Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky

Datová analýza uniklých dokumentů v oblasti investigativní žurnalistiky

DIPLOMOVÁ PRÁCE

Studijní program: Informační systémy a technologie

Specializace: Business Intelligence

Autor: Bc. Mai Phuong Bui

Vedoucí diplomové práce: PhDr. Jan Černý, Ph.D.

Praha, duben 2023

Prohlášení

Prohlašuji, že jsem diplomovou práci Datová analýza uniklých dokumentů v oblasti investigativní žurnalistiky vypracovala samostatně s použitím pramenů a literatury, které jsou v práci uvedeny.

V Praze dne 25. dubna 2023 .....

Mai Phuong Bui

Poděkování

Ráda bych poděkovala PhDr. Janu Černému, Ph.D., za odborné vedení mé diplomové práce, za ochotu, jeho čas, cenné rady a podporu při jejím zpracování.

Dále bych chtěla poděkovat paní Pavle Holcové za poskytnutí času pro náš rozhovor.

Abstrakt

Tato diplomová práce se zabývá analýzou a tvorbou interaktivní databáze zobrazující síť offshore firem z pěti velkých souborů uniklých dokumentů, které byly zpracované a poskytnuté veřejnosti Mezinárodním konsorciem investigativních žurnalistů (ICIJ) jako podpora procesů Open Source Intelligence (OSINT). ICIJ z těchto dat vytvořili na svém webu vlastní databázi, vyhledávání informací o jednotlivých firmách zde však probíhá formou vyhledávacího pole a výsledky se zobrazí jako seznam odkazů odpovídající vyhledávání, což může v tak velkém množství dat být nepřehledné a uživateli tak můžou některé informace uniknout. Cílem práce je vytvořit interaktivní databázi, která bude pomocí filtrů na mapě a doplňující tabulce zobrazovat všechny relevantní informace o jednotlivých firmách, a to vše v jednom okně.

Práce je rozdělena na teoretickou a praktickou část. V teoretické části jsou popsány základní principy investigativní žurnalistiky, metodiky sběru, klasifikace, verifikace a ochrany dat. V praktické části je pak popsán proces přípravy a čištění dat, který je nutný pro následnou tvorbu interaktivní databáze s využitím vizualizačního nástroje Tableau a s cílem vytvořit uživatelsky přívětivý dashboard, který bude veřejně dostupný na portálu Tableau Public.

Klíčová slova

datová analýza, investigativní žurnalistika, data, ICIJ, offshore firmy, offshore úniky, Pandora Papers, Panama Papers, Bahamas Leaks, Paradise Papers, Tableau, Open Source Intelligence, OSINT

Abstract

This thesis concerns the analysis and creation of an interactive database showcasing a network of offshore companies based on five extensive sets of leaked documents that have been processed and made publicly available by the International Consortium of Investigative Journalists (ICIJ) in support of Open Source Intelligence (OSINT) processes. While ICIJ has developed its own database using this data, the search for information about individual companies is carried out through a search box and the results are displayed as a list of links matching the search. This can be confusing when dealing with such a large amount of data, leading users to overlook important information. The aim of this project is to create an interactive database that will display all relevant information about individual companies in one window using filters on a map and a supplementary table.

The project is divided into two parts: theoretical and practical. The theoretical part outlines the fundamental principles of investigative journalism, along with methodologies for data collection, classification, verification, and protection. The practical part describes the process of data preparation and cleaning, which is essential for the subsequent creation of an interactive database using a visualization tool called Tableau. The goal is to create a user-friendly dashboard that will be publicly available on the Tableau Public portal.

Keywords

data analysis, investigative journalism, data, ICIJ, offshore companies, offshore leaks, Pandora Papers, Panama Papers, Bahamas Leaks, Paradise Papers, Tableau, Open Source Intelligence, OSINT

Obsah

Seznam obrázků

Seznam výpisů programového kódu

Úvod

Během studia mě začala bavit práce s daty a rozhodla jsem se svoji kariéru směřovat tímto směrem. Po absolvování předmětu Open Source Intelligence se můj zájem rozšířil také o taková data, která jsou veřejně dostupná, konkrétně mě fascinovali lidé, kteří pomocí veřejně dostupných informací a vyhledávání odhalovali zločiny a skutečnosti, které unikaly autoritám. Jednou ze skupin lidí, kteří pomáhají odhalovat zločiny a poukazují na korupci, jsou investigativní novináři. Jejich práce je pro mě velmi zajímavá a psaním této diplomové práce bych se chtěla více dozvědět o té jejich a současně veřejnosti nabídnout pomocí svých nabytých dovedností za studia užitečný a přehledný nástroj pro vizualizaci uniklých dokumentů zpracovaných právě investigativními žurnalisty.

Předmětem této diplomové práce je tvorba interaktivní databáze zobrazující síť offshore firem z Pandora Papers, Paradise Papers, Bahamas Leaks, Panama Papers a Offshore Leaks poskytnutých Mezinárodním konsorciem investigativních novinářů (angl. International Consortium of Investigative Journalists, dále jen ICIJ). Souhrn těchto souborů tvoří přes 810 tisíc záznamů. Vyhledávání informací o jednotlivých entitách v současné době probíhá formou vyhledávacího pole prostřednictvím ICIJ webu a výsledky se zobrazí jako seznam odkazů odpovídající vyhledávání, což může v tak velkém množství dat být nepřehledné a uživateli tak můžou některé informace ujít. Mým cílem je proto vytvořit interaktivní databázi, ve které si může uživatel pomocí filtrů a zobrazením všech offshore firem na mapě zvolením konkrétní entity. Tato databáze pomůže nejen investigativním novinářům pracovat efektivněji a umožní jim jiný pohled na data, ale zároveň přehledně předloží zajímavé informace z uniklých dokumentů veřejnosti ve vizuální podobě, kterou může lépe zpracovat a interpretovat.

Práce je rozdělena na dvě části – teoretickou a praktickou. Teoretická část je popsána ze dvou hledisek – z hlediska oboru, kde se nejprve zabývám samotným pojmem investigativní žurnalistika a jeho hlavními znaky, a poté z hlediska dat a práce novinářů s nimi. Představím nejznámější osobnosti v oblasti investigativní žurnalistiky a popíšu nejznámější případy uniklých dokumentů, ke kterým došlo v minulosti. Poté se zaměřím na problematiku z hlediska dat počínaje metodikou sběru dat, verifikace, klasifikace, uchovávání až po ochranu dat. Kontaktovala jsem také ředitelku Českého centra pro investigativní žurnalistiku Pavlu Holcovou se žádostí o rozhovor, abych získala větší vhled do problematiky. Rozhovor proběhl sice pouze písemně, ale získala jsem cenné informace z praxe, což dodává mé práci přidanou hodnotu.

Praktická část se zabývá samotnou tvorbou interaktivní databáze. Datové zdroje, které slouží jako podklady pro vznik databáze jsou, jak již bylo zmíněno, z webu ICIJ. Databáze byla vytvořena pomocí vizualizací prostřednictvím Tableau a v podobě dashboardu či několika dashboardů zveřejněná na portálu Tableau Public, díky čemuž bude veřejně dostupná pro všechny. Nejprve popíšu jednotlivé datasety a jejich návaznost na sebe. Dále se zaměřím na data jako taková, kde provedu předzpracování dat, jelikož jejich současná podoba nemusí odpovídat mým potřebám pro další zpracování, a nakonec data propojím

v Tableau a provedu samotnou tvorbu databáze. Na závěr hotový dashboard publikuji na Tableau Public.

1 Investigativní žurnalistika

Investigativní žurnalistika začala původem ve Spojených státech amerických. Zahrnuje rozsáhlé aktivity rešerše, sběru informací, důkazů, které je nutné validovat a ověřovat.

Existují různé definice pojmu investigativní žurnalistika. Investigative Reporters and Editors (IRE), světová nezisková asociace investigativní žurnalistiky ji definuje jako *systematický, hloubkový a originální výzkum a reporting, často zahrnující odhalování tajemství, intenzivní využívání veřejných záznamů a počítačové zpravodajství, se zaměřením na sociální spravedlnost a odpovědnost.* [1] [1]

Dle Global Investigative Journalism Network (zkratkou GIJN) se jedná o *systematický, hloubkový a primární výzkum a reportáže, často zahrnující odhalování tajemství.* [2] [2]

Příručka investigativní žurnalistiky, kterou vydalo UNESCO, definuje investigativní žurnalistiku takto:

Investigativní žurnalistika spočívá v tom, že veřejnosti odhaluje věci, které jsou skryty – buď záměrně někým v mocenském postavení, nebo náhodně, za chaotickou masou faktů a okolností, které zastírají porozumění. Vyžaduje použití tajných i otevřených zdrojů a dokumentů. [3] [3]

Profesor žurnalistiky na Missourské univerzitě Steve Weinberg definoval investigativní žurnalistiku jako *zpravodajství, které z vlastní iniciativy a vlastním pracovním výkonem podává zprávy o věcech důležitých pro čtenáře, diváky nebo posluchače.* [4] [4]

Ač je každá definice popsaná trochu jinak, všechny se shodují v tom, že se jedná o druh žurnalistiky, který zahrnuje odhalování důležitých věcí, jež jsou skryté veřejnosti. Někdy je také profese přirovnávána k policii či jinému aparátu, který bojuje proti nelegálním činnostem. Dle Jaroslava Kmenty, českého investigativního žurnalisty, musí každý, kdo chce tuto práci vykonávat, umět jednat s lidmi, aby od nich získal kvalitní a relevantní

---

[1] Původní znění: *systematic, in-depth, and original research and reporting, often involving the unearthing of secrets, heavy use of public records, and computer assisted reporting, with a focus on social justice and accountability*

[2] Původní znění: *systematic, in-depth, and original research and reporting, often involving the unearthing of secrets*

[3] Původní znění: *Investigative journalism involves exposing to the public matters that are concealed–either deliberately by someone in a position of power, or accidentally, behind a chaotic mass of facts and circumstances that obscure understanding. It requires using both secret and open sources and documents.*

[4] Původní znění: *Reporting, through one's own initiative and work product, matters of importance to readers, viewers, or listeners.*

informace. Dále pak musí mít analytické myšlení, aby byl schopen uvést zjištěné informace do souvislostí, a mít všeobecný přehled, zejména z hlediska znalosti zákonů. V neposlední řadě musí být investigativní žurnalista odolný vůči stresu, neboť se často jedná o poměrně nepříjemnou a nebezpečnou práci. [5]

Novinář při své práci využívá jak otevřené, tak neveřejné zdroje. Podmnožinou investigativní žurnalistiky je datová žurnalistika. Novinář musí umět využívat otevřené zdroje (tzv. open source), což jsou například databáze, rejstříky a seznamy, a zároveň najít a umět využít co nejlépe neveřejné zdroje. Pavla Holcová, zakladatelka webu investigace.cz [5] a ředitelka Českého centra pro investigativní žurnalistiku, považuje otevřené zdroje ve své práci za klíčové. Ve chvíli, kdy případ, kterým se investigativní žurnalisté zabývají, skončí u soudního řízení, musí být důkazy založené na informacích z otevřených zdrojů, aby mohly být použité. [6]

1.1 Investigativní žurnalistika v ČR a SR

První skutečná investigativní reportáž v České republice vznikla v roce 1989 v Mladém světě. Reportáž se týkala Ochranného svazku autorského, který rozděloval honoráře na ideologických základech. V roce 1993 Michal Růžička do Lidových novin napsal reportáž, ve které odhalil nezákonný prodej dat o matkách s dětmi společnosti Procter and Gamble, která pak na získané adresy posílala reklamu. V důsledku reportáže dostal kancléř ministerstva vnitra výpověď a začalo se více dbát na ochranu dat. Od té doby se různá média pokoušejí o investigativní žurnalistiku, byť s různými výsledky. [7]

Za jeden z největších úspěchů české investigativní žurnalistiky můžeme považovat pořad Na vlastní oči Josefa Klímy a Radka Johna. V ohledu Slovenska je nutné zmínit jméno Jána Kuciaka.

1.1.1 Josef Klíma

Josef Klíma je jedním ze zakladatelů české investigativní žurnalistiky, spoluzakladatel časopisu Reflex, moderátor pořadů Na vlastní oči, Očima Josefa Klímy a autor zhruba 40 knih inspirovanými skutečnými příběhy.

Zpočátku vydal několik knih založených na investigativní žurnalistice, ale musel je překlopit do fikce, jelikož kvůli tehdejšímu komunistickému režimu nebylo možné psát tento druh novinařiny. Po roce 1989 vznikl časopis Reflex, do kterého napsal velkou reportáž o červených baretech – příslušnících tajné služby ministerstva vnitra. Postupně se na Reflex začali obracet občané se svými tragédiemi, a tak vznikla v ČR investigativní žurnalistika.

[5] https://www.investigace.cz

Témat ke psaní bylo mnoho a z časopisu se Klímovy příběhy překlopily do pořadu Na vlastní oči od TV Nova a Soukromá drama od TV Prima. [8]

V roce 2018 navázal spolupráci se společností Seznam, kde se svým týmem točí investigativní pořad s názvem Záhady Josefa Klímy. [9]

Mezi největší kauzy, kterým se věnoval, patří například kauza Jiřího Kájínka, o které napsal také 2 knihy s Jankem Kroupou. Kauza se týkala zločince odsouzeného za dvojnásobnou vraždu, který utekl z mírovské věznice. V knize *Pravda o Kájínkovi* přináší různé důkazy a výpovědi svědků, které za dobu vyšetřování nasbírali. Celkem natočili o kauze s Kroupou téměř 20 reportáží. [10]

V další kauze se zabývá krádeží novorozenců za doby totality. S příběhem k němu přišla paní Jana Ulrychová, které v roce 1970 nemocnice 2 dny po porodu oznámila, že její miminko zemřelo a nikdy jí tělo neukázali se slovy, že vše potřebné zařídí. O 46 let později se náhodou dozvěděla, že nemocnice pohřby nikdy nezařizovaly, a začala pátrat po své dceři. Po odvysílání reportáže se Klímovi ozvalo téměř 30 dalších žen s podobným příběhem. [11]

Dnes se věnuje spíše příběhům běžných lidí, které mafie připravila o majetek a rozpadla se jim rodina, a které stát nechal napospas osudu bez pomoci, přestože se proti ničemu neprovinili. [12]

1.1.2 Janek Kroupa

Janek Kroupa je další významný investigativní žurnalista, který je znám například svými reportážemi o kauzách Berdychova gangu, manipulaci veřejných zakázek, na které se podílela čínská firma Huawei či korupci týkající se vládního nákupu obrněných transportérů Pandur II. Stejně jako Josef Klíma byl moderátorem pořadu Na vlastní oči, dále pak také vedl investigativní tým Českého rozhlasu. [13] Je autorem několika knih a scénářů k detektivním seriálům – spolu se Zdeňkem Čechem napsal v roce 2006 knihu Zločin jako profese. S Josefem Klímou pak napsal knihu Jiří Kájínek: Vrah, nebo oběť a spolupracovali na seriálu Expozitura, který je inspirovaný děním kolem právě Berdychova gangu, jenž vyšetřoval. [14]

Jeho již zmíněná nejvýznamnější kauza Berdychův gang odhalila korupci z řad policie, která řídila zločinecké organizace provozující kriminální činnosti po celé republice a zároveň je kryli, odváděli pozornost a sabotovali vyšetřování. [15]

Společně se svými polskými kolegy Bertoldem Kittelem a Jaroslawem Jabrzykem napsal reportáž o pašování zbraní do proruské části Ukrajiny, za kterou získal polskou novinářskou cenu Grand Press. V současné době pracuje jako reportér pro Seznam Zprávy. [13]

1.1.3 Jaroslav Kmenta

Jaroslav Kmenta je český investigativní žurnalista, který se věnuje především politickým kauzám, světu mafie a špionáži. V současné době působí v časopisu Reportér. [16] Svou kariéru započal v Mladé frontě Dnes, kde vydal svoji první kauzu Pouštní horečka týkající se nemocných vojáků, kteří v roce 1990 prošli válkou v Perském zálivu.

V roce 2005 založil vlastní nakladatelství a od té doby vydal celkem 15 knih, přičemž některé byly zfilmované, a několik audioknih zabývající se například zločincem, jako je Radovan Krejčíř, který utekl z ČR, aby se vyhnul vězení za miliardové podvody, ale také třeba politickou kariérou Andreje Babiše. [17]

Během své kariéry se podílel na řadě dalších významných kauz. V roce 1997 odhalil tajný účet ODS ve Švýcarsku, což přispělo k pádu Klausovy vlády. V roce 2005 se zabýval majetkovými nesrovnalostmi tehdejšího premiéra Stanislava Grosse, který své finance nedokázal obhájit a v důsledku rezignoval ze své pozice. [16]

1.1.4 Pavla Holcová

Pavla Holcová je zakladatelka a ředitelka Českého centra pro investigativní žurnalistiku a současně zakladatelka a šéfredaktorka webu investigace.cz. Jako jediná novinářka v ČR byla přizvána do investigace Panama Papers od Mezinárodního konsorcia investigativních žurnalistů (ICIJ). Dále pracuje také jako editorka pro mezinárodní síť investigativních novinářů Organized Crime and Corruption Reporting Project (zkratkou OCCRP). [18] Podílela se na kauze o nezákonných obchodních aktivitách ázerbájdžánského prezidenta a jeho rodiny, za kterou získala cenu Global Shining Light Award. Dále se také podílela na mapování několika miliardové dodávky zbraní pro konflikt v Sýrii. [19]

Holcová začínala v organizaci Člověk v tísni, kde vedla programy na podporu žurnalistů na Kubě. Podílela se na školení kubánských novinářů, učila je rozlišovat názorovou žurnalistiku a žurnalistiku založenou na faktech. K investigativní žurnalistice se dostala přes investigativního novináře Paula Radu, který ji zasvětil do své práce v době, kdy byli zavření ve vězení na Kubě během školení. Jeho příběhy o mezinárodních novinářských projektech ji nadchly a rozhodla se v roce 2013 založit České centrum pro investigativní žurnalistiku. [6]

Je držitelkou mnoha významných ocenění, příkladem je celosvětová novinářská cena ICFJ Knight International Journalism Award. [19]

Za svoji nejzajímavější kauzu považuje kauzu Kočnerova knižnica, která se týká souboru dat o velikosti 53 TB představující materiály, se kterými pracovala slovenská policie při vyšetřování vraždy Jána Kuciaka a Martiny Kušnírové. Celý projekt spojil slovenská média do jednoho týmu, aby tento soubor dat prošli a publikovali to, co bylo ve veřejném zájmu. Data získala a technicky zpracovala organizace OOCRP a soubor obsahuje různé dokumenty a důkazní materiály, jako jsou videozáznamy, digitální kopie telefonů, USB, počítačů a další. Původní objem dat byl téměř 70 TB, po očištění dat zbylo k analýze již zmíněných 53 TB. Přístup k datům z důvodu bezpečnosti mohou získat pouze akreditovaní slovenští novináři, jelikož se jedná o vysoce citlivé dokumenty. [18]

1.1.5 Ján Kuciak

Ján Kuciak byl významný slovenský investigativní novinář a analytik, který byl 21. února 2018 zavražděn. Pracoval tehdy pro portál Aktuality.sk a zabýval se kauzami týkající se daňových podvodů, ve kterém figuroval podnikatel Marián Kočner. [20] Kromě toho se v té

době věnoval také působení italské mafie na východě Slovenska a jejího případného propojení do slovenské politiky. Motivem jeho vraždy se stala právě jeho práce a otřásla celým Slovenskem. Na případu na vyšetřování mafie spolupracovala s Kuciakem Pavla Holcová více než 18 měsíců, která po jeho smrti s uveřejněním článku spěchala, protože se zdálo vysoce pravděpodobné, že se na vraždě podílela italská zločinecká skupina 'Ndrangheta. [21]

Kuciak získával informace především prací s otevřenými zdroji a uváděl již známé informace do souvislostí, nazýváno také datová žurnalistika. Na první pohled se tedy nezdálo, že by byl v nějakém nebezpečí. Při vyšetřování daňových podvodů kolem Mariána Kočnera mu bylo však dle portálu Aktuality.sk podnikatelem půl roku před vraždou vyhrožováno. V telefonátu Kuciakovi sdělil, že si může být jistý, že se mu začne osobně věnovat a bude na něj hledat veškerou špínu, kterou poté zveřejní. A nejen jemu, ale i jeho rodině. [22, 23] Ján Kuciak na podnikatele podal 7. září 2017 trestní oznámení, které však ani po 44 dnech nebylo přiřazeno konkrétnímu policistovi k řešení, jak napsal na svém facebookovém profilu. Vyšetřování vázlo a 11. října 2017 Kuciakovi sdělila policistka, že věc vyšetřovat nemůže, neboť hovor zvedl v jiné lokalitě, než podal trestní oznámení, a případ byl přesunut do jiného okresu. Nakonec byl případ uzavřen tak, že se nejednalo ani o trestný čin, ani o přestupek. [24, 25]

Po vraždě Kuciaka zadržela policie čtyři lidi, mimo jiné i Alenu Zsuzsovou, která měla blízko k Marianu Kočnerovi. Mezi další obžalované patřili dále Zoltán Andruskó, blízký přítel Zsuzsové, která ho na vraždu najala a působil jako zprostředkovatel vraždy mezi Zsuzsovou a muži, kteří vraždu vykonali. Těmi byli Tomáš Szabó a Miroslav Marček, bratranci, které si Andruskó najal. Prvním odsouzeným se stal Andruskó, který podepsal dohodu o vině a trestu. Miroslav Marček, bývalý voják z povolání a Tomáš Szabó, bývalý policista, byli ti, kteří samotnou vraždu vykonali, resp. Marček byl ten, který stiskl spoušť a zastřelil novináře i jeho snoubenku Martinu Kušnírovou. Oba pachatelé byli odsouzeni na 25 let. [26, 27] Marián Kočner a Alena Zsuzsová, kteří jsou obvinění z objednání Kuciakovy vraždy, však odsouzeni stále nejsou. Nový rozsudek nad Kočnerem a nad Zsuzsovou by soud mohl podle lednového oznámení šéfky příslušného senátu Ruženy Sabové vyhlásit v dubnu 2023 (v době psaní této práce nebyl rozsudek ještě známý). [28]

O vraždě Kuciaka a jeho snoubenky vznikl dánský koprodukční dokument s názvem Kuciak: Vražda novináře, který se detailně zabývá jak vraždou Jána Kuciaka a Martiny Kušnírové, tak i následky této události, která na Slovensku vyvolala největší protesty od pádu komunismu, včetně vzniku Kočnerovy knižnice. Česká premiéra se uskutečnila 23. března 2023. Jednou z protagonistek filmu je právě Pavla Holcová. [29]

2 Historie uniklých dokumentů

Předchůdci investigativní žurnalistiky byli tzv. *muckrakers* (volně přeloženo do češtiny jako „kydání hnoje“) [30], kteří se snažili o reformy na místní, státní i federální úrovni. Jejich podrobné vyšetřování a odhalování zkorumpované moci, od zneužívání dětské práce po městské politické machinace a železniční a ropné trusty, vedlo k progresivnímu hnutí v celostátní politice. [31]

Pojem investigativní žurnalistika se spojuje především s kauzou Watergate z roku 1972, která inspirovala mnoho lidí k novinářské profesi – jedním z nich je také David Leigh, britský investigativní žurnalista a držitel ceny Daniel Pearl Award od ICIJ. [32, 33] Dva mladí reportéři, Carl Bernstein a Bob Woodwart, svrhli tehdejšího prezidenta Spojených států amerických, Richarda Nixona. Policisté tehdy zatkli v kancelářích Demokratické strany v komplexu Watergate několik mužů, kteří se snažili zprovoznit odposlouchávací zařízení a fotili na místě různé dokumenty. Zmínění reportéři z Washington Post se poté pustili do vyšetřování a přišli s informací, že stopy vedou do Bílého domu. Woodward a Bernstein odhalili finanční vazby mezi Nixonovou volební kampaní a zloději, kteří byli 17. června 1972 zatčeni. S tímto skandálem veřejně také spojili významné washingtonské osobnosti, jako byl Nixonův bývalý generální prokurátor John Mitchell. [34]

Politické vyšetřování začalo v únoru 1973, kdy Senát zřídil výbor pro vyšetřování aféry Watergate. Výbor odhalil existenci tajných nahrávek z Bílého domu, což vyvolalo velkou politickou a právní bitvu mezi Kongresem a prezidentem. Poslední úder přišel s rozhodnutím Nejvyššího soudu, který nařídil Nixonovi zveřejnit další nahrávky z Bílého domu. Jedna z nich se stala známou jako páska *smoking gun* (v překladu kouřící pistole), když odhalila, že se Nixon podílel na utajování aféry Watergate.

Podle Pavly Holcové však moderní investigativní žurnalistika vznikla dříve, a to kauzou nelidských podmínek v psychiatrické léčebně, kterou odhalila novinářka Nellie Bly. Tajná operace odhalující zneužívání v ústavu na Blackwellově ostrově, nyní Rooseveltově ostrově, odstartovala to, co se proměnilo v seriózní investigativní žurnalistiku. Šlo o kauzu, která otřásla celým světem. [35] Mladá novinářka předstírala nepříčetnost, aby se nechala zavřít do ústavu. Zde žila deset dní po boku sebevražedných, násilnických a psychotických žen, ale i zcela zdravých žen, které byly omylem zavřeny do ústavu. [36]

V roce 2006 vznikl webový portál WikiLeaks Julianem Assangem, online knihovna uniklých dokumentů, která v současné době obsahuje více než deset milionů úředních dokumentů a dalších materiálů týkajících se války, špionáže a korupce, většinou spojených se Spojenými státy. V roce 2010 americká aktivistka a whistleblowerka Chelsea Manning vyzradila WikiLeaks utajované informace prostřednictvím zabezpečeného přenosu souborů, když působila jako zpravodajská analytička v Iráku. Manning poskytla serveru WikiLeaks přibližně 250 tisíc amerických diplomatických depeší a 480 tisíc armádních zpráv z terénu z válek v Afghánistánu a Iráku. V roce 2007 WikiLeaks zveřejnil utajované video americké armády, na kterém američtí vojáci z vrtulníku Apache nevybíravě zabíjejí

irácké civilisty poblíž Nového Bagdádu a další soubory, které obsahovaly výpovědi vojáků americké armády a jejich zážitky z pozemní války v Iráku. [37]

Vzhledem k povaze diplomové práce se nyní zaměřím na historii uniklých dokumentů týkající se offshore firem.

V dubnu 2013 vyšel soubor **Offshore Leaks** od ICIJ díky úniku 2,5 milionů soukromých obchodních zápisů, který obsahuje podrobnosti o více než 120 000 offshorových účtech a odhalil skryté obchody politiků, podvodníků a miliardářů po celém světě. Uniklé dokumenty poskytují informace o převodech peněz, dat založení, vazeb mezi společnostmi a jednotlivci, které ilustrují, jak se offshorové finanční tajemství agresivně rozšířilo po celém světě, což umožňuje bohatým a dobře propojeným lidem vyhýbat se placení daní a podporuje korupci a hospodářské problémy v bohatých i chudých zemích. Na analýze dokumentů spolupracovala ICIJ s reportéry britských deníků The Guardian a BBC, kanadské televizní společnosti Canadian Broadcasting Corporation (CBC), francouzského Le Monde, německých Norddeutscher Rundfunk a Süddeutsche Zeitung, deníku The Washington Post, a dalších 31 mediálních partnerů z celého světa. Záznamy podrobně popisují offshorové majetky lidí a společností ve více než 170 zemích a teritoriích. Celková velikost souborů v GB [6] je více než 160krát větší než únik dokumentů amerického ministerstva zahraničí od Chelsea Manning na WikiLeaks. [38]

3. dubna 2016 ICIJ spolu s více než stovkou novin, mezi které patřily například německé Süddeutsche Zeitung, zveřejnilo další soubor uniklých dokumentů známý jako **Panama Papers**, který obsahuje podrobnosti o offshorových účtech světových lídrů a ukazuje, jak spolupracovníci ruského prezidenta Vladimira Putina tajně převáděli až 2 miliardy dolarů přes banky a stínové společnosti. Dokumenty ukázaly, že významné bankovní subjekty napomáhaly tajnému skrývání těžko dohledatelného majetku na místech, jako jsou Britské Panenské ostrovy a Panama. [39]

Uniklé dokumenty, které prověřoval tým více než 370 novinářů z téměř 80 zemí, pocházejí z právní firmy Mossack Fonseca se sídlem v Panamě, která má pobočky v Hongkongu, Miami, Curychu a na více než 35 dalších místech po celém světě a jsou z anonymního zdroje. Jejich vyšetřování odhalilo tajné offshorové majetky 12 světových lídrů, více než 128 dalších politiků a desítky podvodníků, obchodníků s drogami a dalších zločinců, jejichž společnosti byly v USA i jinde zařazeny na černou listinu. Soubor obsahuje 11,5 milionů záznamů o více než 214 tisíc offshorových firmách. Záznamy se týkají téměř 40 let, od roku 1977 do konce roku 2015. Umožňují nevídaný pohled do světa offshorů, neboť poskytují každodenní pohled na to, jak peníze proudí globálním finančním systémem a připravují státní pokladny o daňové příjmy. [40]

Ve stejném roce po zveřejnění Panama Papers v dubnu 2016 předal neznámý zdroj interní údaje z národního registru firem na Bahamách Frederiku Obermaierovi a Bastianu Obermayerovi, kteří je s pomocí ICIJ analyzovali. Tak vznikl soubor dokumentů **Bahamas**

[6] GB = gigabyte

Leaks. Soubor tvoří sice pouze desetinu co do velikosti v porovnání s Panama Papers, ale jsou neméně důležité. [41] Ve spojení s Panama Papers poskytují údaje z Baham nový pohled na offshorové obchody politiků, zločinců a manažerů, stejně jako bankéřů a právníků, kteří pomáhají převádět peníze. V uniklých dokumentech jsou uvedena jména politiků a dalších osob spojených s více než 175 000 bahamskými společnostmi registrovanými v letech 1990-2016. Mimo jiné obsahují jména 539 registrovaných agentů, resp. firemních prostředníků, kteří slouží jako prostředníci mezi bahamskými úřady a zákazníky, kteří chtějí založit offshore společnost. Mezi nimi je i Mossack Fonseca, právní firma, jejíž uniklé spisy se staly základem pro Panama Papers. Tato firma založila na Bahamách 15 915 subjektů, což z nich činí třetí nejvyužívanější jurisdikci Mossack Fonseca. [42]

Další významný soubor uniklých dokumentů offshore firem dostal název **Paradise Papers** a byl zveřejněn v listopadu 2017. Stejně jako v případě Panama Papers získaly uniklá data německé noviny Süddeutsche Zeitung a sdílely je s ICIJ a sítí více než 380 novinářů v 67 zemích. [43] Součástí této sítě byli také novináři z investigace.cz.

Uniklá data pochází ze dvou offshorových firem Appleby a Estera. Součástí souboru uniklých dokumentů jsou i data z 19 obchodních rejstříků z míst, která slouží jako nervová centra globální šedé ekonomiky. [44]

Celkem uniklo 13,4 milionů záznamů, které poukazují na to, jak moc hluboko je propojen systém daňových rájů nejen se světem soukromého majetku a politiky, ale i korporátních gigantů, jako jsou společnosti Nike, Apple, Uber a dalších globálních firem, které se vyhýbají daním stále sofistikovanějšími způsoby. Dokumenty mimo jiné odhalují také například tajné obchody Justina Trudeaua a offshorové aktivity bývalé britské královny, která investovala miliony dolarů do farmaceutických a úvěrových společností, své offshorové investice však nikdy nezveřejnila. [43, 44] Téměř 7 milionů záznamů od společnosti Appleby a přidružených společností pokrývá období od roku 1950 do roku 2016 a zahrnuje e-maily, smlouvy o miliardových půjčkách a bankovní výpisy týkající se nejméně 25 tisíc subjektů spojených s lidmi ve 180 zemích. [43]

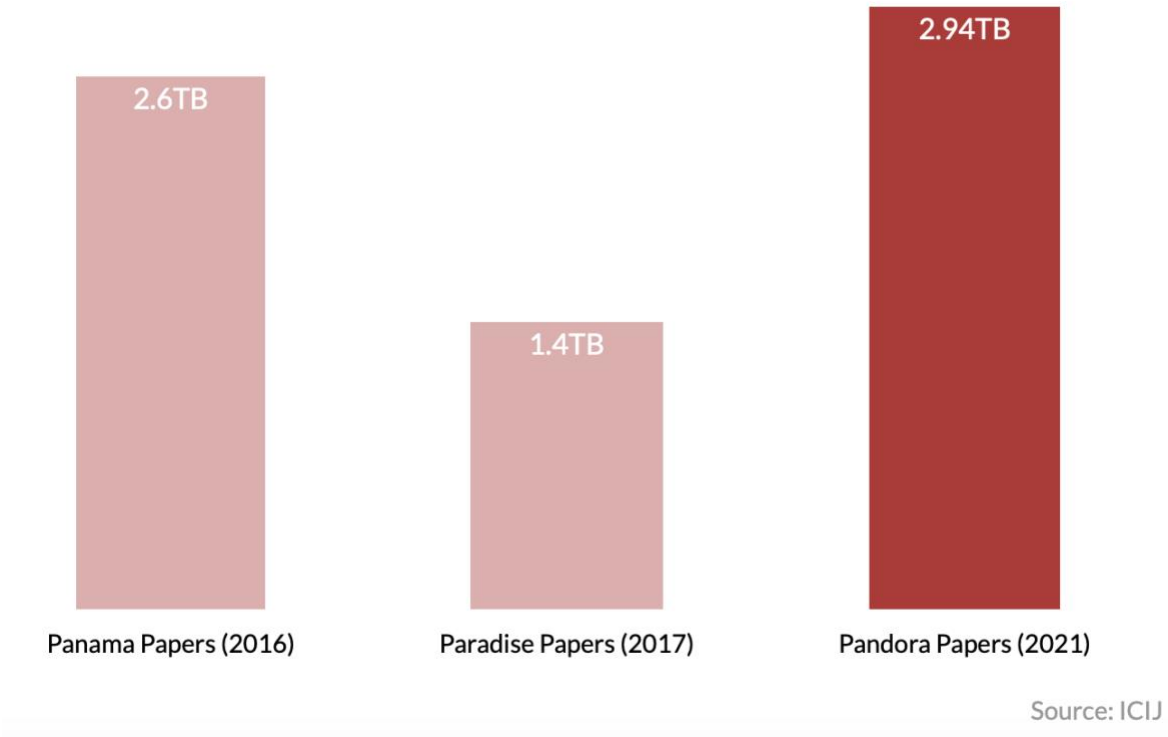
Od prosince 2021 pak došlo k dalšímu velkému úniku dat, tento soubor dostal název **Pandora Papers**, odvozeno od Pandořiny skříňky, ve které bylo zamknuto všechno zlo světa. Pandora Papers odhalují vnitřní fungování stínové ekonomiky, z níž těží bohatí a dobře propojení lidé na úkor všech ostatních.

ICIJ získalo soubor více než 11,9 milionu důvěrných spisů a vedlo tým více než 600 žurnalistů ze 150 novinářských agentur, například The Washington Post, The BBC, The Guardian, Radio France, ale také novináři z investigace.cz. Tento tým žurnalistů dva roky spis procházel, pátral po těžko dostupných zdrojích a prohledával soudní záznamy a další veřejné dokumenty z desítek zemí. Uniklé záznamy pocházejí od 14 právních společností z celého světa, které zakládají fiktivní společnosti a další offshorová zákoutí pro klienty. Záznamy obsahují informace o obchodech téměř třikrát více současných i bývalých představitelů zemí než jakýkoli předchozí únik dokumentů z offshorových rájů. [45]

Jednalo se o největší novinářskou spolupráci v historii, která odhalila 956 společností v offshorových rájích, finanční tajemství 35 současných a bývalých světových vůdců, více

než 330 politiků, kteří jsou s offshorovými firmy spojení a mezi kterými byl mimo jiné také bývalý premiér České republiky Andrej Babiš, a veřejných činitelů v 91 zemích a teritoriích a řadu uprchlíků, podvodníků a vrahů. Více než dvě třetiny těchto společností byly založeny na Britských Panenských ostrovech, jurisdikci, která je již dlouho známá jako klíčový článek offshoreového systému. Data byla zpřístupněná veřejnosti ve třech várkách. Poslední várka byla zveřejněná v květnu 2022. [46]

V porovnání s předchozími úniky, soubor Pandora Papers je větší než jakýkoliv předchozí únik, který ICIJ vyšetřovala. Na *Obrázek 1* můžeme vidět porovnání velikostí tří největších souborů uniklých dokumentů Panama Papers, Paradise Papers a Pandora Papers v terabytech [7].



