeoData.org[4] nebo datové sady produkované Ordnance Survey[5]. Důležitý je i podíl prostorových dat v největších Linked Data datových sadách Wikidata[6] a DBpedia[7] (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,

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

- jsou navázané na sémantiku prvků,
- propojují data (různé datové reprezentace jednoho objektu),
- 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 Quality Assessment for Linked Open Data: A Survey [12], Crowdsourcing Linked Data quality assessment [13], owl: sameAs and Linked Data: An empirical study [14] nebo SameAs networks and beyond: analyzing deployment status and implications of owl: sameAs in linked data [15].

Navíc podle článku Linked Data – The Story So Far [8] patří mezi hlavní směry výzkumu (Research Challenges) také „Schema Mapping and Data Fusion“ (především oblast „Data Fusion“ závisí na identických a podobnostních vazbách), „Link maintenance“ (včetně aktualizace vazeb a nefunkčních linků) a především (z pohledu této práce) „Trust, Quality and Relevance“. 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 Why Linked Data is not Enough for Scientists [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.

Cíle výzkumu

Vzhledem k informacím uvedeným v předchozí části Aktuální problémy prostorových dat a propojených dat je tento výzkum zaměřený na kvalitu identických vazeb pro prostorová data publikovaná ve formě Linked Data. Jak bylo uvedeno v předchozí části, propojená data představují trend i budoucnost informačních technologií (viz [11]), včetně technologií geoinformačních.

Klíčovou vlastností propojených dat jsou vazby na slovníky nebo na externí datové položky. To ukazuje i pětihvězdičkový klasifikační systém LOD ([7]), v němž požadavek na externí vazby představuje pátý, a tedy nejvyšší stupeň propojených otevřených dat. Existuje sice velké množství různých typů vazeb (například vztahy hierarchické nebo propojení charakteru část-celek, tzv. meronymická vazba), ale relace identické a podobnostní jsou velice důležité, především z toho hlediska, že umožňují propojení různých, z pohledu původu a správy nezávislých datových sad. Kombinace dat z různých zdrojů bez nutnosti taková data přímo vlastnit a spravovat je rozhodně hlavní výhodou Linked Data přístupu (další přednosti propojených dat jsou zmíněné v části Linked Data v kapitole Základní pojmy). Při propojování různých datových sad je však nutné mít na paměti i rizika spojená s tímto přístupem (podrobnější výčet potenciálních problémů souvisejících s propojenými daty obsahuje podobně jako v předchozím případě část Linked Data v kapitole Základní pojmy). Z pohledu identických a podobnostních vazeb mohou taková rizika spočívat například v absenci vazeb mezi existujícími prvky nebo v chybném zavedení vazby (ať už z hlediska typu vazby, neexistujícího prvku na jedné straně vazby nebo propojení nesouvisejících objektů). Přítomnost takových nedostatků však není argumentem proti využívání propojených dat, ale spíše důvodem pro změnu paradigmatu správy a nakládání s takovými daty.

Cílem habilitační práce je přispět k diskuzi o využívání identických a podobnostních vazeb mezi objekty propojených dat, především v doméně dat prostorových. První část výzkumu je zaměřena na stanovení kritérií pro hodnocení jednotlivých aspektů kvality identických a podobnostních vazeb a způsobů (metrik) jejich kvantitativního hodnocení. V dalším kroku jsou metody hodnocení aplikovány na vybrané vzorky prostorových dat (případové studie), přičemž získané výsledky jsou postupně ověřovány a následně zobecněny. Posledním cílem této práce je 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)[8]. V jednotlivých fázích jsou ověřovány, testovány, srovnávány a aplikovány na testovací vzorek dat metody (metriky) pocházející z různých vědeckých oborů (například teorie grafů nebo geografie dopravy). Vazby a jejich typy jsou zkoumány z mnoha pohledů, jakými jsou například jejich existence, charakter, vzájemná provázanost prvků nebo homogenita (budování vazeb na základě jednotného logické postupu). Na základě získaných informací jsou navržena zlepšení v oblasti publikování, ale také reálného využívání Linked Data v oblasti prostorových dat jako přístupu pro získávání nových informací.

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 koncepty (například hlavní města), typy konceptů (například koncepty týkající se politické geografie) nebo typy datových objektů (vybraná hlavní města).
- 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.

Kvalita identických a podobnostních vazeb je testovaný na doméně

[8] Vedle tohoto náhledu na kvalitu existuje ještě podle autora „social trust“ a „trust based on provenance information“.

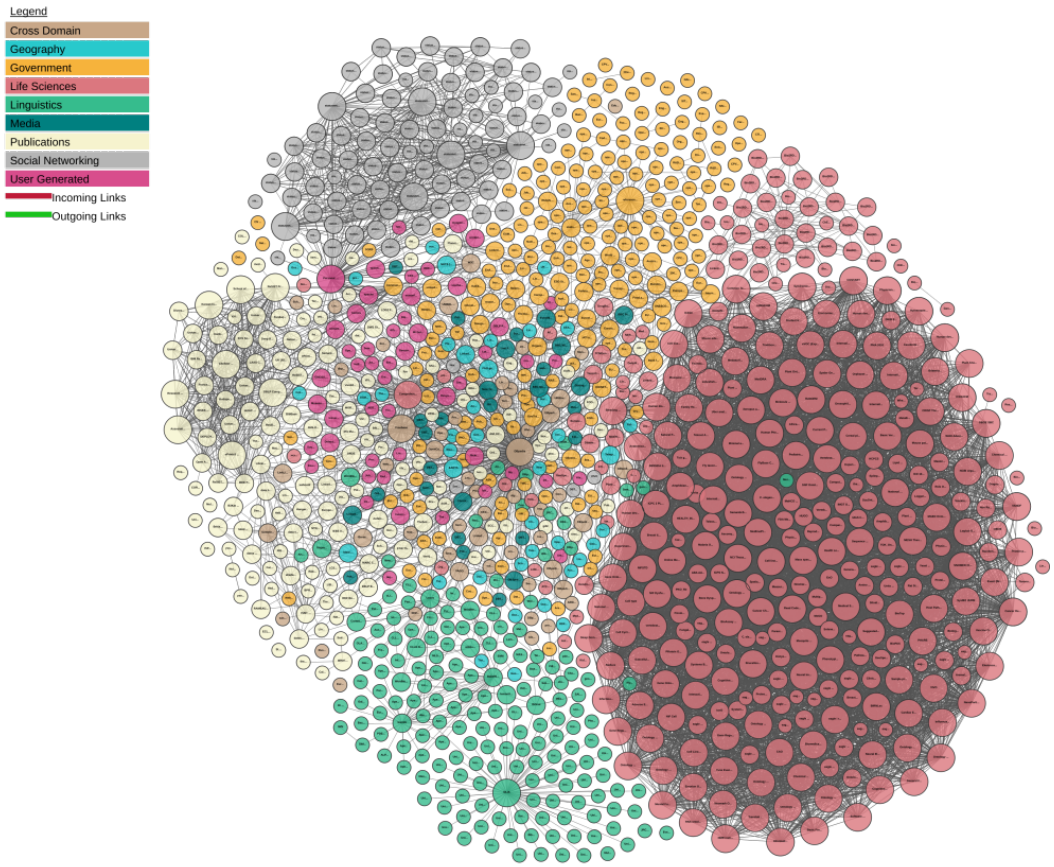
prostorových (geografických, geoprostorových, geo-) propojených dat. Tyto oblasti jsou voleny jako pilotní nebo ilustrační ze čtyř hlavních důvodů:

- 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 [17]).
- Publikace [11] uvádí, že „geografie je další faktor často propojující tematické domény“.
- Jak je patrné z Linking Open Data cloud diagram[9] – Obrázek 1, 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.

Publikované výstupy a získané poznatky respektují

- nejnovější poznatky v oblasti modelování sémantických dat, sémantického webu, ontologického inženýrství i hodnocení kvality dat, včetně dat prostorových,
- doporučení příslušných pracovních skupin odborných organizací (například Geospatial Semantic Web Community Group[10]),
- mezinárodní, všeobecně respektované standardy (např. jazyky OWL, SKOS, SPARQL, RDFS nebo formát RDF),
- existující nástroje a systémy (slovníky, tezaury a ontologie) pro podporu sémantiky v oblasti prostorových dat a informací,
- výsledky mezinárodních projektů a iniciativ, které se v minulosti zabývaly otázkami spojenými s tématem tohoto výzkumu nebo je řeší v současnosti (například LinkedGeoData.org, GeoKnow[11], LOD2[12],

[9] http://lod-cloud.net/
[10] https://www.w3.org/community/geosemweb/
[11] http://geoknow.eu/
[12] http://aksw.org/Projects/LOD2.html



Obrázek 1: Linking Open Data cloud diagram, únor 2017, autoři: Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch a Richard Cyganiak.

LOD Around the Clock, MELODIES[13], SmartOpenData[14], SDI4apps – Uptake of open geographic information through innovative services based on Linked Data[15] a další),
- trend tvorby otevřených a vzájemně propojených dat prosazovaných nejen v rámci Evropské Unie.

Hlavní rozdíly oproti existujícím studiím (zmíněným v kapitolách Základní pojmy, Rešerše a Metodika) spočívají především

- ve výše zdůvodněném zaměření na prostorová data a termíny souvisejícími s vědními obory téměř výhradně pracujícími s prostorovými daty (například geografie, vědy o Zemi, geomatika nebo geoinformatika),
- ve zdůraznění kvality, nikoli kvantity při studiu vazeb propojených dat, včetně manuální kontroly relevance vazeb (existující výzkumy jsou spíše zaměřeny na automatické odhalování chyb ve vazbách a jejich statistické vyhodnocení, například [18], autor se však domnívá, že je nutné na základě vhodných případových studií najít konkrétní nedostatky, chyby a typy chyb, ty odstranit nebo zohlednit při implementaci Linked Data).
- v orientaci na konkrétní entity, a nikoli na datové sady, jako například v publikacích [19] nebo [20].

Výzkum kvality identických a podobnostních vazeb pro prostorová data publikovaná jako Linked Data by měl především přispět k efektivnějšímu využívání prostorových dat jako celku i k lepšímu vytěžení předností propojených dat. V širším kontextu může jít o nemalé úspory související se získáváním a správou dat v mnoha praktických oborech od cestovního ruchu nebo správy majetku až po risk management, včetně fungování státní správy, neboť používání prostorových dat se nevztahuje pouze na vědecké obory, ale je každodenní rutinou v mnoha praktických aktivitách.

[13] http://www.melodiesproject.eu/
[14] http://www.smartopendata.eu/
[15] http://sdi4apps.eu/

Motivace

Důvodů pro zpracování tématu kvalita identických a podobnostních vazeb prostorových propojených dat a vytvoření tohoto dokumentu je velké množství:

1. Výše uvedená problematika patří mezi aktuální výzkumné otázky v oblasti geomatiky, geoinformatiky a ostatních příbuzných oborů. Svědčí o tom například výše uvedené projekty, publikace v seznamu literatury nebo velké množství konferencí a odborných sympózií zaměřených na toto téma (například Linking Geospatial Data[16], Linked Data Seminar[17] nebo GeoVoCamp[18]).
2. Dalším důvodem jsou opatření na mezinárodní i národní úrovni, které souvisí s tvorbou infrastruktury prostorových dat (SDI). Jedná se například o INSPIRE [6], GeoInfoStrategii[19] v případě České republiky nebo Global Spatial Data Infrastructure[20]. Právě takové aktivity vyžadují modelování pomocí propojených dat, přičemž důraz je kladen především na sémantiku a přístupnost, které hrají podle expertů podílejících se na tvorbě těchto strategických dokumentů klíčovou roli při zajišťování interoperability prostorových dat a informací. V souvislosti s pronikáním geografických konceptů do oblasti SDI je nutné zmínit například aktivity typu Reusable INSPIRE Reference Platform (ARE3NA)[21], Ordnance Survey Linked Data Platform[22] nebo Geospatial Semantics and Ontology[23] související s tvorbou U.S. National Map.
3. Navrhované téma je vybráno také s ohledem na oblasti výzkumu, kterým se autor dlouhodobě věnoval a věnuje na Katedře geomatiky[24], Fakulty aplikovaných věd Západočeské univerzity v Plzni a také v rámci projektů Nové technologie pro informační společnost (výzkumného programu VP6: Sběr, zpracování a sdílení geoprostorových dat), PUNTIS – Podpora

[16] https://www.w3.org/2014/03/lgd/
[17] http://www.pilod.nl/wiki/Linked_Data_Seminar_-_December_2,_2016
[18] http://vocamp.org/wiki/GeoVoCampSB2015
[19] http://www.mvcr.cz/clanek/geoinfostrategie.aspx
[20] http://gsdiassociation.org/
[21] https://joinup.ec.europa.eu/community/are3na/home
[22] http://data.ordnancesurvey.co.uk/
[23] http://cegis.usgs.gov/ontology.html
[24] dříve na Oddělení geomatiky, Katedry matematiky

udržitelnosti centra NTIS – Nové technologie pro informační společnost a Exliz.

4. V rámci projektu Exliz měl autor na jaře 2014 možnost účastnit se tříměsíční stáže v Office of Knowledge Exchange, Research and Extension (Food and Agriculture Organization of the United Nations), kde spolupracoval na analýze systému AGROVOC[25] (vícejazyčný tezaurus pro oblast zemědělství) z pohledu geografických konceptů, jejich vymezení a vazeb.
5. Téma také úzce souvisí s autorovou disertační prací Ontologie jako nástroj pro návrhy datových modelů vybraných témat příloh směrnice INSPIRE [21], kde byly analyzovány přínosy doménových ontologií (jako jednoho z nástrojů úzce souvisejících se sémantikou dat a informací) při procesu harmonizace prostorových dat a datového modelování.
6. Aktivity spojené s problematikou propojených dat na doméně prostorových dat a informací zasahují také do mezinárodních projektů, které byly a jsou realizovány na výše zmíněných pracovištích ZČU a také v Českém centru pro vědu a společnost a společnosti Help service remote sensing, se kterými autor úzce spolupracuje. Jedná se například o projekty SmartOpenData nebo SDI4apps – Uptake of open geographic information through innovative services based on Linked Data.
7. Absence literatury na téma „prostorová propojená data“ v rámci geomatiky, geoinformatiky a příbuzných vědních oborů. Tento nedostatek je pociťován především v případě zdrojů psaných v českém jazyce.
8. Interdisciplinarita – data a informace, která se dají označit jako prostorová, mají velice silný vliv na každodenní život člověka (mediální informace, navigace, základní i profesní rozhodovací procesy, volnočasové aktivity). Nezanedbatelná je také obecná vazba prostorových dat na různé obory lidské činnosti (například ochrana životního prostředí, stavebnictví, krizový management, zemědělství, lesnictví, doprava, cestovní ruch apod.) a také na vědecké obory (například geomatika, geoinformatika, geografie, geologie, hydrologie, historie, archeologie apod.), protože většina informací, znalostí a dat v těchto oborech je

[25] http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus

spojena s nějakou vazbou na prostor.

Poznámky pro čtenáře

Následující poznámky jsou určeny pro laskavé čtenáře tohoto textu. Ten se částečně zabývá sémantikou, chápáním smyslu a významů. Proto autor považuje za nutné v úvodu vysvětlit některá stanoviska, které mohou na některé čtenáře působit nesrozumitelně a mohou tak způsobit negativní nebo zkreslené vnímání celého textu.

- V autorských pasážích tohoto textu budou používány termíny propojená data a Linked Data jako synonyma. Výraz Linked Open Data (LOD) bude brán jako ekvivalentní k výše uvedenému, protože otevřenost je podle názoru autora přímým a nutným předpokladem pro vznik propojených dat. Termín Linked Open Data bude tedy většinou zmiňován pouze jako součást citací. Podobně nebudou rozlišovány výrazy „propojená prostorová data“ a „prostorová propojená data“. V tomto případě se jedná pouze o úhel pohledu, zda mluvčí považuje tento typ dat primárně za podmnožinu dat prostorových nebo propojených[26]. Vymezení a vazby mezi dalšími pojmy jsou uvedeny v části Základní pojmy.
- Ve fragmentech kódu nejsou zmíněné identifikátory ve formě URI jednotlivých jmenných prostorů. Jsou však využívány standardní zkratky, jako například `skos:` nebo `owl:`, u nichž je možné nalezení konkrétních odkazů v příslušných standardech. Seznam použitých jmenných prostorů je k dispozici v Příloze A.
- Citace z cizích jazyků (především z angličtiny), se kterými není dále v dokumentu nakládáno (tj. mají pouze informativní charakter), nejsou překládány do češtiny (důvodem je možnost zkreslení původního textu). Podobně i převzaté obrázky a schémata budou zachovány v originálním jazyce.
- Pro vytvoření textu byl použit textový editor gedit[27] (od verze 3.10.4),

[26] Z hlediska tradiční matematiky se jedná o průnik, přičemž tato operace je komutativní.
[27] http://www.gedit.org

který patří do skupiny svobodného software (GNU General Public License, verze 2). Pro převod z formátu Markdown[28] do PDF a případně dalších výstupů byly využity knihovny Pandoc[29] (GNU General Public License), LaTeX[30] – svobodný software šířený pod LaTeX Project Public License a šablony, které byly inspirovány šablonami, jejichž autory jsou Tom Pollard[31] a Chia Huei Tan[32]. Software použitý pro sběr a zpracování dat je uvedený v kapitole Metodika.

- Grafické prvky dokumentu jsou vytvářené v aplikacích yEd[33] (od verze 3.16.1; volně dostupná aplikace se specifickou licencí[34]), LibreOffice[35] (Mozilla Public License v2.0.) a Graphviz[36] (Eclipse Public License - v 1.0) – tato knihovna je využitá především pro automaticky generované grafy a schémata.
- Výpočty a zpracování dat – software R[37] (GNU General Public License).
- Přílohy, včetně PDF souboru s textem habilitační práce, dat a zdrojových kódů, jsou k dispozici na webové adrese http://gis.zcu.cz/projekty/Identity_links/.

Struktura práce

Text této práce je rozdělený do sedmi základních kapitol. V Úvodu je popsána současná situace na poli propojených prostorových dat a také jsou definovány cíle výzkumu popisovaného v tomto dokumentu. Následuje kapitola představující základní termíny výzkumu. Konkrétně se jedná o pojmy geografický koncept, geografická data a propojená data. Dále je popsaný koncept propojených dat (Linked Data) s důrazem na identické a podobnostní vazby. Poslední část kapitoly Základní pojmy je věnovaná

[28] http://daringfireball.net/projects/markdown/
[29] http://pandoc.org/
[30] https://www.latex-project.org/
[31] https://github.com/tompollard/phd_thesis_markdown
[32] https://chiakaivalya.wordpress.com/2014/04/23/using-markdown-pandoc-to-write-my-biology-phd-thesis/
[33] http://www.yworks.com/products/yed
[34] file:///home/cerba/Programy/yEd/license.html
[35] https://www.libreoffice.org/
[36] http://www.graphviz.org/
[37] https://cran.r-project.org/

grafům jako klíčovým strukturám pro popis identických vazeb propojených dat. Třetí kapitola uvádí výzkum publikovaný v této práci do celosvětového kontextu. V první části rešerše jsou zmíněny především publikace zabývající se hodnocením kvality propojených dat především s ohledem na vlastnosti propojující ekvivalentní entity. Poté následuje výčet metrik pro kvantitativní analýzu sítí, jejichž výběr je implementován v metodice, která je jádrem čtvrté kapitoly Metodika. Její struktura koresponduje s postupem pro posuzování kvality ekvivalentních vazeb, který se skládá ze čtyř procesů – vyhledávání, sběr a formalizace informací o identických vazbách; výběr metrik pro hodnocení vazeb, jejich kompozice a deklarace vhodných parametrů; konkrétní implementace metodiky. Pátá kapitola převádí teoretické poznatky získané v předchozích částech práce do praxe. V rámci jednotlivých experimentů jsou publikovány jednotlivé případové studie ilustrující dříve publikovanou metodiku. V následující šesté kapitole (Výsledky) jsou výsledky experimentů shrnuty, diskutovány, komentovány a zobecňovány. Součástí této kapitoly je také naznačení směrů dalšího výzkumu. V Závěru jsou pak veškeré poznatky získané v rámci tohoto výzkumu přehledně uvedeny a srovnány dosažené výsledky s cíli práce. K práci jsou připojeny tři tištěné přílohy – seznam jmenných prostorů (Příloha A), soubor `resources.xml` (Příloha B) se seznam zdrojů propojených dat a zkratek používaných v celé práci, SPARQL dotazy vybírající z databáze DBpedia objekty patřící do testovaných skupin dat (Příloha C). Přílohy v elektronické podobě, včetně PDF souboru s textem habilitační práce, dat a zdrojových kódů, jsou k dispozici na webové adrese http://gis.zcu.cz/projekty/Identity_links/.

Kapitola 2

Základní pojmy

V této kapitole habilitační práce jsou postupně představené základní pojmy používané v celém dokumentu. V první řadě jsou stručně definována prostorová data a geografické koncepty. Největší pozornost je věnována pojmu propojená data (Linked Data), který je pro celý výzkum klíčový. Zmíněny jsou definice, specifika a výhody Linked Data přístupu, dále jsou popsány formáty a slovníky vhodné pro prostorová propojená data. Poté se text zaměřuje na vazby používané v Linked Data, na jejich typy a definice i popisy uvedené v jednotlivých standardech, přičemž hlavní důraz je kladen na identické a podobnostní vazby. Vzhledem k tomu, že prezentovaný výzkum kvality identických a podobnostních vazeb prostorových propojených dat je založený na grafových strukturách, jsou krátce představeny i základní pojmy z teorie grafů v míře odpovídající rozsahu této práce. Tato část textu může být používána i samostatně, bez ohledu na zbytek dokumentu, jako stručný úvod do problematiky prostorových propojených dat.

Geografický koncept, geografický prvek a prostorová data

V oblasti sémantického webu a jeho interakce s disciplínami jako geomatika, geoinformatika, GIScience a dalšími se objevují dva základní pojmy –

geografický koncept a geografický prvek (objekt, entita, položka dat). Geomatika a další výše zmíněné obory pracují s termínem prostorová data. Tato část textu by měla objasnit vztah mezi těmito základními pojmy.

Podle [22] je fundamentální otázkou současné geomatiky a příbuzných oborů způsob, jak se převádí (geografická) realita do formalizované podoby tak, aby co možná nejpřesněji reprezentoval svět a naše znalosti o něm. Tento proces se obecně označuje jako konceptualizace. Pojem „**geografický koncept**[38]“ („geo-koncept[39]“) je podrobně vysvětlený v publikaci Theories of Geographic Concepts [22]. Termín je zde diskutovaný z mnoha hledisek – specifické vlastnosti, které odlišují geo-koncepty od běžných konceptů, vymezení a definice hranic, typologie (rozčlenění na základě společných vlastností), metodologie tvorby a popisu jednotlivých geo-konceptů nebo samotné vymezení termínu geo-koncept.

V současné době existuje více definic termínu „geografický koncept“, které nejsou navzájem v rozporu a akcentují především přítomnost lokalizační informace:

- Podle [22] je geo-koncept definovaný na základě stanovení kontextů. V první řadě se jedná o koncepty, které obsahují prostorový rozměr nebo vlastnost(i) popisující lokalizaci v prostoru.
- V publikaci [24] je uvedeno, že termín „geo-koncept“ odpovídá entitě se zděděnou nebo nepřímo vyjádřenou prostorovou dimenzí[40].
- „Geoprostorový koncept je idea, která charakterizuje typ geografického prvku“ [25].

Geografické koncepty je možné dělit na „fyzické“ a „ne-fyzické“[41] [26].

[38]Koncept je definován jako „základní konstrukt v teorii mysli“ [23]. Podle [22] jsou koncepty „klíčové především z hlediska schopností kategorizace (klasifikace) a porozumění, proto tvoří základní stavební kámen ontologických i lingvistických systémů“.

[39]V dalším textu budou používána označení „geo-koncept“, „geoprostorový koncept“ a „geografický koncept“ (případně jejich anglické ekvivalenty) jako synonyma.

[40]Pod pojmem prostorová dimenze autoři zřejmě míní prostorovou vazbu. Není tedy možné v tomto konkrétním případě zaměňovat s dimensionalitou, tak jak je standardně chápána v geomatice. Z hlediska specifické geomatické terminologie se samozřejmě jedná o nesprávné použití termínu dimenze, ale úkolem této práce není rozbor terminologické kvality použitých zdrojů.

[41]České, poněkud kostrbaté termíny vznikly překladem z anglického originálu – „physical“ a „non-physical“ [26].

Příkladem konceptů spadajících do první skupiny mohou být například kategorie `budova`, `řeka` nebo `pohoří`. Mezi „ne-fyzické“ koncepty mohou patřit `parcela`, `kraj` nebo `obec`. Hlavní odlišnost obou typů spočívá v tom, že koncepty z první skupiny jsou fyzické rozlišitelné v krajině, zatímco objekty „ne-fyzických“ konceptů jsou reprezentovány nějakým (většinou abstraktním) elementem (jako například hranice).

V poslední uvedené definici figuruje také další důležitý pojem – „**geografický prvek**“. Ten má charakter jednotlivých geografických objektů[42], které se mohou sdružovat do konceptů. [27] uvádí, že každý geografický objekt je nějakým způsobem umístěný v prostorou pomocí souřadnicového systému[43]. V případě geografického konceptu může být vazba mnohem obecnější a nikoli exaktně vyjádřitelná. Příkladem geografického konceptu je například `město`, `stát` nebo `sopka`, zatímco příslušnými geografickými objekty mohou být například `Praha`, `Německo` nebo `Etna`. Tento princip je podobný jako v případě ontologií, kdy jsou rozlišovány třídy a instance (viz část věnovaná formátu OWL).

Geografické koncepty jsou důležité z pohledu sémantického webu, neboť jsou využívány ve slovnících a tezaurech, přičemž právě identické a podobnostní vazby umožňují propojení původně izolovaných sémantických nástrojů. Geografické prvky pak lze chápat jako prostorová data. Podobný přístup je publikovaný v článku Features, Objects, and other Things: Ontological Distinctions in the Geographic Domain [28]. Geografický koncept představuje jednu z pěti tzv. geografických kategorií. Tato množina geografických kategorií zahrnuje „geographic features, geographic objects, geographic concepts, something geographic, and something that could be portrayed on a map“.

Jak vyplývá z předchozích odstavců, je možné označit geografický prvek jako objekt (prvek) prostorových dat. Pod pojmem „**prostorová data**“ jsou

[42]V této práci budou výrazy jako „prvek prostorových dat“, „objekt prostorových dat“, „položka prostorových dat“ a „entita“ reprezentující prostorová data fungovat jako synonyma (pokud nebude explicitně uvedeno jinak).

[43]Poznámka autora: U geografických objektů bychom měli spíše hovořit o umístitelnosti pomocí souřadnicového systému, protože reálně mohou být objekty umístěny primárně nikoli pomocí souřadnic, ale prostřednictvím nepřímé lokalizace (například adresy), která však může být převedená do podoby souřadnic.

chápána data[44], která nesou prostorovou informaci. Ta může být uvedena ve formě souřadnic, což je nejčastější řešení. Existují však i jiné možnosti, jako například udávání prostorové informace ve formě adresy nebo topologického vztahu k jiným objektům (nepřímé georeferencování). Obojí (adresy i topologické vazby) lze realizovat i pomocí vazeb propojených dat.

Chápání termínu „prostorová data“, jak je popsáno v předchozím odstavci, nejlépe odpovídají dvě definice uvedené v [29]:

- „Prostorová data jsou jakákoliv data, která obsahují formální polohovou referenci, např. odkaz na buňku gridu. Jedná se např. o data DPZ nebo mapy.“
- „Prostorová data (anglicky spatial data) jsou data, která se vztahují k určitým místům v prostoru a pro která jsou na potřebné úrovni rozlišení známé polohy těchto míst.“

Vzhledem k tomu, že se prostorová data používána v této práci týkají planety Země a úzce se vztahují k oboru geografie, jsou termíny „geografická data“, „geodata“ a „geoprostorová data“ používána jako synonyma.

Propojená data (Linked Data)

Propojená data (**Linked Data**), podrobnější informace jsou k dispozici v publikacích [31] 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, standardizované, sémanticky popsané) vazby umožňují jednodušší sdílení a kombinování datových sad a jejich částí.

Podle [32] nebo [11] jsou Linked Data součástí globální databáze Web of Data,

[44]Publikace [29] definuje pojem „data“ jako „reprezentaci skutečností, pojmů nebo instrukcí (návodů, pokynů) ve formalizované podobě vhodné pro komunikaci, interpretaci a zpracování lidmi nebo automatickými prostředky“. Podobná definice se vyskytuje také v publikaci [30] - „Data jsou opakovatelná reprezentace informace formalizovaným způsobem, vhodným pro komunikaci, interpretaci nebo zpracování.“

přičemž článek [10] uvádí, že právě Linked Data umožňují přechod z tradičního webu (World Wide Web) na Web of Data. Více informací o Web of Data je k dispozici například v knize [11].

Vlastnosti propojených dat

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[45]:

1. Používej URI pro pojmenovávání jednotlivých prvků. – „Use URIs as names for things."
2. Používej HTTP URI, aby názvy byly dohledatelné. – „Use HTTP URIs, so that people can look up those names."
3. Těm, kdo vyhledávají URI, poskytni užitečné informace pomocí standardů (RDF, SPARQL). – „When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)."
4. Zařaď i vazby na jiná URI, protože tak umožníš objevení dalších věcí. – „Include links to other URIs, so that they can discover more things."

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.[46]

[45] Originální seznam [7] je často reprodukován v mnoha publikacích jako například [33].
[46] Poznámka autora: Je třeba si uvědomit, že schéma je kumulativní. To znamená, že bez

Mezi hlavní přednosti Linked Data přístupu patří podle [10] především kombinovatelnost s dalšími daty za účelem vytváření a získávání nových znalostí[47]. 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 datových sad a jejich prvků. Podobně i článek [34] také uvádí kombinovatelnost dat jako hlavní přednost Linked Data přístupu, ale zdůrazňuje ji především v souvislosti možností vývoje nových aplikací.

Vazby v propojených datech

Jak již 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 [10] „Links play a central role in Linked Data. A Linked Dataset should (indeed, must) link to other Linked Data on the Web.“. Podle [35] jsou vazby důležité z hlediska obohacení sémantiky, poukazování na nové zdroje informace a propojování datových sad.

Autoři publikace [11] rozlišují tři základní typy vazeb (příklady viz Obrázek 2):

splnění předchozí podmínky není možné splnit následující.

[47] „Linked Data has one amazing property: it may be easily combined with other Linked Data to form new knowledge. That is the best reason to explore and use Linked Data. Traditional data-management techniques have resulted in separation of most of our data into silos that can't be readily recombined. We need to write programs to find, access, convert, and combine data from silos before we can get on with any particular job. Linked Data makes that sort of work much easier because it's easy to combine Linked Data from multiple sources.“ [10]

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 (například `leží v`), které jsou charakteristické pro prostorová data. Do tohoto typu vztahů spadají i tzv. meronymické vazby (propojení části a celku).
2. Identity Links[48] 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 [36] 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[49].
- 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,

[48] 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 [36], 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 [37], [38] nebo [39].

[49] Poznámka autora: Tento jev lze chápat i negativně, budeme-li usilovat především o jednoznačnost poskytované informace. K jednomu pojmu totiž často existuje velké množství vzájemně si odporujících definic a popisů.

byl by celý systém velice náchylný k celkovému kolapsu, nemluvě o vysokých nákladech na koordinaci, administrativu a byrokracii.

Na obrázku 2[50] jsou ilustrovány příklady vazeb geografického objektu `Milešovka`. Na schématu jsou uvedeny vazby na slovníky, které určují především typ objektu. Dále jsou zmíněny dva typy identických vazeb - na nestrukturovaná[51] nebo strukturovaná, ale strojově nečitelná data (například webové stránky nebo obrázky) a na totožné nebo podobné objekty v jiných sadách Linked Data (rozdíl mezi těmito vazbami a zdůvodnění, proč se jedná o identické vazby, je uveden v následující kapitole. Posledním typem jsou topologické vazby, které reprezentují Relationship Links. V tomto případě se jedná například o vazby na okolní vrcholy, územní jednotky nebo pohoří.

Formáty, jazyky a slovníky používané pro Linked Data

RDF a RDFS

Resource Descriptor Framework[52] poskytuje mechanismus pro explicitní, formalizované a standardizované vyjádření sémantických informací [21]. RDF bylo původně vytvořené v roce 1999 jako standard založený na bázi XML pro zápis metadat (dat o datech) [40].

Základním principem formátu RDF je popis jevů a objektů (zdrojů) pomocí takzvaných trojic (triples). Ty se skládají ze subjektu (podle terminologie běžné v oblasti ontologií, která bude využívána i v tomto textu, jde o třídy nebo individuály), predikátu (vlastnost, která představuje binární relaci mezi oběma zbylými prvky) a objektu (třída, individuál, datový typ, hodnota)[53].

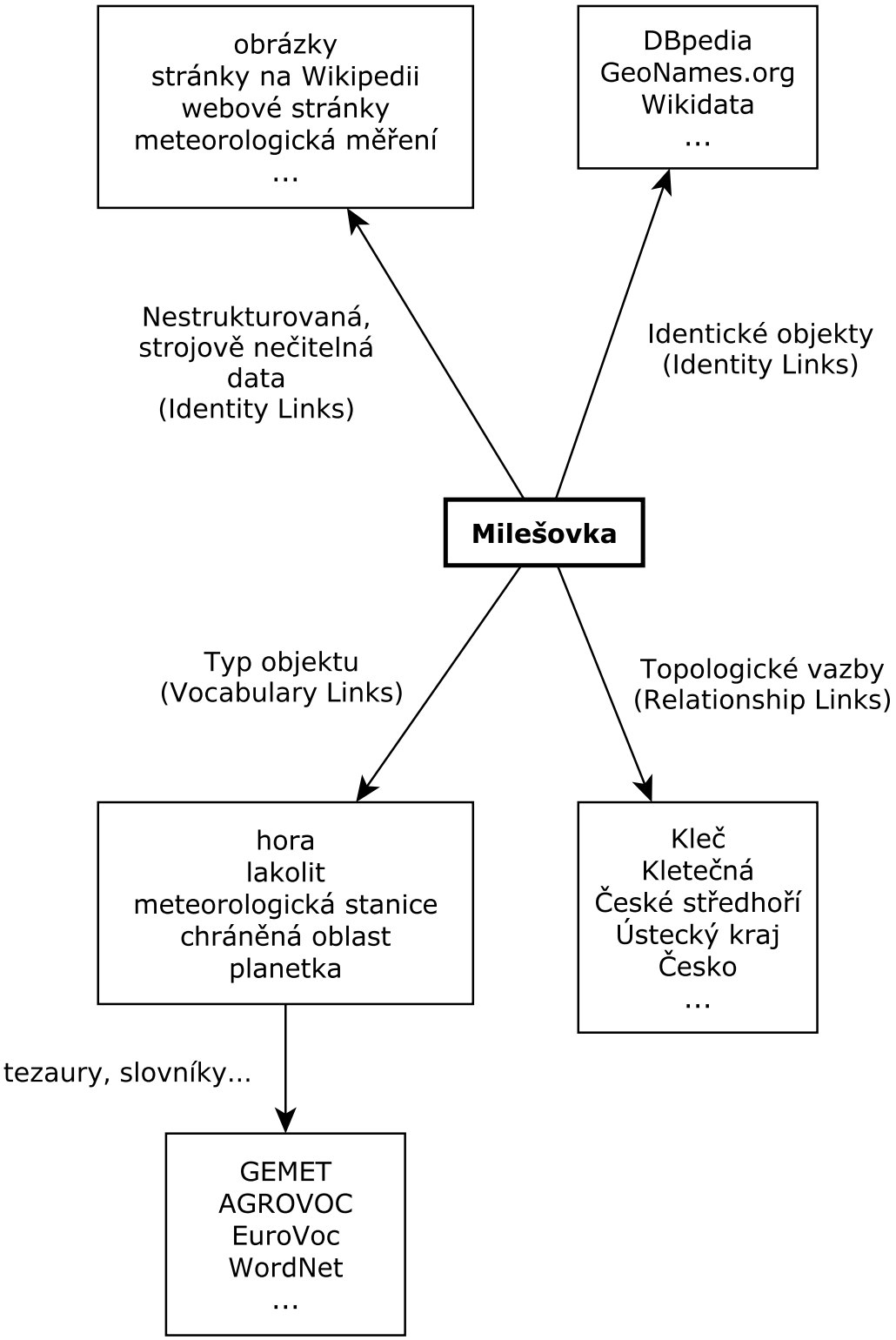
Taková trojice víceméně odpovídá struktuře přirozeného jazyka, kdy jednoduchá věta zpravidla obsahuje také tři základní prvky – podmět, přísudek a předmět (případně přívlastek nebo příslovečné určení). Převedeme-li

[50]Poznámka autora: Originál schématu byl prezentován na konferenci EGU 2016, modifikovaná verze pak na sympoziu Aktivity v kartografii 2016. Oba příspěvky byly publikovány pouze ve formě abstraktů, proto nejsou odcitované a zmíněné v seznamu literatury. Pro účely tohoto textu došlo k dílčím úpravám.

[51]Nejedná se přesně o identické vazby.

[52]https://www.w3.org/RDF/

[53]Podle [41] se pro objekty a subjekty používají také označení koncept, entita nebo zdroj.



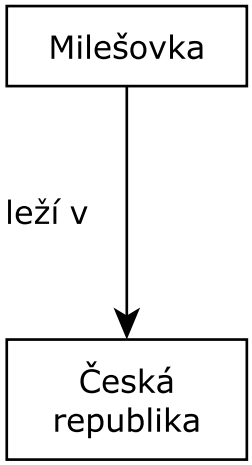
Obrázek 2: Příklady typů vazeb.

strukturu trojice do oblasti větné stavby a přirozeného jazyka, pak na místě subjektu bývá zpravidla substantivum (podmět věty), predikát je reprezentován slovesem (verbum, přísudek ve větné stavbě) a objekt má nejčastěji formu substantiva nebo adjektiva (předmětu věty nebo další fakultativní větné členy). Propojením prvků v RDF vzniká grafová struktura, jejíž hrany tvoří predikáty a uzly subjekty a objekty [41].

Pro jednotlivé části trojice mohou být používány jednoznačné identifikátory ve formě Uniform Resource Identifier (URI)[54], který představuje běžný mechanismus v prostředí internetu a který navíc našel své nezastupitelné místo i v přístupu Linked Data.

Informaci „Milešovka leží v České republice“ (grafické vyjádření viz Obrázek 3) můžeme rozdělit následujícím způsobem:

- Subjekt: `Milešovka`
- Predikát: `leží v`
- Objekt: `Česká republika`

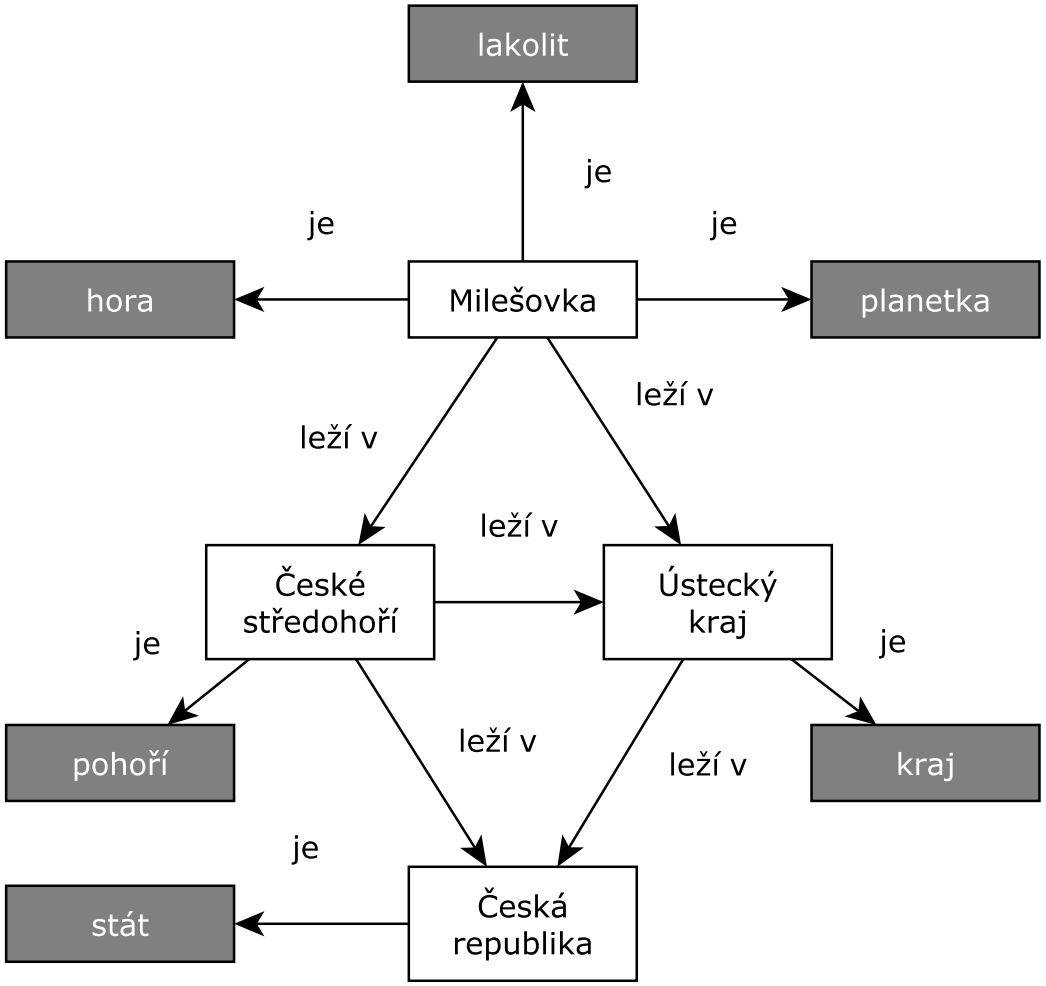


Obrázek 3: Trojice (objekt - predikát - subjekt).

Grafová struktura formátu RDF představuje vhodnou formu pro zápis prostorových dat. Je schopná vyjádřit de facto libovolnou vazbu mezi

[54]Kromě URI mohou být využity i literály, ty však nemohou představovat cíle v rámci odkazů ani subjekty.

jednotlivými geografickými objekty i koncepty, jak ilustruje následující příklad (Obrázek 4)[55].



Obrázek 4: Ukázka grafové struktury.

RDF trojice představuje datový model popisu informace. Je však nutné zavést také formalizovaný způsob zápisu, aby bylo možné RDF data strojově zpracovávat. RDF využívá značkovacích jazyků (XML). Výše uvedená trojice může být reprezentována následujícím kódem.

```
<rdf:description rdf:about="Milešovka">
    <lezi_v>Česká republika</lezi_v>
</rdf:description>
```

[55]Poznámka autora: Na schématech 3 a 4 uzly s bílým pozadím vyjadřují geografické objekty, uzly se šedým pozadím obecné geografické koncepty a šipky vazby.

V předchozí větě je výraz „může být reprezentován“ použit zcela záměrně. Nemá poukázat ani na vágnost RDF standardu, ani na fakt, že na místě objektu byl použitý pouze textový řetězec (literál), ale na to, že pro RDF data existuje několik rovnocenných syntaxí (například RDF/XML, Turtle, N-Triples, N3 a další). Hlavním důvodem pro více způsobů syntaxe je různý způsob využívání RDF dokumentů. V některých případech je důležitá kompatibilita s XML (na úkor větší velikosti souboru), jindy spolupráce s konkrétním softwarovým produktem pracujícím pouze s jedním formátem nebo je důležitá velikost souboru kvůli rychlosti přenosu.

Postupem času se ukázalo, že svoboda, kterou RDF nabízí svým uživatelům (například libovolné pojmenování predikátů), je spíše kontraproduktivní, protože v mnoha případech omezuje hlavní účel RDF – sdílení dat a informací. Uživatel mohl prvky v trojici nazvat libovolným způsobem, a tím docházelo ke zhoršování interoperability.

Proto došlo k zavedení dalších standardů, které mají některé typy vazeb předdefinované. Příkladem může být RDF Schema (RDFS)[56]. RDFS má standardizované některé běžné vlastnosti, typy objektů a subjektů (například `rdfs:subClassOf`, `rdfs:range`, `rdfs:domain`, `rdfs:Class`, `rdfs:Datatype` a další). RDFS je tedy nadstavbou (podmnožinou) RDF, která umožňuje standardizovaným způsobem definovat například hierarchie prvků nebo obor hodnot a definiční obor vlastností.

RDF a RDFS představují velice užitečné nástroje pro popis dat a informací. Mají ovšem také určité nedostatky, jako chybějící vyjádření kardinality nebo detailní specifikace typů vlastností. Proto byly vytvořeny jejich další nadstavby, které využívají principu trojic a předdefinovaných vlastností z RDFS, ale nabízejí širší možnosti například v oblastech deskripční logiky. Touto nadstavbou jsou například ontologické jazyky, především OWL.

Článek [41] uvádí několik předností formátu RDF:

- Standard RDF je spravován silnou a respektovanou organizací W3C.
- Formát RDF je provázán s dalšími nástroji, které zvyšují jeho vyjadřovací schopnosti (například SPARQL, GRDDL nebo RIF).

[56] http://www.w3.org/TR/rdf-schema/

- RDF může být používáno společně s dalšími formáty založenými na XML. Díky tomu RDF může být validováno pomocí automatických nástrojů (validátorů).
- Struktura trojic (triples) je jednoduchá a snadno pochopitelná. Na druhou stranu je pro zpracování trojic (například vyhledání) možné využít existující sofistikované grafové algoritmy vyvinuté matematiky nebo kybernetiky.
- Grafová reprezentace je srozumitelnější než tabulky (v mnoha případech)[57].
- RDF je schopné popsat nejen plochá data (například tabulky), ale i komplikované struktury jako například multihierarchické grafy.

Mezi další výhody patří i existence velkého množství návodů, tutoriálů a další dokumentace, například na portálech linkeddatatools.com[58], zvon.org[59] nebo w3schools.com[60].

OWL

Jazyk Web Ontology Language s poměrně nelogicky vytvořenou zkratkou OWL[61] je vyvíjen v rámci již několikrát zmíněné standardizační organizace World Wide Web Consortium (W3C). Z historického hlediska je jazyk OWL nejvíce svázaný se starším formátem pro popis ontologií DAML+OIL, na který víceméně navazuje[62]. V současné době (od roku 2012) existuje již druhá verze OWL označovaná jako OWL 2[63].

Podobně jako RDFS i jazyk OWL představuje zúžení původního RDF, které

[57]Tato vlastnost je důležitá zejména pro prostorová data reprezentující objekty reálného světa, jehož struktura spíše odpovídá grafům nežli tabulkám.

[58]http://www.linkeddatatools.com

[59]http://www.zvon.org

[60]http://www.w3schools.com

[61]Nabízející se zkratka WOL je využívána v jiných významech. Například Wake On LAN, World Online apod. Server http://acronyms.thefreedictionary.com eviduje celkem 16 významů zkratky WOL. Hypotézu o „obsazenosti“ zkratky WOL poněkud narušuje fakt, že stejný server po zadání zkratky OWL vyhledá celkem 24 výsledky. V anglické verzi online encyklopedie Wikipedia jsou pod příslušným heslem publikovány další informace o vytvoření zkratky OWL.

[62]Vývoj jazyků pro popis ontologií je popsán například v [21].

[63]https://www.w3.org/TR/owl2-overview/

vzniká definováním nových konstrukcí pro reprezentaci znalostí. OWL pro popis těchto znalostí využívá především metody deskripční logiky, které jsou zapisovány pomocí RDF trojic. Využívání deskripční logiky je důležité především při odvozování nových informací pomocí tzv. „reasoningu“. Prvky deskripční logiky (například kvantifikátory) jsou klíčové při vytváření restrikcí pro jednotlivé objekty ontologie. Dalším specifikem OWL je využívání tzv. Open world assumption. Tento přístup lze charakterizovat větou „co není explicitně zakázáno, je povoleno“ nebo „u čeho není jasně prokázána neexistence, může existovat“. Tím se odlišují ontologie zapsané pomocí jazyka OWL od jiných přístupů užívaných při zpracování informací, které ty používají tzv. Close world assumption, kde se předpokládá, že informace, data nebo znalosti jsou kompletní, a tudíž jestliže nějaká informace není popsána, pak ani nemůže existovat.

Podobně jako RDF i OWL využívá pro popis informací a znalostí několik typů syntaxe jako například RDF syntaxi, OWL/XML syntaxi, Manchester syntaxi a další. Sémantika jazyka OWL využívá tři základní typy prvků:

1. Třídy – představují skupiny objektů, které mají společné vlastnosti. Využijeme-li příklad na obrázku 4, pak může být třídou například objekt `hora` nebo `stát`. Hierarchie tříd je realizována pomocí vazby `is-a`, která propojuje obecnou a specifickou třídu. Tímto typem vazby mohou být například spojeny třídy `stát` a `evropský stát` nebo `stát` a `království`.
2. Instance – podobně jako třída i instance reprezentuje prvek ontologie. Zatímco třídy je možné dále dělit, instance (individuály) jsou v dané ontologii nedělitelné (příkladem z Obrázku 4 mohou být koncepty `Milešovka` nebo `Česká republika`).
3. Vlastnosti – konstrukce propojující a specifikující třídy a instance. Vlastnosti se dělí na objektové (vztahy mezi třídami nebo instancemi; příkladem objektové vlastnosti hojně využívané v tomto textu je `owl:sameAs`), datatypové (přiřazují objektu hodnoty na základě datového typu) a anotační (využívané pro popis; typickou anotační vlastností je například `rdfs:label`).

Více informací o standardu OWL je možné vyhledat v dokumentu [42] a v

textech, na které odkazuje. Disertační práce autora [21] popisuje možnosti při ukládání prostorových dat a souvisejících objektů a termínů do ontologií.

