ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

FAKULTA BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

Katedra zdravotnických oborů a ochrany obyvatelstva

Otevřené zdroje dat v síti Internet a možnosti jejich vytěžování

Open Data Sources in the Internet and Possibilities of Their Extraction

Diplomová práce

Studijní program: Civilní nouzové plánování

Autor diplomové práce: Bc. Jan Tisančín, DiS.

Vedoucí diplomové práce: Ing. Václav Navrátil

Kladno 2020

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení:	**Tisančín**	Jméno:	**Jan**	Osobní číslo:	**484198**
Fakulta:	**Fakulta biomedicínského inženýrství**				
Garantující katedra:	**Katedra zdravotnických oborů a ochrany obyvatelstva**				
Studijní program:	**Ochrana obyvatelstva**				
Studijní obor:	**Civilní nouzové plánování**				

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Otevřené zdroje dat v síti Internet a možnosti jejich vytěžování

Název diplomové práce anglicky:

Open Data Sources in the Internet and Possibilities of Their Extraction

Pokyny pro vypracování:

Cílem diplomové práce bude zmapovat a vyhodnotit některé zdroje otevřených dat v prostředí sítě Internet. V teoretické části bude popsána legislativa související s ochranou osobních údajů a potenciální zdroje dat specifické pro Českou republiku. Objasněna bude také problematika vytěžování dat z otevřených zdrojů (OSINT) spolu s jednotlivými metodami, které lze použít. V praktické části bude proveden výzkum, při kterém bude po dobu 3 měsíců sledováno prostředí vybraných webových serverů (diskusní fóra, inzertní portály, sociální sítě atd.), odkud budou průběžně náhodným výběrem zjišťována vstupní data k nejméně 100 profilům. Na tyto profily budou v průběhu výzkumu aplikovány metody OSINT za účelem zjištění maximálního možného množství relevantních informací. Výsledky výzkumu budou v závěru vyhodnoceny pomocí analýzy MCDA.

Seznam doporučené literatury:

[1] BERTRAM, Stewart, The Tao of Open Source Intelligence, Cambridgeshire: IT Governance Publishing Ltd, 2015, ISBN 9781849287296
[2] HASSAN, Nihad, A., HIJAZI, Rami, Open Source Intelligence Methods and Tools: A Practical Guide to Online Intelligence, New York: Apress, 2018, ISBN 9781484232132
[3] KOLOUCH, Jan, CyberCrime, Praha: CZ.NIC, 2016, ISBN 978-80- 88168-15-7

Jméno a příjmení vedoucí(ho) diplomové práce:

Ing. Václav Navrátil

Jméno a příjmení konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **23.09.2019**
Platnost zadání diplomové práce: **18.09.2021**

prof. MUDr. Leoš Navrátil, CSc., MBA, dr.h.c.
podpis vedoucí(ho) katedry

prof. MUDr. Ivan Dylevský, DrSc.
podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Student(ka) bere na vědomí, že je povinnen(a) vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

27.9.2019
Datum převzetí zadání

Podpis studenta(ky)

PROHLÁŠENÍ

Prohlašuji, že jsem diplomovou práci s názvem Otevřené zdroje dat v síti Internet a možnosti jejich vytěžování vypracoval samostatně pouze s použitím pramenů, které uvádím v seznamu bibliografických odkazů.

Nemám závažný důvod proti užití tohoto školního díla ve smyslu § 60 zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů.

Dále prohlašuji, že veškeré osobní údaje, které byly v souvislosti s touto prací shromážděny, byly využity výhradně pro výzkumné účely této práce a po jejím dokončení byly smazány.

V Praze dne 18.05.2020

…….
Bc. Jan Tisančín, DiS.

PODĚKOVÁNÍ

Tímto bych chtěl poděkovat vedoucímu diplomové práce Ing. Václavu Navrátilovi za věcné připomínky a osobní přístup při vedení mé práce. Mé díky patří také doc. JUDr. Janu Kolouchovi, Ph.D. a Ing. Tomáši Kratinovi za odborné konzultace a doporučení literatury.

ABSTRAKT

Diplomová práce zkoumá možnosti vytěžování otevřených zdrojů z prostředí sítě Internet ve vztahu k uživatelům. V teoretické části tyto možnosti komparuje s právním rámcem GDPR, který zajišťuje ochranu osobních údajů uživatelů. Popisuje zdroje výskytu údajů o uživatelích a možná rizika, která z uveřejňování dat pramení. Poslední kapitola teoretické části představuje metody pro vytěžování dat.

V praktické části práce je provedena případová studie na vzorku 100 uživatelů, k nimž s využitím popsaných metod OSINT proběhl sběr dat. Výsledky studie byly dále vyhodnoceny dle rizik, která uživatelům hrozí a analyzovány formou MCDA. Analýza na problém nahlíží z pohledu možného útočníka a demonstruje jeho možnost realizovat vybrané typy hrozeb. Kazuistiky jsou orientovány na weby, které využívají uživatelé z České republiky zejména pro inzerci.

Výzkumem zároveň došlo k vyvrácení obou stanovených hypotéz. Byla zjištěna souvislost mezi počtem úniků uživatelských dat a obdobím, kdy na prostředí mohlo mít vliv nařízení GDPR. Bylo také zjištěno, že uživatelé aktivně obchází prvky, které je mají chránit před možným sběrem dat, což může mít za následek archivaci jejich údajů. Aplikace práva být zapomenut, které nařízení GDPR zavedlo, je pak značně ztížená až nemožná.

Zjištěné nedostatky hodnotí závěrečná diskuzní část, ve které jsou zároveň navrženy způsoby jejich řešení. Diskuze zároveň provádí rozbor situací, se kterými jsem se v rámci výzkumu setkal a demonstruje možná rizika.

Klíčová slova

OSINT; zpravodajství z otevřených zdrojů; sociální sítě; data mining; vytěžování dat; Google hacking; osobní údaje; Internet

ABSTRACT

The diploma thesis examines the possibilities of extracting data about users from open sources on the Internet. In the theoretical part, the possibilities are compared with the legal framework (GDPR) that ensures the protection of users' personal data. It describes the sources of user data and the potential risks of publishing data. The last chapter of the theoretical part presents methods for data mining.

In the practical part of the work, a case study was performed on 100 users. For these, data collection was performed using OSINT methods. The results of the study were evaluated according to the risks they pose to users and further analyzed using MCDA. The analysis uses the perspective of a potential attacker, thus demonstrating the possibility of carrying out some of the attacks which are described here in the theoretical part. The case studies are focused on websites that are used by users from the Czech Republic, especially for auctioning.

The performed research also refuted both hypotheses. A link was found between the number of user data leaks and the period when the environment could have been affected by the GDPR Regulation. It has also been observed that users actively bypass the elements designed to protect them from possible data collection. Such behaviour can result in the archiving of their data and the application of the right to be forgotten, which the GDPR regulation introduced, is then considerably more difficult or even impossible.

The identified shortcomings are evaluated in the final discussion part, in which ways of their solution are proposed. The discussion also analyzes the situations I encountered in the research and demonstrates the possible risks.

Keywords

OSINT; open source intelligence; social networks; data mining; data extraction; Google hacking; personal data; Internet

Obsah

1 ÚVOD

Diplomová práce se zabývá možnostmi vytěžování dat z otevřených zdrojů v souvislosti s riziky, která pramení z výskytu uživatelských údajů v otevřené části sítě Internet. Ze zkušeností vím, že mnoho lidí tuto problematiku podceňuje a častým argumentem je, že nevidí důvod, proč by se zrovna oni měli stát cílem, když jsou jen obyčejní lidé. Své údaje poskytujeme například obchodům kvůli slevovým kartičkám, telefonní čísla dáváme k dispozici firmám, které ve smluvních podmínkách jasně uvádí, že je poskytnou třetím stranám aj. Zároveň nám však vadí, pokud se na dokumentech objevují naše rodná čísla. Sami vlastně nevíme, kde všude jsou o nás shromažďována data a nemůžeme se proto divit, pokud se dříve či později někde objeví. Pokud se to navíc stane na síti Internet, mohou zde taková data zůstat navždy.

Tuto práci jsem z velké části zpracovával na základě vlastních poznatků. První zkušenosti s otevřenými zdroji jsem získával již před více než 15 lety, ačkoli jsem tehdy označení OSINT neznal. V problematice vytěžování otevřených zdrojů existují konvenční nástroje a metody, od kterých jsem se snažil v maximální možné míře oprostit. Často se totiž setkávám se snahou o unifikaci postupů, což dle mého názoru není zcela správné. Postupy a nástroje by měly odpovídat prostředí a potřebě, proto jsou v práci použity i méně obvyklé metody. Ty mají primárně ukázat, jakým způsobem mohou nástroje pro sběr dat fungovat. Jistě existují specializované nástroje, které mnohdy jednotlivé dílčí kroky zvládnou lépe, či v kratším čase, avšak v práci se jsem spojil nástroje spolu s myšlením a zkušenostmi, které mohou čtenáře inspirovat. Přínosem může být i skutečnost, že některé uvedené příklady jsou z prostředí, které je v České republice dobře známé. Do práce jsem zařadil i zkoumání problémů, které jsem doposud sám neznal, nebo kterým jsem se záměrně vyhýbal. Zpracování diplomové práce bylo ideální příležitostí se k některým z nich vrátit a pokusit se na ně najít odpověď.

2 CÍLE PRÁCE A HYPOTÉZY

Cílem diplomové práce je vyhodnotit stav prostředí sítě Internet ve vztahu k ochraně uživatelských dat a hrozbám vyplývajícím z jejich zneužití.