Obrázek 1. Porovnání velikosti souborů uniklých dokumentů vyšetřovaných ICIJ. [45]

[7] Terabyte, zkratkou TB, 1 TB = 1 000 gigabyte

3 Data v oblasti investigativní žurnalistiky

Jak již bylo zmíněno v první kapitole, investigativní žurnalistika spočívá v odhalování důležitých věcí, které jsou skryty veřejnosti.

Práce investigativního novináře je stejně jako práce běžného novináře závislá na informacích. Běžné zpravodajství je z velké části a někdy zcela závislé na materiálech, které poskytují jiní (např. policie, vlády, firmy atd.), je v zásadě reaktivní, ne-li pasivní. Investigativní zpravodajství naproti tomu závisí na materiálech shromážděných nebo vytvořených z vlastní iniciativy reportéra, a to vyžaduje použití tajných i otevřených zdrojů a dokumentů. [3]

Zdroj v žurnalistice chápeme jako zdroj informací. Pokud novinář referuje o skutečnosti či události, při které nebyl osobně přítomen, je při tvorbě reportáže odkázán právě na něj. [30] Následující kapitoly se budou zabývat metodikou sběru dat při vyšetřování investigativních žurnalistů, následným ověřováním pravdivosti dat, jejich klasifikací a systému třídění při zpracování a analýze, uchovávání dat, a nakonec ochranou dat, neboť se často jedná o citlivé informace.

Abychom mohli začít sbírat data ke konkrétnímu případu či příběhu, je nutné si nejprve zvolit, jaký příběh budeme psát. Jak se však rozhodnout? V našem rozhovoru Pavla Holcová uvedla, že podněty k zahájení vyšetřování dostávají například od zahraničních kolegů, kteří se na ně obrací se žádostí o pomoc v konkrétních kauzách. V jiných situacích sledují a reflektují konkrétní dění v ČR či Evropě.

Jednou z možností je tedy sledování médií. Obecně je vhodné sledovat dané odvětví, kterému se chceme věnovat, abychom se naučili identifikovat vzorce, a tak si uvědomit, kdy se objeví něco neobvyklého. Další možností je naslouchat stížnostem lidí nebo sledovat své okolí a změny v okolí. Pokud se zastavíme u otázky „Proč to tak musí být?“ či „Proč se to děje?“, je dost možné, že je nutné danou problematiku blíže prozkoumat. A zde přichází nutnost sběru dat – jak řekl Jaroslav Kmenta, bez zdrojů to nejde. [3, 5]

3.1 Metodika sběru dat

Dle Karla Hvížďaly je kvalita informace podmíněna kvalitou rešeršování – a kvalita rešerše záleží na důvěryhodnosti média. [7] Rešerší je myšlen proces, při kterém dochází k vyhledávání, shromažďování a ověřování faktů a informací. Informace můžeme čerpat z různých zdrojů. Nejrozšířenějším zdrojem informací v dnešní době je internet, jejich kvalita je však různorodá a je proto nutné informace a podklady po shromáždění zhodnotit a ověřit jejich pravdivost a spolehlivost. Zdroje můžeme dělit na otevřené a vlastní. [7, 30]

3.1.1 Otevřené zdroje

Otevřeným zdrojem chápeme takové zdroje, které jsou veřejně přístupné a které byly volně publikovány v jakémkoliv dostupném médiu. Obvykle je můžeme dohledat a získat ve veřejných knihovnách, rejstřících, databázích nebo v archivu daných médií, jako jsou například odborné a vědecké publikace, zpravodajství v televizi, novinách, časopisech, rozhlasu či na internetu, se kterým jsou spojená mimo jiné například sociální média zainteresovaných stran.

Univerzitní knihovny jsou otevřené zdroje, které mohou disponovat širší škálou informací a aktualizovanou technikou ve srovnání s veřejnými knihovnami. Neméně důležitým zdrojem jsou dále **katastrální úřady**, které shromažďují údaje o vlastnictví nemovitostí a často i o nesplacených úvěrech souvisejících s těmito nemovitostmi. Například ve Francii byly informace o majetku politiků využity k prokázání toho, že získali mnohem větší majetek, než by odpovídalo jejich zveřejněným příjmům. Kromě toho veřejné zprávy a dokumenty nám mohou poskytnout informace o společnostech, jejich strategiích a dalších činnostech. Může jít o výroční zprávy, regulatorní dokumenty a tiskové zprávy, které často vysvětlují důvody rozhodnutí společnosti. Navíc pokud společnost působí v zahraničí, mohou být informace dostupné v zahraničních spisech snadněji dostupné než v domácích. [3]

Dalším cenným zdrojem jsou například soudy a **soudní záznamy**. Soudy vedou minimálně záznamy o právních rozhodnutích. V některých zemích, například ve Spojených státech, jsou soudní záznamy přístupné veřejnosti, včetně všech důkazů předložených v soudním řízení. Je důležité shromáždit všechny soudní dokumenty týkající se vyšetřovaných osob nebo organizací v každé zemi, kde působí. Svědectví podaná v soudních řízeních jsou obvykle chráněna před trestním stíháním. Pokud je investigativní novinář přítomen soudnímu jednání, měl by si pořizovat podrobné poznámky, zejména pokud není přítomen soudní zapisovatel. [3] U nás však vydání soudního rozsudku na žádost často záleží na konkrétním soudci. Pavla Holcová v rozhovoru s Terezou Dubinovou, kulturoložkou a hebraistkou, která na svém webu Ohel Adom publikuje rozhovory se zajímavými osobnostmi, zmínila, že by ráda zákon o svobodném přístupu k informacím rozšířila tak, aby bylo možné získat jednodušeji přístup k soudním rozsudkům, byť anonymizovaným. V současné době to přirovnává k loterii, buď rozsudek získá anonymizovaný, nebo musí dodat jednací číslo, které ve většině případů nelze zjistit. [6]

Například případ týkající se albánského organizovaného zločinu, který pro nás zapadá do větší mozaiky organizovaného zločineckého uskupení. Víme, že nějaký člověk byl odsouzený, ale nejsme schopní dohledat kde, nejsme schopni dohledat číslo jednací, a tím pádem ani nevíme, na koho se obracet, aby nám dali nahlédnout do rozsudku. [6]

V neposlední řadě jsou velmi důležitými zdroji **obchodní, živnostenský a insolvenční rejstříky**. V každé zemi existuje vládní úřad, který vede záznamy o vlastnictví společností a o tom, zda jsou veřejně obchodovatelné. Množství informací, které musí vlastníci společností zveřejnit, se může lišit, ale obvykle je to více, než očekávají zpravodajové, kteří tyto zdroje nevyužívají. [3]

Seznam otevřených zdrojů je velmi rozmanitý a nelze vypsat všechny existující možnosti, neboť každý novinář informace hledá dle typu projektu na různých místech a svůj seznam dle potřeb upravuje. Každá kauza totiž začíná rešerší pomocí otevřených zdrojů. To potvrzuje také Pavla Holcová v našem rozhovoru (viz Příloha B)

Na začátku bývá ve většině případů jméno člověka nebo firmy. Většinou toto jméno prověříme ve veřejných databázích jako obchodní rejstřík, rejstřík konečných majitelů, katastr – a pak v našich archivech a neveřejných databázích. Podíváme se na obchodní i jiné vazby, co už se o tom člověku ví – jestli vůbec něco – a na základě těchto dat definujeme prvotní hypotézu, kterou se pak snažíme ověřit.

Hypotéza definuje konkrétní otázky, na které je třeba odpovědět, pokud chceme zjistit, zda má smysl. To se děje prostřednictvím procesu, v němž hypotézu rozebíráme a zjišťujeme, jaká jednotlivá konkrétní tvrzení obsahuje. Poté můžeme postupně ověřit každé z těchto tvrzení. Díky hypotéze získáme něco, co můžeme ověřit a zvýšit tak naše šance na odhalení skrytých tajemství. [3]

Otevřené zdroje nám umožňují odvodit potenciální tajemství z veřejně dostupných informací, místo abychom se spoléhali na zdroje, které tvrdí, že mají přístup k tajným informacím. [3] Pomocí otevřených zdrojů ověřujeme také informace z jiných zdrojů. V rozhovoru s Terezou Dubinovou sdílela Pavla Holcová, že drtivá většina kauz jejího týmu je postavena právě na otevřených zdrojích, jelikož všechny publikované informace musejí být zpětně dohledatelné. [6] Aby byl proto investigativní žurnalista úspěšný, musí umět využívat zákon o svobodném přístupu k informacím. [5] Výše uvedené zdroje využívají samozřejmě také tradiční novináři. U investigativní žurnalistiky je však specifické to, že data analyzují v detailnějším měřítku, a zjištění pak uvádějí do mnohem širšího kontextu.

Open Source Intelligence (OSINT)

S investigativní žurnalistikou se dnes bezesporu pojí pojem Open Source Intelligence (zkratkou OSINT) neboli zpravodajské informace z otevřených zdrojů. Metody a nástroje OSINT se staly nedílnou součástí vytěžování informací pro novinářské účely, a ačkoliv se jedná o informace výhradně z veřejně dostupných zdrojů, získání relevantních informací bývá často otázkou dlouhých hodin rešerše a ověřování relevance informací. [47]

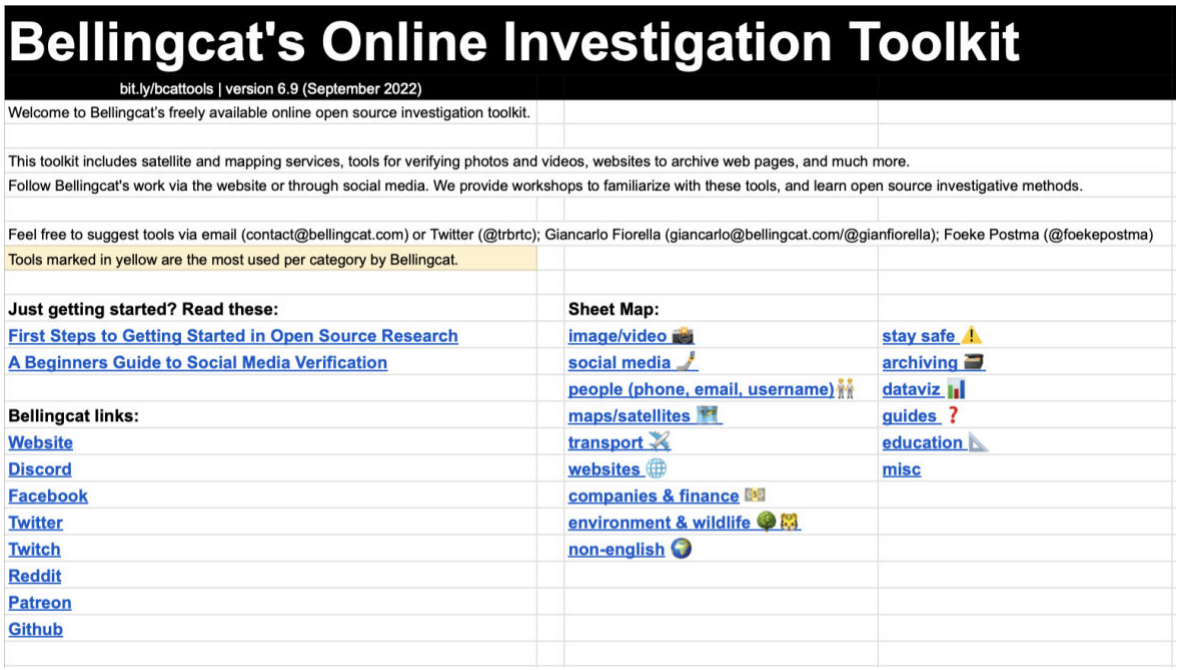
Abych uvedla konkrétní definici, společnost Maltego OSINT definuje jako *zpravodajské informace získané z veřejně dostupných informací, které jsou shromažďovány, analyzovány a sdíleny za účelem pomoci při konkrétním vyšetřování.* [8] [48]

Dalo by se říct, že se jedná o jeden z postupů, jakým se řídí investigativní žurnalisté při vyšetřování různých kauz. A nejen ti. Metody a nástroje OSINT se využívají také například

[8] Původní znění: *We define OSINT as intelligence produced from publicly available information that is collected, analyzed, and shared for the purpose of aiding a specific investigation.*

v oblasti kybernetické bezpečnosti, při vyšetřování podvodů či další různé orgány činné v trestním řízení.

Existuje mnoho OSINT nástrojů, které lze využít při prohledávání otevřených zdrojů. Jedním z nejkomplexnějších seznamů nástrojů vytvořilo sdružení občanských novinářů, kteří pracují s otevřenými zdroji, Bellingcat, mimo jiné také ve spolupráci s českými novináři z investigace.cz. Seznam má v současné době několik set položek a je rozdělen do kategorií dle typů média, jako jsou sociální sítě, vyhledávání pomocí fotografií či videí, vyhledávání informací z map, satelitů a lokací, monitorování leteckého provozu a další. [9]



Bellingcat's Online Investigation Toolkit

bit.ly/bcattools | version 6.9 (September 2022)

Welcome to Bellingcat's freely available online open source investigation toolkit.

This toolkit includes satellite and mapping services, tools for verifying photos and videos, websites to archive web pages, and much more.
Follow Bellingcat's work via the website or through social media. We provide workshops to familiarize with these tools, and learn open source investigative methods.

Feel free to suggest tools via email (contact@bellingcat.com) or Twitter (@trbrtc); Giancarlo Fiorella (giancarlo@bellingcat.com/@gianfiorella); Foeke Postma (@foekepostma)
Tools marked in yellow are the most used per category by Bellingcat.

Just getting started? Read these:	Sheet Map:	
First Steps to Getting Started in Open Source Research	image/video	stay safe
A Beginners Guide to Social Media Verification	social media	archiving
	people (phone, email, username)	dataviz
Bellingcat links:	maps/satellites	guides
Website	transport	education
Discord	websites	misc
Facebook	companies & finance	
Twitter	environment & wildlife	
Twitch	non-english	
Reddit		
Patreon		
Github		

Obrázek 2. Náhled dokumentu se seznamem OSINT nástrojů od Bellingcat. [49]

Aktivní užívání OSINT nástrojů potvrdila v našem rozhovoru Pavla Holcová, stejně jako občasnou spolupráci s organizací Bellingcat.

Bellingcat

Organizace Bellingcat hraje v kontextu s OSINT metodami velkou roli. Jak již bylo naznačeno, jedná se o nezávislé sdružení občanských novinářů, výzkumníků a vyšetřovatelů, které využívá otevřené zdroje a sociální média k vyšetřování různých témat. Za dobu svého působení byly její aktivity obzvláště významné z hlediska šíření informací o konfliktech, zločinnosti a porušování lidských práv. Jedním z jejich největších úspěchů spočívá například v kauze sestřelení letu MH-17 nad Ukrajinou, kdy během 4 let po incidentu lokalizovali pole v Rusku, odkud byla raketa vystřelena a podařilo se jim identifikovat řadu podezřelých osob zapojených do incidentu včetně vysokých důstojníků

[9] Odkaz na seznam OSINT nástrojů od Bellingcat: bit.ly/bcattools

ruského ministerstva obrany. Kromě toho zjistili, že ruská armáda byla do případu zapojena roky předtím, než to potvrdili evropští představitelé. [50]

3.1.2 Vlastní zdroje

Kromě otevřených zdrojů může žurnalista čerpat také z vlastních zdrojů. Jedním z takových zdrojů je například anonymní či tajný zdroj. S takovými zdroji musí však žurnalista zacházet opatrně, neboť se jedná o jeden z nejnebezpečnějších zdrojů a neměly by se použít bez důkladného prověření jejich pravdivosti. Ideální je, když je schopen informace z takových zdrojů ověřit právě otevřenými zdroji. Žurnalista by neměl celou kauzu stavět pouze na tajných zdrojích, anonymní tipy jsou spíše podnětem k novému příběhu a mohou upozornit na zajímavý problém, na který je třeba se více zaměřit. Pokud by případ skončil u soudu, nemůže tato anonymní tvrzení použít jako zdroj, jelikož dané osoby nemůže identifikovat. [51] To potvrzuje také investigativní novinář Jaroslav Kmenta:

Je chiméra si myslet, že investigativní novinář je závislý jen na svých tajných zdrojích. Není to pravda. Samozřejmě, když je nemáte, nevíte, co se děje, a to je špatně. Ale jestli chcete dostat nějakou kauzu do novin, nemůžete ji zveřejnit jen s odkazem na nejmenovaný zdroj. Musíte si najít další důkazy, dokumenty, svědky, indicie. Pracujete skoro jako policajt. [5]

Naopak jedním z nejspolehlivějších zdrojů informací je očité svědectví novináře. Svědkem se novinář stává, když je přítomen při nějaké události, ať už veřejné či neveřejné. S tím se pojí další možný zdroj informací, a to vlastní zkušenosti novináře nebo lidí v jeho okolí. Nemusí to však být pouze svědectví samotného novináře. Zdrojem může být kdokoliv, kdo byl v centru dění nebo problému. Takové lidi označujeme za primární zdroje. Může to být napadená oběť, zaměstnanec vyšetřované firmy, vedoucí odboru, který vede jednání o mzdách nebo zločinec, jehož se kauza týká. Takoví lidé jsou obvykle nejlepším zdrojem informací o své části události, svým úhlem pohledu a rozhovorem můžeme získat podnětné informace ke kauze. Je však nutné brát v potaz, že jejich svědectví nemusí být objektivní nebo přesné, a proto je nutné překontrolovat a porovnat fakta s jinými zdroji.

Dalším zdrojem mohou být různé novinářovy kontakty, například jiní novináři – kolegové. Pokud jsou zkušení, jejich zprávy a informace budou obvykle přesné a lze jim věřit. Byť jsou jedním z nejspolehlivějších zdrojů, neznamená to, že jsou bezchybné. Může se stát chyba, novinář špatně interpretuje to, co si myslí, že viděl, a pak to prezentuje jako skutečnost. Příkladem je třeba interpretace velikosti davu, který novinář vidí. Pokud je na místě velké množství lidí a novinář do své zprávy napíše „Na místě bylo 20 tisíc lidí“, velmi pravděpodobně se bude jednat o odhad a taková zpráva by se měla brát s rezervou. [51]

3.2 Verifikace dat

Povinností investigativního žurnalisty je přinášet čtenářům pravdivé informace, které jsou podložené fakty. Naneštěstí pro nás je velmi těžké rozpoznat fakta od lží. Každou informaci, kterou proto získáme, je nutné verifikovat, ideálně z více zdrojů. Prvním pravidlem při ověřování dat je zpochybňovat všechno a všechny. Data mají reprezentovat realitu v určitém

časovém okamžiku. Abychom ověřili, že soubor dat odpovídá skutečnosti, je potřeba provést proces verifikace dvakrát –poprvé musí proběhnout bezprostředně po získání dat a zjištění, tzv. insights, která jsme zjistili z dat, musí být ověřeny na konci fáze analýzy. [52]

Obrázek 3. Jednotlivé fáze procesu práce s daty při investigaci. [52]

Při hodnocení autenticity zdroje musíme být skeptičtí. V případě komunikace s lidskými zdroji bychom měli posoudit, zda je daná osoba skutečně tím, za koho se vydává – je schopna prokázat svoji totožnost? Má nějaký záznam v trestním rejstříku? Jak adekvátní jsou informace, které nám poskytují? Poskytuje zdroj úplné vysvětlení? Jaký je její motiv pro výpověď? V předchozí kapitole jsem zmínila, že zdroje mohou vypovídat nepřesně a neobjektivně, jejich výpovědi mohou být ovlivněny osobními pocity. V jiných případech můžou prostě udělat chybu nebo si nepamatují věci do detailu. Možná se nás snaží oklamat. Měli bychom sledovat, jak se osoby při rozhovoru chovají, zda působí věrohodně. Kromě toho bychom se měli snažit získat důkazy, které jejich tvrzení podpoří – nebo naopak vyvrátí. Jádrem ověřování je otázka: „Jak to víte?“. [53]

Úkolem novináře je zpochybňovat zdroje a materiály, které mu poskytují, porovnávat je s jinými zdroji a vyřazovat to, co je nepravdivé nebo nedostatečně ověřené. Otevřené zdroje nejsou pouze výborným zdrojem dat, ale také prostředkem k ověření správnosti a přesnosti získaných informací. Každou informaci a jakékoliv tvrzení, které použijeme, musíme podložit průkaznými materiály. K tomu nemohou sloužit tajné zdroje. Můžeme však využít již zmíněné OSINT nástroje, jichž je celá řada a jsou kategorizovány dle typu médií, případně dle typu informací, které potřebujeme ověřit, například ověření identity osoby, lokace či ověřování obsahu vytvořeného uživatelem neboli user-generated content.

Proces verifikace dat je velmi vyčerpávající a časově velmi náročný. V našem rozhovoru Pavla Holcová uvedla, že mají k ověřování dat sestavený externí tým ověřující fakty, tzv. fact-checking tým, který nad ověřováním jednotlivých faktů může strávit až pět dní. Níže představím vybrané nástroje, které lze použít pro ověření faktů.

3.2.1 Vyhledávače

První nástroj, který by nejen novináři měli využít při ověřování správnosti informací, jsou vyhledávače. Správné, a především efektivní využití možností a síly vyhledávačů však spočívá v mnohem více faktorech, než jen obyčejné vyhledávání pomocí klíčových slov. Každý vyhledávač nabízí spoustu operátorů, které umožňují přesnější vyhledávání a zúžení výsledků na ty opravdu relevantní.

Mezi ty základní a nejčastěji využívané operátory patří:

- **AND** – vyhledání takových výsledků, které obsahují obě klíčová slova – zúží tím výsledky.
- **OR** – vyhledá jedno nebo obě klíčová slova – zajišťuje širší výsledky vyhledávání než AND.
- **NOT** – zobrazí výsledky vyhledávání, které obsahují první klíčové slovo, ale ne druhé.

Existuje však mnoho dalších operátorů, pomocí nichž můžeme například zúžit výsledky vyhledávání pouze na ty z konkrétní stránky, konkrétního typu souboru, můžeme vyhledávat klíčová slova buď pouze v názvu nebo naopak pouze v těle dokumentu a mnoho dalšího.

Zápisy operátorů se mohou lišit podle jednotlivých vyhledávačů. Před jejich využití je proto nutné podívat se na dokumentaci příslušného vyhledávače či informačního systému, abychom operátory využili správně.

Dále je nutné také brát v potaz, že existují různé vyhledávače a každý z nich vrací trochu jiné výsledky. Je to z toho důvodu, že každý vyhledávač používá vlastní algoritmus, podle kterého se rozhoduje, která stránka je relevantní. Když potřebujeme nějaký specializovaný druh výsledku, je vhodné sáhnout po takovém specializovaném vyhledávači. Například vyhledávač Dogpile prohledává ostatní vyhledávače a vrací výsledky bez reklam. DuckDuckGo je vyhledávač, který na rozdíl od ostatních nesleduje uživatelovu digitální stopu. Dalším takovým vyhledávačem je například StartPage. Dále existují vyhledávače, které vyhledávají speciálně informace o cestování, videa nebo obrázky. [54]

V tabulce níže lze vidět ukázky použití některých pokročilých operátorů ve vybraných vyhledávačích. Můžeme zde vidět rozdíly v syntaxu operátorů, případně chybějící operátory v určitých vyhledávačích. To naznačuje, že je každý vyhledávač vhodný na různé typy dotazů.

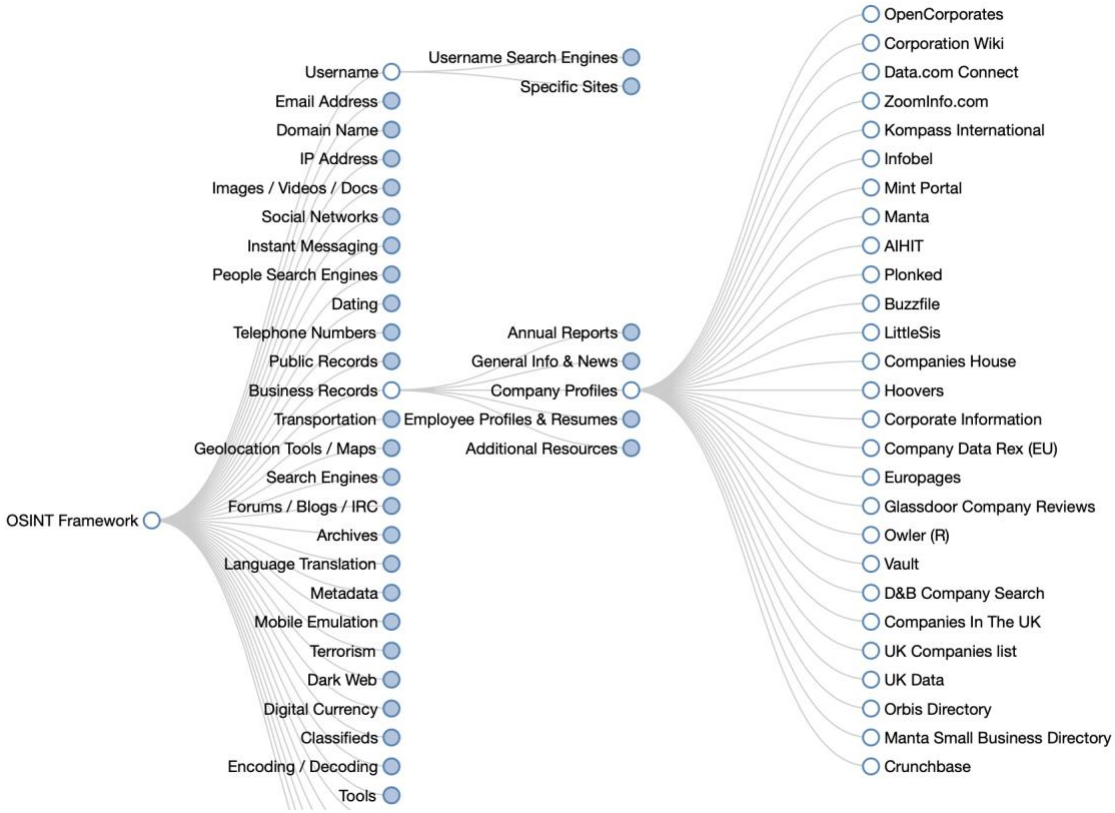
Tabulka 1. Ukázka vybraných pokročilých vyhledávacích operátorů. [55–57]

Funkce operátoru	Google	Bing	Yandex	DuckDuckGo
Přesná shoda fráze	“investigative journalism“	“investigative journalism“	“investigative journalism“	“investigative journalism“
Výsledky, které obsahují konkrétní slovo v názvu	intitle:	intitle:	title:	intitle:
Výsledky, které obsahují konkrétní slovo v těle dokumentu	intext:	inbody:	intext:	-
Omezení výsledků na určité typy souborů obsahující klíčové slovo	filetype:	filetype:	mime:	filetype:
Vyhledání výsledků na určité doméně	site:	site: domain:	domain:	site:
Zobrazení nejnovější verze webu uloženou v mezipaměti	cache:	-	-	-
Výskyt klíčového slova v adrese URL	inurl:	-	-	inurl:
Kontrola, zda je uvedená doména v indexu vyhledávače	-	url:	url:	url:
Výsledky v konkrétním jazyce	lang:	language:	lang:	-
Zástupný znak, odpovídá jakémukoli slovu nebo frázi	*	*	*	*

3.2.2 Příklady datových zdrojů pro verifikaci dat

Kromě vyhledávačů existují specializované zdroje, které se zaměřují na konkrétní typ informací. Příkladem specializovaného zdroje může být například LinkedIn, který slouží pro vyhledávání potenciálních zaměstnanců, či naopak zaměstnavatelů, kteří právě nabírají. Při vyhledávání konkrétních informací jsou tyto zdroje mnohem přesnější než vyhledávače. Patří mezi ně všechny OSINT nástroje, o kterých píšu v kapitole 3.1.1. Z toho vyplývá, že jsou OSINT nástroje užitečné a hojně využívané nejen při sběru dat, ale také při ověřování dat. Vzhledem k rozmanitosti a proměnlivosti OSINT nástrojů neexistuje žádný seznam, o kterém by se dalo říct, že obsahuje všechny existující nástroje. Tyto seznamy se neustále mění, nástroje vznikají a zanikají každou chvíli, každý seznam je různě detailní a aktuální.

Kromě již zmíněného Bellingcat seznamu patří mezi jedny nejznámější OSINT Framework [10], který obsahuje komplexní mapu otevřených zdrojů.



Obrázek 4. Ukázka mapy nástrojů z webu OSINT Framework. [58]

Níže jsou vybrány příklady některých nástrojů, které jsou využívány nejen investigativními novináři při verifikaci dat. Jelikož je nástrojů mnoho, při výběru jsem se zaměřila kromě vlastní znalosti nástrojů také na to, aby jednotlivé nástroje sloužili pro získávání informací z různých oblastí, jako jsou nástroje pro verifikaci letových dat, námořních dat a nástroje pro ověření důvěryhodnosti on-line médií, neboť sociální sítě jsou obrovským zdrojem dat s velkým potenciálem, ale zároveň nemusí být vždy pravdivé. V neposlední řadě je mezi příklady uveden také zdroj zabývající se informacemi o společnostech, což je pro investigativní žurnalisty velmi významným zdrojem – velmi často se totiž zabývají případy týkající se právě korupce firem. V následujícím přehledu uvádím jedny z nejvýznamnějších OSINTových nástrojů, které se používají mimo jiné pro sběr a verifikaci dat nejen hospodářské kriminality.

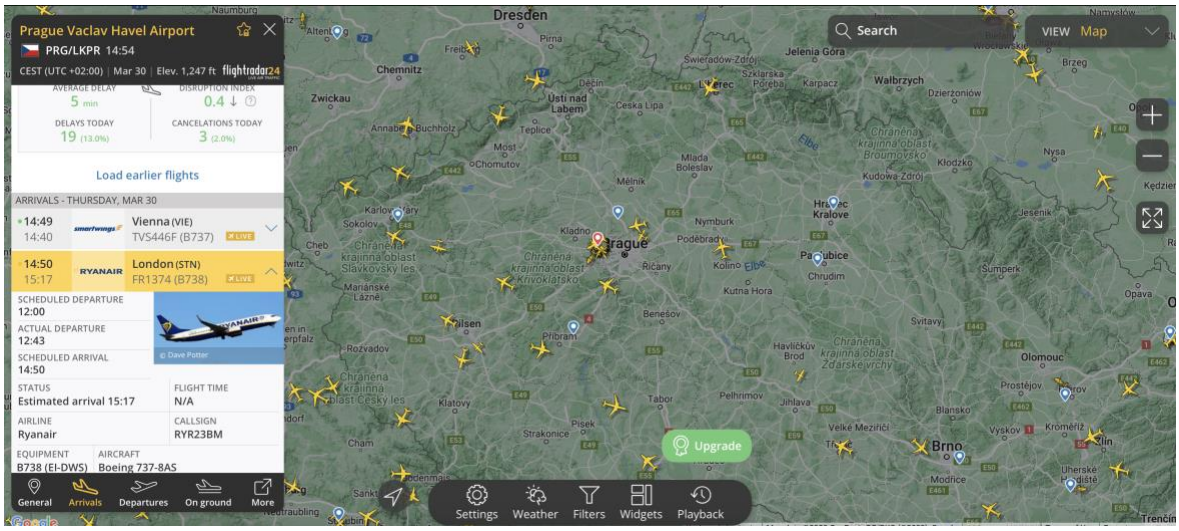
[10] https://osintframework.com/

Flight Radar 24

FlightRadar24 [11] je nástroj pro sledování letů, který v reálném čase poskytuje informace o tisících letadel po celém světě. Jak již název napovídá, služba je dostupná ve dne v noci. Téměř všechny komerční lety, které jsou na cestě nebo právě vzlétají či přistávají, lze sledovat živě. Službu lze využít například v situaci, kdy se do letadla chystá nastoupit někdo z našich blízkých a my chceme sledovat a ověřit si, zda bezpečně dorazili do cíle. Nástroj také poskytuje zajímavé informace o aktuálním stavu leteckého provozu ve městě či jiných oblastech zájmu. Jeho použití je velmi snadné, vyžaduje pouze internetové připojení a zobrazovací zařízení, jako je telefon, počítač či tablet, a nakonec jakýkoliv prohlížeč. [59]

Službu využili novináři z investigace.cz, když sledovali letadla přistávající v Sýrii během vyšetřování kauzy o tom, že Češi dovážejí zbraně do syrského konfliktu.

Udělali jsme tým zhruba třiceti lidí, měli jsme služby a pořád někdo seděl u monitoru. Díval se, kde přistává letadlo v Sýrii, co je to za letadlo, jaké má označení, jaký má tzv. volací kód a odkud tam letěli. Z údajů jsme udělali obrovskou excelovou tabulku, kam jsme všechno zapsali, a pak jsme se začali ptát na letovém provozu, co to bylo za lety. [6]



Obrázek 5. Sledování příletů do Prahy pomocí nástroje FlightRadar24. [60]

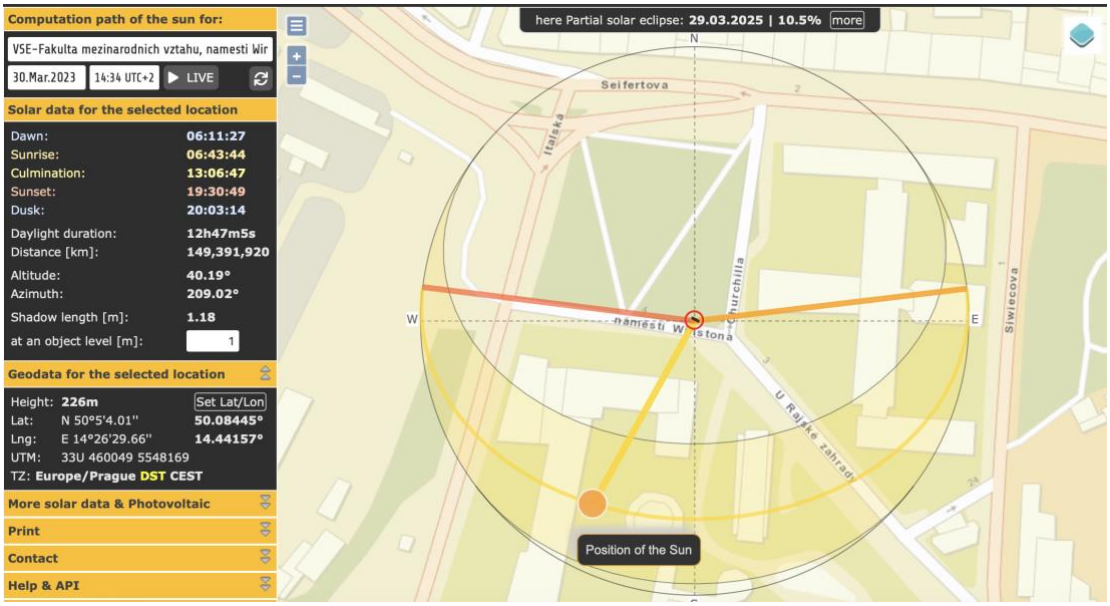
SunCalc

Suncalc[12] je webový nástroj, který poskytuje informace o poloze Slunce v libovolném místě a čase. Pomáhá vizualizovat údaje týkající se Slunce, například časy východu a západu Slunce, sluneční poledne, úhly azimutu a elevace Slunce a délku dne. SunCalc mohou používat například fotografové, krajináři a filmaři při plánování venkovních aktivit

[11] https://www.flightradar24.com
[12] https://www.suncalc.org/

a posuzování vlivu slunce na jejich práci. Nástroj zobrazuje informace v grafické podobě, což usnadňuje jejich pochopení a interpretaci.