SKOS

Podobně jako OWL i SKOS[64] (Simple Knowledge Organization System) představuje specifický formát pro popis informace pomocí standardů RDF a RDFS. SKOS je tedy zúžením dvou dříve jmenovaných standardů. Za jeho vývojem stojí organizace W3C, SKOS je standardizován jako W3C Recommendation [43] z roku 2009. Oproti OWL je SKOS jednodušší, protože se nezaměřuje na struktury typu ontologie využívající pro popis informace i deskriptivní logiku. SKOS je určen především pro tvorbu znalostních systémů, tezaurů, slovníků, taxonomií nebo heslářů, proto je jeho role v sémantickém webu velice důležitá. Mezi příklady aplikace SKOS patří slovníky AGROVOC nebo EuroVoc[65].

Pro Linked Data jsou důležité především vlastnosti poukazující na shodné nebo podobné prvky – `skos:exactMatch` a `skos:closeMatch`. Kromě těchto vlastností je důležitá i výstavba hierarchie pojmů pomocí vlastností propojující obecnější a specifičtější termíny (`skos:broader` a `skos:narrower`, případně jejich tranzitivní ekvivalenty). Kromě nich je možné vytvořit vazby na příbuzné prvky nebo na pojmy, které používají daný termín. Pro tezaury a podobné produkty jsou výhodné i možnosti různých označení (popisků) pro jeden termín.

Identické a podobnostní vazby v propojených datech

V kontextu této části textu má smysl zmínit také termín „identita“ (resp. od něj odvozený přívlastek „identický“). Ve zkoumaných slovnících[66] je tento

[64] https://www.w3.org/2004/02/skos/

[65] http://eurovoc.europa.eu/

[66] Z webových stránek Internetové jazykové příručky[67] jsou dostupná hesla ve Slovníku spisovného jazyka českého, Slovníku spisovné češtiny a Akademickém slovníku cizích slov.

termín popisován jako „shoda ve všech vlastnostech“, „totožnost“ nebo „úplná stejnost“[68]. Popis identity pomocí logických pravidel je k dispozici v článku [44]. Definice identity instancí, která je založena na totožných hodnotách vlastností, je publikována v článku [45].

Tato kapitola vychází především ze standardů OWL [42], SKOS [43] a dalších studií jako například [46] nebo [47]. Následující přehled ukazuje, jak jsou definované jednotlivé vlastnosti propojující identické (shodné, ekvivalentní) a podobné objekty. Seznam vlastností je abecední, přičemž řazení i popis relace využívají i standardní prefixy (například `skos:` nebo `owl:`).

`owl:sameAs` Vlastnost `owl:sameAs` je speciálním případem relace `rdf:Property`[69]. Podle [46] vlastnost propojuje dva individuály[70] (tedy nikoli třídy[71]), proto se často využívá k mapování mezi ontologiemi. Referenční příručka [46] tvrdí, že vlastnost `owl:sameAs` poukazuje na to, „že dva identifikátory (URI) odkazují na stejný objekt, tyto individuály mají stejnou identitu“. Na tomto místě je potřeba poznamenat, že v originální definici je poslední slovo předchozího citátu, tedy „identita“, psáno v uvozovkách, což zcela jistě znamená, že se jedná o poměrně vágní pojem, především ve světě komunikace a přenosu informací.

`rdfs:seeAlso` V některých řídkých případech (například GeoNames.org) se pro odkazy na příbuzné (tedy i ekvivalentní a podobné) prvky používá i tato velice obecná vlastnost stadardizovaná v rámci RDFS [48]. Vlastnost `rdfs:seeAlso` má podle standardu [48] „poukazovat na zdroj, který může poskytovat doplňkové informace o subjektu“. Při výzkumu identických a podobných objektů je v případě použití vlastnosti `rdfs:seeAlso` přesvědčit se o charakteru datového zdroje a způsobu použití výše jmenované vlastnosti.

`skos:closeMatch` Publikace [43] uvádí, že vlastnost `skos:closeMatch` „je

[68] Výjimkou je význam slova „identita“ ve smyslu „národní svébytnosti“.

[69] Tento vztah mimo jiné ukazuje návaznost jazyka OWL na formát RDF.

[70] Standard OWL disponuje i opačnou vlastností `owl:differentFrom`, která by se v případě prostorových dat hodila pro odlišení různých prvků se stejným zeměpisným jménem.

[71] Výjimkou je dialekt OWL Full, kde se s třídami zachází stejně jako s instancemi, a vlastnost `owl:sameAs` může tudíž definovat ekvivalenci tříd. Pro propojení dvou stejných tříd se obvykle používá vlastnost `owl:equivalentClass`.

používána k propojení dvou konceptů, které jsou dostatečně podobné, takže mohou být zaměnitelně použity v aplikacích pro vyhledávání informací“. Vlastnost `skos:closeMatch` není definována jako tranzitivní (především z důvodu toho, že „dostatečná podobnost“ je velice vágní a neurčité kritérium pro propojování konceptů napříč znalostními bázemi).

`skos:exactMatch` Podle Referenční příručky standardu SKOS [43] vlastnost `skos:exactMatch` je použita „k propojení dvou konceptů vykazujících vysoký stupeň důvěry, že tyto koncepty mohou být využívány nezaměnitelně ve velkém množství aplikací pro vyhledávání informací“. Vlastnost `skos:exactMatch` je tranzitivní a je pod-vlastností `skos:closeMatch`.

Standard SKOS [43] ještě uvádí další vlastnosti, které slouží ke vzájemnému mapování zdrojů. Konkrétně se jedná o relace `skos:broadMatch`, `skos:narrowMatch` a `skos:relatedMatch`. V tomto dokumentu však nejsou více rozváděny, protože jejich využití je sporé a nevyskytují se ani v případových studiích.

Kromě čtyř výše uvedených standardizovaných vlastností pro zápis vazeb mezi identickými nebo podobnými prvky používají různé znalostní báze a další podobné nástroje využívající princip Linked Data ještě další vazby, které můžeme označit jako produktově specifické (viz tabulka 1 v článku [34]). Mezi nejdůležitější patří Wikidata Identifiers. V této sekci se vyskytují položky, které reprezentují odkazy na ekvivalentní nebo vysoce podobné objekty v jiných datových sadách. Problémem je, že každý externí zdroj má definovanou specifickou vlastnost. Například vlastnost P1566[72] představuje identifikátor v GeoNames.org.

Z předchozích řádků jsou patrné následující závěry:

1. V oblasti Linked Data neexistuje jedna dominantní vlastnost propojující stejné a podobné prvky.
2. Definice vlastností obsahují řadu vágních pojmů, které znesnadňují rozhodování, kdy jsou objekty skutečně identické nebo podobné.

[72] https://www.wikidata.org/entity/Property:P1566

Grafy

Vzhledem k tomu, že analýza vazeb bude probíhat s použitím grafových struktur, je nutné zabývat se také problematikou grafů. To je ještě zvýrazněno tím, že otázky terminologie z oblasti grafů nejsou v oboru geomatiky a geoinformatiky zcela intuitivně zavedeny. Proto se tato kapitola zabývá grafy od úplných elementárních začátků. Pokud není uvedeno jinak, veškeré informace v této části vycházejí z publikací [49] a [50]. Další pojmy z teorie grafů (například stupeň uzlu nebo typy uzlů) jsou uvedeny také v rešerši metrik.

Graf G je dvojice $G = (V, E)$, kde V je konečná množina a $E \subset \binom{V}{2} \wedge V \times V$, přičemž

$$\binom{V}{2} = \{\{x, y\} : x, y \in V \text{ a } x \neq y\}$$

je množina všech dvouprvkových množin (neuspořádaných dvojic) prvků množiny V. Prvky množiny V nazýváme **vrcholy** (často také uzly) grafu – $V(G)$, prvky množiny E pak **hrany** grafu G – $E(G)$. Vrcholy $x, y \in V$ jsou **sousední**, pokud $\{x, y\} \in E$. Tato definice nezahrnuje tzv. smyčky (tj. spojení uzlu se sebou samým) ani vícenásobné hrany. Obojí nemá význam pro popis vazeb v Linked Data.

Pro popis symetrických[73] a nesymetrických (asymetrických)[74] vazeb v Linked Data je zapotřebí rozlišovat mezi orientovanými (například obrázky 4 a 5) a neorientovanými grafy. V předchozí definici je množina hran popsána dvěma způsoby. Platí, že

$$E = \begin{cases} \binom{V}{2,} & \text{pro neorientované grafy,} \\ V \times V, & \text{pro orientované grafy.} \end{cases}$$

Jinými slovy hrana v orientovaných grafech je uspořádanou dvojicí uzlů, na

[73] Vazba R mezi prvky a, b z množiny X se označuje jako symetrická, pokud platí $\forall a, b \in X, aRb \Rightarrow bRa$.

[74] Vzhledem k rozsahu a zaměření tohoto textu nebude rozlišována silně a slabě asymetrická relace.

rozdíl od grafů neorientovaných, kde na pořadí uzlů nezáleží.

Z hlediska hodnocení vazeb Linked Data je důležitá souvislost grafu. Podle definice graf G je souvislý, pokud pro každé dva vrcholy x, y existuje v grafu G cesta z x do y. V opačném případě je graf G nesouvislý. Cesta z x do y v grafu G je sled $(x = y_0, y_1, ..., y_k = y)$, ve kterém se každý vrchol y_i objevuje pouze jednou. V případě orientovaných grafů existuje slabá souvislost, kdy souvislý je symetrický graf (graf vzniklý odstraněním orientace hran), a silná souvislost, kdy pro každé dva vrcholy x a y existují cesty x do y i z y do x.

Kvůli rozlišování typů vazeb (například vazby podobnostní a identické nebo vazby podle jednotlivých standardů) je možné používat ohodnocené grafy. Publikace [49] definuje ohodnocený orientovaný graf (G, w) je orientovaný graf G spolu s reálnou funkcí $w : E(G)(0,)$. Je-li e hrana grafu G, číslo $w(e)$ se nazývá její ohodnocení nebo váha.

Matice sousednosti (Adjacent matrix) je jeden ze způsobů, jak popsat graf, včetně propojení mezi jednotlivými vrcholy. Jedná se o čtvercovou matici, kdy počet řádků i sloupců odpovídá počtu vrcholů v grafu. Publikace [49] pro tuto matici používá následující definici. Nechť G je orientovaný graf na vrcholech $v_1, ..., v_n$. Matice sousednosti grafu G je reálná matice o rozměrech $n \times n$, definovaná předpisem $S(G) = (\sigma_{ij})$, kde

$$\sigma_{ij} = \begin{cases} 1 & \text{pokud } v_i v_j \in E(G), \\ 0 & \text{jinak,} \end{cases}$$

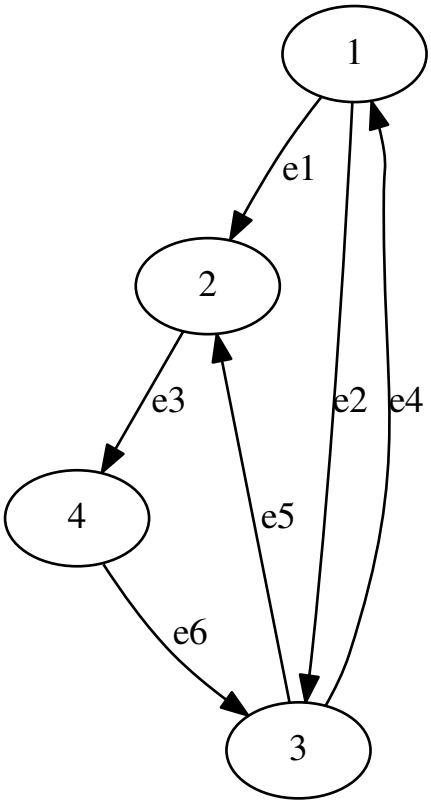
pro $i, j = 1, ..., n$.

Tato definice se týká orientovaných neohodnocených grafů. V případě grafů orientovaných je maticí sousednosti takového grafu matice sousednosti jeho symetrické orientace, kdy je každá hrana nahrazena dvojicí protichůdných orientovaných hran. Publikace [49] uvádí i výpočet matice sousednosti pro ohodnocené grafy (tzv. vážená matice sousednosti, str. 130).

V procesu zjišťování kvality vazeb propojených prostorových dat je možné matici sousednosti využít dvěma způsoby. V první řadě matice celistvě popisuje vazby mezi jednotlivými uzly grafu (tedy mezi výskyty reprezentací jednoho

geografického objektu v různých datových zdrojích). Matice sousednosti má však jednu zajímavou vlastnost. její jednotlivé mocniny ukazují cesty mezi uzly grafu s délkou, která odpovídá hodnotě mocniny. V případě propojených dat je důležité především nalezení nejkratší cesty mezi dvěma uzly.

Následující příklad (Obrázek 5 a příslušné matice) ukazují matici sousednosti a její mocniny pro ilustrační graf G se čtyřmi vrcholy (Obrázek 5). Matice jednotlivých mocnin ukazují počty a délky cest mezi jednotlivými vrcholy.



Obrázek 5: Ilustrační orientovaný graf.

$$S(G) = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$(S(G))^2 = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

$$(S(G))^3 = \begin{bmatrix} 0 & 1 & 2 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

Poslední matice této sekce ukazuje délky nejkratších cest mezi jednotlivými uzly grafu. Je třeba si uvědomit, že pro propojená data nemá význam spojení zdroje (uzlu) se sebou samým, proto hodnoty na diagonále nemají praktický význam.

$$\begin{bmatrix} 2 & 1 & 1 & 2 \\ 3 & 3 & 2 & 1 \\ 1 & 1 & 2 & 2 \\ 2 & 2 & 1 & 3 \end{bmatrix}$$

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 (definice viz kapitola Grafy). Zmínku o grafech znázorňujících `owl:sameAs` vazbu má ve svém článku [34]. Tyto grafy jsou však neorientované a předpokládá se u nich symetričnost vazby,

která však reálně neexistuje ([14]). Z těchto důvodů nejsou vhodné pro účely vyhodnocování a popisu kvality vazeb.

Publikace [35] uvádí tzv. **datovou síť** (Data Network). Ta je definovaná[75] jako orientovaný, označený graf $G = (V, E, L)$, kde V je množina uzlů, E množina hran a L množina popisků[76]. 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[77]. Pomocí grafových struktur vyjadřují vazby mezi propojenými daty (nikoli pouze identické a podobnostní) také [51]. V tomto případě je graf označován jako Linked data graph. Další vyjádření propojení dat pomocí orientovaného grafu je k dispozici například v [52]. V tomto případě se však nejedná o propojená data ve smyslu Linked Data.

Článek [35] nabízí také další dvě definice související s grafovým vyjádřením struktury propojených dat:

Sousedství Přímé sousedství uzlu $v_i \in G$ je množina uzlů $v \in G$, které jsou s uzlem v_i přímo propojeny sledovanou vazbou. Hrana této vazby může směřovat do uzlu v_i i z uzlu v_i. V prvním případě se jedná o vstupní hranu (resp. množinu vstupních hran N_i^-), ve druhém případě o hranu výstupní (resp. množinu vstupních hran N_i^+). Platí, že $N_i = N_i^+ \cup N_i^- = \{v_j | e_{ij} \in E\} \cup \{v_j | e_{ji} \in E\}$. Rozšířené sousedství zahrnuje také přímé sousedy sousedů: $N_i^* = N_i \cup \bigcup_{v_j \in N_i} N_j$.

Lokální síť Lokální síť $G_i = (V_i, E_i, L_i)$ uzlu $v_i \in V$ je orientovaný, označený graf rozšířeného sousedství uzlu v_i. Množina uzlů je definovaná jako $V_i = N_i^*$, hrany jako $E_i = \{e_{jk} \in E | (v_j, v_k) \in N_i^* \times N_i^*\}$ a popisky jako $L_i = \{l(e_{jk}) | e_{jk} \in E_i\}$.

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

[76] Poslední prvek L představuje odlišnost oproti definici uváděné v textu výše. V principu používání grafů a celkovém smyslu definice však nedochází k žádným změnám v souvislosti s jeho zavedením.

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

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. Konkrétní kvantitativní vyjádření počtu, rozmístění a topologie jsou dále podrobně rozebrána v následujících částech této kapitoly. Navíc jsou do grafové struktury a jejího vyhodnocení zařazeny typy chyb a kvalitativní aspekty z předchozí podkapitoly.

Kapitola 3

Rešerše

Hlavním úkolem této kapitoly je začlenit popisovaný výzkum hodnocení kvality identických vazeb propojených prostorových dat do širšího kontextu souvisejících aktivit. Rešerše je rozdělená do dvou částí. První z nich se věnuje obecně problematice hodnocení kvality propojených dat, především s ohledem na vlastnosti popisující ekvivalentní vztahy mezi objekty. Druhá podkapitola je pak věnována výčtu dílčích metrik, které umožňují exaktně popisovat vlastnosti grafových struktur, které reprezentují identické vazby propojených dat.

Kvalita propojených dat

O tom, že hodnocení obecné kvality dat je velice aktuální i v oblasti propojených dat, svědčí řada publikací, jako například [11] (především kapitola 6.3.5), [53], [54], [55], [56] nebo [57] (zaměřeno především na otázky neurčitosti). Existuje také řada článků orientovaných přímo na téma kvality vazeb v propojených datech (většina z nich je citována v dalším textu práce), jako jsou například

- Quality Assessment for Linked Open Data: A Survey [12],
- Crowdsourcing Linked Data quality assessment [13],
- On the Likelihood of an Equivalence [58],

- owl: sameAs and Linked Data: An empirical study [14],
- Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora [47],
- SameAs networks and beyond: analyzing deployment status and implications of owl: sameAs in linked data [15].

Kvalita propojených dat s ohledem na identické a podobnostní vazby vychází nejen z výhod Linked Data přístupu, ale i z rizik a problémů, které s sebou nesou vlastnosti typu „Identity Links“. Například článek When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web [36] poukazuje na fakt, že vlastnost `owl:sameAs` může být chápána různými způsoby – stejný objekt v různém kontextu, stejný objekt různě vymezený, objekt a jeho reprezentace a také velmi podobné objekty (viz tabulka 1), z čehož mohou vyplynout komunikační problémy. Stejný článek hovoří o tzv. krizi identity v Linked Data. Totožný termín je použitý přímo v názvu článku [59], kde je jako návrh řešení představený systém OKKAM3 – služba pro podporu transparentní integrace znalostí o entitách.

Je však nutné si uvědomit, a autoři jako Halpin [36] nebo Wood [10] to také naznačují, že problém nespočívá v Linked Data přístupu, který nemůže být brán jako samospasitelný[78]. Hlavní problém z hlediska tradičního chápání kvality dat spočívá podle [35] v decentralizaci a samosprávnosti webu jako celku. Propojení jednotlivých prvků může kvalitu zvýšit, ale také ji degradovat. Jedním z cílů této práce je právě upozornit na situace, kdy využívání vazeb může uživatele „poškodit“ a naznačit cesty (spojené s tvorbou a korektním využíváním identických vazeb), které umožní tomuto problému se vyhnout. Další nedostatek propojování ekvivalentních datových i slovníkových objektů tkví ve faktu, že pro vyjádření jednoznačné identity není možné stoprocentně použít ani jednu z existujících forem popisu entit a fenoménů, které podporují explicitní sémantiku. O nedostatku expresivity (různých způsobů vyjádření) se zmiňuje také dokument [60]. V článku [15] jsou zmíněny vybrané publikace [61], [36] nebo [62], které rozebírají způsoby využívání relace `owl:sameAs`, včetně její symetričnosti a tranzitivity. Například články [61] nebo [20] popisují problém co-reference (vzájemného odkazování identických objektů a rizik, která jsou s

[78] V originále „silver bullet“.

tímto procesem spojena) jako jednu z hlavních oblastí výzkumu výskytu chyb v oblasti propojených dat.

Je tedy jasné, že jakékoli modely, tedy i slovníky používající Linked Data, jsou do jisté míry omezené. Na příkladu vlastnosti `owl:sameAs` to ukazuje i dokument [36]. Ačkoli ve standardu OWL má vlastnost `owl:sameAs` jednoznačnou definici „indicates that two URI references actually refer to the same thing: the individuals have the same "identity"", její reálné používání může být mnohem různorodější a často chybné. Zdroj [15] poznamenává, že reálné používání vlastnosti `owl:sameAs` je značně subjektivní, Sleeman [63] ji dokonce z mnoha důvodů uvedených v této části textu považuje za nebezpečnou. Následující tabulka (Tabulka 1) vytvořená podle [36] ukazuje různá chápání vazby `owl:sameAs`.

Tabulka 1: Způsoby chápání vazby `owl:sameAs`.

Způsob chápání `owl:sameAs`	Korektní vyjádření
Stejný prvek, ale v jiném kontextu - příkladem může být propojení dvou stejně se jmenujících prvků (Velká Británie jako stát a jako ostrov).	Využívat RDF konstrukt pro popis kontextu tzv. named graphs.
Stejný prvek, ale nesrozumitelně nebo nepřesně referencovaný - například územní s celky s různým vymezením.	V případě, že oba prvky jsou skutečně propojitelné, je vhodnější zvolit nějakou topologickou, mereologickou nebo podobnostní vazbu.
Reprezentace - například propojení emailové adresy a konkrétní osoby, kterou adresa reprezentuje.	Pokusit se najít nebo vytvořit vlastnost, která by vztah vystihovala přesněji, například v příkladu z vedlejšího sloupce by se mohlo jednat o vlastnost `foaf:mbox`.

Způsob chápání `owl:sameAs`	Korektní vyjádření
Vysoká podobnost (může svým způsobem zahrnovat všechny výše uvedené příklady)	Pro tyto účely mohou být použity některé vlastnosti ze standardu SKOS, jako například `skos:broaderMatch`, `skos:narrowerMatch` nebo `skos:closeMatch`.

Otázky kvality identických a podobnostních vazeb v rámci propojených dat se nemohou zužovat pouze na význam jednotlivých vlastností určujících tyto vazby. Článek [18] zmiňuje jako značný problém i nalezení skutečně identických prvků, které se na první pohled zdá být velmi jednoduché, ale ve skutečnosti jsou potřeba určité, poměrně rozsáhlé znalosti dané domény[79]. Pro automatické extrahování subjektů a objektů spojených identickou relací navrhuje Hogan [18] hledání shodných (nebo alespoň velice podobných) párů vlastností a příslušných hodnot pro jednotlivé výskyty instancí objektů reálného světa v různých datových sadách a znalostních bázích.

Podobně také článek Not Quite the Same: Identity Constraints for the Web of Linked Data [34] ukazuje, že signifikantní množství `owl:sameAs` vazeb na webu neodpovídá oficiálnímu významu propojených objektů. Tento článek také definuje kritéria (reflexivitu a nerozlišitelnost) pro identitu vycházející z Leibnitzova zákona identity [64]. Vedle těchto striktních pravidel jsou v publikaci [34] publikována i kritéria pro blízkou identitu a podobnost. Tyto úrovně vztahu jsou podle autora ovlivněny především kontextem, metodou pro zjišťování shody a lidským faktorem (úrovní expertních znalostí).

Publikace zmíněné v předchozích dvou odstavcích se zaměřují na popis a evaluaci používání vlastnosti `owl:sameAs` (jako dominantního predikátu identických vazeb), ale cílí především na statické a probabilistické vyhodnocování velkého vzorku dat. Tento přístup na jedné straně přináší zajímavé výsledky, ale vzhledem k decentralizaci Linked Data se získané

[79] O tom se autor přesvědčil při analýze geografických objektů a konceptů tezauru AGROVOC, kterou realizoval v první polovině roku 2014 pro FAO.

výsledky velice obtížně implementují, a tudíž slouží jako jisté varování nebo úvod do detailnějších studií.

Článek Assessing Linked Data Mappings using Network Measures [35] se podobně jako tato práce zabývá kvantitativním hodnocením propojených dat (na rozdíl od tohoto textu se nezabývá pouze daty prostorovými). Pro testování a určování kvality používá grafové struktury a metriky pro jejich popis. Konkrétně jsou v textu představeny stupně uzlu, centralita, clustering coeficient, Open SameAs řetězce (chains) a Description Richness (podrobnější popis těch metod využitých v této práci je k dispozici v následující části rešerše a v kapitole Metriky), přičemž první tři metriky se zaměřují na robustnost grafu (odolnost proti náhodným chybám) a další tři na fragmentaci sítě.

Pro popis identických vazeb a jejich následné hodnocení jsou v této práci použity grafové struktury a příslušné metriky. Podobná řešení jsou uvedena například v článcích [65] a [35] (postupy uvedené v tomto článku jsou detailně rozebírány v následující části Metodika, především v podkapitole Metodika a metriky), [66] nebo [15], kde jsou zmíněné strukturální parametry grafů (velikost grafu, průměr grafu nebo stupně uzlů) a sémantické vlastnosti jako je reflexivita, symetrie nebo tranzitivita.

Přehled metrik

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 uvedena v následujícím textu[80]) ukazuje řadu vědních disciplín a oborů lidské činnosti, kde tyto postupy nalézají své uplatnění. Jako příklady lze jmenovat sociologii [73], informační vědy [74], literaturu [75], biologii [70], medicínu [76], zdravotní péči [77], sport [78, 79] a především v současnosti velmi moderní sociální sítě[81] [67, 72, 80–85].

[80] 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 [67–72] nebo metody využívající aktivní chování prvků sítě [67] nebo velmi specifické biologické metriky [70].

[81] Původní výzkum sociálních sítí nebyl zaměřený na služby typu Facebook, ale na vztahy ve společnosti obecně.

Prvky grafu

Nejjednoduššími metrikami (tzv. strukturálními vlastnostmi sítí) jsou počet základních komponent grafu, tedy uzlů (size, network size) a hran (ties). S těmito hodnotami pracují například analýzy [77, 84, 86, 87]. Počty se mohou vztahovat i na specifické typy uzlů – izolované, kořenové, listové a vnitřní vrcholy (viz část věnovaná stupni uzlů). 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) [69, 74, 77, 79, 84, 86, 87], 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 [67] popisována jako průměrný počet propojení mezi dvěma prvky grafu.

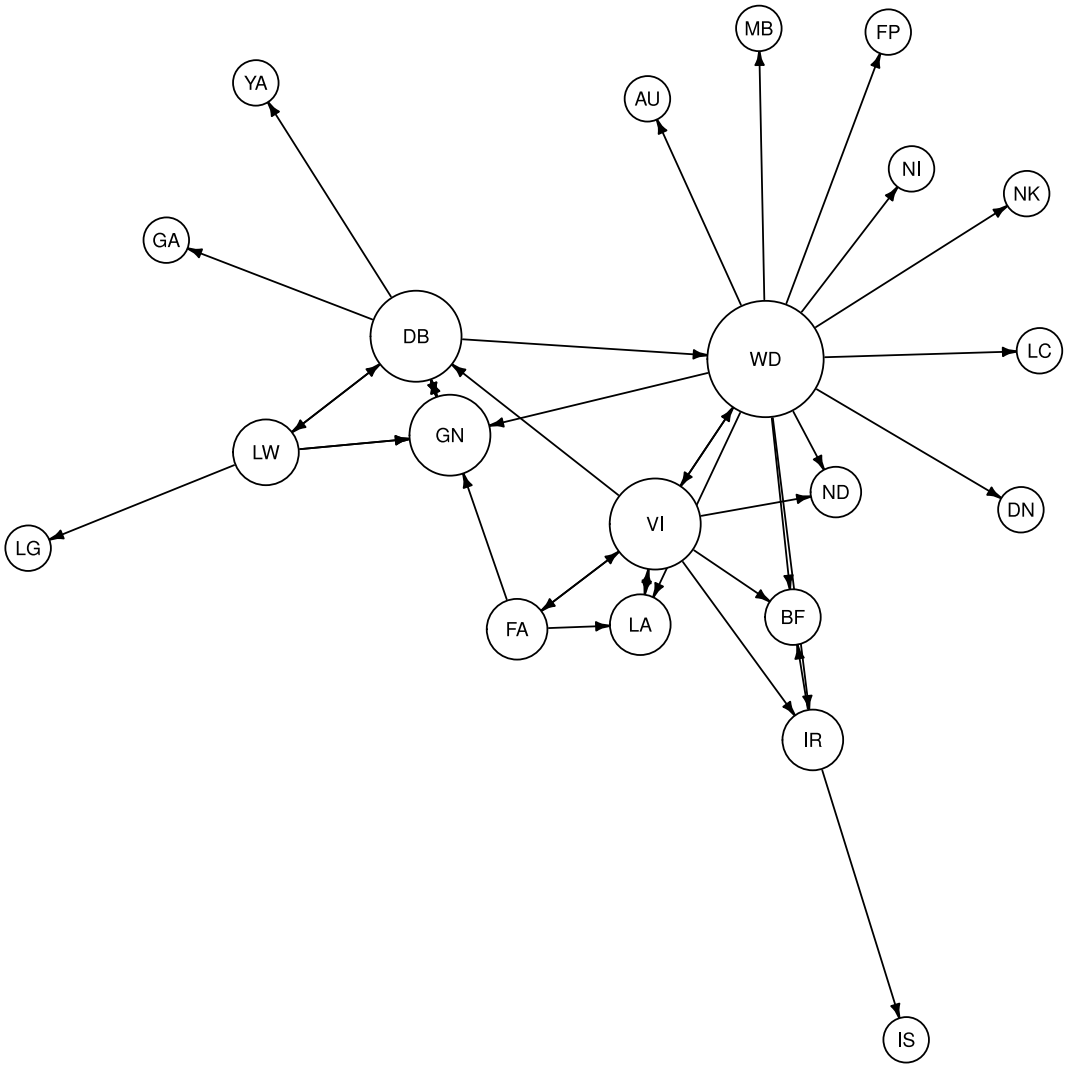
Stupeň uzlu

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 [77, 80, 84, 87–90], včetně sítí v prostředí webu i propojených dat, jako například [35, 91].

Publikace [49] 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.

Graf na obrázku 6 ukazuje normované stupně uzlů pro zdroje propojených

dat, které obsahují objekt `Prague`. Podobně jako v následujících grafech, které ilustrují jednotlivé metriky, jsou i v tomto případě velikosti (poloměry) kruhů, které vyjadřují datové zdroje, přímo úměrné vizualizované hodnotě.



Obrázek 6: Normovaný stupeň uzlu pro datové zdroje obsahující objekt `Prague`.

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^+[82]. Grafické srovnání hodnot typů stupňů uzlů pro datové zdroje obsahující ilustrační objekt `Prague` nabízí obrázek 7. Na základě hodnoty vstupních a

[82]Matematická definice celkového stupně jako součtu vstupních a výstupních stupňů uzlů je k dispozici v [35].

výstupních stupňů uzlu je možné vrchol označit jako

- izolovaný (isolated) – z něj ani do něj nevede žádná vazba (hodnoty obou stupňů jsou nula)[83],
- 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 [35] by cílem sledování stupňů uzlů a jejich následné modifikace (posilování, doplňování vazeb) mělo být přiblížení se k tzv. bezškálovým sítím[84] (scale-free networks, power law networks[85]), kde pro pravděpodobnost P nalezení uzlu stupně k platí $P(k) \sim k^{-\gamma}$, kde γ je parametr distribuce větší než 1. V experimentech zaměřených na provázanost různých částí sémantického webu [91] se hodnota γ pohybuje mezi 1,19 a 2,35[86]. Vztahem mocninné distribuce stupňů uzlů a vazeb v propojených datech se zabývá například [89]. Tyto sítě představují ideální variantu, která je díky existenci hubů mnohem odolnější (robustnější) vůči vlivům náhodných chyb.

V souvislosti se stupněm uzlu uvádí [15, 70, 78, 80, 88, 91] tzv. distribuci stupně uzlu (degree distribution, podle [91] Cumulative Distribution Function). Jedná se o graf, který na vodorovné ose zobrazuje hodnotu stupně uzlu. Na svislé ose jsou pak uvedeny počty uzlů s daným stupněm. Grafy je možné konstruovat pro celkový stupeň uzlu, ale také pro vstupní a výstupní uzly. [89] používá podobné grafické vyjádření, v němž jsou absolutní čísla nahrazeny logaritmickými hodnotami.

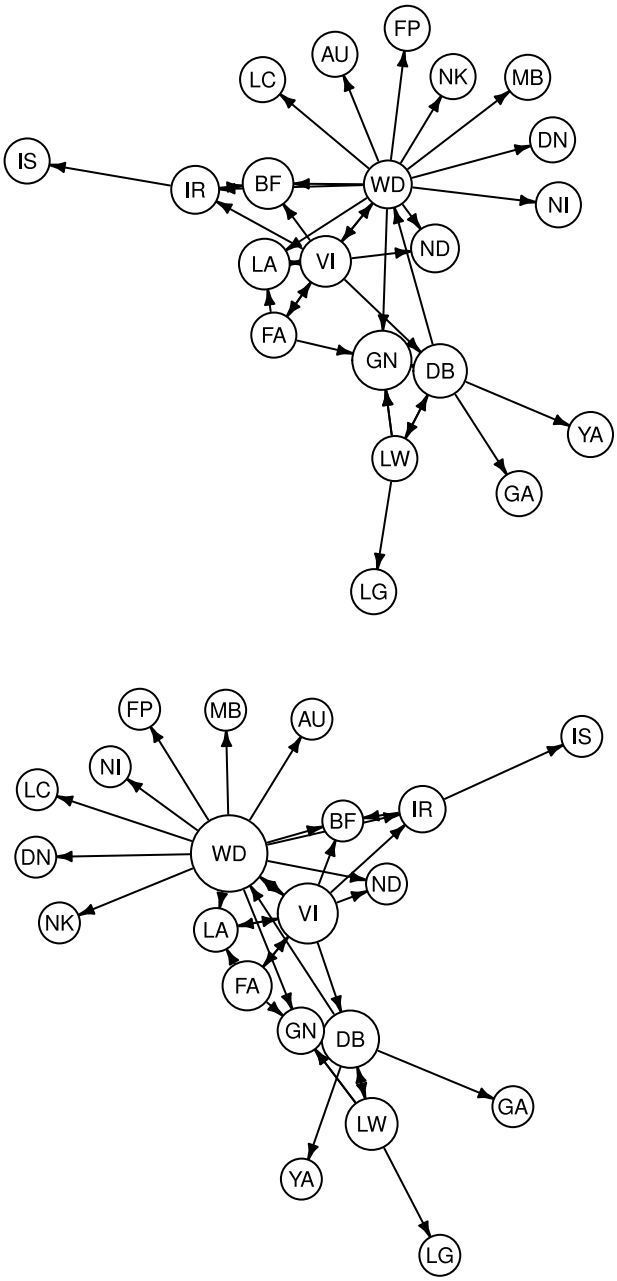
Síť grafu jako celek je možné hodnotit podle stupňů uzlů. V tomto případě je možné použít skóre (posloupnost) grafu, které je dáno sestupným seřazením jednotlivých stupňů vrcholů. Graf lze také popsat pomocí maximálního, minimálního nebo průměrného stupně uzlu.

[83] Izolované uzly jsou spojené s dalším parametrem – existencí sousedních uzlů[92].

[84] Více o bezškálových sítích viz [93] nebo [94].

[85] Podle [89] jsou na stejném principu založené i sítě jako například obecná topologie internetu (viz [95]).

[86] Článek Emergence of scaling in complex networks [96] uvádí hodnoty 2,5-3.



Obrázek 7: Normované hodnoty vstupních (nahoře) a výstupních (dole) stupňů uzlu pro datové zdroje obsahující objekt `Prague`.

Souvislost grafu

Dalším kritériem pro hodnocení ekvivalentních a podobnostních vazeb propojených dat je souvislost grafu (connectivity) [70]. Pokud bude graf souvislý, pak bude existovat propojení mezi všemi uzly, a tudíž bude možné propojit veškeré informace o každém konceptu. Pro zjištění souvislosti grafu je možné použít větu z publikace [49]: „Jsou-li stupně všech vrcholů v grafu G alespoň $n/2$, pak je G souvislý graf.“ Hodnota n znamená počet uzlů v grafů. Jak je patrné z jednotlivých případových studií, většina sítí obsahují izolované uzly, proto jsou převážně nesouvislé. Nebude-li graf obsahovat izolované vrcholy, pak bude minimálně slabě souvislý.

Vzdálenost

Další metriky 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 [76, 84, 90, 97].

$$l(G) = \frac{1}{n.(n-1)} \sum_{i \neq j} d(v_i, v_j),$$

kde n je počet uzlů grafu G a $d(v_i, v_j)$ je vzdálenost mezi uzly.

Inverzní hodnota vzdálenosti $l(G)$ se označuje jako Global Efficiency [76, 90, 98]. Global efficiency podává informaci o blízkosti uzlu k ostatním vrcholům sítě. Zdroje [76] a [90] 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 [78] nebo [84].

Centralita

Podle Guereta [32], [35] „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 [67–69, 71, 73, 76, 77, 97, 99–101] 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 [102]. Rolí centrality (jako jedné z metrik) na poli prostorových dat se zabývají například publikace [103] a [104].

V publikacích jsou nejčastěji zmiňovány tyto základní typy centrality [100], které budou 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 [72, 86, 87, 92], Barycenter Centrality [92], Information centrality [86], [87], Katz centrality [105, 106] nebo Reachability centrality [86], [87].

Následující matematické vzorce ilustrují tři základní typy centrality uzlu v, který je součástí orientovaného grafu $G = (V, E)$, kde V je množina všech uzlů v a E je množina všech hran e.

Centralita stupně C_d [69, 72, 73, 75, 82, 86, 100, 107–109] se určuje jako stupeň uzlu (viz výše; Obrázek 6)[87]. V případě orientovaných grafů lze hovořit o tzv. outdegree C_{od} a indegree C_{id} centrality [77], [75] (Obrázek 7).

[87]Rozšíření definic centrality a příklady různých způsobů výpočtů jsou k dispozici například v [110] nebo [107].

$$C_d(v) = \sum d_G(v)$$

$$C_{od}(v) = \sum d_G^+(v)$$

$$C_{id}(v) = \sum d_G^-(v)$$

Práce [111] 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 [112] a použitý v analýzách publikovaných v [71] nebo [72].

Centralita blízkosti [69, 72, 73, 82, 85–87, 92, 100, 107] 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[88]. 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 [111] 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.

[88]Poznámka autora: U tohoto typu hodnocení by bylo zajímavé zjistit nejen průměrnou hodnotu, ale také směrodatnou odchylku, aby byla získána informace o tom, zda si jsou hodnoty jednotlivých vzdáleností podobné, nebo zda dochází ke zprůměrování extrémů.

Centralita mezilehlosti [32, 35, 72, 73, 77, 82, 85, 87, 92, 100, 107, 113] hodnotí uzly z hlediska „mezilehlosti“, tj. specifické polohy, kdy daným uzlem prochází velké množství cest mezi ostatními uzly [99]. Vysoké hodnoty centrality mezilehlosti indikují, že daný uzel tvoří „most“ uvnitř grafu, to znamená, že propojují do jisté míry samostatné nebo izolované podgrafy.

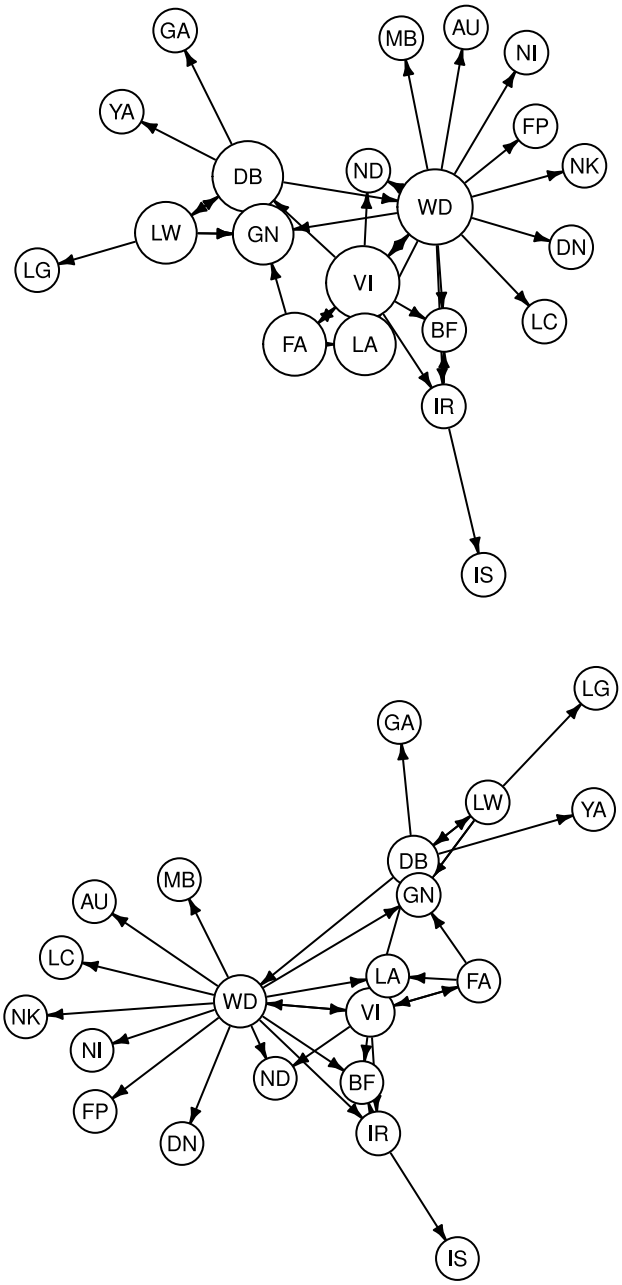
$$C_b(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

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 [35] 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 [32] a [35] 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).

Obrázek 8 porovnává centrality blízkosti (horní schéma) a mezilehlosti (dole) pro zdroje propojených dat, které obsahují reprezentace prostorového objektu `Prague`. Podobně jako u následující schémat (například obrázky 9 a 10 je velikost kruhu vyjadřujícího zdroj dat přímo úměrná hodnotě sledované metriky. Na první pohled jsou vidět odlišnosti obou hodnot, kdy hodnoty centrality blízkosti jsou poměrně diverzifikované (minimální vzdálenost mezi jednotlivými uzly nabývá různých hodnot), zatímco velikost normalizovaných hodnot centrality mezilehlosti si jsou velice podobné (v grafu neexistují výrazně oddělené komponenty).



Obrázek 8: Grafy ukazující rozdíly v centralitě blízkosti (nahoře) a centralitě mezilehlosti (dole) pro datové zdroje obsahující objekt `Prague`.

Autority a středy

Autority (authorities) a středy (hubs) [75, 82, 85, 92] představují dva specifické typy vrcholů grafu. Autorita je takový uzel, k němuž směřuje velké množství vazeb. Má tedy vysokou hodnotu vstupního stupně d_G^-. Naopak střed je vrcholem, z něhož vychází velké množství propojení směrem ke zbytku sítě. V tomto případě je naopak vysoká hodnota výstupního stupně uzlu d_G^+. Oba pojmy jsou úzce spojené s algoritmem HITS [114]. Tento iterativní algoritmus umožňující výpočet tzv. authority score $A(v)$ and hub score $H(v)$ vychází z předpokladu ideální sítě, v níž kvalitní autorita je odkazována z mnoha hubů a naopak hodnotný střed je propojený na mnoho vrcholů typu autorita. Pro každý uzel v platí vztahy

$$A(v) = \sum_y H(y),$$

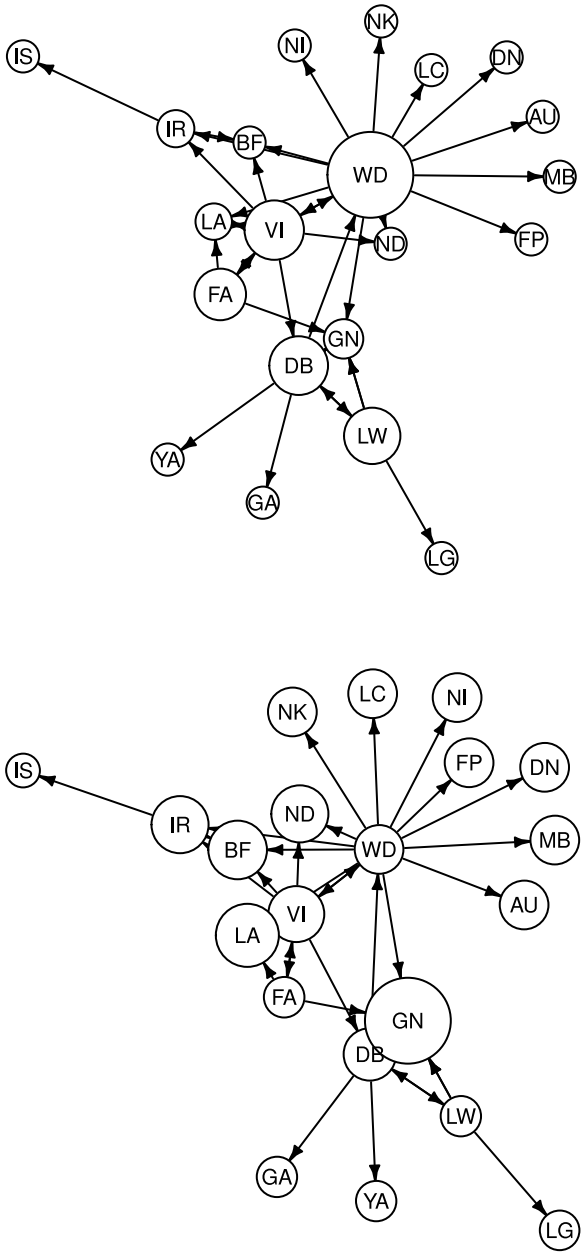
$$H(v) = \sum_y A(y),$$

kde y jsou uzly přímo sousedící s vrcholem v.

Na obrázku 9 jsou graficky vyjádřeny středy (nahoře) a autority (dole) pro zdroje reprezentující prostorový objekt `Prague`. Čím vyšší je hodnota příslušného parametru, tím větší je velikost uzlu v grafu. Porovnáním obou metrik je patrné, že se jedná o komplementární metody hodnocení vrcholů v grafu.

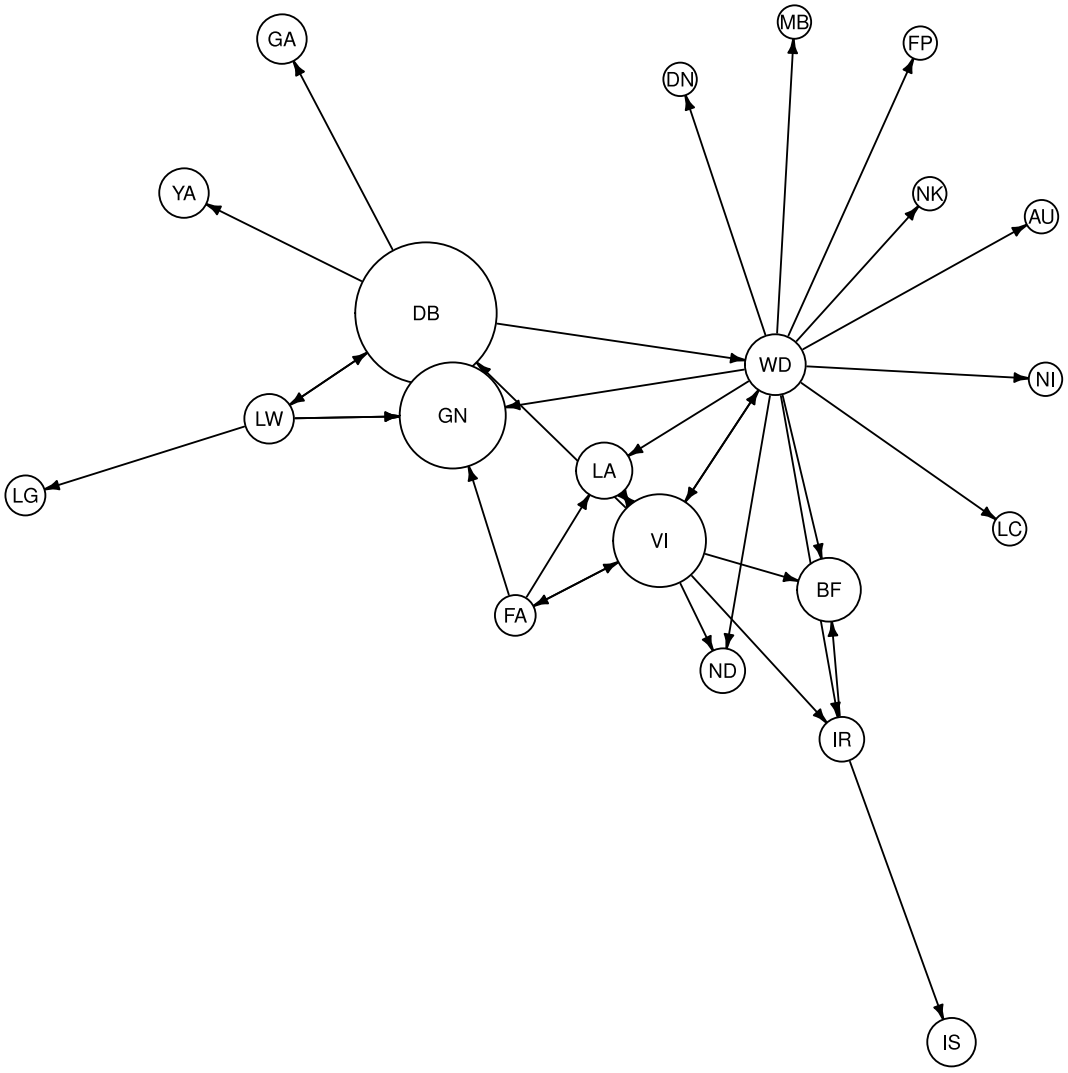
Page Rank

Podobně jako v předchozím případě i zde se jedná o iterační algoritmus (navržený Larry Pagem a Sergeyem Brinem) počítající významnost jednotlivých vrcholů v síti. Základní princip algoritmu PageRank [115] spočívá v tom, že pro každý uzel se zjišťuje počet a také významnost (z pohledu PageRank) vrcholů, které jsou s ním přímo spojeny (příklad na obrázku 10, kde je velikost kruhů vyjadřujících uzly přímo úměrná hodnotě



Obrázek 9: Středy (nahoře) a autority (dole) mezi zdroji obsahujícími objekt `Prague`.

koeficientu Page Rank). Pro síťové analýzy je tento algoritmus používán jako metrika například v publikacích [82, 83, 85, 92].



