Mendelova univerzita v Brně
Provozně ekonomická fakulta

Detekce duplicit v geoprostorových datech

Diplomová práce

Vedoucí práce:
Ing. Pavel Turčínek, Ph.D.

Bc. Adam Prchal

Brno 2024

Poděkování

Velké poděkování patří vedoucímu diplomové práce Ing. Pavlovi Turčínkovi, Ph.D. za užitečné rady, vedení a ochotu konzultovat v jakoukoliv hodinu. V neposlední řadě patří poděkování také všem, kteří se jakkoliv podíleli na zlepšení kvality této práce.

Čestné prohlášení

Prohlašuji, že jsem práci **Detekce duplicit v geoprostorových datech** vypracoval samostatně a veškeré použité prameny a informace uvádím v seznamu použité literatury. Souhlasím, aby moje práce byla zveřejněna v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách ve znění pozdějších předpisů a v souladu s platnou Směrnicí o zveřejňování závěrečných prací.

Jsem si vědom, že se na moji práci vztahuje zákon č. 121/2000 Sb., autorský zákon, a že Mendelova univerzita v Brně má právo na uzavření licenční smlouvy a užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

Dále se zavazuji, že před sepsáním licenční smlouvy o využití díla jinou osobou (subjektem) si vyžádám písemné stanovisko univerzity, že předmětná licenční smlouva není v rozporu s oprávněnými zájmy univerzity, a zavazuji se uhradit případný příspěvek na úhradu nákladů spojených se vznikem díla, a to až do jejich skutečné výše.

Brno 2024 .....

podpis

Abstract

.

.

Abstrakt

.

.

Obsah

Todo list

Seznam obrázků

1 Úvod a cíl

1.1 Úvod

Společnosti poskytující služby v oblasti Location Intelligence využívají své platformy pro komplexní geoprostorové analýzy, což je zásadní pro efektivní rozhodování v různých typech podnikání. Jako součást těchto platforem, některé společnosti, včetně CleverMaps, nabízí služby zvané jako Data Marketplace, které umožňují uživatelům získávat a integrovat různorodé datové sady. Tyto sady zahrnují např. demografické informace, obchodní statistiky a infrastrukturu měst, což pomáhá klientům lépe cílit své služby a strategická rozhodnutí. (CleverMaps, 2024)

Vzhledem k tomu, že při sběru dat a následnou manipulací s geoprostorovými daty dochází často ke kombinování dat z různých zdrojů (pro co největší úplnost konečných datových sad), vzniká problém s výskytem duplicitních záznamů. Duplicity v datech mohou mít různé podoby a vznikají z několika důvodů, některé z nich jsou:

- **Rozdílné názvy stejných míst** – např. *"Velký městský park"* vs. *"Park v centru města"*.
- **Typografické chyby a nejednotné formáty** – např. ulice *"Masarykova"* vs. *"Masarykova tř."*.
- **Rozdílné souřadnicové systémy** – jeden dataset může používat *WGS84*, zatímco jiný *S-JTSK*.

Tyto duplicity snižují kvalitu dat a mohou způsobit chyby v rozhodovacích procesech, např. při plánování dopravy, marketingových analýzách nebo při geokódování obchodních poboček. Proto je detekce a eliminace duplicit v geoprostorových datech klíčová. K identifikaci duplicit mohou být využity různé metody, od jednoduchých textových porovnání až po složitější techniky strojového učení, které analyzují podobnost dat v širším kontextu. (Nauman a Herschel, 2022; Christen, 2012)

Existují komerční i open-source nástroje pro detekci duplicit, které často využívají cloudové technologie a pokročilé algoritmy. Mezi ně patří např. Data Ladder, Tilores nebo Melissa. Ačkoliv tyto nástroje zlepšují kvalitu dat, často představují vysoké náklady nebo nejsou dostatečně přizpůsobitelné konkrétním datovým sadám. To motivuje společnosti k hledání vlastních řešení. (Christen, 2012)

1.2 Cíl

Cílem této práce je prozkoumat a otestovat různé metody detekce duplicit na geoprostorových datech, včetně metod založených na strojovém učení. Na základě analýzy výsledků testů doporučit nejvhodnější metody pro konkrétní typy sad geoprostorových dat, přičemž ověření těchto metod proběhne na datových sadách poskytnutých společností CleverMaps.