Pomocí tohoto nástroje můžeme ale také určit či ověřit čas pořízení fotografie či videa. To může být občas velmi frustrujícím úkolem. Nabízí se řešení čas odhadnout pomocí úhlu a směru stínů, ne vždy je však ve videu či fotografii úhel těchto stínů jasný nebo tam nemusí být dostatek referenčních bodů pro přesné posouzení úhlu. SunCalc můžeme použít k potvrzení denní doby na fotografiích a videozáznamech a k určení, zda je v souladu s dalšími důkazy. Stejně tak můžeme také pomocí nástroje zjistit, zda se slunce v době události nacházelo na určitém místě, což může pomoci ověřit výpovědi svědků a určit, zda byla osoba nebo předmět ve stínu nebo na přímém slunci. [61]



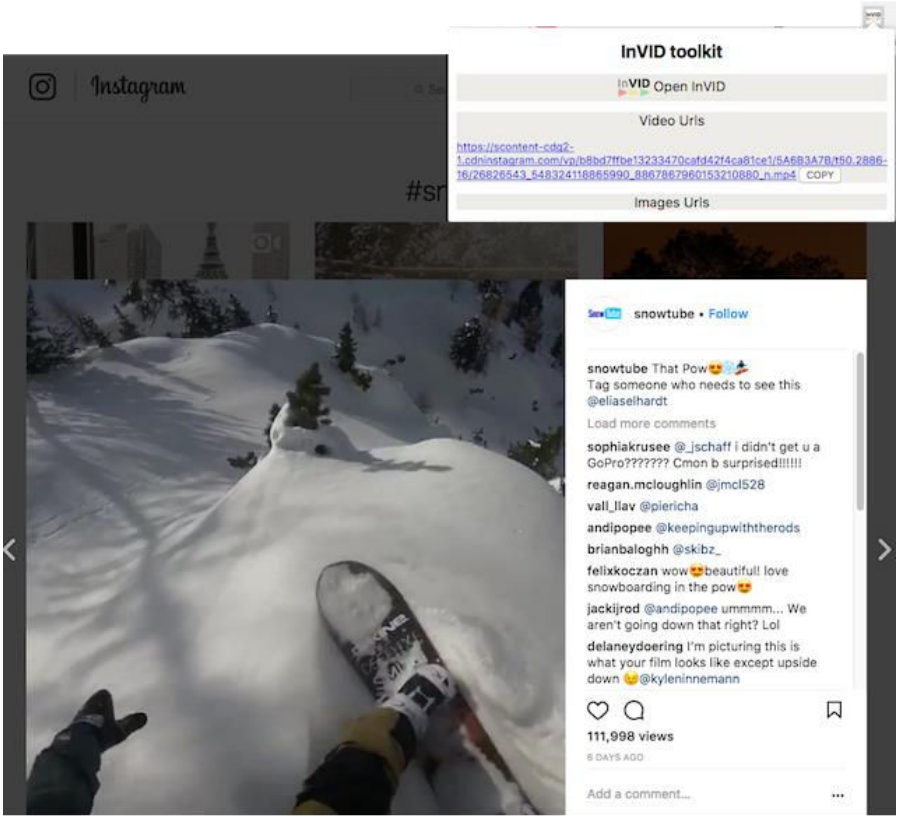
Obrázek 6. Zobrazení polohy Slunce a dalších údajů v nástroji SunCalc. [62]

InVID Verification

InVid Verification[13] je nástroj pro ověřování pravosti online videa a obrazového obsahu. Tuto sadu nástrojů poskytuje evropský projekt InVID, který pomáhá novinářům ověřovat obsah na sociálních sítích. Byl vytvořen s cílem ušetřit novinářům čas a zefektivnit jejich práci při ověřování faktů a vyvracení informací na sociálních sítích, zejména při ověřování videí a obrázků. Umožňuje uživatelům rychle a snadno zkontrolovat, zda video nebo obrázek nebyly zfalšovány, pozměněny nebo jakkoli zmanipulovány. Mezi nástroje, které poskytuje, patří zpětné vyhledávání obrázků ve vyhledávačích Google, Yandex nebo Baidu, fragmentování videí z různých platforem (Facebook, Instagram, YouTube, Twitter, Daily Motion) do klíčových snímků a analýza zvuku, které lze použít ke kontrole nesrovnalostí

[13] https://www.invid-project.eu/tools-and-services/invid-verification-plugin/

a anomálií v obsahu. Dále umožňuje zkoumat klíčové snímky a obrázky pomocí lupy, efektivněji se dotazovat na Twitteru pomocí časových intervalů a mnoha dalších filtrů. Lze je také použít ke kontrole přítomnosti vodoznaků, metadat a dalších ukazatelů pravosti. Nástroj často využívají právě novináři, výzkumní pracovníci a fact-checking týmy, aby pomohli ověřit informace a odhalit nepravdivý nebo zavádějící obsah na internetu. [63, 64]



Obrázek 7. Získání URL adresy videa na Instagramu pomocí InVID Verification. [65]

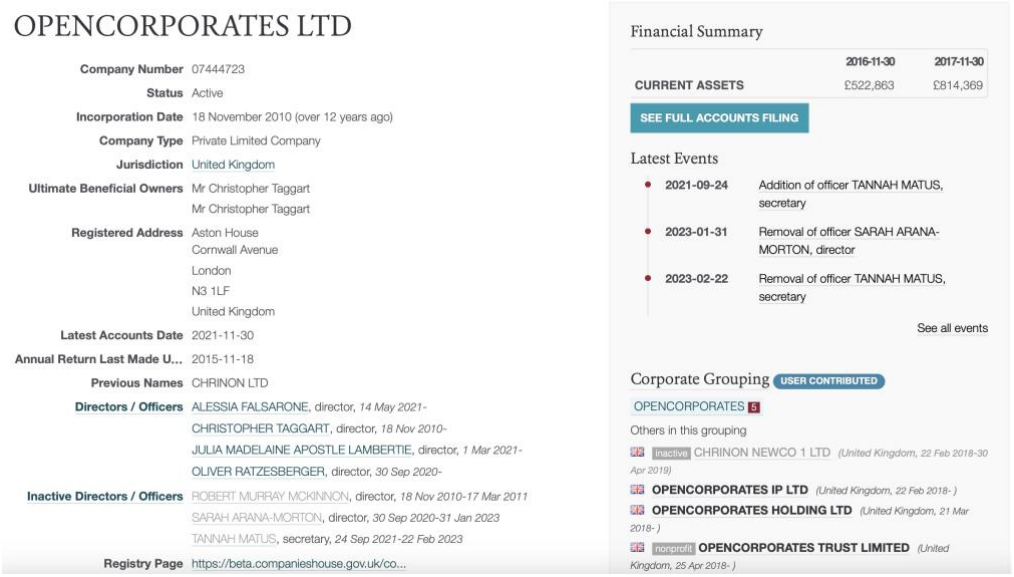
Open Corporates

Open Corporates[14] je největší a nejobsáhlejší otevřenou databází společností a údajů o společnostech na světě. Platforma vznikla v roce 2010 a sídlí ve Velké Británii. [66] OpenCorporates shromažďuje údaje z celé řady zdrojů, včetně rejstříků společností, vládních databází a dalších veřejných zdrojů. Veškerá data jsou shromažďovaná z primárních zdrojů, což činí databázi více než spolehlivou. Data jsou strukturovaná, aktuální a ověřená. Platforma v současné době obsahuje informace o více než 200 milionech společností, 140 jurisdikcí a denně ji využívá přes 7 milionů uživatelů. [67, 68]

Údaje na OpenCorporates zahrnují základní informace o společnostech, jako je název, adresa a registrační číslo, ale i podrobnější informace, jako jsou ředitelé společností, akcionáři a finanční informace. Uživatelé mohou vyhledávat společnosti podle názvu,

[14] https://opencorporates.com

umístění, odvětví a dalších kritérií. Tyto informace jsou užitečné zejména pro žurnalisty, výzkumníky a další odborníky, kteří potřebují prozkoumat společnosti a jejich aktivity. Rozsáhlá databáze informací o společnostech na platformě může pomoct odhalit potenciální střety zájmů, identifikovat skryté vztahy mezi společnostmi a jednotlivci a sledovat vlastnictví a kontrolu společností. Svou Offshore Leaks databázi ICIJ propojilo mimo jiné také právě s Open Corporates.



Obrázek 8. Informace o firmě OpenCorporates na platformě OpenCorporates. [66]

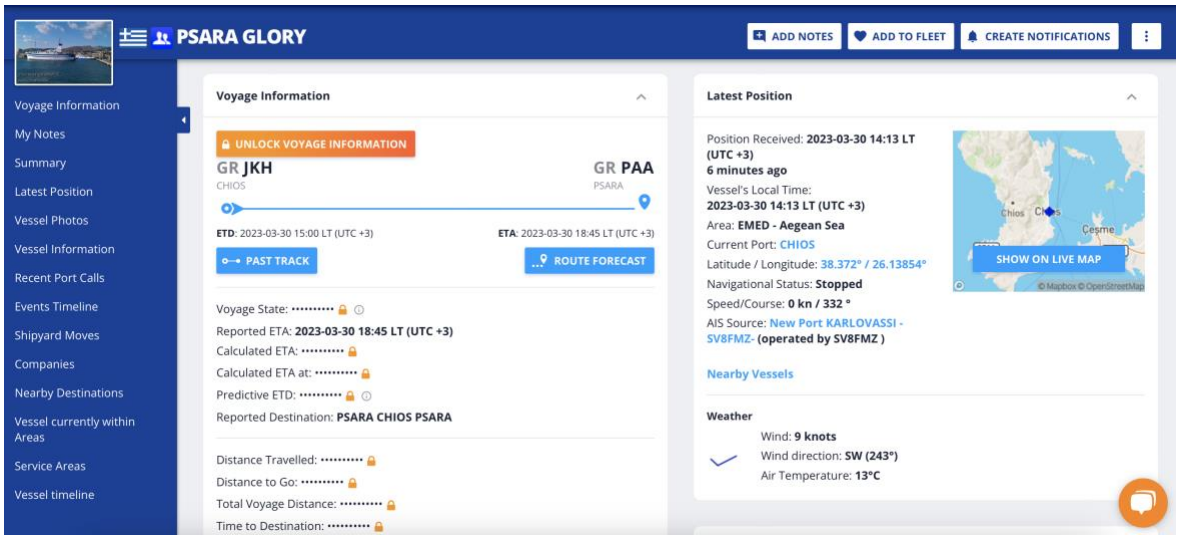
MarineTraffic

MarineTraffic[15] je webová a mobilní aplikace, která slouží pro sledování lodí a námořního zpravodajství v reálném čase. Poskytuje informace o pohybu lodí a plavidel po celém světě. Platforma byla spuštěna v roce 2007 a je všeobecně uznávána jako nejobsáhlejší námořní databáze na světě s více než 6 miliony uživateli měsíčně. [69]

MarineTraffic využívá síť pobřežních přijímačů Automatic Identification System (zkratkou AIS) ke sledování polohy lodí a dalších plavidel v reálném čase. AIS označuje sledovací systém, který se používá na lodích a plavidlech. Od roku 2004 musí být všechny osobní lodě jakékoliv velikosti a všechny mezinárodní lodě s hrubou prostorností 300 a více tun povinně vybaveny transpondérem AIS, který je schopen vysílat a přijímat údaje AIS. Transpondér AIS vysílá informace GPS a řadu dalších údajů, které přijímají přijímací stanice AIS. Služba MarineTraffic využívá více než 3 000 těchto stanic rozmístěných po celém světě. [69] Platforma také integruje údaje z dalších zdrojů, jako jsou satelitní snímky a údaje o lodích od přístavních úřadů, a poskytuje tak uživatelům komplexní přehled o námořní dopravě. Kromě toho některé informace o poloze lodí podávají sami lidé na palubě. Tyto informace

[15] https://www.marinetraffic.com

jsou pak použity k poskytnutí přesnější a aktuálnější živé mapy polohy plavidel. [69, 70] Uživatelé mohou vyhledávat lodě podle názvu, čísla IMO, čísla MMSI nebo volacího znaku a mohou si zobrazit řadu informací o každé lodi, včetně její aktuální polohy, rychlosti, kurzu, cíle a odhadovaného času příjezdu. Platforma rovněž poskytuje fotografie plavidel, podrobnosti o jejich majiteli a provozovateli a historické údaje o jejich pohybu.



Obrázek 9. Informace o lodi PSARA GLORY na platformě MarineTraffic. [71]

Dále také spolupracují s předními světovými přístavy, námořními společnostmi a hlavními ropnými společnostmi na projektech zaměřených na zvýšení efektivity a snížení dopadu na životní prostředí. [70] Jedná se o cenný nástroj pro každého, kdo se zajímá o sledování pohybu lodí a plavidel po celém světě, a pro profesionály, kteří potřebují být informováni o námořní dopravě pro obchodní nebo výzkumné účely. Služba je k dispozici k bezplatnému používání, existuje však i placená verze, která přidává pokročilejší funkce.

3.3 Klasifikace a zpracování dat

Na základě předchozích kapitol je zjevné, že při investigaci kauz získají novináři velké množství informací, které je třeba nějakým způsobem třídit, aby měli přehled o tom, jaké dokumenty a data našli, co obsahují a kde se nacházejí. Díky dobré organizaci podkladů mohou novináři potřebná data v případě potřeby ihned vyhledat a použít, uvést nové informace do kontextu s těmi již získanými a dokázali tak získat co největší detaily v rámci vyšetřování. Klasifikace pomáhá rozlišit, do jaké kategorie, tématu či množiny data patří a jak na sebe navazují časově.

Jelikož proces shromažďování dat byl již pokrytý v kapitole 3.1, předpokládejme, že již máme k dispozici materiály, které je nyní třeba roztřídit. Sestavíme si seznam otázek, na které chceme zodpovědět během analýzy, to nám pomůže při návrhu databáze. Získané dokumenty nejprve prozkoumáme, abychom posoudili jejich obsah. Podtrhneme či zvýrazníme pasáže, které se zdají být obzvlášť důležité a zjistíme, zda se v datech vyskytují nějaké opakující se vzory, které nám mohou napovědět, jak strukturovat naši databázi. Dokument označíme názvem, pokud ho ještě nemá, číslem nebo jiným způsobem, který

jsme si zvolili ke klasifikaci. Název může být jakýkoliv, pokud nám připomene, co obsahuje za informace. U rozhovorů je například vhodné použít jméno subjektu, v případě tajného zdroje můžeme použít krycí jméno. [3, 72]

Dokumenty po označení tagem, indexem či jiným způsobem založíme do složky dle vybrané klasifikace a seřadíme je dle vlastního uvážení a požadavků, nejčastěji se využívá abecední řazení. Tento postup můžeme aplikovat jak v případě elektronické databáze – rozřazování do složek – tak při řazení do fyzických papírových složek. Pro lepší přehlednost můžeme třídit dokumenty do různých předmětových složek, resp. spisů. V rámci těchto složek je řadíme ideálně chronologicky od nejnovějších dat po ty nejstarší.

Existují různé způsoby, jak data klasifikovat. Jde především o to, abychom vybrali takový způsob, který se nejvíce hodí na náš případ. Mezi jedny z nejběžnějších typů patří klasifikace založená na obsahu, která kontroluje a interpretuje soubory a hledá citlivé informace. Ty pak můžeme rozdělit následovně: [73, 74]

- **Veřejné dokumenty** jsou takové, které pocházejí z veřejných zdrojů. Jde o dostupný materiál, který je volně přístupný lidem ke čtení, zkoumání, prohlížení a ukládání.
- **Soukromé dokumenty** obsahují informace, které je rozumné chránit před veřejným přístupem, aby byla co nejlépe chráněna integrita informací a přístup k dalším datům jejich prostřednictvím. Mezi příklady soukromých údajů mohou patřit například osobní kontaktní údaje, jako jsou e-mailové adresy a telefonní čísla či údaje o výzkumu nebo historii procházení online.
- **Interní dokumenty** jsou striktně přístupné interním pracovníkům společnosti nebo interním zaměstnancům, kteří mají povolen přístup. Může se jednat o poznámky nebo jiná sdělení určená pouze pro interní účely, obchodní plány atd.
- **Citlivé dokumenty** představují údaje omezené na použití určitými osobami nebo skupinami a vyžadují zvláštní oprávnění. Příkladem citlivých údajů je například duševní vlastnictví.
- **Restriktivní dokumenty** jsou nejcitlivější z klasifikací dat. Často mají přísné bezpečnostní kontroly, které omezují množství osob s přístupem k datům. V případě narušení nebo kompromitace mohou omezená data představovat riziko pro veřejné zdraví nebo pro vlastnické informace společnosti či organizace. Příkladem jsou data chráněná dohodami o důvěrnosti, federální daňové informace či zdravotní údaje. Klasifikace citlivých a restriktivních dat se občas používá zaměnitelně.

Rozřazování můžeme provést také na základě typu dat – zda se jedná o zákaznická data, finanční údaje, duševní vlastnictví atd.

Dalším způsobem je klasifikace založená na kontextu, která využívá metadata a další informace o prostředí dokumentu k použití klasifikačních značek na data. Například dokumenty vytvořené určitým zaměstnancem nebo aplikací mohou být automaticky klasifikovány jako finanční data. [75]

V neposlední řadě je zde klasifikace založená na uživateli, která se spoléhá na úsudek znalého uživatele, v našem případě tedy žurnalisty, a závisí na ručním výběru každého

dokumentu danou osobou. Jeho přístup ke klasifikaci pak závisí na dané situaci a kauze, kterou vyšetřuje. [74]

Data máme přehledně označená a klasifikovaná, nyní je však klíčové vytvořit tzv. master souboru, lze také přeložit jako kmenový či hlavní soubor, ve kterém budou veškeré informace a podklady, které jsme získali. Tento hlavní soubor bude obsahovat hypotézu novinářova šetření, časovou osu událostí, informace a citace ze zdrojů, otázky, které budou dále směřovat jeho vyšetřování, a shrnutí jeho pozorování. [76] Cílem je mít veškeré informace, které možná do budoucna využijeme na jednom místě a v jedné podobě, a které jsou nějakým způsobem organizovány, nejčastěji v podobě databáze. Master soubor můžeme vytvořit buď jeden souhrnný pro všechny zdroje, nebo ho můžeme segmentovat do několika listů dle typů informací – například dokumenty, kontakty, osoby, rozhovory a podobně. Můžeme mít také zvlášť list, který chronologicky zobrazuje proběhlé události. V této kapitole je pojem databáze a master soubor užíván zaměnitelně a označuje to samé, jelikož se jedná prakticky o stejně strukturovaný produkt, jež se tvoří stejně.

Při strukturování databáze strukturuje žurnalista současně také svůj příběh. Jako příklad lze uvést databázi vytvořenou organizací Rutas del Conflicto, která zaznamenávala úmrtí spojená s protesty, dále také obvinění týkající se zneužívání moci, násilí a zadržování protestujících. Na základě shromažďování a ověřování informací z každodenních událostí, zpráv z tisku, nevládních organizací a rozhovorů se svědky a příbuznými obětí se týmu novináře Óscara Parry podařilo vytvořit ověřenou databázi násilí. To jim umožnilo zmapovat události a prozkoumat, kdo byly oběti, a zároveň analyzovat okolnosti jejich úmrtí a odhalit, že za mnoha incidenty stála policejní brutalita. [72]

Přestože se investigativní novináři mohou setkat s informacemi v nejrůznějších formátech – zprávy ve formátu PDF, chaotické papírové záznamy, rozhovory a pozorování žurnalistů, naskenované soubory, ručně psané dokumenty, staré archivy – se správnými odbornými znalostmi je lze všechny transformovat do zpracovatelných databází. [72] Tvorbu databáze nebo archivu lze realizovat pomocí papírových složek, elektronických dat nebo kombinací obojího.

Před tvorbou databáze bychom se měli seznámit s nasbíranými daty a dokumenty a zjistit, zda se v nich dají identifikovat nějaké opakující se vzory. Na základě toho se rozhodneme, jak databázi strukturovat a jak budeme případně segmentovat jednotlivé listy informací. Následně si musíme určit, co budou představovat jednotlivé záznamy, respektive řádky. Stejně tak si musíme určit, jakými prvky budou jednotlivé záznamy identifikovány – to budou naše sloupce. To znamená, že pokud tvoříme například databázi se seznamem osob, každý řádek bude obsahovat jednu osobu, a ta je identifikována například jménem, příjmením, věkem, profesí, adresou a tak dále. [72]

Každému záznamu je nutné přidělit unikátní identifikační klíč, pomocí kterého můžeme odkazovat na informace mezi soubory, například pokud máme tabulek či souborů více, ale i v rámci jednoho souboru. Kromě identifikačního pole by nemělo chybět pole pro poznámky či doplnění detailu a v neposlední řadě pole pro klasifikaci daného záznamu, ať už v podobě indexace, tagování, anotace či jiným označením, o němž píšu výše. Mezi další důležitá pole patří identifikace autora, který záznam přidal, v případě, že na databázi spolupracuje více novinářů. Nicméně i přesto, že novinář zpočátku na projektu pracuje sám,

měl by brát v potaz možnost, že se jeho návrh může rozrůst do budoucího velkého projektu, a tvorbu databáze by měl realizovat tak, aby byla škálovatelná. [72]

Při tvorbě bychom se také měli snažit o jednotnou formu databáze, například datum by se měl zadávat vždy ve stejném formátu, čísla by tak měla být označená a kategorie by měly být vždy napsány stejně, aby se nám v ní lépe vyhledávalo. [76]

Po zhotovení návrhu je na místě otestovat vzniklou databázi vyplněním několika záznamů, abychom si vyzkoušeli, zda funguje tak, jak očekáváme. Dále bychom měli zhodnotit spolehlivost a konzistenci nejdůležitějších polí, případně zda nám chybí některá pole pro doplnění informací. Funkční databázi poté začneme plnit záznamy. Pokud do databáze bude přidávat záznamy více uživatelů, je nutné všechny se strukturou databáze seznámit a poskytnout jim základní školení, aby všichni chápali pojmy a kategorie stejným způsobem a nedošlo ke zmatenému vyplňování, což by ztížilo vyhledávání a orientaci v záznamech.

Vytvoření databáze je pouze prvním krokem šetření. Před analýzou dat a vyvozením závěrů je třeba je potvrdit u původních zdrojů, ať už jde o dokumenty, nebo o hlavní postavy příběhů. Dále musíme také provést audit dat. To, jakým způsobem audit provedeme závisí na velikosti naší databáze a rozsahu projektu. Při auditu bychom se měli zaměřit na překlepy, čísla, data, duplicity a záznamy, které nesplňují kritéria, například nepatří do dané kategorie. Můžeme zkontrolovat každý jednotlivý záznam křížovým porovnáním s původními dokumenty nebo můžeme provést náhodné namátkové kontroly, které by však měly pokrýt značný počet záznamů v databázi. V obou případech by osoba, která údaje kontroluje, neměla být osobou, která je zadala. Databáze nebude připravena k použití, dokud neprojde jak kontrolou faktů, auditem dat, konfrontací s osobními zdroji, tak právní kontrolou. [72]

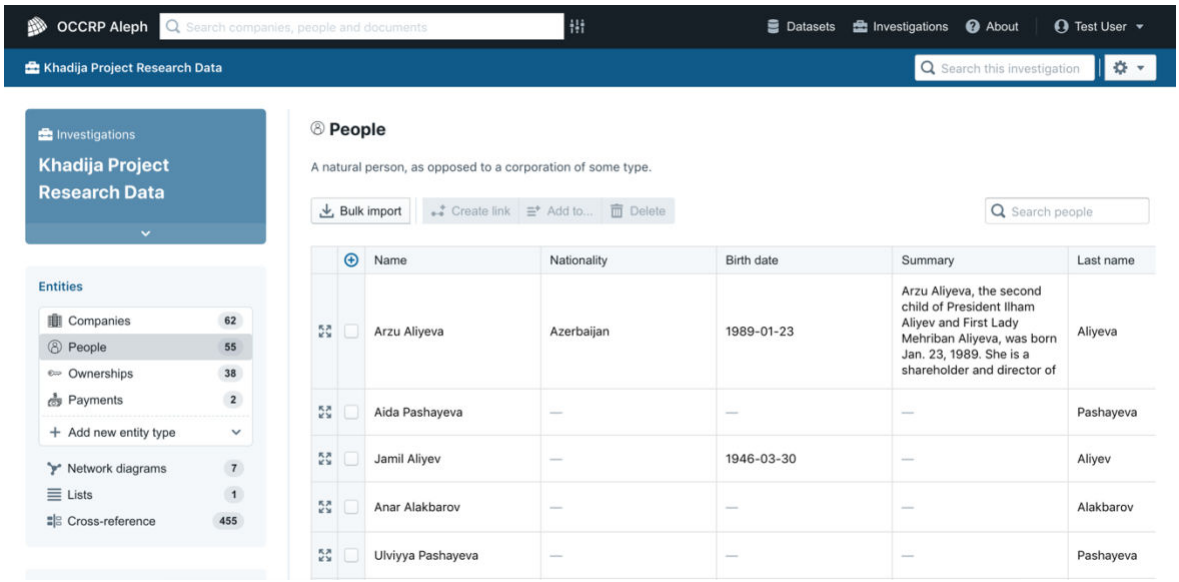
3.3.1 Platforma Aleph

Existuje mnoho nástrojů, které mohou usnadnit proces klasifikace a třídění dat, jako jsou nástroje pro skenování s OCR (Optical Character Recognition), resp. optické rozeznávání znaků, online formuláře pro vyplňování databáze, konvertory PDF či nástroje pro scraping a zpracování velkého množství textu. Novinář nemusí být současně i vývojářem, aby tvorbu databáze realizoval. Stačí spolupracovat s někým, kdo tyto schopnosti má, případně může využít již existující nástroje. OCCRP (Organized Crime and Corruption Reporting Project) vytvořili platformu Aleph, která zvládá téměř všechny tyto úkoly. OCCRP je organizace, která sdružuje investigativní novináře, jenž se zabývají korupcí, praním špinavých peněz a dalšími druhy organizovaného zločinu. Projekt zveřejňuje články investigativních novinářů s cílem šířit informace o trestných činech a postavit zúčastněné osoby před soud. Platforma Aleph slouží k tomu, aby usnadňovala výzkum, vyšetřování a konkurenční zpravodajství tím, že uživatelům poskytuje přizpůsobitelný pracovní prostor, kam mohou nahrávat, organizovat a analyzovat svá data. Platforma podporuje širokou škálu zdrojů dat, včetně veřejných záznamů, sociálních médií, zpravodajských článků a další. Uživatelé mohou nahrávat i svá vlastní data a také přistupovat k různým datovým souborům třetích stran. Jakmile jsou data do Alephu importována, lze je označit, klasifikovat a analyzovat pomocí různých nástrojů a funkcí. Platforma obsahuje pokročilé funkce vyhledávání a filtrování a také nástroje pro vizualizaci dat, které uživatelům umožňují vytvářet grafy,

diagramy a síťové diagramy. Jednou z klíčových předností systému Aleph jsou funkce pro spolupráci, což je nejen v investigativní žurnalistice mnohdy nezbytné. Uživatelé mohou pozvat členy týmu, aby se připojili k jejich pracovním prostorům a společně pracovali na šetřeních a výzkumných projektech. [77] Celkově je Aleph výkonným nástrojem pro žurnalisty a organizace, které potřebují shromažďovat a analyzovat velké množství dat z různých zdrojů.

Na vývoji platformy se mimo jiné podílela také Pavla Holcová a nástroj tak využívají i novináři v investigaci.cz.

Většinu dat, která získáme, se snažíme indexovat a cross-checkovat s ostatními dokumenty, které už máme k dispozici. Používáme k tomu softwarový nástroj, který jsme sami vyvinuli a jmenuje se Aleph. Zbytek děláme ručně. Do budoucna chceme některé tyto procesy automatizovat s pomocí AI. – Pavla Holcová z investigace.cz (viz Příloha B)



Obrázek 10. Ukázka pracovního prostoru platformy Aleph. [77]

K výše zmíněným účelům, jako je tvorba master souboru či databáze, filtrování a třídění dat či zpracování a analýzy dat však může posloužit také obyčejný Microsoft Excel či Google Sheets. Pro čištění a transformaci dat lze použít například open-source nástroj OpenRefine. [16] Pro vizualizaci dat lze využít platformu Tableau, která umožňuje vytvářet interaktivní grafy, mapy a diagramy. Dalším nástrojem pro vizualizaci dat a analýzu vazeb je například Maltego [17], který lze použít k analýze vztahů a vazeb mezi lidmi, organizacemi a dalšími subjekty. Je obzvláště užitečný pro vyhledávání skrytých nebo nejasných vazeb ve velkých souborech dat. Podobných nástrojů je nespočet a záleží pouze na konkrétním

[16] https://openrefine.org
[17] https://www.maltego.com

novináři či organizaci, který zvolí při svém vyšetřování, jelikož se budou lišit na základě typů analyzovaných dat a povaze jednotlivých případů.

3.4 Uchovávání dat

Předchozí kapitola se může s touto v některých aspektech prolínat, což je pochopitelné, neboť proces práce s daty probíhá průběžně a jednotlivé kroky navazují na sebe a probíhají víceméně souběžně, jelikož existují samozřejmě nástroje umožňující všechny činnosti na jednom místě. Investigativní novináři mohou k ukládání dokumentů a dat používat různé nástroje a metody v závislosti na svých specifických potřebách a preferencích.

První možností je lokální úložiště, kdy novináři ukládají dokumenty u sebe na počítači či na externím disku. Lokální úložiště můžeme chápat jako fyzický archiv (papírové kopie, disky či jiné fyzické médium, jako je záznamové zařízení) i digitální archiv (již zmíněný externí disk či pevný disk). Externí pevné disky lze pro větší bezpečnost šifrovat a lze je snadno přenášet mezi jednotlivými místy. Tento přístup může zajistit větší kontrolu a zabezpečení dat, ale může být také náchylnější ke ztrátě nebo krádeži.

Naopak další možností jsou cloudová úložiště, která mohou být využitá jak k ukládání dokumentů a dat, tak i sdílení se spolupracovníky. Mezi nejznámější cloudové služby patří například Google Drive, Dropbox nebo Microsoft OneDrive. Tyto služby poskytují pohodlný způsob přístupu k souborům odkudkoli s připojením k internetu a často nabízejí funkce, jako je šifrování a kontrola verzí, které chrání data.

Investigativní žurnalisté mohou používat také systémy pro správu dokumentů, jako je již zmíněný Aleph, DocumentCloud nebo Hederis, k ukládání a organizaci velkého množství dokumentů, což usnadňuje jejich vyhledávání a přístup k nim. Některé nástroje lze stáhnout a používat lokálně, jiné zas fungují na cloudové bázi a jsou poskytovatelem hostované na jejich vlastním serveru. Pro přepis, uložení a správu poznámek v elektronické podobě můžou posloužit nástroje, jako je Evernote, OneNote či Notion.

Velké organizace mohou používat databázové systémy a datové sklady k ukládání a správě velkého množství dat. Investigativní novináři, kteří mají přístup k těmto skladům, mohou používat nástroje pro analýzu dat a získávat z nich poznatky a vzorce. Příkladem takových databázových systémů je například MySQL, MongoDB, Firebase nebo PostgreSQL.

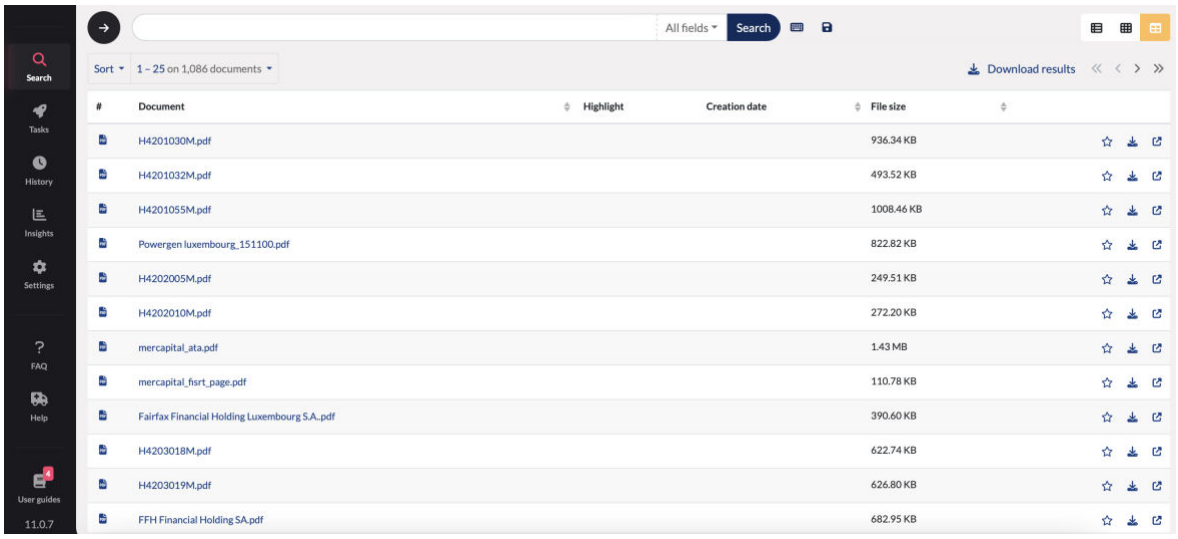
Ať už novinář či organizace zvolí jakýkoliv způsob uložení dat, ideální je kombinovat více způsobů a data zálohovat na několika místech. Ve většině případů se v investigativní žurnalistice jedná o citlivá data a bylo by nepříjemné o ně přijít. Data je tedy nutné zálohovat, ideálně víckrát a na více místech. O zálohování dat píšu více v kapitole 3.5, neboť souvisí s ochranou dat.

3.4.1 Datashare

Datashare [18] je bezplatná open-source aplikace vyvinutá ICIJ, která umožňuje novinářům spolupracovat na rozsáhlých investigativních projektech a sdílet mezi sebou citlivé údaje při zachování důvěrnosti a bezpečnosti. Tento nástroj by se dal zařadit jak do kategorie klasifikace, uchovávání, tak ochrany dat, jelikož nabízí možnost indexace a vyhledávání a filtrování v datech, ale také slouží jako úložiště dokumentů s možností kolaborace. Jelikož je nástroj vyvinutý přímo pro investigativní žurnalisty, je zde kladen velký důraz i na bezpečnost.

Platforma umožňuje investigativním novinářům nahrávat a sdílet dokumenty, tabulky a další data a poskytuje nástroje pro analýzu dat, vizualizaci a spolupráci a zároveň je dokáže zabezpečit před potenciálními zásahy třetích stran. Dále pak nabízí možnost současného prohledávání dokumentů ve formátu PDF, obrázky, texty, tabulky, diapozitivy a mnoho dalších a během toho automaticky detekovat a filtrovat data podle osob, organizací a míst. [78] Užívání platformy je bezpečné, jelikož ji lze používat na místním počítači a v případě potřeby ji lze používat i offline. Žádná data v rámci Datashare nejsou předávána ICIJ ani žádné jiné třetí straně. To znamená, že riziko zachycení dat je výrazně sníženo, což je důležité zejména při práci s citlivými dokumenty. Vzhledem k tomu, že Datashare není online služba pro extrakci textu, je riziko zachycení výrazně omezeno. [79]

Pomocí tohoto nástroje mohou novináři spolupracovat efektivněji a účinněji a mohou vzájemně sdílet informace, aniž by byla ohrožena bezpečnost nebo důvěrnost dat. To může být užitečné zejména u investigativních projektů, které zahrnují velké množství dat nebo více zdrojů, protože to novinářům umožňuje spojit své zdroje a efektivněji spolupracovat.