Obrázek 10: Page Rank skóre zdrojů obsahujících objekt `Prague`.

Se zjišťováním významu vrcholů souvisí i další méně rozšířené metriky publikované v [92]:

- Important Neighbor Proportion – procento významných uzlů mezi přímými sousedy.
- Unknown Neighbor 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

Shlukový koeficient (koeficient shlukování, Clustering coefficient) je publikován například v dokumentech [35, 69, 77, 78, 84, 90, 91, 116–118]. Tento lokální koeficient popisuje uzavřenost částí grafu a hustotu sítě kolem konkrétního uzlu. Podle publikací [116] i [35] se koeficient pro jednotlivé uzly grafu počítá jako poměr vazby mezi sousedy uzlu a maximálním možným počtem vazeb mezi těmito sousedy:

1. Pro daný uzel se zjistí počet sousedů.
2. Vypočítá se maximální možný počet vazeb mezi těmito sousedy (dvoučlenná variace bez opakování – $n \times (n-1)$, kde n je počet přímo sousedících uzlů).
3. Identifikuje se reálný počet propojení mezi sousedy.
4. Clustering coefficient uzlu je podíl mezi hodnotami zjištěnými v krocích 3 a 2.

Celkový shlukový koeficient grafu se vypočítá jako průměrný koeficient všech uzlů grafu [84]. Podle [35] 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 [119] nebo [120]. Ideální hodnota koeficientu je rovna 1. Této hodnoty je dosaženo, když je spojený každý uzel s každým. Gil [91] poznamenává, že „graf blízký ideálu“ by měl mít hodnotu tohoto koeficientu výrazně vyšší než náhodně vytvořený graf. Gueret [35] vypočítává vzdálenost reálného koeficientu grafu (číslo menší nebo rovno jedné) od ideálního stavu jako rozdíl mezi oběma hodnotami.

Se shluky (klastry) souvisí také metrika nazývaná Number of Connected Components [84]. Ta představuje počet zřetelných klastrů obsažených v grafu.

Sítě jednotlivého aktéra

Sítě jednotlivého aktéra (tzv. ego networks) [121–123] představují podgrafy vybraného uzlu, jeho přímých sousedů a jejich vzájemných propojení. Tento typ grafové struktury se používá především v oblasti sociálních sítí. Podle publikací [86] a [87] 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[89].
- 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 [69] je tento parametr úzce spojený s centralitou mezilehlosti.

Reciprocita

Tato vlastnost je důležitá především pro orientované grafy [67, 101, 124]. Reciprocita vyjadřuje míru vzájemného propojení mezi dvojicemi uzlů. Vyjadřuje se zpravidla jako poměr mezi počtem vzájemně propojených uzlů a všech propojených uzlů [124]. Další možnosti výpočtu reciprocity jsou k dispozici na příklad v publikaci [125].

Indexy používané v geografii dopravy

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

[89]Tato hodnota se normalizuje vydělením počtem uzlů – jedná se o tzv. Reach Efficiency [86, 87].

řeší síťové analýzy a dostupnost jednotlivých sídel (uzlů) v rámci komunikační sítě.

Podle publikace [126] je základním přístupem pro hodnocení dostupnosti (accessibility) a konektivity (connectivity) tzv. matice konektivity (connectivity matrix). Matice je čtvercová, přičemž počet řádků a sloupců závisí na počtu uzlů v síti. V matici se vyskytují pouze hodnoty 1 (pro propojené prvky) a 0 pro ostatní prvky, včetně diagonály. V případě orientovaného grafu se jedná o matici asymetrickou. V terminologii používané v teorii grafů jde o matici sousednosti [127].

Zdroj [126] uvádí, že indexy jsou důležitější metody pro reprezentování strukturálních vlastností grafů než standardní statické hodnocení sítě. Navíc se jedná o typy veličin, které umožňují srovnání více grafů. Pro hodnocení sítí identických vazeb v Linked Data není vhodné používat indexy pracující s poměrem přímé a dopravní vzdálenosti (Detour index, deviatilita), přepravou (theta index) nebo plošnou rozlohu sítě (hustota sítě). Následující seznam indexů vychází z publikací [126, 128–130].

- Beta index (β) – představuje nejjednodušší metodu hodnocení dopravní sítě. Jedná se o podíl počtu hran (e) a počtu uzlů (v).

$$\beta = \frac{e}{v}$$

- Gama index[90] (γ) – jedná se o poměr mezi skutečným počtem spojení (hran, e) a maximálně možným počtem spojení. Maximálně možný počet hran v grafu se získá jako kombinace druhé třídy pro počet vrcholů (v). Hodnota se pohybuje mezi 0 a 1 a podle [126] indikuje kompletnost sítě. Podle stejného zdroje je tento typ indexu vhodný pro sledování časových změn v síti.

$$\gamma = \frac{e}{\binom{v}{2}} = \frac{e}{\frac{v!}{2!(v-2)!}} = \frac{2e}{v(v-1)}$$

[90]Výpočty Gama indexu a Alfa indexu se liší u planárních a neplanárních grafů. Pro účely verifikace grafů znázorňujících identické a podobnostní vazby v Linked Data byly zvoleny varianty pro neplanární grafy, protože většina většina sítí nemá charakter rovinného grafu.

- Alfa index (α) – 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 (e - počet hran, v - počet uzlů, p - počet podgrafů).

$$\alpha = \frac{2(e - v + p)}{(v - 1)(v - 2)}$$

- Eta index (η) – má význam pouze v případě ohodnoceného grafu, protože se vypočítá jako podíl celkové délky hran grafu ($L(G)$) a počtu hran (e).

$$\eta = \frac{L(G)}{e}$$

Kapitola 4

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 (viz kapitola Vyhledávání, sběr a formalizace informací o identických vazbách, kde jsou popsány i omezení vyplývající ze strojového testování kvality vazeb). Poté následuje nasazení metrik ukazující míru propojení geografických objektů nebo konceptů (viz kapitola Metriky).

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 (například hlavní města, evropské řeky, pojmy týkající se konkrétního vědního oboru apod.). 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).

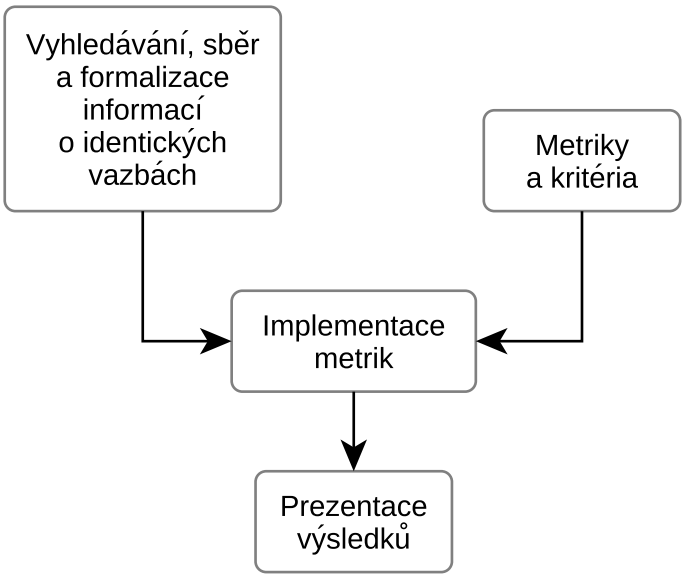
Je samozřejmě nutné poznamenat, že ani jeden z výše uvedených přístupů nelze aplikovat samostatně, protože mezi nimi dochází k výrazným průnikům. Také data a informace sbírané pro kvantitativní způsoby hodnocení, používání datových sítí (grafického znázornění vazeb a zdrojů) i jednotlivé metriky se dají využít pro testování objektů, zdrojů i vazeb.

Navržení způsobu hodnocení identických vazeb mezi objekty prostorových propojených dat je založené na čtyřech základních krocích (Obrázek 11), které lze formulovat pomocí jednoduchých otázek:

1. Jakým způsobem lze získat a formalizovat informace o identických vazbách?
2. Jaké aspekty kvality budou v hodnocení zohledněny?
3. Jak bude hodnocení probíhat? Jaké metody a jaká kritéria budou použity?
4. Jak budou výsledky prezentovány a interpretovány?

Následující schéma (Obrázek 11) ukazuje architekturu navrhovaného řešení, která je částečně inspirovaná pojetím v publikaci [35]. V článku [51] je publikovaný také velmi podobný postup skládající se z následujících kroků – extrakce dat, tvorba grafu, výpočet míry zkreslení informací, kvantifikace a komparace výsledků. Odpovědi na výše uvedené otázky se skrývají v jednotlivých uzlech schématu a také v částech textu Metodika, Experimenty a Výsledky. Jednotlivé komponenty grafu (s výjimkou prezentace výsledků) jsou dále detailně popsány v následujících sekcích této kapitoly.

V první podkapitole této části je publikovaná problematika sběru a prvotního zpracování (formalizace) dat o identických vazbách. Konkrétně se jedná o postup sběru dat (přístup „Follow Your Nose“), včetně popisu skriptů, identifikaci chyb vazeb a zdrojů (aspekty kvality) a také volba vhodných



Obrázek 11: Architektura procesu hodnocení identických vazeb prostorových propojených dat.

zdrojů, od nichž začíná prohledávání sítě propojených dat. Poté následuje popis jednotlivých metrik pocházejících z různých vědních oborů a jejich kompozice do jednotné metodiky. V této pasáži je zohledňován také vhodný výběr parametrů pro metriky i zhodnocení vhodnosti metrik pro konkrétní účely. Třetí podkapitola je zaměřená na implementaci metodiky (souhrny hodnot metrik, normalizace číselných údajů z důvodů porovnávání jednotlivých sledovaných geografických objektů a jejich skupin). Interpretace výsledků je součástí následujících dvou kapitol (Experimenty a Výsledky). Získané informace jsou použity jako zpětná vazba pro zpřesnění metodiky (vhodných metrik, jejich kombinace a nastavení parametrů).

Vyhledávání, sběr a formalizace informací o identických vazbách

Sběr dat

Data 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[91][92]
3. Automaticky pomocí skriptu vytvořeného v rámci této práce

Zásadní rozdíly jsou především v množství získaných údajů, rychlosti jejich získání a v kvalitě získaných informací. Automatické služby jsou výrazně rychlejší (je-li pominut čas nutný k vytvoření takové služby), ale získané výsledky nejsou zcela uspokojivé (malý počet a nízká kvalita oproti manuálnímu prohledávání zkušeným expertem). Navíc v případě existujících služeb není možné ovlivnit seznam prohledávaných datových zdrojů, který je často poplatný době vzniku takového produktu. Na druhou stranu automatické prostředky jsou schopné hromadně zpracovat větší množství dat. U jednotlivých případových studií v kapitole Případové studie bude vždy (s výjimkou prvního experimentu, kde budou všechny přístupy sběru dat porovnány na konkrétních datech) použitý **způsob sběru pomocí skriptu**, který data nejen z příslušných zdrojů stahuje, ale převádí je do formalizované podoby (datové soubory v XML a CSV).

[91] http://sameas.org/

[92] Dále byly na základě informací ze zdroje [18] zkoumány služby a nástroje Sig.Ma, rkbexplorer nebo ObjectCoref. Dále byly bez většího úspěchu (funkčnost nebyla dostačující pro potřeby tohoto výzkumu) testovány nástroje jako Silk, LDspider [131], Any23 nebo Falcons.

Aspekty kvality vazeb

V následující části jsou typy chyb identických a podobnostních vazeb ilustrovány na modelovém příkladu (elementární krok popisující vazbu mezi dvěma výskyty jednoho geografického objektu ve dvou sadách propojených prostorových dat), který vychází z následujících předpokladů:

1. Existují datové nebo znalostní báze `D1` a `D2`, které obsahují propojená data.
2. Existují instance `O1` a `O2`, které reprezentují stejný reálný nebo abstraktní prostorový objekt.
3. Objekt `O1` je součástí datové báze `D1` a objekt `O2` náleží do datové sady `D2`.
4. Objekty jsou `O1` a `O2` jsou propojeny identickou vazbou[93], takže tvoří RDF trojici, kde
   - `O1` je subjektem,
   - identická vazba je predikátem,
   - `O2`je objektem.
5. Objekt `O1` (subjekt RDF trojice) je považován za správný. To znamená, že případné chyby se budou týkat pouze predikátu a/nebo subjektu.

Případné chyby, které ovlivňují vazbu mezi instancemi reprezentujícími geografické objekty, lze rozdělit na chyby[94]

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 (resp. ji zapříčinil správce, tvůrce, administrátor nebo editor datové sady `D1`) při zápisu identické vazby a jejího cíle. Může se jednat například o drobnou vadu (například záměnu písmen) v názvu vazby nebo identifikátoru objektu `O2`. Syntaktická chyba se tedy týká jak objektu (neboli uzlu v grafickém vyjádření, které je popsáno v následující části dokumentu),

[93]Teoreticky nemusí být typ vazby nutně omezován pouze na identitu a podobnost, ale vzhledem k zaměření tohoto textu se jedná o výše uvedený typ propojení dat.

[94]V následujícím textu a především ve zpracovávaných datech jsou typy chyb označovány zkratkami X - syntaktická chyba, T - technická chyba a S - sémantická chyba.

tak predikátu (hrany grafu). Je nutné si uvědomit, že identifikace takové chyby je velice obtížná, protože ve většině případů nebude možné rozlišit, zda došlo k omylu při zápisu URI nebo zda identifikátor objektu relace nebyl perzistentní. Podobně bude také velice složité (i když jednodušší než v předchozím případu) oddělit, kdy je vazba nestandardní nebo syntakticky nesprávně zapsaná. Z těchto důvodů nejsou syntaktické chyby v jednotlivých částech experimentů uvedených v této práci evidovány (to se týká především dat o identických relacích sbíraných pomocí skriptů).

technické Chyba se může projevit nefunkčností (nedostupností) objektu `O2`. Jinými slovy se nepodařilo získat cíl vazby. Mezi technické chyby je možné řadit situace, kdy

- Datová sada `S2` není v době výzkumu nebo v době potřeby získání dat dostupná, funkční (ať už zcela nebo dočasně), případně byla zcela odstraněna (neexistuje).
- Objekt `O2` byl z datové sady `S2` odstraněný[95].
- Došlo ke změně identifikátoru objektu `O2` v datové sadě `S2` (identifikátor nebyl perzistentní).

sémantické Chyba má za následek nedostatečné (nekorektní) fungování vazby. Nalezené výsledky mohou být například nepřesné nebo nejsou jednoznačné. V případě sémantických chyb je však určení místa projevení chyby (nebo zodpovědnosti za chybu) velice problematické. Na základě dostupných informací většinou nelze určit, zda se chyba prvně objevila ve vazbě (a tudíž na straně vstupní datové sady) nebo na úrovni cíle vazby (objektu, výstupní datové sady). Proto v dalším textu bude popsána sémantická chyba, která se (pokud nebude uvedeno jinak) bude týkat kombinace predikátu a objektu RDF trojice, která představuje identickou nebo prostorovou objektovou vazbu. Mezi sémantické nedostatky identických vazeb patří například následující případy:

[95] V tomto případě může nastat situace, že objekt v datové sadě nikdy neexistoval. V tomto značně hypotetickém případě byla vazba špatně zavedena, takže se jedná spíše o sémantickou chybu. V případě chyby v zápisu URI objektu vazby jde o chybu syntaktickou.

- Vazba neodpovídá skutečnému vztahu objektu a subjektu. V tomto případě je nutné akceptovat především sémantiku (význam) vycházející z explicitní definice vazby a explicitní (formálně popsané) sémantiky objektu a subjektu, nikoli z pohledu (názoru) uživatele nebo tvůrce datové sady (implicitní, nesdílená sémantika). Příkladem může být použití identické vazby pro podobné objekty. Zde je nutné si uvědomit, že ačkoli vlastnosti jako například `skos:exactMatch` a `owl:sameAs` jsou označovány jako identické nebo ekvivalentní, „síla vyjádření ekvivalence“ v definicích obou vlastností je rozdílná. Neodůvodněná kombinace různých vlastností by se dala označit jako budování vazeb na základě nejednotného logické principu. Tento typ chyby je popsán také v publikaci [36] jako propojování stejných objektů v různém kontextu, vazby mezi podobnými a nikoli identickými objekty nebo linky mezi nejasně popsanými objekty, které na základě dostupných informací vypadají jako ekvivalentní.
- Objekt RDF trojice (cíl vazby) neodpovídá subjektu a vazbě (například propojení geografického objektu `Česká republika` a termínu `Administrativní členění České republiky` identickou vazbou). V tomto případě však jde nejen o sémantiku (podobně jako v předchozím bodu), ale tento problém se týká i dereferencovatelnosti objektu (dohledatelnosti dat a další možnosti jejich strojového zpracování pouze na základě znalosti identifikátoru ve formě URI). Jinými slovy tato chyba nastane, když identifikátor objektu vazby odkazuje nikoli na data, ale pouze na jejich reprezentaci například ve formě mapy nebo webové stránky nebo na data, která nejsou k dispozici v RDF/XML syntaxi jako základním tvaru pro propojená data.
- Není dostatek informací a vazbě (predikátu), protože se jedná o nestandardizované (ani explicitně nepopsané, případně uzavřené) řešení. V tomto případě nejde o fatální chybu, která by znemožnila používání propojených dat. Používání standardů však usnadňuje implementaci dat do aplikací, automatické vytěžování Linked Data zdrojů a strojové zpracování těchto dat.

- Poslední možnou sémantickou chybou, která se může v propojených datech vyskytovat, je absence vazby. Tedy oba objekty `O1` a `O2` jsou sice identické nebo velmi podobné, ale nejsou spojeny příslušnou vazbou, ačkoli z hlediska sémantiky by taková vazba měla existovat. Jedná se o nedostatek v konektivitě sítě propojených dat. Tento druh chyby je důležitý především pro proces mapování a propojování datových sad. Tento nedostatek však zároveň paradoxně představuje jednu ze silných stránek propojených dat a vazeb mezi nimi. Jedná se o již dříve zmíněnou sociální funkci, která v kombinaci s kontextovou (hlediskovou) sémantikou umožňuje prezentaci různých pohledů a názorů na jeden objekt. Proto tento druh chyby nebude v následných analýzách zmiňován, ani graficky vizualizován. Na druhou stranu, zjištění četnosti této chyby je velice jednoduché. Jde pouze o rozdíl mezi skutečným počtem hran v reálném grafu a počtem hran v úplném grafu se stejným počtem uzlů.

Při hodnocení chyb pomocí různých metrik (viz následující část této kapitoly) budou zohledněny nejen výše uvedené typy chyb, ale také následující pravidla, která deklarují jednotný přístup a zpracování informací o identických vazbách:

1. Technické a syntaktické chyby mají pro celou vazbu fatální následky. V případě sémantické chyby je většinou možné získat data pomocí identifikátoru objektu vazby, ale tato data mohou být nepřesná, nekorektní, obtížně využitelná nebo interpretovatelná (dokonce na základě jejich interpretace může docházet k mylným závěrům). Rozhodnutí, zda a případně jak taková data dále používat, je plně na straně uživatele (zpracovatele). Proto budou technické a syntaktické chyby chápány jako vážnější, což však nijak nezlehčuje důsledky vyplývající z potenciálních sémantických chyb.
2. Může nastat situace, kdy výsledek interpretace vazby není jednoznačný. Například, pokud existuje více různých typů vazeb mezi dvěma objekty – `O1` a `O2` jsou zároveň podobné i identické, což je sice z hlediska logiky možné, protože identita se dá chápat jako specifická (maximální) podobnost, ale dochází minimálně k redundantnímu zápisu. Nebo

naopak může jedna vazba propojovat více objektů – více identických objektů (často z jedné datové sady) spojených stejnou vazbou s jedním subjektem. Zvláště v případě prostorových dat (geografických objektů) by k takovému jevu nemělo docházet, protože díky informaci o poloze (v tomto případě samozřejmě záleží na podrobnosti a přesnosti této informace) téměř neexistují zcela identické objekty, včetně jejich polohy. Na druhou stranu, právě polohová informace je velice důležitá při analýze skutečně identických objektů nebo pouze objektů stejně pojmenovaných (například sídla `Praha` v České republice a na Slovensku, prvky `London` a `City of London`, které mohou, ale také nemusí být ekvivalentní, nebo geografické entity `Gruzie` a `Georgia`, která mají v angličtině shodné pojmenování, viz příslušná případová studie).

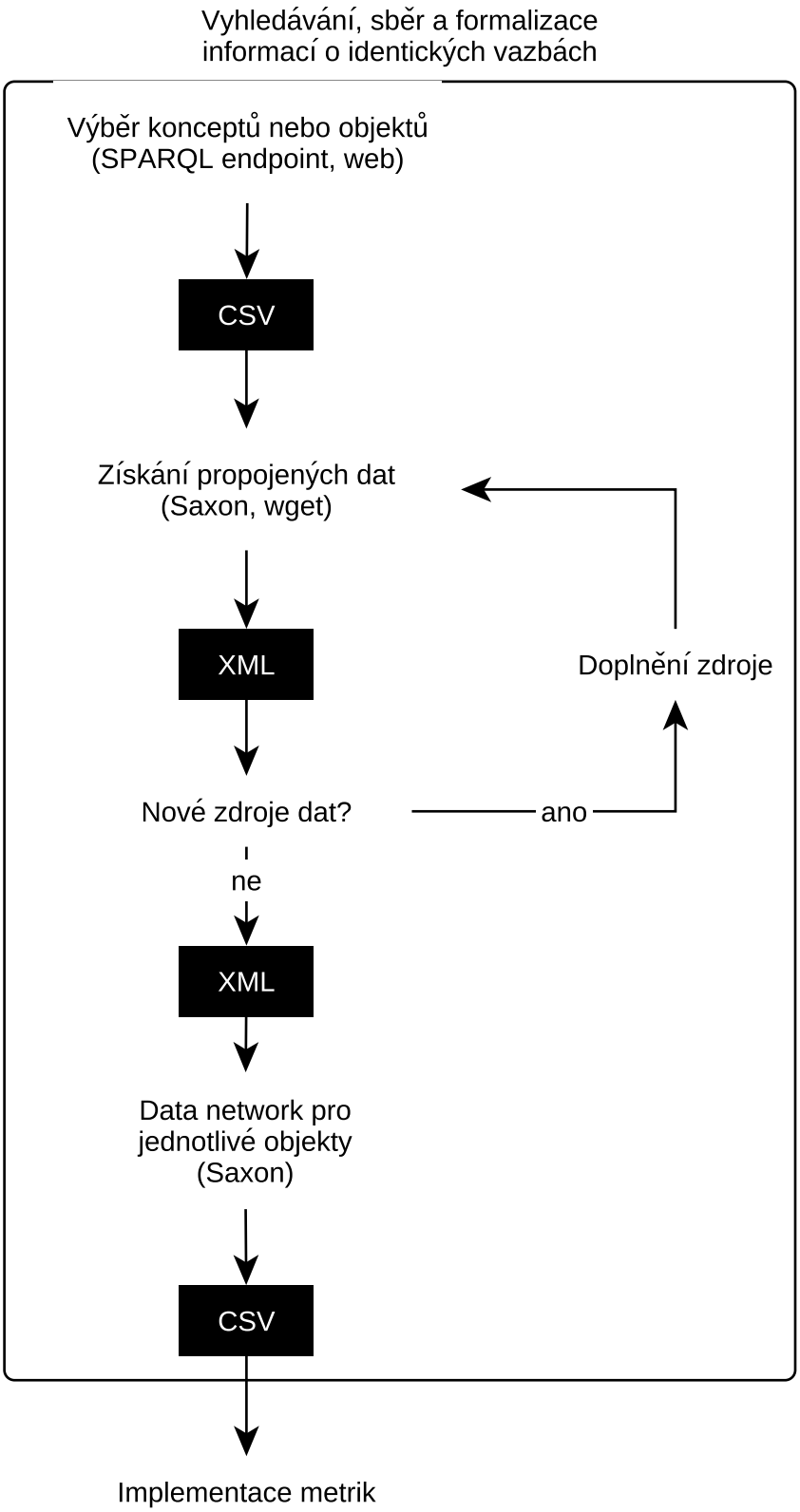
3. Ideální (bezchybnou, plně odpovídající požadavkům na propojená data) situací je stav, kdy se objekt (`O2`) podařilo pomocí vazby uvedené v predikátu RDF trojice získat, tento objekt odpovídá významu vazby a obsahuje strojově čitelná data.

Skript pro automatický sběr a předzpracování dat

Vyhledávání, sběr a formalizace informací o identických vazbách probíhá podle následujícího postupu (Obrázek 12), který je zpracován ve formě série skriptů v jazycích Bash s využitím prostředí UNIX shell (řídící proces celého postupu), XSLT a programovacího jazyka R. Na obrázku 12 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.

1. Nejprve jsou získána URI reprezentací geografických objektů ze zdroje, který je zvolen jako počáteční. Tímto zdrojem je ve většině případů znalostní báze DBpedia[96] [132], která je považována za centrum prostoru propojených dat (viz Linking Open Data cloud diagram[97]). V případě sběru většího množství dat se s výhodou využívají příslušné SPARQL

[96] dbpedia.org
[97] http://lod-cloud.net/



Obrázek 12: Architektura – vyhledávání, sběr a formalizace informací o identických vazbách.

endpointy (například DBpedia SPARQL endpoint[98], viz Příloha C). Identifikátory jsou uloženy v CSV souboru, který je dále zpracovaný pomocí XSLT skriptu s využitím nástrojů Saxon (XSLT procesor) a wget (program pro stahování dat pomocí protokolů HTTP, HTTPS a FTP).

Skript prohledává příslušné RDF soubory, v nichž nalézá identické vazby (v současné době jsou zpracovávány relace `rdfs:seeAlso`, `owl:sameAs`, `skos:exactMatch`, `skos:closeMatch` a vybrané Wikidata identifiers, viz následující fragment XSLT kódu).

```
<xsl:for-each select="//owl:sameAs|
//skos:exactMatch|//skos:closeMatch|
//rdfs:seeAlso|//wdt:P227|//wdt:P646|
//wdt:P349|//wdt:P214|//wdt:P244|
//wdt:P409|//wdt:P508|//wdt:P691|
//wdt:P906|//wdt:P949|//wdt:P950|
//wdt:P982|//wdt:P99|//wdt:P1014|
//wdt:P268|//wdt:P269|//wdt:P1566|
//wdt:P1670|//wdt:P1997|//wdt:P2163|
//wdt:P2503|//schema:sameAs">
```

Objekty identických vazeb jsou uloženy do dočasného souboru, který lze procházet v dalších krocích. Zároveň je vytvářený také druhý dočasný soubor obsahující již zpracované zdroje, aby nedocházelo k jejich opakovanému stahování a načítání. Na počátku procesu je možná volba hloubky procházení. Na základě experimentů se jako dostačující ukazuje hodnota 6, neboť v žádném dosavadním testu nedošlo k nalezení dosud nezpracovaných zdrojů v šesté úrovni.

2. V případě, že dojde k nalezení nového, dosud neznámého zdroje, který zatím není zapsaný v seznamu zdrojů (XML soubor), dojde k manuálnímu doplnění takového zdroje a novému spuštění skriptu.

3. Posledním krokem této fáze je vytvoření datového souboru ve formátu

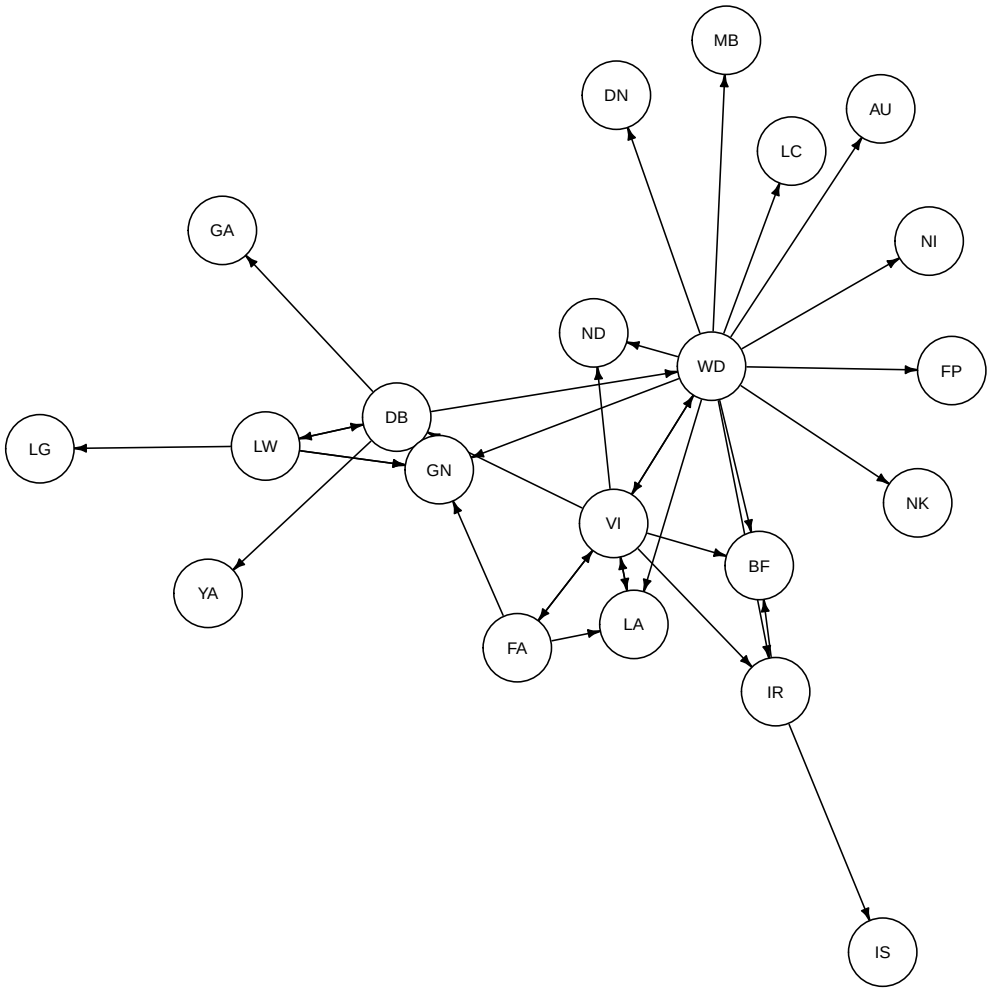
[98] http://dbpedia.org/sparql

XML, který je základem pro další zpracování, včetně konstrukce příslušného grafu. Následující ukázka představuje fragment datového souboru, který popisuje jednu vazbu typu `owl:sameAs` mezi dvěma reprezentacemi (v databázích Yago a DBpedia) geografického objektu Německo. Jednoduché schéma datového souboru obsahuje kořenový element `<term>` s atributem `name` a `date`, které udávají název zpracovávaného prvku a datum zpracování ve formátu RRRR-MM-DD (tento atribut je nepovinný). Každá identická nebo prostorová vazba mezi sémantickými nástroji nebo systémy pro organizaci znalostí je zaznamenaná v elementu `<link>`. V souladu s konstrukcí RDF trojic jsou součástí každého elementu `<link>` další vnořené elementy `<subject>`, `<predicate>` a `<object>`. Element `<subject>` představuje geografický objekt, ze kterého vazba vychází (v terminologii teorie grafů se jedná o výstupní uzel). Element `<object>` je cíl vazby (vstupní uzel). Oba tyto elementy mají dva atributy – `acronym-2` (dvoupísmenná zkratka zdroje, kompletní číselník je k dispozici v Příloze B) a `uri` (identifikátor geografického objektu; tento atribut je nepovinný). Uvedení objektu vazby také není povinné, protože je v některých případech potřeba evidovat také výskyty geografického prvku, i když ten nemá žádné vazby na externí zdroje. Následujícím prvkem (`<predicate>`), který se vyskytuje uvnitř elementu `<link>`, je označení typu vazby. Typ vazby je uložený v atributu `type`. Posledním elementem je označení případné chyby ve vazbě – element `<error>`. Chyba je popsána pomocí dvou atributů – `type` (typ chyby popsaný v části Metodika) a `note` (poznámka; nepovinný atribut).

```
<link>
<object acronym-2="YA"
uri="http://yago-knowledge.org/resource/
 Federal_Republic_of_Germany"/>
<subject acronym-2="DB" uri="http://dbpedia.org/
 data/Germany.rdf"/>
<predicate type="sameAs"/>
<error type="S"/>
```

```
</link>
```

XML soubor může být vizualizován ve formě grafu (například Obrázek 13). Dvoupísmenné zkratky uvnitř jednotlivých vrcholů reprezentují zdroje propojených dat. Jejich vysvětlení je k dispozici v Příloze B.



Obrázek 13: Identické vazby prvku `Prague`.

Následující kód představuje XML schéma (datový model) souboru pro ukládání vazeb. Schéma je zapsané v kompaktní syntaxi formátu RELAX NG. Tato forma zápisu je jednoduchá a stručná a navíc pomocí programového vybavení jako například Trang[99] je možné formát RELAX NG

[99]http://www.thaiopensource.com/relaxng/trang.html

přetransformovat do jiných formátů pro popis syntaxe dokumentu, kterými jsou například DTD nebo W3C XML Schema. Soubory pro evidenci vazeb v propojených datech jsou popsány také v publikacích [133] nebo [56].

```
default namespace = ""

start =
  element term {
    attribute date { xsd:date }?,
    attribute name { text },
    element link {
      element subject {
        attribute acronym-2 { xsd:NCName },
        attribute uri { xsd:anyURI }?
      },
      element object {
        attribute acronym-2 { xsd:NCName },
        attribute uri { xsd:anyURI }?
      }?,
      element predicate {
        attribute type { "owl:sameas" | "skos:exactMatch" |
         "skos:closeMatch" | "rdfs:seeAlso" |
         "Wikidata identifier" | "other" }
      }?,
      element error {
        attribute note { text }?,
        attribute type { "S" | "T" | }
      }?
    }
  }
```

Kromě XML souborů s daty o vazbách jednotlivých konceptů, existuje v systému ukládání dat další XML soubor (resources.xml, viz Příloha B). Ten obsahuje seznam zdrojů, které byly analyzovány v rámci jednotlivých testů.

Kořenový element `<resources>` obsahuje informace o jednotlivých zdrojích (element `<resource>`). Každý zdroj může být charakterizován šesti atributy (z nichž jsou pouze první tři povinné):

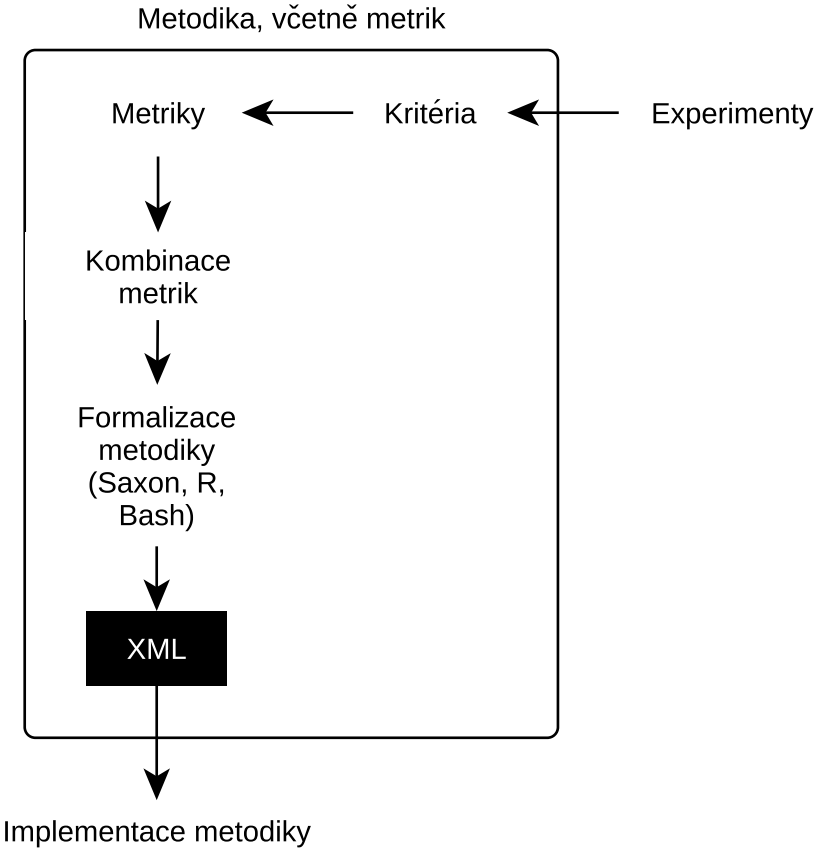
- `acronym` – dvoupísmenná zkratka zdroje,
- `name` – jméno zdroje,
- `key` – řetězec charakteristický pro URI daného zdroje,
- `special` – označení (hodnoty `true` nebo `false`), zda je zdroj speciálně zpracovávaný v XSLT stylu pro vytěžování dat,
- `add_end` – řetězec, který se přidává na konec URI, pro získání souboru s daty,
- `error` – typ chyby zdroje.

Na tomto místě textu je nutné upozornit, že publikovaný postup sběru dat je využitelný pouze pro získání vzorku dat pro účely testování identických vazeb. V žádném případě se nejedná o komplexní postup procházení sítě propojených dat, který jsou publikované například v [56] (v této publikaci je uvedená také jiná varianta datového modelu souboru pro správu vazeb v Linked Data), [134] nebo [135]. Implementace nebo tvorba sofistikovaného postupu pro prohledání prostoru Linked Data by zcela jistě vedla k získání většího množství testovaných dat a tento proces by byl zřejmě po všech stránkách efektivnější. Jedná se však o aktivitu, která je zcela mimo rozsah této práce.

Metriky

Metodika pro hodnocení identických vazeb mezi objekty prostorových propojených dat (Linked Data) (Obrázek 14) vzniká propojením dílčích metrik pro popis a hodnocení orientovaných grafů využívaných v různých vědních oborech, jako je teorie grafů (například stupně grafu), sociologie (například centralita) nebo geografie dopravy (například síťové indexy). 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



Obrázek 14: Architektura – metriky.

reprezentují relace mezi jednotlivými výskyty jednoho objektu v různých datových sadách. Metriky[100] 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. Následující kapitoly popisují metriky pro hodnocení uzlů i sítí, přičemž jako hlavní klasifikační kritérium je použitý vědní obor, ze kterého daný konkrétní postup pochází.

V souladu s [35] byly použité metriky, které mají následující vlastnosti:

- Jsou spočitatelné v lokální síti[101].
- Výsledky se dají zobecnit na celou (globální) síť.
- Jsou schopné identifikovat ideální části sítě (grafu).
- Výsledné hodnoty jsou kladná reálná čísla[102].

Metriky pro hodnocení uzlu v rámci jednoho grafu

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. Ten je zajištěný skriptem pro vyhledávání, sběr, kontrolu a formalizaci informací o identických vazbách popsaných v předchozí sekci.

Optimální uzel v rámci grafu by měl mít následující vlastnosti, které souvisí s významnou pozicí vrcholu v rámci grafu (jinými slovy zdroje propojených dat, který obsahuje reprezentaci sledovaného prvku). 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

[100]Termín metriky pro tento typ metod byl převzatý z publikací [136], [65] a [35].
[101]Poznámka autora: Pro výpočty v této práci byl zvolený program R s knihovnou igraph.
[102]Poznámka autora: Výsledné hodnoty jsou v případě potřeby normovány tak, aby umožnily korektní porovnávání mezi více sítěmi (grafy). Normalizace byla realizována vydělením originálních hodnot číslem $n - 1$, kde n je počet uzlů sítě grafu [[100];].

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.

Předchozí seznam lze snadno vyjádřit pomocí vybraných metrik:

- Centralita stupně[103] – zdroj je pomocí identické vazby propojený na mnoho dalších datových sad.
- Centralita blízkosti[104] – zdroj leží blízko ostatních uzlů (z hlediska délky nejkratší cesty).
- Centralita mezilehlosti[105] – zdroj propojuje nezávislé části Linked Data prostoru.
- Autorita[106] – zdroj je pomocí identických relací dosažitelný z mnoha uzlů (dalších zdrojů propojených dat).
- Hub[107] – 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.

Metriky pro hodnocení uzlů napříč grafy

Uživatele propojených dat zajímá zdroj takových dat nejen z pohledu jediného konkrétní objektu, ale také z hlediska skupiny více či méně homogenních objektů. Uživatel může například položit otázku „Jaký je nevhodnější zdroj prostorových propojených dat, pokud budu chtít maximální množství informací o zemích Sahelu? “. Pro tyto účely je možné implementovat metriky publikované v předchozí části, tedy centralitu stupně, centralitu blízkosti,

[103]Tato metrika má stejný význam jako stupeň uzlu.
[104]Tato metrika zastupuje celkovou vzdálenost a metodu Global Efficiency, které mají podobný význam.
[105]Centralita mezilehlosti má podobný význam jako metoda EgoBetween.
[106]Pro zjednodušení může být nahrazena vstupním stupněm uzlu, případně odpovídajícím typem centrality stupně.
[107]Pro zjednodušení může být nahrazena výstupním stupněm uzlu, případně odpovídajícím typem centrality stupně.

centralitu mezilehlosti, Page Rank, koeficient autority a středu. Souhrnné informace pak mohou být získány ve formě průměru normovaných hodnot (v případě různého ohodnocení významu metrik, ve formě váženého průměru). Druhou možností je sdružování celkových hodnot (například váženého součtu).

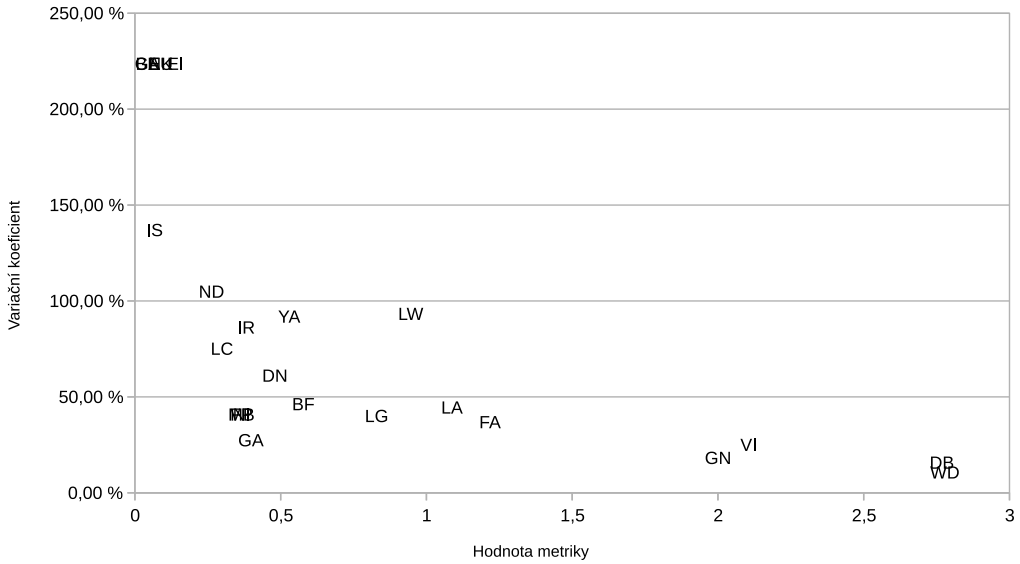
Pro hodnocení uzlů napříč grafy, resp. hodnocení zdrojů propojených dat z hlediska entit patřících do společné domény, jsou vhodné nejen vlastní hodnoty metrik, ale také jejich stabilita. Je vhodné uživatele informovat, zda jsou vysoké průměry vypočteny z podobných hodnot, nebo zda jsou ovlivněné jedním či několika málo extrémními případy, které celý výsledek zkreslují. Jinými slovy, aby mělo zprůměrování vypovídací hodnotu, mělo by rozložení průměrovaných veličin odpovídat Gaussovu normálnímu rozdělení.

Stabilita, resp. variabilita jako její opak, může být vyjádřena několika statistickými veličinami. Kvůli srovnání se jako nejvhodnější jeví variační koeficient (Coefficient of Variation), který slouží ke srovnání variability souborů s rozdílnou velikostí hodnot. Hodnota variačního koeficientu, který někdy bývá označován jako relativní směrodatná odchylka, se získá vydělením směrodatné odchylky aritmetickým průměrem. Výsledky variačního koeficientu jsou zpravidla udávány v procentech.

Na základě průměrných hodnot získaných z metriky (nebo metrik) a variačního koeficientu (Obrázek 15) je možné rozdělit uzly do čtyř základních skupin:

1. Uzly, které se nevyskytují ve všech zkoumaných vzorcích. Jinými slovy slovy zdroje, které neobsahují reprezentanty všech testovaných objektů. Takové zdroje jsou z dalšího hodnocení vyřazeny, k čemuž by pravděpodobně došlo v dalších krocích, neboť vykazují značnou variabilitu.
2. Kvalitní uzly, které mají příznivou (zpravidla vysokou) hodnoty vypočtenou na základě metriky.
3. Stabilní uzly – vrcholy s nízkým variačním koeficientem (u dat testovaných v rámci experimentů se za nízkou dá označit hodnota nižší než 50%; na grafu 15 v dolní pětině).
4. Kvalitní a zároveň stabilní uzly – průnik dvou předchozích množin (na grafu 15 v pravém dolní rohu – v tomto konkrétním případě jako nejvýhodnější vychází uzly DB – DBpedia, WD – Wikidata, GN –

GeoNames.org a VI – VIAF).



Obrázek 15: Hodnoty metriky a variační koeficient uzlů grafu.

Metriky pro hodnocení grafu

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. Uzly vyjadřují zdroje a hrany grafu jejich propojení. 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 kapitole 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[108]),
   (b) reciprocita (do jaké míry existují vzájemné vazby mezi reprezentacemi objektu),
4. Shlukový koeficient – koeficient deklarující existenci nezávislých komponent grafu, které snižují jeho homogenitu.

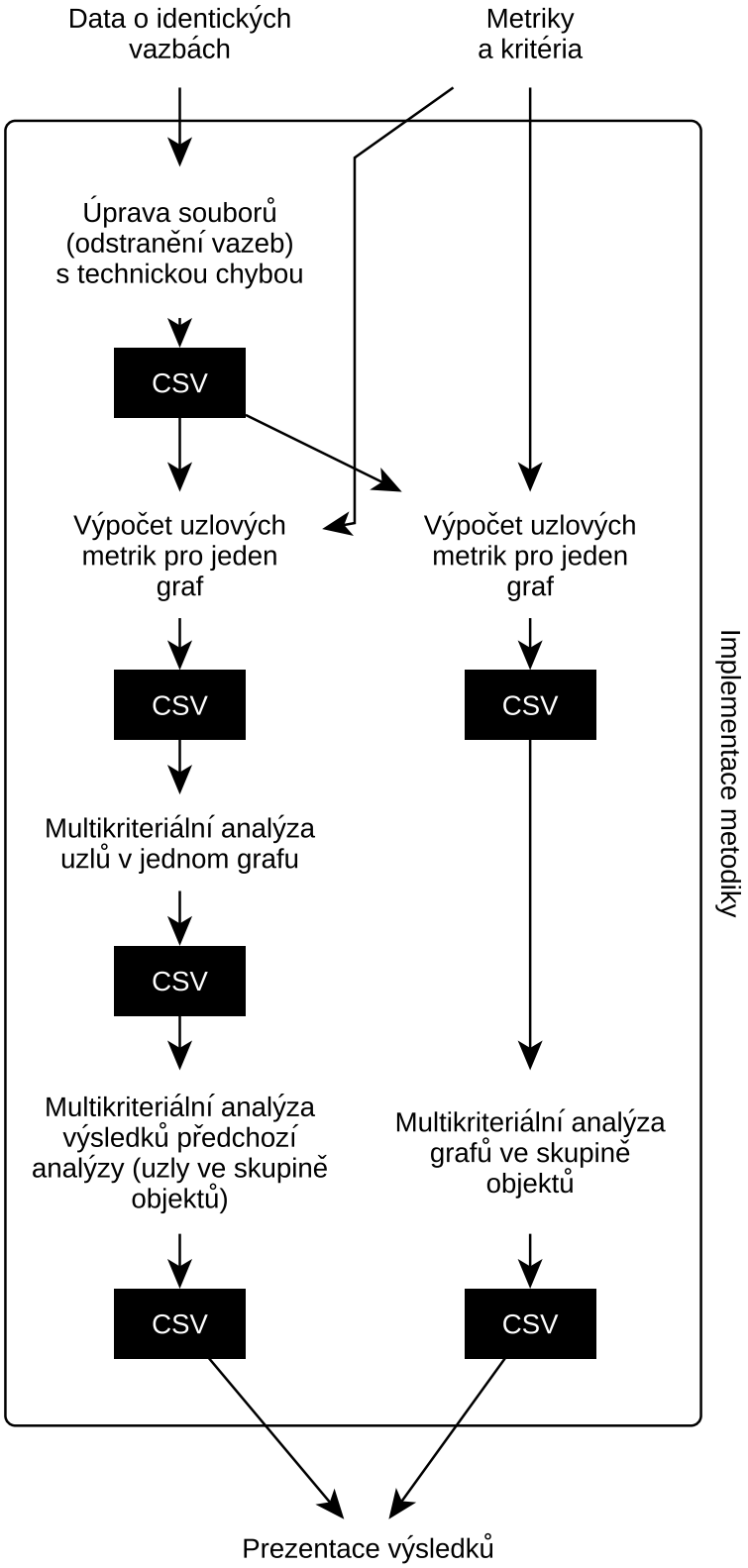
Tyto metriky mohou být použity pro srovnání parametrů grafů (prvků propojených dat), přičemž je použitá stejná váha pro každou metriku. V rámci experimentů jsou porovnávány grafy ukazující pojmy, které spolu souvisí (patří do skupiny podobných objektů, jako jsou hlavní města evropských států nebo stratovulkány).