V průběhu práce bude za tímto účelem prováděn výzkum nad souborem případových studií, při kterém budou sbírána data z otevřených zdrojů v síti Internet ke stovce náhodně zvolených uživatelů sítě Internet. Data přitom sama o sobě nemusí mít charakter osobních nebo citlivých údajů, avšak v kontextu dalších veřejně přístupných informací by se o takové údaje jednat mohlo.

Pro kazuistiky budou použity vyhledávací techniky uvedené v teoretické části práce. Budou při něm vybrána náhodná vstupní data (telefonní čísla, e-maily, jména s fotografiemi) z různých zdrojů. Po získání požadovaného množství dat bude provedeno vyhodnocení s přihlédnutím ke skutečnosti, zda jsou data vědomě zveřejněna samotnými uživateli, či jejich zpracovatelem. Vyhodnocení rizik bude provedeno formou MCDA (Multiple-criteria Decision Analysis – vícekriteriální rozhodovací analýza), která kategorizuje zjištěné údaje dle potenciálních hrozeb.

2.1 Stanovení hypotéz diplomové práce

Pro zpracování práce byly stanoveny následující hypotézy:

Hypotéza 1: Neexistuje korelace mezi počtem úniků uživatelských dat a směrnicí GDPR.

Hypotéza 2: Občané ČR se v souvislosti s ochranou osobních údajů v síti Internet chovají zodpovědně.

3 PŘEHLED SOUČASNÉHO STAVU

3.1 OSINT – definice, historie

Zpravodajství z otevřených zdrojů neboli OSINT (z angl. Open Source INTelligence) je jedna z metod získávání a využívání poznatků o určitém objektu zájmu, a to za využití zdrojů dat, které jsou pro běžného uživatele přístupné bez zvláštního oprávnění, nebo k nim lze získat přístup např. za úplatu [1]. Práce se primárně zabývá vztahem mezi OSINT a běžnými uživateli sítě Internet, nikoli penetračními testy.

Přístup k datům může být v různé míře ovlivněn schopnostmi, či zkušenostmi osoby, která vytěžování provádí. Do kategorie otevřených zdrojů dat tak spadají i takové informace, jež se nachází na uzavřených webových fórech, přestože se k obsahu dostane jen omezený okruh uživatelů. Naopak data v státem spravovaných evidencích (např. registr obyvatel aj.) nelze v žádném případě považovat za otevřený zdroj, byť by k němu osoba mohla z pozice svého povolání mít přístup.

Otevřenými daty se tedy pro účely této práce rozumí data neutajená a přístupná bez zvláštního (zákonného) oprávnění. Má diplomová práce je dále omezena jen na zdroje dat, které lze nalézt v síti Internet.

V širším smyslu může pojem OSINT zahrnovat též analýzu např. denního tisku, časopisů, rádiového vysílání nebo také záznamů z matrik a další podklady, které nebyly digitalizovány. Tyto materiály byly prvními v novodobé historii, jež byly pro účely zpravodajství z otevřených zdrojů používány.

První použití otevřených zdrojů se objevilo u zpravodajských služeb v průběhu 2. světové války. V tomto období OSS (The Office of Strategic Services – předchůdce dnešní CIA) využíval těchto metod proti Německu, sledoval nekrology v regionálním tisku o úmrtí významných nacistů

a shromažďoval obrazové materiály o bitevních lodích, bombových kráterech a letadlech, aby mohl posoudit stav německého vojska [2].

Problém samotného dohledání zájmové informace v dnešním světě setrval, a navíc se objevil i problém nový, kterým je nutnost zpracování velkého množství informací. Díky relativně nízkým cenám zařízení umožňujícím připojení k síti Internet a samotné dostupnosti internetového připojení je dnes každou minutu na sociální síť Facebook vloženo více než půl miliónu komentářů, na portál YouTube nahráno více než 300 hodin audiovizuálních záznamů a na síť Instagram je každý den vloženo přes 100 miliónů fotografií a videí [3].

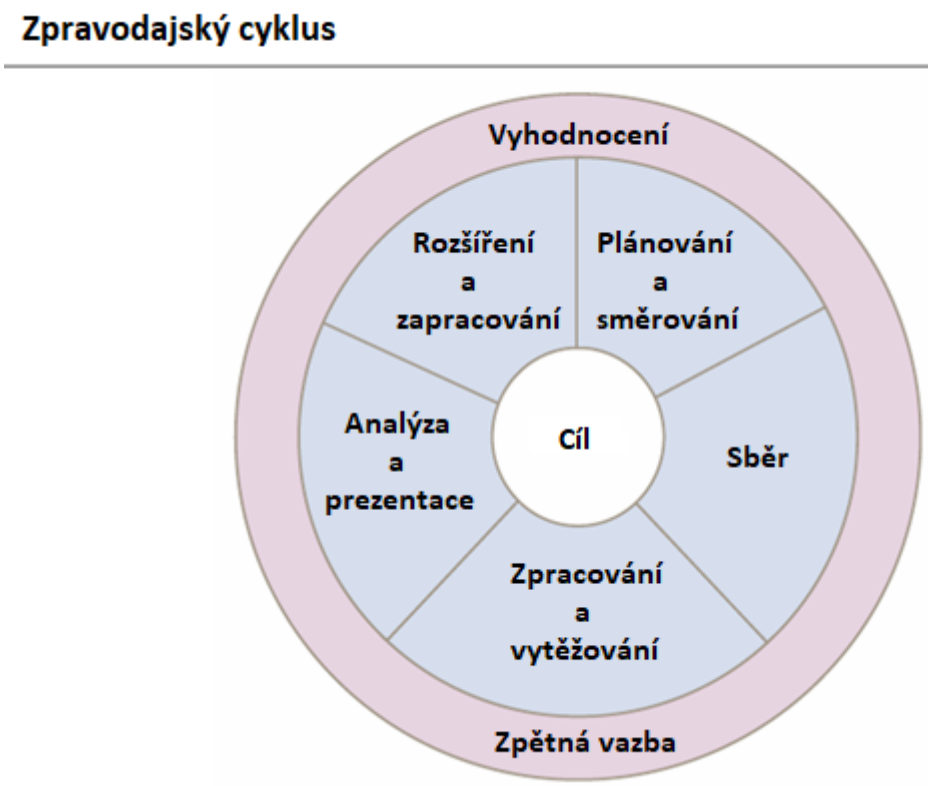
Množství dat, které je na síť Internet vkládáno, způsobuje jev zvaný informační smog. Obecně je jako informační smog chápán irelevantní obsah, který odvádí pozornost. Černohlávková informační smog popisuje následovně: *„Toto „informační smetí" člověka pouze obtěžuje a nic mu nepřináší. Ztěžuje nám také využití cenných informací, které skrze mlhu nepotřebných nevidíme"* [4, s. 35].

Kategorii irelevantního obsahu Internetu mohou tvořit weby, jejichž obsah je generován automatizovaně nebo také blogy bez informační hodnoty, které pouze duplikují již existující obsah. Podobně však fungují i mediální servery, které plní přední příčky díky vysokým počtům návštěv a častým aktualizacím, avšak publikují jen lehce pozměněný obsah převážně z jednoho zdroje. Hledáním zájmových informací v tomto balastu uživatel zbytečně stráví více času, protože je nucen filtrovat nepodstatný obsah.

S přehlcením úzce souvisí i potřeba vyhodnocení relevance a pertinence dat. Pojmy definuje Černohlávková takto: *„Relevance je důležitost či závažnost; vyjadřuje, nakolik se vyhledaný dokument shoduje s původním požadavkem. Pertinence je subjektivní dojem uživatele; vyjadřuje jeho spokojenost s vyhledaným dokumentem na základě původního požadavku"* [4, s. 11].

Osoba provádějící sběr (data-mining) musí všechny informace hodnotit nezaujatě a oprostit se v maximální možné míře od subjektivních názorů. Vnášení subjektivního názoru snižuje relevanci dat a může způsobit zkreslení výsledků, neboť v průběhu sběru nejsme schopni posoudit význam dílčích informací. Subjektivní pocit lze aplikovat až při následném vytěžování, kde je naopak jistá míra pertinence žádoucí, a to např. při posuzování kvality zdroje. V případě úspěchu jsou extrahována relevantní a využitelná data, která mohou být dále zpracována, analyzována a prezentována.

