

DATOVÁ ANALÝZA UNIKLÝCH DOKUMENTŮ V OBLASTI INVESTIGATIVNÍ ŽURNALISTIKY



Autor: Ing. Mai Phuong Bui

Vedoucí práce: PhDr. Jan Černý, Ph.D.

Fakulta informatiky a statistiky, Vysoká škola ekonomická v Praze

MOTIVACE A CÍL

- Únik dat o offshore firmách (Pandora Papers, Paradise Papers, Bahamas Leaks, Panama Papers a další)
- Zpracování a zveřejnění Mezinárodním konsorciem investigativních novinářů (ICIJ) v podobě databáze.
- Vyhledávání entit pomocí vyhledávacího pole na webu.
- Výsledky zobrazeny jako seznam odkazů, každý výsledek se otevře v novém okně.
- Uživatelé mohou mít problémy s přehledností a zmatením při velkém množství výsledků.**

Addresses

323

Obrázek 1. Počet adres v ICIJ databázi pro klíčové slovo "Czech Republic" je 323, pro zobrazení detailu je nutné otevřít každou adresu v novém okně

Cíl

- Vytvořit interaktivní geolokační databázi s možností filtrování a zobrazení informací o firmách na mapě.** Mapu bude doplňovat tabulka s detaily o jednotlivých entitách, a to vše v jednom okně.
- Databáze tak nabídne uživatelům alternativní pohled na data, než ta od ICIJ.

DATOVÝ ZDROJ

- ICIJ zveřejnilo soubory se surovými daty, ale struktura tabulek nevyhovovala potřebám práce - bylo potřeba vytvořit jednotný zdroj.
- Byla provedena úprava a čištění dat pro zobrazení geolokačních informací pomocí Pythonu.
- Problémy s chybějícími nebo nesprávnými hodnotami u geolokačních sloupců a adresou.

SUROVÁ DATA SE SKLÁDALA Z

5 různých uniklých dokumentů
6 csv souborů
3M řádků reprezentující vztahy mezi subjekty

FINÁLNÍ DATABÁZE

Entity Name	Company Type	Related Officer	Relationship to Entity	Address	Active From	To	Linked Country	Country Code	Registered In	Incorporation Date	Struck Off Date	Inactivation Date	Dorm Date	Data From	Valid Until	Linked Country
ANTI-GRAVITY HOLDINGS	Business	GAVIN MULL	intermediary	Null	Null	Null	Hong Kong	HKG	British Virg.	1/12/1999	Null	Null	Null	Panama Pa.	The Pan.	All
ANTI-GRAVITY HOLDINGS	Business	WONG KIN HAY FELIX	beneficiary	Null	Null	Null	Hong Kong	HKG	British Virg.	1/12/1999	Null	Null	Null	Panama Pa.	The Pan.	All
ANTI-GRAVITY INC	Business	HAY FELIX	shareholder	Null	Null	Null	Hong Kong	HKG	British Virg.	1/12/1999	Null	Null	Null	Offshore Le.	The Off.	All
ANTI-GRAVITY INC	Business	LAM LAP KO	director of	Null	6/28/2006	Null	Not identifi.	XXX	Undetermin.	6/28/2006	Null	Null	Null	Offshore Le.	The Off.	All
ANTI-GRAVITY INC	Business	SINOMIX B	intermediary	Null	Null	Null	Hong Kong	HKG	Undetermin.	6/28/2006	Null	Null	Null	Offshore Le.	The Off.	All
AZURE	Business	Address	registered	Z206-19 Jar	Null	Null	Hong Kong	HKG	Hong Kong	8/28/2014	Null	Null	Null	Paradise Pa.	Appleb.	All

Obrázek 2. Odkaz na finální databázi

Výsledná databáze je **veřejná a volně k použití na platformě Tableau Public** a lze vyhledat pod názvem ICIJ Offshore Leaks Geolocation Database. Díky přehlednému vizuálnímu zpracování získají nejen investigativní novináři, ale i další uživatelé, kteří nejsou obvykle zvyklí pracovat s velkým množstvím dat, přehledně zobrazené

zajímavé informace z uniklých dokumentů o offshore firmách v takové podobě, kterou mohou lépe zpracovat a interpretovat.

V případě potřeby lze zobrazené informace stáhnout na lokální úložiště, kde s nimi může uživatel dále pracovat díky možnosti zobrazení detailních informací o jednotlivých subjektech v tabulce. **Tato tabulka existuje také pro zobrazení entit, ke kterým není přiřazená žádná adresa či země,** jelikož se nepodařilo data vyčistit kompletně a stále jsou tu chybějící hodnoty, které však mohou uživateli přinést užitek v podobě dalších dostupných informací k danému subjektu. Vzhledem k povaze dat (uniklé dokumenty) je možnost chybějících dat logická.

PŘÍNOS PRÁCE

V současné době existuje na **veřejné platformě Tableau Public interaktivní geolokační databáze zobrazující síť offshore firem** z pěti velkých souborů uniklých dokumentů poskytnutých ICIJ, která slouží jako podpora procesů Open Source Intelligence (OSINT) nejen pro investigativní žurnalisty. Jeden z hlavních přínosů práce spočívá v datovém zdroji, který vznikl během procesu přípravy dat, aby mohla databáze vzniknout. Vlivem úpravy a čištění dat obsahuje výsledný datový zdroj v mnoho případech **detailnější informace z hlediska geolokace než oficiální databáze ICIJ.**

Existence této práce a databáze na veřejné datové platformě, jako je Tableau Public může pomoci zvýšit povědomí veřejnosti o práci ICIJ a investigativních novinářů. Bez nich a podpory veřejnosti by tato data neexistovala.