Implementace

Implementace (Obrázek 16) 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 uložených jako XML soubory, kdy jeden graf odpovídá jednomu datovému objektu. Uzly v takovém grafu symbolizují zdroje propojených dat, které obsahují identifikátor reprezentace daného objektu v konkrétní datové sadě propojených dat (například objekt `Prague` v databázi DBpedia) a orientované hrany propojení uzlů pomocí identických a podobnostních vazeb.

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

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



Obrázek 16: Architektura – implementace metodiky.

vzniklých izolovaných uzlů, které nevstupují do následující analýzy. V tomto kroku je například odstraněný zdroj Freebase, který je sice velice často odkazovaný, ale v současné době již není dostupný. Dále jsou vstupní data přetransformována do formátu CSV, který vstupuje do fáze dalšího zpracování pomocí software R.

Před výpočtem metrik jsou data převedena do grafové struktury. Kalkulace grafových i uzlových metrik je zajištěna R skriptem `metrics.r`. Výsledek je vyexportovaný pro každý graf (reprezentující objekt propojených prostorových dat) do dvou typů tabulek[109]. Tabulka s uzlovými metrikami (například `Prague_node_metrics.csv`) obsahuje v řádcích jednotlivé uzly (zdroje)[110], k nimž jsou ve sloupcích připojené absolutní hodnoty šesti zjišťovaných metrik – centralita stupně, centralita blízkosti, centralita mezilehlosti, skóre autority, skóre středu, Page Rank. Tento druh tabulky obecně slouží k posuzování kvality zdrojů propojených prostorových dat – jak jsou jednotlivé datové zdroje zapojeny do šíření informací na základě principu Linked Data. Tabulka pro grafové metriky (například `Prague_graph_metrics.csv`) je jednodušší. Opět obsahuje ve sloupcích absolutní hodnoty metrik (počet uzlů, hustota grafu, reciprocita, shlukový koeficient), ale vyskytuje se v ní pouze jediný řádek (kromě hlavičky), který reprezentuje celý graf (objekt). Objekty jsou v tabulce vyjádřené v prvním sloupci pomocí zkráceného identifikátoru v datové sadě DBpedia (například objekt `Prague`). Tento typ tabulky je určený ke zjišťování kvality popisu datových objektů v prostoru propojených dat z hlediska identických vazeb.

Souhrnné hodnoty, v případě uzlových metrik pro celý graf, v případě grafových metrik pro skupinu objektů, jsou zjišťovány pomocí multikriteriální (vícekriteriální) analýzy [137]. 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

[109]Tabulky jsou publikované ve formátu CSV. Jako oddělovač buněk se používá středník. Pro oddělení desetinných míst je využitá tečka.

[110]První sloupec tabulky obsahuje dvoupísmenné zkratky zdrojů jednotně používané v celé práci

kvantitativními hodnotami. Z důvodů jednoduchosti a srozumitelnosti byl pro první fáze výzkumu zvolený tzv. vážený součet (weighted sum) [138]. Výše uvedené tabulky mají charakter kriteriálních matic, které do váženého součtu vstupují. Před výpočtem váženého průměru pro uzlové i grafové metriky na úrovni skupin objektů (více grafů) je nutné dílčí tabulky propojit.

Technicky je implementační fáze zajištěna pomocí sady skriptů, které využívají jazyk R skript pro výpočet hodnot jednotlivých metrik, pro jejich normování a výpočet hodnot váženého součtu a případně dalších metod multikriteriální analýzy. Původně byl ještě využitý formát XSLT – značkovací jazyk pro vytváření transformačních šablon, který sloužil především pro výpočet souhrnných statistik a transformaci mezi formáty (XML a CSV). V průběhu testování případových studií a optimalizace skriptů došlo k nahrazení XSLT šablon skripty v jazyce R. Důvodem byla především rychlost zpracování rozsáhlých dat. Skripty jsou zpracovávány pomocí programu R (především knihovny igraph, MCDA). Spouštění a řízení celého procesu zajišťuje skript pro unixový příkazový procesor (shell) Bash doplněný o další externí programy (například AWK pro zpracování textů).

Nastíněný postup implementace metrik bude stejný ve všech případových studiích s výjimkou první, která ilustruje především chyby v identických vazbách a způsoby sběru dat. Jednotlivé experimenty se odlišují pouze ve vyvozených závěrech, způsobech prezentace výsledků (různé druhy grafů, schémat, tabulek, map apod.), podskupinách a vlastnostech určených pro vzájemné porovnání.

Kapitola 5

Experimenty

Kapitola Experimenty představuje praktickou část této studie, která slouží především k ověřování poznatků získaných studiem, analýzou a syntézou materiálů v rámci rešerše a také během sestavování metodiky pro testování jednotlivých aspektů kvality identických a podobnostních vazeb propojených prostorových dat (viz příslušné kapitoly práce). V jednotlivých případových studiích byla aplikována metodika, případně dílčí metriky, prezentované v předchozí části práce na reálná propojená prostorová data. Výsledky jsou komentovány, shrnuty a zobecněny v následující kapitole Diskuze, případně v závěru práce.

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.

Všechny publikované experimenty mají totožnou strukturu, přičemž v konkrétních případech může docházet k jistým odchylkám například v rozsahu jednotlivých částí nebo způsobu prezentace výsledků:

- Popis dat využívaných v případové studii (téma, specifika, testované výchozí předpoklady, počet testovaných objektů, použitý SPARQL dotaz pro získání dat apod.),
- výpočet a publikování výsledků dílčích uzlových a grafových metrik,
- multikriteriální analýza pomocí hodnot metrik s využitím metody váženého součtu,
- vyhodnocení výsledků pro celou skupinu analyzovaných dat, včetně potvrzení nebo odmítnutí výchozích předpokladů.

Následující seznam obsahuje výčet případových studií, které jsou součástí experimentální fáze výzkumu[111]. Kromě názvu skupiny objektů propojených prostorových dat, který je zároveň označením celé případové studie, uvádí i výchozí předpoklady, které jsou v rámci experimentu ověřovány.

- Hlavní města ve střední Evropě
  - Navrhované metriky a metody bude možné použít pro hodnocení kvality propojených prostorových dat z hlediska identických vazeb.
  - Výsledky hodnocení budou reflektovat geografické odlišnosti zkoumaných objektů
- Evropské mezinárodní silnice
  - Metriky, metody a jejich implementace se dají použít pro rozsáhlejší datový soubor.
- Hraniční řeky

[111]Poznámka autora: Ve skutečnosti bylo realizováno více případových studií především ve fázi tvorby a prvotního ověřování metodiky. Tyto studie, například testování objektu `Georgia` z pohledu různých způsobů sběru dat o identických vazbách nebo z hlediska sémantické správnosti, nakonec nebyly do práce zařazeny především proto, že narušovaly homogenitu předloženého výzkumu. Může se stát, že budou později publikovány jako samostatné studie.

  - V případě dat nehomogenních z hlediska lokalizace (na rozdíl od předchozí datové sady, která byla omezená pouze na území Evropy) je očekávána větší variabilita metrik.
  - Vzhledem k tomu, že všechny prvky této skupiny leží na území minimálně dvou států, je zde odůvodněný předpoklad, že v datech budou vznikat shluky obsahující lokálně omezené zdroje.
- Uzlová letiště
  - Při porovnání výsledků získaných pomocí kritérií s jednotkovou váhou a kritérií rozdělených podle významu (na základě Saatyho metody) nedojde k významným rozdílům, protože zvolené metriky (jako kritéria) si jsou navzájem rovné.
- Republiky
  - Na základě dílčích metrik je možné vybrat datové sady zdrojů prostorových propojených dat vhodných pro konkrétní účely.
- Srovnávací studie
  - Další typy případových studií jsou zaměřené na srovnání výsledků metrik a multikriteriální analýzy pro dvě skupiny, které pokrývají mezi sebou související téma, ale zároveň se odlišují v území, k němuž se vztahují. Porovnávána budou data globální s lokálními daty z České republiky, data z Evropy a Afriky (z geopolitického hlediska se jedná o vztah centra a periférie [139]) a data z Evropy a Severní Ameriky, převážně z USA (obě území jsou srovnatelná z pohledu politické a socioekonomické geografie). Cílem těchto experimentů je identifikace rozdílů v zavádění identických vazeb mezi zdroji prostorových propojených dat.
  - Jednotlivé studie jsou zaměřené na stratovulkány ve světě, hory v České republice, evropská hlavní města, hlavní města v Africe, závodní okruhy, které hostily velké ceny Formule 1, a okruhy, které jsou spojené se sérií IndyCar.
  - Pro všechny studie jsou používány váhy odvozené v rámci zpracování uzlových letišť.

Účelem této části práce je

- ověření metod definovaných v předchozí kapitole,

- poskytnutí zpětných vazeb pro konkrétní testované datové entity a jejich skupiny,
- zobecnění výsledků na úroveň popisu prostorových propojených dat a vyvození námětů pro zlepšení situace propojených prostorových dat především z hlediska identických vazeb.

Hlavní města ve střední Evropě

Tato případová studie ukazuje implementaci základních metrik a metod hodnocení kvality identických a podobnostních vazeb propojených prostorových dat. Proto bude popis jednotlivých kroků rozsáhlejší než v následujících experimentech. Testovaný výchozí předpoklad je možné zformulovat následujícím způsobem – „Navrhované metriky a metody bude možné použít pro hodnocení kvality propojených prostorových dat z hlediska identických vazeb“.

Metodika je ilustrovaná na příkladu pěti geografických objektů, které reprezentují středoevropská hlavní města[112]. Vybrané objekty lze považovat za poměrně homogenní, protože jsou velmi podobné z hlediska prostorové lokalizace, společné historie, velikosti a počtu obyvatel nebo významu. Přesto se mezi pěticí zkoumaných objektů dají najít některé odlišnosti. Právě na ty se zaměřuje druhý výchozí předpoklad – „Výsledky hodnocení budou reflektovat geografické odlišnosti zkoumaných objektů“.

Informace o vazbách (Obrázek 17) jsou stejně jako v ostatních experimentech vyhledávány pomocí skriptu založeného na principu „follow your nose“. Průzkum sítě propojených dat začíná v příslušném prvku v databázi DBpedia (viz následující seznam). Hloubka prohledávání byla nastavena na hodnotu 6.

- Berlin – Berlin[113]
- Bratislava – Bratislava[114]
- Praha – Prague[115]

[112]Názvy měst byly ponechány ve formě anglických endonym.
[113]http://dbpedia.org/data/Berlin
[114]http://dbpedia.org/data/Bratislava
[115]http://dbpedia.org/data/Prague

- Warszawa – Warsaw[116]
- Wien – Vienna[117]

Podobně jako v dalších případových studiích i do tohoto experimentu vstupují data o identických vazbách zbavená všech vazeb a vrcholů, která jsou ovlivněná technickou chybou. Ostatní chyby odstraněné nejsou, protože sice znemožňují strojové zpracování dat s využitím všech výhod Linked Data přístupu, ale i přesto umožňují přístup k dalším informacím.

První část analýzy je zaměřená na vyhledání nejvhodnějšího zdroje propojených dat pro každý datový objekt. Vhodnost zdroje z hlediska identických vazeb je vyjádřena šesti metrikami, kterými jsou centralita stupně (v tabulkách označení jako C_d), centralita blízkosti (C_c), centralita mezilehlosti (C_b), skóre autority (A), skóre středu (H) a metoda PageRank (PR). Pro každou datovou entitu je generována tabulka, kde ve sloupcích jsou absolutní hodnoty jednotlivých metrik a v řádcích datové sady propojených dat (příklad objektu `Prague` ukazuje tabulka 2).

Tabulka 2: Hodnoty metrik pro objekt `Prague`.

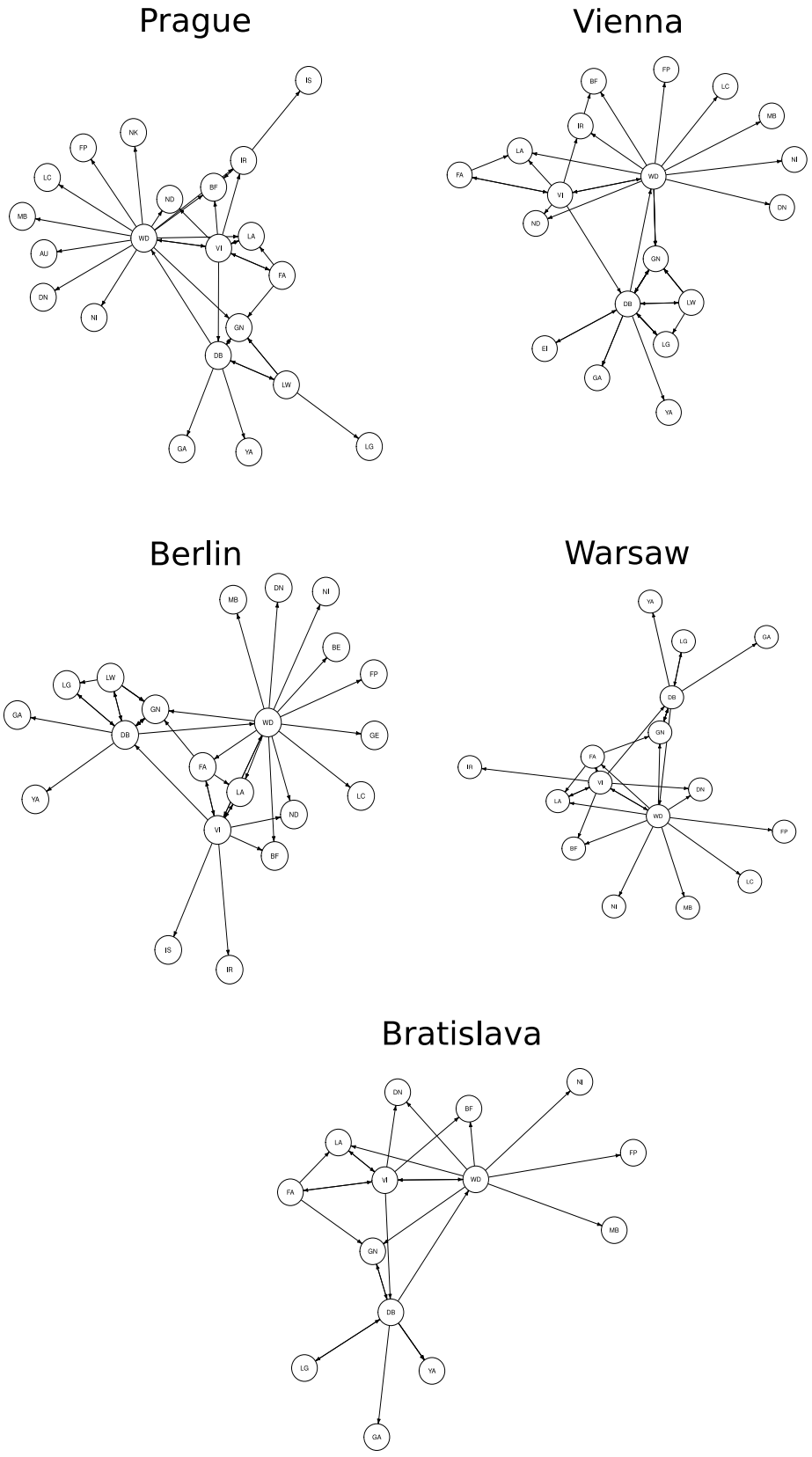
Zdroj	C_d	C_c	C_b	A	H	PR
AU	1	0,002	0	0,31	0,00	0,03
BF	3	0,002	0	0,49	0,00	0,06
DB	10	0,027	53	0,38	0,49	0,13
DN	1	0,002	0	0,31	0,00	0,03
FA	4	0,020	0,5	0,16	0,36	0,04
FP	1	0,002	0	0,31	0,00	0,03
GA	1	0,002	0	0,15	0,00	0,04
GN	8	0,018	6,67	1,00	0,14	0,09
IR	4	0,002	7	0,46	0,09	0,04
IS	1	0,002	0	0,03	0,00	0,04
LA	4	0,019	0	0,57	0,08	0,05
LC	1	0,002	0	0,31	0,00	0,03
LG	1	0,002	0	0,14	0,00	0,04

[116]http://dbpedia.org/data/Warsaw
[117]http://dbpedia.org/data/Vienna

Zdroj	C_d	C_c	C_b	A	H	PR
LW	5	0,019	6	0,15	0,45	0,04
MB	1	0,002	0	0,31	0,00	0,03
ND	2	0,002	0	0,46	0,00	0,04
NI	1	0,002	0	0,31	0,00	0,03
NK	1	0,002	0	0,31	0,00	0,03
VI	10	0,029	38,3	0,44	0,51	0,08
WD	15	0,031	63,5	0,31	1,00	0,05
YA	1	0,002	0	0,15	0,00	0,04

Z tabulky 2 je možné vyvodit následující závěry týkající se zdrojů výskytu objektu `Prague` v různých datových sadách založených na principu Linked Data, které jsou vzájemně propojené pomocí identických vazeb:

- Nejvíce je do sítě propojených dat integrovaná reprezentace entity `Prague` v datové sadě Wikidata (WD), která vykazuje nejvyšší hodnotu centrality stupně.
- Datová sada GeoNames.org (GN) představuje do jisté míry autoritu na poli prostorových propojených dat, o čemž svědčí vysoké skóre autority (1). Z grafu (Obrázek 13) je patrné, že existuje více tzv. listových uzlů, na které je pouze odkazováno z jiných vrcholů, ale vrchol s označením GN je odkazován v nejvyšší možné míře.
- Wikidata mají obecně vysoké postavení v pořadí jednotlivých metrik. To znamená, že představují ideální datovou sadu, z níž má smysl začít sběr informací o daném objektu a dále postupovat po identických vazbách. To odpovídá i schématu na obrázku 13, kde uzel označený jako WD leží zhruba ve středu sítě, dosažitelnost okolních vrcholů je poměrně krátká (vysoká centralita blízkosti). Navíc přímí sousedi tohoto uzlu jsou poměrně hustě provázáni a jedná se o kvalitní uzly ve smyslu hodnocení pomocí metrik (především PageRank).
- Výše uvedené propojení na kvalitní uzly je nejvíce patrné v případě zdroje DBpedia (DB), který vykazuje nejvyšší hodnotu koeficientu PageRank.
- Z pohledu využívání Linked Data je zajímavá druhá pozice zdroje VIAF (VI). Tato nepříliš v praxi využívaná datová sada je vysoce hodnocena



Obrázek 17: Identické vazby prvků reprezentujících hlavní města ve střední Evropě.

ze tří důvodů – (1) mnoho propojení na ostatní datové sady; (2) velké množství recipročních vazeb; (3) propojení s důležitými zdroji (všechna spojení jsou s uzly hodnocenými v první desítce).

Podobné závěry by bylo možné získat i pro další testované objekty. Účelem případových studií však není apriorní testování jednotlivých prvků, ale hodnocení dat v rámci zvolené skupiny. Teprve při výskytu konkrétních anomálií jsou analyzovány i dílčí objekty.

Souhrnná analýza uzlových metrik pro každý graf (objekt) je získána pomocí metody váženého součtu jako jedné z technik multikriteriální analýzy. Tabulka 3 ukazuje výsledné hodnoty vícekriteriální analýzy všech objektů skupiny středoevropská hlavní města. Jednotlivé zdroje (řádky tabulky) představují alternativy (v tomto konkrétním případě alternativy zdrojů propojených dat nejlépe umístěného a prolinkovaného geografického objektu náležícího do zvolené skupiny). Jako kritéria slouží uzlové metriky vypočítané pro jednotlivé objekty v předchozím kroku. Výsledné hodnoty multikriteriální analýzy jsou pro objekty publikované ve sloupcích tabulky.

Tabulka 3: Výsledky multikriteriální analýzy uzlových metrik pro objekty ze skupiny středoevropských hlavních měst.

Zdroj	Berlin	Bratislava	Prague	Vienna	Warsaw
BE	0,61	0,00	0,00	0,00	0,00
BF	0,75	0,96	1,18	0,73	1,36
DB	2,97	2,95	2,21	3,37	2,57
DN	0,61	0,96	0,82	0,63	1,36
FA	1,27	1,18	1,17	0,66	1,99
FP	0,61	0,72	0,82	0,63	0,97
GA	0,63	0,87	0,62	0,82	0,72
GE	0,61	0,00	0,00	0,00	0,00
GN	2,44	1,58	2,19	2,44	2,19
IR	0,52	0,00	1,30	0,83	0,79
IS	0,52	0,00	0,46	0,00	0,00
LA	0,96	1,29	1,37	0,83	1,79
LC	0,61	0,00	0,82	0,63	0,97

Zdroj	Berlin	Bratislava	Prague	Vienna	Warsaw
LG	0,95	0,98	0,51	1,18	0,86
LW	1,81	0,00	1,31	1,67	0,00
MB	0,61	0,72	0,82	0,63	0,97
ND	0,75	0,00	1,06	0,74	0,00
NI	0,61	0,72	0,82	0,63	0,97
VI	1,81	2,22	2,22	1,44	2,84
WD	2,57	2,73	3,02	2,89	3,34
YA	0,63	1,73	0,62	0,62	0,72
AU	0,00	0,00	0,82	0,00	0,00
NK	0,00	0,00	0,82	0,00	0,00
EI	0,00	0,00	0,00	0,72	0,00

Z tabulky 3 jasně vyplývá, že nejvhodnějšími zdroji pro propojená prostorová data reprezentující středoevropská hlavní města jsou z hlediska identických vazeb Wikidata (WD) a DBpedia (DB). Wikidata vykazují nejlepší hodnoty pro uzly (objekty) `Prague` a `Warsaw`. DBpedia je nejvhodnější zdroj pro objekty `Berlin`, `Bratislava` a `Vienna`. Mezi další nadprůměrně hodnocené zdroje patří VIAF (VI), GeoNames.org (GN).

Z tabulky 3 jsou také zřetelné zdroje, které se objevují pouze pro jediný objekt.

- Libraries Australia (AU)
- Biblioteca nacional de Espaňa (BE)
- Eionet (EI)
- Genealogy.net (GE)
- Databáze Národní knihovny ČR (NK)

Zajímavé je, že se v tomto případě až na Databázi Národní knihovny ČR nejedná o žádný lokální zdroj, od kterého by se dalo očekávat, že bude obsahovat pouze prvky z jednoho státu. Jednotlivá města se liší i svým zastoupením v Linked Data prostoru, které do jisté míry odpovídá významu města (a například i počtu obyvatel), kdy Berlín je zastoupený v největším počtu zdrojů propojených dat a naopak reprezentace entity `Bratislava` se vyskytuje v nejnižším počtu případů.

Poslední tabulka týkající se uzlových metrik (Tabulka 4 ukazuje souhrnné výsledky pro uzly (zdroje) za celou skupinu. V tomto případě nejsou zdroje seřazeny abecedně jako v předchozích tabulkách (2, 3), ale sestupně podle hodnoty ve druhém sloupci. Čísla prezentovaná jako „Celková hodnota“ byla vypočítána opět technikou váženého součtu z dat publikovaných v tabulce 3. Kritérii pro analýzu se tedy stávají dílčí výsledky multikriteriální analýzy pro jednotlivé objekty skupiny. Jako alternativy opět slouží zdroje propojených prostorových dat, které jsou v grafech zastoupeny uzly.

Tabulka 4: Výsledky souhrnné multikriteriální analýzy uzlových metrik pro skupinu středoevropských hlavních měst.

Zdroj	Vážený součet
WD	14,55
DB	14,07
GN	10,84
VI	10,53
FA	6,27
LA	6,24
BF	4,98
LW	4,79
LG	4,48
DN	4,38
YA	4,32
FP	3,75
MB	3,75
NI	3,75
GA	3,66
IR	3,44
LC	3,03
ND	2,55
IS	0,98
AU	0,82
NK	0,82
EI	0,72

Zdroj	Vážený součet
BE	0,61
GE	0,61

Z tabulky 4 jsou patrné následující poznatky. V testované skupině

- jsou zdroje DBpedia a Wikidata srovnatelné a naprosto dominantní,
- poměrně vysoké hodnoty vykazuje zdroj VIAF a GeoNames.org,
- jsou mezi jednotlivými hodnotami značné rozdíly (variační koeficient je 84,22%, což znamená, že směrodatná odchylka se velice výrazně podílí na hodnotě průměru).

Druhým aplikovaným typem metrik jsou metriky grafové, které se týkají grafů jako celku. Jejich výpočty a zpracování je podobné jako v případě uzlů grafu. V části Metodika byla vybrána pětice grafových metrik, jejich detailnější přehled je součástí kapitoly Rešerše. Tento soubor metrik tvoří počet uzlů v grafu (N), hustota grafu (D), reciprocita vazeb (R) a shlukový koeficient (CC). Podobně jako metriky uzlů jsou i grafové metriky chápané jako maximalizační (nejvyšší hodnota byla chápána nejpozitivněji).

Tabulka 5 ukazuje přehled hodnot grafových (síťových) metrik pro všechny objekty (grafy) skupiny středoevropská hlavní města. Objekty jsou umístěné do řádků, absolutní hodnoty metrik jsou ve sloupcích.

Tabulka 5: Výsledné hodnoty grafových metrik pro skupinu středoevropských hlavních měst.

Objekt	S	D	R	CC
Berlin	21	0,10	0,38	0,20
Bratislava	14	0,15	0,37	0,23
Prague	21	0,09	0,32	0,20
Vienna	19	0,13	0,41	0,22
Warsaw	16	0,12	0,40	0,26

Z tabulky 5 vyplývá, že existují malé rozdíly ve výsledcích reciprocity (tyto

hodnoty indikující oboustranné vzájemné vazby mezi objekty jsou obecně nízké) a shlukového koeficientu (v grafech víceméně neexistují izolované skupiny, viz schémata na obrázku 17, kde grafy jsou vizuálně rozlišitelné dva typy grafů – `Berlin` a `Prague`, kde sice kolem uzlu WD vzniká podobný klastr jako u ostatních grafů, ale tyto skupiny jsou se zbytkem grafu lépe provázané než je tomu u grafů s označením `Bratislava`, `Vienna` a `Warsaw`). Zbylé dvě metriky vykazují výraznější rozdíly – počet uzlů v grafu do jisté míry koresponduje s významem měst z hlediska příslušnosti k tzv. západní Evropě ve smyslu zemí, které nebyly zcela součástí východního (komunistického) bloku. Sídla ležící ve východních oblastech střední Evropy jsou popsaná v menším počtu zdrojů než hlavní města tzv. post-socialistických států[118]. Hustota grafu je na základě dat v tabulce 5 závislá na počtu uzlů – grafy, které obsahují méně vrcholů, jsou lépe provázané.

Poslední tabulka této případové studie (Tabulka 6) obsahuje pouze dva sloupce – seznam objektů náležejících do skupiny (seřazené sestupně podle hodnoty v následujícím sloupci) a hodnoty vycházející z příslušné metody multikriteriální analýzy. Tyto hodnoty byly podobně jako v předchozích případech zjištěny metodou váženého součtu se shodnou váhou pro všechna kritéria (grafové metriky).

Tabulka 6: Výsledky multikriteriální analýzy grafových metrik pro skupinu středoevropských hlavních měst.

Objekt	Vážený součet
Berlin	21,68
Prague	21,61
Vienna	19,76
Warsaw	16,78
Bratislava	14,75

Na rozdíl od tabulky 4 jsou celkové hodnoty vycházející z multikriteriální analýzy (Tabulka 6) mnohem více stabilní. Minimální hodnota představuje

[118]V tomto případě je možné diskutovat o termínu `Berlin`, který byl hlavním městem socialistického státu, ale je potřeba si uvědomit existenci Západního Berlína a Německa jsou sloučeného státu.

68% z hodnoty maximální. Lze tedy prohlásit, že podobnost propojení reprezentací testovaných objektů v různých zdrojích obsahujících propojená prostorová data je pro skupinu středoevropských měst poměrně vysoká. Dílčí odlišnosti jsou identifikovány a popsány výše. Naopak jednotlivé zdroje (jejich zastoupení a vzájemné propojení pro testované objekty) jsou velice různorodé. Lze vybrat dvě klíčové datové sady (Wikidata, DBpedia), významnou datovou sadu (VIAF), zdroj, který představuje autoritu (často referencovanou datovou sadu) pro zvolené prvky (GeoNames.org) a marginální zdroje, s nimiž zřejmě nemá smysl pracovat, pokud v nich nebude identifikovaný specifický a pro daný užitečný obsah.

Na počátku této podkapitoly byly definovány dva výchozí předpoklady: - „Navrhované metriky a metody bude možné použít pro hodnocení kvality propojených prostorových dat z hlediska identických vazeb.“, - „Výsledky hodnocení budou reflektovat geografické odlišnosti zkoumaných objektů.“

První z předpokladů se podařilo potvrdit. Zvolené metriky, metodika jejich propojení (pomocí multikriteriální analýzy) a technické zpracování umožňují získat kvantitativní údaje o testovaných vlastnostech propojených prostorových dat. Výsledky jsou patrné především při srovnání grafových sítí na obrázku 17 a výsledných hodnot v tabulkách 6 a 6, kde vizuální dojem a kvantifikované výsledky jsou v naprosté shodě. Je samozřejmě možné diskutovat především váhy jednotlivých metrik, což je úkolem některých dalších experimentů.

Druhý výchozí předpoklad také koresponduje s výsledky zkoumání. Z tabulky 6, kde jsou objekty seřazeny podle celkových hodnot produkovaných multikriteriální analýzou, je zřetelný rozdíl mezi městy, která jsou často (především v oblastech na západ od regionu střední Evropy) řazena do spíše do východní Evropy (Bratislava a Varšava), metropolemi západoevropských zemí (Berlín, Vídeň). Výjimkou je Praha, jejíž postavení v tabulce 6 pravděpodobně evokuje nikoli politicko-geografickou polohu, ale spíše její faktické umístění. Je však potřeba si uvědomit, že tvrzení předložená v předešlém textu vychází z ověřování na velice malém vzorku dat. Proto budou testovány v dalších případových studiích jako výchozí předpoklady.

Evropské mezinárodní silnice

Účelem této případové studie je otestovat první výchozí předpoklad potvrzený v předchozím experimentu na rozsáhlejším datovém vzorku. Výchozí předpoklad v této případové studii tedy zní – „Metriky, metody a jejich implementace se dají použít pro rozsáhlejší datový soubor.“. Posuzování se bude týkat nejen robustnosti metodiky, ale také technického řešení (například únosné rychlosti zpracování).

Evropské mezinárodní silnice představují evropské páteřní dopravní trasy. Vedou převážně po dálničních komunikacích a rychlostních silnicích (pokud jsou dostupné). Procházejí většinu zemí Evropy (kromě ostrovních a malých států typu Vatikán). Silnice mají jednotné označení, které začíná písmenem E, a čísla jsou přidělovaná separátně ve směru sever-jih a západ-východ. Číslování bylo vytvořené organizací UNECE (Evropská hospodářská komise OSN). Datová sada získaná pomocí dotazu na SPARQL endpoint datové sady DBpedia obsahuje celkem 231 objekt, přičemž evropských silnic je zhruba 250.

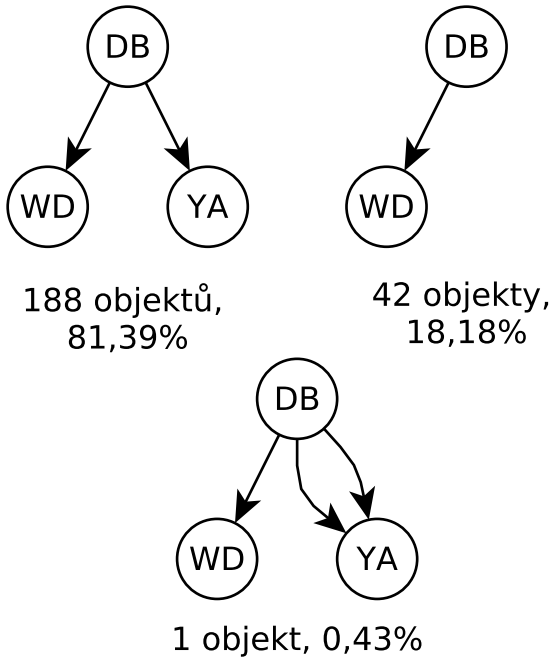
Analýza uzlů je v tomto konkrétním případě velice jednoduchá. Celá skupina prostorových dat obsahuje pouze tři různé uzly (Tabulka 7, Obrázek 18) – DBpedia (DB), Wikidata (WD) a Yago (YA). Tyto tři uzly se objevují pouze ve třech kombinacích (Tabulka 8, Obrázek 18). Ve všech případech je DB jako kořenový uzel a vrcholy WD a YA jsou listové. Sítě se liší pouze v propojení na zdroj Yago, který se ve 42 případech (18,18%) nevyskytuje, v 1 grafu (0,43%) existují dvě propojení z DBpedia na dvě různé reprezentace v Yago (objekt `Autopista\_AP-68`) a zbylé případy (188; 81,39%) obsahují jedno propojení mezi sadami DBpedia a Yago.

Tabulka 7: Výsledky souhrnné multikriteriální analýzy uzlových metrik pro skupinu evropských mezinárodních silnic.

Zdroj	Celková hodnota
DB	799,34
WD	590,82
YA	480,59

Tabulka 8: Kombinace hodnot multikriteriální analýzy ve skupině evropských mezinárodních silnic.

DB	WD	YA	Počet uzlů
3,26	2,54	2,54	188
4,35	2,65	0	42
4,76	2,00	3,57	1



Obrázek 18: Identické vazby prvků reprezentujících hlavní města ve střední Evropě.

Jak vyplývá z obou tabulek (7 a 8), ale v tomto případě je tento fakt patrný i na první pohled z grafů (Obrázek 18), dominantním zdrojem pro téma evropských mezinárodních silnic je DBpedia, kterou je možné považovat za nejlepší (i když rozhodně ne ideální) datovou sadu Linked Data, co se týká výše uvedeného tématu.

Podobně jednoduché je i srovnání metrik mezi jednotlivými grafy (objekty zařazenými do skupiny evropských mezinárodních silnic). Metriky reciprocita a shlukový koeficient nabývají ve všech případech jediné hodnoty – reciprocita a shlukový koeficient jsou nulové. Hustota a velikost grafu se odvozují od počtu

uzlů a hran, proto vzhledem k předchozímu odstavci existují dvě různé hodnoty, ale tři kombinace těchto parametrů.

Evropské mezinárodní silnice představují příklad velmi specifických prostorových dat, jejichž využití v běžné praxi geoinformatiky není časté. Důvodem tohoto stavu může být i fragmentace Evropy, kde stále převažují národní zájmy nad kontinentálními. Proto existují minimální požadavky na tento typ dat, který v případě potřeby bývá nahrazován spojením a výběrem národních datových sad, případně výběrem z OpenStreetMap [140]. Z tohoto důvodu je zajímavá absence vazeb na zdroj LinkedGeoData, který představuje „obraz“ OpenStreetMap ve světě propojených dat[119]). Od toho se odvíjí i jejich postavení v síti propojených dat, které lze označit jako velice omezené z hlediska identických vazeb. Z hlediska využívání identických vazeb není patrná ani hierarchie evropských mezinárodních silnic, kdy by silnice významnější kategorie byly obsaženy ve více propojených zdrojích než silnice druhé kategorie.

Výchozí předpoklad se podařilo potvrdit, protože pomocí zvolené metodiky se podařilo zpracovat více než padesátinásobné množství dat než v předchozím případě. Na běžném notebooku trval výpočet zhruba jednu minutu, což nepředstavuje velkou zátěž vzhledem k tomu, že zpracování dat se provádí jednorázově. Kromě možnosti zpracování rozsáhlejších dat se ukázal další zajímavý fakt. Ačkoli data o identických vazbách si byla poměrně podobná, metodika citlivě zareagovala na každou dílčí změnu a drobné odlišnosti ve vstupních datech byly ve výsledku jasně zřetelné.

Hraniční řeky

Hraniční řeky jsou zajímavým geografickým fenoménem. Na jedné straně tvoří tradiční a přirozenou hranici mezi dvěma územními celky. Společně s dalšími nepřístupnými nebo těžko přístupnými prvky krajiny, jako jsou například horská pásma, představují zřejmě nejstarší typ hranice ve smyslu vymezení území. Na straně druhé se jedná o klíčovou složku krajiny

[119]Bohužel se v současné době vyskytují problémy s aktualizacemi a fungováním tohoto zdroje.

pro mnoho oborů lidské činnosti, jako je například získávání pitné vody, zemědělství (zavlažování), doprava, vojenství (přirozená ochrana území), průmysl (především odvětví náročná na spotřebu vody), cestovní ruch a podobně. Z těchto důvodů existuje velké množství odborných studií na téma hraniční řeky (například [141][120], [142] nebo [143]). Tento typ objektu je zajímavý i z hlediska propojených dat, neboť je zde odůvodněný předpoklad, že by konkrétní objekt mohl být obsažený v datových sadách pocházejících z obou stran hranice a díky tomu by mohly grafy sítě obsahovat minimálně dva výrazné shluky.

Pro tuto skupinu dat jsou připraveny dva vstupní předpoklady:

1. V případě dat nehomogenních z hlediska lokalizace (na rozdíl od předchozí datové sady, která byla omezená pouze na území Evropy) je očekávána větší variabilita metrik.
2. Vzhledem k tomu, že všechny prvky této skupiny leží na území minimálně dvou států, je zde odůvodněný předpoklad, že v datech budou vznikat shluky obsahující lokálně omezené zdroje.

SPARQL dotaz vygeneroval celkem 82 záznamy, které se týkaly hraničních řek na celém světě. Před vlastním zpracováním byly dva soubory vypuštěny, neboť se jednalo o seznam hraničních řek a nikoli vlastní geografické objekty. Dalších 11 objektů nebylo zpracováno při procesu vyhledávání identických vazeb, protože tyto prvky neobsahovaly žádné standardizované identické nebo podobnostní vazby. Analýza se týká 69 objektů.

Objekty jsou popsané celkem ve 14 zdrojích (Tabulka 9). Z toho se pouze 2 zdroje (DBpedia, Wikidata) vyskytují u všech 69 objektů. Znalostní báze Yago chybí pouze v jednom případě (`Pigeon_Bay`[121]). Další v pořadí z hlediska výskytu je geografická databáze GeoNames.org. Ta však absentuje téměř v jedné třetině testovaných objektů.

[120] V Austrálii existuje skupina řek s názvem Border Rivers, které tvoří hranici mezi Novým Jižním Walesem a Queenslandem.
[121] Hraniční řeka mezi Ontariem v Kanadě a Minnesotou (USA).

Tabulka 9: Výsledky souhrnné multikriteriální analýzy uzlových metrik pro skupinu hraničních řek.

Zdroj	Celková hodnota
DB	229,46
WD	200,16
GN	122,61
YA	110,68
VI	91,55
DN	53,98
FA	21,78
LA	16,20
ND	13,84
LG	5,45
IR	3,90
BF	2,69
NI	2,46
GA	1,13

Tabulka 9 ukazuje výsledky multikriteriální analýzy pro zdroje (uzly) vyskytující se v propojených datech reprezentující hraniční řeky. Z tabulky je zřetelná (podobně jako v předchozích studiích) dominance datových sad DBpedia a Wikidata. Poměrně vysokou celkovou hodnotu váženého součtu mají také GeoNames.org, Yago (i přes to, že představuje listový uzel bez odkazů na další zdroje) a VIAF.

Význam zdroje DBpedia pro téma hraniční řeky je patrný také z hodnot jednotlivých metrik průměrovaných přes grafy reprezentující konkrétní datové objekty. DBpedia v šesti ze sedmi metrik (včetně výsledků multikriteriální analýzy) dosahuje maximální průměrné hodnoty. Výjimkou je skóre autority, což vypovídá o tom, že DBpedia je ideální zdroj pro počátek vyhledávání informací o hraničních řekách, ale mnohé relevantní informace se uživatel dozví až z jiných datových sad, kterými v tomto případě jsou Wikidata (maximální hodnota skóre autority), Yago (83% z maximální průměrné hodnoty skóre autority) a GeoNames.org (74%).

Z hlediska grafových metrik je zajímavá hned ta první – počet uzlů (velikost grafu). Ačkoli bylo během analýzy detekováno 14 zdrojů, maximální počet uzlů v grafu je 12 (objekt `Ussuri_River^`[Řeka Ussuri tvoří hranici mezi Ruskem a Čínou.). To znamená, že žádný objekt ze skupiny hraničních řek se nevyskytuje ve všech zdrojích, které obsahují alespoň jeden prvek skupiny. Mezi další objekty, které jsou reprezentované ve větším počtu (z hlediska počtu uzlů v grafu se pohybují ve zkoumané skupině na prvních třech místech) zdrojů patří řeky Jordán[122] – 11 zdrojů, Torne[123] – 9 zdrojů a Zambezi[124] – 9 zdrojů. Minimální počet uzlů v grafu je 3. Tohoto čísla dosahuje 7 objektů – řeky Atrek, Cuiari, Kong, Moei, Pigeon Bay, Sharda, Tumen[125]. Až na jedinou výjimkou (objekt `Pigeon_Bay`) se jedná kombinaci zdrojů DBpedia, Wikidata a Yago (v případě řeky Pigeon Bay je Yago nahrazené GeoNames.org). Z hlediska lokalizace tyto v oblasti propojených dat méně zmiňované vodní toky leží především v Asii, pouze Cuiari protéká Jižní Amerikou, jak již bylo uvedeno výše, Pigeon Bay je severoamerickou řekou. Průměrný počet uzlů v grafu je 5. Z tohoto čísla je zřetelné, že ačkoli je rozdíl mezi maximem a minimem poměrně velký (viz Obrázek 19, který ukazuje srovnání grafů pro nejlépe a nejhůře popsanou hraniční řeku z hlediska počtu zdrojů propojených dat), hodnoty velikosti grafu nesplňují normální rozdělení a ve vzorku převažují spíše grafy s menší velikostí.

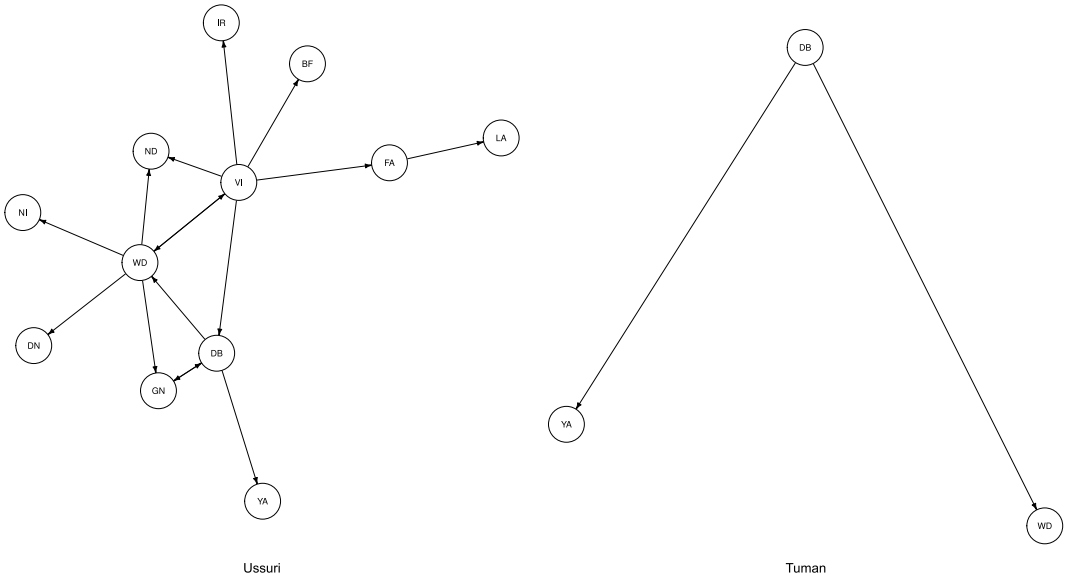
Rozdělení hraničních vodních toků podle velikosti grafů poukazuje na stav v oblasti propojených dat, kdy tyto datové sady jsou připravovány především v Evropě, případně ve Spojených státech amerických, a tudíž obsahují především objekty z těchto oblastí nebo jsou takové prvky popsané podrobněji. To je patrné z faktu, že mezi nejméně popsanými vodními toky je například řeka Tuman, jejíž délka toku a rozloha povodí je mnohem větší než u některých podrobněji popsaných evropských řek (například Tana mezi Norskem a Finskem). Na druhou stranu je potřeba uvést, že většina hraničních řek, jejichž grafy dosahují vysokého počtu uzlů, jsou nějakým způsobem výjimečné

[122]Hranice mezi Izraelem, Jordánskem a Palestinským státem.

[123]Hranice mezi Švédskem a Finskem. Řeka Torne je proslulá svojí birfurkací – rozdělení koryta a odváděním vody do dvou různých říčních systémů.

[124]Hraniční řeka oddělující Zimbabwe, Botswanu, Namibii a Zambii.

[125]V češtině se tento vodní tok tvořící přirozenou hranici mezi KLDR, Čínou a Ruskem nazývá Tuman.



Obrázek 19: Ukázka grafů pro řeky Ussuri a Tuman.

– Jordán z hlediska historie a významu pro různá náboženství, Zambezi z pohledu Viktoriiných vodopádů a Torne kvůli výše zmíněné bifurkaci.

Hustota a průměr grafu jako další grafové metriky dosahují hodnot, které na první pohled korespondují s počtem uzlů v grafech. K potvrzení tohoto faktu byl vypočtený Pearsonův korelační koeficient (Tabulka 10), který ukazuje závislost mezi dvěma skupinami dat. Hodnota 1 znamená přímou závislost, hodnota -1 nepřímou závislost a hodnoty kolem 0 vyjadřují nezávislost.

Tabulka 10: Pearsonův korelační koeficient pro potenciálně závislé grafové metriky.

Grafové metriky	Pearsonův korelační koeficient
Velikost a hustota	-0,78
Velikost a reciprocita	0,01

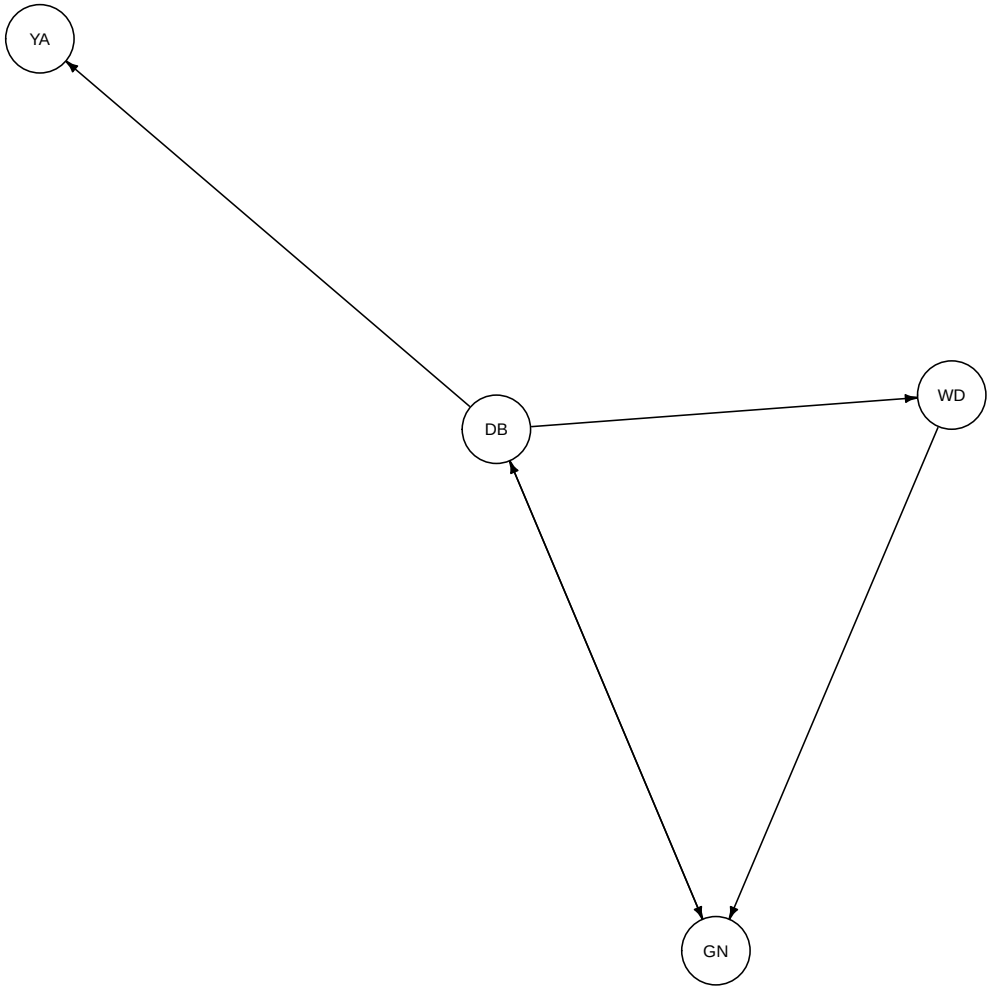
Z tabulky 10 je patrná nepřímá závislost podle hodnot Pearsonova korelačního koeficientu mezi velikostí grafu a hustotou (počet vazeb vztažený k maximálnímu možnému počtu vazeb). To znamená, že grafy s nižším počtem uzlů vykazují vyšší hustotu vazeb než rozsáhlejší grafové struktury. Hodnota 0,01 ukazuje na nezávislost reciprocity a velikosti grafu. Maximální i minimální hodnoty reciprocity (0,67 a 0) dosahují grafy o třech uzlech.

Podobně na velikosti je nezávislý i shlukový koeficient (hodnota 0,01). Tento koeficient se pohybuje mezi hodnotami 0 (řeky Salween – 8 uzlů a Tumen – 3 uzly) a 0,6 (20 vodních toků s počtem uzlů 4 a 5, shluky se obvykle vytváří mezi uzly DB, WD a GN, jak je ukázáno na obrázku 20). Oproti skupinám dat v předchozích experimentech je průměrná hodnota shlukového koeficientu skutečně vyšší (Tabulka 11). To se týká i maximálních hodnot (Tabulka 11).