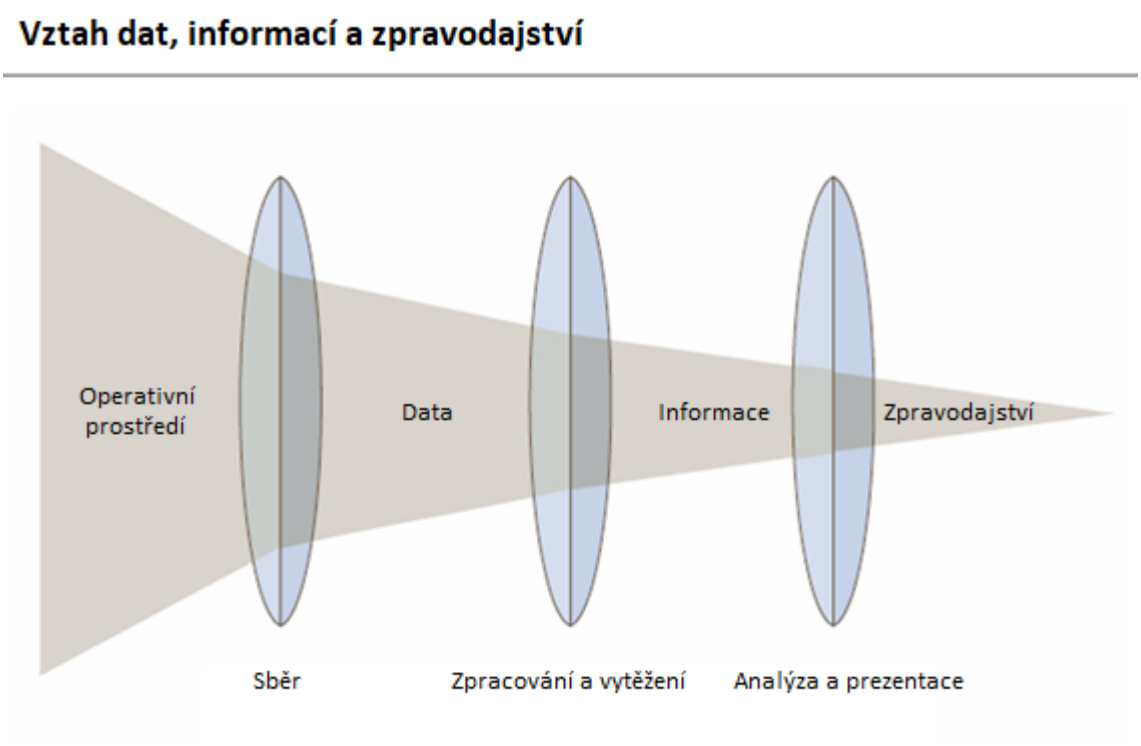
Proces celého OSINTu je možné znázornit jako zpravodajský cyklus pomocí obrázku č.1.



Obrázek 1 – Zpravodajský cyklus [5]

V cyklu je vytyčen cíl a poté zvolen vhodný postup. Posléze je zahájen sběr dat s jejich následným zpracování dat a vytěžením zájmových informací (filtrace zbytečných dat, překlady). Tyto informace jsou v další fázi analyzovány a je zpravidla také provedena jejich vizualizace (např. vztahová analýza). Výsledky analýzy jsou zapracovány a v případě potřeby je rozhodnuto o opakování celého cyklu s cílem získat další informace. Výsledné zpravodajské informace jsou na závěr předány k vyhodnocení. Důležité je též předání zpětné vazby, aby měl analytik odezvu o využitelnosti informací.

Objem zpracovávaných dat a jejich využitelnost v průběhu zpravodajského cyklu znázorňuje obrázek č. 2.



Obrázek 2 – Vztah dat, informací a zpravodajství [5]

Obrázek č. 2 vizualizuje a pojmenovává kategorie dat na základě jejich využitelnosti. Operativní prostředí představuje okolní svět, kde se všechna data nachází. Při jejich sběru je důležité dohledat maximální množství dat k danému cíli a následně zpracovat pouze relevantní data. Zpracování irelevantních dat je neefektivní, neekonomické a defacto nemožné. Po určení relevantních dat a jejich stažení, je tato možné zpracovat a vytěžit. Produktem je informace, která už určitým způsobem souvisí s vytyčeným cílem, ale ještě nemusí být zcela zřejmé jak. Výstup závěrečné analýzy získaných informací je možné pojmenovat jako zpravodajskou informaci [5].

Při vytěžování dat je nutné vyhodnocovat důvěryhodnost zdrojů. Každý, kdo provádí extrakci, by měl být schopen určit kvalitu dat (viz dále). Šíření dezinformací je v prostředí sítě Internet velmi častým jevem a existují i celé weby založené jen na publikování mylných nebo zkreslených informací, které se pak někdy šíří dále, např. prostřednictvím sociálních sítí. Příklad webu, který bývá často označován za dezinformační je Aeronet.cz [6]. Existují však také weby, které se snaží fake news (falešné zprávy) identifikovat. Projekt Manipulátoři.cz v souvislosti s pandemií COVID-19 upozornil na několik takových zpráv a právě web Aeronet.cz je jeho častým terčem.

Při určování kredibility zdroje je třeba ověřovat několik faktorů. Příkladem pro vyhodnocení může být např. metoda „The CRAP Test“. Jak již její název napovídá, slouží pro identifikaci dat, která nemá smysl vytěžovat a skládá se ze 4 kroků – Currency (aktuálnost), Reliability (spolehlivost), Authority (důvěryhodnost autora), Purpose/Point of View (úhel pohledu, zkreslení) [7]. Jednotlivé kroky je možné obodovat obdobně, jak tomu je např. v tabulce č. 1.

Tabulka 1 – Příklad hodnocení zdroje pomocí CRAP Testu [8]

	0 bodů	1 bod	2 body	3 body
Aktuálnost	zdroj neaktualizuje	aktualizace dlouho neproběhly	není úplně aktuální, ale je aktivní	pravidelné, aktuální aktualizace
Spolehlivost	nepřesné nebo neúplné informace	bez citací, vykrádá cizí zdroje bez odkazování	cituje své zdroje a uvádí původ informací	cituje další důvěryhodné zdroje, odkazuje na ně, má vědecké kvality a originální zjištění
Autor	bez autora, bez vydavatele	autor nebo skupina autorů, bez vydavatele nebo záštity sdružení	autor nebo skupina autorů s vydavatelem nebo podporou organizace/sdružení	odborný autor nebo skupina autorů, známé vydavatelství
Zkreslení	jednoznačně ovlivněné a propagující jednostranný názor	zkreslené údaje, publikace osobních názorů bez faktických podkladů	lehká míra zkreslení, snaží se o názorovou vyváženost	bez propagování osobních zájmů, srovnává názory nezaujatě, minimální zkreslenost údajů

Stanovení kredibility zdroje je klíčové při vytěžování dat. Přestože by bylo možné data z takových zdrojů nesbírat vůbec, mohou se v textech nacházet i další zdroje, které hodnotu mají.

V dnešní době si OSINT získal pozornost médií i veřejnosti ne díky ohromnému množství příspěvků, obrázků, videí a vůbec veškerých dat, která uživatelé na síť Internet dobrovolně sami nahrávají, ale zejména zásluhou skandálů.

Známým případem byla kauza Cambridge Analytica, která vyvolala poprask v roce 2018. Společnost Cambridge Analytica pracovala s analytickým nástrojem, který byl mj. využit pro shromažďování údajů o voličích při prezidentské kampani v USA. Údaje měly sloužit pro cílenou politickou agitaci a ovlivnění voličů. Ačkoli společnost porušila smluvní podmínky sociální sítě Facebook a shromážděná data monetarizovala, bylo shromáždění těchto dat a jejich

následná analýza z trestněprávního hlediska naprosto legální. Data obsahovala identitu uživatelů, sítě přátel a „like“. Záměrem bylo zmapovat osobnostní rysy uživatelů a informace poté použít podstrčení cílené reklamy [9].

Pro oblast marketingu jsou otevřené zdroje nekonečnou studnicí cenných informací. Přestože je zpracování dat stále poměrně složité, prochází velmi rychlým vývojem díky obrovským investicím a technologiím jako jsou strojové učení a umělá inteligence.

V současnosti je OSINT využíván nejen v komerční sféře, ale samozřejmě i v rámci policejní činnosti, pro vojenské účely a také pro páchání trestné činnosti.

3.2 Legislativní úprava

3.2.1 GDPR

Obecné Nařízení Evropského parlamentu a Rady (EU) 2016/679 ze dne 27. dubna 2016 o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů a o zrušení směrnice 95/46/ES (Obecné nařízení o ochraně osobních údajů – dále jen GDPR) představuje právní rámec stanovující pravidla pro zpracování osobních údajů [10].

„Cílem Obecného nařízení je přizpůsobení právního rámce ochrany osobních údajů dnešní době, dosažení větší jednoty právního rámce ve všech zemích, na které dopadá, posílení práv subjektu údajů a v neposlední řadě je snahou dosáhnout sjednoceného výkladu Obecného nařízení a dozoru jednotlivými dozorovými úřady“ [10].

Nařízení GDPR bylo do českého právního řádu adaptováno zákonem č. 110/2019 Sb., o zpracování osobních údajů (dále ZoZOÚ).

3.2.2 Osobní údaj

Osobním údajem se rozumí celá řada informací. Podle článku 4 GDPR se osobními údaji rozumí *„veškeré informace o určené nebo určitelné fyzické osobě a to tehdy, pokud lze na jejich základě subjekt přímo či nepřímo identifikovat."* Osobní údaj tak může tvořit kombinace několika údajů, které samy o sobě osobními údaji být nemusí. Konkrétní výčet, co vše se osobním údajem rozumí, tedy neexistuje. Z logiky věci také vyplývá, že co není pro jednoho osobním údajem, může pro druhého osobním údajem být.

Příkladem může být pan Jan Novák. Takové jméno konkrétní osobu neoznačí, pokud však souběžně s jménem uvedeme, kde pan Novák pracuje, o osobní údaj už se jedná. Totéž může platit o e-mailové schránce, pokud se v ní vyskytuje jméno. Pokud bude e-mailová schránka jmeno.prijmeni@seznam.cz moc toho sama o sobě nevypoví. Pokud však bude jmeno.prijmeni@cvut.cz, mohlo by se jednat o konkrétní určitelnou osobu [11]. Pojmenování schránky jmeno.prijmeni@seznam.cz však pochopitelně při vytěžování dat z otevřených zdrojů svou hodnotu má, což bude zohledněno mj. ve výzkumné části práci.