Obrázek 11. Snímek obrazovky pracovního prostoru platformy Datashare. Zdroj: autor

[18] https://datashare.icij.org

Od roku 2020 spolupracuje ICIJ se švýcarskou univerzitou EPFL na vývoji bezpečného DataShareNetwork systému, který umožní uživatelům Datashare bezpečně sdílet data s dalšími důvěryhodnými uživateli nebo organizacemi prostřednictvím privátní sítě. Platforma by měla usnadnit novinářům podněcování nových, spontánních nebo ad-hoc projektů a spolupráce tím, že kombinuje decentralizovaný vyhledávač se zabezpečeným systémem zpráv, který mohou novináři používat k bezpečnému a anonymnímu vyhledávání a výměně informací. Základem systému je samozřejmě ochrana soukromí. Novináři budou moci vyhledávat a vyměňovat si informace anonymně. ICIJ plánuje vydávat svým členům virtuální tokeny, které mohou připojovat ke zprávám jako důkaz jejich pravosti. ICIJ však nebude poskytovat žádné informace, ani nebude vědět, co novináři hledají. Tento decentralizovaný přístup minimalizuje riziko hackerských útoků, protože dokumenty zůstanou na jednotlivých počítačích, nikoli na centrálním serveru. Přístup k informacím bude navíc omezen na konkrétní vyhledávací dotazy a bude prováděn na základě individuálního přístupu. Pokud vyhledávání přinese relevantní výsledky, může novinář navázat kontakt přímo s anonymním kolegou, který informace sdílel. Vyhledávací dotazy budou zašifrované a bude na novináři, který má dokumenty k dispozici, aby se rozhodl, zda se bude případným dotazem zabývat. [80] V současné době nejsou žádné informace ohledně data spuštění platformy, v roce 2021 měla vzniknout testovací verze systému, lze tedy předpokládat, že je projekt stále ve fázi vývoje a testování.

3.5 Ochrana dat

Bezpečnost dat je v investigativní žurnalistice zásadním faktorem, zejména pokud se pracuje s citlivými nebo důvěrnými informacemi. Investigativní novináři často pracují s velkým množstvím údajů, včetně osobních a finančních informací, které by mohly být škodlivé nebo dokonce nebezpečné, kdyby se dostaly do nesprávných rukou. Stejně tak by bylo velice nepříjemné, kdyby o dané informace přišli. Data mohou být kdykoliv ztracena v důsledku technických problémů, jako jsou selhání disků nebo havárie počítačů. Proto je důležité přijmout vhodná opatření na ochranu bezpečnosti a soukromí údajů a zdrojů, které je poskytly a ke všemu mít záložní kopie.

Nejprve se zaměříme na ochranu dat před ztrátou. Toho docílíme již zmíněným zálohováním. Vzhledem k tomu, že digitální soubory jsou tak snadno replikovatelné, je jednoduché vytvořit několik kopií, které nám mohou pomoct v případě ztráty dat. Zásadní je uložit alespoň jednu kopii dokumentu či celého pevného disku na jiné fyzické místo – ideálně externí disk, a ten uložit na jiné fyzické místo, než je originál. U souborů, které nejsou citlivé, lze zvolit možnost uložení kopie na cloudové úložiště. Pokud se jedná o zabezpečený cloudový systém vyvinutý přímo pro investigativní žurnalisty (např. Aleph), můžeme nahrát všechny dostupné materiály na jedno místo. Zálohu bychom si měli vytvořit i u fyzických materiálů, ke každému dokumentu mít alespoň jednu kopii a tu uložit, stejně jako v případě zálohy disků, na jiné místo, abychom nepřišli o všechna data například v případě povodně, hurikánu či jiné katastrofy. [81]

Dále jsou uvedeny některé běžné postupy zabezpečení dat používané v investigativní žurnalistice za účelem ochrany soukromí.

Prvním je zabezpečená komunikace. Investigativní novináři často komunikují se zdroji a kolegy prostřednictvím zabezpečených komunikačních kanálů, jako jsou aplikace pro šifrované zprávy nebo zabezpečené e-mailové služby. Novináři se při své každodenní práci často spoléhají na telefonní hovory a digitální zprávy jako na primární komunikační prostředky. Používání běžných komunikačních kanálů však může způsobit, že obsah těchto konverzací bude zranitelný vůči hackerům. I když samotný obsah není zachycen, hacker může mít přístup k souvisejícím metadatům, jako jsou účastníci a načasování konverzace. Naštěstí je nyní k dispozici mnoho možností, které novinářům pomáhají komunikovat bezpečně a s větší mírou jistoty. Jasným favoritem pro zabezpečené hlasové hovory a zasílání zpráv mezi novináři, jejich redaktory a někdy i jejich zdroji je v současné době aplikace Signal, která nabízí tzv. end-to-end šifrování, což znamená, že komunikaci lze dešifrovat pouze na fyzických zařízeních komunikujících uživatelů. I kdyby se vláda pokusila přimět skupinu vývojářů, která spravuje službu, k předání obsahu komunikace, nemohla by informace poskytnout: Signal jednoduše nemá možnost zjistit, co přesně na jeho platformě uživatelé dělají. End-to-end šifrování využívá stále více platforem, jako je například WhatsApp či Wire, některé z těchto platforem se však od služby Signal liší v jednom klíčovém ohledu: byť tyto platformy nemají přístup k obsahu uživatelovy komunikace, často mají přístup k cenným metadatům, která mohou odhalit, s kým a kdy uživatel naposledy komunikoval. [82]

Novináři by si měli být vědomi nebezpečí digitálních útoků, včetně hackerských útoků, phishingu a sledování. V rámci zabezpečené komunikace bychom si tedy měli také dát pozor a zkoumat pečlivě legitimitu kontaktních údajů odesílatelů. Správce hesel nikdy nevyplní heslo na phishingové stránce. Je vhodné zachovat si zdravou míru podezřívavosti vůči zprávám, a to i těm, které se zdají být ze známých zdrojů. [83] Novináři by si měli být vědomi různých taktik podvodných e-mailů a telefonátů, aby byli schopni phishingový útok odhalit včas.

Abychom dále minimalizovali riziko útoků, měli bychom si zabezpečit své účty pomocí silných hesel. Novináři by měli pro všechny své účty používat silná a jedinečná hesla, aby se zabránilo neoprávněnému přístupu. K vytváření a ukládání silných hesel lze použít správce hesel, jako je LastPass, 1Password nebo KeePass. Pro větší bezpečnost by měl žurnalista používat dvoufaktorové ověřování, tzv. 2FA (angl. two-factor authentication). Funkce 2FA chrání uživatelovy účty i v případě, že se někdo dozví jeho heslo. Mezi možnosti patří fyzický bezpečnostní klíč, který nelze podvrhnout, aplikace pro generování kódů v telefonu a kódy zasílané prostřednictvím SMS/e-mailu. [83]

Z výše uvedených informací je jasné, že je otázka bezpečnosti dat pro investigativní novináře důležitým aspektem při jejich práci. Novináři z týmu Pavly Holcové získávají několik školení ročně ohledně digitální bezpečnosti a měsíčně dostávají informace o nově odhalených bezpečnostních rizicích, kterými se řídí. S citlivými dokumenty pracují v databázích na dedikovaných serverech a platformách. Pro ty nejcitlivější dokumenty pak používají tzv. air-gapped počítač, na kterém dokumenty procházejí a vůbec s nimi nepracují online (viz Příloha B). „Air-gapped“ počítač je izolován od všech ostatních sítí, což znamená, že není přímo připojen k internetu ani k žádnému jinému systému, který je připojen k internetu. Tato izolace má zabránit jakémukoli neoprávněnému přístupu nebo přenosu dat do počítače nebo z počítače. [84]

4 Tvorba interaktivní databáze offshore společností

4.1 ICIJ Offshore Leaks Database

V současné době již existuje v jisté podobě databáze offshore společností na webu Mezinárodního konsorcia investigativních žurnalistů. Vyhledávání v databázi probíhá prostřednictvím vyhledávacího pole, kde lze vyhledávat podle jmen nebo klíčových slov, hledání spojení pro jednotlivé země nebo procházení offshore subjektů podle jurisdikce. Výsledky se pak zobrazí jako seznam odkazů a po zvolení konkrétního subjektu se zobrazí nové okno, jehož součástí je také vizualizace vztahů entit. Tento způsob zobrazení výsledků může pro některé být matoucí a nutnost otevírat nová okna pro různé subjekty není příliš uživatelsky přívětivé. Cílem mého projektu je vytvořit mapu zobrazující přehled offshore firem, který umožňuje zobrazení informací k různým subjektům v jednom dashboardu bez nutnosti překlikávat mezi několika okny a odkazy a současně umožňuje filtrování jak prostřednictvím interakce s mapou, tak ze seznamu filtrů. Díky vizuálnímu zobrazení dat je uživatel schopen vstřebat a pochopit informace a souvislosti mnohem rychleji a jednodušeji.

4.2 Seznámení s datovým zdrojem

Data, se kterými budu pracovat pocházejí z webu ICIJ [19], která svoji databázi převedla do několika souborů, tzv. uzlů a vztahů mezi nimi, aby k datům měli přístup všichni bez ohledu na technické prostředky. Jedná se o zveřejněný souhrn uniklých dokumentů týkající se offshorových firem z pěti souborů, a to Offshore Leaks, Panama Papers, Bahamas Leaks, Paradise Papers a Pandora Papers.

Datový zdroj se skládá ze šesti souborů ve formátu .csv. Obsahuje data týkající se více než 810 tisíc offshore společností a trustů a odkazuje na osoby a společnosti ve více než 200 zemích a teritoriích. Soubory jsou pojmenovány následovně:

- nodes-addresses.csv
- nodes-entities.csv
- nodes-intermediaries.csv
- nodes-officers.csv
- nodes-others.csv

[19] Odkaz na zdroj dat: https://offshoreleaks.icij.org/pages/database

- relationships.csv

Jak jsem již naznačila, datový zdroj je rozdělen, a tedy i pojmenován podle několika hlavních typů uzlů, které se v datech vyskytují:

- **Entity** – offshore právnická osoba. Může to být společnost, trust, nadace nebo jiná právnická osoba vytvořená v zemi s nízkým zdaněním, tzv. daňový ráj.
- **Address** – Adresa sídla, jak je uvedena v původních databázích, které získala ICIJ.
- **Intermediaries** – Zprostředkovatel mezi zájemcem o offshore společnost a poskytovatelem offshore služeb – obvykle advokátní kancelář, banka nebo prostředník, který požádá poskytovatele offshore služeb o vytvoření offshore společnosti.
- **Officers** – Osoba nebo společnost, která hraje v offshore subjektu určitou roli, například beneficienta, ředitele nebo akcionáře.
- **Others** – Další subjekty nalezené v datech.
- **Relationships** – Zobrazuje vztahy mezi jednotlivými uzly.

Níže je detailní popis atributů jednotlivých souborů – uzlů.

4.2.1 Entity | nodes-entities.csv

814 344 řádků × 21 sloupců

Tabulka 2. Popis sloupců v tabulce Entity. Zdroj: autor

atribut	popis	příklad
node_id	unikátní identifikátor přiřazený každé entitě v databázi	10000001
name	aktuální název entity/subjektu	8808 HOLDING LIMITED
original_name	původní název entity, pokud byl změněn	8808 HOLDING LIMITED (EX-DIAMOND LIMITED)
former_name	veškeré dřívější názvy, které entita používala	DIAMOND LIMITED
jurisdiction	země nebo území, kde je entita zaregistrována	SAM
jurisdiction_description	popis jurisdikce, v níž je entita zapsána, včetně informací o jejích zákonech a předpisech	Samoa
company_type	typ subjektu, například společnost, trust nebo nadace	International Trust
address	registrované sídlo entity	ORION HOUSE SERVICES (HK) LIMITED ROOM 1401;

		14/F.; WORLD COMMERCE CENTRE; HARBOUR CITY; 7-11 CANTON ROAD; TSIM SHA TSUI; KOWLOON; HONG KONG
internal_id	interní identifikátor přidělený subjektu poskytovatelem údajů	1000916
incorporation_date	datum, kdy byl subjekt založen nebo zaregistrován	05-JAN-2006
inactivation_date	datum, kdy klient sdělil agentovi, aby deaktivoval offshore subjekt, který může být později znovu aktivován	18-FEB-2013
struck_off_date	datum, kdy byl subjekt vyřazen z úředního rejstříku (pokud neplatí licenční poplatky do obchodního rejstříku)	15-FEB-2013
dorm_date	datum, kdy byl subjekt uveden do stavu nečinnosti	18-FEB-2013
status	aktuální stav subjektu, například aktivní nebo zrušený	Active
service_provider	název poskytovatele služeb, který pomohl subjekt založit a spravovat	Mossack Fonseca
ibcRUC	identifikátor používaný některými jurisdikcemi ke sledování subjektů	25249
country_codes	kódy ISO 3166-1 alpha-2 pro země, kde má subjekt spojení	HKG
countries	názvy zemí, ve kterých má subjekt spojení	Hong Kong
sourceID	soubor uniklých dokumentů, ze kterých záznam pochází	Panama Papers
valid_until	datum, do kterého jsou údaje o entitě platné	The Panama Papers data is current through 2015

note	jakékoli další poznámky nebo komentáře k entitě	Holt Global Limited was dissolved on December 29th 2016, based on documents provided by a representative of the company (Updated on June 28th, 2021

4.2.2 Address | nodes-addresses.csv

402 246 řádků × 8 sloupců

Tabulka 3. Popis sloupců v tabulce Address. Zdroj: autor

atribut	popis	příklad
node_id	unikátní identifikátor přiřazený každé entitě v databázi	24000001
address	adresa offshore subjektu nebo zprostředkovatele, jak je uvedena v uniklých dokumentech	KABANBAI BATYR STREET, 122A, APT. 3, ALMATY, 050000, KAZAKHSTAN
name	také obsahuje adresu, stejně jako sloupec address	6B Chenyu Court; 22-24 Kennedy Road; Hong Kong
countries	země, v nichž se adresa nachází nebo v nichž je offshore subjekt či zprostředkovatel registrován či působí, jak je uvedeno v uniklých dokumentech	Bahamas
country_codes	kódy zemí ISO 3166-1 alpha-2 pro země uvedené v atributu "countries	BHS
sourceID	soubor uniklých dokumentů, ze kterých záznam pochází	Bahamas Leaks
valid_until	datum, do kterého jsou údaje platné	The Bahamas Leaks data is current through early 2016
note	jakékoli další poznámky nebo komentáře	In August 2022, ICIJ was provided with an official document showing that another business is currently linked to this address.

4.2.3 Intermediaries | nodes-intermediaries.csv

26 768 řádků × 10 sloupců

Tabulka 4. Popis sloupců v tabulce Intermediaries. Zdroj: autor

atribut	popis	příklad
node_id	unikátní identifikátor přiřazený každé entitě v databázi	11000001
name	název zprostředkovatele, jak je uveden v uniklých dokumentech	MICHAEL PAPAGEORGE, MR.
status	aktuální stav subjektu, například aktivní nebo zrušený	ACTIVE
internal_id	interní identifikátor přidělený subjektu poskytovatelem údajů	10001
address	adresa zprostředkovatele, jak je uvedena v uniklých dokumentech	MICHAEL PAPAGEORGE; MR. 106 NICHOLSON STREET BROOKLYN PRETORIA 0002; GAUTENG (PWV) SOUTH AFRICA
countries	země, ve kterých je zprostředkovatel registrován nebo působí, jak je uvedeno v uniklých dokumentech	South Africa
country_codes	kódy zemí ISO 3166-1 alpha-2 pro země uvedené v atributu "countries"	ZAF
sourceID	soubor uniklých dokumentů, ze kterých záznam pochází	Panama Papers
valid_until	datum, do kterého jsou údaje platné	The Panama Papers data is current through 2015
note	jakékoli další poznámky nebo komentáře	Not all beneficiaries are aware of offshore trusts in which they are named because the

		settlor may select beneficiaries without their knowledge.

4.2.4 Officers | nodes-officers.csv

771 315 řádků × 7 sloupců

Tabulka 5. Popis sloupců v tabulce Officers. Zdroj: autor

atribut	popis	příklad
node_id	unikátní identifikátor přiřazený každé entitě v databázi	50135
name	jméno úředníka, jak je uvedeno v uniklých dokumentech	THE BEARER
countries	země, v nichž se úředník nachází nebo působí, jak je uvedeno v uniklých dokumentech	Hong Kong
country_codes	kódy zemí ISO 3166-1 alpha-2 pro země uvedené v atributu "countries"	HKG
sourceID	soubor uniklých dokumentů, ze kterých záznam pochází	Offshore Leaks
valid_until	datum, do kterého jsou údaje platné	The Offshore Leaks data is current through 2010
note	jakékoli další poznámky nebo komentáře	ICIJ previously published the names and addresses of officers of Ploomo Ltd based on data contained in the Maltese Business Registry. ICIJ subsequently learned that this data was incorrect, and that the listed officers were not associated with the company. ICIJ therefore removed the officers' names and addresses from the database (7th July 2022)

4.2.5 Others | nodes-others.csv

2 989 řádků × 13 sloupců

Tabulka 6. Popis sloupců v tabulce Others. Zdroj: autor

atribut	popis	příklad
node_id	unikátní identifikátor přiřazený každé entitě v databázi	85004929
name	aktuální název entity/subjektu	ANTAM ENTERPRISES N.V.
type	typ subjektu, například společnost, trust nebo nadace	LIMITED LIABILITY COMPANY
incorporation_date	datum, kdy byl subjekt založen nebo zaregistrován	18-MAY-1983
struck_off_date	datum, kdy byl subjekt vyřazen z úředního rejstříku	31-DEC-2002
closed_date	datum, kdy byla společnost, svěřenský fond nebo nadace oficiálně uzavřena nebo zrušena	28-NOV-2012
jurisdiction	země nebo území, kde je entita zaregistrována	AW
jurisdiction_description	popis jurisdikce, v níž je entita zapsána, včetně informací o jejích zákonech a předpisech	Aruba
countries	země, v nichž se úředník nachází nebo působí, jak je uvedeno v uniklých dokumentech	Hong Kong
country_codes	kódy zemí ISO 3166-1 alpha-2 pro země uvedené v atributu "countries"	HKG
sourceID	soubor uniklých dokumentů, ze kterých záznam pochází	Paradise Papers - Aruba corporate registry
valid_until	datum, do kterého jsou údaje platné	Aruba corporate registry data is current through 2016

note	jakékoli další poznámky nebo komentáře	Closed date stands for Cancelled date

4.2.6 Relationships | relationships.csv

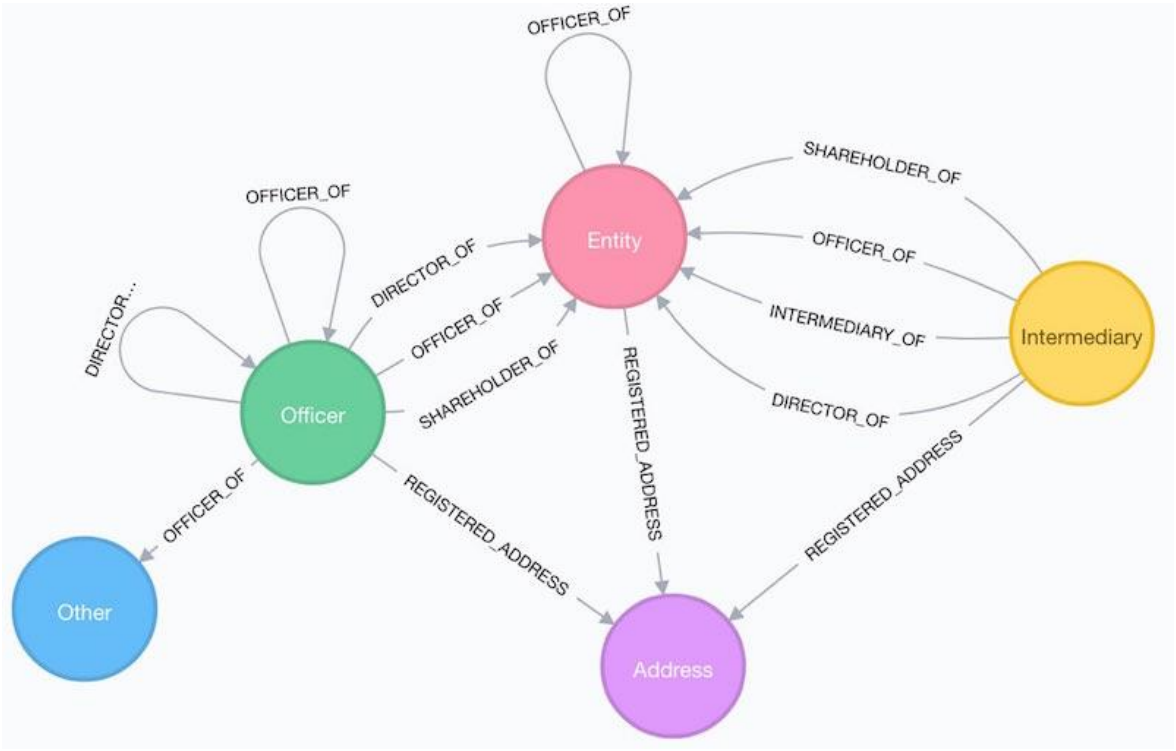
3 336 913 řádků × 8 sloupců

Tabulka 7. Popis sloupců v tabulce Relationships. Zdroj: autor

atribut	popis	příklad
node_id_start	ID uzlu počáteční entity nebo prostředníka ve vztahu	12160432
node_id_end	ID uzlu koncového subjektu nebo prostředníka ve vztahu	10000001
rel_type	typ vztahu mezi dvěma uzly, například "ředitel" nebo "akcionář"	officer_of
link	popis vazby nebo spojení mezi oběma uzly	shareholder of
status	aktuální stav vztahu	Appointed
start_date	datum začátku vztahu	23-MAR-2006
end_date	datum ukončení vztahu	27-MAR-2007
sourceID	soubor uniklých dokumentů, ze kterých záznam pochází	Panama Papers

4.3 Datový model

Společně s daty zveřejnilo ICIJ také diagram vztahů neboli datový model, který zobrazuje, jakým způsobem jsou jednotlivé uzly propojené. Vazby mezi uzly jsou definované v tabulce relationships.csv. Každý uzel je definován unikátním ID (tzv. *node_id*) a vazební tabulka relationships pak obsahuje primární klíč jako kombinaci *node_id_start* a *node_id_end* napříč všemi uzly. Každý řádek dále definuje, o který typ vazby se jedná, například zda se jedná o registrovanou adresu, zprostředkovatele či akcionáře daného uzlu. O jaký typ vazby se může jednat zobrazuje *Obrázek 12* níže.



Obrázek 12. Datový model Offshore databáze od ICIJ. [85]

4.4 Předzpracování dat

Před tvorbou dashboardu se musím s daty blíže seznámit. Obsah jednotlivých tabulek již trochu znám z předchozí kapitoly, nevím však, jak moc jsou data kompletní a správné. Jelikož se jedná o uniklé dokumenty, je zcela jasné, že v datech budou chybějící hodnoty, neboť je nemožné získat kompletní údaje ke všem subjektům.

ICIJ zveřejňuje jména i adresy tak, jak byly napsány v uniklých záznamech. To znamená, že v seznamu jmen se může vyskytovat více pravopisných chyb nebo dokonce překlepů a jedno jméno může být napsáno několika způsoby a adresa, která se může jevit jako totožná, je v datech napsána několika způsoby. Každý řádek v jednotlivých souborech obsahuje své unikátní ID, která jsou propojená pomocí tabulky *relationship.csv*. Níže se na každou tabulku zaměřím individuálně, neboť v každé bude potřeba udělat trochu jiné úpravy.

Předzpracování dat probíhá pomocí Pythonu v prostředí Jupyter Notebook, protože má na rozdíl od Tableau mnohem větší možnosti úprav. Naopak v Tableau však můžu mnohem rychleji, efektivněji a přehledněji zobrazit propojené tabulky a pomocí jednoduchých vizualizací zjistit na první pohled ne příliš patrné chyby v datech, jelikož se jedná o velký soubor dat. V průběhu čištění dat proto používám souběžně oba nástroje, Python pro čištění dat a Tableau pro kontrolu, zda bude mnou vyžadovaná vizualizace fungovat s daty v této podobě a díky čemuž najdu případné chyby.

4.4.1 Kompletace tabulky address

Jelikož je hlavní částí dashboardu mapa zobrazující lokaci subjektů, tabulka nodes-address je pro mě jednou z klíčových, aby mapa zobrazovala geolokaci entit tak, jak má. Nejprve jsem si prohlédla strukturu jednotlivých sloupců.

Po použití metody info u tabulky nodes-address, kterou jsem uložila do proměnné df_address, jsem zjistila následující informace o její struktuře a počtu null hodnot (viz *Výpis 1*).

Výpis 1. Zobrazení základních informací o struktuře tabulky df_address

```
df_address.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 402246 entries, 0 to 402245
Data columns (total 8 columns):
 #   Column         Non-Null Count   Dtype 
---  ------         --------------   ----- 
 0   node_id        402246 non-null  object
 1   address        382314 non-null  object
 2   name           223348 non-null  object
 3   countries      276918 non-null  object
 4   country_codes  276918 non-null  object
 5   sourceID       402246 non-null  object
 6   valid_until    401529 non-null  object
 7   note           32 non-null      object
dtypes: object(8)
memory usage: 24.6+ MB
```

Snažila jsem se nejprve přijít na rozdíl mezi sloupcem *address* a *name*, jelikož se na první pohled zdálo, že obsahují stejnou informaci, nicméně ve sloupci *address* občas chyběla hodnota, která naopak byla vyplněná ve sloupci *name*. Po kontaktování a konzultaci s ICIJ jsem zjistila, že se jedná o nekonzistenci v datech a oba sloupce skutečně mají obsahovat stejné údaje, jen nebyly nikdy organizací sjednoceny. Rozhodla jsem se doplnit chybějící údaje ve sloupci *address* údaji ze sloupce *name*, pokud tam byly. Z chybějících hodnot bylo ze sloupce *name* doplněno celkem 19 930 hodnot.

```
df_address['address'] = df_address['address'].fillna(df_address['name'])
```

Dále jsem zjistila, že ačkoliv jsou v obou sloupcích hodnoty, občas je hodnota ve sloupci *name* detailnější a obsahuje celou adresu na rozdíl od hodnoty ve sloupci *address*. Porovnávám tedy délku řetězce v obou sloupcích a pokud je v *name* delší, nahrazuji adresu ve sloupci *address* tou delší.

```
df_address['address'] = df_address.apply(lambda x: x['name'] if len(str(x['address'])) < len(str(x['name'])) else x['address'], axis=1)
```

Odstranila jsem sloupec *note*, jelikož obsahuje pouze 32 hodnot a pro tvorbu dashboardu nebude potřeba. Odstraním také sloupec *name*, jelikož jsem důležité hodnoty již propsala do sloupce *address*, a tak tento již nepotřebuji. Rozhodla jsem se také smazat sloupce *sourceID* a *valid_until,* jelikož se tyto sloupce nacházejí i v ostatních tabulkách, primárně v tabulce entity, a u adresy pro mě není podstatné vědět informaci, z jakého zdroje pochází, jelikož tuto informaci získám po propojení s tabulkou entity.

```
df_address = df_address.drop(['note','name','sourceID','valid_until'],axis=1)
```

Ačkoliv je sloupec *address* téměř kompletní, sloupec *countries* obsahuje 125 328 chybějících hodnot. Jelikož byl tento sloupec organizací ICIJ doplněn automatickou identifikaci názvů zemí v rámci všech adres obsažených v databázi, je možné, že pokud je adresa nekompletní či napsána s chybou, nebylo možné identifikovat název země.

```
In [14]: df_address[(df_address['countries'].isnull())]
```

Out[14]:

	node_id	address	countries	country_codes
626	14000097	#25 Mason Complex; Stoney Ground; P.O. Box	NaN	NaN
627	14000098	#25 Mason Complex, Stoney Ground P.O. Box 193 Stoney Ground	NaN	NaN
632	14000103	#25 Mason Complex Stoney Ground; P.O. Box 193; The Vall	NaN	NaN
751	14000222	: UAE; DUBAI; JBR; BAHAR 1; Apt. 1904; P.O. Box 191884	NaN	NaN
764	14000235	002.GRD FLR; SAHIL BUNGALOW; DR. ANNIE BESANT ROAD; WORLI; MUMBAI ¿ 400 018	NaN	NaN
...	...	...	...	...
401827	240492512	17 ESPLANADE, ST HELIER, JE1 1B, JERSEY	NaN	NaN
401831	240492522	APARTMENT 42F. ESTORIL COURT. 55 GARDEN ROAD, COUNTRY, ZUG, COUNTRY	NaN	NaN
401832	240492523	ANARTMENT 42F. ESTORIL COURT. GARDEN ROAD, COUNTRY	NaN	NaN
401839	240492533	P.O. BOX 42211, APARTMENT 856. TAJ PALACE HOTEL, AL RIGGA ROAD, MIRAT	NaN	NaN
401939	240491639	BNP PARIBAS HOUSE, ANLEY STREET, ST HELIER, JE2 3QE, JERSEY, CHANNEL ISLANDS	NaN	NaN

125328 rows × 4 columns

Obrázek 13. Ukázka záznamů z tabulky df_address, které obsahují null hodnotu ve sloupci countries. Zdroj: autor

Každý z dílčích souborů obsahoval sloupec *country* a *country_codes*, rozhodla jsem se proto porovnat jejich hodnoty se tabulkou address, zda jsou hodnoty 1:1 a zjistila jsem, že velká část řádků s chybějícími hodnotami *countries* a *country_codes* v první tabulce jsou k dispozici v tabulkách entity a officer. Rozhodla jsem se tyto prázdné buňky doplnit hodnotou ze zmíněných tabulek.

Countries (Address.Csv)	Name	Rel Type	Address (Address.Csv)	Countries (Entities.csv)	Country Codes (Entities.csv)
Null	1 LEVER LTD	registered_address	CMS HOUSE, THIRD FLOO..	Malta	MLT
	1 MEDIA PROC..	registered_address	24, WINDSOR STREET, SLI..	Malta	MLT
	1 SHIPPING AN..	registered_address	45/1 TRIQ L-ISQOF F.S. CA..	Malta	MLT
	1 STEP AHEAD ..	registered_address	'THE PENTHOUSE, CAROLI..	Malta	MLT
	1 Stop Propert..	registered_address	5, TRIQ XEHDA, NAXXAR, ..	Malta	MLT
	1A Trading Ltd	registered_address	14/19, VINCENTI BUILDIN..	Malta	MLT
	1FACH HIMMLI..	registered_address	137 SPINOLA ROAD, ST. J..	Malta	MLT
	1GAMING (MA..	registered_address	VINCENTI BLDGS, SUITE 3..	Malta	MLT
	1o PARZECI LI..	registered_address	LEVEL 1, BLUE HARBOUR ..	Malta	MLT
	1SON LTD.	registered_address	3RD FLOOR, MERLIN HOU..	Malta	MLT
	1ST RESIDENTI..	registered_address	'VILLA MAURAMY' MONS. ..	Malta	MLT
	1ST SPADE LIM..	registered_address	5/2 MERCHANTS STREET, ..	Malta	MLT
	1STONE EIR LTD	registered_address	34, WINDSOR TERRACE, S..	Malta	MLT
	1StoneConnect..	registered_address	34, WINDSOR TERRACE, S..	Malta	MLT

Obrázek 14. Ukázka záznamů v Tableau, kde sloupec countries v tabulce address je null, ale nikoliv v tabulce entity. Zdroj: autor

V Jupyter notebooku jsem propojila tabulky entity a address prostřednictvím tabulky relationships a výsledek vložila do proměnné temp_table. Zobrazila jsem si takové řádky, které mají vyplněnou zemi z tabulky entity, ale nikoliv v tabulce address (viz *Výpis 2*). Během tohoto kroku jsem narazila na poměrně závažný problém, a to, že v tabulkách officer a entity je sice vyplněná hodnota země, ale ne jenom jedna. Sloupec country pravděpodobně slouží v těchto tabulkách k tomu, aby zobrazil všechny země, ve kterých daný subjekt jakýmkoliv způsobem figuruje. Při propojení s adresou však nelze přesně říct, která z adres patří ke které zemi. Musela jsem proto takové hodnoty vynechat, jinak by sloupec v Tableau nebyl rozpoznán jako geografická dimenze a nebylo by možné vytvořit vizualizaci v podobě mapy. Pro lepší přehlednost jsem přejmenovala sloupec *countries* a *country_codes* z tabulky address na *countries_ad* a *country_codes_ad,* jelikož jsou v ostatních tabulkách pojmenované stejně.

Countries Ad	Node Id	Node Id Ad	Address Ad	Countries	Country Codes
Null	56071925	58058260	AVENIDA DOM NUNO ALVARES PEREIRA, 24 RC ESTORIL 2765-260	Portugal;Spain	PRT;ESP
	56071928	58033788	4, AVENUE DES GUELFES APT B 053, MONACO MC98000	Monaco;Spain	MCO;ESP
	56071930	58093788	MILTON HILL HOUSE 309, STEVENTON-ABIGDON OXFORDSHIRE OX13 6AF	United Kingdom;Spain	GBR;ESP
	56071934	58051108	AV. DEL FENER 007, A, 65, EDIF. EL FORESTAL, ANDORRA LA VELLA	Andorra;Spain	AND;ESP
	56071937	58070743	DISCOVER GARDENS, BLDG 226, STUDIO 209, DUBAI	United Arab Emirates;..	ARE;ESP
	56071942	58047331	6 IPOH LANE 15-04 IMPERIAL HEIGHTS 438620	Singapore;Spain	SGP;ESP
	56071960	58013242	25, BOULEVARD DE BELGIQUE, EDEN TOWER 98000	Monaco;Spain	MCO;ESP
		58025277	25, BOULEVARD DE BELGIQUE, MONACO MC98000	Monaco;Spain	MCO;ESP
	56071961	58028536	25, BOULEVARD DE BELGIQUE, MONACO	Monaco;Spain	MCO;ESP
	56071962	58016520	3, STARI VAROS STREET, PODGORICA 81000	Montenegro;Spain	MNE;ESP
	56071966	58090797	LARGO DO PELORINHO 3 ALJEZUR 8670-085	Portugal;Spain	PRT;ESP
	56071967	58090797	LARGO DO PELORINHO 3 ALJEZUR 8670-085	Portugal;Spain	PRT;ESP
	56071975	58039233	8, QUEENS' TERRACE, KINGS ROAD, WINDSOR, BERKSHIRE SL4 2AR	United Kingdom;Spain	GBR;ESP
	56071981	58096194	RUA FREI NICOLAU OLIVEIRA 33, GANDARINHA, CASCAIS 2750-641	Spain;Portugal	ESP;PRT
	56072005	58053863	9A, VIA RIVIERA, CASTAGNOLA 6976	Switzerland;Spain	CHE;ESP
	56072007	58055996	88/235 EIGHT THONGLOR RESIDENCES SUKHUMVIT 55 BANGKOK 10100	Thailand;Spain	THA;ESP
	56072094	58046899	95, BOULVERAD GENERAL DE GAULLE, ST JEAN CAP FERRET, 21630	France;Greece	FRA;GRC

Obrázek 15. Ukázka záznamů, které obsahují více než 1 přiřazenou zemi k adrese. Zdroj: autor

Výpis 2. Tvorba tabulky temp_table spojením df_entity a df_address pomocí tabulky relationships

```
temp_table = pd.merge(pd.merge(
          df_entity[['node_id','countries','country_codes']],
          relationship_df[['node_id_start','node_id_end']],left_on='node_id',
          right_on='node_id_start'), df_address,left_on='node_id_end',
                    right_on='node_id_ad')

temp_table[(temp_table['countries'].notnull()) &
          (temp_table['countries_ad'].isnull())]

temp_table[(temp_table['countries'].notnull()) &
          (temp_table['countries_ad'].isnull()) &
          (temp_table['country_codes'].str.len()<=3)]
```

Out[34]:

	node_id	countries	country_codes	node_id_start	node_id_end	node_id_ad	address_ad	countries_ad	country_codes_ad
50691	130574	British Virgin Islands	VGB	130574	234443	234443	20 Raffles Place\n#27-03/08 Ocean Towers	NaN	NaN
55153	174608	Samoa	WSM	174608	171868	171868	HOLD FOR COLLECTION THE OCEAN TRUST	NaN	NaN
142609	82005027	Bermuda	BMU	82005027	81030824	81030824	814 Yuksam-dong; Kangnam-gu; Seoul; Democratic People's Republic of Korea	NaN	NaN
142666	82004733	Bermuda	BMU	82004733	81030741	81030741	Yumi LEE, Associate, Neoplux Capital, Co. Ltd., Doosan Tower, 26 FL 18-12, Euljiro - 6ga Joong-gu; Seoul; 100-730	NaN	NaN
146275	82017861	Cayman Islands	CYM	82017861	81011232	81011232	Prisco S.A.; Mariano de Los Santos 141; SAN ISIDRO LIMA	NaN	NaN
...	...	...	...	...	...	...	...	...	...
329421	55034096	Malta	MLT	55034096	58035752	58035752	42, GERONIMO, ST. JOHN STREET, FGURA, MALTA	NaN	NaN
329422	55034099	Malta	MLT	55034099	58068257	58068257	COUNTRY FLOWERS, CANNON ROAD, QORMI, MALTA	NaN	NaN
329423	55034100	Malta	MLT	55034100	58117301	58117301	VINCENTI BUILDINGS, SUITE 415, 28/15, STRAIT STREET, VALLETTAVLT 08, MALTA	NaN	NaN
329424	55034101	Malta	MLT	55034101	58104668	58104668	VINCENTI BUILDINGS, SUITE 406, 14/19, STRAIT STREET, VALLETTA, MALTA	NaN	NaN
329425	55034106	Malta	MLT	55034106	58009082	58009082	121, TRIQ IL-KARMNU, LUQA, MALTA	NaN	NaN

83860 rows × 9 columns

Obrázek 16. Zobrazení výsledků na základě skriptu z Výpisu 2. Zdroj: autor

Z výsledků vidím, že řádků, které obsahují pouze jednu hodnotu ve sloupci *countries*, nikoliv ve sloupci *countries_ad,* je 83 860. Tyto řádky vyberu a null hodnoty nahradím těmi ze sloupce *countries* (viz *Výpis 3*).