Výsledná doporučení by měla společnosti CleverMaps pomoci v rámci zvyšování automatizace a zkvalitnění procesů kontrol kvality dat.

2 Literární rešerše

2.1 Problematika detekce duplicit v datech

Data jsou v rámci různých ekosystémů ukládána v různých formách. Liší se ve struktuře, pojmenování atributů a hlavně ve způsobu využívání unikátních identifikátorů datových entit. Ve chvíli, kdy se pokoušíme data z takových různých zdrojů agregovat například do centrální databáze, vzniká problém. Tím je rozhodnout, které záznamy jsou unikátní a které se ve skutečnosti opakují a mají pouze jinou podobu kvůli odlišné struktuře dat nebo způsobu jejich získání.

Nekvalitní data jsou komplikací, která může vést k zásadním chybám v datové analýze a rozhodovacích procesech. Například ve zdravotnictví může mít existence duplicitních záznamů o pacientech vážné důsledky – od nesprávně předepsaných léků až po chybné zdravotní statistiky.(Bess, 2024) V oblasti e-commerce může duplicita zákaznických účtů znamenat špatně cílené marketingové kampaně nebo mylné vyhodnocení chování uživatelů.(Brown, 2019) A v geoprostorových datech mohou duplicity vést k nesprávné identifikaci lokací, chybným navigačním trasám nebo nesrovnalostem v mapových podkladech.

Abychom mohli přesně popsat proces detekce duplicit, je nutné nejprve vyjasnit základní pojmy související s datovými entitami a jejich reprezentací v databázích.

Entita je obecný pojem označující reálný objekt nebo koncept, který má své vlastnosti a může být reprezentován v datech. V závislosti na kontextu může entitou být např. osoba, firma, geografický bod zájmu (POI) nebo administrativní oblast.

Záznam představuje konkrétní reprezentaci entity v databázi. Jedna entita tak může být v různých databázích reprezentován více různými záznamy, například s mírně odlišnými údaji nebo formátem.

Atribut je konkrétní vlastnost nebo charakteristika entity, která ji popisuje a umožňuje její identifikaci v databázích. Každý záznam obsahuje jeden nebo více atributů, přičemž některé atributy mohou být klíčové pro rozpoznání duplicit. Atributy mohou být různých typů – například textové, číselné, kategorické nebo prostorové. V geolokačních datech bývají klíčovými atributy například souřadnice (zeměpisná šířka a délka), adresa, název místa nebo jeho typ (např. restaurace, park, obchodní centrum).

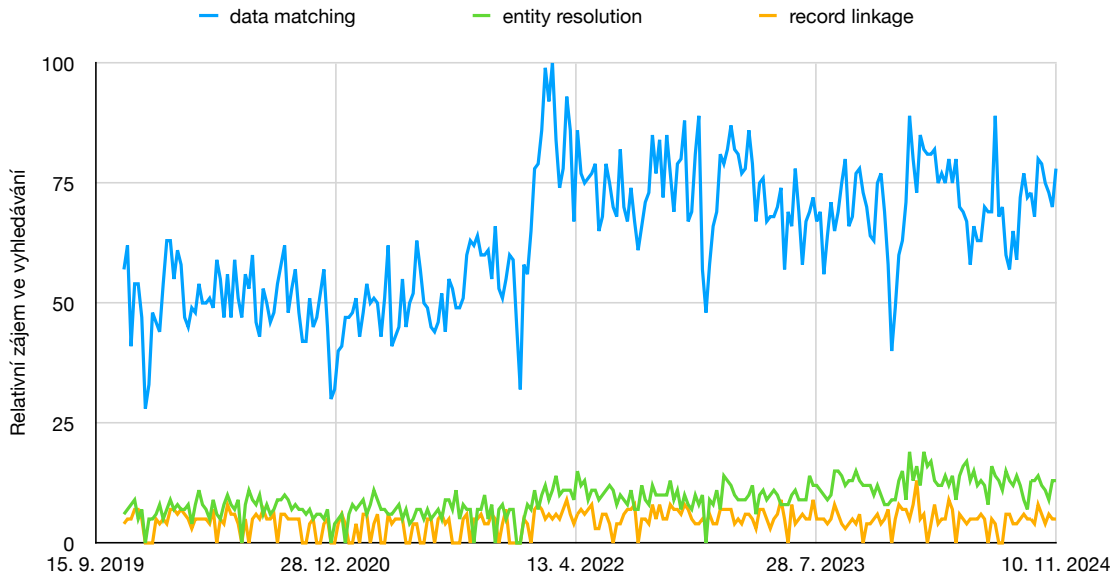
Duplicitní záznamy vznikají tehdy, když databáze obsahuje více reprezentací stejné entity, ať už kvůli chybám v zápisu, rozdílnému formátu dat, nebo rozdílným zdrojům dat. V literatuře se však termíny označující proces identifikace těchto duplicit ne vždy používají jednotně a často se jejich významy překrývají. Zatímco některé zdroje považují následující pojmy za synonyma, jiné mezi nimi rozlišují:

- **Data matching** – proces porovnávání záznamů za účelem nalezení těch, které odpovídají stejné entitě, i když mají různé atributy nebo formáty. (Christen, 2012)

- **Entity resolution** – proces slučování duplicitních záznamů do jednoho sjednoceného záznamu reprezentujícího danou entitu. (Quantexa, 2024)
- **Record linkage** – propojení odpovídajících záznamů napříč různými databázemi, aniž by došlo ke sloučení do jedné reprezentace. (Stepanenko, 2024)

To, že existují různé termíny pro podobné procesy, ukazuje, že problematika detekce duplicit vznikala nezávisle v různých oborech, a proto je důležitější se soustředit na samotné postupy, než na přesné označení. (Christen, 2012)

V této práci bude hlavní pozornost věnována *data matching*, tedy samotnému procesu hledání duplicitních záznamů v datech, spíše než jejich následnému slučování (*entity resolution*). Pro zajímavost, podle Google Trends je termín *data matching* v posledních pěti letech globálně vyhledávanější než ostatní zmiňované termíny. Viz obr. 2.



Obrázek 1: Zájem o termíny ”data matching”, ”entity resolution” a ”record linkage”. *Zdroj*: (Google Trends, 2024)

2.2 Příčiny vzniku duplicit

Zazněli zde již některé důvody vzniku duplicit v datech, pojdmě se blíže podívat na nejčastější z nich. Rozdělil jsem je zde na duplicity v datech, které mohou vzniknout v rámci jednoho datového zdroje, a duplicity vznikající spojením více datových zdrojů.

Duplicity v samostatném zdroji

Přestože by se na první pohled zdálo, že v rámci jednoho datového zdroje šance na vznik duplicity minimiální, je zde i tak několik případů, které vedou ke vzniku duplicit. Tady jsou některé z nich:

- Lidský faktor – při zadávání dat člověkem ručně je velká pravděpodobnost zadání typografických chyb, různých nekonzistentních formátů a nebo přímo neplatných informací. Při zadávání dalších záznamů si pak daný člověk nemusí všimnout již existujícího záznamu vloženého dříve.
- Nedostatečné standardizace formátů – různé formuláře, dotazníky, nebo části/-moduly jednoho systému mohou využívat různé formáty pro datumy, telefonní čísla, nebo rodná čísla a dalších typů dat. Pokud není správně ošetřen převod těchto dat do jednotného formátu před vložením záznamu, může opět docházet ke vzniku duplicit.
- Chybějící integritní omezení databáze – v případě, že daná databáze nemá nastavená integritní omezení (např. na unikátní hodnoty v rámci rodného čísla, IČO nebo třeba unikátnosti dvou a více atributů v rámci jednoho záznamu), bude každý vložený záznam brán jako nový, doposud neexistující, záznam, přestože se může jednat o identickou duplicitu.

Duplicity ve sloučeném zdroji

Při pokusu o spojení více zdrojů do jednoho vzniká několik výzev. V rámci business inteligence se tento proces spojování různých datových zdrojů nazývá ETL[1]a v rámci něj může docházet k duplicitám v několika případech:

- Integrace různých systémů – v případě systémů, které mají data o entitách kvůli různým důvodům (např. inventární systém školy ukládá záznamy o vybavení, online bazar má záznamy o prodávaném vybavení), může vzniknout situace, kdy záznamy z více systémů nelze spojit na základě společného unikátního klíče, protože má každý systém svůj unikátní pro danou entitu.
- Průběžné spojování dat – některé datasety jsou pravidelně aktualizované, aby co nejpřesněji odráželi realitu. Pokud však entita z realného světa v rámci aktuality změnila své atributy zásadně (např. pekárna se přestěhovala na jinou adresu

[1]Extract, transform, load

a upravila své jméno), může vzniknout situace, kdy některý ze zdrojů bude mít stále staré údaje, a jiný nové, a tím pádem nepůjde tento rozdíl spojit s původním záznamem a bude se milně jevit, že jde o novou entitu.