3.2.3 Zvláštní kategorie osobních údajů (též citlivé údaje)

Zvláštní kategorií osobních údajů se rozumí údaje, které v případě zneužití mohou subjekt ve společnosti poškodit nebo způsobit diskriminaci. Jedná se např. o údaje o rasovém nebo etnickém původu, politickém přesvědčení, náboženském vyznání, sexuální orientaci, nebo trestní minulosti osoby. Do této kategorie jsou také zahrnuty genetické a biometrické údaje, tedy údaje, které jsou neměnné a vyžadují přísnější opatření při jejich zpracování [12].

3.2.4 Zpracování osobních údajů

V souvislosti s tématem práce je nutné definovat pojmy shromažďování a zpracování osobních údajů a také pojmy správce a zpracovatel osobních údajů, se kterými legislativa pracuje. Zpracováním osobních údajů se rozumí operace s osobními údaji, a to od jejich shromáždění (systematický postup s cílem získat osobní údaje a tyto uložit na datový nosič za účelem jejich zpracování) až po jejich využití, tedy v případě mé práce analýzu, a závěrečnou likvidaci.

Defacto jsem se tedy stal správcem osobních údajů, neboť jsem určil účel a prostředky pro zpracování osobních údajů a zároveň i zpracovatelem, který určenou činnost vykonal.

Vzhledem k tomu, že výzkumná část této práce je zaměřena na sběr a vyhodnocení údajů pomocí OSINT, je vhodné zmínit, že v souladu s § 16 odst. 1 ZoZOÚ je vědecký výzkum právním důvodem pro zpracování údajů a dle § 16 odst. 3 ZoZOÚ není třeba dotčené subjekty o zpracování informovat, neboť by poskytnutí informací vyžadovalo nepřiměřené úsilí.

Problematiku zpracování osobních údajů pro účely diplomové práce dosud pravděpodobně nikdo neřešil a je zde prostor pro různé názory. Aplikovat by tak bylo možné i výjimku pro účely akademického projevu dle § 17 odst. 1 ZoZOÚ, kdy navíc podle § 17 odst. 2 písm. b) ZoZOÚ není třeba poskytnout subjektům identitu správce, pokud lze zpracování osobních údajů oprávněně očekávat. Což pochopitelně platí, neboť se jedná o otevřená data.

Data, která se na síti Internet nachází, mohla být zveřejněna vědomě samotným uživatelem nebo s jeho souhlasem uživatele. Pokud byla uveřejněna bez vědomí a souhlasu uživatele, nejsem tuto skutečnost schopen rozpoznat, a proto v dobré víře pracuji s tím, že data byla shromážděna a zveřejněna oprávněně.

3.2.5 Právo být zapomenut

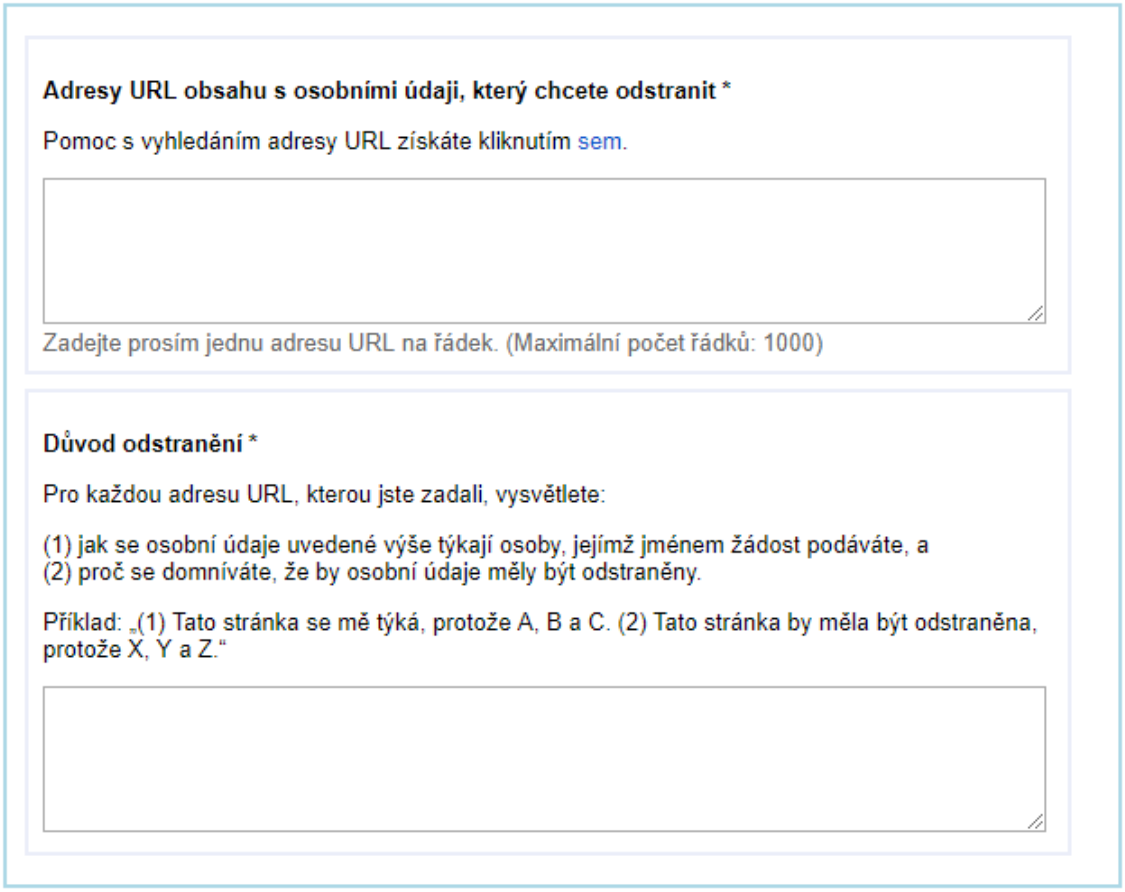
Z hlediska problematiky OSINT je právo být zapomenut jedním z pilířů ochrany uživatelů na síti Internet. Jakýmsi milníkem v této problematice byl Rozsudek Soudního dvora EU C-131/12 ze dne 13. 5. 2014 [13]. Kolouch ve své knize CyberCrime uvádí:

„Mario Costeja González si v roce 2010 úřadu na ochranu osobních údajů ve Španělsku postěžoval, že je mu ve výsledcích hledání v Googlu po zadání jeho jména zobrazován odkaz na novinové články z roku 1998, kde se psalo o dražbě jeho majetku kvůli dluhům na sociálním pojištění" [14, s. 174].

Po společnosti Google tedy požadoval odstranění výsledků, které si o něm mohl najít kdokoli na světě. Vzhledem k tomu, že Google odmítal výsledky odstranit, došel tento spor až před soudní dvůr EU, který dal za pravdu Gonzálezovi.

„Právo být zapomenut je podle Rozsudku (zejména pak podle jeho čl. 91) jednou z esencí práva na ochranu soukromí jednotlivce, spočívající zejména v tom, že tento jedinec má právo, aby bylo ve webovém vyhledávači vymazáno zobrazení seznamu výsledků vyhledávání provedeného na základě jména osoby a/nebo informace týkající se této osoby, a to také v případě, kdy toto jméno nebo tyto informace o dané osobě nebyly předtím nebo současně vymazány z uvedených webových stránek" [15].

Toto právo lze tedy díky rozsudku aplikovat (nejen) ve vyhledávači Google. Formulář, který zde byl pro účely výmazu vytvořen, zahrnuje mimo kontaktní údaje, zmocnění a vztah také pole pro zadání URL adres, které mají být z vyhledávání odstraněny. Důvod žádost zde uvádí osoba podávající žádost, a to na základě dvou stěžejních informací viz obrázek č. 3.



Obrázek 3 – Formulář Google k uplatnění práva být zapomenut [16]; snímek autora

Podobně lze nechat obsah odebrat i u ostatních služeb Google a stejné plnění lze pochopitelně požadovat i od všech dalších subjektů, pokud je účel zpracování údajů v rozporu s GDPR.

O legislativě je v této práci pojednáno ze dvou důvodů. Tím prvním je, že v průběhu práce bude ve výzkumné části s osobními údaji pracováno, druhým důvodem je skutečnost, že s účinností nařízení GDPR došlo k viditelným změnám v prostředí sítě Internet. Na straně jedné GDPR chrání uživatele, na straně druhé však byly zásluhou tohoto nařízení odstraněny některé zdroje dat, které mohly být důležité při prověřování a vyšetřování trestné činnosti. Dalším viditelnou změnou jsou obtěžující upozornění o zpracování osobních údajů ze strany provozovatelů různých webů, kteří plní své úkoly v oblasti ochrany osobních údajů nad rámec povinností.

3.3 Hrozby plynoucí z OSINT

V předchozí kapitole bylo zpravodajství z otevřených zdrojů (OSINT) představeno jako možný nástroj pro získávání informací. V dalších kapitolách se budu věnovat problematice OSINT z jiného hlediska, a to jako potenciální hrozbě pro každého jednotlivce. Zde je důležité vzít v potaz i skutečnost, že kvalita získaných dat přímo souvisí s potřebou potenciálního útočníka. Potřeba proto ne vždy musí souviset s množstvím dat, naopak jediný citlivý údaj může vést k viktimizaci. Pro sexting tak může postačit nalezení telefonního čísla na objekt zájmu, pro vydírání a šikanu mohou být zneužity intimní fotografie, stalker může zneužít údaje o bydlišti a pracovišti nebo fotografii na sociální síti. Zloděj ocení informaci o odjezdu na dovolenou, případně fotografie z letiště při odletu, které v ideálním případě předcházela dokumentace vybavení domácnosti [17]. Řešením by bylo provést upload fotek až po návratu a případně využít službu HootSuite, která umožní plánovat budoucí nahrání, a to i na několik sítí nebo profilů zároveň.