Výpis 3. Nahrazení prázdných hodnot sloupců tabulky df_address hodnotami z tabulky df_entity, pokud není délka řetězce delší než 3, případně neobsahuje v buňce středník

```
temp_table.loc[(temp_table['countries'].notnull()) &
(~temp_table['countries'].str.contains(';', na=False)), 'countries_ad'] =
temp_table.loc[(temp_table['countries'].notnull()) &
(~temp_table['countries'].str.contains(';', na=False)),
'countries_ad'].fillna(temp_table.loc[(temp_table['countries'].notnull()) &
(~temp_table['countries'].str.contains(';', na=False)), 'countries'])

temp_table.loc[(temp_table['country_codes'].notnull()) &
(temp_table['country_codes'].str.len() <= 3), 'country_codes_ad'] =
temp_table.loc[(temp_table['country_codes'].notnull()) &
(temp_table['country_codes'].str.len() <= 3),
'country_codes_ad'].fillna(temp_table.loc[(temp_table['country_codes'].notnull())
& (temp_table['country_codes'].str.len() <= 3), 'country_codes'])
```

Po tomto kroku jsem hodnoty vložila do původní tabulky, která je uložená v proměnné df_address. Abych tento krok mohla podniknout, odstranila jsem nejprve z dočasné tabulky všechny sloupce kromě *node_id*, který bude sloužit jako primární klíč pro propojení s původní tabulkou), *countries_ad* a *country_codes_ad*, které jsou nyní obohacené o hodnoty z tabulky entity. Jelikož jedna adresa může být přiřazena více entitám, před propojením tabulek je nutné také odstranit řádky s duplicitním *node_id*, které vznikly po propojení tabulek address a entity. S tím přichází také nutnost nejprve seřadit všechny řádky dle *node_id_ad* a *country_codes_ad*, jelikož se název a kód země mohl napsat pouze do jednoho z několika řádků, a já si chci nechat právě takový řádek, který má *countries_ad* vyplněn. Seřazením hodnot jsem přesunula null hodnoty u *countries* nakonec a při spuštění metody *drop.duplicates()* jsem přidala parametr, který zachová vždy první hodnotu ze všech duplikací, což bude ta s nenulovou hodnotou, pokud tam taková je. Nakonec mohu tabulku temp_table spojit s původní tabulkou df_address. Celý proces zobrazuje *Výpis 4*.

Výpis 4. Úprava tabulky temp_table, aby bylo možné ji spojit s tabulkou df_address

```
temp_table.drop(['node_id','countries','country_codes','node_id_start',
          'node_id_end'],axis=1, inplace=True)
temp_table = temp_table.sort_values(by=['node_id_ad', 'country_codes_ad'],
          na_position='last')
temp_table = temp_table.drop_duplicates(subset='node_id_ad', keep='first',
          ignore_index=True)

#joinuju temp_table s původní tabulkou
df_address = pd.merge(df_address,temp_table[['node_id_ad','countries_ad',
          'country_codes_ad']],how='left',on='node_id_ad')
```

Nyní již zbývalo pouze nahradit prázdné hodnoty u původních sloupců těmi, které jsou k dispozici v nových, dočasných sloupcích a poté ty dočasné smazat, viz *Výpis 5*.

Výpis 5. Vložení hodnot z nových sloupců v tabulce df_address do původních sloupců tam, kde jsou null

```
#nové hodnoty vložím do původní address tabulky tam, kde původní countries a
country_codes je null
df_address['countries_ad_x'] =
df_address['countries_ad_x'].fillna(df_address['countries_ad_y'])
df_address['country_codes_ad_x'] =
df_address['country_codes_ad_x'].fillna(df_address['country_codes_ad_y'])

df_address.drop(['country_codes_ad_y','countries_ad_y'],axis=1, inplace=True)
df_address.rename(columns={'countries_ad_x':'countries_ad',
          'country_codes_ad_x':'country_codes_ad'},inplace=True)
```

Stejný postup jsem aplikovala i u dalších tabulek a každou porovnala s tabulkou address. Zjistila jsem, že kromě tabulky entity obsahuje i tabulka officers vyplněnou zemi u ID, které jsou v tabulce address prázdné, a to u 124 511, resp. 106 269 bez řádků s více než 1 vyplněnou zemí. Zopakovala jsem tedy stejné kroky a vytvořila dočasnou tabulku, ve které jsem vyplnila chybějící hodnoty v tabulce address, a ty poté vložila do té původní.

Out[50]:

	node_id	name	countries	country_codes	node_id_start	node_id_end	node_id_ad	address_ad	countries_ad	country
1183	56039863	"ALPHA SHIPPING COMPANY" SIA	Latvia	LVA	56039863	58121826	58121826	VISBIJAS PROSPEKTS 7, RIGA LV-1014	NaN	
1184	56047202	"AWPG" ARTHUR WORLD PARTICIPATION GROUP	France	FRA	56047202	58050320	58050320	78, AVENUE MARCEAU, PARIS 75008	NaN	
1185	56051171	AW RESORT	France	FRA	56051171	58050320	58050320	78, AVENUE MARCEAU, PARIS 75008	NaN	
1310	56095670	"ELSEBETH" SCHIFFFAHRTSGESELLSCHAFT MBH & CO. KG	Germany	DEU	56095670	58035787	58035787	67, PALMAILLE HAMBURG 22767	NaN	
1311	56095668	"EMERALD" SCHIFFFAHRTSGESELLSCHAFT MBH & CO. KG	Germany	DEU	56095668	58035787	58035787	67, PALMAILLE HAMBURG 22767	NaN	
...	...	...	...	...	...	...	...	...	...	...
484814	56095980	morimedia GmbH	Germany	DEU	56095980	58122930	58122930	Wibbeltstrasse 10 Muenster 48147	NaN	
484822	56081059	netX International Limited	Malta	MLT	56081059	58094692	58094692	THE MARINE BUSINESS CENTRE, LEVEL 2, TRIQ ABATE RIGORD, TA' XBIEX	NaN	
484833	56014187	seeTelligence International B.V.	Netherlands	NLD	56014187	58083949	58083949	KONINGIN WILHELMINAPLEIN 13, 1062 HH, AMSTERDAM	NaN	
484952	240410749	湯 海 東	China	CHN	240410749	240390164	240390164	中國上海市嘉定區 菊園新區永靖路 288 弄 73 號 201 室	NaN	
484953	240410750	湯海東	China	CHN	240410750	240390164	240390164	中國上海市嘉定區 菊園新區永靖路 288 弄 73 號 201 室	NaN	

106269 rows × 10 columns

Obrázek 17. Ukázka záznamů, kde sloupce countries a country codes z tabulky df_officer obsahují hodnotu, ale v tabulce df_address jsou prázdné. Zdroj: autor

Z původních 125 328 chybějících hodnot ve sloupci *countries* a *country_codes* jich nyní chybí 15 654. Zjistila jsem, že u některých řádků chybí název země, přestože je vyplněn kód, a byť jich není příliš, rozhodla jsem se země doplnit na základě již existujících hodnot v tabulce.

Out[59]:

	node_id_ad	address_ad	countries_ad	country_codes_ad
396966	240450412	21 FARROW CRES, AJAX, ON LIS 4X - CANADA	NaN	CAN
397087	240450774	L08A SAN COMBE ROAD - RICHMOND TW9 3YG - UNITED K!IYGDOM	NaN	DEU
397331	240451384	ESTRADA JM DE 3546 (1636) LA LUCILLA VICENTE LOPEZ BSAS	NaN	ARG
397511	240452054	PANAMA	NaN	PAN
397580	240452086	GARAY JUAN DE 1225 – (1686) HURLINGHAM HURLINGHAM BSAS – E/C JAURETCHE ARTURO Y CRUCERO GRAL BELGRANO	NaN	ARG
399607	240453084	AV 105-C, APT. TA 0502 EDF., MENORCA ENTRE CA, BUILDING, 136 Y CA, 137 URB, PREBO 1	NaN	VEN
400189	240452647	OYSTERBAY HOUSE # 372/2, CHOLE ROAD	NaN	TZA
400465	240452822	012	NaN	UGA

Obrázek 18. Ukázka záznamů z tabulky df_address, kde chybí název země, ale znám kód země. Zdroj: autor

Kód níže nejprve vytvoří logickou masku, která obsahuje informaci o tom, kde je *countries_ad* null, resp. tam, kde je nulová hodnota, bude *True*. Dále vyberu všechny unikátní hodnoty ve sloupci *country_codes* a vytvořím slovník, který mapuje každý kód země na název země. Pomocí map() funkce nahrazuju null hodnoty ve sloupci *countries_ad* odpovídajícími názvy zemí, jak je uvedeno ve slovníku *country_name_dict* (viz *Výpis 6*).

Výpis 6. Doplnění chybějících hodnot ve sloupci countries pomocí mapování hodnot ze slovníku

```
mask = df_address['countries_ad'].isnull()
unique_codes = df_address['country_codes_ad'].unique()
country_name_dict =
dict(df_address.dropna(subset=['countries_ad']).groupby('country_codes_ad')
          ['countries_ad'].first())

# mapuju dictionary na null hodnoty v countries_ad
df_address.loc[mask, 'countries_ad'] = df_address.loc[mask,
          'country_codes_ad'].map(country_name_dict)
```

Po propojení s tabulkami entity a officer jsem zjistila, že některé řádky těchto souborů měly naopak nejspíše vyplněný název země, ale ne kód. Stejným postupem jako výše jsem tedy vyplnila kód země do řádků s chybějící hodnotou dle existující hodnoty v ostatních řádcích.

Out[67]:

	node_id_ad	address_ad	countries_ad	country_codes_ad
255495	240007953	BLVD. KARBISMEVA, GENERAL, HOUSE 18, APP 213 RUSSIAN FEDERATION	Russian Federation	NaN
255713	240003743	ANDREA ZARDA STREET, SEA WAY COURT, OFFICE/FLAT 57	Cyprus	NaN
256299	240000065	UNIT 3O-01-1560, DMCC-PH2-P&G PLEXS	United Arab Emirates	NaN
256312	240004353	43/45 LA MOTTE STREET	Jersey	NaN
256344	240004369	NO. 9 ROTHERFILED PLACE	Sri Lanka	NaN
256364	240004380	PERENLOK KERAMICHESKIJ 7A-7B	Ukraine	NaN
256378	240004387	STAVOKONUSHUNY PEREULOK 10	Russian Federation	NaN
257487	240010356	FLAT 3, 40 BUKHORO STREET	Tajikistan	NaN
257501	240010370	47 MOLODAYA GVARDIA STREET	Kazakhstan	NaN

Obrázek 19. Zobrazení záznamů z tabulky df_address, kde chybí kód země, ale znám název země. Zdroj: autor

Při propojení s tabulkou intermediaries v Tableau jsem zjistila, že obsahuje pouze 2 hodnoty, které v tabulce address chybí. Jelikož se jedná o tak nízký počet, rozhodla jsem se hodnoty přepsat ručně (viz *Výpis 7*).

node_id_ad	countries_ad	country_co..	address_ad	Name (Nodes-I..	Countries (..	Country Co..
33000034	Null	Null	PO BOX F-2682, FREEPORT GRAND BAHAMA	CONSTANCE M..	Bahamas	BHS
33000044	Null	Null	SUITE 14 & 15, PIONEER'S WAY FREEPORT	BRIDGEWATER..	Bahamas	BHS

Obrázek 20. Ukázka záznamů v Tableau, kde chybí země v tabulce address, ale hodnota je známa v tabulce intermediary. Zdroj: autor

Výpis 7. Manuální nahrazení hodnot ve sloupcích countries_ad a country_codes_ad.

```
#manuálně přepisuju 2 hodnoty country, které jsou k dispozici v tabulce
intermediaries, ale ne v address
df_address.loc[df_address['node_id_ad'] == '33000034', 'countries_ad'] =
                    'Bahamas'
df_address.loc[df_address['node_id_ad'] == '33000034', 'country_codes_ad'] =
                    'BHS'

df_address.loc[df_address['node_id_ad'] == '33000044', 'countries_ad'] =
                    'Bahamas'
df_address.loc[df_address['node_id_ad'] == '33000044', 'country_codes_ad'] =
                    'BHS'

df_address[(df_address['node_id_ad']=='33000034') |
          (df_address['node_id_ad']=='33000044')]
```

Out[67]:

	node_id_ad	address_ad	countries_ad	country_codes_ad
228768	33000034	PO BOX F-2682, FREEPORT GRAND BAHAMA	Bahamas	BHS
228778	33000044	SUITE 14 & 15, PIONEER'S WAY FREEPORT	Bahamas	BHS

Obrázek 21. Zobrazení úspěšného manuálního nahrazení hodnot ve vybraných záznamech. Zdroj: autor

Po výše provedených úpravách mi v datech zbývá 15 171 prázdných hodnot. Vzhledem k představě mého dashboardu, který by měl ukazovat počet entit na základě geolokace, se mi prázdné hodnoty příliš nehodí, jelikož se v mapě nezobrazí. Je zřejmé, že se mi nepodaří doplnit všechny státy k adresám, neboť jsou některé napsané špatně, nekompletně či vůbec nedávají co se obsahu týče smysl, viz *Obrázek 22*. Zkusila jsem část adres doplnit pomocí geokodéru Nominatim z python knihovny geopy, který převádí adresy nebo názvu místa na souřadnice zeměpisné šířky a délky, resp. umístění na mapě (viz *Výpis 8*), nicméně žádná adresa nebyla doplněna.

countries_ad	address_ad	country_co..	Name
Null	CAPITAL AUTORIZADO CAMBIO	Null	EL PORTADOR
	CARR. A PUERTA PARADA VILLA 1-3 PIEDRA PARAD..	Null	Marta Carolina Saadeh Di..
	CARRERA 7 126-30 TORRE 8, APTO. 1230.	Null	INGRID CAROLINA PAZ M..
	CARRERA 7; No.113-43 OF. 302	Null	NELSON JULIAN BONILLA ..
	CARRERA 11 NO. 82-01 OF-92	Null	THE BEARER
	CARRERA 11 NO. 82-01 OF. 902	Null	THE BEARER
	CARRERA 11 NO. 82-11 OF. 902	Null	THE BEARER
	Cascajal 205 (ex-Casuarinas)	Null	EDUARDO EMILIO AUZA N..
	CASE 1796766	Null	The client keeps the infor..
	CASE NO. 1796762	Null	the client keeps the infor..
	CASE NO. 1796764	Null	The client keeps the infor..
	CASE NO. 1796768	Null	The client keeps the infor..
	CASE NO. 1796772	Null	the client keep the inform..
	CASE POSTALE 348	Null	H2R NOMINEE FOUNDATI..
	Castellum Fiducia Trust Reg.	Null	Castellum Fiducia Trust R..
	CENTRAL WING, MEZZANINE FLOOR	Null	ADAM SMITH ASSOCIATE..
	Centro Gerencial Mohedano; Piso 3; Ofc. 3-B	Null	FERNANDO LAURÍA ROME..
	Cerrito 517, Of. 603	Null	Alberto Santiago Santoro

Obrázek 22. Ukázka záznamů v Tableau, kde hodnota ve sloupci address_ad není skutečná adresa. Zdroj: autor

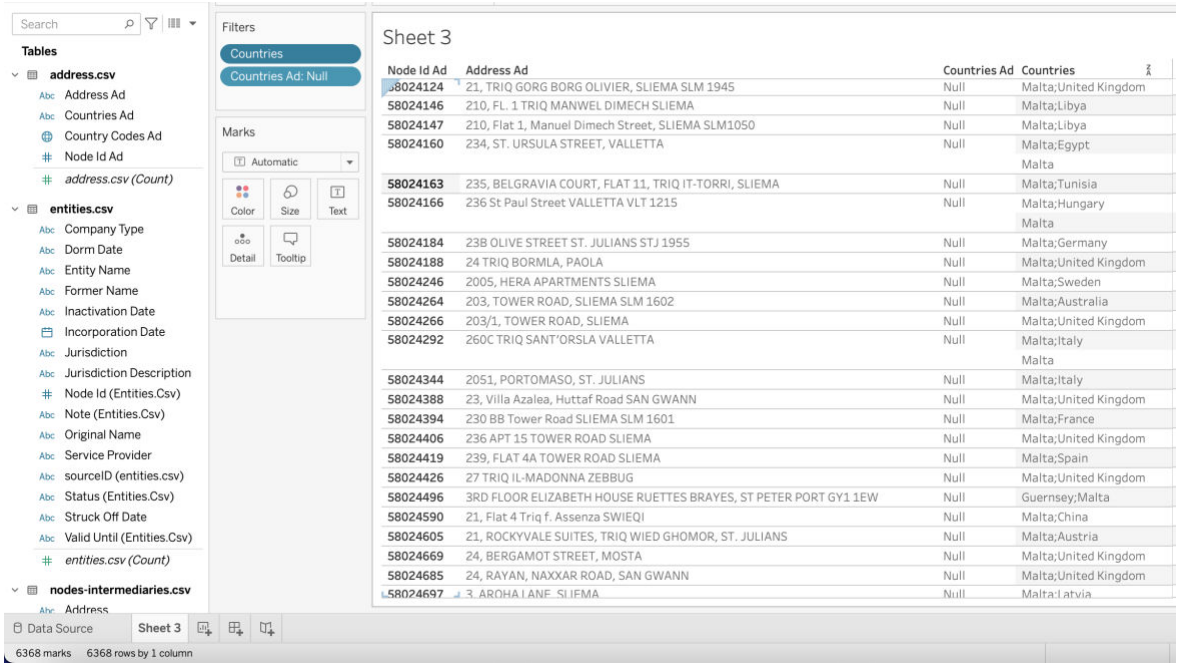
Výpis 8. Využití geokodéru pro automatické doplnění názvu a kódu země

```
from geopy.geocoders import Nominatim
for i, row in df_address.iterrows():
 if not row['countries_ad'] and not row['country_codes_ad']:
        address = row['address_ad']
        location = geolocator.geocode(address, timeout=10)
        if location is not None:
            country = location.raw.get('address', {}).get('country')
            country_code = location.raw.get('address', {}).get('country_code')
            df_address.at[i, 'countries_ad'] = country
            df_address.at[i, 'country_codes_ad'] = country_code
```

Ačkoliv s chybějícími hodnotami počítám, tento počet se mi zdá stále příliš vysoký. Rozhodla jsem se proto prozkoumat hodnoty manuálně. Tabulku address jsem nahrála do Tableau Deskop a zobrazila si seznam veškerých adres, které obsahují chybějící název země. Dále jsem k ní připojila tabulku officers, abych mohla pomocí názvu zemí z této tabulky, především pak hodnot, které obsahují více zemí v jednom, adresy více seskupit. Přestože nemůžu jednoznačně určit, do které země adresa připojená k danému seznamu zemí patří, minimálně to seskupilo adresy do určitých oblastí a můžu si díky tomu všimnout opakujících se klíčových slov, ze kterých vytvořím slovník a pomocí něj můžu vyplnit název země k co nejvíce adresám.

Jednou z nejčastěji vyskytujících se zemí přiřazených k adresám je Malta, rozhodla jsem se proto začít adresami, které se dle sloupce countries tabulky officer nachází na Maltě. Počet

adres, které potenciálně patří na Maltu, je 6 368, téměř polovina chybějících hodnot. Pokud se alespoň polovina těchto adres nachází zde, snížila bych počet chybějících hodnot o poměrně významnou část.



Obrázek 23. Zobrazení takových adres, ke kterým chybí název země, ale patří do jedné ze zemí uvedené v tabulce officer. Zdroj: autor

Klíčová slova volím především dle názvu měst, jelikož se jedná o nejpravděpodobněji opakované slovo ve všech adresách, pokud se nachází ve stejném městě. Pokud se název města opakuje často v adresách, je pro mě vhodným kandidátem do seznamu, podle kterého se vyplní název a kód země (viz Výpis 9). Mým cílem je získat s každým klíčovým slovem alespoň 10 zemí, jelikož není v mých možnostech manuálně sepsat všechna vyskytující se města v seznamu adres. Částečným důvodem je i fakt, že se název města ne vždy v datech vyskytuje, některé záznamy vypadají nekompletně, a není tak podle čeho název země identifikovat.

Výpis 9. Funkce na doplnění konkrétních hodnot dle seznamu klíčových slov

```
def contains_substring(s, substrings):
    if isinstance(s, str):
        return any(substring in s for substring in substrings)
    else:
        return None

malta = ['SLIEMA', 'SWIEQI', 'ST. JULIANS', 'ST PAUL\'S BAY', 'MELLIHEA',
         'LIJA', 'VALETTA', 'MOSTA', 'GZIRA', 'ATTARD', 'MARSASCALA',
         'FLORIANA', 'SAN GWANN', 'RABAT','BALZAN','NAXXAR','ZEBBUG','PEMBROKE',
         'BIRKIRKARA','MSIDA','ZABBAR','PAOLA','QORMI','KERCEM','TA\'XBIEX',
         'ZURRIEQ','MALTA','XAGHRA','GHARB','SIGGIEWI']

mask3 = df_address['country_codes_ad'].isnull()

df_address.loc[mask, 'address_ad'] = df_address.loc[mask3,
'address_ad'].str.upper()
df_address.loc[mask, 'countries_ad'] = df_address.loc[mask3,
'address_ad'].apply(lambda x: 'Malta' if contains_substring(x, malta) else None)
df_address.loc[mask, 'country_codes_ad'] = df_address.loc[mask3,
'address_ad'].apply(lambda x: 'MLT' if contains_substring(x, malta) else None)
```

Stejným způsobem, jako u Malty, jsem doplnila názvy zemí pro adresy v České republice, Rakousku, Německu, Švýcarsku, Belgie, Francii, Britských panenských ostrovech, Anglii, Itálii, Gruzii, Španělsku, Kanadě, Austrálii, Brazílii, Libyi, Libérii, Monaku, Lucembursku, Koreji, Číně, na Kypru a Bahamách.

Po těchto úpravách mi zbývá 7 569 chybějících hodnot, tedy o více než polovinu méně než před manuálním doplňováním.

Ostatní prázdné hodnoty v souboru jsem se rozhodla přejmenovat na „Not identified“, aniž bych jakýkoliv řádek smazala. Důvod je ten, že přestože v některých sloupcích chybí hodnota, ostatní se budou hodit jako doplňující informace na jiný druh pohledu a investigativní novinář uvítá při vyšetřování jakoukoliv informaci o hledaném subjektu, byť nekompletní. Některé entity se sice nezobrazí na mapě z důvodu chybějícího názvu či kódu země, ale v pohledu detailu o konkrétně zvolené entitě či osobě může stále zjistit mnoho informací a na základě uvedené adresy sám odvodit zemi, jelikož při bližší práci s dashboardem předpokládám, že uživatel hledá konkrétní informace o daném subjektu, a musí při další práci získaná data tak či tak verifikovat.

4.4.2 Odstranění nepotřebných a duplicitních sloupců

Jelikož se ve všech souborech vyskytují určité sloupce, které jsou pro mou práci nepodstatné, případně se jedná o sloupce, které obsahují stejnou infomaci a po propojení by tedy byly duplicitní, smazala jsem je, viz *Výpis 10*. Nejenže se tím zmenší velikost už tak

opravdu velké tabulky, ale také tím předejdu nepřehlednosti po propojení tabulek (duplikace názvů sloupců). Konkrétně se jedná o sloupce *internal_id*, *ibRUC* v tabulkách entity a intermediaries, *sourceID* a *valid_until* v tabulkách intermediaries, officer a relationships (tento sloupec zůstává pouze v tabulce entity, neboť tvoří hlavní komponentu celého modelu).

Výpis 10. Odstranění nepotřebných sloupců ze všech dílčích tabulek

```
df_entity = df_entity.drop(['internal_id','ibcRUC'],axis=1)

#odstranění nepotřebných sloupců - sourceID, valid_until zůstanou pouze v entity
tabulce
df_inter = df_inter.drop(['internal_id','sourceID','valid_until','note'],axis=1)
df_officer = df_officer.drop(['sourceID','valid_until','note'],axis=1)
relationship_df = relationship_df.drop(['sourceID'],axis=1)
```

Po doplnění hodnot ze sloupců *countries* a *country_codes* do tabulky address se jeví jako logická možnost smazat i tyto sloupce ze všech dílčích tabulek, nicméně jsem po detailnějším pátrání došla k závěru, že subjekty v těchto souborech mohou mít vyplněnou zemi, ale nemají k sobě přiřazenou žádnou adresu. Celkový počet záznamů v tabulce address je totiž pouze 402 246, kdežto například v tabulce entity je celkových záznamů 814 344 a i když beru v potaz možnost, že jedna adresa může být přiřazena k více entitám (protože patří například jednomu subjektu), jde o dvojnásobný počet záznamů více. Sloupec *country_codes* a *countries* v této tabulce obsahuje 504 991 nenulových záznamů. Tyto sloupce proto nechávám až do poslední chvíle k dispozici a budu s nimi dále pracovat i po spojení tabulek do jedné.

```
df_entity.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 814344 entries, 0 to 814343
Data columns (total 21 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   node_id                  814344 non-null  object
 1   name                     814315 non-null  object
 2   original_name            389522 non-null  object
 3   former_name              6837 non-null    object
 4   jurisdiction             806780 non-null  object
 5   jurisdiction_description 806782 non-null  object
 6   company_type             138751 non-null  object
 7   address                  299323 non-null  object
 8   internal_id              389522 non-null  object
 9   incorporation_date       788470 non-null  object
 10  inactivation_date        144773 non-null  object
 11  struck_off_date          343843 non-null  object
 12  dorm_date                20207 non-null   object
 13  status                   358032 non-null  object
 14  service_provider         344086 non-null  object
 15  ibcRUC                   562475 non-null  object
 16  country_codes            504991 non-null  object
 17  countries                504991 non-null  object
 18  sourceID                 814344 non-null  object
 19  valid_until              814137 non-null  object
 20  note                     41765 non-null   object
dtypes: object(21)
memory usage: 130.5+ MB
```

Obrázek 24. Zobrazení informací o sloupcích v tabulce entity. Zdroj: autor

4.4.3 Sjednocení tabulek entity a others

Jelikož je struktura tabulek entity i others téměř totožná a označuje tu samou věc – tedy nějaký subjekt – rozhodla jsem se tabulky sjednotit do jedné. Jelikož je pro mě mnohem přirozenější propojování tabulek jak vertikálně, tak horizontálně, provádět pomocí SQL, využila jsem knihovnu pandasql, která umožňuje pracovat s pandas DataFrames pomocí SQL dotazů. Tabulky jsem spojila pomocí operátoru UNION, který vyžaduje, aby obě tabulky měl stejný počet sloupců. Tabulka entity jednoznačně obsahuje více sloupců, proto abych vyrovnala počet sloupců, vytvořila jsem pro tabulku others stejnojmenné sloupce s NULL hodnotami a některé sloupce obsahující stejnou informaci přejmenovala tak, aby seděly s názvy z tabulky entity, viz *Výpis 11*.

Výpis 11. Sjednocení tabulky entity a others pomocí SQL v Jupyter Notebooku

```
from pandasql import sqldf
df_merged = sqldf(
    "SELECT * FROM df_entity
     UNION
     SELECT
     node_id,
     name,
     null as original_name,
     null as former_name,
     jurisdiction,
     jurisdiction_description,
     type as company_type,
     incorporation_date,
     null as inactivation_date,
     struck_off_date,
     closed_date as dorm_date,
     null as status,
     countries as entity_countries,
     country_codes as entity_country_codes,
     null as service_provider,
     sourceID,
     valid_until,
     note
     FROM df_others"
     )
```

4.4.4 Redukce řádků tabulky relationships

V tabulce relationship jsou vazby v rámci jedné tabulky, například v tabulce entity jedna entita má vazbu na jinou. Často se jedná o vazbu „same name as“, „similar name“ nebo „same_company_as“. ICIJ tento typ vazby do dat přidal, protože název úředníků ani entit nijak neupravovali ani nesjednotili, a tak může být název subjektu v datech napsaný různě.

Výsledný datový zdroj byl poměrně náročný na moji výpočetní techniku, jelikož obsahoval přes 3 miliony řádků, a tak jsem se rozhodla tyto vazby smazat, abych tabulku alespoň trochu zmenšila. Důvodem je mimo jiné také fakt, že Tableau nabízí poměrně chytré filtrování, a pokud se název entity či úředníka skutečně vyskytuje v datech různě napsaný, s pravopisnou chybou či částečně chybějícím názvem, případný investigativní žurnalista, který databázi bude využívat, tuto skutečnost uvidí – všechny výsledky částečně vyhovující názvu, který uživatel hledá, bude zobrazen ve filtru při zadání klíčového slova, v případě filtrování přes wildcard filtr Tableau vrátí všechny výsledky, které částečně odpovídají vyhledávanému slovu. V případě filtrování ze seznamu jsou pak jednotlivé hodnoty seřazené abecedně. Vzhledem k tomu, že při vyšetřování novinář hledá jakoukoliv stopu a každou informaci poté ověřuje, jsem přesvědčená o tom, že by si prohlédl všechny podobně pojmenované záznamy bez ohledu na to, zda je mezi nimi vazba „same_name_as“ či nikoliv. Celkem jsem tímto krokem odstranila 171 831 řádků.

```
values_to_drop = ['same_as', 'same_company_as', 'similar_company_as', 'same_name_as',
                  'same_intermediary_as', 'similar', 'probably_same_officer_as',
'same_id_as']
relationship_df = relationship_df[~relationship_df["rel_type"].isin(values_to_drop)]
```

4.4.5 Propojení tabulek a finální úpravy

Před provedením několika finálních úprav, které se budou vztahovat na všechny tabulky, je nejprve propojím a vytvořím jednotný datový zdroj. Při práci s úpravou dat jsem různě jednotlivé tabulky propojovala v nástroji Tableau, brzy jsem ale přišla na to, že je systém relationships v Tableau v případě existence vazební tabulky, která propojuje více než 2 tabulky, trochu problematická a vazby nefungují tak, jak by měly. Usoudila jsem, že bude propojení tabulek přes SQL query, se kterým jsem lépe obeznámená, mnohem komfortnější. Jelikož jsem si již stáhla pandasql knihovnu, rozhodla jsem se tabulky propojit rovnou v Jupyter Notebooku.

Tabulky jsem propojila postupně, základ datasetu tvoří tabulka relationships, jelikož obsahuje vazby mezi všemi tabulkami. Z datového modelu v kapitole *4.3* vidím, že mají všechny tabulky vazbu na entity, přičemž *node_id* v tabulce entity v takovém případě odpovídá sloupci *node_id_end* v tabulce relationships, jelikož šipka směřuje **do** tohoto uzlu.

Nejprve jsem tedy propojila tabulku entity s officer, která zobrazí ke všem entitám příslušného úředníka, pokud existuje. Díky LEFT JOIN vazbě ponechám ve výsledku celou tabulku relationship bez ohledu na to, zda měl záznam odpovídající záznam v tabulce officer. Výsledek jsem uložila do proměnné off_to_entity (viz Výpis 12) a v dalších krocích, kdy k ní budu připojovat další tabulky, o ni budu hovořit jako o hlavní tabulce.

Výpis 12. Propojení tabulky entity s officer pomocí SQL v Jupyter Notebooku

```
#propojení tabulky entity s officer - office id = id_start, entity_id = id_end
off_to_entity = sqldf(
    "SELECT
       M.*,
       O.node_id as officer_id,
       O.officer_name,
       officer_countries,
       officer_country_codes,
       R.*
     FROM relationship_df R LEFT JOIN df_officer O ON O.node_id=R.node_id_start
     LEFT JOIN df_merged M ON R.node_id_end=M.node_id")
```

Nyní mám vyplněné všechny řádky, jejichž *node_id_end* korespondovalo s *node_id* v tabulce entity. Problém je v tom, že *node_id* v této tabulce může v tabulce relationships představovat jak *node_id_start*, tak *node_id_end* v závislosti na tom, s jakou tabulkou je právě spojována. Při vazbách mezi entity a officer/intermediaries je entity na koncové straně vazby (*node_id_end*), při vazbě s address je však počáteční vazbou (*node_id_start*), neboť v modelu vychází šipka **od** ní. Bez úprav by řádky, které zobrazují vazbu mezi entity a address (konkrétně vazba „registered_address“) obsahovaly prázdné hodnoty ve sloupcích korespondující původní tabulce entity, resp. by byly vyplněné pouze řádky, kde entity odpovídá *node_id_end*.

Vytvořila jsem nejprve tabulku entity_to_rel, kde jsem propojila přes INNER JOIN s tabulkou relationship ty záznamy, které odpovídaly vztahu *node_id=node_id_start*. Do tabulky zahrnuji také null sloupce s názvem *officer_id* a officer_name, neboť jsem v předchozím kroku už propojila entity s tabulkou officer a jsou tam tedy navíc. Jelikož v dalším kroku tabulky spojuji pomocí operátoru UNION, potřebuji, aby měly obě tabulky stejný počet sloupců. Cílem je spojit řádky, které obsahují informaci o entity, když odpovídá *node_id_start* s hlavní tabulkou, ve které jsou všechny vztahy, a tu jsem vložila do proměnné entity2, viz *Výpis 13*.