Tabulka 11: Srovnání shlukového koeficientu.

Shlukový koeficient	Středoevropská hlavní města	Evropské mezinárodní silnice	Hraniční řeky
Průměr	0,22	0	0,37
Maximum	0,26	0	0,60

První vstupní předpoklad se týkal variability hodnot metrik, která by měla být v případě hraničních řek vyšší než u obou předchozích experimentů. Důvodem je fakt, že hraniční řeky tvoří globální datovou sadu, která s velkou pravděpodobností na rozdíl od evropských mezinárodních silnic a středoevropských hlavních měst, které jsou lokálně omezené, nebyla do zdrojů propojených dat importována jako celek, a proto by se úroveň popisu, včetně zavedení identických vazeb, měla mezi jednotlivými objekty více lišit. Tento dohad potvrzuje tabulka 13 zaměřená na grafové metriky – počtu uzlů v grafu (N), hustota grafu (D), reciprocita vazeb (R) a shlukový koeficient (CC). Zatímco tabulka 12, která obsahuje celkové hodnoty váženého součtu po provedení vícekriteriální analýzy aplikované na uzlové metriky, obsahuje mnohem vyšší hodnotu variačního koeficientu v případě hlavních měst střední Evropy.



Obrázek 20: Datová síť objektu `Iguazu_River`.

Tabulka 12: Hodnoty variačního koeficientu váženého součtu uzlových metrik ve skupinách dat – středoevropská hlavní města, evropské mezinárodní silnice a hraniční řeky.

Skupina dat	Variační koeficient
Středoevropská hlavní města	0,84
Evropské mezinárodní silnice	0,26
Hraniční řeky	1,23

Tabulka 13: Hodnoty variačního koeficientu pro grafové metriky ve skupinách dat – středoevropská hlavní města, evropské mezinárodní silnice a hraniční řeky.

Skupina dat	S	D	R	CC
Středoevropská hlavní města	0,17	0,20	0,09	0,11
Evropské mezinárodní silnice	0,14	0,18	-	-
Hraniční řeky	0,36	0,31	0,56	0,61

Na základě tabulky 11 lze potvrdit vyšší výskyt shluků (viz druhý vstupní předpoklad). Není však možné výskyt shluků připisovat na úkor lokálních datových sad pocházejících z různých států okolo hraničního vodního toku. Ze 14 datových zdrojů je možné 5 označit jako lokální. Většinou se jedná o národní knihovny nebo knihovny významných institucí. Tyto zdroje pocházejí z USA (zdroj označeny jako LA), Německa (DN), Francie (BF), Izraele (NI) a Japonska (ND). Pomocí některého z těchto „národních“ zdrojů jsou popsány pouze 3 objekty z dvacítky grafů s nejvyšším shlukovým koeficientem, přičemž ani v jednom případě neexistuje vztah mezi státem, kterým vodní tok protéká, a zemí, kde byla datová báze vytvořená. Oba výchozí předpoklady se tedy podařilo potvrdit pouze částečně.

Uzlová letiště

Téma uzlových letišť (hub airports) je velice úzce spojené se síťovými analýzami. Jedná se o dominantní uzly v síti letišť, které jsou jednak spojené s ostatními uzlovými letišti přímými linkami, a jednak poskytují dopravu (přímá spojení) do méně významných míst ve svém okolí. Jinými slovy, uzlová letiště jsou důležitými přestupními body, především při dálkových letech. Jednotlivé aerolinie (případně tzv. aliance leteckých společností) využívají pro své účely síť uzlových letiště optimalizovanou tak, aby minimalizovaly náklady.

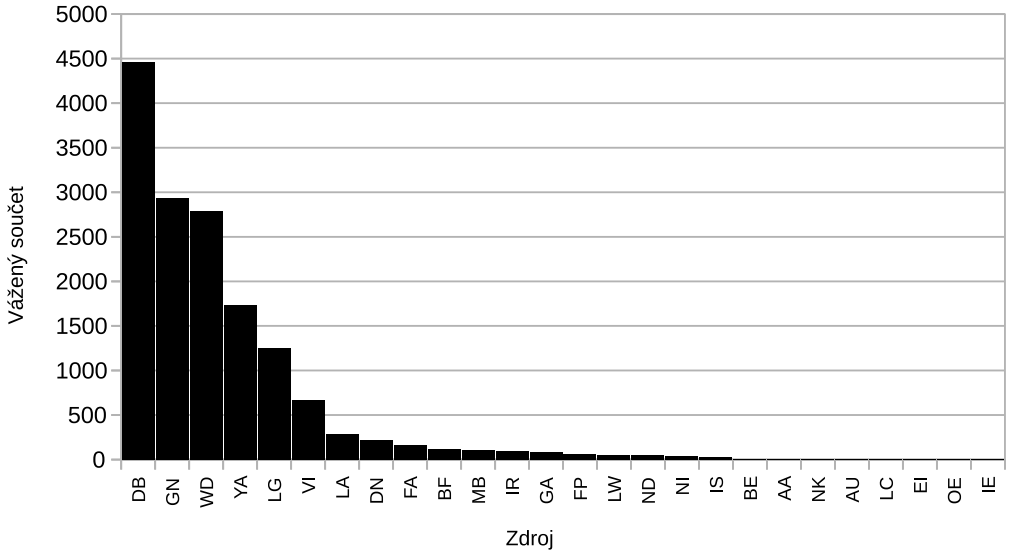
Sada uzlových letišť získaných na základě SPARQL dotazu během extrakce dat o identických vazbách mezi prvky čítá 1300 objektů[126] rozložených po celém světě. Jedná se tedy o poměrně rozsáhlou datovou sadu, jejíž zpracování trvá řádově jednotky minut s použitím běžné výpočetní techniky.

Výchozí předpoklad se v tomto případě týká váhy jednotlivých kritérií (metrik) – „Při porovnání výsledků získaných pomocí kritérií s jednotkovou váhou a kritérií rozdělenými podle významu (na základě Saatyho metody) nedojde k významným rozdílům, protože zvolené metriky (jako kritéria) si jsou navzájem rovné.“ V rámci studie jsou tedy nejprve klasifikovány metodou váženého součtu s jednotkovými vahami všechny uzly (zdroje propojených dat) a grafy (objekty propojených prostorových dat). Poté jsou pomocí Saatyho metody zvoleny váhy jednotlivých kritérií a následně opět provedena multikriteriální analýza. V závěru jsou oba výsledky porovnány.

Při analýze uzlů bylo objeveno celkem 26 různých zdrojů prostorových propojených dat, které obsahovaly reprezentace objektů se skupiny uzlových letišť. Z grafu na obrázku 21 je patrné dominantní postavení databáze DBpedia, která je podobně jako v předchozích studiích následována datovou sadou Wikidata. Rozdílem oproti dříve provedeným experimentům je dvoutřetinová hodnota celkového váženého součtu při srovnání produktů Wikidata a DBpedia (dříve realizované studie ukazují, že si jsou obě hodnoty velice blízké). Dokonce vyšší hodnota než v případě zdroje Wikidata byla

[126]Výsledky analýzy byly získány pouze pro 1072 objekty, které obsahovaly identické vazby mezi zdroji.

zjištěna i pro databázi GeoNames.org. Další parametry z analýzy opět nejsou překvapivé. Na čelních místech, ale již s nízkými hodnotami ve vícekriteriální analýze se nachází zdroje jako Yago a VIAF. Z hlediska tohoto hodnocení může být zajímavá pouze pozice datové sady LinkedGeoData (LG), která v případě uzlových letišť dosahuje vyšších hodnot než VIAF. Z obrázku 21 je také patrné, že tvar grafu připomíná exponenciálu, což je opět společný rys pro již realizované studie.



Obrázek 21: Hodnoty váženého součtu pro zdroje dat ve skupině uzlových letišť.

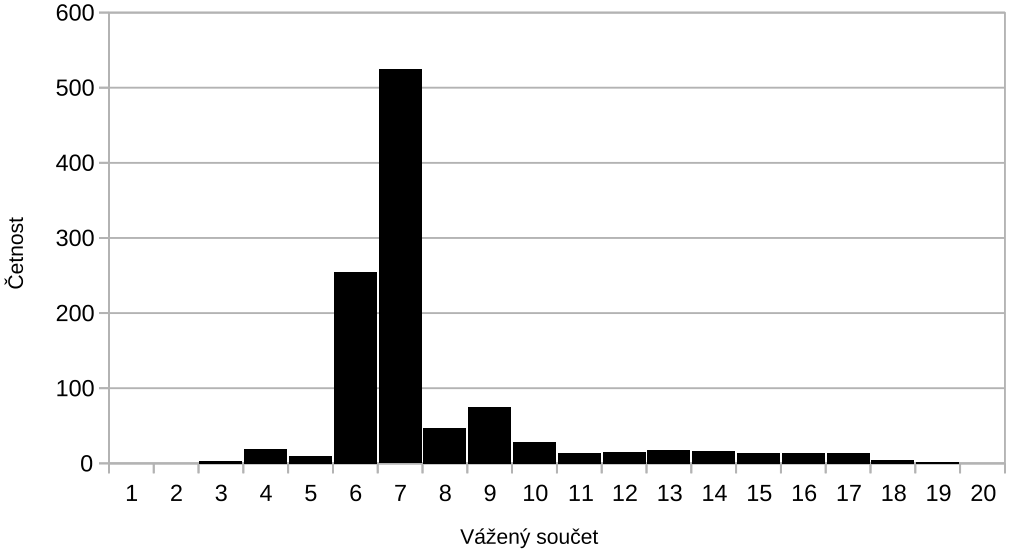
Z hlediska grafových metrik, tedy úrovně propojení reprezentací prvků dat pomocí identických vazeb, je možné dojít k následujícím závěrům:

- Existují malé rozdíly mezi nejlépe a nejhůře propojenými objekty. Maximální hodnota váženého součtu je 16,68 a minimální 2,50.
- Hodnota variačního koeficientu je 0,37, je tedy možné říct, že se směrodatná odchylka podílí na průměru z více než jedné třetiny.
- Jak ukazuje graf četnosti (Obrázek 22), data neodpovídají normálnímu rozdělení (na vině je především vysoký počet hodnot mezi 5 a 7).
- Nejvyšších hodnot dosahují významná letiště ve světových metropolích (Paříž, New York, Tel Aviv, Sidney, Londýn[127]). Mezi tato města se

[127]Ve skutečnosti se nejedná o data letišť v těchto městech, ale přímo o data měst, která jsou z nějakých (evidentně chybných) důvodů řazena do kategorie uzlových letišť.

vklínila (na 4. místě z hlediska pořadí podle celkového váženého součtu) také Budapešť.

- Datový vzorek a jeho analýzy reflektuje hierarchii letišť i na opačném konci, kde jsou výstupní hodnoty multikriteriální analýzy nejnižší.



Obrázek 22: Graf četnosti hodnot váženého součtu grafových metrik ve skupině uzlových letišť.

Dalším krokem je stanovení vah jednotlivých kritérií použitých v rámci vícekriteriální analýzy. Pro tento účel je použita Saatyho metoda [144] (její vazba na multikriteriální analýzu je zmíněna například v publikacích [145–147]). Tato metoda spočívá v sestavení Saatyho matice (matice relativních důležitostí, S). Tato čtvercová matice, v níž sloupce i řádky vyjadřují jednotlivá kritéria (ve stejném pořadí), obsahuje hodnoty znamenající vztah jednotlivých kritérií. Pro matici platí následující pravidla:

- Prvky na diagonále mají hodnotu 1.
- Vztah mezi libovolnými prvky $s_{i,j}$ a $s_{j,i}$ (inverzní prvky matice) je

$$s_{i,j} = \frac{1}{s_{j,i}}.$$

- Hodnota $s_{i,j}$ vyjadřuje vztah mezi i-tým a j-tým kritériem. Tento vztah vychází z expertního hodnocení, v němž by měly být identifikovány

následující vazby mezi kritérii (české termíny podle [148] a [149]):

- Kritéria jsou stejně významná – rovnocennost – hodnota $s_{i,j} = 1$.
- Kritérium s_i je slabě významnější než kritérium s_j – slabá preference – hodnota $s_{i,j} = 3$.
- Kritérium s_i je dosti významnější než kritérium s_j – silná preference – hodnota $s_{i,j} = 5$.
- Kritérium s_i je prokazatelně významnější než kritérium s_j – velmi silná preference – hodnota $s_{i,j} = 7$.
- Kritérium s_i je absolutně významnější než kritérium s_j – absolutní preference – hodnota $s_{i,j} = 9$[128].

Po sestavení matice následuje vypočítání geometrického průměru pro každý řádek a následná normalizace těchto hodnot. Normalizované veličiny pak představují váhy pro jednotlivá kritéria.

Pro zjištění váhy kritérií reprezentujících uzlové i grafové metriky jsou použity tři hlavní vstupy:

1. Korelace zjištěné v předchozích experimentech – Jsou-li na sobě nějaké hodnoty závislé, pak by minimálně jedna měla být při rozhodování oslabena, aby nedošlo k nadhodnocování jednoho parametru grafu na úkor jiných.
2. Metriky použité v podobných studiích [15, 32, 35] – tyto otestované metriky by měly mít vyšší váhu.
3. Zkušenosti autora – metriky důležité pro použití v oblasti prostorových dat[129].

Výsledkem prvního kroku aplikace Saatyho metody jsou dvě matice – S_n (pro uzlové metriky – centralita stupně, centralita blízkosti, centralita mezilehlosti, skóre autority, skóre středu a Pege Rank) a S_g (pro grafové metriky – počet

[128] Pro jemnější rozlišení vztahů parametrů se dají použít i sudé hodnoty 2, 4, 6 a 8.

[129] V tomto případě došlo především k upřednostnění následujících metrik: Page Rank (vazba na důležité zdroje dat), centralita mezilehlosti (existence propojení mezi do značné míry izolovanými skupinami zdrojů), skóre autority (zdroj s vysokou hodnotou bude patrně respektovaný ve sféře prostorových dat), reciprocita (vlastnost důležitá pro spojitost grafu, tedy pro získávání informací napříč zdroji), shlukový koeficient (existence různých pohledů na jeden prostorový objekt, jež mohou vést ke získání nových informací a souvislostí) a počet uzlů v grafu (do jaké míry je objekt začleněný do struktury propojených dat).

uzlů, hustota, reciprocita a shlukový koeficient).

$$S_n = \begin{bmatrix} 1 & 1 & \frac{1}{5} & 3 & 7 & \frac{1}{3} \\ 1 & 1 & \frac{1}{5} & 3 & 7 & \frac{1}{3} \\ 5 & 5 & 1 & 7 & 9 & 3 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{7} & 1 & 5 & \frac{1}{5} \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{9} & \frac{1}{5} & 1 & \frac{1}{8} \\ 3 & 3 & \frac{1}{3} & 5 & 8 & 1 \end{bmatrix}$$

$$S_g = \begin{bmatrix} 1 & 3 & \frac{1}{3} & \frac{1}{5} \\ \frac{1}{3} & 1 & \frac{1}{5} & \frac{1}{7} \\ 3 & 5 & 1 & \frac{1}{3} \\ 5 & 7 & 3 & 1 \end{bmatrix}$$

Po vypočítání geometrického průměru v řádcích obou matic a normalizaci hodnot jsou získány vektory vah, které ukazují tabulky 14 (pro uzlové metriky) a 15 pro grafové metriky.

Tabulka 14: Hodnoty vah pro uzlové metriky.

Metrika	Váha
Centralita stupně	0,12
Centralita blízkosti	0,12
Centralita mezilehlosti	0,45
Autorita	0,05
Střed	0,02
Page Rank	0,24

Tabulka 15: Hodnoty vah pro grafové metriky.

Metrika	Váha
Velikost grafu (počet uzlů)	0,12
Hustota grafu	0,06
Reciprocita	0,26

Metrika	Váha
Shlukový koeficient	0,56

Tabulka 16 poskytuje srovnání výsledků váženého součtu uzlových metrik po (sloupec Saatyho metoda) a před (sloupec Váha 1) aplikacích vah získaných v předchozím kroku. Kromě změn v absolutních hodnotách, které jsou dány především normalizací výsledků, nedošlo k žádným zásadním proměnám výstupů z multikriteriální analýzy pro uzly. Při srovnání obou sloupců lze zaznamenat drobné odchylky v pořadí (absolutní hodnoty se pochopitelně odlišují). Změny se však týkají pouze zdrojů marginálních z pohledu testované skupiny dat.

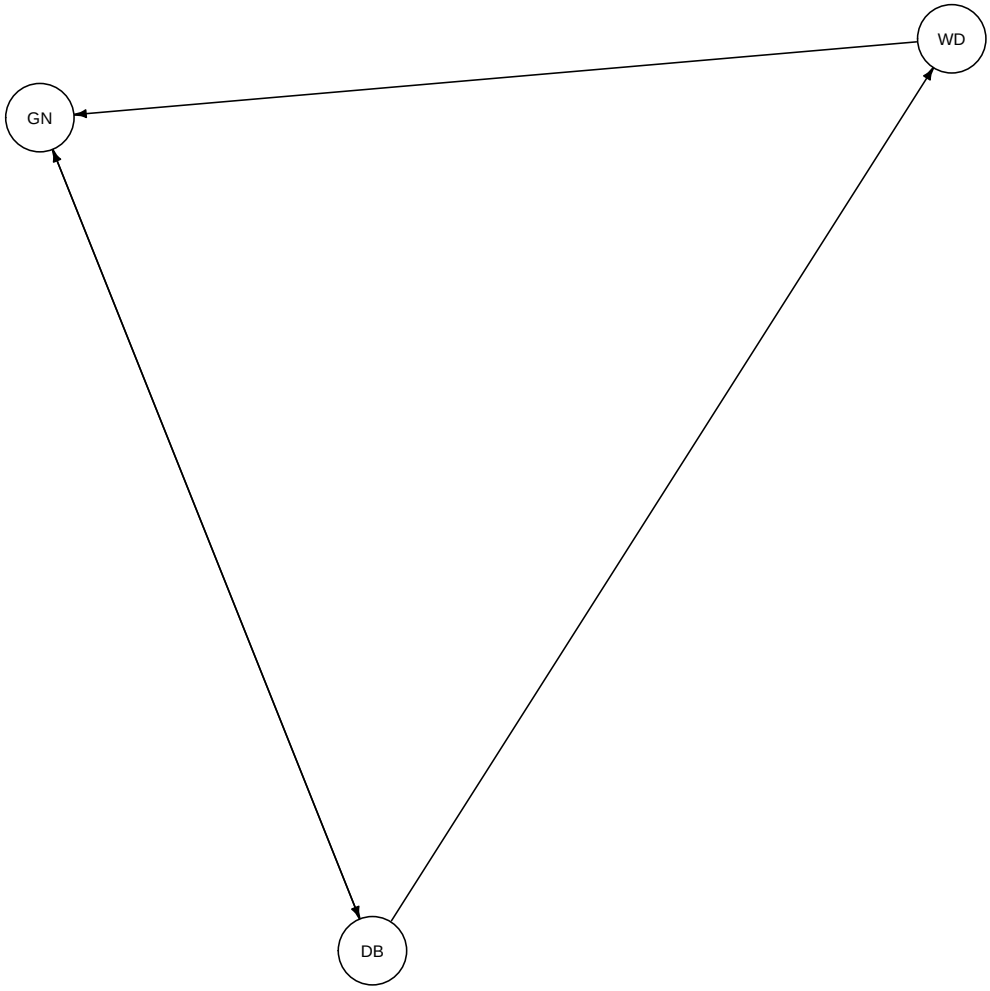
Tabulka 16: Srovnání výsledků aplikace vah pro kritéria do výpočtu celkové hodnoty váženého součtu uzlových metrik.

Zdroj	Saatyho metoda	Bez vah
DB	4459,39	1087,23
GN	2936,35	700,14
WD	2788,83	581,71
YA	1736,09	357,62
LG	1247,48	203,82
VI	664,13	150,17
LA	283,57	62,45
DN	221,44	46,81
FA	165,20	40,62
BF	114,97	23,44
MB	111,32	28,11
IR	90,56	16,30
GA	87,69	20,30
FP	59,58	14,35
LW	51,33	11,05
ND	47,01	7,21
NI	37,43	8,16
IS	26,36	3,91

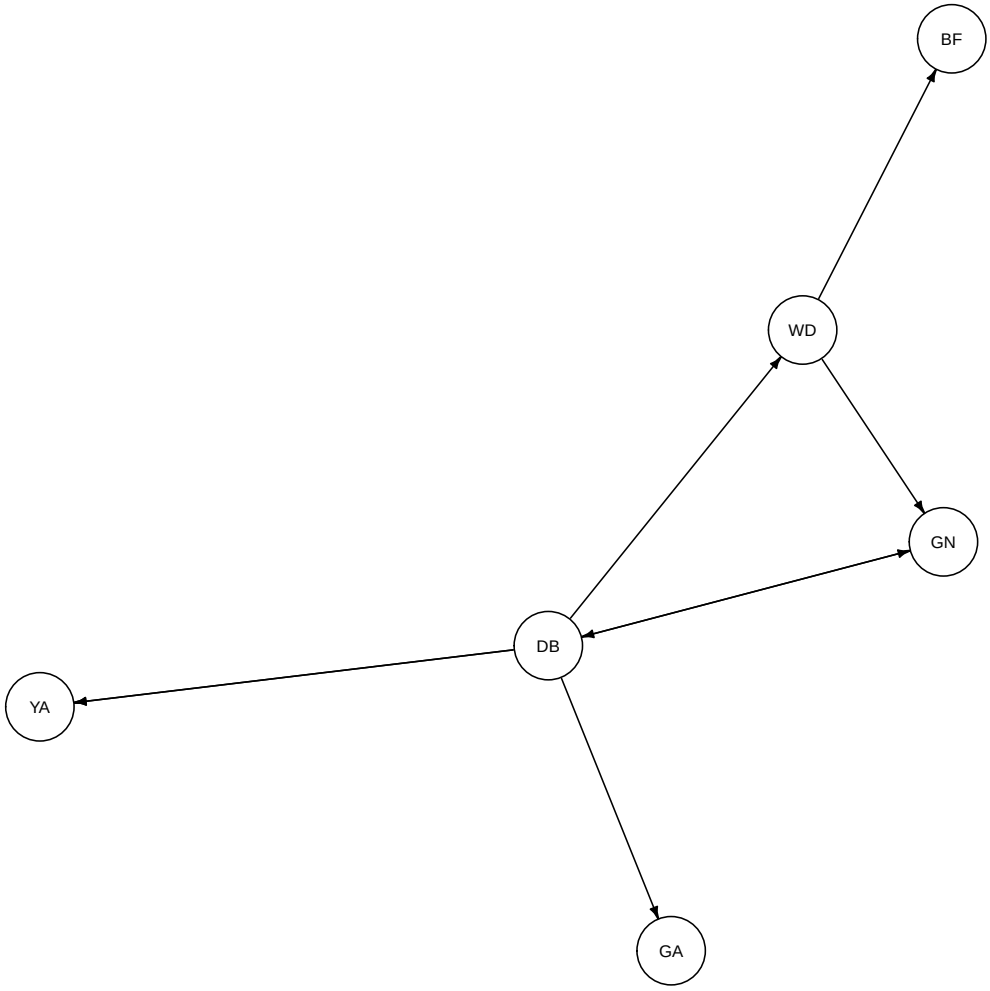
Zdroj	Saatyho metoda	Bez vah
BE	5,24	2,12
AA	4,28	0,45
NK	3,97	1,59
AU	2,54	1,08
LC	0,97	0,97
EI	0,81	0,09
OE	0,62	0,07
IE	0,51	0,07

Větší odchylky byly zaznamenány při analýze grafů. Zde měly změny vah za následek přesun v pořadí až 95% objektů. Maximální změna v pořadí byla 786 příček a o 18% vzhledem k maximu. Obrázky 23 a 24 ukazují význam posílení shlukového koeficientu ve výpočtu hodnot multikriteriální analýzy. Graf 23 zaznamenal největší vzestup z hlediska pořadí, protože se jedná o souvislou strukturu obsahující velké procento recipročních vazeb. Naopak druhý graf 24 v celkovém pořadí klesl díky tomu, že obsahuje 50% listových uzlů.

Úvodní předpoklad této studie, který tvrdil, že po zavedení metrik nedojde k výrazným změnám, se potvrdil. Sice pojem „výrazná změna“ nebyl přesně specifikován, ale z předchozích odstavců a tabulky 16 je patrné, že změny byly pouze drobné a nedotýkaly se extrémních hodnot. Jinými slovy, výsledky pro prvky (zdroje, objekty), které by mohly být považovány za příklady dobré praxe, byly pomocí obou druhů vah vyhodnoceny totožně.



Obrázek 23: Datová síť objektu `Nabire_Airport`.



Obrázek 24: Datová síť objektu `Jambi`.

Republiky

Tato případová studie pracuje se skupinou dat, které mají jako společný jmenovatel státní zřízení ve formě republiky. SPARQL dotaz nalezl v databázi DBpedia celkem 130 takových položek[130], které byly dále zpracovány. Účelem této studie je najít vhodné dílčí uzlové metriky pro konkrétní aspekty kvality prostorových propojených dat. Jinými slovy, výsledkem by měly být zdroje propojených dat, u nichž hodnoty uzlových metrik predikují určité vlastnosti (například spolehlivost nebo různorodost poskytovaných informací). Tento experiment je poněkud atypický a vymyká se ostatním případovým studiím publikovaným v této práci. Neobsahuje totiž zavedení a zhodnocení výchozího předpokladu ani zpracování grafových metrik. Pro výpočet váženého součtu v rámci multikriteriální analýzy jsou použity váhy zjištěné v části Uzlová letiště.

Dílčí uzlové metriky byly vybrány na základě toho, že indikují důležité vlastnosti pro vrcholy sítě, přičemž tyto vlastnosti se dají transponovat na vlastnosti zdrojů propojených prostorových dat. Následující seznam obsahuje výčet metrik a vliv jejich hodnot na zdroje dat:

Centralita stupně (C_d) Určuje míru 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. Existuje velká pravděpodobnost, že s využitím těchto zdrojů lze dosáhnout na velké množství různorodých informací o objektu.

Centralita blízkosti (C_c) Jak vyplývá z tabulky 14, centralitě blízkosti byla po realizaci expertního hodnocení přiřazená poměrně nízká váha. Je to dáno tím, že centralita blízkosti je spojená s rychlostí dosažitelnosti uzlů v grafech. Vzhledem k tomu, že počet zdrojů propojených prostorových dat se pohybuje zhruba v desítkách (ve výzkumu provedeném v rámci této práce bylo nalezeno 66 různých zdrojů) a navíc grafy nejsou příliš složité, nehraje hodnocení prostupnosti sítě příliš velkou roli. Je však nutné tuto metriku vést v patrnosti, především pro případy složitějších grafů nebo získávání většího množství dat.

[130]Nejedná se ve všech případech o republiky jako nezávislé státy.

Centralita mezilehlosti (C_b) Centralita mezilehlosti indikuje uzly, které tvoří „mosty“ mezi více či méně izolovanými součástmi grafu. Z hlediska využívání propojených dat nemá tato metrika velký význam pro uživatele z hlediska toho, zda je a pro jaké účely je vhodné zdroj používat. Je však nutné si uvědomit, že takový zdroj představuje tzv. „úzké hrdlo“ a jakékoli jeho narušení může vést až ke kolapsu celého systému. Jinými slovy v případě technické chyby, kdy zdroj nebude dostupný, nebude ani možné prohledávání datové sítě pro daný objekt. Je nutné zmínit i důraz na kvalitu obsahu takového datového zdroje. Vzhledem k jeho poloze v síti ho budou uživatelé jistě často zpracovávat, a proto je nutné si uvědomit, že případné chyby budou ve velké míře přebírány i do uživatelských řešení.

Autorita (skóre autority, A) Skóre autority je vysoké pro takové vrcholy grafu, na které přímo odkazuje velké množství ostatních uzlů. V případě sítě propojených dat je možné zdroje s vysokým skóre autority označit jako populární, často využívané a vzhledem k tomu, že i při používání dat platí tržní principy, budou takové zdroje s vysokou mírou pravděpodobnosti i kvalitnější než ostatní poskytovatelé dat. V případě prostorových dat je možné za aspekty kvality brát především dostupnost a přesnost souřadnic, homogenitu dat, kompletnost a vazbu na jiné (nepropojené) datové zdroje.

Střed (hub, skóre středu, H) Střed je opakem autority, tedy takovým uzlem, z něhož vychází velké množství hran směrem k ostatním vrcholům. Zdroj propojených prostorových dat (reprezentovaný uzlem v síťovém grafu) s vysokým skóre středu lze považovat za vhodný začátek pro procházení sítě propojených dat. S jeho pomocí je možné vytvořit si o studovaném prostorovém objektu komplexní představu, neboť může odkazovat na datové sady poskytující jiné (doplňkové, rozšiřující) informace.

Page Rank (PR) Algoritmus Page Rank kladně hodnotí přímá propojení s „kvalitními“ uzly. Kvalita v tomto případě znamená, že zdroje jsou silně integrované do sítě propojených dat. To znamená, že jsou přímo spojené s dalšími takovými zdroji dat. Jinými slovy Page Rank bude

lépe hodnotit uzel spojený s několika málo vrcholy, které budou mít velké množství dalších vazeb, než uzel propojený s velkým množstvím listových vrcholů. Tento parametr může být nápomocný především pro orientaci mezi ostatními metrikami stejně nebo mezi podobně hodnocenými zdroji.

Tabulka 17: Průměrné hodnoty síťových metrik ve skupině data Republiky.

Zdroj	C_d	C_c	C_d	A	H	PR
AA	0,00	0,02	0,00	0,01	0,00	0,00
AU	0,00	0,02	0,00	0,01	0,00	0,00
BE	0,00	0,04	0,00	0,03	0,00	0,00
BF	0,10	0,36	0,00	0,37	0,00	0,04
DB	0,77	0,70	0,27	0,69	0,81	0,16
DN	0,09	0,41	0,00	0,36	0,00	0,04
EI	0,01	0,06	0,00	0,04	0,02	0,00
ES	0,16	0,41	0,00	0,55	0,00	0,06
FA	0,15	0,34	0,00	0,20	0,22	0,03
FP	0,00	0,00	0,00	0,00	0,00	0,00
GA	0,05	0,37	0,00	0,22	0,00	0,04
GN	0,35	0,54	0,03	0,91	0,17	0,09
IE	0,00	0,00	0,00	0,00	0,00	0,00
IR	0,11	0,35	0,00	0,31	0,07	0,03
IS	0,05	0,26	0,00	0,17	0,00	0,03
LA	0,15	0,39	0,00	0,41	0,03	0,04
LC	0,00	0,01	0,00	0,01	0,00	0,00
LG	0,16	0,39	0,00	0,24	0,35	0,04
LW	0,05	0,10	0,00	0,06	0,10	0,01
MB	0,05	0,39	0,00	0,21	0,00	0,03
ND	0,09	0,41	0,00	0,39	0,00	0,04
NI	0,04	0,27	0,00	0,16	0,00	0,02
NK	0,00	0,03	0,00	0,03	0,00	0,00
NM	0,00	0,00	0,00	0,00	0,00	0,00
OE	0,05	0,38	0,00	0,23	0,00	0,04
TI	0,30	0,41	0,02	0,37	0,49	0,05
VI	0,49	0,58	0,10	0,30	0,66	0,06

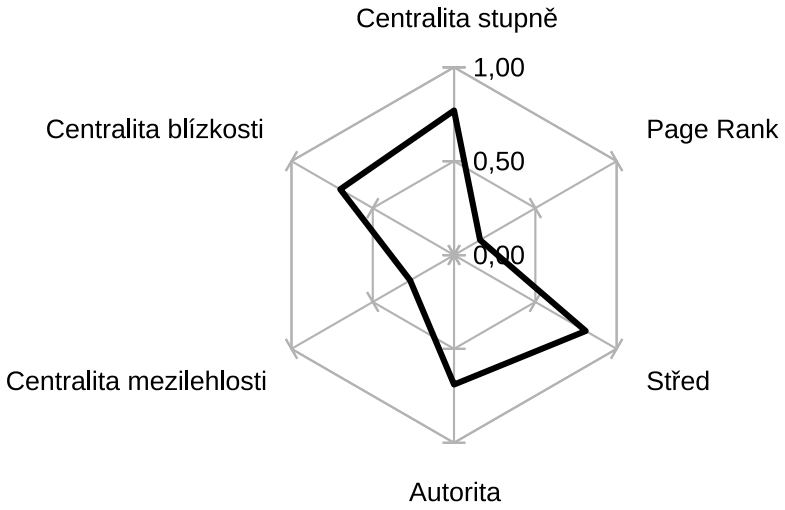
Zdroj	C_d	C_c	C_d	A	H	PR
WB	0,25	0,41	0,00	0,14	0,54	0,03
WD	0,55	0,67	0,16	0,47	0,71	0,06
YA	0,08	0,41	0,00	0,29	0,00	0,05

Průměrné výsledky uzlových metrik pro skupiny zahrnující republiky jsou shrnuty v tabulce 17. Z ní vyplývají následující fakta:

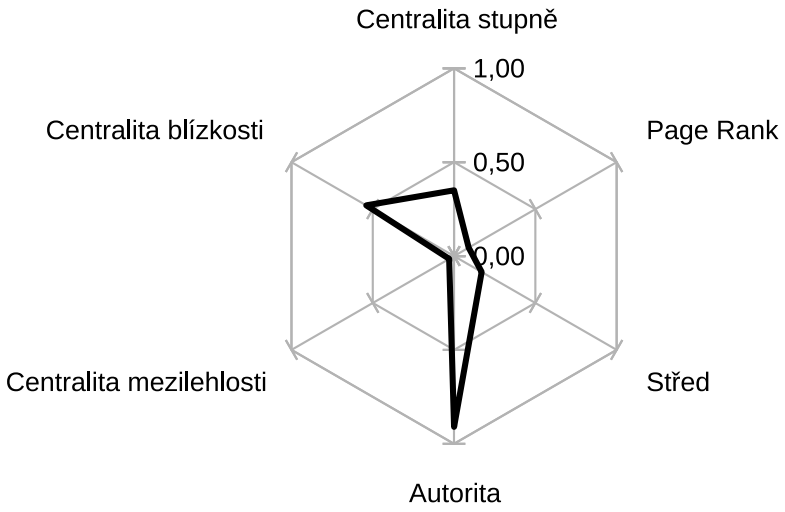
- Pouze 5 zdrojů (ze 30 zastoupených) má nenulové hodnoty ve všech metrikách.
- 10 zdrojů se umístilo v pořadí metrik na prvních pěti místech.
- Z hlediska identických vazeb, ve kterých je zdroj objektem nebo subjektem, se mezi nejlepší zdroje řadí DBpedia (DB), Wikidata (WD), VIAF (VI), GeoNames.org (GN) a Transparency International (TI) (v tomto pořadí).
- Podle blízkosti k ostatním zdrojům patří mezi nejvhodnější zdroje DB, WD, VI, GN a na páté příčce se umístily Deutschen Nationalbibliothek (DN), Eurostat Linked Statistics (ES) National Diet Library (ND) a TI.
- Z hlediska mezilehlosti jsou dominantními zdroji DB, WD, VI, GN a TI. Toto pořadí odpovídá i pořadí alternativ po aplikaci multikriteriální analýzy.
- Největší autoritu mezi zdroji představují GN, DB, ES, WD, Library of Congrees Authority File (LA) (v tomto pořadí). Zvlášť v případě GN je z obrázku 26 vidět, že se jedná o typickou autoritu – zdroj, který je často odkazován, ale sám poskytuje minimum referencí.
- Nejvyšší průměrné skóre středu je zaznamenáno u zdrojů DB, WD, VI (z obrázku 28 je vidět, že tento zdroj se specializuje na poskytování odkazů), World Bank (WB) a TI.
- Z hlediska hodnocení Page Rank se na předních místech umístily zdroje DB, GN a dále se shodným skóre ES, VI a WD.

V první řadě je třeba poznamenat, že státy představují specifickou skupinu propojených prostorových dat. To je zřetelné jednak ze zastoupení zdrojů, které poskytují převážně statistické informace o jednotlivých zemích (ES, TI,

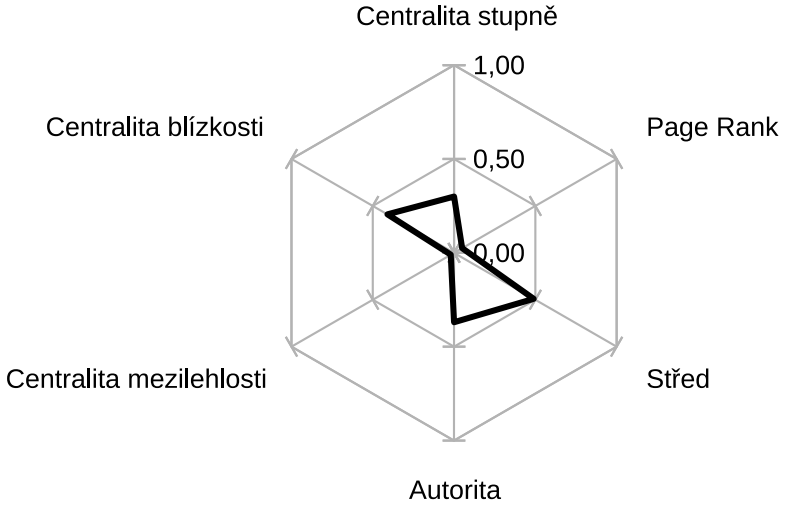
WB). Jejich provozovateli bývají respektované organizace jako Světová banka nebo Eurostat. To, že se nejedná o marginální data (jako v případě jiných typů propojených prostorových dat zkoumaných v předchozích experimentech), dokládají přední pozice těchto zdrojů ve sloupcích tabulky 17. Vysoké hodnoty byly samozřejmě zaznamenány i u stejných zdrojů jako v předešlých případových studiích, kterými jsou DB, WD (na grafech na obrázcích 25 a 29 je zřetelná vyrovnanost ve všech metrikách, to platí i o zdroji TI – obrázek 27, ale ten nedosahuje maximálních hodnot), VI nebo GN. Porovnáním paprskových grafů 25, 29 a 27 vynikne stejný charakter (poměr) hodnot jednotlivých metrik – grafy se liší pouze v absolutních číslech, ale tvar lomené čáry je téměř totožný.



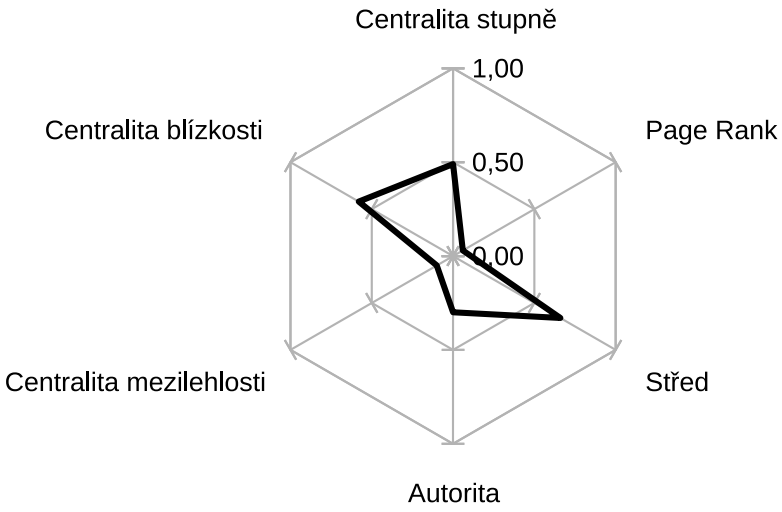
Obrázek 25: DBpedia – hodnoty uzlových metrik.



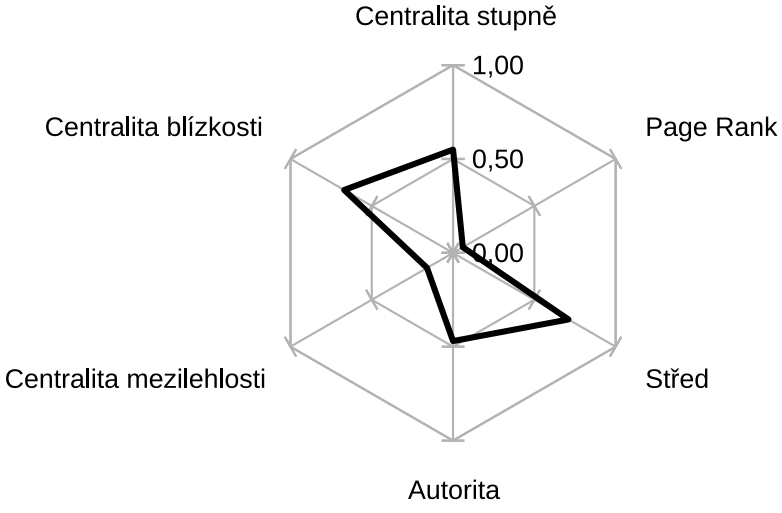
Obrázek 26: GeoNames.org – hodnoty uzlových metrik.



Obrázek 27: Transparency International – hodnoty uzlových metrik.



Obrázek 28: VIAF – hodnoty uzlových metrik.



Obrázek 29: Wikidata – hodnoty uzlových metrik.

Srovnání lokálních a globálních dat

Účelem této studie je ověřit, zda identické vazby jsou stejným způsobem využity pro data, která mají globální rozměr, a data, která jsou spíše místní. Jako odůvodněný vstupní předpoklad se jeví tvrzení – „Globální data, vzhledem ke svému významu, jsou popsána ve více zdrojích Linked Data a také jsou lépe provázána z hlediska identických vazeb.“. Druhé tvrzení, které je v tomto experimentu ověřováno, se týká zdrojů propojených dat – „Zdroje propojených dat a jejich charakteristiky vycházející z uzlových metrik si jsou podobné pro globální i lokální data, která popisují podobné téma.“

Pro testování obou výchozích předpokladů byly zvoleny dvě datové sady. Lokální data reprezentují hory v České republice[131]. Datový vzorek obsahuje celkem 60 položek, ale 44 z nich neobsahuje žádné identické vazby na reprezentace objektu v dalších datových zdrojích propojených dat. Z geografického hlediska je zajímavé, že vazby obsahují především hory v bývalých Sudetech, tedy na hranicích s dnešním Německem a Polskem.

Globální datovou sadu tvoří také podmnožina typu objektu `hora`. V tomto případě nebylo omezení výběru stanoveno lokalizačně, ale tematicky. Zvoleny byly světové stratovulkány, tedy specifické typy sopek, které jsou charakteristické svojí velikostí a kuželovým tvarem. Oproti prvnímu vzorku je tato datová sada rozsáhlejší. Obsahuje celkem 743 objektů (po zpracování zůstalo 664 prvků). I počet zkoumaných prvků odráží vztah mezi lokálním a globálním, kde nutně musí být zaznamenaná vyšší kvantita. Vzhledem k tomu, že se ve studiích pracuje s normovanými hodnotami, není nutné vybírat stejně početné vzorky dat.

Porovnávány jsou (a v dalších dvou studiích budou) nejen výsledky multikriteriální analýzy, ale také vybrané dílčí metriky. V případě uzlů se jedná o Page Rank a centralitu mezilehlosti (jako dvě metriky s nejvyššími

[131]Podobně jako v ostatních případech byla i tato datová sada vybrána pomocí SPARQL dotazu do datové sady DBpedia. Při zběžném pohledu je jasné, že výčet není kompletní (tento problém byl zaznamenaný již v předchozí studii věnované republikám). Účelem této práce však není hodnocení kvality obsahu jednotlivých sad propojených dat, proto tento nedostatek není dále popisován a hodnocen. Pro účely testování vazeb je získaný vzorek dostačující.

vahami, a tudíž s největší důležitostí). K nim byly přidány autorita a střed, které mají význam pro používání propojených prostorových dat.

Tabulka 18: Srovnání parametrů skupin dat Hory v ČR a Stratovulkány.

Kritérium	Hory v ČR	Stratovulkány
Počet zdrojů	7	15
Centralita mezilehlosti	0,44 (DB)	0,30 (DB)
Skóre autority	0,84 (GN)	0,81 (WD)
Skóre středu	1,00 (DB)	0,97 (DB)
Page Rank	0,33 (DB)	0,30 (DB)

Výsledky srovnání uvádí tabulka 18. V řádcích vyjadřujících uzlové metriky jsou uvedeny maximální hodnoty průměru metriky pro jednotlivé uzly.

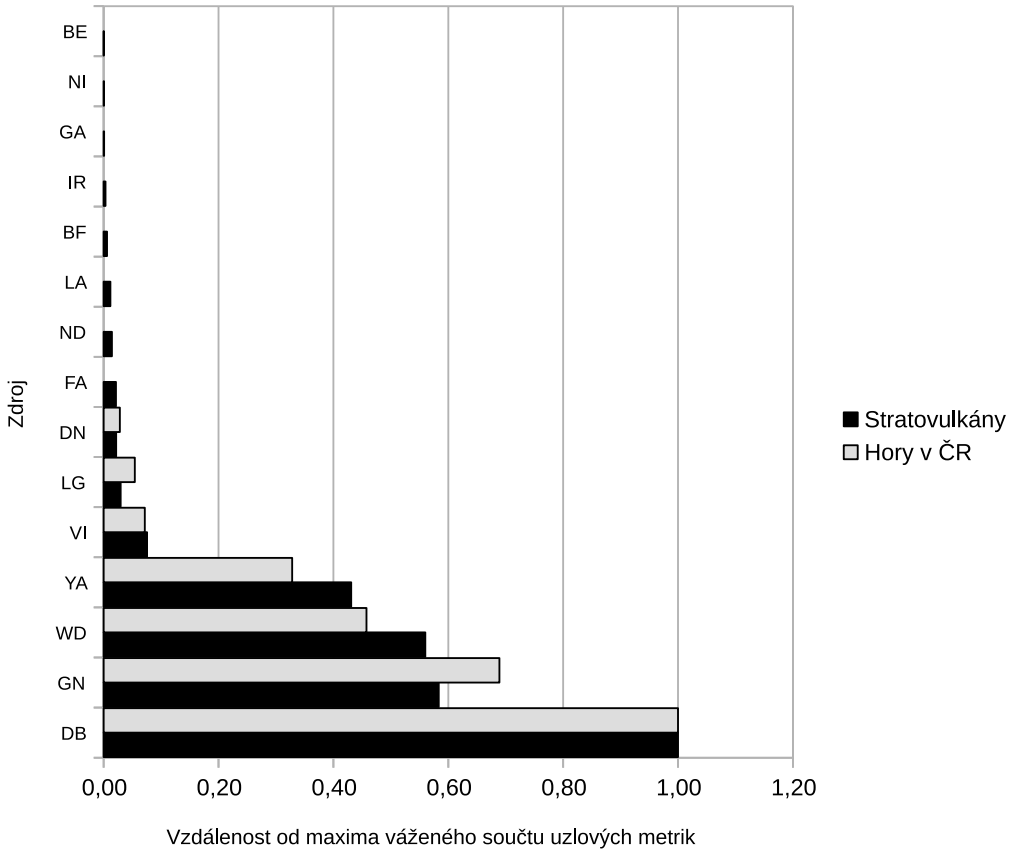
- Hory v Česku se vyskytují pouze v 7 základních zdrojích propojených dat (DB, GN, WD, VI, YA, LG a DN). Zajímavé je, že tato data se nevyskytují ani v jednom zdroji propojených dat, který pochází z České republiky (NK nebo LW). Zastoupeny jsou oba klíčové zdroje pro propojená prostorová data (GN a LG). To platí i pro skupinu stratovulkánů. V ní se objevují zdroje reprezentující národní knihovny a podobné autority (LA, NI, ND, BF nebo DN - Německá národní knihovna obsahuje i údaje o českých horách).
- Maximální hodnota průměrné centrality mezilehlosti je vyšší v případě lokální datové sady, přičemž v obou případech se jedná o zdroj DB. To znamená, že role uzlu DB jako prvku klíčového z hlediska propojení různých částí grafů, je v případě první skupiny dat důležitější než u dat vyjadřujících stratovulkány. V obou datových sadách je stejné pořadí a podobné hodnoty i na dalších místech – GN (0,17 – Hory v ČR / 0,25 – Stratovulkány), WD (0,03 / 0,03).
- Zatímco ve většině předchozích studií se jako největší autorita na poli propojených prostorových dat profilovala datová sada GeoNames.org, v případě stratovulkánů mají vyšší hodnotu skóre autority Wikidata a Yago (0,77). Tyto zdroje představují důležité autority i v tématu Hory

v ČR. Také z hlediska absolutních hodnot si jsou obě témata velice podobná, i když absolutní hodnoty jsou mírně (v řádu setin) vyšší v lokálních datech.

- Z hlediska středu (tedy uzlu / zdroje, který odkazuje na další vrcholy grafu) je výsledek v případě stratovulkánů opět atypický ve srovnání s druhou testovanou skupinou dat, ale i s předchozími experimenty. Nejde ani tak o uzly s maximálními průměrnými hodnotami, tím jsou v obou případech DB a WD, ale o to, že v případě stratovulkánů jsou důležitými středy obě geografické databáze (GN a LG) s hodnotou skóre středu 0,03. Ve všech předchozích studiích GN (a částečně i LG) představovaly autority, ale nikoli huby. Tato situace zřejmě nastala proto, že skupina globálních dat (stratovulkány) obecně obsahuje málo středů (hubů). To je patrné i z absolutních hodnot skóre středu u prvků, které se umístily za dvojicí DB a WD na třetím místě – Hory v ČR – VI (0,12) / Stratovulkány – GN, LG (0,03).
- Trend zaznamenaný u předešlých uzlových metrik pokračuje i případě metody Page Rank. Hodnoty i zdroje dat na prvních třech místech se v obou skupinách víceméně shodují (jde tedy o potvrzení druhého vstupního předpokladu – „Zdroje propojených dat a jejich charakteristiky vycházející z uzlových metrik si jsou podobné pro globální i lokální data, která popisují podobné téma.“). Jediné odlišnosti představují mírně nižší absolutní čísla pro stratovulkány a změna v pořadí, kdy ve skupině Hory v ČR GN má vyšší skóre Page Rank než WD a YA, které v obou případech mají totožnou velikost Page Rank. Lehce nižší průměrné hodnoty všech testovaných uzlových metrik u stratovulkánů (nehrající významnou roli při využívání propojených prostorových dat) jsou pravděpodobně způsobené výrazně větší velikostí vzorku.