Často se setkávám s lhostejností k tomu, jak je s našimi údaji nakládáno. Osobně přikládám vinu tomu, že si málokdo uvědomuje souvislost mezi následky (např. dokonaný trestný čin), a prvotní příčinou problému. Vyhnout se možnému zneužití údajů preventivně je přitom mnohem snazší, než se pokoušet o nápravu a dožadovat se smazání dat, která již nemáme pod kontrolou a o kterých už ani nevíme, kde všude se vyskytují. Chyb se přitom dopouští nejen samotní uživatelé, ale také orgány veřejné moci a korporace, které uveřejňují více než nezbytně nutné množství informací.

Údaje dohledatelné na síti, ať už se jedná o údaje vložené samotnými uživateli, zpracovatelem údajů nebo následkem úniku informací, mohou být kdykoli zneužity. Vše záleží jen na kreativitě a motivaci útočníka.

Z praktických důvodů jsem k tématice uveřejňování dat na síti Internet zvolil hrozby, protože si pod nimi každý dovede představit konkrétní problém. OSINT bude v této kapitole představen jako možný nástroj pro potenciálního útočníka.

Z hlediska způsobu sběru dat a jejich využití jsem hrozby rozdělil do dvou kategorií na plošné a individuální, a to dle vektoru prvotního útoku. U hrozeb uvedu i konkrétní příklady, rizika a také vazbu na problematiku OSINT. Ve výzkumné části budou tyto hrozby zohledněny při kategorizaci zjištěných údajů.

3.3.1 Plošné hrozby

Kategorie zahrnuje hrozby, které nemají ze své podstaty cílit na konkrétní osobu. Rozhodujícím kritériem je kvantita a útočníkovým úmyslem ani není získat data od všech cílů. Neznamená to však, že by se jednalo o méně sofistikované a promyšlené útoky.

Útočník může využívat například databáze e-mailových adres, které lze získat za úplatu nebo stažením z Internetu, popř. si je lze i vlastnoručně vytvořit. Data je možné sbírat i použitím automatizovaných nástrojů a dotazů ve vyhledávačích [18].

Tyto nástroje budou představeny v kapitole 3.5.2. Vyloučen není ani sběr dat z vytipovaných oblastí, jako jsou například rejstříky firem nebo prostředí internetových seznamek.

3.3.1.1 Spam

Spamem se rozumí jakákoli nevyžádaná pošta, která je doručena prostřednictvím elektronické pošty. Jejím obsahem bývají nejčastěji reklamy na produkty a služby. Strana rozesílající spam profituje v případě, kdy adresát poštu otevře a proklikne odkaz, nebo objedná zboží. Původní masové rozesílání e-mailů již poměrně úspěšně mnohé společnosti blokují a e-mail tak sice dorazí, nicméně je kategorizován jako spam a zpravidla dochází k jeho automatickému smazání. Většina lidí si je navíc vědoma rizik vyplývajících z otevírání

irelevantních e-mailů a tyto maže nebo ignoruje. Dříve jsme se mohli často setkat s e-maily, které např. nabízeli prodej léku Viagra. Anti-spamová ochrana takové zprávy často blokovala podle výskytu „závadových" slov. Spammeři pak samozřejmě e-maily upravili, aby těmito filtry prošly a písmena nahrazovali tak, aby slova neztratila čitelnost – např. I/l/1 (velké i, číslice jedna a malé písmeno L). Použití takové techniky se nazývá obfuskace a může posloužit i v boji proti spamu a pro ochranu uživatelských dat.

Z trestně-právního hlediska není spam nijak postižitelný. S ohledem na právo svobody projevu dle čl. 17 Listiny základních práv a svobod je omezení spamu omezením tohoto práva ve prospěch práva na ochranu osobní integrity dle čl. 10 odst. 2 Listiny. Spam je podle Koloucha postižitelný podle zákona č. 127/2005 Sb., o elektronických komunikacích (dále ZoEK):

„Mimo trestní právo je možné postihnout spamera dle § 119 odst. 1 písm. h) či i) ZoEK, kde se osoba dopustí přestupku, pokud v rozporu s § 93 použije adresu elektronické pošty pro odeslání zprávy nebo zpráv třetím osobám bez souhlasu držitele adresy elektronické pošty, nebo pokud v rozporu s § 96 odst. 1 ZoEK nabídne marketingovou reklamu nebo jiný obdobný způsob nabídky zboží nebo služeb účastníkovi nebo uživateli, který uvedl, že si nepřeje být kontaktován za účelem marketingu" [14, s. 234].

Pokud se tedy útočníkovi nějaký spam přeci jen podaří protlačit do doručené pošty, je třeba toho využít lepším způsobem, než jen doufat v provizi za zobrazení reklamy nebo nákup zboží. Prvním takovým příkladem „lepšího způsobu" jsou nigerijské dopisy.

3.3.1.2 Scam (podvod), nigerijské dopisy

V souvislosti s OSINT a rizikem plošného sběru e-mailových adres jsou dnes již poměrně dobře známé nigerijské dopisy (angl. též Nigerian scam nebo 419 scam). Podvod tohoto druhu je znám už dlouhou dobu, a zatímco médium pro šíření zpráv se z dopisů transformovalo do e-mailu, princip zůstal stále stejný

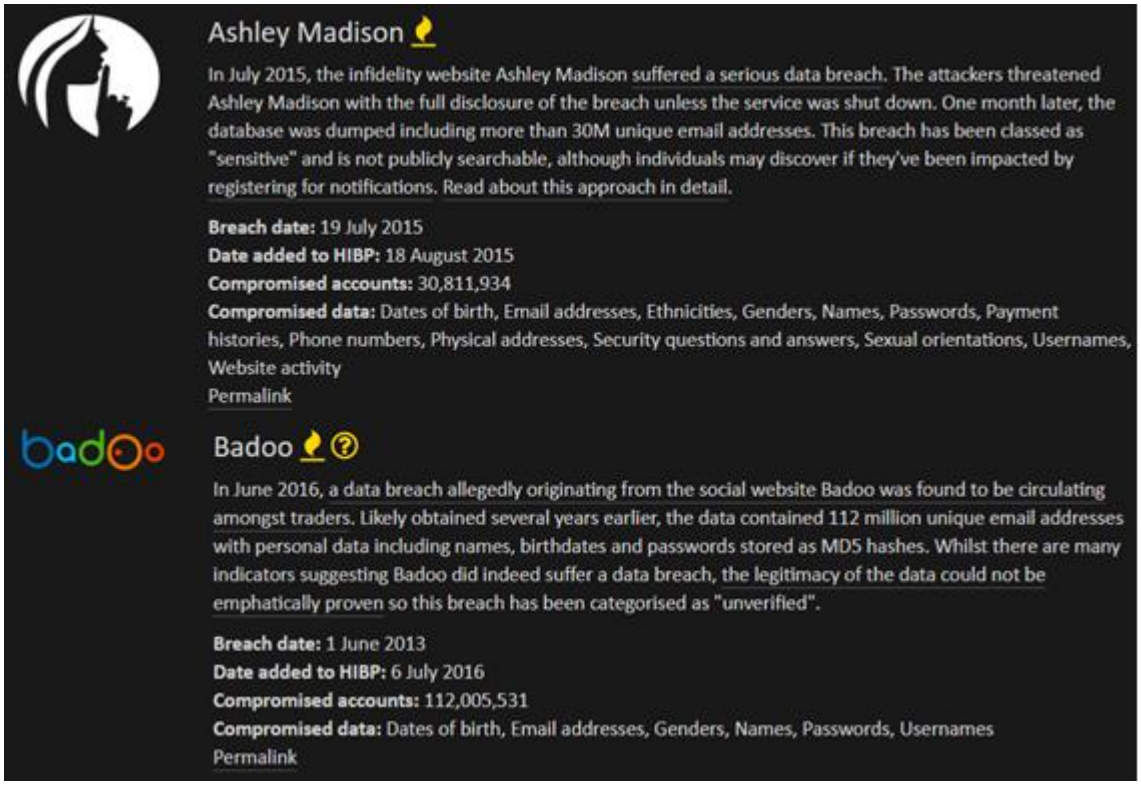
[19]. Vazba na Nigérii je zde kvůli první vlně zpráv, která z Nigérie pocházela [20] a příběhy se navíc často odehrávají právě v Africe. Finanční prostředky také mnohdy míří právě do Nigérie.

Oběť je útočníkem oslovena a pokud zareaguje, je s ní po krátkou dobu udržována korespondence. Na tu po několika týdnech navazuje žádost o finance, která má u oběti vzbudit citovou soustrast nebo vidinu zisku. Jednat se může o půjčku na cestu na pohřeb blízkého, úhrady za právní služby nebo proclení cenného nákladu apod. V příběhu jako hlavní postava figuruje dobře postavený a důvěryhodný člověk, kterým může být doktor, právník nebo armádní generál. Nejprve se jedná o menší sumu peněz, která je, pokud oběť příběhu uvěří, dále postupně navyšována. Finance jsou zpravidla převáděny netransparentním způsobem prostřednictvím služeb jako je MoneyGram nebo Western Union.

Zprávy jsou důkladně propracované nejen z psychologického hlediska, bývají také psány záměrně špatnou angličtinou, čímž útočník provádí prvotní separaci vhodných obětí náchylnějších k viktimizaci, aby neztrácel zbytečně čas. Škody způsobené tímto typem podvodu postupně klesají, nicméně v USA stále dosahují výše 700.000 $ ročně [21].