Výpis 13. Tvorba tabulka entity_to_rel a následně entity2

```
entity_to_rel = sqldf(
    "SELECT
       M.*,
       NULL as officer_id,
       NULL as officer_name,
       R.*
     FROM relationship_df R JOIN df_merged M ON M.node_id=R.node_id_start"
       )
entity2 = sqldf(
    "SELECT * FROM off_to_entity UNION SELECT * FROM entity_to_rel"
       )
```

Nyní mám v nové tabulce duplikované řádky, jeden s prázdnou hodnotou u entity a jeden s vyplněnými údaji. Musím proto tyto hodnoty smazat. Nejprve jsem si hodnoty v tabulce seřadila dle kombinace *node_id_start* a node_id_end, dále pak podle *entity_name* a *officer_name,* jelikož chci řádky, které obsahují null v jednom z těchto dvou sloupců dát nakonec. V druhém kroku jsem odstranila duplicity, přičemž jsem zachovala vždy první řádek, jelikož jsem ty s null hodnotami seřadila na konec (viz *Výpis 14*).

Výpis 14. Odstranění duplikovaných hodnot v tabulce entity2 vzniklé při jejím vzniku

```
entity2 = entity2.sort_values(by=['node_id_start',
'node_id_end','entity_name','officer_name'], na_position='last')
entity2 = entity2.drop_duplicates(subset=['node_id_start',
'node_id_end','rel_type','link','status','start_date','end_date'], keep='first',
ignore_index=True)
```

Během tohoto kroku jsem zjistila, že jsou v tabulce i další duplicitní hodnoty než jen ty, které jsem způsobila já spojením předchozích dvou tabulek. Rozšířila jsem odstranění duplicit na seznam všech sloupců z tabulky relationship, aby se vymazaly pouze takové duplicitní řádky, které mají stejný jak *node_id_start a node_id_end,* tak i *rel_type, link, status, start_date* a *end_date*. Kromě mnou vytvořených duplicitních záznamů bylo odstraněno dalších 8 012 řádků. Pro jistotu jsem zkontrolovala, zda duplicity existují v původní tabulce relationships.csv (viz *Výpis 15*) nebo jsem je způsobila vlastní chybou, a skutečně zde byly. Rozhodla jsem se tedy smazat i tyto řádky.

Výpis 15. Seřazení a zobrazení duplicitních hodnot v tabulce relationships

```
relationship_df = relationship_df.sort_values(by=['node_id_start',
'node_id_end','rel_type','link','status','start_date','end_date'])
relationship_df[relationship_df.duplicated(['node_id_start',
'node_id_end','rel_type','link','status','start_date','end_date'])]
```

Out[122]:

	node_id_start	node_id_end	rel_type	link	status	start_date	end_date
925510	115663	167551	officer_of	shareholder of	NaN	NaN	NaN
2398688	240380003	240370004	officer_of	Ultimate Beneficial Owner	NaN	25-AUG-2016	NaN
2398691	240380004	240370005	officer_of	Ultimate Beneficial Owner	NaN	08-JAN-2001	NaN
2398692	240380004	240370005	officer_of	Ultimate Beneficial Owner	NaN	08-JAN-2001	NaN
2398695	240380005	240370006	officer_of	Ultimate Beneficial Owner	NaN	24-AUG-2016	NaN
...	...	...	...	...	...	...	...
2040048	86030936	88012581	registered_address	registered address	NaN	NaN	NaN
2040049	86030936	88012581	registered_address	registered address	NaN	NaN	NaN
2040052	86030936	88012581	registered_address	registered address	NaN	NaN	NaN
2055218	86030956	88017556	registered_address	registered address	NaN	NaN	NaN
2004544	86031214	88001935	registered_address	registered address	NaN	NaN	NaN

8012 rows × 7 columns

Obrázek 25. Ukázka duplicitních záznamů v tabulce relationships. Zdroj: autor

Tabulku v nové proměnné jsem poté vzala a k ní stejným způsobem jako u officer připojila tabulku intermediary, viz *Výpis 16*.

Výpis 16. Propojení hlavní tabulky s intermediary pomocí SQL v Jupyter Notebooku

```
#propojení tabulky entity2 s intermediary - inter id = id_start, entity2 = id_end
int_to_df = sqldf(
"SELECT
    I.node_id as inter_id,
    I.intermediary_name,
    I.status AS intermediary_status,
    I.inter_countries,
    I.inter_country_codes,
    DF.*
 FROM entity2 DF
 LEFT JOIN df_inter I ON I.node_id=DF.node_id_start"
 )
```

Nyní jsou vyplněné vazby mezi entity = officer a entity = intermediary. Tím bych měla propojené základní subjekty. Všechny tyto subjekty mají nějakou registrovanou adresu, přichází tedy na řadu připojení tabulky address, k jejímž řádkům jsou nyní připravené i příslušné údaje o entitách.

```
address_to_df = sqldf(
"SELECT
    DF.*,
    A .address_ad,
    A .countries_ad,
    A .country_codes_ad
FROM int_to_df DF
LEFT JOIN df_address A ON DF.node_id_end=A .node_id_ad"
)
```

Přestože jsem z tabulky relationships odstranila self join vazby, jako jsou „same_name_as“ či „similar_name_as“, existují zde vazby, které zobrazují například typ vazby nominee_shareholder či nominee director, což jsou osoby, které jsou za pravidelný roční poplatek pověřené držením akcií společnosti, zatímco klient je finálním vlastníkem nebo vykonávají funkci statutárního orgánu společnosti. [86] Jelikož mi tyto vazby přišly důležité a pro případného novináře užitečné, rozhodla jsem se tyto vazby ponechat.

Officer Name	Rel Type	Link	Related Name
Asia Consulting and Inves..	underlying	Nominee Shareholder of	Dewi Suryati Liauw
ASIA INVESTMENT CAPIT..	underlying	Nominee Shareholder of	Irawan Kurniadi Tjandra
Atherton Investment Gro..	underlying	Nominee Shareholder of	Toshio Nakama
AxisInvest Corporation	underlying	Nominee Shareholder of	Budi Enijati Maria Soedja..
Barlaw Limited	underlying	Nominee Director of	Cheng Yong Kwang
Bill Irving Murray	underlying	Nominee Trust Settlor of	Dr Yeung, Tsun Man Eric &..
BNB Brothers Limited	underlying	Nominee Shareholder of	Ng Tje Suang
			Tjiu Thomas Effendy
Bobby Iman Satrio	underlying	Nominee Shareholder of	Supartono Suparto
Bret John Gibson	underlying	Nominee Director of	Marcus Leanard Magnus I..
Bright City Group Corpora..	underlying	Nominee Shareholder of	Michael Steven and Ingrid..
CAPEWELL INTERNATION..	underlying	Nominee Shareholder of	JIALIPTO JIARAVANON
Capital Reserves Ltd	underlying	Nominee Shareholder of	HARTADI ANGKOSUBROTO
Caribbean Technology Inv..	underlying	Nominee Shareholder of	Wu, Yu-Mei
Chan Chung, Chin Hoi	underlying	Nominee Trust Settlor of	CHAN Pak Kee
Chandler Capital Inc.	underlying	Nominee Shareholder of	Lambertus Somar
Charterhouse Limited	underlying	Nominee Shareholder of	Le Purna Harjani
CHEN, KUEI-JUNG	underlying	Nominee Shareholder of	HUANG, PI-CHU
Cheston Investments Ove..	underlying	Nominee Shareholder of	Lim Choi Hwee

Obrázek 26. Ukázka záznamů, kdy úředník může mít vazbu na jiného úředníka. Zdroj: autor

To zahrnuje opětovné propojení tabulek officer k mému finálnímu datasetu, tentokrát však jen sloupce officer_name. V datovém modelu je zobrazena SELF vazba také na tabulky entity, neobsahovala však žádné důležité informace, většinou se jednalo právě o „same_name_as“ vazby, a z výkonnostních důvodů mé výpočetní techniky jsem se rozhodla tuto vazbu kompletně vynechat, abych trochu zkrátila dobu běhu skriptu. Připojila jsem tedy pouze tabulku officer a nový sloupec pojmenovala related_name, viz *Výpis 17*.

Výpis 17. Opětovné připojení tabulky officer k hlavní tabulce pomocí SQL

```
self_join_off = sqldf(
"SELECT
    DF.*,
    B.officer_name as related_name
 FROM address_to_df DF
LEFT JOIN df_officer B ON DF.node_id_end=B.node_id"
)
```

Nyní mám ve své nové tabulce 4 různé sloupce, které obsahují název země (pojmenovány *entity_country, officer_countries, intermediary_countries* a *countries_ad*) a 4, které obsahují kód země (*entity_country_codes, officer_country_codes, inter_country_codes* a *country_codes_ad*). Postupně jsem vzala sloupce z jednotlivých tabulek a stejným způsobem, jako jsem se v kapitole *4.4.1* snažila doplnit tabulku address_ad, jsem doplnila název a kód země ze sloupců *entity_countries* a *entity_country_codes* do sloupců

countries_ad a *country_codes_ad*, pokud je v těchto sloupcích v daném řádku null. Jelikož každý řádek nyní reprezentuje jeden typ vztahu, bude hodnot, které je třeba doplnit hodně. V případě sloupců country z tabulky entity jde o 1 939 126 řádků. Opět však doplňuji pouze ty záznamy, které obsahují pouze jednu zemi, nikoliv více zemí oddělené středníkem, viz *Výpis 18*.

	node_id	entity_countries	entity_country_codes	countries_ad	country_codes_ad
0	200521	Not identified;Canada	XXX;CAN	None	None
1	218909	British Virgin Islands	VGB	None	None
2	217070	British Virgin Islands;Jersey	VGB;JEY	None	None
3	213050	Hong Kong;British Virgin Islands	HKG;VGB	None	None
4	199001	British Virgin Islands;United States	VGB;USA	None	None
...	...	...	...	...	...
3157063	157161	Samoa	WSM	None	None
3157064	157161	Samoa	WSM	None	None
3157065	157161	Samoa	WSM	None	None
3157067	164980	British Virgin Islands	VGB	None	None
3157068	164980	British Virgin Islands	VGB	None	None

1939126 rows × 5 columns

Obrázek 27. Porovnání sloupců zemí z tabulky entity s vyplněnou zemí vs hlavní sloupce z tabulky address, které jsou null. Zdroj: autor

Výpis 18. Doplnění názvu a kódu země z entity u řádků, které nemají adresu

```
#doplnění countries a country_codes z entity u řádků, které nejspíš nemají adresu
self_join_off.loc[(self_join_off['entity_countries'].notnull()) &
(~self_join_off['entity_countries'].str.contains(';', na=False)), 'countries_ad']
          = self_join_off.loc[(self_join_off['entity_countries'].notnull()) &
          (~self_join_off['entity_countries'].str.contains(';', na=False)),
'countries_ad'].fillna(self_join_off.loc[(self_join_off['entity_countries']
.notnull()) & (~self_join_off['entity_countries'].str.contains(';', na=False)),
          'entity_countries'])
self_join_off.loc[(self_join_off['entity_country_codes'].notnull()) &
(self_join_off['entity_country_codes'].str.len() <= 3), 'country_codes_ad'] =
self_join_off.loc[(self_join_off['entity_country_codes'].notnull()) &
(self_join_off['entity_country_codes'].str.len() <= 3),
'country_codes_ad'].fillna(self_join_off.loc[(self_join_off
          ['entity_country_codes'].notnull()) &
          (self_join_off['entity_country_codes'].str.len() <= 3),
          'entity_country_codes'])
```

Stejný postup jsem aplikovala u sloupců zemí z tabulek intermediary a officer.

Abych si byla jistá, že jsou názvy subjektů vyplněné ve všech řádcích u odpovídajících ID, rozhodla jsem se pomocí funkce forward_fill (*ffill*), která doplňuje chybějící hodnoty poslední nenulovou hodnotou, viz *Výpis 19*. Záznamy jsem nejprve seskupila podle odpovídajících ID. Jelikož považuji jurisdikci za důležitý sloupec, ve kterém však chybí poměrně hodně záznamů, které nemám odkud doplnit, rozhodla jsem se v tomto kroku zahrnout i tento sloupec, aby se případně doplnilo vše, co šlo.

Výpis 19. Seskupení a následné doplnění první nenulové hodnoty pomocí ffil() v jednotlivých skupinách, pokud zde nějaká existuje

```
self_join_off["entity_name"] =
          self_join_off.groupby("node_id")["entity_name"].ffill()

self_join_off["officer_name"] =
          self_join_off.groupby("node_id_start")["officer_name"].ffill()

self_join_off["intermediary_name"] =
          self_join_off.groupby("node_id_start")["intermediary_name"].ffill()

self_join_off["jurisdiction"] =
          self_join_off.groupby("node_id")["jurisdiction"].ffill()

self_join_off["jurisdiction_description"] =
          self_join_off.groupby("node_id")["jurisdiction_description"].ffill()
```

Když jsem chtěla začít tvořit vizualizaci v Tableau s daty, které jsem transformovala, zjistila jsem, že jsou ve sloupci countries země, které jsou jednou napsané celým názvem a podruhé zas pouze kódem země.

Countries	Country Codes ..
CYP	CYP
Cyprus	CYP
CZE	CZE
Czech Republic	CZE
Denmark	DNK
DEU	DEU
Djibouti	DJI
DNK	DNK
Dominica	DMA
Dominican Republic	DOM
DR Congo	COD
DZA	DZA
ECU	ECU
Ecuador	ECU

Obrázek 28. Ukázka názvů zemí, kdy je jedna země označená různým způsobem. Zdroj: autor

Jelikož jsem tento sloupec chtěla použít jako filtr pro uživatele, byly hodnoty velmi matoucí a uživatel by musel zvolit několik hodnot odpovídající jedné zemi, aby si mohl zobrazit, co potřebuje. Snažila jsem se takové hodnoty co nejvíce redukovat pomocí funkce *replace*, kde jsem po manuální revizi hodnot nahrazovala hodnoty kódů země jejím názvem, viz Výpis 20.

Výpis 20. Ukázka kódu pro nahrazení hodnot [20]

```
self_join_off['countries_ad'] = self_join_off['countries_ad'].replace({'CZE':
'Czech Republic', 'ARE': 'United Arab Emirates','AUT':'Austria',
          'BEL':'Belgium','BLZ':'Belize','BRA':'Brazil','CAN':'Canada',
          'CHL':'Chile', 'CYP':'Cyprus','DEU':'Germany','DNK':'Denmark',
          'EGY':'Egypt','ESP':'Spain', 'FRA':'France','GHA':'Ghana',
          'GRC':'Greece','HKG':'Hong Kong','HUN':'Hungary','IND':'India',
          'ITA':'Italy','JEY':'Jersey', 'KAZ':'Kazakhstan','KEN':'Kenya',
          'KNA':'Saint Kitts and Nevis','KOR':'South Korea',
          'Korea, Republic of':'South Korea', 'Republic of Korea':'South Korea',
          'LBN':'Lebanon','LCA':'Saint Lucia','LTU':'Lithuania',
          'LUX':'Luxembourg','LVA':'Latvia','MAR':'Maroco',
          'MDG':'Madagascar','MEX':'Mexico', 'MLT':'Malta','MNG':'Mongolia',
          'MWI':'Malawi','MYS':'Malaysia','NAM':'Namibia','NIC':'Nicaragua',
          'NOR':'Norway','NLD':'Netherlands','NZL':'New Zealand','OMN':'Oman',
          'PAK':'Pakistan','PAN':'Panama','PHL':'Philippines','POL':'Poland',
          'PRT':'Portugal','PSE':'Palestine','QAT':'Quatar','ROU':'Romania',
          ……})
```

Jedním z mých posledních kroků je zopakování kroku, kde doplním název země na základě hodnot ze sloupce *country_codes_ad* (viz Výpis 21) a naopak, jelikož po všech úpravách výše se mi v tabulce objevilo 459 156 řádků, kde název země chybí, ale mám k dispozici kód země a naopak 2 219 řádků, kde název země vyplněn je, ale není k němu přiřazen kód.

[20] Z důvodu délky celého kódu je zde zobrazena pouze část hodnot.

Výpis 21. Zopakování doplnění řádků, kde je prázdný název země, ale znám kód

```
# doplním countries na základě hodnot ze sloupce country_codes, jelikož po
úpravách výše mám v datech 459 156 řádků, kde je vyplněn country, ale ne
country_code
mask_countries = self_join_off['countries_ad'].isnull()

# seznam unikátních country_codes
unique_codes = self_join_off['country_codes_ad'].unique()

# dictionary mapping country codes na country names
country_name_dict = dict(self_join_off.dropna(subset=
['countries_ad']).groupby('country_codes_ad')['countries_ad'].first())

# mapuju dictionary na null hodnoty v countries_ad
self_join_off.loc[mask_countries, 'countries_ad'] =
self_join_off.loc[mask_countries, 'country_codes_ad'].map(country_name_dict)
```

Nakonec převedu všechny názvy subjektů na verzálky, jelikož se v datech občas objeví stejný název či jméno, ale kvůli odlišné podobě jsou zobrazené jako různé hodnoty, což ovlivňuje například filtry.

```
self_join_off['entity_name'] = self_join_off['entity_name'].str.upper()
self_join_off['officer_name'] = self_join_off['officer_name'].str.upper()
self_join_off['intermediary_name'] = self_join_off['intermediary_name'].str.upper()
```

V neposlední řadě odstraním nyní již přebytečné sloupce se zeměmi a tabulku uložím k dalšímu použití (viz *Výpis 22*).

Výpis 22. Odstranění nepotřebných sloupců a uložení finální tabulky

```
self_join_off.drop(
          ['entity_countries','entity_country_codes','officer_countries',
          'officer_country_codes','inter_countries',
          'inter_country_codes'],axis=1,inplace=True)

self_join_off.to_csv('final_dataset.csv', index=False)
```

4.5 Tvorba dashboardu

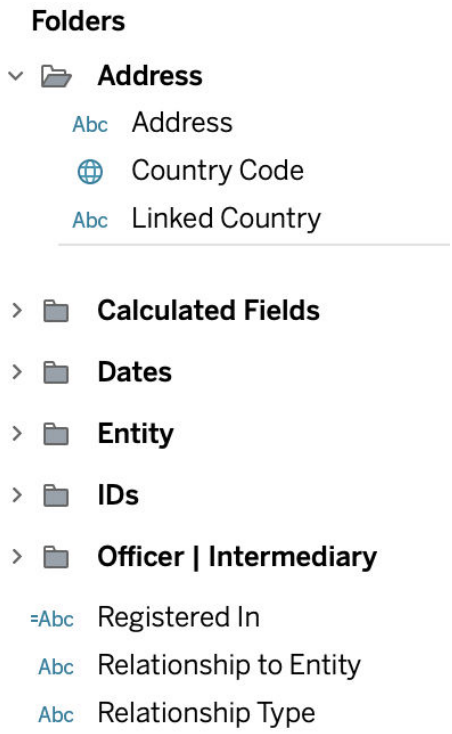
Můj dashboard se bude skládat ze dvou pohledů, resp. reportů. První report bude tvořit mapu, která bude zobrazovat lokaci jednotlivých entit. Druhý pohled bude doplňovat ten první v podobě tabulky s detaily o dané entitě. V další části popíšu postup při tvorbě jednotlivých reportů a jejich propojení v jeden dashboard.

Po nahrání finálního datového zdroje do Tableau Public jsem přejmenovala všechny pracovní názvy sloupců na jejich finální. Původní název sloupce a jeho nové znění lze vidět v tabulce níže. U nezmíněných sloupců došlo pouze k odstranění podtržítka v názvu sloupce.

Tabulka 8. Seznam sloupců, které mají změněný název v Tableau. Zdroj: autor

PŮVODNÍ NÁZEV	NOVÝ NÁZEV
inter_id	Intermediary ID
sourceID	Data From
jurisdiction	Jurisdiction Code
jurisdiction Description	Jurisdiction
valid_until	Valid Until
address_ad	Address
countries_ad	Linked Country
country_codes_ad	Country Code
start_date	Active From
end_date	To
link	Registered In
rel_type	Relationship Type
former_name	Former Name Orig
company_type	Company Type Orig

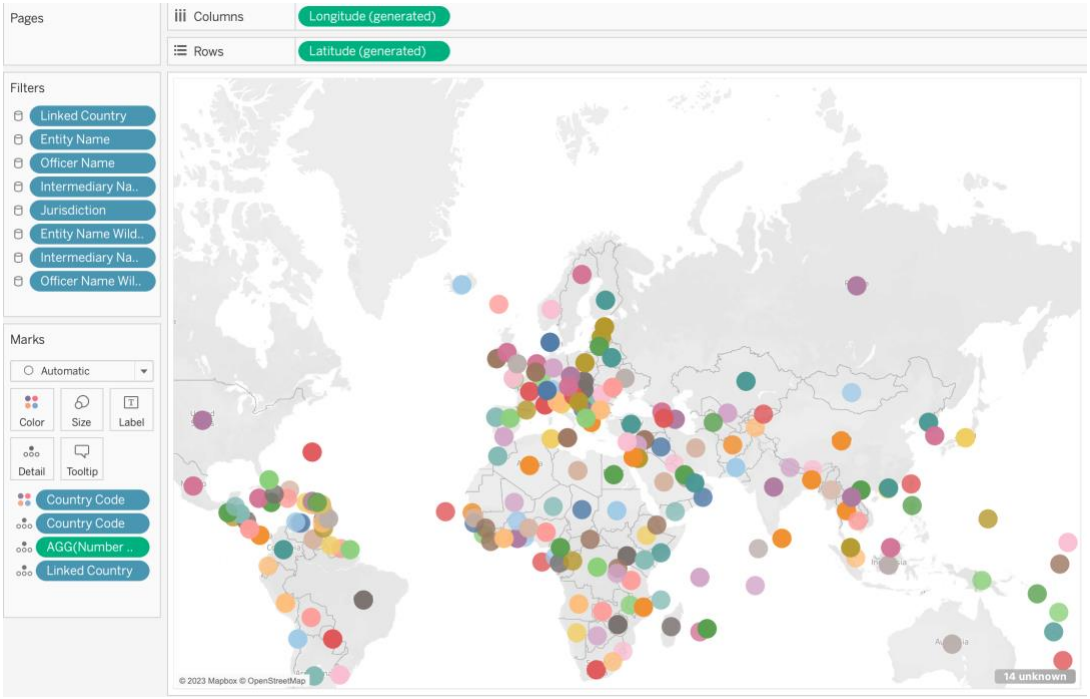
Dále jsem jednotlivé dimenze rozdělila do složek pro lepší orientaci.



Obrázek 29. Rozdělení dimenzí v Tableau do složek. Zdroj: autor

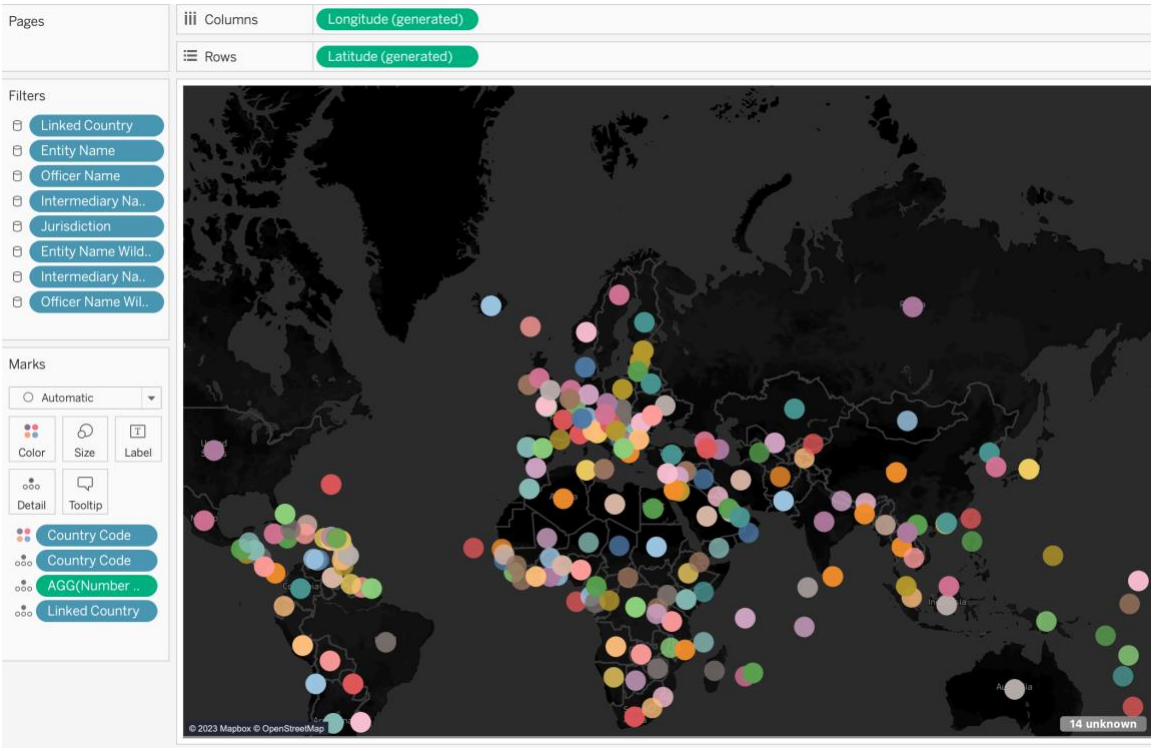
4.5.1 Mapa

Při tvorbě reportu v podobě mapy je pro mě nejdůležitější dimenze pro zobrazení geolokace, v mém případě se jedná o sloupec Country, resp. Country Code. Ačkoliv se mi podařilo minimalizovat ve sloupci Country názvy zemí, které byly jednou napsané celým názvem a podruhé pouze částečně či kódem země, některé hodnoty jsou zde stále duplicitní. Rozhodla jsem se proto zvolit pro zobrazení geolokace sloupec Country Code. Tableau dokáže automaticky na základě zadaného kódu země vygenerovat jeho zeměpisnou délku a šířku, tudíž tvorba samotné mapy nebyla složitá. Do detailu jsem vložila obě dimenze, a to jak Country, tak Country Code. Kromě toho jsem vytvořila kalkulované pole s názvem Number of Entities, které zobrazuje počet entit vyskytující se v dané zemi. Barevně rozlišuji jednotlivé země pro lepší přehlednost při filtrování.



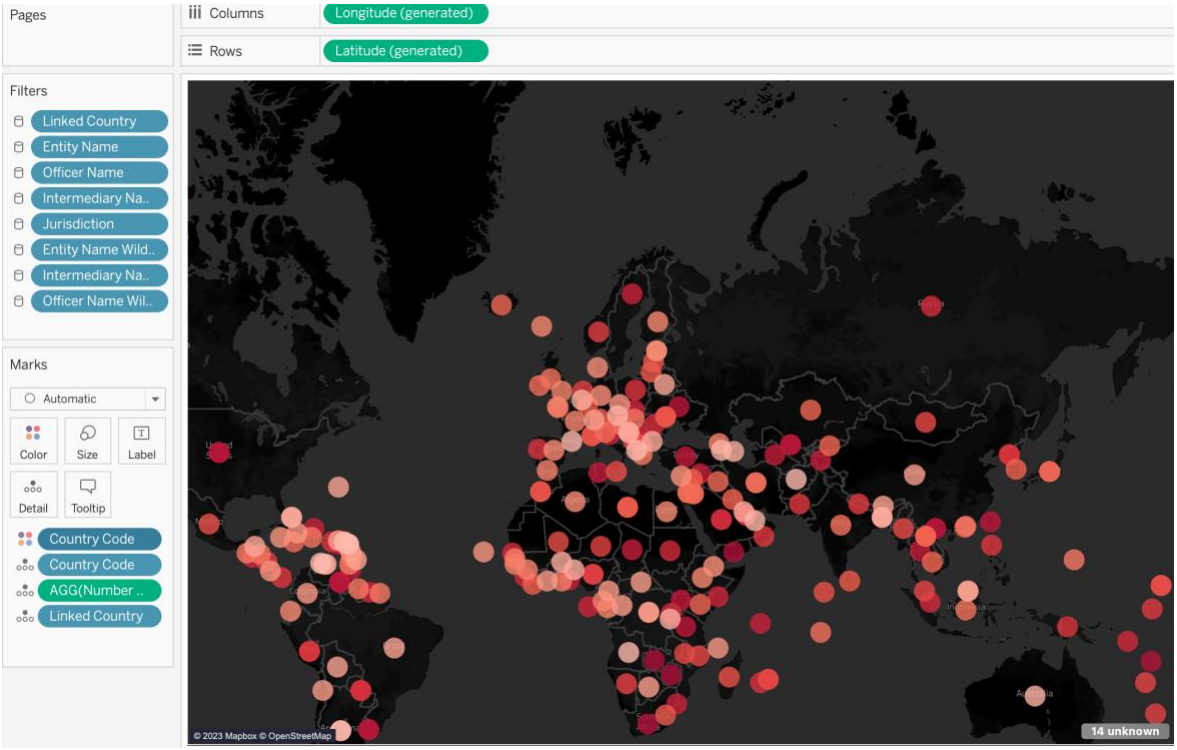
Obrázek 30. Tvorba základní mapy v Tableau. Zdroj: autor

Co se týče vizuální stránky reportu, a posléze i celého dashboardu, rozhodla jsem se vyjít z barev ICIJ a jejich loga, jelikož pracuji s jejich daty a chci, aby můj výsledný produkt ladil s jejich image. Tableau neumožňuje manuální úpravu vizuálu mapy, pouze přednastavené šablony, a tak jsem zvolila jejich tmavou variantu – Dark.



Obrázek 31. Úprava vizuálu mapy v Tableau na tmavou variantu. Zdroj: autor

Hlavními barvami ICIJ jsou černá a červená, barvy rozlišující jednotlivé země jsem změnila na odstíny červené. Na obrázku níže jsou zobrazeny všechny země bez filtrace, v praxi se bude jednat pouze o několik zemí dle toho, na který subjekt se chce uživatel zaměřit.



Obrázek 32. Další úprava vizuálu mapy, aby ladila s barvami ICIJ. Zdroj: autor

Tvorba a úprava tooltipu

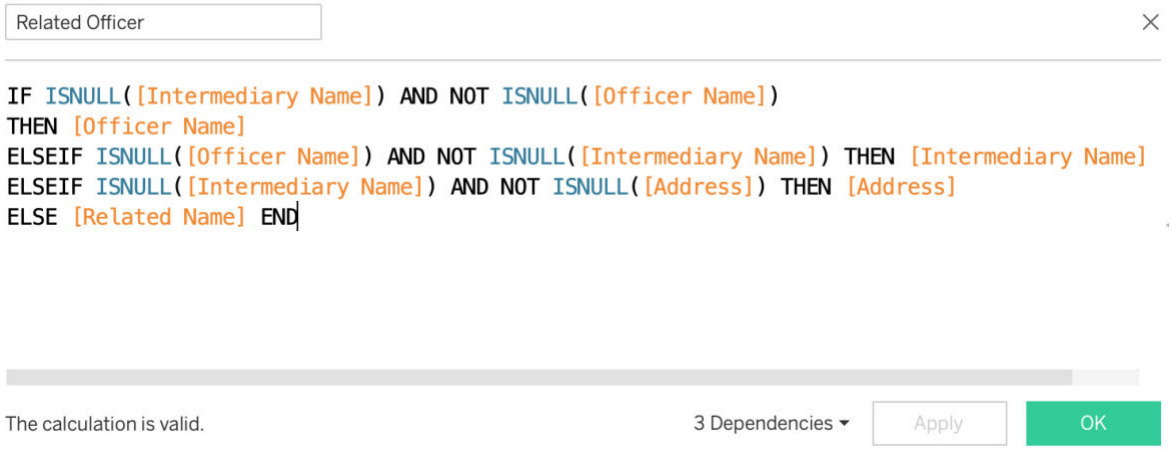
Při najetí kurzoru na jednotlivé země chci zobrazit následující údaje v tooltipu: počet entit v této zemi odpovídající filtrům, název a kód země, seznam entit v této zemi a všechny možné subjekty pojící se k této entitě, například ředitel, akcionář či zprostředkovatel. Počet entit, název a kód země lze zobrazit přehledně velmi jednoduše. Seznam entit a informace spojené s ní jsem se rozhodla podat v přehledné tabulce, kterou vytvořím zvlášť.

V tabulce bude název entity, subjekt, který s entitou jakýmkoliv způsobem souvisí a detail vztahu mezi subjektem a entitou. Vzhledem k tomu, že vztah mezi entitou a jiným uzlem může tvořit jak sloupec Officer Name, tak i Intermediary Name, Related Name či Address, znamenalo by to, že bych musela přidat všechny tyto sloupce zvlášť vedle sebe. To však způsobí, že pokud půjde o typ vztahu například „intermediary of“, objeví se nenulová hodnota pouze ve sloupci Intermediary Name, nikoliv ve sloupcích Address či Officer Name. V takovém případě by tyto sloupce byly zbytečné a zbytečně nafukují šířku tabulky.

Entity Name	Relationship to Entity	Intermediary Name	Officer Name	Address
ANTI-GRAVITY HOLDINGS LIMITED	beneficiary of	Null	WONG KIN HAY FELIX	Null
	intermediary of	GAWINA LIMITED	Null	Null
	shareholder of	Null	WONG KIN HAY FELIX	Null
ANTIGRAVITY INC.	director of	Null	LAM LAP KO	Null
	intermediary of	SINOMIX BUSINESS SOLU..	Null	Null
	registered address	Null	Null	SinoMix Business Solutio..

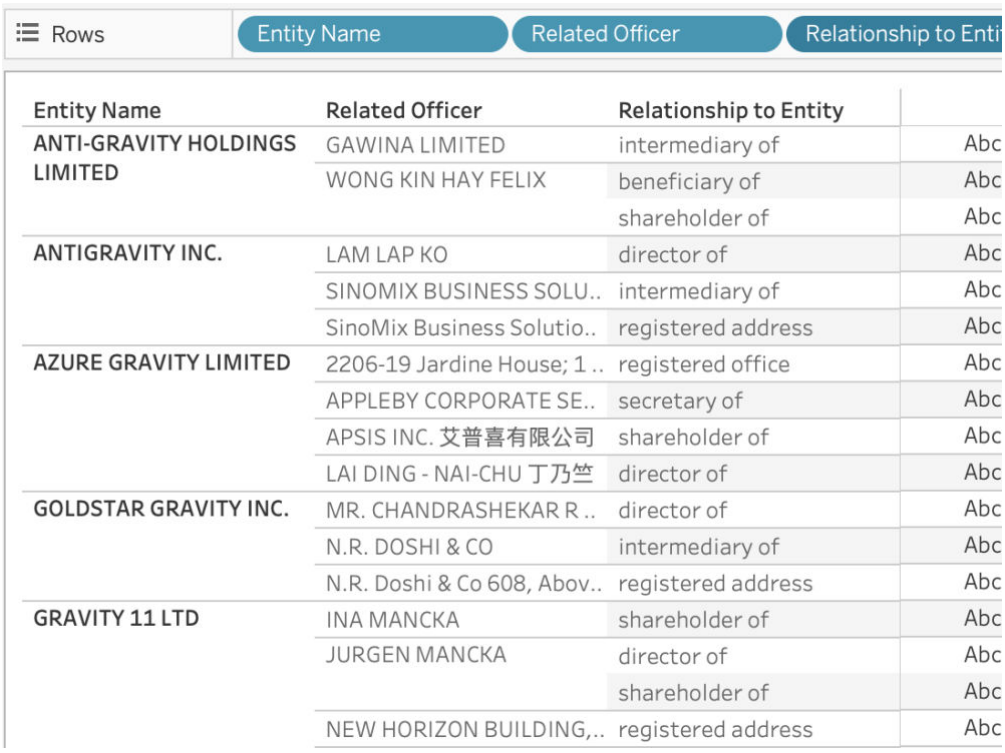
Obrázek 33. Ukázka, jak se v tabulce zobrazují názvy subjektů v několika sloupcích. Zdroj: autor

Vytvořila jsem proto kalkulované pole s názvem Related Officer, které bude vždy zobrazovat tu hodnotu, která je nenulová. Pomocí funkce IF kontroluji, zda je hodnota v jednom sloupci null, a pokud ano, najdu sloupec, který obsahuje nenulovou hodnotu a tu vrátím, viz *Obrázek 34*. Místo 4 sloupců mám nyní 1 sloupec.