- Rozdílné definice entit a identifikační metody – při integraci dat z různých zdrojů se může stát, že jednotlivé systémy definují a identifikují stejné entity odlišně. Jeden systém může například využívat kombinaci jména, adresy a rodného čísla k jednoznačné identifikaci osoby, zatímco jiný systém pracuje s interním identifikačním číslem nebo používá jiné atributy (např. e-mailovou adresu). Tato nejednotnost vede k tomu, že algoritmy pro slučování dat nemusí správně rozpoznat, že se jedná o stejnou entitu, a následně dochází k vytvoření duplicitních záznamů v cílovém datovém skladu.

Toto zajisté nejsou všechny možné příčiny vzniku duplicit v datech, měli by však nastínit, jak se jednotlivé příčiny mohou lišit, a proč se jedná o problém, který nelze vyřešit univerzálním způsobem.

2.3 Důsledky duplicitních dat

Duplicitní data mohou mít na organizaci zásadní a mnohostranné dopady. Především se výrazně snižuje kvalita dat, informace se stávají nepřesnými, neúplnými a méně aktuálními. To ovlivňuje nejen tvorbu reportů, ale i analytické procesy, které jsou základem strategických rozhodnutí. Rozhodnutí vycházející z chybně interpretovaných dat mohou vést k nesprávným závěrům, což se může projevit například ve špatném plánování, neefektivním řízení zdrojů nebo dokonce v riziku finančních ztrát.

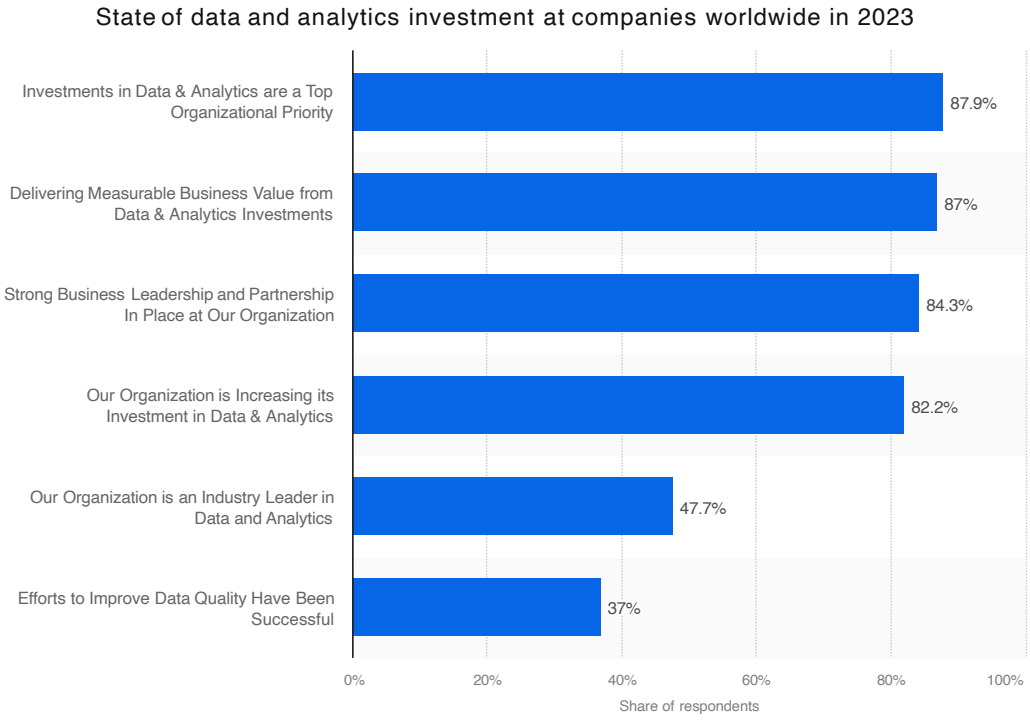
Dalším významným důsledkem je snížení efektivity pracovníků. V situaci, kdy je třeba trávit spoustu času identifikací a odstraňováním duplicit, se zaměstnanci odklánějí od jiných úkolů, které by mohly přinést přidanou hodnotu. Kromě rutinní práce, jako je kontrola a aktualizace záznamů, je také nutné dohledávat různé verze stejných dat, což výrazně komplikuje správu informací. Takový stav nejenže zvyšuje provozní náklady, ale také vede k vyšší chybovosti a celkovému zhoršení kvality práce.