Po provedení multikriteriální analýzy jsou opět výsledky v obou skupinách velice podobné (Obrázek 30). Zajímavé je, že sedm zdrojů, ve kterých se objevují české hory, je zároveň (a to ve stejném pořadí) sedm nejlépe hodnocených zdrojů pro stratovulkány. To potvrzuje domněnku získanou v předchozích experimentech, že až na několik skutečně globálních sad propojených dat jako jsou DBpedia, Wikidata nebo GeoNames.org, je

obsah ostatních databází poplatný potřebám institucí, které data vytváří a komplexnost informací a obsahová integrita je podružnou záležitostí.



Obrázek 30: Výsledky multikriteriální analýzy pro uzly skupin Hory v ČR a Stratovulkány.

Z výsledků vícekriteriální analýzy aplikované na grafové metriky jsou vidět jasné rozdíly potvrzující první vstupní předpoklad experimentu – „Globální data, vzhledem ke svému významu, jsou popsána ve více zdrojích Linked Data a také jsou lépe provázána z hlediska identických vazeb.“ Maximální absolutní hodnota 1,93 byla v případě stratovulkánů u sopky Mount St. Helens (Obrázek 32). Nejvyšší hodnota váženého součtu pro české hory je 1,11, ovšem paradoxně ji zaznamenaly dva vrcholy ležící mimo české území: Luzný (Obrázek 31) a Třístoličník[132]. Pokud by byly obě skupiny dat sloučeny, hodnota 1,36 by

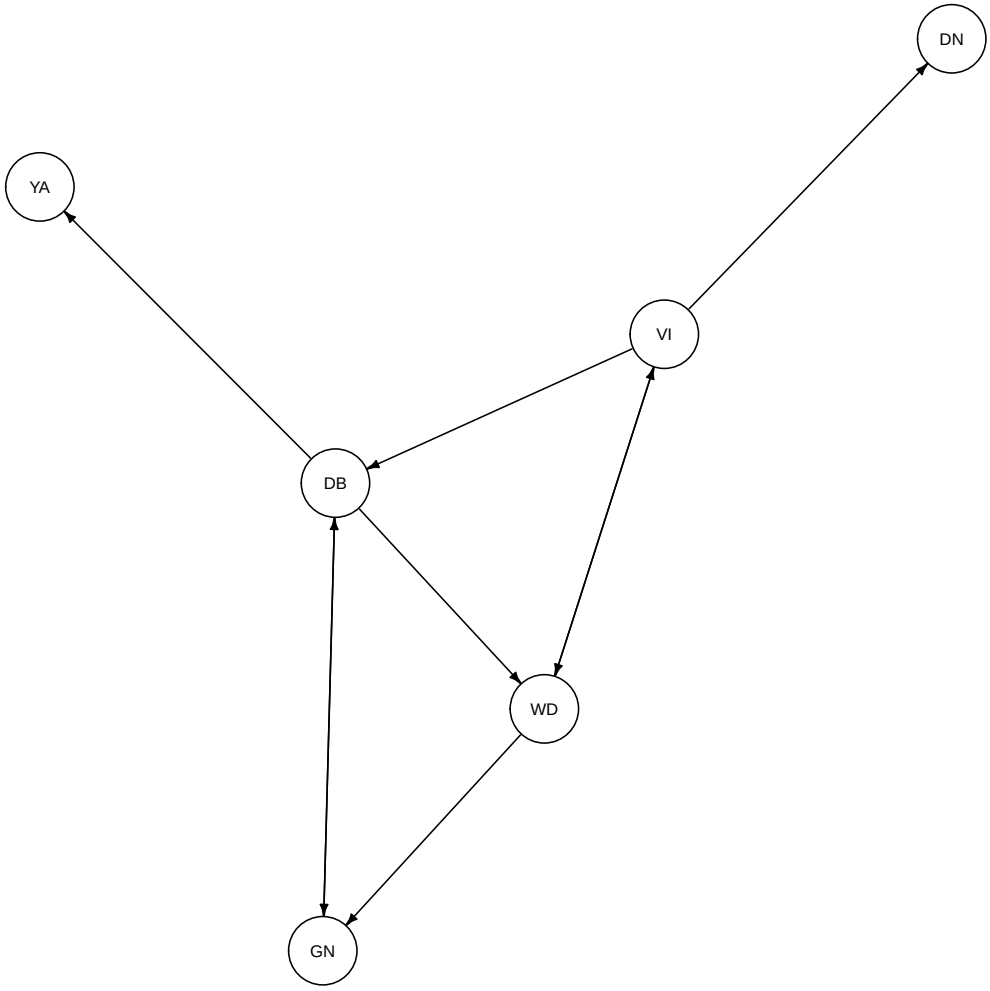
[132]První horou z hlediska pořadí váženého součtu pro grafové metriky, tedy z hlediska propojenosti zdrojů dat obsahujících reprezentaci tohoto objektu, je Svorová hora v Krkonoších na hranici s Polskem s hodnotou 0,95. Stejnou hodnotu má i Sokol v Lužických horách, jehož vrchol leží uvnitř území České republiky.

stačila na 52. místo (ze 664).

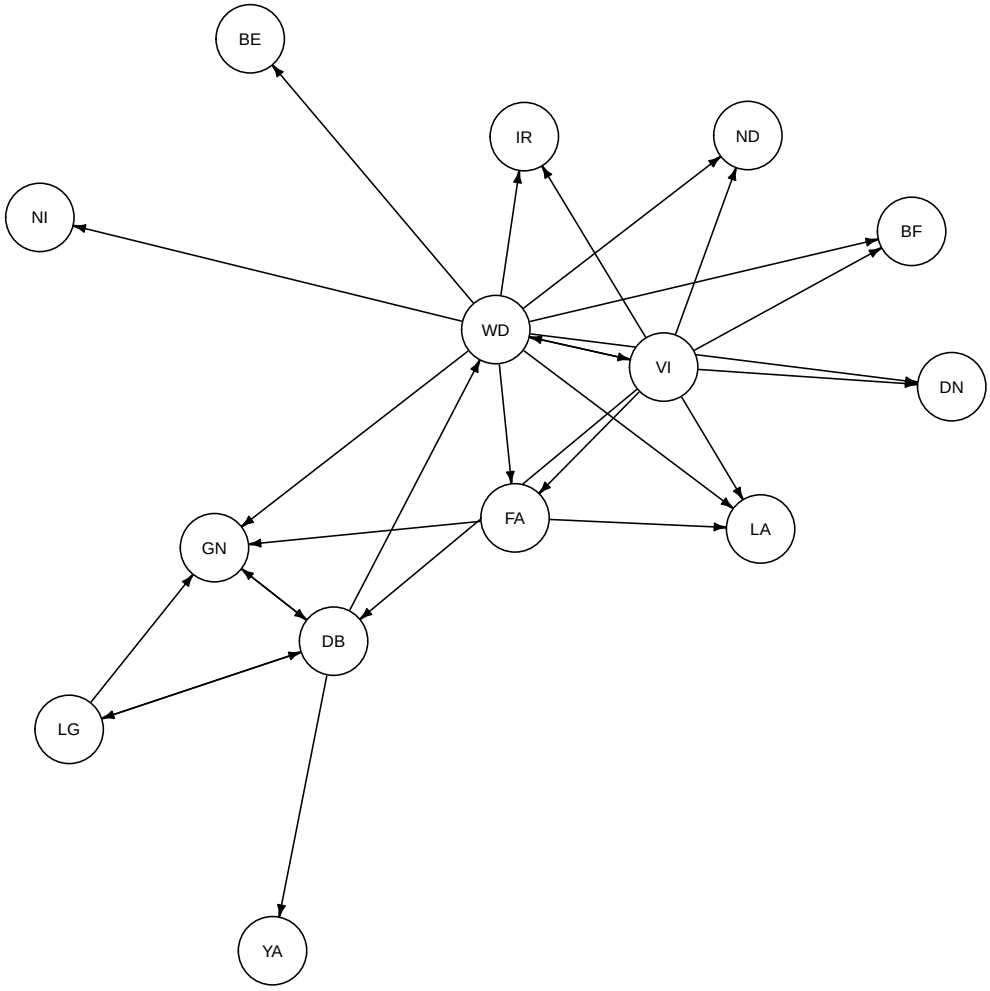
Nejvyšší zaznamenaný shlukový koeficient měl hodnotu 1. Týkal se sopky Black Peak na americké Aljašce. Jeho datová síť byla stejná jako v případě Nabire Airport na obrázku 23 vlevo. Maximální hodnoty reciprocity (jako dalšího důležitého grafového parametru) činí 0,5 (hory v ČR) a 0,67 (stratovulkány). Obě hodnoty byly zaznamenány u velmi malých grafů – hodnota 0,67 znamená, že graf (objekt `The_Black_Tusk`[133]) má tři vrcholy, z nichž dva jsou spojené mezi sebou recipročně (v případě hodnoty 0,5 jde o 4 uzly a oboustranné propojení dvou z nich).

První vstupní předpoklad této studie také potvrzují datové sítě na obrázcích 31 a 32. Příklad hor Luzný a Mount St. Helens, které v obou skupinách dat představují objekty nejlépe zakotvené v systému propojených dat, ukazuje rozdíl mezi lokálními a globálními propojenými daty.

[133]Sopka v Britské Kolumbii v Kanadě.



Obrázek 31: Datová síť objektu `Lusen__Bavaria_`.



Obrázek 32: Datová síť objektu `Mount_St._Helens`.

Srovnání dat s odlišnou geografickou lokalizací I.

V této studii jsou pomocí uzlových a grafových metrik i multikriteriální analýzy porovnávány dva vzorky dat, které se shodují svým tématem a podrobností, ale liší se geografickou lokalizací. Výchozí předpoklad se tedy týká právě odlišné geografické lokalizace a lze ho formulovat například následovně: „Ve vyhodnocení datových sad popisujících identické vazby mezi reprezentacemi v různých Linked Data zdrojích budou zřetelné rozdíly závislé na poloze jednotlivých skupin“.

Jako ilustrační data pro tuto případovou studii jsou vybraná hlavní města v Evropě a Africe. První vzorek zahrnuje 59 prvků (evropská hlavní města) a druhý 62 položek (africká hlavní města). Na rozdíl od předchozího experimentu se tedy jedná o zhruba stejně početné vzorky.

Postup je stejný jako v předchozí studii Srovnání lokálních a globálních dat – vyhodnocení vybraných uzlových metrik (absolutní hodnoty, pořadí uzlů / zdrojů) a vícekriteriální analýza uskutečněná na základě vyhodnocení uzlových i grafových metrik s využitím vah jednotlivých kritérií publikovaných ve studii Uzlová letiště.

Tabulka 19: Srovnání parametrů skupin dat hlavní města evropských a afrických států.

Kritérium	Evropa	Afrika
Počet zdrojů	27	18
Centralita mezilehlosti	0,20 (DB)	0,22 (DB)
Skóre autority	0,84 (GN)	0,85 (GN)
Skóre středu	0,84 (WD)	0,84 (WD)
Page Rank	0,18 (DB)	0,19 (DB)

Tabulka 19 potvrzuje úvodní předpoklad jen zčásti. Hlavní rozdíl je patrný v počtu datových sad propojených dat, které obsahují zkoumané objekty. V případě evropských hlavních měst jde o 27 zdrojů, které zmiňují minimálně jeden objekt této skupiny. To je v celém kontextu této práce nadprůměrné číslo. Druhý nejvyšší rozdíl je patrný ve druhém řádku, který vyjadřuje centralitu mezilehlosti. V obou případech je na prvním místě DBpedia (DB), ale v rámci

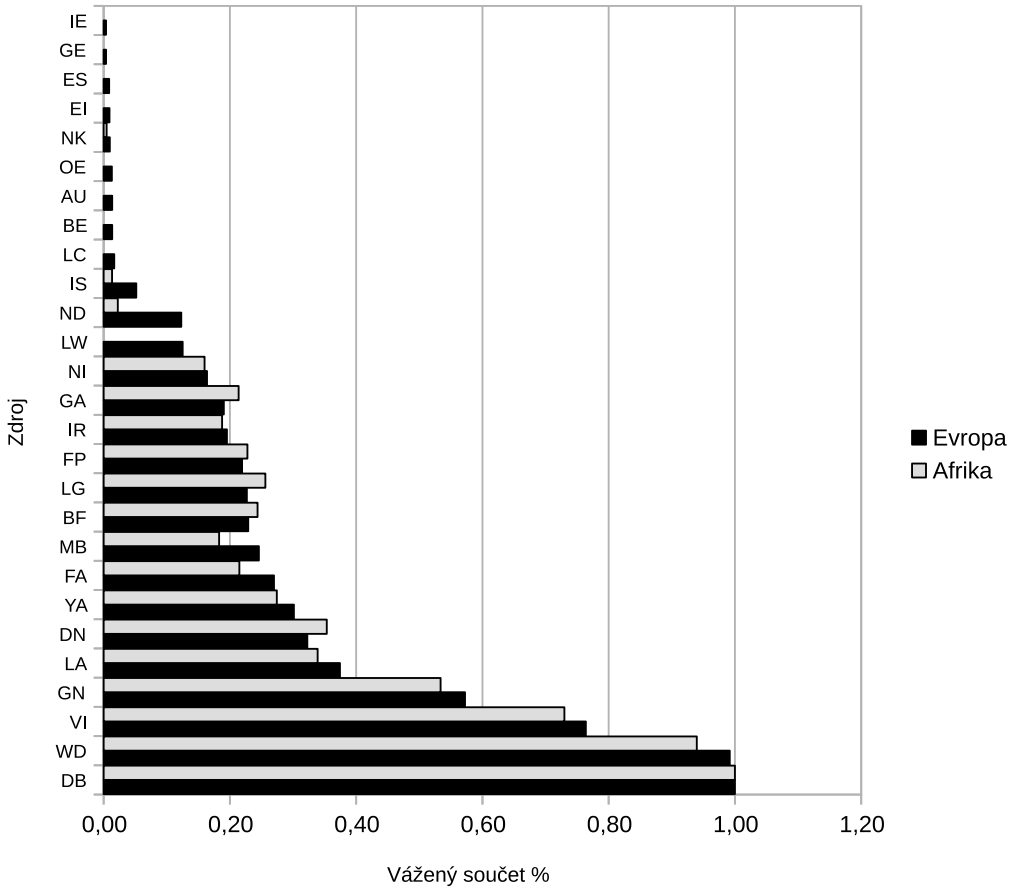
afrických metropolí je její role z hlediska integrace ostatních datových zdrojů mírně důležitější.

V případě ostatních metrik jsou jejich hodnoty i nositelé takřka totožní. Pozici autority v oblasti propojených prostorových dat opět obsadila datová sada GeoNames.org (GN), zdrojem spíše odkazujícím než odkazovaným jsou Wikidata. Tento fakt může být dán tím, že Wikidata představují relativně mladý produkt, podle článku [150] byla představena v roce 2012, a tudíž ještě nebyla do mnoha produktů integrována ve formě identických vazeb. Konečně vazbu na největší množství dobře ohodnocených zdrojů (Page Rank) má DBpedia. Příliš velké rozdíly nejsou v pořadí podle jednotlivých metrik ani dále. Mezi další významné autority pro obě skupiny dat patří datové sady Knihovny Kongresu USA (LA) a Německá národní knihovny (DN). Naopak důležitými huby jsou DBpedia a VIAF (VI). Stejná shoda panuje i na dalších místech hodnocení podle algoritmu Page Rank, kde se vyskytují uzly označené jako WD, VI a GN.

Výsledky multikriteriální analýzy s použitím uzlových metrik jako kritérií a zdrojů dat jako alternativ (Obrázek 33) je možné shrnout do následujících vět:

- Nejlepší alternativou pro oba typy dat je DBpedia.
- DBpedia dosáhla v případě evropských metropolí vyššího celkového skóre než je tomu u afrických měst (příčinou mohou být i rozdílné hodnoty centrality mezilehlosti, viz tabulka 19).
- Africká hlavní města vykazují větší rozdíl mezi hodnotami získanými pro Wikidata.
- Datová sada Linked Web APIs (LW) se objevuje ve vazbách týkajících se evropských měst, ale nikoli měst afrických. Bohužel se z důvodu nedostupnosti dat nepodařilo ověřit, zda se jedná o absenci vlastních objektů nebo pouze identických vazeb.

Z hlediska hodnocení výstupů z multikriteriální analýzy, kde jako kritéria sloužily grafové metriky a jako alternativy jednotlivé objekty, výsledky podporují úvodní předpoklad. Maximální hodnota v případě evropských měst byla 2,74 (Berlín), zatímco v Africe bylo dosaženo pouze 2,12 (Addis Abeba). Tato hodnota by v Evropě stačila na 18. místo. Také průměrná hodnota je vyšší v Evropě než v Africe (1,90 vs. 1,79).



Obrázek 33: Výsledky multikriteriální analýzy pro uzly skupin hlavní města evropských a afrických států.

Dá se říct, že úvodní předpoklad byl potvrzen. Opravdu zřetelné rozdíly byly zaznamenány tři – počet zdrojů Linked Data a oba výstupy z multikriteriální analýzy. Hlavní příčinu zřejmě nemá smysl hledat v geopolitickém rozdělení, protože téma metropolí je natolik globalizované, že vztah periférie-jádro se ho zřejmě nedotýká (to však neznamená, že u jiných typů dat se není možné s jeho vlivem setkat)[134]. Důvodem bude fakt, že mnoho databází propojených dat vzniká na území Evropy (často jako součást projektů podporovaných národními orgány nebo Evropskou Unií), a proto obsahuje výhradně nebo převážně data z území Evropy (například data z Eurostatu). Nejlépe popsaným městem podle kritérií stanovených v této práci je Berlín, což je mimo jiné (například kromě značného a stále rostoucího významu Berlína) důsledkem faktu, že na území Německa je vytvořeno a spravováno velké množství datových sad Linked Data (například DBpedia, LinkedGeoData nebo Deutsche Nationalbibliothek – Linked Data Service).

Srovnání dat s odlišnou geografickou lokalizací II.

Zatímco v předchozí studii byla odlišnost v geografické lokalizaci dána spíše ekonomickým rozvojem nebo geopolitickým postavením, v tomto experimentu se sice také pracuje s daty dvou odlišných kontinentů, ale jedná se o teritoria, která se z hlediska hospodářství nebo geopolitické role významně neliší. Rozdíly mezi Severní Amerikou (převážně reprezentovanou Spojenými státy americkými) a Evropou bývá spatřován převážně v kulturně-geografické sféře. Úvodní předpoklad testovaný v této studii může znít stejně jako v předchozím experimentu – „Ve vyhodnocení datových sad popisujících identické vazby mezi reprezentacemi v různých Linked Data zdrojích budou zřetelné rozdíly závislé na poloze jednotlivých skupin“.

Obě datové sady se týkají závodních okruhů, kde se odehrávají nebo odehrávaly velké ceny dvou dominantních sérií závodů formulových vozů. Obě série dnes již překročily hranice svých „domovských“ kontinentů, ale navzdory expanzi je

[134]Z hlediska geopolitických teorií je zajímavé, že tzv. globální města, jako je Londýn, ze zjištěných hodnot také výrazně nevybočují a umisťují se sice na špičce, ale nikoli na prvních místech.

stále možné považovat Formuli 1 (F1) za evropský fenomén, zatímco IndyCar za záležitost převážně americkou (ve smyslu USA). Rozsah informací získaných pomocí SPARQL dotazu je následující – objekty spojené s okruhy, které se týkají závodů F1, čítají 80 položek, zatímco počet prvků souvisejících s IndyCar je přesně poloviční. Vyšší počet je daný délkou historie obou závodních seriálů. IndyCar byla odstartována v roce 1996, zatímco závody Formule 1 se jezdí od roku 1950, a tudíž bylo vystřídáno mnoho závodních okruhů.

Postup srovnání je shodný s předchozími dvěma studiemi Srovnání lokálních a globálních dat a Srovnání dat s odlišnou geografickou lokalizací I. – vyhodnocení vybraných uzlových metrik (absolutní hodnoty, pořadí uzlů / zdrojů) a vícekriteriální analýza uskutečněná na základě vyhodnocení uzlových i grafových metrik s využitím vah jednotlivých kritérií publikovaných ve studii Uzlová letiště.

Tabulka 20: Srovnání parametrů skupin dat Formule 1 a IndyCar.

Kritérium	Formule 1	IndyCar
Počet zdrojů	11	7
Centralita mezilehlosti	0,27 (DB)	0,21 (DB)
Skóre autority	0,79 (YA)	0,81 (YA)
Skóre středu	0,97 (DB)	0,99 (DB)
Page Rank	0,29 (DB)	0,30 (DB)

Z výpočtů metrik a tabulky 20 vyplývají zajímavé závěry, jak ve srovnání obou typ dat v této studii, tak z hlediska obou předchozích komparativních experimentů.

- První místo uzlu DB v metrikách, jako jsou centralita mezilehlosti, skóre středu a Page Rank, odpovídá s jednou výjimkou výsledkům předchozích studií. To platí i o vypočtených hodnotách, které si jsou velice blízké, i o zdrojích, které se v jednotlivých metrikách umístily v dalším pořadí. Především se jedná o datové sady Wikidata, GeoNames.org, VIAF a Yago.
- S výjimkou prvních dvou řádků tabulky si jsou velice podobné i hodnoty metriky v obou datových sadách zkoumaných v této studii. Navíc zde

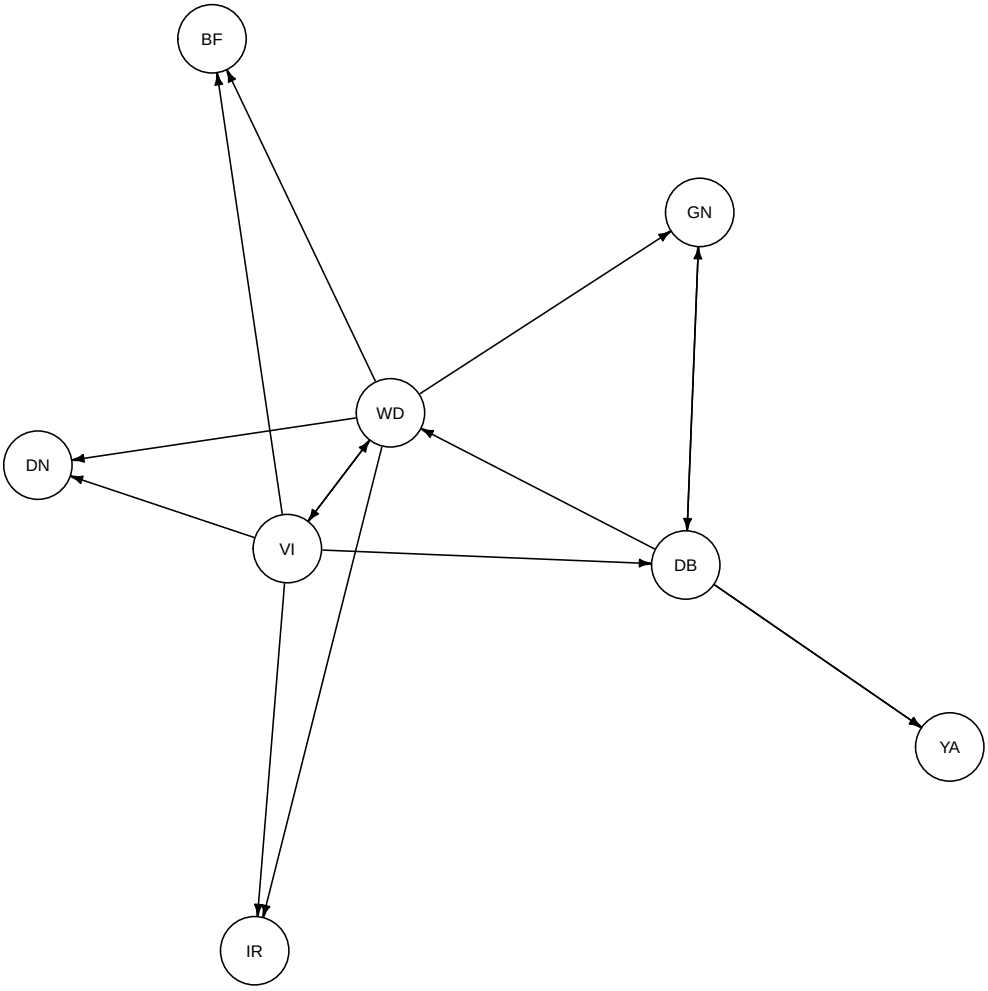
nejsou žádné odchylky z hlediska pořadí na prvních místech. V případě centrality mezilehlosti je uzel DB následovaný vrcholy GN, WD a VI (přičemž hodnota centrality je pro DB více než dvojnásobná ve srovnání s GN). V případě středu je pořadí DB, WD, VI (zde jsou rozdíly mezi prvním a druhým uzlem téměř 80%). Z hlediska skóre Page Rank, který má jistý vyrovnávací charakter, jsou rozdíly minimální, ale pořadí v obou případech stejné – DB, WD, YA.

- Výjimečná je situace v případě skóre autority. Téměř ve všech předchozích experimentech byla za nejvýznamnější autoritu zkoumaných propojených prostorových dat označena datová sada GeoNames.org (GN). V tomto případě se však se značnou ztrátou umístila až za zdroji Yago a Wikidata. Důvodem je zřejmě fakt, že se nejedná o typický příklad prostorových dat.

Ani souhrnné výsledky multikriteriální analýzy pro uzly nepotvrdily větší rozdíly mezi zkoumanými skupinami dat. Obě skupiny vykazují stejné pořadí vhodných zdrojů na prvních třech místech (DB, GN, a WD). Stejně tak se v obou skupinách procentuální rozdíly od maxima pohybují na velmi podobných hodnotách. S výjimkou počtu zdrojů a hodnot centrality mezilehlosti nejsou ve vzorcích zaznamenány rozdíly, které by ukazovaly na jiný původ dat.

To samé platí o výsledcích vícekriteriální analýzy založené na grafových metrikách. V obou skupinách dat je maximální hodnota 1,17 a minimální 0,27. Pouze ze sady okruhů F1 vyčnívá prvek `Circuit_de_Spa-Francorchamps`, který v rámci multikriteriální analýzy získal hodnotu váženého součtu 1,29. Příčinou tohoto vysokého hodnocení je velmi propojený graf, obsahující i reciproční vazby (Obrázek 34). Paradoxní je, že po odstranění zmíněné položky reprezentující slavný belgický okruh, je nejlépe popsaný prvek v obou skupinách stejný. Jedná se o geografický objekt `Indianapolis_Motor_Speedway` reprezentující slavný okruh v Indianapolis (Indiana, USA), po kterém je pojmenovaná série IndyCar, neboť se na něm odehrává slavný závod Indianapolis 500 (v češtině 500 mil v Indianapolis), ale zároveň hostil i závody Formule 1.

Jak vyplývá z předchozího textu, úvodní předpoklad zaměřený na rozdíly plynoucí z různého původu dat, nebyl, až na dílčí kritéria (počet uzlů a



Obrázek 34: Datová síť objektu `Circuit_de_Spa-Francorchamps`.

centralita mezilehlosti), potvrzený.

Kapitola 6

Výsledky

Tato kapitola se skládá ze tří částí. Nejprve jsou předloženy výsledky souhrnného zpracování všech dílčích prostorových dat, které byly součástí předchozích experimentů, ale k nim jsou zároveň připojené i nové objekty. Následně jsou výsledky veškerého zpracování dat shrnuty, komentovány, diskutovány a poté jsou získané poznatky zobecněné a publikované jako návrhy pro zlepšení situace týkající se identických vazeb propojených prostorových dat a také jako příklady dobré praxe. Poslední, třetí část se věnuje směrům budoucího výzkumu na poli propojených prostorových dat, především z hlediska identických vazeb.

Analýza kompletního vzorku prostorových dat

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 4502 objektech. Mezi jednotlivými skupinami dat se objevilo 22 duplicit (0,43%), konkrétně 13 mezi letišti a hlavními městy, 5 duplicit tvořila středoevropská hlavní města, která byla zároveň ve skupině hlavních měst Evropy, dvě duplicity se objevily v obou skupinách zaměřených na závodní okruhy a dvojice stejných objektů se

vyskytly i mezi stratovulkány a letišti a mezi stratovulkány a národními parky.

Do zpracování tedy vstoupilo 4480 objektů (99,51%). Výsledky byly získány pro 3984 prvků (88,49% ze získaných dat, 88,93% z dat vstupujících do procesu zpracování).

Do celkového souboru dat byly zahrnuty prvky ze skupin testovaných v rámci předchozích studií. Doplněny byly dvě nové skupiny – mezinárodní organizace a národní parky. Zastoupení skupin ve vzorku extrahovaných dat je prezentováno v tabulce 21.

Tabulka 21: Zastoupení skupin dat v celkové vzorku.

Skupina dat	Počet objektů	%
Uzlová letiště	1300	28,88 %
Mezinárodní organizace	1186	26,34 %
Stratovulkány	743	16,50 %
Národní parky	579	12,86 %
Evropské mezinárodní silnice	233	5,18 %
Republiky	130	2,89 %
Okruhy F1	80	1,78 %
Hraniční řeky	69	1,53 %
Hlavní města Afrika	62	1,38 %
Hlavní města Evropa	59	1,31 %
Okruhy IndyCar	40	0,89 %
Hory v ČR	16	0,36 %
Hlavní města ve střední Evropě	5	0,11 %

Zdroje

V souboru dat se celkem objevily 32 zdroje (Tabulka 22). Pouze DBpedia (DB) se z pochopitelných důvodů (viz způsob získávání informací o identických vazbách) vyskytuje v každém zkoumaném objektu. Vyskytuje se tedy ve 3984 případech. Naopak minimální výskyt byl zaznamenaný u zdrojů U.S. National

Library of Medicine (NM) a Genealogy.net (GE), které se ve zkoumaném vzorku objevily pouze jednou – NM obsahuje reprezentaci objektu Rovníková Guinea (`Equatorial_Guinea`) a GE tvoří uzel grafu objektu Berlín (`Berlin`).

Tabulka 22: Výskyt zdrojů v popisu datových objektů.

Zdroj	Výskyt	Výskyt %
DB	3984	100,00 %
WD	3982	99,95 %
YA	3918	98,34 %
GN	2057	51,63 %
VI	990	24,85 %
LG	919	23,07 %
DN	879	22,06 %
LA	828	20,78 %
BF	500	12,55 %
IR	486	12,20 %
FA	391	9,81 %
IS	362	9,09 %
ND	316	7,93 %
MB	312	7,83 %
GA	286	7,18 %
NI	200	5,02 %
FP	143	3,59 %
OE	134	3,36 %
ES	124	3,11 %
TI	116	2,91 %
WB	114	2,86 %
AU	82	2,06 %
NK	82	2,06 %
LW	78	1,96 %
BE	70	1,76 %
AA	31	0,78 %
EI	19	0,48 %
LC	13	0,33 %

Zdroj	Výskyt	Výskyt %
BC	4	0,10 %
IE	2	0,05 %
GE	1	0,03 %
NM	1	0,03 %

Podobně jako v předchozích experimentech lze zdroje hodnotit podle šesti vybraných metrik určených pro hodnocení uzlů grafu. Těmito metrikami jsou centralita stupně (Tabulka 23, Obrázek 35), centralita blízkosti (Tabulka 24, Obrázek 36), centralita mezilehlosti (Tabulka 25, Obrázek 37), autorita (Tabulka 26, Obrázek 38), střed neboli hub (Tabulka 27, Obrázek 39) a Page Rank (Tabulka 28, Obrázek 40)[135].

Grafy na obrázcích 35, 36, 37, 38, 39, 40, 41 a 44 jsou konstruovány jako zjednodušené krabicové grafy [151]. Jejich účelem je ukázat rozptyl hodnot dané veličiny a hodnoty ležící mezi prvním a třetím kvartilem. Z těchto grafů je dobře patrná stabilita hodnot jednotlivých metrik i výsledků multikriteriální analýzy. Čím větší je část grafu vyjádřená obdélníkem oproti koncovým a počátečním úsečkám, tím více se zkoumaná data blíží normálnímu rozdělení. Stabilita veličin je dále vyjádřena koeficientem variace v tabulkách, který představuje relativní hodnotu, která je srovnatelná napříč soubory dat.

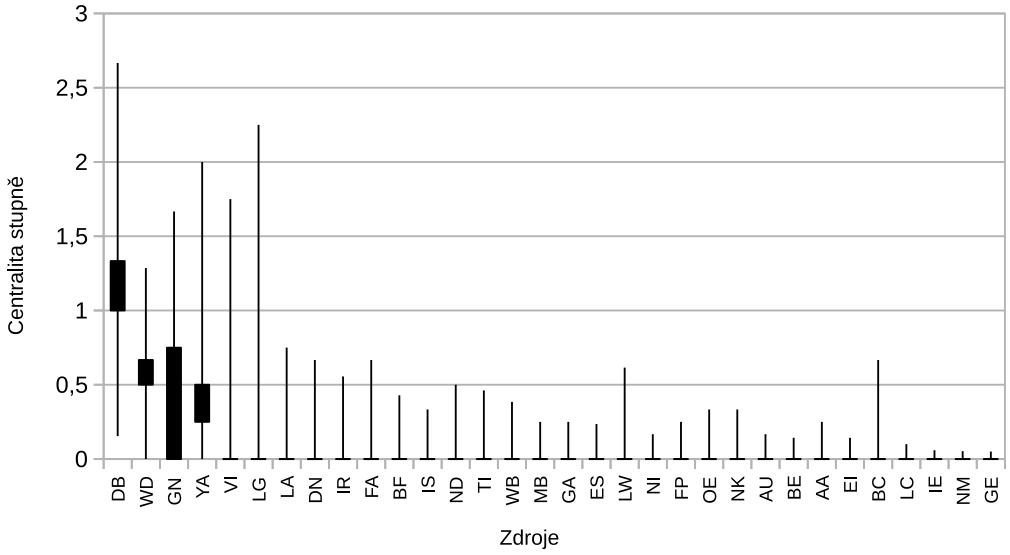
V této části jsou uvedené pouze výsledky analýz. Jejich diskuze a interpretace následují v podkapitole Shrnutí a interpretace výsledků experimentů.

Tabulka 23: Statistické hodnoty popisující centralitu stupně pro zdroje dat.

Zdroj	Maximum	Minimum	Průměr	Medián	Koeficient variace
DB	2,67	0,15	1,05	1,00	0,33
WD	1,29	0,00	0,60	0,50	0,23
GN	1,67	0,00	0,37	0,27	1,09
YA	2,00	0,00	0,34	0,33	0,53

[135]V tabulkách i grafech jsou publikované normalizované hodnoty, aby bylo možné provádět srovnání napříč datovými sítěmi.

Zdroj	Maximum	Minimum	Průměr	Medián	Koeficient variace
VI	1,75	0,00	0,19	0,00	1,84
LG	2,25	0,00	0,09	0,00	2,11
LA	0,75	0,00	0,05	0,00	2,23
DN	0,67	0,00	0,04	0,00	2,17
IR	0,56	0,00	0,03	0,00	3,23
FA	0,67	0,00	0,03	0,00	3,32
BF	0,43	0,00	0,02	0,00	2,99
IS	0,33	0,00	0,01	0,00	3,48
ND	0,50	0,00	0,01	0,00	3,98
TI	0,46	0,00	0,01	0,00	5,82
WB	0,38	0,00	0,01	0,00	5,87
MB	0,25	0,00	0,01	0,00	3,84
GA	0,25	0,00	0,01	0,00	4,02
ES	0,24	0,00	0,01	0,00	5,70
LW	0,62	0,00	0,01	0,00	7,56
NI	0,17	0,00	0,00	0,00	4,52
FP	0,25	0,00	0,00	0,00	5,38
OE	0,33	0,00	0,00	0,00	6,48
NK	0,33	0,00	0,00	0,00	7,54
AU	0,17	0,00	0,00	0,00	7,12
BE	0,14	0,00	0,00	0,00	7,67
AA	0,25	0,00	0,00	0,00	12,37
EI	0,14	0,00	0,00	0,00	14,54
BC	0,67	0,00	0,00	0,00	40,43
LC	0,10	0,00	0,00	0,00	18,11
IE	0,06	0,00	0,00	0,00	44,87
NM	0,05	0,00	0,00	0,00	63,11
GE	0,05	0,00	0,00	0,00	63,11



Obrázek 35: Centralita stupně zdrojů – maximum, minimum, horní a dolní kvartil.

Tabulka 24: Statistické hodnoty popisující centralitu blízkosti pro zdroje dat.

Zdroj	Maximum	Minimum	Průměr	Medián	Koeficient variace
DB	1,00	0,42	0,91	1,00	0,16
WD	1,00	0,00	0,70	0,67	0,10
YA	1,00	0,00	0,56	0,60	0,22
GN	1,00	0,00	0,34	0,50	0,99
VI	0,91	0,00	0,18	0,00	1,77
LG	0,67	0,00	0,12	0,00	1,85
DN	0,57	0,00	0,11	0,00	1,89
LA	0,60	0,00	0,10	0,00	1,97
BF	0,58	0,00	0,06	0,00	2,66
IR	0,62	0,00	0,06	0,00	2,72
FA	0,60	0,00	0,05	0,00	3,06
IS	0,53	0,00	0,04	0,00	3,19
ND	0,57	0,00	0,04	0,00	3,42
MB	0,50	0,00	0,03	0,00	3,44
GA	0,57	0,00	0,03	0,00	3,61

Zdroj	Maximum	Minimum	Průměr	Medián	Koeficient variace
NI	0,50	0,00	0,02	0,00	4,36
FP	0,50	0,00	0,02	0,00	5,19
OE	0,60	0,00	0,01	0,00	5,38
ES	0,53	0,00	0,01	0,00	5,59
TI	0,53	0,00	0,01	0,00	5,78
WB	0,52	0,00	0,01	0,00	5,83
AU	0,50	0,00	0,01	0,00	6,90
NK	0,50	0,00	0,01	0,00	6,91
LW	0,47	0,00	0,01	0,00	7,08
BE	0,50	0,00	0,01	0,00	7,49
AA	0,50	0,00	0,00	0,00	11,33
EI	0,45	0,00	0,00	0,00	14,45
LC	0,48	0,00	0,00	0,00	17,50
BC	0,75	0,00	0,00	0,00	32,20
IE	0,38	0,00	0,00	0,00	44,63
GE	0,44	0,00	0,00	0,00	63,11
NM	0,32	0,00	0,00	0,00	63,11

Tabulka 25: Statistické hodnoty popisující centralitu mezilehlosti pro zdroje dat.

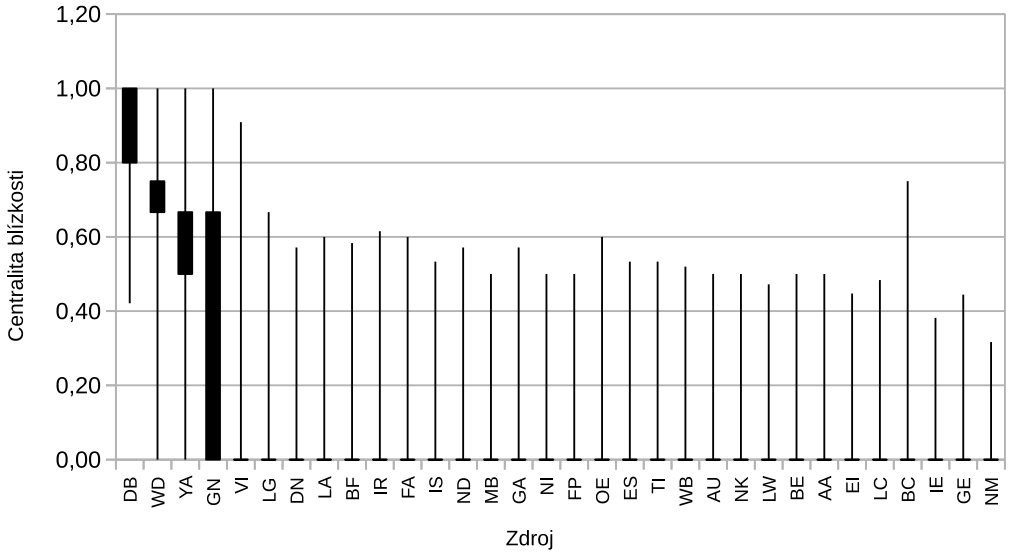
Zdroj	Maximum	Minimum	Průměr	Medián	Koeficient variace
DB	0,67	0,00	0,22	0,05	1,15
GN	0,50	0,00	0,09	0,00	1,43
WD	0,50	0,00	0,05	0,00	1,64
VI	0,45	0,00	0,04	0,00	2,04
FA	0,20	0,00	0,00	0,00	5,94
IR	0,08	0,00	0,00	0,00	5,39
TI	0,06	0,00	0,00	0,00	6,03
BC	0,33	0,00	0,00	0,00	47,04
LW	0,05	0,00	0,00	0,00	22,31
LG	0,08	0,00	0,00	0,00	44,02

Zdroj	Maximum	Minimum	Průměr	Medián	Koeficient variace
LA	0,07	0,00	0,00	0,00	47,66
AA	0,00	0,00	0,00	0,00	0,00
AU	0,00	0,00	0,00	0,00	0,00
BE	0,00	0,00	0,00	0,00	0,00
BF	0,00	0,00	0,00	0,00	0,00
DN	0,00	0,00	0,00	0,00	0,00
EI	0,00	0,00	0,00	0,00	0,00
ES	0,00	0,00	0,00	0,00	0,00
FP	0,00	0,00	0,00	0,00	0,00
GA	0,00	0,00	0,00	0,00	0,00
GE	0,00	0,00	0,00	0,00	0,00
IE	0,00	0,00	0,00	0,00	0,00
IS	0,00	0,00	0,00	0,00	0,00
LC	0,00	0,00	0,00	0,00	0,00
MB	0,00	0,00	0,00	0,00	0,00
ND	0,00	0,00	0,00	0,00	0,00
NI	0,00	0,00	0,00	0,00	0,00
NK	0,00	0,00	0,00	0,00	0,00
NM	0,00	0,00	0,00	0,00	0,00
OE	0,00	0,00	0,00	0,00	0,00
WB	0,00	0,00	0,00	0,00	0,00
YA	0,00	0,00	0,00	0,00	0,00

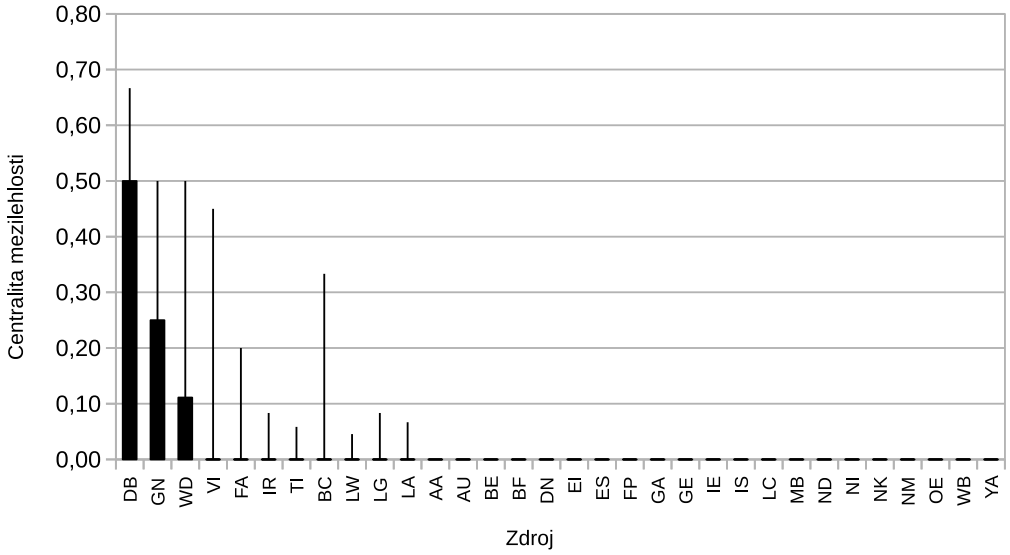
Tabulka 26: Statistické hodnoty popisující skóre autority pro zdroje dat.

Zdroj	Maximum	Minimum	Průměr	Medián	Koeficient variace
WD	1,00	0,00	0,79	0,77	0,31
YA	1,00	0,00	0,72	0,77	0,44
GN	1,00	0,00	0,46	0,00	1,04
DB	1,00	0,00	0,14	0,00	1,91
DN	1,00	0,00	0,13	0,00	2,19

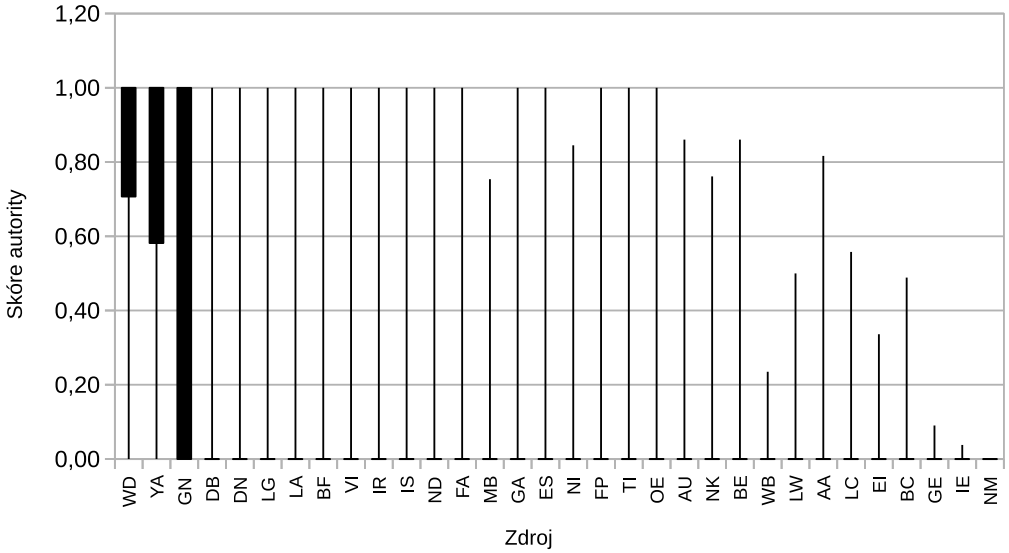
Zdroj	Maximum	Minimum	Průměr	Medián	Koeficient variace
LG	1,00	0,00	0,13	0,00	2,04
LA	1,00	0,00	0,12	0,00	2,29
BF	1,00	0,00	0,08	0,00	2,94
VI	1,00	0,00	0,08	0,00	2,17
IR	1,00	0,00	0,07	0,00	3,01
IS	1,00	0,00	0,04	0,00	3,52
ND	1,00	0,00	0,04	0,00	3,86
FA	1,00	0,00	0,03	0,00	3,56
MB	0,75	0,00	0,02	0,00	3,88
GA	1,00	0,00	0,02	0,00	4,13
ES	1,00	0,00	0,02	0,00	5,90
NI	0,85	0,00	0,02	0,00	4,82
FP	1,00	0,00	0,01	0,00	5,74
TI	1,00	0,00	0,01	0,00	5,99
OE	1,00	0,00	0,01	0,00	6,48
AU	0,86	0,00	0,01	0,00	7,32
NK	0,76	0,00	0,01	0,00	7,35
BE	0,86	0,00	0,01	0,00	7,93
WB	0,23	0,00	0,00	0,00	5,96
LW	0,50	0,00	0,00	0,00	7,55
AA	0,82	0,00	0,00	0,00	12,04
LC	0,56	0,00	0,00	0,00	18,66
EI	0,34	0,00	0,00	0,00	15,39
BC	0,49	0,00	0,00	0,00	45,25
GE	0,09	0,00	0,00	0,00	63,11
IE	0,04	0,00	0,00	0,00	45,57
NM	0,00	0,00	0,00	0,00	0,00



Obrázek 36: Centralita blízkosti zdrojů – maximum, minimum, horní a dolní kvartil.



Obrázek 37: Centralita mezilehlosti zdrojů – maximum, minimum, horní a dolní kvartil.



Obrázek 38: Skóre autority zdrojů – maximum, minimum, horní a dolní kvartil.

Tabulka 27: Statistické hodnoty popisující skóre středu pro zdroje dat.