V české legislativě je toto jednání kodifikováno jako podvod dle § 209 zákona č. 40/2009 Sb., trestního zákoníku (dále jen trestní zákoník). Útočník může OSINT v souvislosti s tímto typem podvodů použít pro zvýšení efektivity útoku. Namísto toho, aby rozeslal a následně vyhodnocoval odpovědi z e-mailových adres, o jejichž oprávněném uživateli nic netuší, může použít např. uniklou databázi e-mailů z některé ze seznamek. Pro e-mailové adresy z takové databáze je pak výrazně vyšší šance, že jejich uživatel bude mít o seznámení zájem. Útočník pak může své zprávy vhodně upravit dle informací, které zná o své oběti již předem. Pokud navíc únik obsahuje i další údaje jako pohlaví, věk, děti a rodinný stav, může útočník vybrat specifický okruh osob, které osloví – např. se může zaměřit na kategorii rozvedených žen ve věku 40–50 let. Známé případy

úniků databází seznamek, ze kterých by mohl útočník čerpat informace jsou např. Ashley Madison (byť zde bylo žen jen minimum) a Badoo. Detaily o těchto a dalších únicích jsou k dispozici na stránkách Haveibeenpwned.com, odkud byl pořízen snímek v obrázku č. 4.



Obrázek 4 – Snímek úniků ze seznamovacích portálů [22]; snímek autora (upravený)

Příklad postupu extrakce e-mailových adres pro cílení takových podvodů, je popsán v kapitole 6.3.

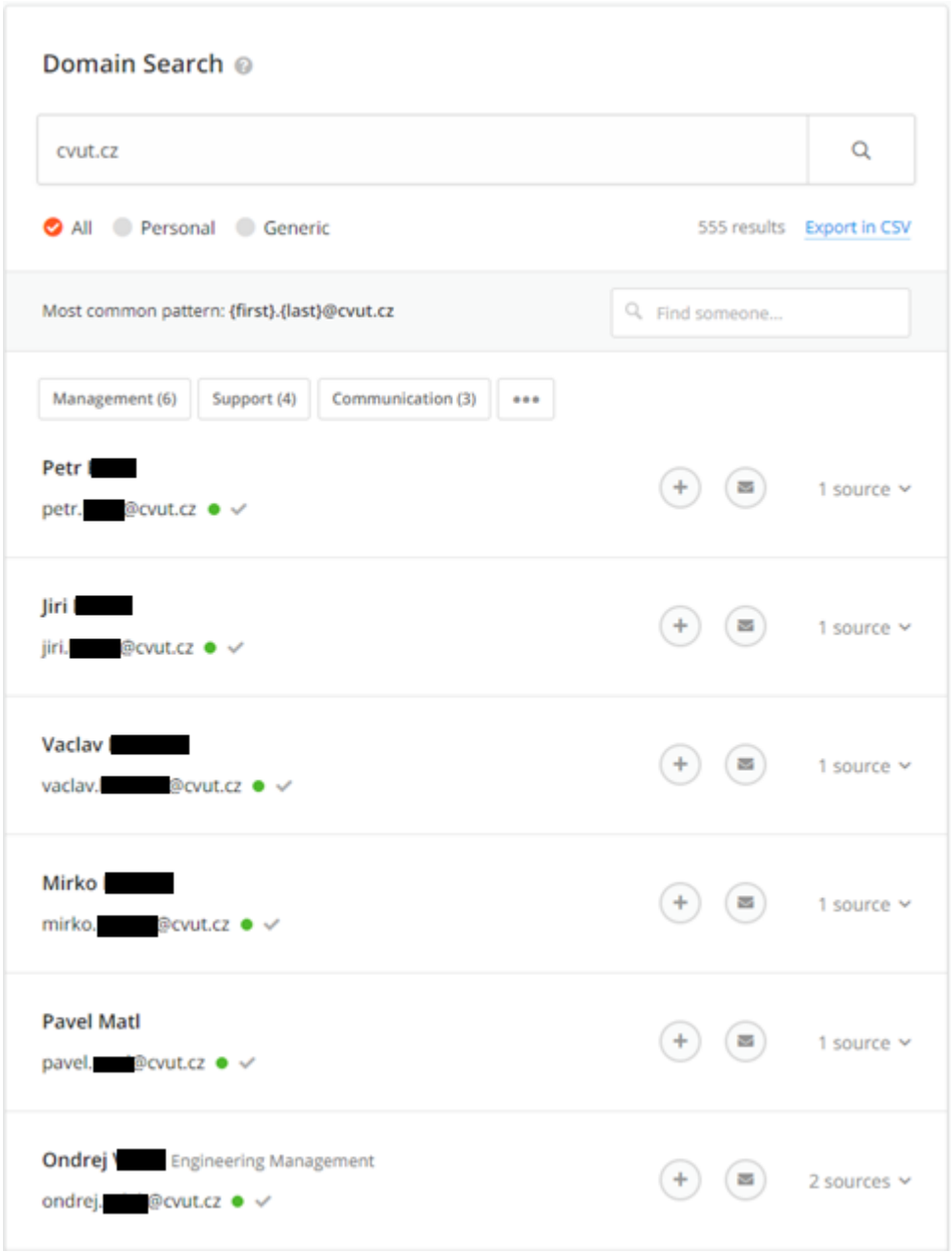
3.3.1.3 Šíření virů, ransomware

Ransomware (z anglických slov: ransom – výkupné a software – program) je škodlivý kód, který útočníkovi poskytuje možnost vydírat oběť, a to nejčastěji pod pohrůžkou ztráty nebo znepřístupnění dat.

První fází útoku je rozšíření infikovaných souborů, nejčastěji právě využitím různých mail-listů. V případě úspěšného útoku na korporátní e-mailovou adresu mohou být dopady pro společnost likvidační.

Vektorem šíření ransomware je nejčastěji e-mail. I přes vysokou úspěšnost antispamových a antivirových řešení se občas podaří některému z hromadných e-mailů projít. Pokud se tak stane a uživatel není dostatečně znalý, snadno se stane obětí. Na vině je pak velmi často zastaralé programové vybavení bez aktualizací. V případě cílených útoků, nelze klást uživateli za vinu, že např. otevře přílohu nazvanou „faktura“, když se navíc tváří, jako by byla odeslána z e-mailové adresy obchodního partnera (tzv. „spoofing“ – podvržení adresy odesílatele).

E-mailové adresy, na které je ransomware zasílán ve větších ransomwarových kampaních (WannaCry, Petya, NotPetya [23]) jsou v řadě případů útočníkem získávány obdobným způsobem, jako pro jakýkoli jiný spam. Jak bylo demonstrováno v předchozím bodu, je možné poměrně snadno extrahovat relevantní e-mailové adresy. Zda útočník zvolí mezi cíleným útokem na korporace, nebo na osobní e-mailové schránky, je už na jeho rozhodnutí. Pokud je cílem konkrétní osoba či korporace, je možné využít další nástroje z kategorie OSINT. Jedním z nich je např. Hunter.io, pomocí tohoto nástroje lze získat seznam e-mailových adres, které jsou hostovány na konkrétní doméně a objevily se v otevřených zdrojích. Pro ukázku jsem použil doménu cvut.cz, jak je zachyceno na obrázku č. 5.



Obrázek 5 – Snímek z nástroje Hunter.io. ***Most common pattern*** *– nejčastější formát e-mailové adresy;* ***source*** *– zdroje a jejich počet [24]; snímek autora (upravený)*

E-mailové adresy nástroj získává automatickým prohlížením různých webových stránek. Zdroje dat, pokud útočníka zajímají, lze na této stránce zjistit rozkliknutím menu u každého jména. Veškerá data pak lze exportovat do formátu CSV a stáhnout. Nástroj navíc zobrazuje role uživatelů, tedy například, zda se jedná o člověka z managementu, což může usnadnit útočníkovi práci, při sofistikovanějších útocích.

Jeden z prvních známých a méně závažných případů ransomware v novodobé historii byl tzv. policejní vir, který se v různých podobách objevuje dodnes. Jedna z podob je zachycena na obrázku č. 6.

Obrázek 6 – Snímek „policejního viru“ [25]

Ransomware je malwarem (škodlivý program), který blokuje používání PC nebo prohlížeče a vyžaduje uhrazení nepříliš velké částky (řádově několik tisíc korun) jako pokuty a za spáchání trestného činu. Trestná činnost, která byla uživateli kladena za vinu přitom byla cílena buď na porušování autorského práva, nebo na trestnou činnost související s pornografií s cílem navodit pocit strachu a následným nabídnutím možnosti řešit situaci „bez ostudy“. Přestože je pro odborníka zřejmé, že se jedná o nesmysl (ustanovení trestního zákoníku neodpovídají nebo ani neexistují, trestní věc nelze řešit zaplacením pokuty přes PaySafe), mohla i laická veřejnost odhalit, že něco není v pořádku díky nedokonalému překladu do českého jazyka. Ransomware však v minulosti způsobil nemálo případů, kdy se zmatení občané dotazovali, co mají dělat [26].

Z této původně nepříjemné, nicméně nepříliš závažné kriminality, se vyvinula jedna z největších kybernetických hrozeb dnešní doby a postrach všech IT oddělení. Jen v USA se škody způsobené v souvislosti s ransomware za loňský rok odhadují na 7,5 miliardy amerických dolarů [27].