Obrázek 34. Tvorba kalkulovaného pole pro Related Officer. Zdroj: autor

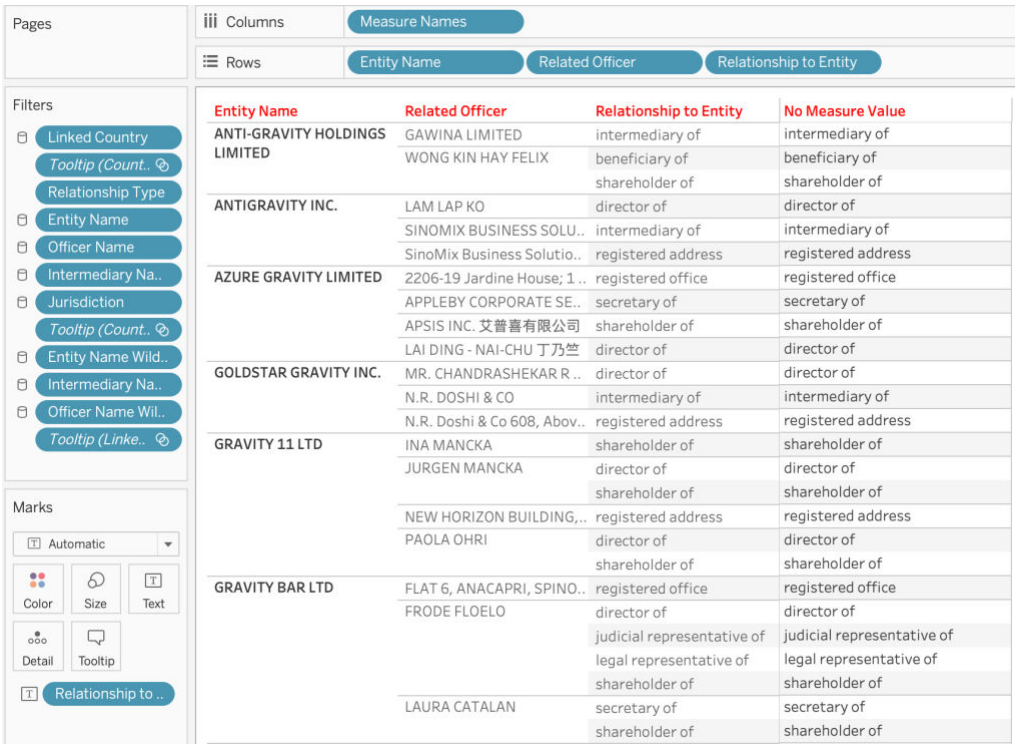
Jelikož Tableau počítá vždy alespoň s jednou metrikou v tabulce, při přidávání jednotlivých dimenzí se vždy objeví automaticky na konci prázdný sloupec, který obsahuje „abc“ zástupné symboly. Jelikož se moje tabulka bude skládat pouze z dimenzí, potřebuji tento sloupec vymazat. Z mého popisu obsahu tabulky vyplývá, že bude posledním sloupcem sloupec Relationship to Entity. To však způsobí již zmíněný problém, viz *Obrázek 35* níže.



Entity Name	Related Officer	Relationship to Entity	
ANTI-GRAVITY HOLDINGS LIMITED	GAWINA LIMITED	intermediary of	Abc
	WONG KIN HAY FELIX	beneficiary of	Abc
		shareholder of	Abc
ANTIGRAVITY INC.	LAM LAP KO	director of	Abc
	SINOMIX BUSINESS SOLU..	intermediary of	Abc
	SinoMix Business Solutio..	registered address	Abc
AZURE GRAVITY LIMITED	2206-19 Jardine House; 1 ..	registered office	Abc
	APPLEBY CORPORATE SE..	secretary of	Abc
	APSIS INC. 艾普喜有限公司	shareholder of	Abc
	LAI DING - NAI-CHU 丁乃竺	director of	Abc
GOLDSTAR GRAVITY INC.	MR. CHANDRASHEKAR R ..	director of	Abc
	N.R. DOSHI & CO	intermediary of	Abc
	N.R. Doshi & Co 608, Abov..	registered address	Abc
GRAVITY 11 LTD	INA MANCKA	shareholder of	Abc
	JURGEN MANCKA	director of	Abc
		shareholder of	Abc
	NEW HORIZON BUILDING,..	registered address	Abc

Obrázek 35. Tabulka v Tableau bez metriky, která obsahuje zástupný sloupec. Zdroj: autor

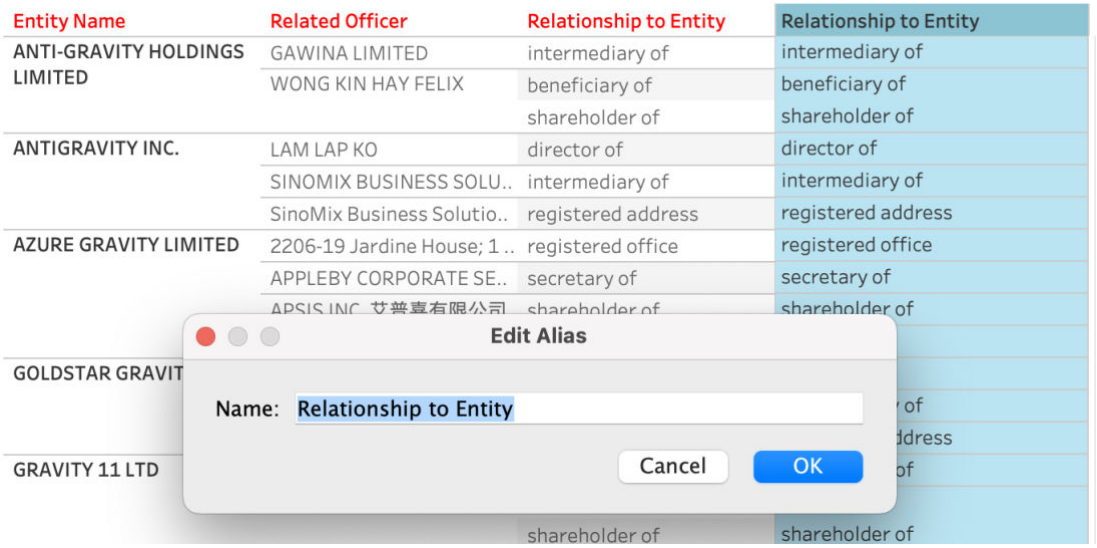
Tento sloupec jde skrýt následujícím způsobem: dimenzi Relationship To Entity kromě řádků vložím také do pole pro metriky jako text. Dále do sekce sloupců vložím dimenzi Measure Names. Poslední zástupný sloupec sice zmizel, nicméně mám teď poslední sloupec duplikovaný s názvem No Measure Value.



Entity Name	Related Officer	Relationship to Entity	No Measure Value
ANTI-GRAVITY HOLDINGS LIMITED	GAWINA LIMITED	intermediary of	intermediary of
	WONG KIN HAY FELIX	beneficiary of	beneficiary of
		shareholder of	shareholder of
ANTIGRAVITY INC.	LAM LAP KO	director of	director of
	SINOMIX BUSINESS SOLU..	intermediary of	intermediary of
	SinoMix Business Solutio..	registered address	registered address
AZURE GRAVITY LIMITED	2206-19 Jardine House; 1 ..	registered office	registered office
	APPLEBY CORPORATE SE..	secretary of	secretary of
	APSIS INC. 艾普喜有限公司	shareholder of	shareholder of
	LAI DING - NAI-CHU 丁乃竺	director of	director of
GOLDSTAR GRAVITY INC.	MR. CHANDRASHEKAR R ..	director of	director of
	N.R. DOSHI & CO	intermediary of	intermediary of
	N.R. Doshi & Co 608, Abov..	registered address	registered address
GRAVITY 11 LTD	INA MANCKA	shareholder of	shareholder of
	JURGEN MANCKA	director of	director of
		shareholder of	shareholder of
	NEW HORIZON BUILDING,..	registered address	registered address
	PAOLA OHRI	director of	director of
		shareholder of	shareholder of
GRAVITY BAR LTD	FLAT 6, ANACAPRI, SPINO..	registered office	registered office
	FRODE FLOELO	director of	director of
		judicial representative of	judicial representative of
		legal representative of	legal representative of
		shareholder of	shareholder of
	LAURA CATALAN	secretary of	secretary of
		shareholder of	shareholder of

Obrázek 36. Odstranění zástupného sloupce v Tableau přidáním Measure Names mezi sloupce. Zdroj: autor

Tento problém spravím tak, že nastavím poslednímu sloupci alias.



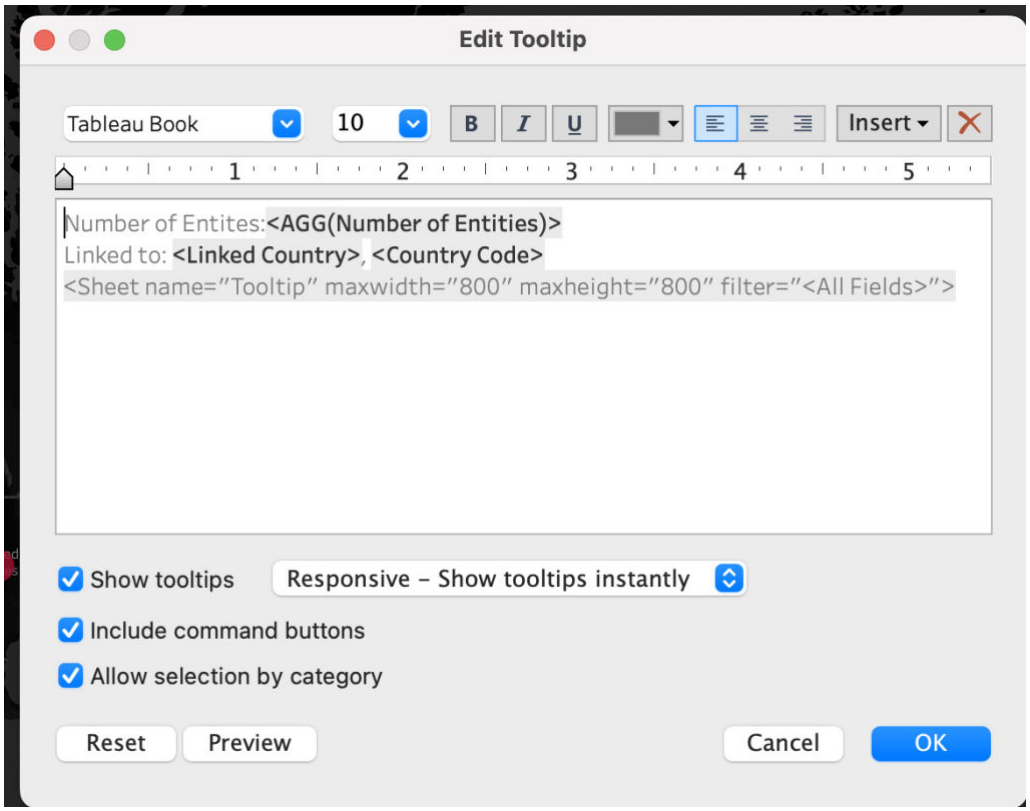
Obrázek 37. Přejmenování sloupce Measure Names, aby korespondoval se zobrazovanou informací. Zdroj: autor

Nyní vyberu a otevřu možnosti v sekci řádků dimenze Relationship to Entity a odkliknu možnost Show Header. Tím duplikovaný sloupec zmizí.

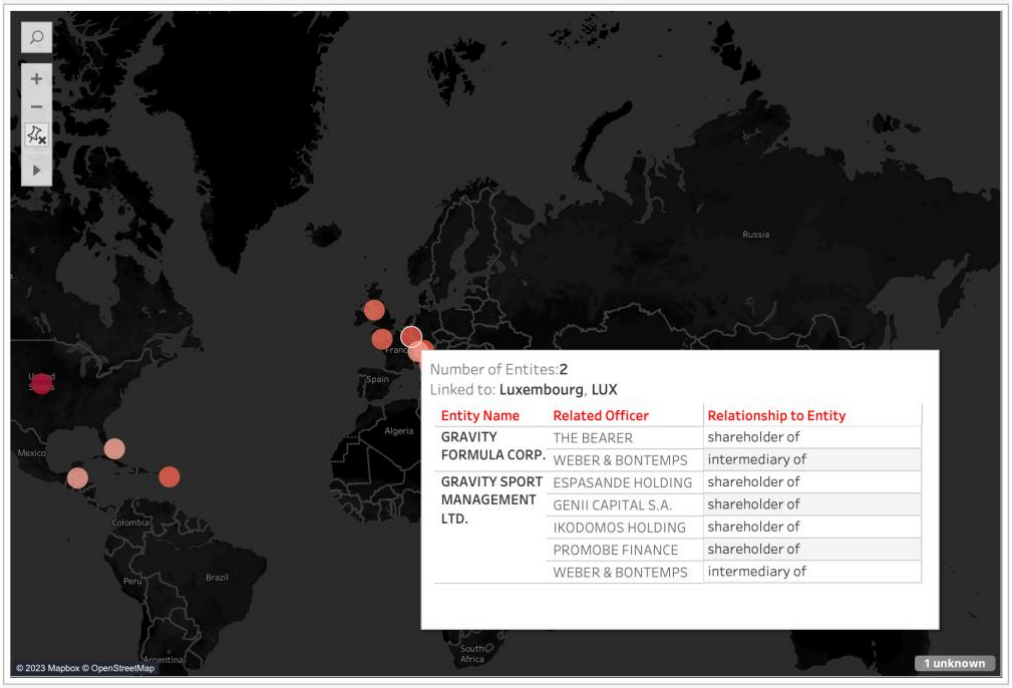
Entity Name	Related Officer	Relationship to Entity
ANTI-GRAVITY HOLDINGS LIMITED	GAWINA LIMITED	intermediary of
	WONG KIN HAY FELIX	beneficiary of
		shareholder of
ANTIGRAVITY INC.	LAM LAP KO	director of
	SINOMIX BUSINESS SOLU..	intermediary of
	SinoMix Business Solutio..	registered address
AZURE GRAVITY LIMITED	2206-19 Jardine House; 1 ..	registered office
	APPLEBY CORPORATE SE..	secretary of
	APSIS INC. 艾普喜有限公司	shareholder of
	LAI DING - NAI-CHU 丁乃竺	director of
GOLDSTAR GRAVITY INC.	MR. CHANDRASHEKAR R ..	director of
	N.R. DOSHI & CO	intermediary of
	N.R. Doshi & Co 608, Abov..	registered address
GRAVITY 11 LTD	INA MANCKA	shareholder of

Obrázek 38. Finální podoba tooltip tabulky. Zdroj: autor

Hotovou tabulku nyní použiji jako součást tooltipu k již vytvořené mapě.



Obrázek 39. Vložení tabulky do tooltipu mapy. Zdroj: autor

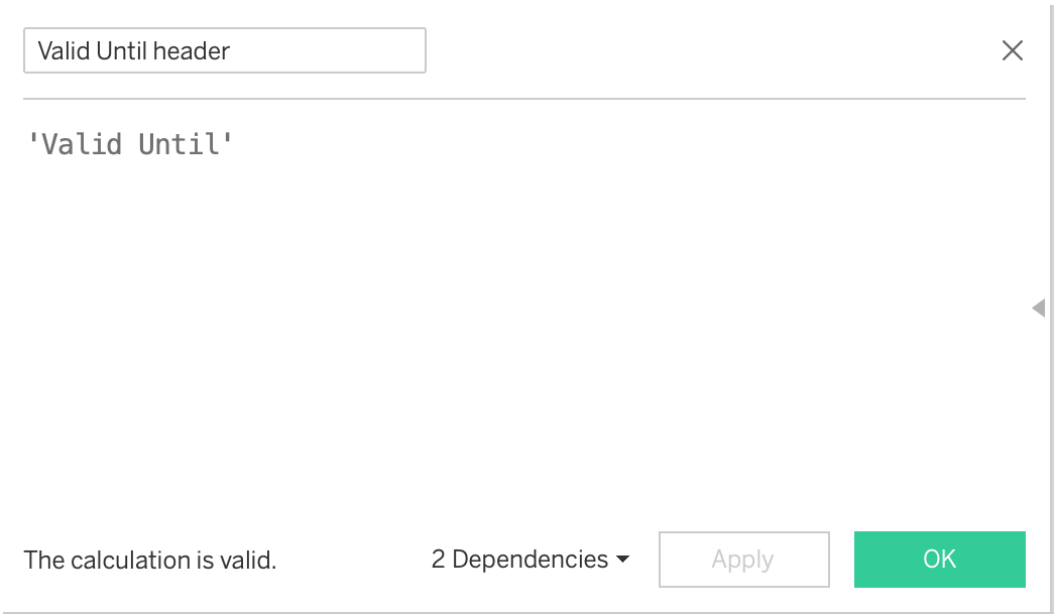


Obrázek 40. Ukázka finální podoby tooltipu při užívání mapy. Zdroj: autor

4.5.2 Detailní tabulka

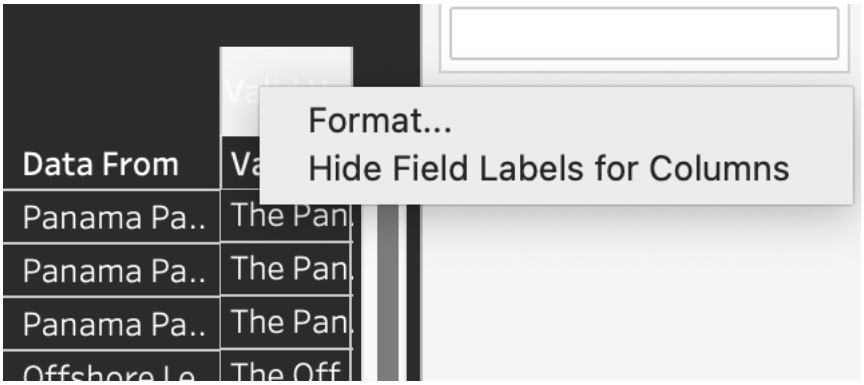
Jelikož ke každé entitě existuje mnohem více informací, než je možné zobrazit v tooltipu, rozhodla jsem se vytvořit detailní tabulku, která bude doplňovat informace z prvního reportu, ale zároveň může být použita samostatně pro případný export a další užití. Zároveň bude k jednotlivým entitám zobrazovat informace i přesto, že bude k entitě chybět název či kód země.

V tabulce chci zobrazit ke každé entitě její typ, adresu a samozřejmě všechny osoby, které s danou entitou mají nějaký vztah. Dále chci zobrazit dobu, od které je daný vztah aktivní, příp. kdy byl ukončen. Zobrazím také informaci o názvu a kódu státu, přestože by měla být tato informace viditelná v mapě, ze dvou důvodů – pokud se v mapě daná entita nezobrazí, chci, aby uživatel věděl, že je to z důvodu chybějící hodnoty v názvu/kódu země. Druhým důvodem je přehled uživatele v záznamech v případě, že by potřeboval zobrazenou tabulku exportovat například do excelové tabulky. Jako další informaci chci zobrazit jurisdikci, v rámci které je entita zaregistrována a datum jejího vzniku. V neposlední řadě chci zobrazit informace o zdroji úniku dat, ze kterého je daný záznam a informace o případném zániku entity či její deaktivace. Tabulku tvořím jednoduchým přidáváním dimenzí do sekce rows. Stejně jako při tvorbě tabulky pro tooltip řeším problém s prázdným sloupcem a jejími zástupnými symboly „abc“. Nemohu však použít stejné řešení, jako v tabulce Tooltip, jelikož jsem zde zadávala alias pro název sloupce „Measure Names“, který je globální, a tedy může být pojmenovaný pouze pod jedním názvem. V tabulce tooltip je ale posledním sloupec Relationship to Entity, kdežto v detailní tabulce tvoří poslední sloupec Valid Until a název by tedy neodpovídal pro oba sloupce. Vytvořila jsem si proto kalkulované pole s názvem Valid Until header, který obsahuje pouze string s názvem sloupce tak, jak chci, aby byl zobrazený v tabulce.



Obrázek 41. Vytvoření kalkulovaného pole s názvem Valid Until header. Zdroj: autor

Poté sloupec Valid Until (tedy ten původní s hodnotami) vložím do tabulky jako Text a kalkukované pole naopak vložím do sekce Columns. Tímto se v posledním sloupci zobrazí název kalkulovaného pole, ale i hodnota, kterou jsem do něj napsala. Název kalkulovaného pole můžu vypnout pomocí možnosti Hide Field Labels for Columns.



Obrázek 42. Vypnutí názvu kalkulovaného pole v reportu. Zdroj: autor

Nyní mám tabulku kompletní a správně. Na závěr vizuál tabulky naformátuji tak, aby ladila barevně s mapou.

DETAIL ABOUT CHOSEN ENTITY

Entity Name	Company Type	Related Officer	Relationship to Entity	Active From	To	Linked Country	Country Code	Registered In	Incorporation Date	Struck Off Date	Inactivation Date	Dorm Date	Data From	Valid Until
ANTI-GRAVITY HOLDINGS ..	-	GAWINA LI..	intermediar..	Null	Null	Hong Kong	HKG	British Virg..	12/01/1999	Null	Null	Null	Panama Pa..	The Pan..
		WONG KIN HAY FELIX	beneficiary ..	Null	Null	Hong Kong	HKG	British Virg..	12/01/1999	Null	Null	Null	Panama Pa..	The Pan..
			shareholde..	Null	Null	Hong Kong	HKG	British Virg..	12/01/1999	Null	Null	Null	Panama Pa..	The Pan..
ANTIGRAVITY INC.	Business Company Limited by ..	LAM LAP KO	director of	28/06/2006	Null	Not identifi..	XXX	Undetermin..	28/06/2006	Null	Null	Null	Offshore Le..	The Off..
		SINOMIX B..	intermediar..	Null	Null	Hong Kong	HKG	Undetermin..	28/06/2006	Null	Null	Null	Offshore Le..	The Off..
		SinoMix Bu..	registered ..	Null	Null	Hong Kong	HKG	Undetermin..	28/06/2006	Null	Null	Null	Offshore Le..	The Off..
AZURE GRAVITY LIMITED	-	2206-19 Jar..	registered ..	Null	Null	Hong Kong	HKG	Hong Kong	28/08/2014	Null	Null	Null	Paradise Pa..	Appleb..
		APPLEBY C..	secretary of	Null	Null	Hong Kong	HKG	Hong Kong	28/08/2014	Null	Null	Null	Paradise Pa..	Appleb..
		APSIS INC. ..	shareholde..	Null	Null	Hong Kong	HKG	Hong Kong	28/08/2014	Null	Null	Null	Paradise Pa..	Appleb..
		LAI DING - ..	director of	Null	Null	Hong Kong	HKG	Hong Kong	28/08/2014	Null	Null	Null	Paradise Pa..	Appleb..
GOLDSTAR GRAVITY INC.	Business Company Limited by ..	MR. CHAND..	director of	19/09/2006	Null	Not identifi..	XXX	Undetermin..	19/09/2006	Null	Null	01/05/2008	Offshore Le..	The Off..
		N.R. DOSHI ..	intermediar..	Null	Null	United Arab..	ARE	Undetermin..	19/09/2006	Null	Null	01/05/2008	Offshore Le..	The Off..
		N.R. Doshi ..	registered ..	Null	Null	United Arab..	ARE	Undetermin..	19/09/2006	Null	Null	01/05/2008	Offshore Le..	The Off..
GRAVITY 11 LTD	-	INA MANCKA	shareholde..	Null	Null	Switzerland	CHE	Belize	Null	Null	Null	Null	Pandora Pa..	Provide..
		JURGEN MANCKA	director of	Null	Null	Italy	ITA	Belize	Null	Null	Null	Null	Pandora Pa..	Provide..
			shareholde..	Null	Null	Italy	ITA	Belize	Null	Null	Null	Null	Pandora Pa..	Provide..
		NEW HORIZ..	registered ..	Null	Null	Belize	BLZ	Belize	Null	Null	Null	Null	Pandora Pa..	Provide..
		PAOLA OHRI	director of	Null	Null	Albania	ALB	Belize	Null	Null	Null	Null	Pandora Pa..	Provide..
			shareholde..	Null	Null	Albania	ALB	Belize	Null	Null	Null	Null	Pandora Pa..	Provide..
GRAVITY BAR LTD	-	FLAT 6, AN..	registered ..	Null	Null	Malta	MLT	Malta	22/05/2009	Null	Null	Null	Paradise Pa..	Malta c..
		FRODE	director of	Null	Null	Malta	MLT	Malta	22/05/2009	Null	Null	Null	Paradise Pa..	Malta c..

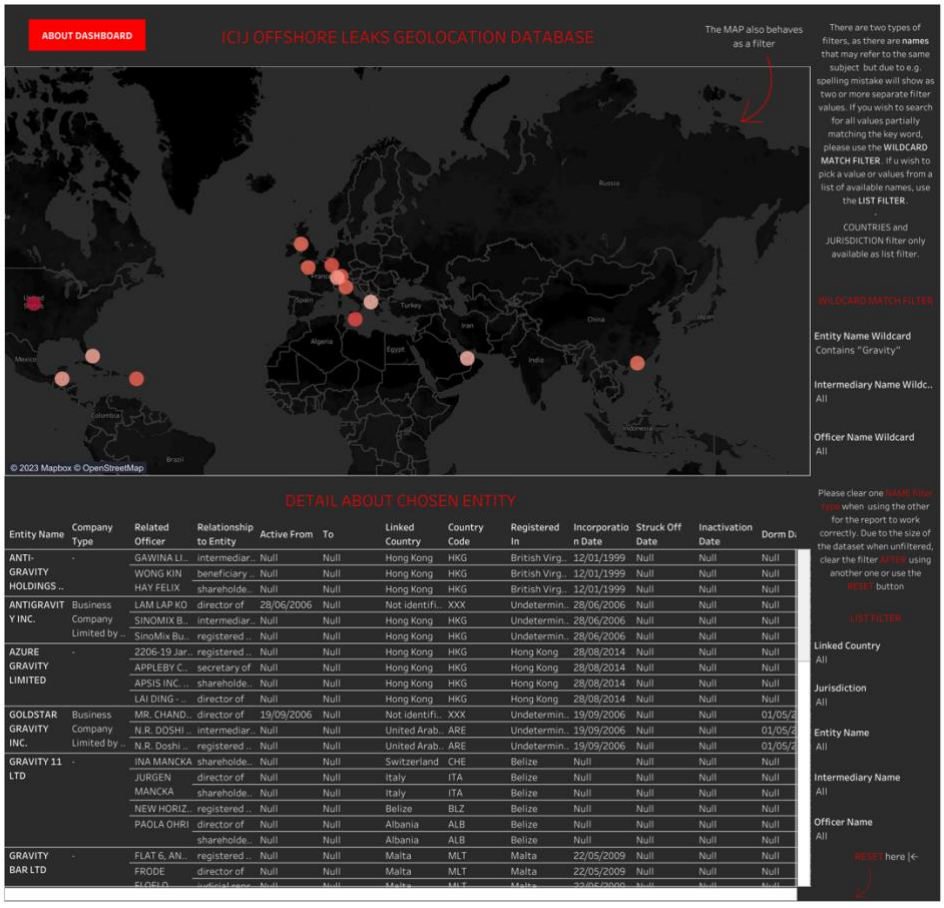
Obrázek 43. Finální podoba reportu s detailní tabulkou. Zdroj: autor

4.5.3 Finální dashboard

Základní pohledy mám vytvořené, nyní zbývá sestavit finální dashboard, resp. databázi. Horní část dashboardu bude tvořit můj hlavní report, tedy mapa zobrazující polohu subjektů. Pod ní bude tabulka s detaily doplňující údaje z mapy. Po pravé straně pak budou filtry umožňující uživatelům vyhledávat dle entit, úředníků, zprostředkovatelů i podle státu a jurisdikce.

Filtrování entit, úředníků a zprostředkovatelů umožňuji ve dvou formách – filtrování ze seznamu a wildcard filtrování, které vrátí všechny hodnoty částečně odpovídající vyhledávanému slovu. Vybrala jsem více způsobů, jelikož chci co nejvíce vyhovět potřebám uživatele. Filtrování ze seznamu považuji za užitečné, pokud se jedná o uživatele, který chce

vědět, jaké možné hodnoty se objevují v databázi. Wildcard filtrování naopak považuji za užitečné, pokud již uživatel hledá někoho konkrétního, ale nevybaví si celý název, případně chce vidět širší oblast odpovídajících záznamů. I při filtrování ze seznamu lze zobrazit vyhledáváním všechny hodnoty odpovídající hledané frázi, ale pokud by šlo o velké množství hodnot, uživatel by musel všechny zaklikávat ručně a není to tedy příliš uživatelsky přívětivé. Nabízím tímto uživatelům variabilitu ve vyhledávání. Oba typy filtrů však fungují závisle na sobě a vrací průnik hodnot všech filtrů, a proto je třeba nejprve jeden typ filtru odstranit poté, co je použit druhý, aby byl dashboard zafiltrován správně, ale zároveň byl vždy alespoň jeden filtr aktivní vzhledem k velikosti datového zdroje. K seznamu filtrů přibude proto také návod, jak filtry používat. *Obrázek 44* zobrazuje výsledný produkt. Dashboard je k dispozici na webu Tableau Public viz *Příloha* .



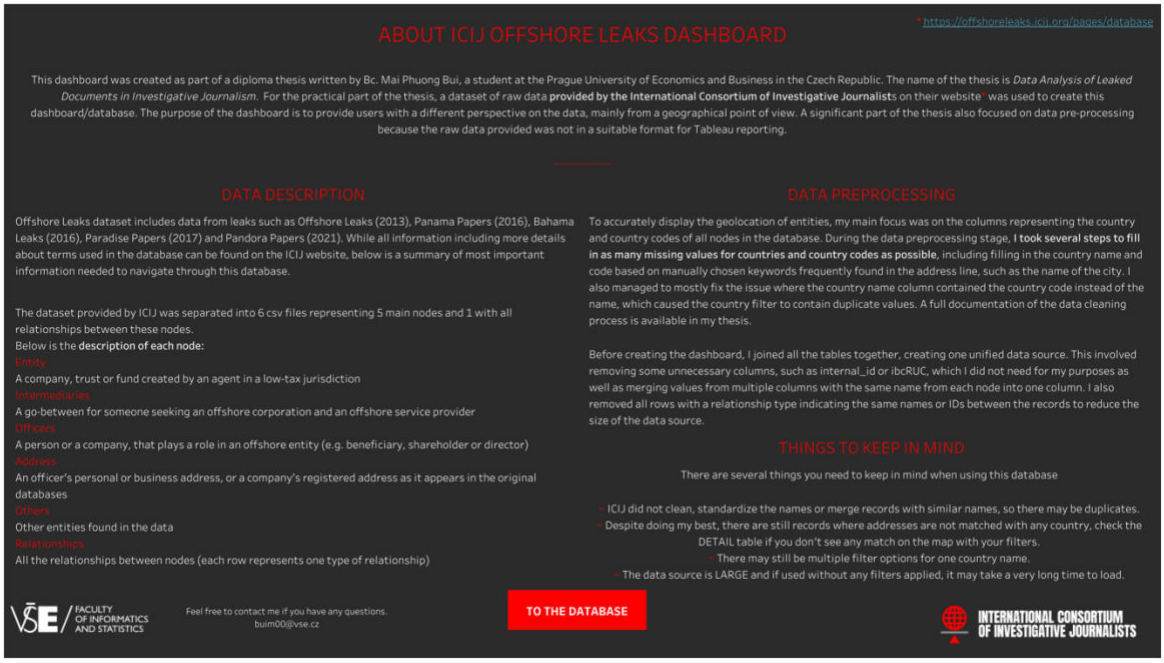
Obrázek 44. Podoba finálního dashboardu. Zdroj: autor

V levém horním rohu je tlačítko, které uživatele přenese na kartu O dashboardu (angl. About Dashboard).

O dashboardu

Soubor Offshore Leaks je obrovský a plný komplexních dat, se kterými jsem provedla poměrně hodně úprav a která navíc patří jiné organizaci. Původ dat a jejich současný stav a účel by mohl být proto matoucí. Rozhodla jsem se vytvořit kartu, ve které popisuji zdroj dat, důvod vzniku tohoto dashboardu a základní informace o datech, které by uživatel měl znát, než s nimi začne pracovat. Téměř všechny informace z této karty jsou dostupné na

webu ICIJ, přidanou hodnotu tvoří právě v tom, že je uživatel nemusí nikde hledat, a pokud ho zajímají detaily, přikládám zde také odkaz přímo k původním datům organizace. Ve zkratce také uživateli sděluji, že jsou původní data mnou upravená a vyčištěná, vzhledem k velkému množství úprav a limitovanému prostoru jsem však příliš nespecifikovala konkrétní úpravy. Uživateli však minimálně dávám tímto sdělením najevo, že mohou má data obsahovat, alespoň co se týče geolokace, o něco více informací než původní data.



Obrázek 45. Finální podoba karty O Dashboardu. Zdroj: autor

Závěr

Cílem této diplomové práce byla tvorba interaktivní databáze, která zobrazuje síť offshore firem z pěti velkých souborů uniklých dokumentů, a to Pandora Papers, Paradise Papers, Bahamas Leaks, Panama Papers a Offshore Leaks poskytnutých ICIJ.

Pro dosažení tohoto cíle bylo třeba v teoretické části nejprve seznámit čtenáře s principy investigativní žurnalistiky a metodiky sběru, klasifikace, verifikace a ochrany dat v této oblasti. Díky této části získá čtenář lepší povědomí nejen o investigativních novinářích, díky kterým existuje soubor dat, se kterými pracuji v praktické části, ale také o tom, jakým způsobem pracují s daty.

V praktické části se pak zabývám datovými zdroji a tvorbou samotné databáze. Velkou část zde tvoří proces předzpracování dat, jelikož bylo potřeba podniknout mnoho kroků k tomu, aby bylo možné pracovat s daty z pohledu geolokace. Hlavním problémem zde byly chybějící hodnoty u sloupců indikující geolokaci, jako je název země a kód země, případně chybějící či špatně napsaná adresa.

Data jsem čistila prostřednictvím Pythonu v Jupyter Notebooku. Jelikož byla data od ICIJ rozdělená do 6 souborů a každý reprezentoval jeden z hlavních uzlů vystupujících v databázi, čistila jsem nejprve jednotlivé soubory zvlášť a poté tabulky postupně propojila dohromady. Jeden ze souborů obsahoval přes 3 miliony záznamů představující všechny existující vazby mezi ostatními uzly, a tvořil tak vazební tabulku mezi všemi ostatními tabulkami navzájem.

Při čištění dat v souboru address jsem narážela na různé problémy, první problém jsem řešila v momentě, kdy jsem zjistila, že existují v souboru dva různé pojmenované sloupce (address a name), které obsahují stejnou informaci – adresu – ale občas se nachází hodnota v jednom sloupci a občas v druhém. Rozhodla jsem se ponechat sloupec address, a pokud ve sloupci chyběla hodnota, která byla k dispozici v druhém, doplnila jsem ho do sloupce address. Stejně tak jsem nahradila hodnoty ve všech řádcích, kde byla délka řetězce ze sloupce name delší než ve sloupci address, jelikož to znamenalo, že obsahuje detailnější adresu.

Na další problém jsem narazila, když jsem zjistila, že přestože existovala adresa, chyběl název a kód země. Důvodem mohla být chybně napsaná adresa, což znemožnilo geokódéru identifikovat automaticky název a kód země, nebo se ve sloupci vůbec nejednalo o adresu, ale jiný údaj, který s ní vůbec nesouvisí. U takových záznamů pak nebylo možné doplnit název a kód země ani manuálně, jelikož byla adresa prakticky neznámá. Pokud byla adresa pouze chybně napsaná, ale stále obsahovala určitá klíčová slova, pomocí kterých jsem mohla identifikovat název a kód země, pokusila jsem se hodnoty vyplnit manuálně, resp. jsem provedla manuální revizi záznamů a pokud se v adrese vyskytovalo určité klíčové slovo (nejčastěji název města), přiřadila jsem veškeré záznamy obsahující toto klíčové slovo k určité zemi. Pomocí několika dalších úprav se mi podařilo snížit chybějící hodnoty ve sloupci s názvem a kódem země ze 125 328 na 7 569.