Zdroj	Maximum	Minimum	Průměr	Medián	Koeficient variace
DB	1,00	0,00	0,89	1,00	0,30
WD	1,00	0,00	0,27	0,30	1,11
VI	1,00	0,00	0,19	0,00	1,90
FA	1,00	0,00	0,03	0,00	3,95
GN	1,00	0,00	0,03	0,00	3,23
LG	1,00	0,00	0,03	0,00	4,12
WB	1,00	0,00	0,02	0,00	6,05
TI	0,93	0,00	0,02	0,00	6,03
IR	0,73	0,00	0,01	0,00	4,30
LW	1,00	0,00	0,01	0,00	8,01
LA	0,50	0,00	0,01	0,00	3,84
EI	0,18	0,00	0,00	0,00	15,59
YA	0,00	0,00	0,00	0,00	1,79
DN	0,00	0,00	0,00	0,00	4,32
BF	0,00	0,00	0,00	0,00	5,21

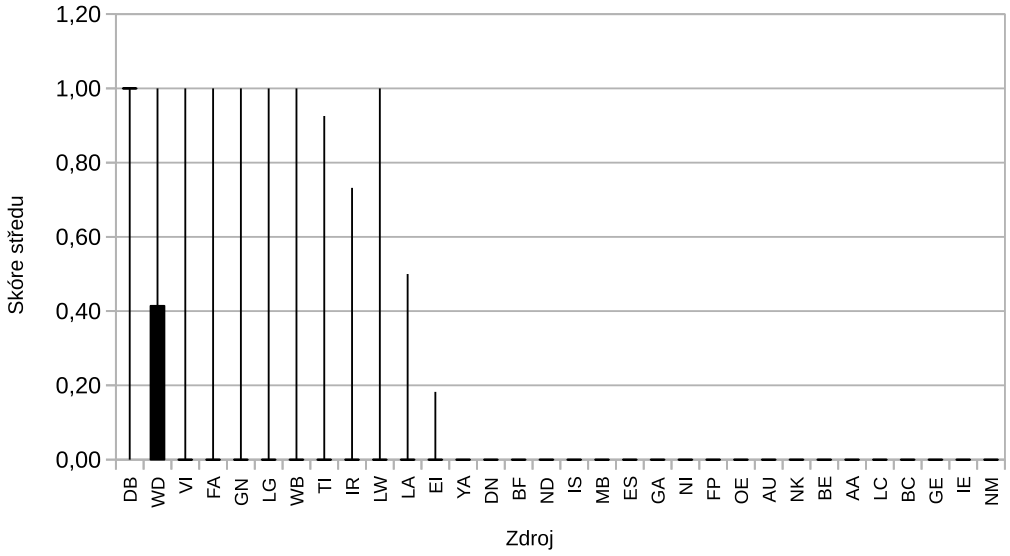
Zdroj	Maximum	Minimum	Průměr	Medián	Koeficient variace
ND	0,00	0,00	0,00	0,00	6,98
IS	0,00	0,00	0,00	0,00	6,39
MB	0,00	0,00	0,00	0,00	6,81
ES	0,00	0,00	0,00	0,00	10,68
GA	0,00	0,00	0,00	0,00	6,62
NI	0,00	0,00	0,00	0,00	8,22
FP	0,00	0,00	0,00	0,00	9,43
OE	0,00	0,00	0,00	0,00	10,94
AU	0,00	0,00	0,00	0,00	11,48
NK	0,00	0,00	0,00	0,00	11,53
BE	0,00	0,00	0,00	0,00	12,59
AA	0,00	0,00	0,00	0,00	18,96
LC	0,00	0,00	0,00	0,00	24,24
BC	0,00	0,00	0,00	0,00	39,17
GE	0,00	0,00	0,00	0,00	0,00
IE	0,00	0,00	0,00	0,00	0,00
NM	0,00	0,00	0,00	0,00	0,00

Tabulka 28: Statistické hodnoty popisující Page Rank pro zdroje dat.

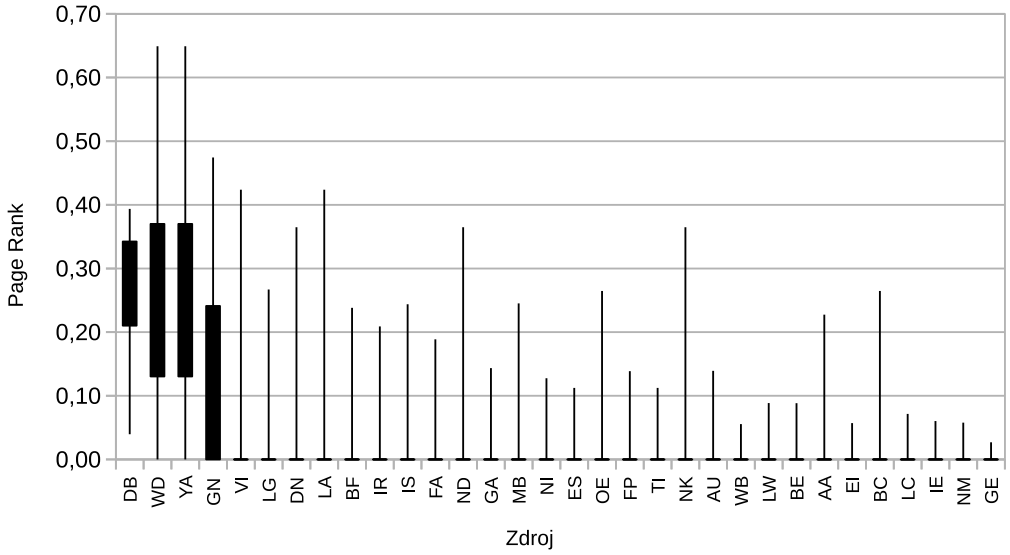
Zdroj	Maximum	Minimum	Průměr	Medián	Koeficient variace
DB	0,39	0,04	0,26	0,26	0,32
WD	0,65	0,00	0,23	0,17	0,53
YA	0,65	0,00	0,22	0,17	0,55
GN	0,47	0,00	0,13	0,08	1,09
VI	0,42	0,00	0,03	0,00	1,98
LG	0,27	0,00	0,02	0,00	1,96
DN	0,36	0,00	0,02	0,00	2,43
LA	0,42	0,00	0,02	0,00	2,21
BF	0,24	0,00	0,01	0,00	2,99
IR	0,21	0,00	0,01	0,00	3,01

Zdroj	Maximum	Minimum	Průměr	Medián	Koeficient variace
IS	0,24	0,00	0,01	0,00	3,43
FA	0,19	0,00	0,01	0,00	3,41
ND	0,36	0,00	0,01	0,00	4,44
GA	0,14	0,00	0,00	0,00	3,86
MB	0,25	0,00	0,00	0,00	3,98
NI	0,13	0,00	0,00	0,00	4,66
ES	0,11	0,00	0,00	0,00	5,64
OE	0,26	0,00	0,00	0,00	6,52
FP	0,14	0,00	0,00	0,00	5,41
TI	0,11	0,00	0,00	0,00	5,85
NK	0,36	0,00	0,00	0,00	8,27
AU	0,14	0,00	0,00	0,00	7,18
WB	0,06	0,00	0,00	0,00	5,87
LW	0,09	0,00	0,00	0,00	7,32
BE	0,09	0,00	0,00	0,00	7,79
AA	0,23	0,00	0,00	0,00	12,82
EI	0,06	0,00	0,00	0,00	14,57
BC	0,26	0,00	0,00	0,00	37,80
LC	0,07	0,00	0,00	0,00	18,66
IE	0,06	0,00	0,00	0,00	44,62
NM	0,06	0,00	0,00	0,00	63,11
GE	0,03	0,00	0,00	0,00	63,11

Stejné tabulky a grafy jako pro jednotlivé metriky jsou zpracovány i pro výsledky multikriteriální analýzy (Tabulka 29, Obrázek 41), která byla realizována metodou váženého součtu. Do procesu hodnocení alternativ (zdrojů) vícekriteriální analýzy vstupovaly váhy určené v případové studii Uzlová letiště a hodnoty vybraných šesti uzlových metrik jako kritéria analýzy.



Obrázek 39: Skóre středu zdrojů – maximum, minimum, horní a dolní kvartil.

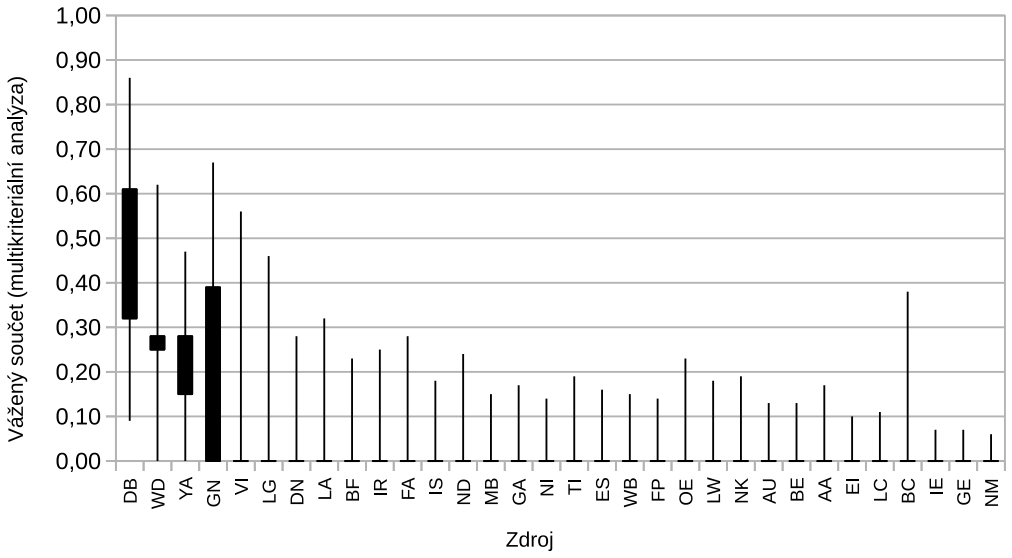


Obrázek 40: Page Rank zdrojů – maximum, minimum, horní a dolní kvartil.

Tabulka 29: Statistické hodnoty popisující výsledky multikriteriální analýzy pro zdroje dat.

Zdroj	Maximum	Minimum	Průměr	Medián	Koeficient variace
DB	0,86	0,09	0,42	0,32	0,40
WD	0,62	0,00	0,28	0,28	0,17
YA	0,47	0,00	0,20	0,19	0,38
GN	0,67	0,00	0,18	0,16	1,08
VI	0,56	0,00	0,08	0,00	1,81
LG	0,46	0,00	0,04	0,00	1,93
DN	0,28	0,00	0,03	0,00	1,97
LA	0,32	0,00	0,03	0,00	2,04
BF	0,23	0,00	0,02	0,00	2,76
IR	0,25	0,00	0,02	0,00	2,86
FA	0,28	0,00	0,01	0,00	3,21
IS	0,18	0,00	0,01	0,00	3,27
ND	0,24	0,00	0,01	0,00	3,55
MB	0,15	0,00	0,01	0,00	3,50
GA	0,17	0,00	0,01	0,00	3,69
NI	0,14	0,00	0,00	0,00	4,41
TI	0,19	0,00	0,00	0,00	5,80
ES	0,16	0,00	0,00	0,00	5,64
WB	0,15	0,00	0,00	0,00	5,85
FP	0,14	0,00	0,00	0,00	5,23
OE	0,23	0,00	0,00	0,00	5,60
LW	0,18	0,00	0,00	0,00	7,18
NK	0,19	0,00	0,00	0,00	6,99
AU	0,13	0,00	0,00	0,00	6,96
BE	0,13	0,00	0,00	0,00	7,54
AA	0,17	0,00	0,00	0,00	11,56
EI	0,10	0,00	0,00	0,00	14,48
LC	0,11	0,00	0,00	0,00	17,70
BC	0,38	0,00	0,00	0,00	36,18
IE	0,07	0,00	0,00	0,00	44,62

Zdroj	Maximum	Minimum	Průměr	Medián	Koeficient variace
GE	0,07	0,00	0,00	0,00	63,11
NM	0,06	0,00	0,00	0,00	63,11



Obrázek 41: Vážený součet (multikriteriální analýza zdrojů) – maximum, minimum, horní a dolní kvartil.

Tabulka 30 ilustruje vztahy mezi uzlovými metrikami i výsledkem vícekriteriální analýzy[136]. Tento vztah je kvantifikován pomocí Pearsonova korelačního koeficientu, u nějž hodnoty blížící se k maximu (1) signalizují přímou závislost obou jevů, zatímco hodnoty okolo 0 představují nezávislost. Záporná čísla se v tabulce 30 nevyskytují, protože všechny metriky, a tudíž i výsledek multikriteriální analýzy, představují maximalizační kritéria.

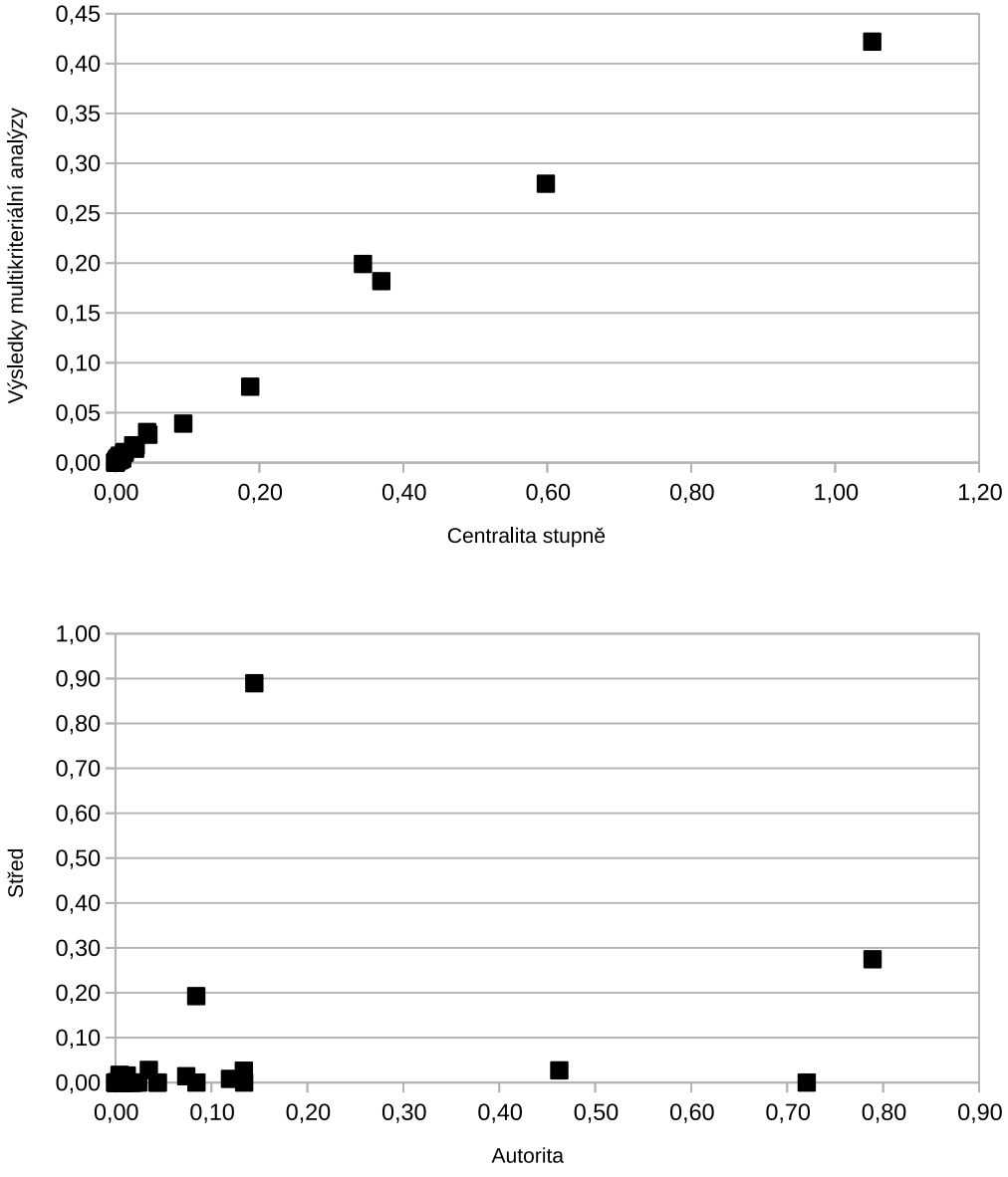
[136]Zkratky pro označení metrik v tabulce 30 jsou použité v souladu se jmennou konvencí užívanou v celém textu. Označení MCDA vychází z anglického názvu multikriteriální analýzy.

Tabulka 30: Korelace (Pearsonův korelační koeficient) mezi uzlovými metrikami a výsledkem multikriteriální analýzy.

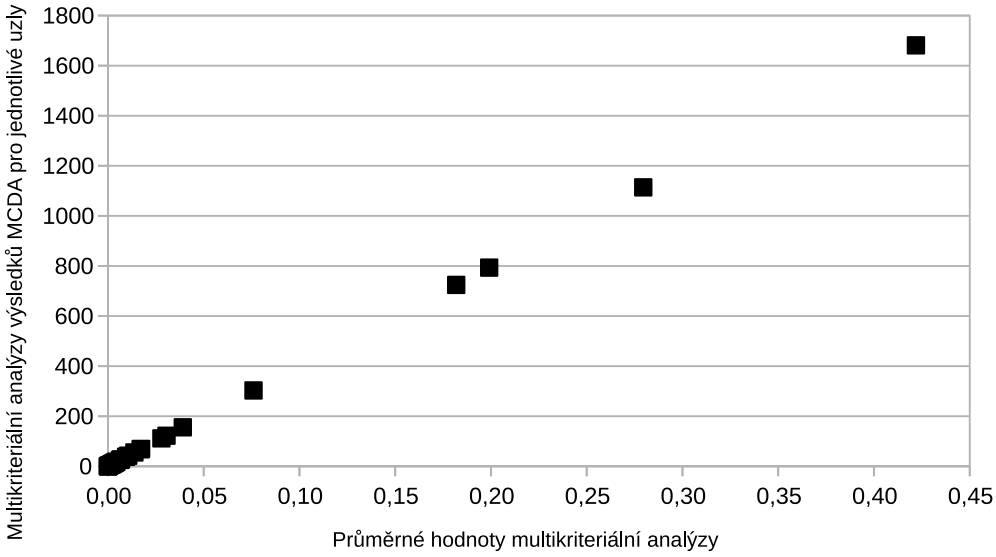
Metrika	C_d	C_c	C_b	A	H	PR	$MCDA$
C_d	-	0,97	0,92	0,62	0,90	0,93	0,99
C_c	0,97	-	0,80	0,76	0,80	0,99	0,99
C_b	0,92	0,80	-	0,31	0,93	0,73	0,87
A	0,62	0,76	0,31	-	0,24	0,84	0,71
H	0,90	0,80	0,93	0,24	-	0,71	0,85
PR	0,93	0,99	0,73	0,84	0,71	-	0,97
$MCDA$	0,99	0,99	0,87	0,71	0,85	0,97	-

Grafy na obrázku 42 ukazují příklady dvou dvojic hodnot s maximální (nahoře) a minimální (dole) korelací. Již výše bylo uvedeno, že autorita a střed (hub) představují do jisté míry dva protikladné jevy. To vyplývá i z tabulky 30 a grafu na obrázku 42. Na druhou stranu výsledek vícekriteriální analýzy je velmi podobný datům popisujícím centralitu stupně. A to přes to, že centralita stupně do výpočtu váženého součtu vstupuje s poměrně nízkou vahou (0,12). Příčina je zřejmě v tom, že centralita stupně podobně silně koreluje i s dalšími metrikami, které se na výpočtu výsledků multikriteriální analýzy podílí větší měrou.

Poslední test určený pro zdroje (uzly grafu) je ověření, zda zvolená „dvojitá“ vícekriteriální analýza (vícekriteriální analýzy znovu aplikovány na výsledky multikriteriální analýzy pro jednotlivé grafy), která byla aplikovaná v předchozích případových studiích, koreluje s průměrnou hodnotou váženého součtu za jednotlivé uzly. Z grafu 44 je zcela jasně patrné, že oba soubory hodnot korelují velmi silně (hodnota Pearsonova korelačního koeficientu je 0,999999995044041), a tudíž lze souhrnné výsledky multikriteriální analýzy získávat oběma způsoby.



Obrázek 42: Příklady korelace: centralita stupně a vážený součet multikriteriální analýzy (maximální korelace ve vzorku); autorita a střed (minimální korelace).



Obrázek 43: Korelace dvou souhrnných vyhodnocení multikriteriální analýzy.

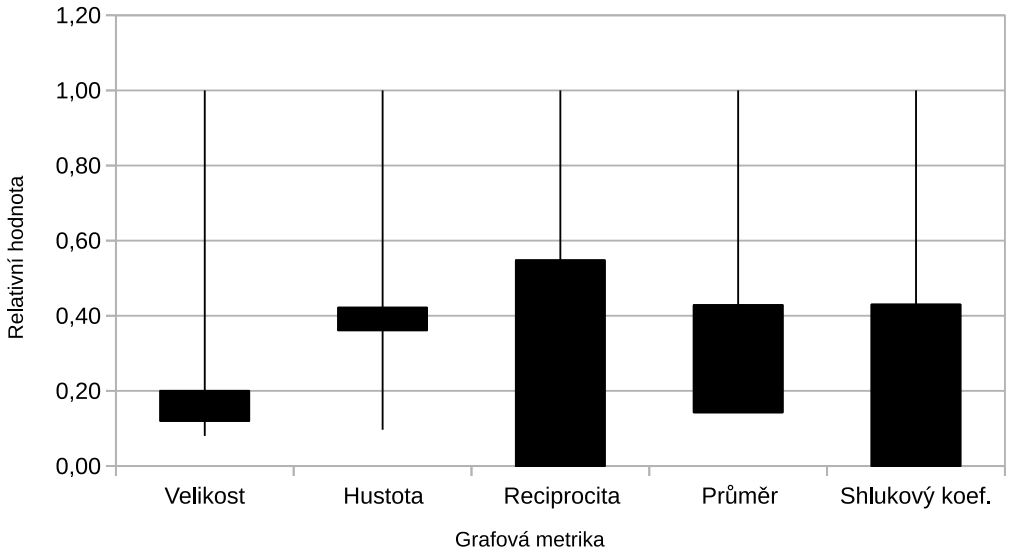
Objekty

Tabulka 31 a graf 44 ukazují statistické parametry a krabicový graf pro všechny grafové metriky. U všech metrik je zřetelný velký rozptyl a šikmost (tedy asymetrické rozložení) hodnot. Přičemž se jedná o kladnou šikmost, což znamená, že většina hodnot se nachází pod průměrem, zatímco extrémy se týkají především vysokých (ovšem v případě průměru a shlukového koeficientu negativně chápaných hodnot). Variační koeficient ukazující míru variability (tedy jak výrazně se jednotlivé hodnoty odlišují) ve vzorku je minimální v případě hustoty (z grafu 44 je zřetelné, že 50% všech hodnot se pohybuje v poměrně úzkém intervalu okolo 0,40), zatímco hodnoty reciprocity a shlukového koeficientu kolísají mnohem výrazněji (směrodatná odchylka tvoří téměř 100% průměru, viz rozsáhlé intervaly mezi prvním a třetím kvartilem v grafu 44).

Tabulka 31: Statistické parametry absolutních hodnot grafových kritérií.

	Velikost	Hustota	Reciprocita	Shlukový koef.
Maximum	25	0,83	0,73	1
Minimum	2	0,08	0	0

	Velikost	Hustota	Reciprocita	Shlukový koef.
Průměr	5,38	0,32	0,22	0,26
Směrodatná odchylka	3,76	0,09	0,22	0,24
Variační koeficient	0,70	0,30	1,00	0,94
Dolní kvartil	3	0,30	0	0
Medián	4	0,33	0,2	0,3
Horní kvartil	5	0,35	0,4	0,43



Obrázek 44: Statistické parametry (maximum, minimum, horní a dolní kvartil) relativních hodnot grafových kritérií.

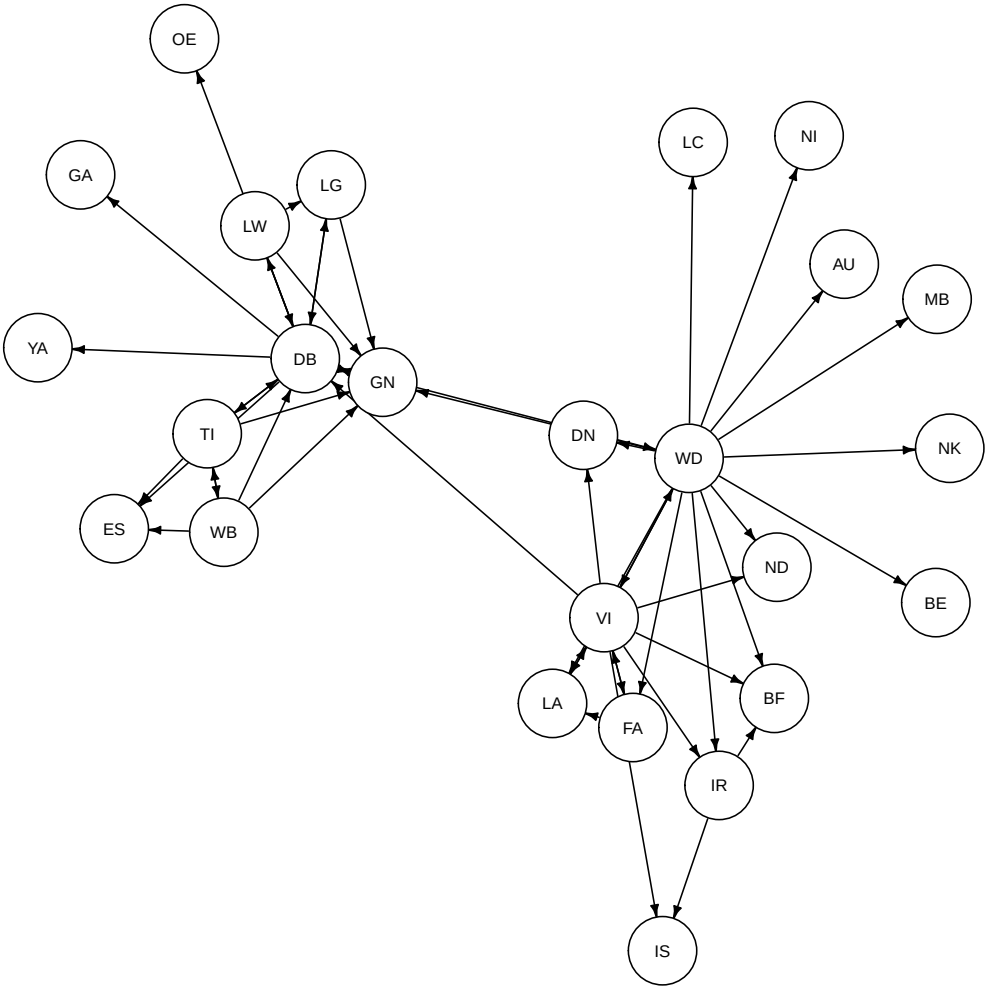
V úvodu této práce bylo zmíněno, že testování identických vazeb propojených prostorových dat by mělo vést k nalezení vhodných příkladů dobré praxe, které by mohly inspirovat tvůrce a správce datových sad k lepšímu využívání identických a podobnostních vazeb mezi reprezentacemi objektů v různých databázích publikovaných podle principů Linked Data. Následující seznam a k němu přiřazené datové sítě ukazuje příklady objektů, které se jeví jako

optimální (nejlepší) z hlediska grafových metrik a také podle celkového zhodnocení pomocí multikriteriální analýzy. Ukazuje se, že ne všechny publikované objekty jsou skutečně příklady dobré praxe, protože některých příznivých hodnot metrik bylo dosaženo především minimalizací počtu uzlů.

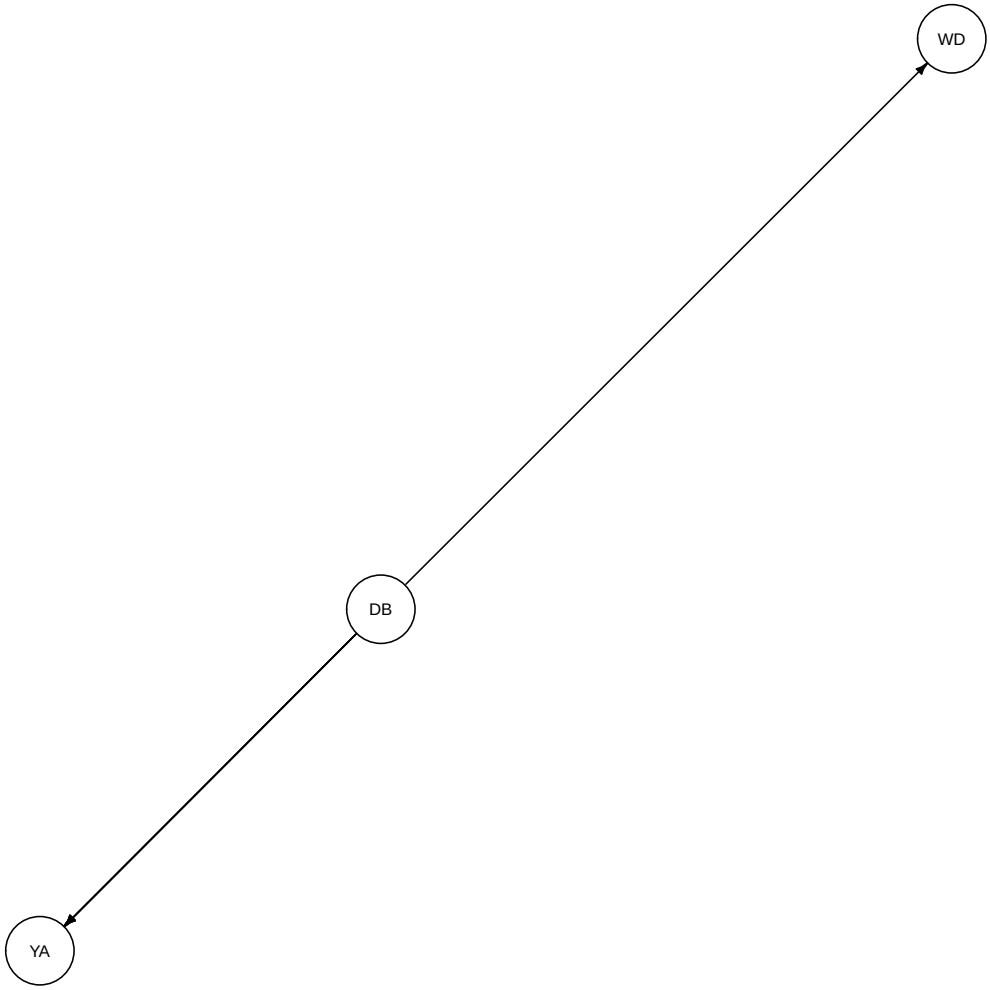
- Maximální velikost grafu (počet uzlů) a také nejlepší hodnocení podle multikriteriální analýzy dosáhl prvek `Israel`[137] (Obrázek 45).
- Objekt `South_Asian_Association_for_Regional_Cooperation` (Obrázek 46) vyniká hned z hlediska maximální hustoty. Jde však o graf s pouhými třemi vrcholy a dvěma hranami.
- Minimální shlukový koeficient (obojí při nadprůměrném počtu uzlů) má díky atypickému hvězdicovitému tvaru své sítě geografický objekt `Georgia_country_`[138] (Obrázek 47).
- Nejlepší postavení z hlediska shlukového koeficientu z hlediska vazeb má pozoruhodná a komplikovaná datová síť objektu `Niger` (Obrázek 48, která obsahuje několik uzavřených podgrafů.
- Poslední metrikou, jejíž extrémní případ ještě nebyl zmíněný, je reciprocita. V tomto případě se nejlepší (tedy nejvyšší) hodnota objevuje u prvku `John_Wyane_Airport` (Obrázek 49), který obsahuje pět uzlů a sedm hran, přičemž dvě dvojice jsou reciproční.

[137] Z hlediska velikosti grafu se na stejném místě umístil objekt `Bulgaria`, který však v multikriteriální analýze dosáhl o jednu setinu horšího počtu bodů.

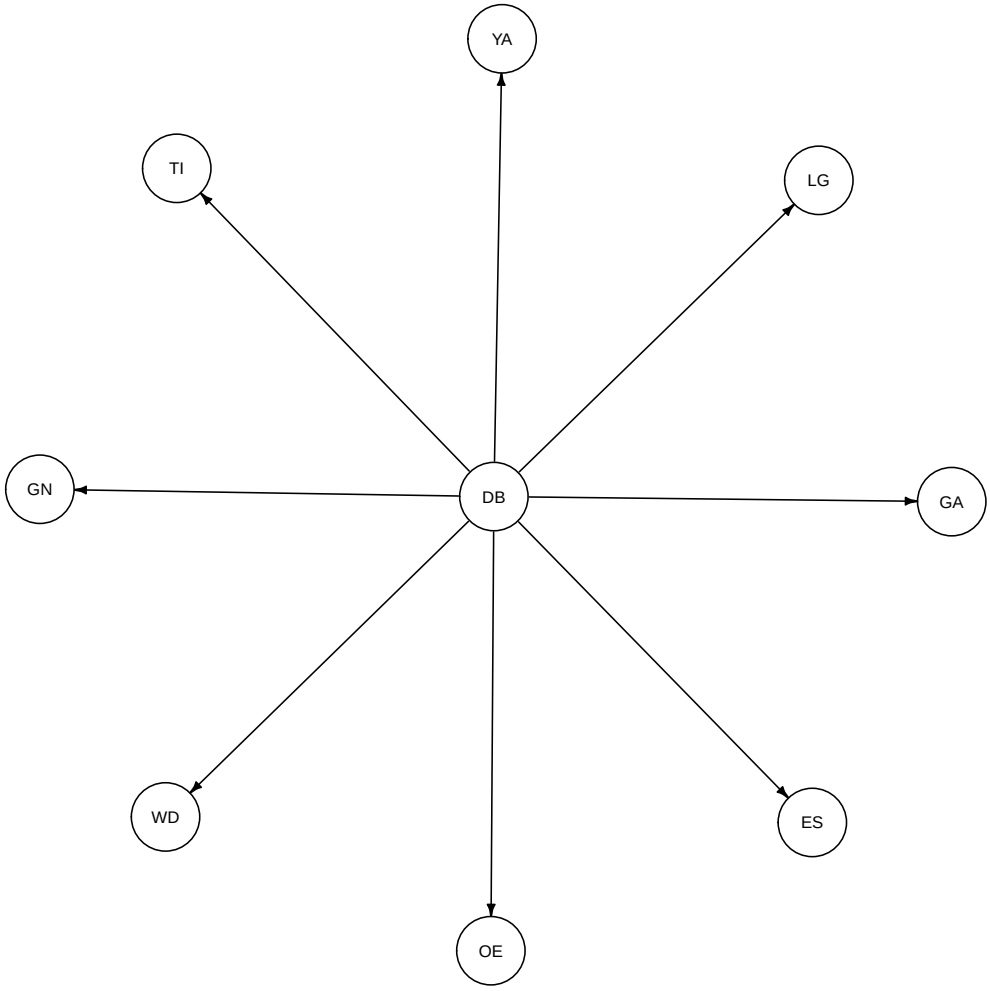
[138] Přívlastek „country“ je k prvku doplněný z důvodu odlišení Gruzie od amerického státu Georgia.



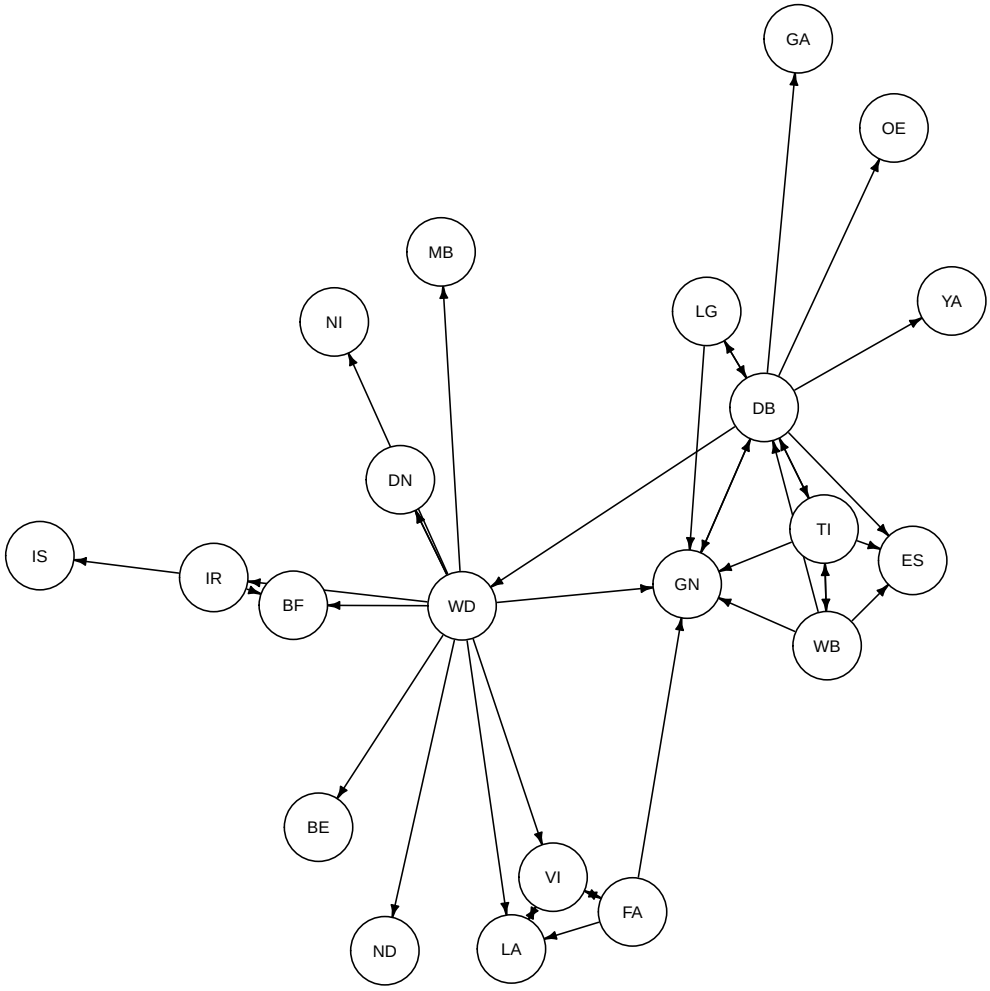
Obrázek 45: Datová síť objektu `Israel`.



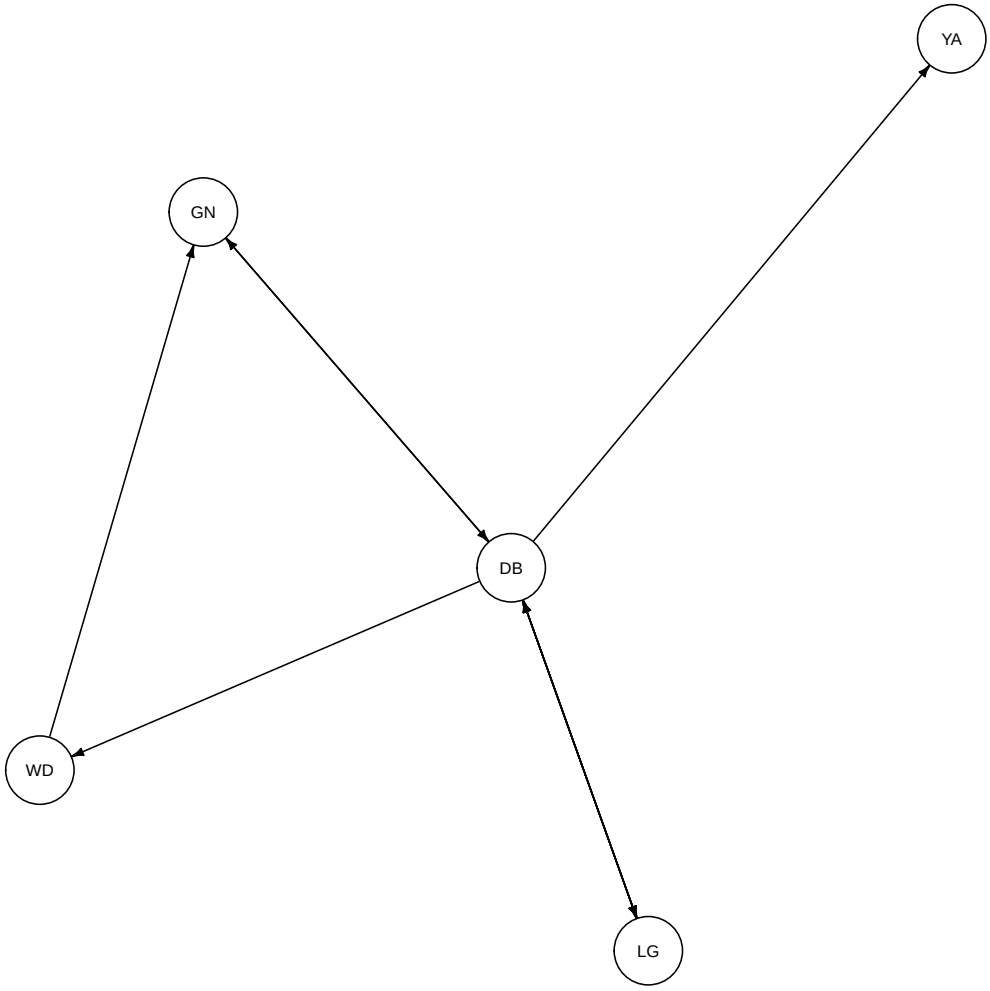
Obrázek 46: Datová síť objektu reprezentujícího SAARC.



Obrázek 47: Datová síť objektu `Georgia_country_`.



Obrázek 48: Datová síť objektu `Niger`.



Obrázek 49: Datová síť objektu `John_Wyane_Airport`.

Shrnutí a interpretace výsledků experimentů

V předchozích částech výzkumu byly pomocí metodiky založené na metrikách a následné multikriteriální analýze popisovány a hodnoceny grafové struktury. Tyto grafy reprezentují objekty propojených prostorových dat. Uzly představují zdroje (datové sady) poskytující propojená prostorová data (tedy reprezentace nebo konkrétní instance objektů). Hrany propojující uzly zastupují identické a podobnostní vazby (v současné fázi výzkumu dále nečleněné).

Z výsledků získaných v dílčích případových studiích a také během zpracování celého vzorku dat lze získat informace o zdrojích propojených dat, které obsahují prostorová data, a také o způsobu, jakým jsou v oblasti prostorových dat zavedeny identické vazby. Struktura této části textu je následující:

1. Shrnutí významu jednotlivých metrik pro uzly i grafy,
2. sumarizace výsledků analýz týkajících se uzlů,
3. popis důležitých zdrojů pro propojená prostorová data a jejich vhodnosti pro specifické účely,
4. sumarizace výsledků analýz týkajících se grafů,
5. doporučení týkající se implementace objektových vazeb v oblasti propojených prostorových dat.

Význam jednotlivých metrik

Pro účely analýz bylo v kapitole Metodika vybráno celkem šesti metrik týkajících se uzlů grafu a čtyř metrik zaměřených na grafy jako celky:

Centralita stupně – centralita obecně představuje kritická pozice uzlu v grafu [32, 35]. Centralita stupně (neboli stupeň uzlu) 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. Distribuce hodnot centrality stupně, která v případě této práce není akcentována, je důležitá z hlediska vztahu k bezškálovým sítím, které jsou robustní a odolné vůči náhodným chybám. 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 (z hlediska nejkratší cesty) procházení datové sítě při získání nových informací z reprezentací stejného geografického objektu. Tato metrika nemá pro 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 [111] 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 [111] 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 tato metrika 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 (Obrázek 51). 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).

Shrnutí výsledků uzlových analýz

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 [152]);
- zpracování dat – pro datové sady, které nebyly v době získání validní z pohledu RDF[139] (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[140];
- 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;

[139] Byl poskytnutý například pouze náhled ve formě webových stránek.

[140] Toto omezení je však plně v souladu s Linked Data přístupem, který se opírá právě o strojové zpracování.

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).

Vybrané datové sady obsahující propojená prostorová data

Z předchozího shrnutí vychází, že existuje několik málo zdrojů propojených dat, které hrají významnou roli i na poli dat prostorových. Následující seznam poskytuje abecední výčet a krátký popis těchto datových sad.

DBpedia[141] vznikla za účelem extrahování strojově čitelných informací z Wikipedie a jejich bezplatného zpřístupnění ve formě propojených dat. Podobně jako v případě Wikipedie se jedná o crowdsourcingový projekt, který podporují univerzity v Lipsku a Mannheimu společně s firmou OpenLink Software. První verze dat byla publikována v roce 2007. V současnosti DBpedia obsahuje přibližně 4,5 miliónu prvků (3 biliony RDF trojic), které jsou poskytovány ve formě RDF trojic (syntaxe Turtle)[142]. DBpedia poskytuje pro stahování dat i SPARQL endpoint[143],

[142]Nově jsou k dispozici i tzv. quad-turtle obsahující subjekt, predikát, objekt a grafový kontext.

[143]http://dbpedia.org/sparql

který byl využitý pro získání dat i v této práci. Další informace o tomto datovém zdroji je možné získat z publikací [132, 153, 154].

GeoNames.org[144] se označuje jako geografická databáze. Obsahuje 11 miliónů pojmenovaných geografických objektů a je v souladu s přístupem Linked Open Data publikovaná zdarma. Na rozdíl od projektů jako Wikidata nebo DBpedia data nepochází od uživatelů, ale jedná se o data poskytnutá významnými organizacemi, jako jsou národní statistické úřady, ministerstva, národní mapovací agentury nebo mezinárodní projekty. Data jsou poskytována ve formě dumpů (vyexportovaných souborů z databáze) nebo webových služeb[145]. Databáze začala fungovat v roce 2005. GeoNames.org se nepovažuje za výzkumnou aktivitu, proto je výskyt odborných článků o této datové sadě velice řídký, přičemž se tyto publikace zabývají především dílčími problémy dat uložených v GeoNames.org (například článek [155] hodnotí přesnost GeoNames.org).

Wikidata[146] představují do jisté míry konkurenční projekt k databázi DBpedia. I v tomto případě se jedná o formalizaci nestrukturovaných informací z Wikipedie a o jejich poskytování v podobě Linked Data. Podobně jako DBpedia i Wikidata jsou otevřená, založená na crowdsourcingu a vícejazyčná. Projekt spravuje nadace Wikimedia (Wikimedia Foundation) se sídlem v USA, ale podnět na vytvoření databáze Wikidata pochází z Německa. Projekt začal v roce 2012. Datová sada v současnosti obsahuje více než 26 milionů dat. Data jsou poskytována jako dumpy, přes vlastní API nebo skrze SPARQL endpoint[147]. Další informace jsou k dispozici například v publikacích [150, 156, 157]. Bakalářská práce [9] porovnává datové sady DBpedia a Wikidata jako zdroje prostorových dat.

Yago[148] je znalostní báze, kterou od roku 2008 vyvíjí Max Planck Institute for Computer Science v německém Saarbrückenu (v současné době se na projektu Yago podílí také Telecom ParisTech University)[149]. Data do této datové sady jsou získána především strojovou extrakcí z Wikipedie,

[145] http://www.geonames.org/export/ws-overview.html

[147] https://query.wikidata.org/

[149] Vývoj znalostní databáze Yago je v současnosti zřejmě velice pomalý, o čemž svědčí webové stránky produktu, které obsahují poslední novinku z roku 2015.

obsahově lze tedy Yago označit jako produkt crowdsourcingu. Mezi další zdroje patří WordNet (lexikální databáze angličtiny, Yago používá především taxonomii WordNetu) a GeoNames.org. Yago obsahuje data o více než 10 miliónech prvků. Přístup k datům je omezený pouze na stahovací služby (to je důvod, proč je Yago v datech o identických vazbách získaných pro tuto práci klasifikován jako listový uzel). Další informace o produktu Yago jsou k dispozici například v publikacích [158–161].

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 (v kategorii evropských hlavních měst je výsledek vícekriteriální analýzy pro DBpedii a Wikidata téměř totožný). Posledním zdrojem, který má vysoké hodnoty váženého součtu, jsou GeoNames.org, která představují druhý nejvhodnější zdroj pro kategorie dat hory v ČR, uzlová letiště a stratovulkány.

Na základě výsledků multikriteriální analýzy lze datové zdroje rozdělit na skupiny, které mají podobné hodnoty váženého součtu:

- DB – 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.

- WD, YA, GN – Skupina, která představuje dominantní datové zdroje nejen z hlediska vícekriteriální analýzy, ale také z pohledu jednotlivých metrik (viz odstavce výše).
- VI – 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).
- LG, DN, LA – 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.
- BF, IR, FA, IS, ND, MB, GA – 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).
- NI, TI, ES, FP, WB, OE, LW, NK, AU, BE – 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.

Metriky pro grafové struktury

Jak vyplývá z obrázků 20, 23, 24, 46 nebo 49 mnoho grafových metrik (s výjimkou velikosti grafu) dosahuje svých maximálních hodnot v grafech, které mají malý počet uzlů. Takové grafové struktury však není možné považovat za příklady dobré praxe z hlediska toho, jak popisovat prostorové objekty pomocí identických vazeb. Z tohoto důvodu budou takové objekty, které by mohly

inspirovat editory databází propojených dat, vybírány z množiny prvků, které se z hlediska počtu uzlů umístily v intervalu mezi horním kvartilem a maximem (Obrázek 44).

Po výše popsaném omezení vzorku dat na 950 prvků s nejvyššími hodnotami velikosti grafu došlo k následujícím změnám statistických parametrů absolutních hodnot grafových kritérií oproti tabulce 31 – tabulka 32 ukazuje změny v procentech v jednotlivých kritériích.

Tabulka 32: Statistické parametry absolutních hodnot grafových kritérií pro grafy s velikostí větší než 6.