Známé případy ransomware z poslední doby postihly i instituce v České republice, a to nemocnici Benešov a společnost OKD. Obě má na svědomí ransomware Ryuk. V případě Benešovské nemocnice hejtmanka kraje Jaroslava Pokorná Jermanová uvedla, že nedošlo k žádnému požadavku na výkupné [28], což příliš neodpovídá modu operandi skupiny, která Ryuk šíří. Ta své cíle volí právě na základě bonity a důležitosti [29] (nejedná se tedy o plošné šíření ransomware) a požaduje zaplacení za dešifrování dat, které však nemá v úmyslu provést [30].

3.3.1.4 Phishing

„V českém překladu je možné vyložit tento termín jako „rybaření“. Jde o podvodnou techniku, která je založena na získávání údajů, jimiž mohou být hesla, kreditní karty nebo jiné údaje. Většinou je tato metoda využívána v elektronické komunikaci, kde se pod nějakou záminkou (e-maily ze služby, banky, sociální sítě), snaží získat z uživatelů citlivé údaje“ [31, s. 123].

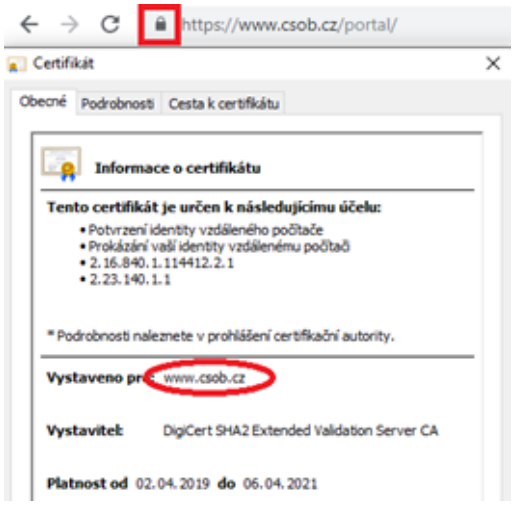
Nejsnazší obranou před phishingem je podezřelé zprávy vůbec neotevírat. Zprávy nejčastěji obsahují zavirovanou přílohu nebo podvodný odkaz. Pokud zpráva působí natolik důvěryhodně, že ji oběť otevře, existuje několik možností, jak postupovat dále.

Přílohu je možné nad rámec běžné ochrany scanovat např. službou VirusTotal.com, kde je prověřena hned několika antiviry. V případě, že chce útočník lákat oběť k návštěvě webu, je vhodné ověřit, že odkaz skutečně směřuje na zamýšlený web.

Je nutné brát zřetel na již zmíněnou možnost obfuskace, kterou může útočník využít. Pro ukázku jsem vytvořil jednoduchý kód, který je na obrázku č. 7, je zde zapsáno písmeno malé „L“ jako číslice „1“.

Obrázek 7 – Obfuskace kódu HTML. ***Dole*** *– reálný odkaz na web s číslicí 1 místo písmene malé „L“;* ***vpravo*** *– odkaz, který vidí uživatel [32]; snímek autora (upravený)*

Další příklady, jak útočník využívá adres URL ke zmatení uživatele, jsou poměrně dobře popsány na webu cleverandsmart.cz [33]. Možností ochrany po návštěvě webu je ověřit správnost certifikátu, tedy že se skutečně nacházíme na webu, na kterém být chceme, a že je spojení šifrováno pomocí SSL/TLS (v adrese se nachází https), jak je patrné z obrázku č. 8.

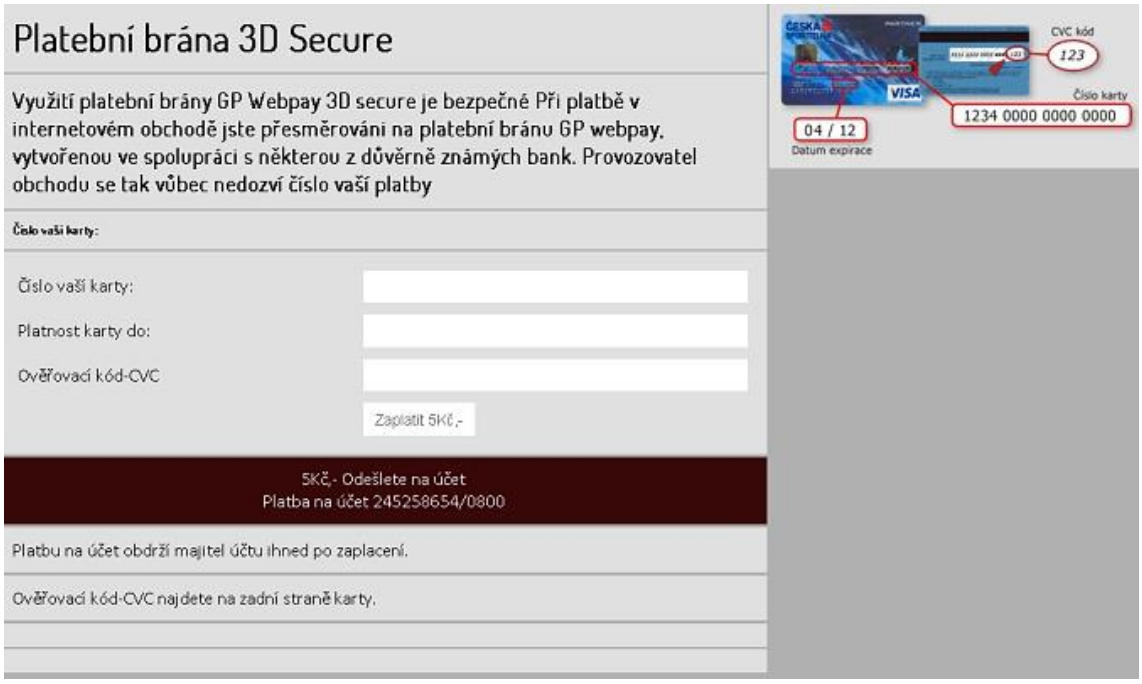


Obrázek 8 – Certifikát internetového bankovnictví. ***Nahoře*** *– ikona symbolizující zapnuté šifrování;* ***vprostřed*** *– potvrzení, že je certifikát skutečně vystaven pro danou doménu [34]; snímek autora (upravený)*

Případy, kdy je phishingový útok cílený na konkrétní uživatele nebo společnosti, nazýváme spear phishing. Tomuto útoku se budu věnovat v kapitole 3.3.2.1.

Poměrně sofistikovaná série phishingových podvodů, se odehrála přibližně v období 2014–2015. Pachatelé nejprve získali přístup k účtů na sociální síti Facebook a následně kontaktovali veškeré přátele s žádostí o půjčku menšího obnosu peněz z urgentních důvodů (řádově několik desítek až sta korun). Podvod však nebyl prováděn kvůli „pár korunám“. Pokud oběť na žádost reagovala, zaslal jí pachatel stránku se vzhledem platební brány. Na té měla osoba předvyplněnou částku k platbě a výzvu k vyplnění dalších údajů o platební kartě (číslo, platnost, kód CVV/CVC).

Pokud osoba údaje vyplnila a potvrdila „platbu“, pachatel tyto údaje získal a obratem zadal transakci mnohem větší finanční částky. Následně si vyžádal od oběti potvrzovací kód a dokončil převod. Pro tyto podvody byly použity stránky podobné stránce vyobrazené na obrázku č. 9.



Obrázek 9 – Snímek phishingového webu [35]

3.3.1.5 Kompromitace účtů / automatizované prolomení hesla

Na síti Internet lze nalézt obrovské množství databází, které obsahují uniklá data v různé kvalitě. Vážná situace nastává v případě, kdy únik obsahuje hesla spolu s e-mailovými adresami. Hesla by konvenčně měla být na straně

provozovatele služby „hashovaná“ (provedení kontrolního součtu, jednosměrná funkce), což však samo o sobě nestačí, pokud se jedná o slabé heslo a pokud je hash provedena bez dodatečného „saltu“. V takovém případě může být heslo reverzně zjištěno mnohem snáze. Salt je řetězec, který k uživatelskému heslo přidává administrátor. Výslednou hash pak zpravidla nelze prolomit slovníkovými útoky.

O tom, že k úniku došlo, se často ani nemusíme dozvědět. Velké korporace se na jednu stranu zpravidla chovají zodpovědně a v případě úspěšného kybernetického útoku na jimi držené údaje vyrozumí dotčené uživatele. Na druhou stranu je mnohdy ve hře jejich pověst. Mohou se tak snažit zastírat, že k nějakému útoku došlo nebo mohou problém zlehčovat.

Stále diskutovanou a připomínanou skutečností je síla hesla. Většina uživatelů si pod sílou hesla představí jeho délku a složitost. Není tomu tak. Dnes se hesla zpravidla neprolamují silou, ale pomocí slovníkových útoků. Prolamování hesel pomocí hrubé síly je ekonomicky náročné a bývá až poslední možností. Pokud útočník neprovádí cílený útok, pravděpodobně ani k této metodě nepřikročí. Silné heslo je tedy takové, které zatím nikdo jiný nepoužil. To platí zejména v případech, kdy heslo opakovaně používá jedna osoba. Jak na svém webu uvádí Špaček: *„Silné heslo nevypadá tak, že obsahuje minimálně jedno malé a jedno velké písmeno, jedno číslo a jeden speciální znak, protože velké dáte na začátek a číslo na konec a za něj dáte vykřičník“* [36].

Velmi užitečnou stránkou je https://haveibeenpwned.com. Zde je po zadání e-mailové adresy možné ověřit, jestli nedošlo k úniku hesla vázanému na zadanou e-mailovou adresu. Pochopitelně služba pouze uvede, v jakém úniku byla daná adresa uveřejněna a samotné heslo neprozradí. Na webu jsou také popsány jednotlivé úniky dat včetně rozsahu uniklých informací.