Nesmíme opomenout ani problémy spojené s dodržováním regulatorních požadavků. Regulace často vyžadují přesné, konzistentní a bezpečné uchovávání dat. Duplicitní záznamy komplikují tvorbu přesných reportů, což může mít za následek sankce, pokuty nebo dokonce ztrátu důvěry ze strany regulačních orgánů, spolupracujících firem a zákazníků. Správa citlivých údajů, jako jsou osobní nebo finanční informace, se také stává náročnější, což představuje další riziko v případě případných narušení bezpečnosti nebo při nutnosti vyhovět zákonným požadavkům na přístup a správu těchto dat.

Duplicitní data také negativně ovlivňují i zákaznický servis a správu IT infrastruktury. Když jsou informace o zákaznících roztříštěny mezi několika duplicitními záznamy, je pro pracovníky obtížné získat ucelený přehled o jejich historii, což

vede ke zhoršení kvality poskytovaných služeb. Současně se tím zvyšuje náročnost správy datových logů, zálohovacích a archivních procesů.

Celkově lze říct, že neřešené duplicitní záznamy mají rozsáhlý negativní vliv na efektivitu, bezpečnost a spolehlivost informací v organizaci. Firmy se snaží aktivně problém s kvalitou dat řešit, a to ukazuje i průzkum "Big Data and AI Executive Survey 2024" od NewVantage Partners za rok 2023. Průzkumu se učasnilo více než 100 firem z globálního Fortune 1000, a ukazuje, že pro většinu organizací (87,9 %) má investice do dat a analytiky vysokou prioritu. Ze stejného výzkumu však i vychází informace, že snaha o zlepšení kvality dat byla úspěšná pouze u 37 % dotázaných firem.



Obrázek 2: Výsledky výzkumu o investicích do dat a analytiky.
Zdroj: (NewVantage Partners, 2024)

2.4 Metody pro rozpoznání duplicitních entit v datech

K řešení detekce duplicit je třeba využít kombinaci několika metod, které se vzájemně doplňují a umožňují zvýšit celkovou přesnost a robustnost systému. Následující podkapitola rozebírá jednotlivé skupiny přístupů, jejich fungování, výhody i nevýhody, a možnosti jejich vzájemného propojení.

Obecné metody detekce duplicitních entit

Pravidlové metody

Jedná se o principielně jednoduchou skupinu metod. Jejich princip spočívá v tom, že se vytváří jednoznačná pravidla nebo sada pravidel, které podle daných kritérií určují duplicitní skupiny záznamů. Taková pravidla bývají navržena experty v dané oblasti, kteří vychází ze znalostí a charakteristik konkrétních dat.

Pravidla mají zpravidla jasně definované atributy, nebo skupinu atributů, které porovnávají a to buď striktně (hodnoty atributů musí být identické), kdy se jedná o tzv. exact matching, nebo volněji s použitím podobnostních metrik (Levenshteinova vzdálenost, Jaro-Winkler podobnost, nebo Jaccardův koeficient) či prahových hodnot. Lze však kombinovat striktní a volnější porovnání v rámci složeného pravidla. To může slovně být popsáno třeba takto: *„Dva záznamy jsou duplicity, pokud se shodují v atributu jméno a příjmení alespoň na 90 % (dle Levenshteinovy vzdálenosti) a zároveň mají shodné datum narození.“*.

K volnějšímu porovnání lze úvest i fonetické metody, které se používají primárně u textových údajů, které zní velice podobně, ale zapisují se různě. Jedná se například o Soundex, Metaphone a Double Metaphone, které z textového vstupu generují fonetické kódy, které lze následně porovnávat, a to buď striktně, nebo s využitím podobnostních metrik.

Výhodou pravidlových metod je snadná implementace a jasná interpretace výsledků. Nicméně vytvořená pravidla jsou jednak velice málo flexibilní, jelikož jsou psaná vždy specificky pro konkrétní typ dat, nebo pro konkrétní odvětví. Pro jednoduché scénáře mohou být dostačující, nicméně ve složitějších situacích mohou sloužit spíš jako vstupní krok pro následné pokročilejší metody, které následují.