Dílčí tabulky jsem po čištění a úpravách spojila v jeden datový zdroj, který byl poté použit pro tvorbu dashboardu v nástroji Tableau. Finální databáze se skládá ze dvou částí, samotným dashboardem a kartou s popisem dashboardu.

Hlavní dashboard se skládá ze dvou reportů. První report je mapa, která zobrazuje lokaci jednotlivých entit, v detailu při najetí na danou zemi jsou poté souhrnné informace o entitách nacházející se v této zemi a všech osobách, které mají vazbu na tuto entitu. Druhý report je detailní tabulka, která obsahuje další informace k daným entitám, jako je datum vzniku, případně zániku, typ společnosti či zdroj uniklých dokumentů, ze kterého byla tato informace získána. Tato tabulka existuje také pro případy 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.

Karta s popisem dashboardu existuje jako úvod a návod pro použití databáze. Bylo pro mě velmi důležité sdělit uživateli, díky komu tato data existují v první řadě, ale také vysvětlit, že jsou data mnou upravená a čištěná, což znamená, že jsou v mé databázi u některých entit informace o zemi, které v databázi ICIJ v současné době chybí. Celá databáze je vizualizována do barev ICIJ, jelikož jsem chtěla ladit s jejich organizací, oceňuji jejich práci a chci, aby bylo znát, že databáze vznikla především díky tomu, co dělají.

Svého cíle jsem v této diplomové práci dosáhla a v současné době existuje na platformě Tableau Public veřejná interaktivní databáze offshore firem, kterou může kdokoliv využít ke svým potřebám pod názvem ICIJ Offshore Leaks Geolocation Database.

Použitá literatura

[1] ISMAIL, Adibah, Mohd Khairie AHMAD a Che Su MUSTAFFA. Conceptualization of Investigative Journalism: The Perspectives of Malaysian Media Practitioners. *Procedia - Social and Behavioral Sciences* [online]. 2014, **155**, 165–170. ISSN 18770428. Dostupné z: doi:10.1016/j.sbspro.2014.10.274

[2] DAVID E. KAPLAN. What Is Investigative Journalism? *Global Investigative Journalism Network* [online]. 2023 [vid. 2023-01-21]. Dostupné z: https://gijn.org/investigative-journalism-defining-the-craft

[3] HUNTER, Mark Lee. *Story-based inquiry: a manual for investigative journalists - UNESCO Digital Library* [online]. Francie: UNESCO, 2011 [vid. 2023-01-21]. ISBN 978-92-3-104189-1. Dostupné z: https://unesdoc.unesco.org/ark:/48223/pf0000193078.nameddest=193103

[4] WEINBERG, Steve a INVESTIGATIVE REPORTERS AND EDITORS, INC, ed. *The reporter's handbook: an investigator's guide to documents and techniques*. 3rd ed. New York: St. Martin's Press, 1996. ISBN 978-0-312-13596-6.

[5] KMENTA, Jaroslav. INVESTIGATIVNÍ ŽURNALISTIKA. *Jaroslav Kmenta* [online]. 2023 [vid. 2023-01-21]. Dostupné z: https://www.kmenta.cz/investigativni-zurnalistika

[6] DUBINOVÁ, Tereza. *S Pavlou Holcovou o investigativní žurnalistice I.* [online]. 19. duben 2022 [vid. 2023-01-21]. Dostupné z: http://www.oheladom.cz/2022/puvodni-rozhovory/s-pavlou-holcovou-o-investigativni-zurnalistice-i/

[7] POLÁCH, Vladimír. Důvěryhodnost média a práce s informacemi. In: *Mediální teorie a praxe*. 1. vyd. Olomouc: niverzita Palackého v Olomouci, 2008. ISBN 978-80-244-2056-1.

[8] SOUŠOVÁ, Monika. Přiznání zuřivého reportéra Josefa Klímy (71) Mně lžou lidi pořád! [online]. 2022. Dostupné z: https://www.ahaonline.cz/clanek/199254/priznani-zuriveho-reportera-josefa-klimy-71-mne-lzou-lidi-porad

[9] HOLUBOVÁ, Hana. Josef Klíma je další novou tváří Televize Seznam. Se svým týmem chystá investigativní pořad. *Blog Seznam.cz* [online]. 3. červenec 2018 [vid. 2023-01-22]. Dostupné z: https://blog.seznam.cz/2018/07/josef-klima-je-dalsi-novou-tvari-televize-seznam-se-svym-tymem-chysta-investigativni-porad/

[10] ČESKÉ PODSVĚTÍ. Epizoda čtvrtá: I z člověka odsouzeného za dvojnásobnou vraždu může být celebrita - Seznam Zprávy. In: [online]. [vid. 2023-01-24]. Dostupné z: https://www.seznamzpravy.cz/clanek/znacka-kajinek-jak-ceska-media-pomohla-ze-zlocince-vyrobit-celebritu-113689

[11] SEZNAM ZPRÁVY. Záhady Josefa Klímy: Vyšetřování údajných úmrtí novorozenců za totality. *Seznam Zprávy* [online]. 27. květen 2022 [vid. 2023-01-24]. Dostupné z: https://www.seznamzpravy.cz/clanek/porady-zahady-zahady-josefa-klimy-vysetrovani-udajnych-umrti-novorozencu-za-totality-204062

[12] KOSTKOVÁ, Tereza. Josef Klíma: Žijeme v právním státě. I když lumpové mají až příliš mnoho možností, jak se bránit. *iROZHLAS* [online]. 1. listopad 2022 [vid. 2023-01-22]. Dostupné z: https://www.irozhlas.cz/zivotni-styl/spolecnost/josef-klima-zlocin-detektivky-pravni-stat-mafie-kniha_2211010911_kac

[13] SEZNAM ZPRÁVY. *Janek Kroupa - Seznam Zprávy* [online]. 2023 [vid. 2023-03-14]. Dostupné z: https://www.seznamzpravy.cz/autor/janek-kroupa-687

[14] TACHECÍ, Barbora. Klíma o Expozituře: Kroupa je otec seriálu, já matka. Porodil jsem ho. *iDNES.cz* [online]. 10. srpen 2011 [vid. 2023-03-14]. Dostupné z: https://www.idnes.cz/kultura/film-televize/klima-o-expozitura-kroupa-je-otec-serialu-ja-matka-porodil-jsem-ho.A110809_184130_televize_jaz

[15] KLÍMA, Josef. *Epizoda sedmá: Ve službách policie. Jak dva detektivové rozbili Berdychův gang - Seznam Zprávy* [online]. 17. srpen 2020 [vid. 2023-03-14]. Dostupné z: https://www.seznamzpravy.cz/clanek/epizoda-sedma-ve-sluzbach-policie-jak-dva-detektivove-rozbili-berdychuv-gang-115997

[16] KMENTA, Jaroslav. O MNĚ. *Jaroslav Kmenta* [online]. 2023 [vid. 2023-01-22]. Dostupné z: https://www.kmenta.cz/o-mne

[17] KMENTA, Jaroslav. KNIHY. *Jaroslav Kmenta* [online]. 2023 [vid. 2023-01-22]. Dostupné z: https://www.kmenta.cz/knihy

[18] ŠIMÁK, Jakub. Co je to „Kočnerova knižnica“ a kdo za ní stojí. *investigace.cz* [online]. 3. únor 2020 [vid. 2023-01-24]. Dostupné z: https://www.investigace.cz/co-je-to-kocnerova-kniznica-a -kdo-za-ni-stoji/

[19] INVESTIGACE. Náš tým. *investigace.cz* [online]. 2023 [vid. 2023-01-23]. Dostupné z: https://www.investigace.cz/kdo-jsme/

[20] VILČEK, Ivan. *Slovensko ve strachu: Vražda novináře měla souviset s jeho prací - Novinky* [online]. 26. únor 2018 [vid. 2023-03-29]. Dostupné z: https://www.novinky.cz/clanek/zahranicni-evropa-slovensko-ve-strachu-vrazda-novinare-mela-souviset-s-jeho-praci-11720

[21] GUMENYUK, Nataliya. *Did Italian Mafia Have Anything To Do With the Murder of Ján Kuciak?* [online]. 28. červen 2019 [vid. 2023-03-29]. Dostupné z: https://hromadske.ua/en/posts/did-italian-mafia-have-anything-to-do-with-the-murder-of-jn-kuciak

[22] ČESKÁ TELEVIZE. Podnikatel Kočner půl roku před vraždou po telefonu vyhrožoval Kuciakovi. *ČT24 - Nejdůvěryhodnější zpravodajský web v ČR - Česká televize* [online]. 2. listopad 2019 [vid. 2023-03-29]. Dostupné

z: https://ct24.ceskatelevize.cz/svet/2731598-podnikatel-kocner-pul-roku-pred-vrazdou-po-telefonu-vyhrozoval-kuciakovi

[23] PETER, Hanák a Hopková DENISA. Ako sa Kočner vyhrážal Kuciakovi. Berie ústavný súd moc politikom? (podcast). *Aktuality.sk* [online]. 11. únor 2019 [vid. 2023-03-29]. Dostupné z: https://www.aktuality.sk/clanok/666401/ako-sa-kocner-vyhrazal-kuciakovi-berie-ustavny-sud-moc-politikom-podcast/

[24] KUCIAK, Ján. Facebookový příspěvek Jána Kuciaka ve věci podání trestního oznámení na Mariána Kočnera. In: [online]. 20. říjen 2017 [vid. 2023-03-29]. Dostupné z: https://www.facebook.com/photo.php?fbid=10203823631506566&set=a .1378356554293.43324.1693287062&type=3&theater

[25] TÓDOVÁ, Monika, Miro KERN a Veronika FOLENTOVÁ. Investigatívneho reportéra Aktualít Jána Kuciaka a jeho partnerku zavraždili, vláda núka milión eur za odhalenie páchateľov. *Denník N* [online]. 26. únor 2018 [vid. 2023-03-29]. Dostupné z: https://dennikn.sk/1040750/investigativneho-reportera-aktualit-jana-kuciaka-a -jeho-partnerku-zavrazdili/

[26] MARTINOVSKÝ, Marek. Miroslav Marček a Tomáš Szabó. *investigace.cz* [online]. 27. únor 2020 [vid. 2023-03-29]. Dostupné z: https://www.investigace.cz/miroslav-marcek-a -tomas-szabo/

[27] ČTK. Slovenský nejvyšší soud zpřísnil trest pro Kuciakova vraha. Marček má strávit ve vězení 25 let. *iROZHLAS* [online]. 2. prosinec 2020 [vid. 2023-03-29]. Dostupné z: https://www.irozhlas.cz/zpravy-svet/slovensko-jan-kuciak-martina-kusnirova-soud-trest-25-let-vezeni-miroslav-marcek_2012021131_kar

[28] FIALA, Adam a Tamara KEJLOVÁ. „Nesmíme rezignovat na slušnost.“ Slovensko si připomíná pět let od vraždy Kuciaka. *ČT24 - Nejdůvěryhodnější zpravodajský web v ČR - Česká televize* [online]. 21. únor 2023 [vid. 2023-03-29]. Dostupné z: https://ct24.ceskatelevize.cz/svet/3566628-nesmime-rezignovat-na-slusnost-slovensko-si-pripomina-pet-let-od-vrazdy-kuciaka

[29] INVESTIGACE. Pět let po šokujícím zločinu míří do kin dokument Kuciak: Vražda novináře s dosud nezveřejněnými materiály. *investigace.cz* [online]. 12. leden 2023 [vid. 2023-03-29]. Dostupné z: https://www.investigace.cz/dokument-kuciak/

[30] OSVALDOVÁ, Barbora a Jan HALADA. *Praktická encyklopedie žurnalistiky a marketingové komunikace*. 3. vyd. Praha: Libri, 2007. ISBN 978-80-7277-266-7.

[31] KOVACH, Bill a Tom ROSENSTIEL. *The elements of journalism: what newspeople should know and the public should expect*. 1st rev. ed., Completely updated and rev. New York: Three Rivers Press, 2007. ISBN 978-0-307-34670-4.

[32] LEIGH, David. *Investigative journalism: a survival guide*. Cham, Switzerland: Palgrave Macmillan, 2019. ISBN 978-3-030-16751-6.

[33] ICIJ. *Daniel Pearl Awards - ICIJ* [online]. 16. říjen 2017 [vid. 2023-01-24].

Dostupné z: https://www.icij.org/about/awards/daniel-pearl-awards/

[34] CAMPBELL, W. Joseph. Woodward and Bernstein didn't bring down a president in Watergate – but the myth that they did lives on. *The Conversation* [online]. 14. červen 2022 [vid. 2023-01-24]. Dostupné z: http://theconversation.com/woodward-and-bernstein-didnt-bring-down-a -president-in-watergate-but-the-myth-that-they-did-lives-on-183290

[35] ČESKÁ TELEVIZE. Aféra Watergate zlomila vztah veřejnosti a politiků. Investigace se za 50 let zásadně proměnila. *Česká televize* [online]. 18. červen 2022 [vid. 2023-01-23]. Dostupné z: https://ct24.ceskatelevize.cz/svet/3508275-afera-watergate-zlomila-vztah-verejnosti-a -politiku-investigace-se-za-50-let-zasadne

[36] BERNARD, Diane. She went undercover to expose an insane asylum's horrors. Now Nellie Bly is getting her due. *Washington Post* [online]. 2019 [vid. 2023-01-24]. ISSN 0190-8286. Dostupné z: https://www.washingtonpost.com/history/2019/07/28/she-went-undercover-expose-an-insane-asylums-horrors-now-nellie-bly-is-getting-her-due/

[37] JAMES RUCHSER. A History of Whistleblowers and Document Leaks. *The Cairo Review of Global Affairs* [online]. 15. září 2022 [vid. 2023-01-24]. Dostupné z: https://www.thecairoreview.com/timelines/a -history-of-whistleblowers-and-document-leaks/

[38] GERARD RYLE, MARINA WALKER GUEVARA, MICHAEL HUDSON, NICKY HAGER, DUNCAN CAMPBELL, STEFAN CANDEA, MAR CABRA, a KIMBERLY PORTEROUS. *Secret Files Expose Offshore's Global Impact - ICIJ* [online]. 2. duben 2013 [vid. 2023-01-24]. Dostupné z: https://www.icij.org/investigations/offshore/secret-files-expose-offshores-global-impact/

[39] ICIJ. *Giant Leak of Offshore Financial Records Exposes Global Array of Crime and Corruption - ICIJ* [online]. 3. duben 2016 [vid. 2023-01-24]. Dostupné z: https://www.icij.org/investigations/panama-papers/20160403-panama-papers-global-overview/

[40] MARINA WALKER GUEVARA. *ICIJ releases Panama Papers offshore company data - ICIJ* [online]. 9. květen 2016 [vid. 2023-01-24]. Dostupné z: https://www.icij.org/inside-icij/2016/05/icij-releases-panama-papers-offshore-company-data/

[41] VICE. What We Know So Far About the Bahamas Leaks. *Vice* [online]. 22. září 2016 [vid. 2023-01-24]. Dostupné z: https://www.vice.com/en/article/7bm87b/what-we-know-so-far-about-the-bahamas-leaks-876

[42] WILL FITZGIBBON, EMILIA DÍAZ-STRUCK, MAR CABRA, RIGOBERTO CARVAJAL, MIGUEL FIANDOR GUTIÉRREZ, BASTIAN OBERMAYER, a FREDERIK OBERMAIER. *Former EU Official Among Politicians Named in New Leak of Offshore Files from The Bahamas - ICIJ* [online]. 20. září 2016 [vid. 2023-01-24]. Dostupné z: https://www.icij.org/investigations/offshore/former-eu-official-among-politicians-named-new-leak-offshore-files-bahamas/

[43] WILL FITZGIBBON, MICHAEL HUDSON, MARINA WALKER GUEVARA, GERARD RYLE, BASTIAN OBERMAYER, FREDERIK OBERMAIER, SIMON BOWERS, a SASHA CHAVKIN. *Paradise Papers Exposes Donald Trump-Russia links and Piggy Banks of the Wealthiest 1 Percent - ICIJ* [online]. 5. listopad 2017 [vid. 2023-01-24]. Dostupné z: https://www.icij.org/investigations/paradise-papers/paradise-papers-exposes-donald-trump-russia-links-and-piggy-banks-of-the-wealthiest-1-percent/

[44] HOLCOVÁ, Pavla. #Paradise Papers. *investigace.cz* [online]. 5. listopad 2017 [vid. 2023-01-25]. Dostupné z: https://www.investigace.cz/paradise-papers/

[45] ICIJ. *Offshore havens and hidden riches of world leaders and billionaires exposed in unprecedented leak - ICIJ* [online]. 3. říjen 2021 [vid. 2023-01-25]. Dostupné z: https://www.icij.org/investigations/pandora-papers/global-investigation-tax-havens-offshore/

[46] ŠOTOVÁ, Zuzana. Pandora Papers: Třetí a poslední várka dat. *investigace.cz* [online]. 4. květen 2022 [vid. 2023-01-25]. Dostupné z: https://www.investigace.cz/pandora-papers-treti-a -posledni-varka-dat/

[47] ČERNÝ, Jan. *Open Source Intelligence (OSINT) | Informační gramotnost | Elektronický časopis ze světa dat a informací* [online]. 3. leden 2017 [vid. 2023-01-26]. Dostupné z: https://www.informacnigramotnost.cz/skoleni-workshopy/open-source-intelligence-osint/

[48] MALTEGO. *Everything about Open Source Intelligence and OSINT Investigations (2021)* [online]. 2021 [vid. 2023-01-26]. Dostupné z: https://www.maltego.com/blog/what-is-open-source-intelligence-and-how-to-conduct-osint-investigations/

[49] WILD, Johanna. These are the Tools Open Source Researchers Say They Need. *bellingcat* [online]. 12. srpen 2022 [vid. 2023-01-26]. Dostupné z: https://www.bellingcat.com/resources/2022/08/12/these-are-the-tools-open-source-researchers-say-they-need/

[50] BELLINGCAT. About. *bellingcat* [online]. 2023 [vid. 2023-01-26]. Dostupné z: https://www.bellingcat.com/about/

[51] PETER HENSHALL a DAVID INGRAM. Chapter 59: Sources of information. In: *The News Manual* [online]. B.m.: Poroman Press, 2008 [vid. 2022-01-27]. Dostupné z: https://www.thenewsmanual.net/Manuals%20Volume%203/volume3_59.htm

[52] SILVERMAN, Craig a Giannina SEGNINI. Chapter 5: Investigating with databases: Verifying data quality. *Verification Handbook: a guide to online search and research techniques for using ugc and open source* [online]. 2014 [vid. 2023-01-29]. Dostupné z: http://verificationhandbook.com/downloads/verification.handbook.2.pdf

[53] MATHEW CHARLES. Investigative Journalism IV: Dealing with Sources. *Medium* [online]. 5. únor 2018 [vid. 2023-01-29]. Dostupné z: https://medium.com/@headlinexplorer/investigative-journalism-ii-dealing-with-

sources-513de04eae68

[54] SOUTHERN POVERTY LAW CENTER. Not All Search Engines Are Built Alike. *https://www.learningforjustice.org* [online]. 2017 [vid. 2023-01-30]. Dostupné z: https://www.learningforjustice.org/sites/default/files/TT_Digital%20Literacy_Not%20All%20Search%20Engines%20Are%20Built%20Alike.pdf

[55] ČERNÝ, Jan. *Pokročilé operátory Bing | Informační gramotnost | Elektronický časopis ze světa dat a informací* [online]. 24. únor 2017 [vid. 2023-01-30]. Dostupné z: https://www.informacnigramotnost.cz/pokrocile-operatory-bing/

[56] MATTEO DUO. Google Search Operators: 40 Commands to Know in 2023 (Improve Research, Competitive Analysis, and SEO). *Kinsta®* [online]. 30. březen 2022 [vid. 2023-01-30]. Dostupné z: https://kinsta.com/blog/google-search-operators/

[57] DUCKDUCKGO. DuckDuckGo Search Syntax. *DuckDuckGo Help Pages* [online]. 2023 [vid. 2023-01-30]. Dostupné z: https://help.duckduckgo.com/duckduckgo-help-pages/results/syntax/

[58] NORDINE, Justin. OSINT Framework. *OSINT Framework* [online]. 2023 [vid. 2023-03-29]. Dostupné z: https://osintframework.com/

[59] FLIGHT-RADAR.ORG. Flightradar24 | Follow any flight online - Simple and free flight tracking. *flight-radar.org* [online]. 2023 [vid. 2023-01-30]. Dostupné z: https://www.flight-radar.org/flightradar24/

[60] FLIGHTRADAR24. Live Flight Tracker - Real-Time Flight Tracker Map. *Flightradar24* [online]. 2023 [vid. 2023-03-30]. Dostupné z: https://www.flightradar24.com/

[61] WEIDE, Youri van der. Using the Sun and the Shadows for Geolocation. *bellingcat* [online]. 3. prosinec 2020 [vid. 2023-01-30]. Dostupné z: https://www.bellingcat.com/resources/2020/12/03/using-the-sun-and-the-shadows-for-geolocation/

[62] SUNCALC. *SunCalc sun position- und sun phases calculator* [online]. 2023 [vid. 2023-03-30]. Dostupné z: https://www.suncalc.org

[63] INVID. Workpackages. *InVID project* [online]. 2023 [vid. 2023-01-30]. Dostupné z: https://www.invid-project.eu/workpackages/

[64] INVID. Description. *InVID project* [online]. 2023 [vid. 2023-01-30]. Dostupné z: https://www.invid-project.eu/description/

[65] INVID. InVID Verification Plugin. *InVID project* [online]. 2023 [vid. 2023-01-30]. Dostupné z: https://www.invid-project.eu/tools-and-services/invid-verification-plugin/

[66] OPENCORPORATES. *OPENCORPORATES LTD :: United Kingdom :: OpenCorporates* [online]. 2023 [vid. 2023-03-30]. Dostupné z: https://opencorporates.com/companies/gb/07444723

[67] OPENCORPORATES. *Our Purpose* [online]. 2023 [vid. 2023-03-30]. Dostupné z: https://opencorporates.com/info/our-purpose/

[68] OPENCORPORATES. *OpenCorporates :: The Open Database Of The Corporate World* [online]. 2023 [vid. 2023-03-30]. Dostupné z: https://opencorporates.com/

[69] SHIPTRACKER. *MarineTraffic - Worldwide Ship And Yacht Tracking In Real-time* [online]. 2023 [vid. 2023-03-30]. Dostupné z: https://shiptracker.live/

[70] MARINETRAFFIC. About Us. *MarineTraffic* [online]. 2023 [vid. 2023-03-30]. Dostupné z: https://www.marinetraffic.com/en/p/company

[71] MARINETRAFFIC. Ship PSARA GLORY (Ro-Ro/Passenger Ship). *MarineTraffic.com* [online]. 2023 [vid. 2023-03-30]. Dostupné z: https://www.marinetraffic.com/en/ais/details/ships/shipid:211352/mmsi:239722800/imo:7909437/vessel:PSARA GLORY

[72] ARIZA, Miriam Forero. Tips for Building a Database for Investigations. *Global Investigative Journalism Network* [online]. 13. červenec 2021 [vid. 2023-02-03]. Dostupné z: https://gijn.org/2021/07/13/tips-for-building-a -database-for-investigations/

[73] INDEED. 5 Types of Data Classification (With Examples). *Indeed Career Guide* [online]. 13. červenec 2021 [vid. 2023-02-04]. Dostupné z: https://www.indeed.com/career-advice/career-development/data-classification-types

[74] ILIA SOTNIKOV. Data Classification: What It Is and How to Implement It. *https://blog.netwrix.com/* [online]. 2. září 2020 [vid. 2023-02-04]. Dostupné z: https://blog.netwrix.com/2020/09/02/data-classification/

[75] NORMAN GUTIÉRREZ. *The Complete Guide to Data Classification | Prey Blog* [online]. 26. duben 2021 [vid. 2023-02-04]. Dostupné z: https://preyproject.com/blog/the-complete-guide-to-data-classification

[76] THAKUR, Meenal. Organise, structure, present: Effective ways of working with information for investigative…. *European Journalism Centre* [online]. 29. duben 2021 [vid. 2023-02-15]. Dostupné z: https://medium.com/we-are-the-european-journalism-centre/organise-structure-present-effective-ways-of-working-with-information-for-investigative-fc624a735b93

[77] OCCRP. *Building out your investigation* [online]. 2022 [vid. 2023-02-19]. Dostupné z: https://docs.alephdata.org/guide/building-out-your-investigation/using-the-table-editor

[78] ICIJ. *Datashare* [online]. Java. B.m.: The International Consortium of Investigative Journalists. 21. únor 2023 [vid. 2023-02-23]. Dostupné z: https://github.com/ICIJ/datashare

[79] SOLINE LEDÉSERT. *What is Datashare? FAQs about our document analysis software - ICIJ* [online]. 4. listopad 2019 [vid. 2023-02-23]. Dostupné z: https://www.icij.org/inside-icij/2019/11/what-is-datashare-frequently-asked-

questions-about-our-document-analysis-software/

[80] DOUGLAS DALBY. *„Trust and technology" at heart of latest collaborative journalism tool - ICIJ* [online]. 17. srpen 2020 [vid. 2023-02-23]. Dostupné z: https://www.icij.org/inside-icij/2020/08/trust-and-technology-at-heart-of-latest-collaborative-journalism-tool/

[81] GIJN. *Why journalists need an archiving system - ICIJ* [online]. 16. srpen 2021 [vid. 2023-02-23]. Dostupné z: https://www.icij.org/inside-icij/2021/08/why-journalists-need-an-archiving-system/

[82] CHRISTOPH SCHOLZ. *5 digital security tools journalists use to protect their work and sources* [online]. 29. leden 2018 [vid. 2023-02-23]. Dostupné z: https://www.icij.org/inside-icij/2018/01/five-digital-security-tools-to-protect-your-work-and-sources/

[83] JERREAT, Jessica. Digital safety. *Committee to Protect Journalists* [online]. 10. září 2018 [vid. 2023-02-23]. Dostupné z: https://cpj.org/2018/09/digital-safety/

[84] NOHE, Patrick. What is an Air Gapped Computer? *Hashed Out by The SSL Store™* [online]. 13. březen 2018 [vid. 2023-02-25]. Dostupné z: https://www.thesslstore.com/blog/air-gapped-computer/

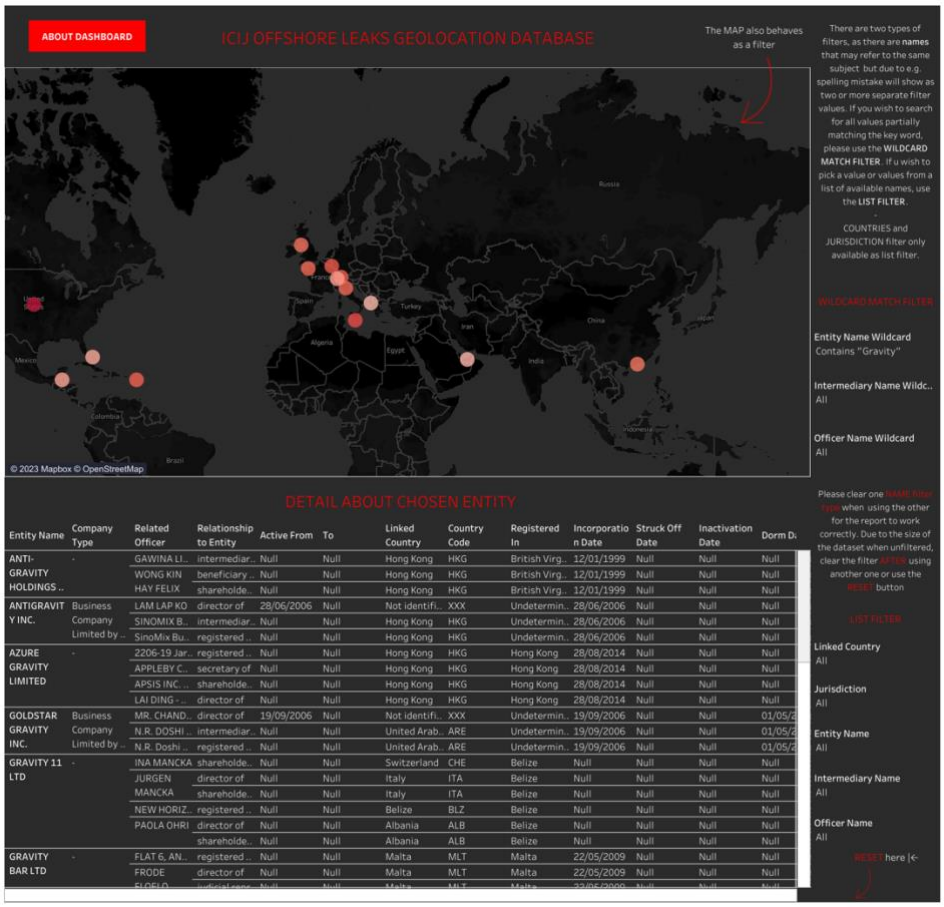
[85] ICIJ. The Shape of the Data. *Offshore Leaks Data* [online]. 2023 [vid. 2023-03-19]. Dostupné z: https://offshoreleaks-data.icij.org/offshoreleaks/neo4j/guide/datashape.html

[86] STEHNO, Pavel. Nominee services - profesionální jednatelé ve vašich službách. *Hospodářské noviny (HN.cz)* [online]. 30. leden 2007 [vid. 2023-03-30]. Dostupné z: https://byznys.hn.cz/c1-20306450-nominee-services-profesionalni-jednatele-ve-vasich-sluzbach

Přílohy

Příloha A: Odkaz na Offshore Leaks Tableau Dashboard

Příloha A obsahuje odkaz k publikovanému dashboardu, který vznikl v rámci praktické části této práce.



Odkaz na dashboard zde

Případně lze dashboard vyhledat pomocí názvu – ICIJ Offshore Leaks Geolocation Database.

Příloha B: Rozhovor s Pavlou Holcovou z investigace.cz

Příloha B obsahuje strukturovaný rozhovor s investigativní novinářkou Pavlou Holcovou, který proběhl písemně prostřednictvím e-mailu.

V rozhovoru pro reflex.cz s Čestmírem Strakatým se vás ptal, zda se bojíte o svoji bezpečnost – zmínila jste, že máte pro komunikaci o citlivých datech vlastní systém, který jste si vyvinuli. Zajímalo by mě, jak takový systém funguje/vypadá? Jakým způsobem je takový systém zabezpečený? Chápu, že nebude možné mi sdělit detailní informace, chtěla bych jen pochopit, jak funguje proces komunikace a předávání citlivých informací mezi novináři. Je možné přirovnat to k nějaké existující komunikační platformě (např. Telegram)?

Je to kombinace platformy na bezpečné sdílení a prohledávání dokumentů, platformy na shromažďování poznatků, na které jsme přišli (typ řazení informací wiki) a instatní komunikační kanál, což je často skupina na Signalu.

Jakým způsobem se předává tak velké množství dat, jako je například soubor Panama Papers?

Online databáze s bezpečným přístupem 2FA login, navázanost přístupu na IP adresu.

Jakými způsoby dostáváte podněty k novým kauzám?

Oslovují nás zahraniční kolegové, že potřebují s něčím pomoct, popřípadě reflektujeme konkrétní dění v Evropě/ČR.

Jak probíhá proces sběru dat při vyšetřování nové kauzy? Jak postupujete při sběru dat?

Na začátku bývá ve většině případů jméno člověka nebo firmy. Většinou toto jméno prověříme ve veřejných databázích jako obchodní rejstřík, rejstřík konečných majitelů, katastr – a pak v našich archivech a neveřejných databázích. Podíváme se na obchodní i jiné vazby, co už se o tom člověku ví (jestli vůbec něco) a na základě těchto dat definujeme prvotní hypotézu, kterou se pak snažíme ověřit. To už je samotné jádro našich kauz.

Jakými způsoby ověřujete správnost těchto dat?

Máme externí fact-chekingový tým. Samotné ověřování faktů, neboli fact-checking většinou trvá cokoliv mezi jedním a pěti dny a bývá dost vyčerpávající - jakékoliv tvrzení v kauze totiž musímě podložit průkaznými materiály.

K projektu Kočnerova knižnica, který obsahuje 57 TB dat – tato data jste roztřídili, zanalyzovali a zpřístupnili ostatním novinářům. Jaké metody používáte při třídění dat, která získáte?

S tímto nám pomohli kolegové z OCCRP data týmu – více detailů zde: https://www.investigace.cz/jak-jsme-porcovali-slona/

Jakým způsobem probíhá analýza získaných dat? Používáte nějaké nástroje/softwary na procházení a analýzu jednotlivých dokumentů, obzvlášť, když jde o tak velké množství dat? Pokud vše procházíte ručně, kolik lidí musí pracovat na projektech, kde je objem dat v rozmezí terabytů?

Většinu dat, která získáme, se snažíme indexovat a cross-checkovat s ostatními dokumenty, které už máme k dispozici. Používáme k tomu softwarový nástroj, který jsme sami vyvinuli a jmenuje se Aleph. Zbytek děláme ručně. Do budoucna chceme některé tyto procesy automatizovat s pomocí AI.

Když v terénu vyšetřujete a zapisujete si informace, jak zpracováváte data, která si zaznamenáte v podobě poznámek pomocí tužky a papíru? Zpracováváte taková data později elektronicky nebo pracujete dál s poznámkami např. v podobě nějaké mapy?

Ano, poznámky přepisujeme poslední den výjezdu do terénu do digitální podoby, aby byly prohledávatelné. Zároveň děláme zálohy všech digitálních materiálů, které jsme nasbírali.

Jakým způsobem a kde ukládáte získaná data? Pokud máte nějaké vlastní vyvinuté systémy, bylo by možné popsat, o jaký typ úložiště jde a jak funguje?

Dokumenty nahráváme do našeho cloudu, popřípadě Alephu, aby byly v budoucnosti dostupné i ostatním.

Jakým způsobem data chráníte, než dojde k jejich zveřejnění, aby nedošlo k jejich úniku či ztrátě?

Na digitální bezpečnost máme několik školení ročně, dostáváme měsíční update o nově objevených bezpečnostních rizicích, kterými se řídíme. Pro databáze obsahující citlivé dokumenty máme dedikované servery a platformy. S těmi nejcitlivějšími dokumenty pak vůbec nepracujeme online, máme air-gapped počítač, na kterém tyto dokumenty můžeme procházet.

Stalo se Vám někdy za Vaši kariéru, že by z neopatrnosti, ať už Vaší či nějakého kolegy, došlo k úniku nějakých důležitých dat?

Je to možné, já o tom ale nevím – většina hackerů, která se chce dostat k vašim citlivým dokumentům, Vám o úspěšném útoku nedá vědět.

Pracujete s nějakými OSINT (Open Source Intelligence) nástroji? Pokud ano, jaké jsou podle vás nejužitečnější nástroje pro práci s otevřenými zdroji?

Ano, je jich celá řada – dohromady s Bellingcat jsme dávali dohromady celý seznam OSINT nástrojů, v současné době má několik set položek a link najdete na webu Bellingcat.

Spolupracujete nějakým způsobem s Bellingcat, když vyšetřujete nějaké kauzy?

Občas ano, jsou to kamarádi.

Vyhledáváte informace ke svým kauzám i na dark webu?

Když je to potřeba, tak ano.

Pokud by to šlo takto říct, máte nějakou oblíbenou kauzu, která byla pro Vás nejzajímavější?

Kočnerova knižnica. Z mnoha profesních i osobních důvodů.

Příloha C: Skript pro předzpracování dat v Jupyter Notebooku

Součástí diplomové práce je soubor obsahující skript pro předzpracování dat s názvem PrilohaC*_DP_preprocessing_skript.ipynb*.