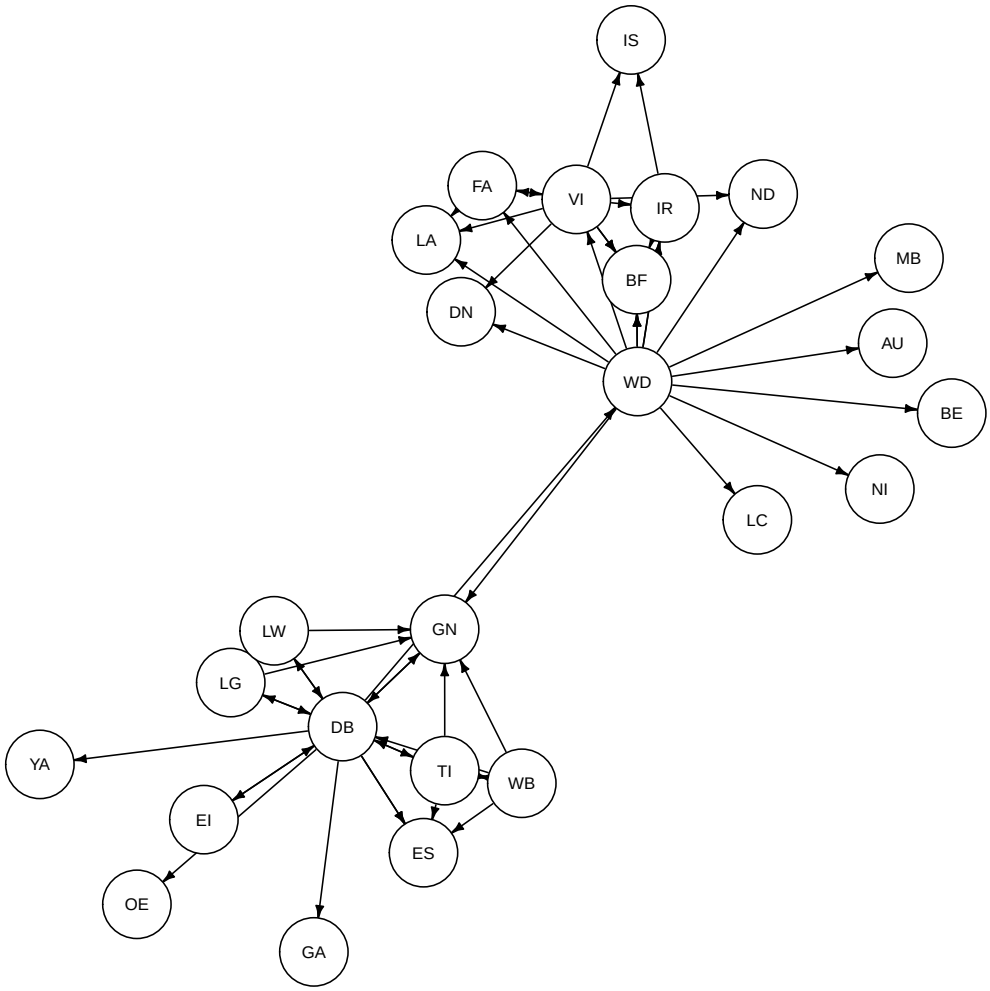
	Hustota	Reciprocita	Shlukový koef.
Maximum	51.81%	84.93%	53,00%
Minimum	100,00%	100.00%	100,00%
Průměr	59.62%	139.91%	125,82%
Variační koeficient	117,00%	43,54%	31,50%

Tabulka 32 ukazuje podobné změny v reciprocitě a shlukovém koeficientu. Se zúžením vzorku v obou příkladech pokleslo maximum, což podporuje původní předpoklad, že vysoké hodnoty těchto metrik dosahují především grafy se třemi, čtyřmi nebo pěti uzly (zdroji). Z hlediska minima nedošlo k žádným změnám ani u jedné z porovnávaných metrik. Naopak se zvedla průměrná hodnota reciprocity i shlukového koeficientu a došlo ke snížení variability hodnot v souboru dat.

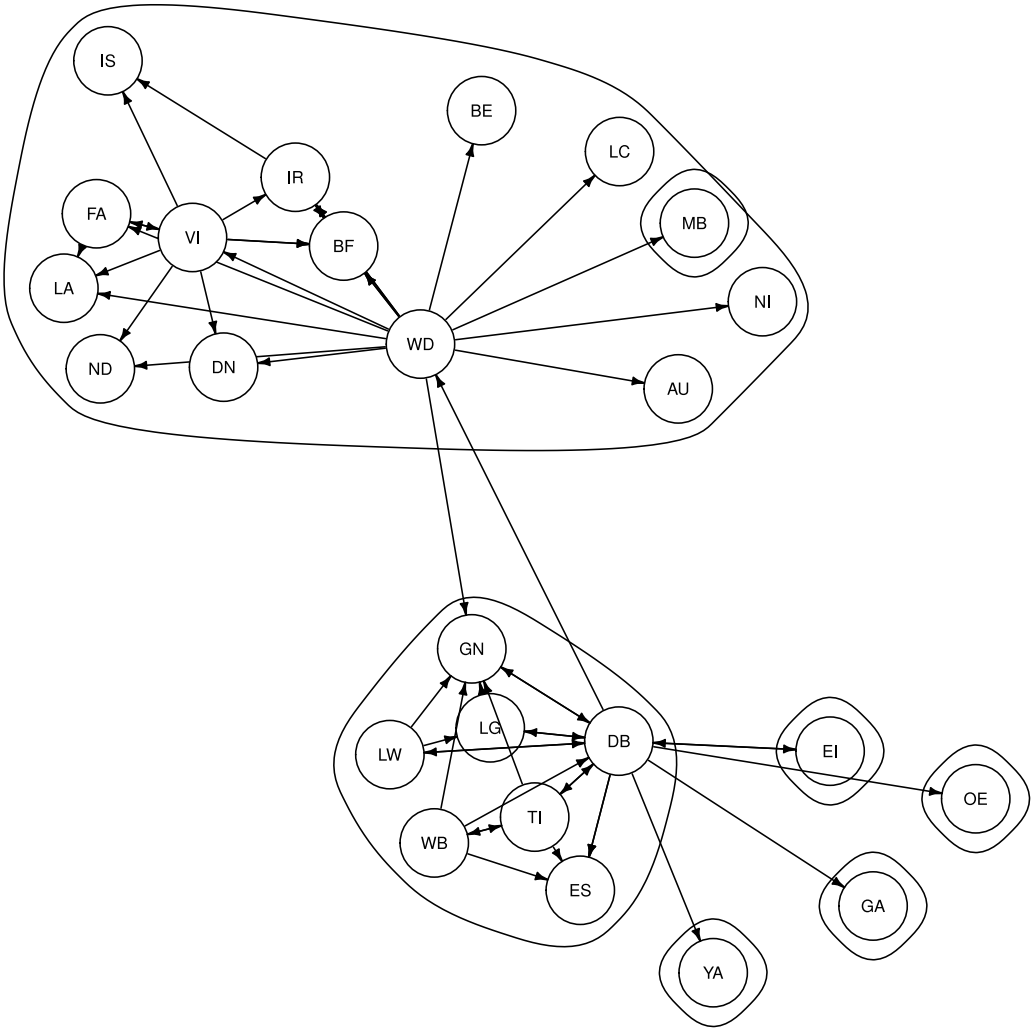
V případě hustoty grafu jsou změny odlišné. Hodnoty průměru a maxima se zmenšily a variační koeficient se naopak oproti celému datovému souboru zvýšil. Pro vysvětlení těchto změn je nutné si uvědomit způsob výpočtu hustoty, kdy pro graf se třemi uzly a čtyřmi hranami bude hustota $\frac{4}{3}$, zatímco pro graf se čtyřmi uzly a čtyřmi hranami $\frac{4}{4}$ a pro graf s pěti uzly a čtyřmi hranami jen $\frac{4}{5}$.

Z hlediska grafů dosahujících maximálních hodnot u jednotlivých metrik jsou výsledky podobné kompletnímu vzorku dat. Obecně se dá říct, že čím je graf menší, tím lepších dosahuje hodnot reciprocity, hustoty i shlukového koeficientu. Proto byl vytvořený koeficient, který přepočítává dosažené

hodnoty metrik na poměr počtu uzlů k maximálnímu počtu vrcholů ve vzorku. Po tomto přepočtu se ukázalo, že hustota grafu je stále nejvyšší pro struktury s nejmenším počtem vrcholů, což může být chápáno jako důkaz závislosti obou metrik (viz případová studie Hraniční řeky, tabulka 10). Nejvyšší reciprocitu zaznamenal graf ilustrující identické vazby pro objekt `Italy` (Obrázek 45) a nejvyšší shlukový koeficient graf popisující objekt `Bulgaria` (Obrázky 50 a 51).



Obrázek 50: Datová síť objektu `Bulgaria`.



Obrázek 51: Datová síť objektu `Bulgaria` s vyjádřením shluků.

Doporučení

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.

Důležitost výběru vhodných uzlů k propojení pomocí identických vazeb ukazuje následující příklad. Na něm je použita datová síť objektu `Iguazu_River` (Obrázek 20). Ke stávajícímu grafu je připojený pomocí identických vazeb další uzel reprezentující nový datový zdroj. Tabulka 33 ukazuje jaký vliv má zavedení identických vazeb mezi novým uzlem a existujícími uzly (použity byly všechny dostupné jedno-, dvou-, tři- a

čtyřprvkové kombinace) na shlukový koeficient.

Tabulka 33: Vliv připojení nové uzlu na shlukový koeficient.

Připojené zdroje	Shlukový koeficient
WD-GN-DB	0,80
GN-WD-YA-DB	0,79
WD-YA-DB	0,64
YA-GN-DB	0,64
DB-YA	0,60
WD-GN	0,60
originální data	**0,60**
DB-GN	0,55
DB-WD	0,55
YA	0,50
YA-GN-WD	0,46
GN	0,43
WD	0,43
DB	0,38
GN-YA	0,33
WD-YA	0,33

Z tabulky 33 je zřetelné, že zlepšení shlukového koeficientu bylo zaznamenáno nejen u připojení nové uzly na všechny vrcholy původního grafu, ale dokonce větší nárůst se projevil v případě trojice uzlů, které s novým vrcholem utvořily uzavřené trojúhelníky. To samé platí i opačně – ačkoli byl zdroj zdroj připojený na tři ze čtyř jiných zdrojů, došlo ke zhoršení shlukového koeficientu, protože rozložení uzlů v grafu nevykazuje dostatečný počet uzvřených trojúhelníků, které jsou vhodné z hlediska robustnosti a odolnosti grafu.

Rozšíření a pokračování výzkumu

Závěry publikované v této prácí rozhodně nepředstavují ukončený výzkum na poli identických vazeb prostorových propojených dat. Existuje několik dalších směrů, kterými by se autor a případně také jeho studenti, chtěli v budoucnosti vydat.

Z praktického hlediska jde především o vylepšení a optimalizaci všech částí práce, které souvisejí s programovým kódem. V první řadě jde o skript, pomocí něhož se získávají data o identických vazbách pro jednotlivé objekty. Jak již bylo uvedeno, skript využívá především jazyk XSLT a pro jeho zpracování procesor Saxon, konkrétně verzi založenou na jazyku Java. Proto (a na vině je samozřejmě také kvalita kódu, protože se autor v žádném případě nepovažuje za programátora) je celý proces získávání informací velice pomalý. Krokem vpřed by tedy měla být zcela nová sada skriptů nebo program, který by stále využíval „Follow Your Nose“ přístup, ale fungoval by mnohem efektivněji než současná verze.

Se stahováním dat souvisí také získávání informací o vazbách. V současné době je omezené pouze na validní RDF soubory, což sice na jedné straně odpovídá principům propojených dat, které by měly být strojově zpracovatelné, ale na straně druhé nejsou získány užitečné informace o identických a podobnostních vazbách, které mohou být skryté v nevalidních souborech nebo například ve webových stránkách, které obsahují vizualizaci originálních RDF dat. Proto by do celého procesu sběru informací mohlo být v budoucnosti zařazené také zpracování jiných než RDF souborů.

V současném výzkumu nebyly akcentovány hrany. Další pokračování se tedy může týkat zavedení vah pro hrany. Tyto váhy by odpovídaly „síle“ konkrétní standardizované vazby. Například identické vazby by měly přiřazené vyšší váhy než vazby podobnostní. Pro testování takových grafů z hlediska uzlů a objektů by bylo zřejmě nutné modifikovat metriky a metodiku popsanou v této práci. Navíc by bylo možné vyhodnocovat identické vazby z pohledu standardů.

Posledním krokem pokračování popisovaného výzkumu je rozšíření vzorku dat a srovnání závěrů s publikovanými výsledky. Nejde však jen o kvantitu

zkoumaného vzorku, zajímavá by mohla být také longitudinální studie zkoumající vývoj identických vazeb určitého vzorku v čase, která by ilustrovala vývoj propojených dat. Získané výsledky by mohly být testovány vzhledem ke geografickým faktorům tak, jak jsou v této práci publikované ukázky týkající se geopolitického rozdělení světa. Těmito faktory by mohly být konkrétní statistiky pro jednotlivé typy objektů (například hrubý domácí produkt pro státy, počet obyvatel pro sídla apod.). Tím by se dostalo pozornosti skutečné sociální roli propojených prostorových dat, což by mohlo být zajímavé především z toho důvodu, že řada sad prostorových dat (například Wikidata nebo LinkedGeoData) je tvořena dobrovolníky podle principů crowdsourcingu. To sice s velkou pravděpodobností snižuje kvalitu dat ve smyslu přesnosti, integrity a podobně, ale na druhou stranu to podporuje různorodost a možnost integrace nových informací.

Důležité však není jen pokračování vlastního výzkumu, ale také implementace jeho výsledků do praxe. V tomto konkrétním případě je plánováno zohlednění výsledků při dalším rozvoji datové sady Smart Points of Interest[150] (SPOI), která je vyvíjena autorem na Západočeské univerzitě v Plzni. Jedná se o největší datovou sadu bodů zájmu na světě (přibližně 28 000 000 bodů), která je publikovaná ve formě propojených dat. Výsledky publikované v této práci budou sloužit jednak pro vyhledávání vhodných zdrojů obsahujících další reprezentace bodů zájmu a také pro návrh takové struktury identických vazeb směřujících ze SPOI, aby došlo k posílení propojených prostorových dat jako celku. SPOI by mohly tvořit „mosty“ mezi nepropojenými nebo řídce propojenými oblastmi prostorových propojených dat. Může se jednat například o posílení vazeb mezi blokem DBpedie a příbuzných sémantických nástrojů a mezi skupinou tzv. zemědělských tezaurů (například AGROVOC), které jsou zatím velice omezené.

Druhou aplikační oblastí může být tvorba tezaurů a ontologií vycházejících z potřeb GeoInfoStrategie. V tomto případě by opět mohlo být užitečné hledání a vyhodnocování propojení na existující sémantické nástroje v zahraniční.

Z hlediska rozšíření nejsou úmyslně zmíněny možné studie, týkající se kvality obsahu propojených prostorových dat a oprávněnosti existence nebo správnosti

[150] http://kgm.zcu.cz/spoi

zavedení identických vazeb. Tyto experimenty bohužel není možné ve velké míře automatizovat. Navíc se jedná o ryze multidisciplinární výzkum, který dalece přesahuje hranice této práce.

Kapitola 7

Závěr

Habilitační práce 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 (jen za poslední dva roky je na portálu Web of Science registrováno více než 48 000 článků, které uvádějí propojená data jako téma příspěvku), 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[151], 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.

[151] V Česku například existuje Polytematický strukturovaný heslář (https://psh.techlib.cz/skos/) vytvořený Národní technickou knihovnou.

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 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 28 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. Právě první dva ze čtyř produktů uvedených v předchozí větě představují hlavní datové zdroje ve světě Linked Data (Obrázek 1), které jsou zaměřené výhradně na propojená data[152].

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

[152]GeoNames.org jsou v rámci této práce popsané v kapitole Výsledky, LinkedGeoData jsou zmíněné v kapitole Experimenty.

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.

Nejprve jsou v textu definovány a popsány základní pojmy z oblasti prostorových a propojených dat a také z teorie grafů, protože kvantifikovaný popis identických vazeb byl realizován pomocí grafových struktur (tzv. datových sítí) a jejich vlastností. Následuje část věnovaná rešerším, které jsou rozděleny na zdroje popisující hodnocení identických vazeb propojených dat a na publikace zabývající se metodami pro kvantitativní popis grafů. Poté byla ze zvolených metrik sestavená metodika pro hodnocení uzlů v grafu (uzly reprezentují zdroje propojených dat) a celých grafů, které vyjadřují jednotlivé prvky prostorových dat. Souhrnné výsledky byly získány pomocí multikriteriální analýzy, kde dílčí metriky tvoří kritéria a zdroje propojených

dat nebo objekty prostorových dat jsou alternativami, z nichž se pomocí váženého součtu vybírá ta nejvhodnější. Celá metodika je testovaná v další části na celkem osmi experimentech zpracovávajících různé podmnožiny prostorových dat, kde se jednak ověřuje a optimalizuje metodika samotná a její funkčnost a jednak se potvrzují nebo zamítají úvodní předpoklady týkající se vlivu základních geografických charakteristik na propojená prostorová data. Předposlední kapitola shrnuje získané poznatky. Nejprve jsou pomocí metodiky zpracovány sady geografických objektů z případových studií doplněné ještě o další data. Tato množina by měla představovat reprezentativní průřez prostorovými daty z hlediska mnoha faktorů jako je pokrytí, způsob georeferencování, charakter geometrie nebo příslušnost k různým geografickým disciplínám. Na základě výsledků všech experimentů jsou navrženy důležité existující zdroje pro propojená prostorová data (s ohledem na konkrétní způsoby využití těchto dat) a také postup pro tvorbu identických vazeb v rámci případné nové sady propojených prostorových dat. Před závěrem práce je ještě uvedena krátká pasáž týkající se směrů následného výzkumu v oblasti identických vazeb propojených prostorových dat.

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 habilitační práce 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 důležité 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[153] 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.

Habilitační 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[154]. 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, ontologii 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.

[153]Autor si plně uvědomuje, že ideální stav není záležitostí blízké budoucnosti, neboť je nutné vyřešit otázky týkající se například garance dat, kvality obsahu, velikosti souborů a podobně.

[154]Veřejná zakázka Technologické agentury České republiky TB0500MV003.

Seznam literatury

1. Devillers R, Stein A, Bédard Y, et al (2010) Thirty years of research on spatial data quality: achievements, failures, and opportunities. Transactions in GIS 14:387–400.

2. Oort PA van (2006) Spatial data quality: from description to application. Wageningen Universiteit

3. Dragland A (2013) Big Data–for better or worse. SINTEF, retrieved on July 22:

4. Gantz J, Reinsel D (2011) Extracting value from chaos. IDC iview 1142:1–12.

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

6. Directive I (2007) 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

7. Berners-Lee T (2006) Linked data: Design Issues.

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

9. Macura J (2016) Porovnání projektů Wikidata a DBpedia jako zdrojů prostorových dat. Bachelor Thesis, University of West Bohemia

10. Wood D, Zaidman M, Ruth L, Hausenblas M (2014) Linked Data. Manning

Publications Co.

11. Heath T, Bizer C (2011) Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology 1:1–136.

12. Zaveri A, Rula A, Maurino A, et al (2015) Quality assessment for linked data: A survey. Semantic Web 7:63–93.

13. Acosta M, Zaveri A, Simperl E, et al (2013) Crowdsourcing linked data quality assessment. In: International Semantic Web Conference. Springer, pp 260–276

14. Ding L, Shinavier J, Finin T, McGuinness DL (2010) owl: sameAs and Linked Data: An empirical study.

15. Ding L, Shinavier J, Shangguan Z, McGuinness DL (2010) SameAs networks and beyond: analyzing deployment status and implications of owl: sameAs in linked data. In: International Semantic Web Conference. Springer, pp 145–160

16. Bechhofer S, Buchan I, De Roure D, et al (2013) Why linked data is not enough for scientists. Future Generation Computer Systems 29:599–611.

17. Estrada E, Bodin Ö (2008) Using network centrality measures to manage landscape connectivity. Ecological Applications 18:1810–1825.

18. Hogan A, Polleres A, Umbrich J, Zimmermann A (2010) Some entities are more equal than others: statistical methods to consolidate linked data. 4th International Workshop on New Forms of Reasoning for the Semantic Web: Scalable and Dynamic (NeFoRS2010)

19. Toupikov N, Umbrich J, Delbru R, et al (2009) DING! Dataset Ranking using Formal Descriptions. LDOW

20. Glaser H, Jaffri A, Millard I (2009) Managing co-reference on the semantic web.

21. Čerba O (2012) Doménová ontologie – nástroj pro harmonizaci prostorových dat. PhD thesis, Univerzita Karlova v Praze

22. Kavouras M, Kokla M (2007) Theories of geographic concepts: ontological

approaches to semantic integration. CRC Press

23. Laurence S, Margolis E (1999) Concepts and cognitive science. Concepts: core readings 3–81.

24. Haav H-M, Kaljuvee A, Luts M, Vajakas T (2009) Ontology-based retrieval of spatially related objects for location based services. In: OTM Confederated International Conferences“ On the Move to Meaningful Internet Systems”. Springer, pp 1010–1024

25. Schwering A (2008) Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. Transactions in GIS 12:5–29.

26. Kovacs K, Dolbear C, Goodwin J (2007) Spatial concepts and OWL issues in a topographic ontology framework. Proc. of the GIS

27. Schwering A, Raubal M (2005) Spatial relations for semantic similarity measurement. In: International Conference on Conceptual Modeling. Springer, pp 259–269

28. Mark DM, Skupin A, Smith B (2001) Features, objects, and other things: Ontological distinctions in the geographic domain. In: International Conference on Spatial Information Theory. Springer, pp 489–502

29. Rapant P (2006) Geoinformatika a geoinformační technologie. 513.

30. Šíma J (2003) Geoinformační terminologie pro geodety a kartografy. Vyzkumny ústav geodeticky, topograficky a kartograficky. Zdiby

31. Bizer C, Heath T, Idehen K, Berners-Lee T (2008) Linked data on the web (LDOW2008). In: Proceedings of the 17th international conference on World Wide Web. ACM, pp 1265–1266

32. Guéret C, Groth P, Van Harmelen F, Schlobach S (2010) Finding the achilles heel of the web of data: using network analysis for link-recommendation. In: International Semantic Web Conference. Springer, pp 289–304

33. Bonatti PA, Hogan A, Polleres A, Sauro L (2011) Robust and scalable linked data reasoning incorporating provenance and trust annotations. Web

Semantics: Science, Services and Agents on the World Wide Web 9:165–201.

34. De Melo G (2013) Not Quite the Same: Identity Constraints for the Web of Linked Data. AAAI

35. Guéret C, Groth P, Stadler C, Lehmann J (2012) Assessing linked data mappings using network measures. In: Extended Semantic Web Conference. Springer, pp 87–102

36. Halpin H, Hayes PJ, McCusker JP, et al (2010) When owl: sameas isn't the same: An analysis of identity in linked data. In: International Semantic Web Conference. Springer, pp 305–320

37. Bennett B (2001) Application of supervaluation semantics to vaguely defined spatial concepts. Springer, pp 108–123

38. Bennett B (2002) Physical objects, identity and vagueness. In: KR. pp 395–408

39. Bennett B, Agarwal P (2007) Semantic categories underlying the meaning of "place". In: International Conference on Spatial Information Theory. Springer, pp 78–95

40. Tauberer J (2006) What is RDF. XML. com 26:

41. Bergman M (2009) Advantages and Myths of RDF. AI3, April

42. W3C (2012) OWL 2 web ontology language document overview.

43. Miles A, Bechhofer S (2009) SKOS simple knowledge organization system reference. W3C recommendation 18:W3C.

44. Halpin H, Hayes PJ, Thompson HS (2015) When owl: sameAs isn't the same redux: towards a theory of identity, context, and inference on the semantic web. In: International and Interdisciplinary Conference on Modeling and Using Context. Springer, pp 47–60

45. Hogan A, Decker S, Harth A (2007) Performing object consolidation on the semantic web data graph.

46. Dean M, Schreiber G, Bechhofer S, et al (2004) OWL web ontology

language reference. W3C Recommendation February 10:

47. Hogan A, Zimmermann A, Umbrich J, et al (2012) Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. Web Semantics: Science, Services and Agents on the World Wide Web 10:76–110.

48. Brickley D, Guha R (2014) RDF Schema 1.1. W3C Recommendation (25 February 2014). World Wide Web Consortium

49. Čada R, Ryjáček Z, Kaiser T (2004) Diskrétní matematika. Západočeská univerzita

50. Ryjácek Z (2007) Teorie grafu, diskrétni optimalizace a vypocetni slozitost.

51. Tiddi I, d'Aquin M, Motta E (2014) Quantifying the bias in data links. In: International Conference on Knowledge Engineering and Knowledge Management. Springer, pp 531–546

52. Lu Q, Getoor L (2003) Link-based classification. In: ICML. pp 496–503

53. Golbeck J, others (2008) Trust on the world wide web: a survey. Foundations and Trends in Web Science 1:131–197.

54. Hartig O (2009) Querying trust in rdf data with tsparql. In: European Semantic Web Conference. Springer, pp 5–20

55. Bizer C, Cyganiak R (2009) Quality-driven information filtering using the WIQA policy framework. Web Semantics: Science, Services and Agents on the World Wide Web 7:1–10.

56. Volz J, Bizer C, Gaedke M, Kobilarov G (2009) Discovering and maintaining links on the web of data. In: International Semantic Web Conference. Springer, pp 650–665

57. Magnani M, Montesi D (2010) A survey on uncertainty management in data integration. Journal of Data and Information Quality (JDIQ) 2:5.

58. Bartolomeo G, Salsano S, Glaser H (2013) On the Likelihood of an Equivalence. In: OTM Confederated International Conferences" On the Move

to Meaningful Internet Systems". Springer, pp 2–11

59. Bouquet P, Stoermer H, Mancioppi M, Giacomuzzi D (2006) OkkaM: Towards a Solution to the"Identity Crisis"on the Semantic Web. SWAP 201:

60. Jain P, Hitzler P, Yeh PZ, et al (2010) Linked Data Is Merely More Data. AAAI Spring Symposium: linked data meets artificial intelligence 11:

61. Jaffri A, Glaser H, Millard I (2008) Uri disambiguation in the context of linked data.

62. McCusker J, McGuinness DL (2010) owl: sameAs considered harmful to provenance. Proceedings of the ISCB Conference on Semantics in Healthcare and Life Sciences

63. Sleeman J, Finin T (2010) Computing foaf co-reference relations with rules and machine learning. Proceedings of the third international workshop on social data on the web

64. Leibniz GW (1989) Discourse on metaphysics. In: Philosophical papers and letters. Springer, pp 303–330

65. Guéret C, Wang S, Schlobach S (2010) The Web of Data is a complex system—first insight into its multi-scale network properties. Proceedings of the ECCS 10:

66. Tiddi I, d'Aquin M, Motta E (2014) Walking Linked Data: a graph traversal approach to explain clusters. In: Proceedings of the 5th International Conference on Consuming Linked Data-Volume 1264. CEUR-WS. org, pp 73–84

67. Tichy NM, Tushman ML, Fombrun C (1979) Social network analysis for organizations. Academy of management review 4:507–519.

68. Emirbayer M, Goodwin J (1994) Network analysis, culture, and the problem of agency. American journal of sociology 99:1411–1454.

69. Haythornthwaite C (1996) Social network analysis: An approach and technique for the study of information exchange. Library & information

science research 18:323–342.

70. Blüthgen N, Fründ J, Vázquez DP, Menzel F (2008) What do interaction network metrics tell us about specialization and biological traits. Ecology 89:3387–3399.

71. Abbasi A, Altmann J, Hossain L (2011) 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 5:594–607.

72. Cimenler O, Reeves KA, Skvoretz J (2014) A regression analysis of researchers' social network metrics on their citation performance in a college of engineering. Journal of Informetrics 8:667–682.

73. Freeman L (2004) The development of social network analysis. A Study in the Sociology of Science

74. Otte E, Rousseau R (2002) Social network analysis: a powerful strategy, also for the information sciences. Journal of information Science 28:441–453.

75. Agarwal A, Corvalan A, Jensen J, Rambow O (2012) Social network analysis of alice in wonderland. In: Workshop on Computational Linguistics for Literature. pp 88–96

76. Rubinov M, Sporns O (2010) Complex network measures of brain connectivity: uses and interpretations. Neuroimage 52:1059–1069.

77. Dunn AG, Westbrook JI (2011) Interpreting social network metrics in healthcare organisations: A review and guide to validating small networks. Social Science & Medicine 72:1064–1068.

78. Vaz de Melo PO, Almeida VA, Loureiro AA (2008) 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. ACM, pp 695–703

79. Clemente FM, Couceiro MS, Martins FML, Mendes RS (2015) Using network metrics in soccer: A macro-analysis. Journal of human kinetics

45:123–134.

80. Kossinets G, Watts DJ (2006) Empirical analysis of an evolving social network. science 311:88–90.

81. Bajaj A, Russell R (2010) AWSM: Allocation of workflows utilizing social network metrics. Decision Support Systems 50:191–202.

82. Varlamis I, Eirinaki M, Louta M (2010) 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. IEEE, pp 168–175

83. Hajian B, White T (2011) 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. IEEE, pp 497–500

84. Spiliotopoulos T, Oakley I (2013) Understanding motivations for Facebook use: Usage metrics, network structure, and privacy. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp 3287–3296

85. Varlamis I, Eirinaki M, Louta M (2013) 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, pp 49–74

86. Zimmermann T, Nagappan N (2009) Predicting defects using network analysis on dependency graphs. In: Software Engineering, 2008. ICSE'08. ACM/IEEE 30th International Conference on. IEEE, pp 531–540

87. Premraj R, Herzig K (2011) Network versus code metrics to predict defects: A replication study. In: Empirical Software Engineering and Measurement (ESEM), 2011 International Symposium on. IEEE, pp 215–224

88. Ding L, Finin T, Joshi A (2004) Analyzing social networks on the semantic web. IEEE Intelligent Systems (Trends & Controversies) 8:815–820.

89. Theoharis Y, Tzitzikas Y, Kotzinos D, Christophides V (2008) On graph

features of semantic web schemas. IEEE Transactions on Knowledge and Data Engineering 20:692–702.

90. Telesford QK, Morgan AR, Hayasaka S, et al (2010) Reproducibility of graph metrics in fMRI networks. Frontiers in neuroinformatics 4:117.

91. Gil R, García R, Delgado J (2004) Measuring the semantic web. AIS SIGSEMIS Bulletin 1:69–72.

92. Thung F, Lo D, Osman MH, Chaudron MR (2014) Condensing class diagrams by analyzing design and network metrics using optimistic classification. In: Proceedings of the 22Nd International Conference on Program Comprehension. ACM, pp 110–121

93. Li L, Alderson D, Doyle JC, Willinger W (2005) Towards a theory of scale-free graphs: Definition, properties, and implications. Internet Mathematics 2:431–523.

94. Barabási A-L (2009) Scale-free networks: a decade and beyond. science 325:412–413.

95. Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the internet topology. In: ACM SIGCOMM computer communication review. ACM, pp 251–262

96. Barabási A-L (2002) Emergence of scaling in complex networks. Handbook of graphs and networks: from the genome to the internet 69–84.

97. Bode M, Burrage K, Possingham HP (2008) Using complex network metrics to predict the persistence of metapopulations with asymmetric connectivity patterns. ecological modelling 214:201–209.

98. Latora V, Marchiori M (2001) Efficient behavior of small-world networks. Physical review letters 87:198701.

99. Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 35–41.

100. Freeman LC (1978) Centrality in social networks conceptual clarification.

Social networks 1:215–239.

101. Wellman B (1983) Network analysis: Some basic principles. Sociological theory 155–200.

102. Ristoski P, Schuhmacher M, Paulheim H (2015) Using graph metrics for linked open data enabled recommender systems. In: International Conference on Electronic Commerce and Web Technologies. Springer, pp 30–41

103. Sinha R, Mihalcea R (2007) Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In: Semantic Computing, 2007. ICSC 2007. International Conference on. IEEE, pp 363–369

104. Coursey K, Mihalcea R (2009) 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. Association for Computational Linguistics, pp 117–120

105. Grindrod P, Parsons MC, Higham DJ, Estrada E (2011) Communicability across evolving networks. Physical Review E 83:046120.

106. Zhao J, Yang T-H, Huang Y, Holme P (2011) Ranking candidate disease genes from gene expression and protein interaction: a Katz-centrality based approach. PloS one 6:e24306.

107. Borgatti SP, Everett MG (2006) A graph-theoretic perspective on centrality. Social networks 28:466–484.

108. Yan E, Ding Y (2009) Applying centrality measures to impact analysis: A coauthorship network analysis. Journal of the Association for Information Science and Technology 60:2107–2118.

109. Varlamis I, Eirinaki M, Louta M (2013) 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, pp 49–74

110. Hanneman RA, Riddle M (2005) Introduction to social network methods.

111. Nykl M (2013) Určování vyznamnosti vrchol grafu: PageRank a jeho

modifikace. Technical report No. DCSE/TR-2013-09, University of West Bohemia

112. Burt RS (2009) Structural holes: The social structure of competition. Harvard university press

113. Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. science 323:892–895.

114. Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM) 46:604–632.

115. Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: Bringing order to the web. Stanford InfoLab

116. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world'networks. nature 393:440–442.

117. Sundaresan SR, Fischhoff IR, Dushoff J, Rubenstein DI (2007) Network metrics reveal differences in social organization between two fission–fusion species, Grevy's zebra and onager. Oecologia 151:140–149.

118. Qu Y, Ge W, Cheng G, Gao Z (2009) Class association structure derived from linked objects.

119. Milgram S (1967) The small world problem. Psychology today 2:60–67.

120. Adamic LA (1999) The small world web. In: International Conference on Theory and Practice of Digital Libraries. Springer, pp 443–452

121. Borgatti S (1997) Structural holes. analytictech com 20:35–38.

122. Everett M, Borgatti SP (2005) Ego network betweenness. Social networks 27:31–38.

123. Arnaboldi V, Conti M, Passarella A, Pezzoni F (2012) 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). IEEE, pp 31–40

124. Newman ME, Forrest S, Balthrop J (2002) Email networks and the spread

of computer viruses. Physical Review E 66:035101.

125. Wasserman SS (1980) A stochastic model for directed graphs with transition rates determined by reciprocity. Sociological methodology 11:392–412.

126. Rodrigue J-P, Comtois C, Slack B (2013) The geography of transport systems. Routledge

127. Tinkler KJ (1977) An introduction to graph theoretical methods in geography.

128. Haggett P, Chorley RJ (2015) Network analysis in geography.

129. Rodrigue J-P, Comtois C, Slack B (2004) Transport Geography on the Web. Dept. of Economics & Geography, Hofstra

130. Xie F, Levinson D (2007) Measuring the structure of road networks. Geographical analysis 39:336–356.

131. Isele R, Umbrich J, Bizer C, Harth A (2010) 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. CEUR-WS. org, pp 29–32

132. Bizer C, Lehmann J, Kobilarov G, et al (2009) DBpedia-A crystallization point for the Web of Data. Web Semantics: science, services and agents on the world wide web 7:154–165.

133. Cudré-Mauroux P, Haghani P, Jost M, et al (2009) idMesh: graph-based disambiguation of linked data. In: Proceedings of the 18th international conference on World wide web. ACM, pp 591–600

134. Volz J, Bizer C, Gaedke M, Kobilarov G (2009) Silk-A Link Discovery Framework for the Web of Data. LDOW 538:

135. Ngomo A-CN, Auer S (2011) Limes-a time-efficient approach for large-scale link discovery on the web of data. integration 15:

136. Tartir S, Arpinar IB, Moore M, et al (2005) OntoQA: Metric-based

ontology quality analysis.

137. Greco S, Figueira J, Ehrgott M (2005) Multiple criteria decision analysis. Springer's International series

138. Jankowski P (1995) Integrating geographical information systems and multiple criteria decision-making methods. International journal of geographical information systems 9:251–273.

139. Huntington SP (2001) Střet civilizací. Praha: Rybka Publishers

140. Jedlička K, Hájek P, Cada V, et al (2016) Open transport map—Routable OpenStreetMap. In: IST-Africa Week Conference, 2016. IEEE, pp 1–11

141. Kingsford R (1999) Managing the water of the Border Rivers in Australia: irrigation, Government and the wetland environment. Wetlands Ecology and Management 7:25–35.

142. Barroso LA, Lino P, Porrua F, et al (2008) Cheap and clean energy: Can Brazil get away with that? In: Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE. IEEE, pp 1–8

143. Jia D-x, Bai J-h, Liang F-c (2010) Perspective Analysis of Cooperative Development of Hydropower Resources of Border Rivers between China and Russia [J]. Energy Technology and Economics 2:

144. Saaty TL (1977) A scaling method for priorities in hierarchical structures. Journal of mathematical psychology 15:234–281.

145. Guitouni A, Martel J-M (1998) Tentative guidelines to help choosing an appropriate MCDA method. European Journal of Operational Research 109:501–521.

146. Triantaphyllou E, Mann SH (1995) Using the analytic hierarchy process for decision making in engineering applications: some challenges. International Journal of Industrial Engineering: Applications and Practice 2:35–44.

147. Hwang C-L, Yoon K (2012) Multiple attribute decision making: methods

and applications a state-of-the-art survey. Springer Science & Business Media

148. Fotr J, Dědina J, Hrůzová H (2000) Manažerské rozhodování. Ekopress

149. Zmeškal Z (2009) Vícekriteriální hodnocení variant a analyza citlivosti při vyběru produkt finančních institucí. Finanční řízení podnik a finančních institucí 7. mezinárodní vědecká conference

150. Vrandečić D, Krötzsch M (2014) Wikidata: a free collaborative knowledgebase. Communications of the ACM 57:78–85.

151. Williamson DF, Parker RA, Kendrick JS (1989) The box plot: a simple visual method to interpret data. Annals of internal medicine 110:916–921.

152. Palma R, Reznik T, Esbrí M, et al (2015) An INSPIRE-based vocabulary for the publication of Agricultural Linked Data. In: International Experiences and Directions Workshop on OWL. Springer, pp 124–133

153. Auer S, Bizer C, Kobilarov G, et al (2007) Dbpedia: A nucleus for a web of open data. The semantic web 722–735.

154. Lehmann J, Isele R, Jakob M, et al (2015) DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web 6:167–195.

155. Ahlers D (2013) Assessment of the accuracy of GeoNames gazetteer data. In: Proceedings of the 7th Workshop on Geographic Information Retrieval. ACM, pp 74–81

156. Erxleben F, Günther M, Krötzsch M, et al (2014) Introducing Wikidata to the linked data web. In: International Semantic Web Conference. Springer, pp 50–65

157. Pellissier Tanon T, Vrandečić D, Schaffert S, et al (2016) From freebase to wikidata: The great migration. In: Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, pp 1419–1428

158. Suchanek FM, Kasneci G, Weikum G (2007) Yago: a core of semantic knowledge. In: Proceedings of the 16th international conference on World Wide

Web. ACM, pp 697–706

159. Suchanek FM, Kasneci G, Weikum G (2008) Yago: A large ontology from wikipedia and wordnet. Web Semantics: Science, Services and Agents on the World Wide Web 6:203–217.

160. Biega J, Kuzey E, Suchanek FM (2013) Inside YAGO2s: A transparent information extraction architecture. In: Proceedings of the 22nd International Conference on World Wide Web. ACM, pp 325–328

161. Mahdisoltani F, Biega J, Suchanek F (2014) Yago3: A knowledge base from multilingual wikipedias. 7th Biennial Conference on Innovative Data Systems Research

Seznam příloh

Ostatní přílohy, včetně PDF souboru s textem práce, dat a zdrojových kódů, jsou k dispozici na webové adrese http://gis.zcu.cz/projekty/Identity_links/.

Příloha A. Jmenné prostory používané v habilitační práci

dbpedia-owl http://dbpedia.org/ontology/
fb http://rdf.freebase.com/ns/
foaf http://xmlns.com/foaf/0.1/
owl http://www.w3.org/2002/07/owl#
rdf http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs http://www.w3.org/2000/01/rdf-schema#
schema http://schema.org/"
skos http://www.w3.org/2004/02/skos/core#
wdt http://www.wikidata.org/prop/direct/
xsl http://www.w3.org/1999/XSL/Transform

Příloha B: Soubor resources.xml – seznam zdrojů propojených dat

```
<?xml version="1.0" encoding="utf-8"?>
<resources>
<resource acronym="AA" name="Getty Art &amp;amp; Architecture Thesaurus"
key="getty.edu"/>
<resource acronym="AU" name="Libraries Australia" key="nla.gov.au"
error="S"/>
<resource acronym="AV" name="AGROVOC" key="fao.org"/>
<resource acronym="BC" name="BNCF - Biblioteca Nazionale
Centrale di Firenze" key="bncf.firenze"/>
<resource acronym="BE" name="Biblioteca nacional de España"
key="datos.bne.es"/>
<resource acronym="BF" name="Bibliothèque nationale de France"
key="bnf.fr" error="S"/>
<resource acronym="BF" name="Bibliothèque nationale de France"
key="stitch.cs.vu.nl/vocabularies/rameau/ark:" error="S"/>
<resource acronym="BM" name="Freie Universität Berlin - University
of Manheim" key="www4.wiwiss.fu-berlin" error="T"/>
<resource acronym="BS" name="Bundesamt für Statistik Linked Data
(BFS)" key="bfs.270a" error="T"/>
<resource acronym="CT" name="Linked Open Data of Chinese
Agricultural Thesaurus(CAT)" key="aii.caas" error="T"/>
<resource acronym="CY" name="Cyc" key="cyc.com" error="T"/>
<resource acronym="DB" name="DBpedia" key="dbpedia"
special="true"/>
<resource acronym="DM" name="DMOZ" key="dmoz.org" error="S"/>
<resource acronym="DN" name="Deutschen Nationalbibliothek"
key="d-nb.info" error="S"/>
<resource acronym="DW" name="Dewey" key="dewey" error="T"/>
<resource acronym="EA" name="EARTh - Enviromental Applications
Reference THesaurus (via LusTRE)" key="EARTh" special="true"/>
<resource acronym="EB" name="ECB (European Central Bank)
```

Linked Data" key="ecb.270" error="T"/>
<resource acronym="EI" name="Eionet" key="eionet"
special="true"/>
<resource acronym="EN" name="eculture.cs.vu.nl"
key="purl.org" error="T"/>
<resource acronym="ES" name="Eurostat Linked Statistics"
key="eurostat.linked-statistics"/>
<resource acronym="EU" name="European Union Open Data Portal"
key="data.europa" error="T"/>
<resource acronym="EV" name="EuroVoc" key="eurovoc.europa"
error="S"/>
<resource acronym="FA" name="FAST Linked Data"
key="worldcat.org/fast" add_end="/rdf.xml"/>
<resource acronym="FB" name="Freebase" key="freebase"
error="T"/>
<resource acronym="FL" name="FAO (Food and Agriculture
Organization) of the United Nations Linked Data"
key="fao.270" error="T"/>
<resource acronym="FP" name="Facebook Places" key="facebook"
error="S"/>
<resource acronym="GA" name="GADM" key="gadm" add_end=".rdf"
error="S"/>
<resource acronym="GE" name="Genealogy.net"
key="genealogy.net" error="S"/>
<resource acronym="GH" name="Global Health Observatory"
key="ghodata" error="T"/>
<resource acronym="GK" name="GeoSpecies Knowledge Base"
key="lod.geospecies.org" error="T"/>
<resource acronym="GM" name="GEMET" key="gemet" error="S"/>
<resource acronym="GN" name="GeoNames.org" key="geonames"
add_end="about.rdf"/>
<resource acronym="GS" name="Gesis" key="gesis" error="T"/>
<resource acronym="HX" name="Humanitarian Response"
key="humanitarian" error="T"/>
<resource acronym="IE" name="Institut national de la statistique

et des études économiques" key="insee" error="S"/>
<resource acronym="IM" name="International Monetary Fund
Linked Data" key="imf.270a" error="T"/>
<resource acronym="IR" name="Identifiants et Référentiels"
key="idref.fr" add_end=".rdf"/>
<resource acronym="IS" name="ISNI - International Standard
Name Identifier" key="isni.org" error="S"/>
<resource acronym="IT" name="data.linkedopendata.it"
key="data.linkedopendata.it" error="T"/>
<resource acronym="LA" name="Library of Congress Name
Authority File" key="loc.gov/authorities"/>
<resource acronym="LC" name="Library and Archives of Canada"
key="collectionscanada.gc.ca" error="S"/>
<resource acronym="LG" name="LinkedGeoData" key="linkedgeodata"
special="true"/>
<resource acronym="LI" name="LIUC Thesauro di economia
e scienze sociali" key="biblio.liuc.it" error="S"/>
<resource acronym="LW" name="Linked Web APIs" key="cvut"
special="true"/>
<resource acronym="MB" name="MusicBrainz" key="musicbrainz"
error="S"/>
<resource acronym="ND" name="NDL" key="ndl.go.jp" error="S"/>
<resource acronym="NI" name="National Library of Israel"
key="nli.org.il" error="S"/>
<resource acronym="NK" name="Databáze Národní knihovny ČR"
key="aleph.nkp.cz" error="S"/>
<resource acronym="NL" name="National Agricultural Library
Thesaurus" key="lod.nal.usda" add_end=".rdf" error="S"/>
<resource acronym="NM" name="U.S. National Library of Medicine"
key="nih.gov" add_end=".rdf" error="S"/>
<resource acronym="NY" name="New York Times" key="nytimes" error="T"/>
<resource acronym="OE" name="OpenEI" key="openei" special="true"/>
<resource acronym="OS" name="Ordnance Survey"
key="ordnancesurvey.co.uk" error="T"/>
<resource acronym="PL" name="Katalogi Biblioteki Narodowej"

```
key="bn.org.pl" error="T"/>
<resource acronym="RA" name="RDF about" key="rdfabout.com"
error="T"/>
<resource acronym="ST" name="STW Thesaurus for Economics"
key=".eu/stw" special="true"/>
<resource acronym="SW" name="National Library of Sweden"
key="libris.kb.se" error="T"/>
<resource acronym="TI" name="Transparency International"
key="transparency" add_end=".rdf"/>
<resource acronym="UM" name="UMTHES" key="uba.de/umt"
special="true" error="T"/>
<resource acronym="US" name="UNESCO Institute for Statistics
Linked Data" key="uis.270" error="T"/>
<resource acronym="VI" name="VIAF" key="viaf.org"/>
<resource acronym="WB" name="World Bank Linked Data"
key="worldbank.270a" add_end=".rdf"/>
<resource acronym="WD" name="Wikidata" key="wikidata" add_end=".rdf"/>
<resource acronym="WK" name="Wolters Kluwer" key="wolterskluwer"
add_end=".rdf"/>
<resource acronym="YA" name="Yago" key="yago-knowledge" error="S"/>
<resource acronym="ZT" name="Zitigist Technologies" key="zitgist"
error="T"/>
</resources>
```

V příloze se vyskytují všechny testované zdroje. Některé z nich se nemusí objevit v textu práce, ani ve vzorcích dat.

Příloha C. SPARQL dotazy používané pro výběr objektů

Publikované SPARQL dotazy jsou využívány pro vyhledávání konceptů, jejichž identické vazby mezi reprezentacemi v různých zdrojích propojených dat jsou následně vyhodnocovány. Dotazy byly zadávány v Virtuoso SPARQL Query Editor připojeném k datové sadě DBpedia[1]. Výsledky dotazů byly ukládány ve formě CSV souborů k dalšímu zpracování. CSV soubory obsahují dva sloupce – anglický popisek prvku a identifikátor (URI) prvku.

Stratovulkány

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?l,?p
where {
?p dbo:type <http://dbpedia.org/resource/Stratovolcano> .
?p rdfs:label ?l.
filter(langMatches(lang(?l),"EN"))
}
```

Republiky

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dct: <http://purl.org/dc/terms/>

SELECT ?label,?uri
where {
?uri dct:subject
<http://dbpedia.org/resource/Category:Republics> .
```

[1] http://dbpedia.org/sparql/

```
?uri rdfs:label ?label.
filter(langMatches(lang(?label),"EN"))
}
```

Hlavní města Afriky

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dct: <http://purl.org/dc/terms/>

SELECT ?label,?uri
where {
?uri dct:subject
<http://dbpedia.org/resource/Category:Capitals_in_Africa> .
?uri rdfs:label ?label.
filter(langMatches(lang(?label),"EN"))
}
```

Hlavní města Evropy

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dct: <http://purl.org/dc/terms/>

SELECT ?label,?uri
where {
?uri dct:subject
<http://dbpedia.org/resource/Category:Capitals_in_Europe> .
?uri rdfs:label ?label.
filter(langMatches(lang(?label),"EN"))
}
```

Hraniční řeky

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
```

```
PREFIX dct: <http://purl.org/dc/terms/>

SELECT ?label,?uri
where {
?uri dct:subject
<http://dbpedia.org/resource/Category:Border_rivers> .
?uri rdfs:label ?label.
filter(langMatches(lang(?label),"EN"))
}
```

Okruhy Formule 1

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dct: <http://purl.org/dc/terms/>

SELECT ?label,?uri
where {
?uri dct:subject
<http://dbpedia.org/resource/Category:Formula_One_circuits> .
?uri rdfs:label ?label.
filter(langMatches(lang(?label),"EN"))
}
```

Okruhy série Indycar

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dct: <http://purl.org/dc/terms/>

SELECT ?label,?uri
where {
?uri dct:subject
<http://dbpedia.org/resource/Category:IndyCar_Series_tracks> .
?uri rdfs:label ?label.
filter(langMatches(lang(?label),"EN"))
```

```
}
```

Památky Světového dědictví UNESCO

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?label,?uri
where {
?uri rdf:type <http://dbpedia.org/ontology/WorldHeritageSite> .
?uri rdfs:label ?label.
filter(langMatches(lang(?label),"EN"))
}
```

Hory v ČR

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT ?label,?uri
where {
?uri rdf:type <http://dbpedia.org/ontology/Mountain> .
?uri dbo:locatedInArea
<http://dbpedia.org/resource/Czech_Republic> .
?uri rdfs:label ?label.
filter(langMatches(lang(?label),"EN"))
}
```

Mezinárodní organizace

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

```
SELECT ?label,?uri
where {
?uri rdf:type
<http://dbpedia.org/class/yago/WorldOrganization108294696> .
?uri rdfs:label ?label.
filter(langMatches(lang(?label),"EN"))
}
```

Uzlová letiště

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT distinct ?label,?uri
where {
?x dbo:hubAirport ?uri .
?uri rdfs:label ?label.
filter(langMatches(lang(?label),"EN"))
}
```

Národní parky

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?label,?uri
where {
?uri rdf:type
<http://dbpedia.org/class/yago/NationalPark108600992> .
?uri rdfs:label ?label.
filter(langMatches(lang(?label),"EN"))
}
```

Evropské silnice

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dct: <http://purl.org/dc/terms/>

SELECT ?label,?uri
where {
?uri dct:subject
<http://dbpedia.org/resource/Category:
International_E-road_network> .
?uri rdfs:label ?label.
filter(langMatches(lang(?label),"EN"))
}
```