Metody založené na podobnosti

Definice podobnosti, používané metriky (textové, numerické, atd.)

Praktické ukázky (Levenshteinova vzdálenost, Jaccard, kosinová podobnost)

Výhody, nevýhody a omezení metody

Pravděpodobnostní metody

Definice a princip Fellegi-Sunterova modelu a jeho variant

Možnosti aplikace bayesovských metod

Výhody, nevýhody a limitace při použití

Metody strojového učení

- Definice a typy (supervizované, semi-supervizované, nesupervizované)
- Typické algoritmy: rozhodovací stromy, Random Forest, Gradient Boosting
- Klíčové aspekty přípravy dat a výběr atributů
- Výhody, nevýhody

Metody založené na hlubokém učení

- Princip neuronových sítí pro entity matching (např. embeddings, transformer modely)
- Scénáře nasazení a případové studie z literatury
- Omezení a nevýhody (black-box přístup, náročnost na data a výpočetní zdroje)

Grafové metody

- Grafové přístupy k modelování entit
- Komunitní detekce, grafové neuronové sítě (GNN)
- Specifika aplikace, výhody, nevýhody

Specifické metody pro rozpoznání duplicit v geolokačních datech

- Specifika geolokačních dat
- Prostorová přesnost a chyby měření
- Problematika výběru vhodných vzdálenostních metrik a projekcí

Metody založené na vzdálenosti

- Popis a využití Haversinovy a Vincentyho vzdálenosti
- Aplikace vzdálenostních metod, výběr prahů

Shlukovací (clusteringové) metody

- DBSCAN, HDBSCAN a OPTICS
- Typické scénáře použití v praxi
- Výhody a omezení metod

Metody založené na prostorové mřížce (grid-based)

- Principy Geohashe a prostorových mřížek
- Scénáře využití a důsledky výběru velikosti mřížky
- Omezení metody a problémy s přesností

Pravděpodobnostní a bayesovské metody

- Modelování prostorové nepřesnosti (GPS chyby)
- Příklady bayesovských přístupů ke geolokaci
- Výhody a limitace těchto metod

Metody strojového učení a hlubokého učení v geolokaci

- Stručně zmínit možnost aplikace ML přístupů, pokud to téma umožňuje
- Zvýraznit případové studie (pokud existují relevantní)

Geokódování a porovnávání adres

- Standardizace adresních údajů
- Geokódovací nástroje a postupy

Literatura

Bess, Oleg (12. led. 2024). *The problem with duplicate and mismatched patient records*. Physicians Practice. URL: https://www.physicianspractice.com/view/the-problem-with-duplicate-and-mismatched-patient-records (cit. 09. 02. 2025).

Brown, Forrest (13. dub. 2019). *8 Problems That Result from Data Duplication • Profisee*. Enterprise Master Data Management • Profisee. URL: https://profisee.com/blog/8-business-process-problems-that-result-from-data-duplication/ (cit. 09. 02. 2025).

CleverMaps (2024). *Location Insights for Business*. CleverMaps. URL: https://www.clevermaps.io (cit. 28. 04. 2024).

Google Trends (2024). *Google Trends*. Google Trends. URL: https://trends.google.com/trends/explore?date=today%205-y&q=data%20matching,entity%20resolution,record%20linkage&hl=cs (cit. 14. 11. 2024).

Christen, Peter (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Data-centric systems and applications. Berlin Heidelberg: Springer. 270 s. ISBN: 978-3-642-43001-5 978-3-642--31163-5.

Nauman, Felix a Melanie Herschel (2022). *An Introduction to Duplicate Detection*. Google-Books-ID: DYdyEAAAQBAJ. Springer Nature. 84 s. ISBN: 978-3-031--01835-0.

NewVantage Partners (2024). *Global state of data/analytics investment 2023*. Statista. URL: https://www.statista.com/statistics/1453262/global-state-of-data-analytics-investment/ (cit. 12. 02. 2025).

Quantexa (2024). *What is Data Matching & How Does it Work?* Quantexa. URL: https://www.quantexa.com/resources/data-matching/ (cit. 13. 11. 2024).

Stepanenko, Roman (21. ún. 2024). *What is Entity Resolution*. RecordLinker. URL: https://recordlinker.com/what-is-entity-resolution/ (cit. 14. 11. 2024).