3.3.2 Individuální hrozby, sociální inženýrství

Samotná možnost profilace cíle je objektivně rizikovým faktorem. Útočník si primárně může vybrat snazší cíle, přičemž jeho atraktivita může být dána i dalšími vlastnostmi jako je společenské postavení, majetkové poměry či pracovní pozice. Mezi viktimizační faktory lze zařadit i další kritéria jako např. pohlaví, věk, vzhled, rodinný stav atd.

3.3.2.1 Spear phishing

Spear phishing, který je možné přeložit jako „rybaření oštěpem“, je poměrně výstižný termín pro tento druh útoku. Narozdíl od phishingu, kde se útočník soustředí spíše na oslovení co největšího počtu obětí, je v tomto případě kladen důraz na zacílení důležitých osob a společností. Motivem přitom nemusí být jen ekonomický zisk, jako tomu běžně je v případě phishingu, ale může jít i o náboženské a politické cíle. Útočník se také dostává do užšího kontaktu s obětí nebo jejím pracovním prostředím, provádí průzkum a často využívá i sociálního inženýrství.

Představme si personalistu zodpovědného za obsazení několika IT pozic. Řediteli IT běží lhůty a zoufale potřebuje dobré pracovníky. Personalista na sociální sítě vyvěsí volná místa s žádostí o rozšíření nabídky. Za pár dní se mu na e-mail ozve budoucí kandidát odpovídající požadavkům. Personalista otevře přiložený životopis a infikuje tak počítač malwarem [37].

Personalista přitom nijak nepochybil a choval se zcela racionálně, když otevřel přílohu od neznámého člověka, je to jeho práce. Takových scénářů může být celá řada. V posledních letech jsou populární zejména e-maily s fakturami.

Za tímto účelem může útočník předem získávat data o oběti použitím OSINT. Jeho aktivity by mohly zahrnovat shromažďování firemních e-mailů, zjišťování pracovních pozic konkrétních zaměstnanců ze sociálních sítí a podobně. Průběh útoku by mohl vypadat stejně jako na obrázku č. 10.

Obrázek 10 – Průběh phishingového útoku [37]

První dva body (průzkum a útok) již ve výše uvedeném případě proběhly. Kompromitace účtu také, pokud společnost neměla dostatečně chráněné prostředí. Útočník se tak nachází „uvnitř" a pokud je toho schopen, sbírá další informace. Samotné informace již mohou být cílem, pokud by šlo o podnikovou špionáž. V opačném případě by útočník mohl vyčkávat na vhodnou situaci, popř. se pokoušet získat další oprávnění. Pokud by se dostal k uživatelským účtům zaměstnanců účtárny, mohl by zde např. pozměnit faktury nebo čísla účtů, na které mají být zasílány výplaty. Možností útoku je celá řada.

3.3.2.2 Stalking

„Stalking je soustavné (systematické) a excesivní (tj. evidentně z normy vybočující) obtěžování druhé osoby projevy nevyžádané a umanuté pozornosti, podnícené v zásadě buď údajným či skutečným obdivem (erotickou náklonností), nebo pociťovanou záští, nenávistí či pomstou" [38, s. 55].

Moderní název pro stalking s využitím informačních technologií je kyberstalking. OSINT s problematikou stalkingu velmi úzce souvisí. Informace, které pachatel o své oběti z tohoto prostředí získá, může velmi snadno zneužít a nezřídka se to i děje.

Stalking tedy nemusí nutně znamenat, že se osoba projevuje agresivním chováním. Pachatelem se může stát i osoba, která bude oběti třeba i vyznávat lásku, pokud tak činí obtěžujícím způsobem a opakovaně i přes odmítnutí. Takové jednání se přesněji označuje jako kyberharašení [39].

Z trestně-právního hlediska je v České republice stalking kodifikován jako nebezpečné pronásledování v § 354 trestního zákoníku. Samotné shromažďování dat samozřejmě trestně postižitelné není.

3.3.2.3 Kyberšikana

Kolouch k problematice kyberšikany uvádí:

„Šikana ve světě reálném spočívá ve snaze útočníka ublížit, ponížit, zesměšnit, urazit jiného, ať fyzicky či psychicky. Kyberšikana pak přenáší „klasickou šikanu" do světa virtuálního a umožňuje útočníkovi použít nástroje a prostředky, které mohou mít mnohem větší dopad na oběť, než by tomu bylo ve světě reálném" [14, s. 309].

Dále popisuje několik případů kyberšikany, které skončili smrtí nebo vážnými psychickými problémy [14].

Tato problematika s OSINTem souvisí z toho důvodu, že kyberšikana může zahrnovat mimo jiné umísťování obsahu do veřejné části sítě Internet a také zneužívání obsahu, který se zde již vyskytuje. Často se jedná o videa nebo fotografie, které mají oběť zesměšnit, nebo je na nich zaznamenána klasická šikana. Nahráním takových záznamů pak útočník zpravidla očekává další zesměšnění oběti, a to často nejen okruhem osob, které oběť zná, ale celým světem. Pachatel také může o oběti vyhledat informace, které proti ní použije.

3.4 Zdroje výskytu dat

3.4.1 Sociální sítě, diskusní fóra

Sociální sítí se v prostředí Internetu rozumí služba umožňující komunikaci a vzájemné sdílení informací s ostatními uživateli. Sociální sítě jsou fenoménem dnešní doby. Díky mobilním aplikacím a téměř všudypřítomné konektivitě k síti Internet mohou milióny uživatelů vytvářet, či konzumovat jejich obsah kdykoli a kdekoli.

„Na sociálních sítích nahlížíme do cizích životů, abychom zjistili, co dělají naši přátelé, sousedé, bývalí partneři a spolužáci, celebrity, vzory a známí. A děláme to často, aby nám nic neuniklo. Je to jakási soukromá reality show, ve které si sami určujeme hlavní postavy“ [40, s. 35].

Sociální sítě mohou sloužit pro šíření obsahu i mimo okruh přátel, čehož se hojně využívá pro komerčních sdělení, ale také pro publikaci sdělení a názorů zájmových, politických a náboženských uskupení. Členové těchto sdružení bývají na sítích různými způsoby provázáni, a přestože nemusí být některé informace zřejmé na první pohled, je použitím vhodných metod OSINT možné odhalit jejich vzájemné vazby.

Jednou z prvních sociálních sítí byl projekt SixDegrees.com již v roce 1997 [41]. Vzhledem k tehdejšímu nízkému počtu uživatelů sítě Internet však nebyl tento projekt příliš úspěšný.

Pojmenování SixDegrees vzniklo z teorie šesti stupňů odloučení, kterou koncem 60. let objevil při svém pokusu Stanley Milgram v souvislosti s hledáním odpovědi na otázku problému malého světa. Milgram náhodně zvolil 160 lidí a každému poslal balíček určený jeho známému v jiném městě. Každý z adresátů byl požádán o to, aby na balík uvedl své jméno a balík poslal dále osobě, o které se domnívá, že by mohla cílového adresáta znát. Při tomto experimentu Milgram zjistil, že balíček byl většinou dodán po 5-6 krocích [42].

„Šest stupňů odloučení neznamená, že od každého ke každému vede spojnice o pouhých šesti krocích. Znamená, že velmi malý počet lidí je v několika krocích spojen se všemi ostatními a my ostatní jsme propojeni se světem skrz těchto pár výjimečných lidí“ [42 s. 34–35].

Pokus v roce 2001 opakovali na Columbia University. Jako médium přenosu byl zvolen e-mail a ten měl být zároveň jediným prostředkem k dosažení cíle (tedy bez použití vyhledávačů).

Projekt zahrnoval 60.000 osob a 18 cílových adresátů v 13 různých zemích. Výsledek byl překvapivě podobný Milgramovu, e-mail dorazil po 5–7 přeposláních [41].

Teorii šesti stupňů odloučení lze použít v prostředí sociálních sítí i webu jako takového. Je podstatné tyto poznatky přenášet i do praxe při analyzování dat [42]. Závěr výzkumu provedeného pod záštitou Facebook Research zjistil, že na jejich síti byla v roce 2016 (při 1,6 miliardách uživatelů) průměrná hodnota odloučení 4,57 a má s počtem přibývajících uživatelů klesající tendencí [43].

Tato tendence má vliv i na relevanci vazeb, které lze ze sítě zjistit. Konvenčně se pracuje s předpokladem, že důležité vazby mají vzdálenost na 1 až 2 kroky. 3 kroky jsou již spíše výjimka [44].

Při provádění vytěžování, stejně jako u jakékoli jiné činnosti, má člověk svůj cíl a očekávání. Zjištěné vazby můžeme při vizualizaci označit dle tabulky č.2.

Tabulka 2 – Hodnocení zjištěných vazeb [42]

	Očekávání	
Přínos	**Pozitivní přínos a očekávané výsledky**	**Pozitivní přínos a neočekávané výsledky**
	Zjištění korespondují s cílem	Zjištění doplňují cíl
	Negativní přínos a očekávané výsledky	**Negativní přínos a neočekávané výsledky**
	Zjištění nejsou relevantní	Zjištění vylučují dosažení cíle

Facebook

Ve vztahu k datům, která nám může sociální síť Facebook použitím OSINT poskytnout, došlo k poměrně výrazným změnám. Facebook byl vytvořen původně jako TheFacebook v roce 2004 a začátkem roku 2009 se stal oficiálně špičkou mezi sociálními sítěmi.

Pro vývojáře různých aplikací nabízí Facebook rozhraní pro programování aplikací (API). Nejzajímavějším z hlediska OSINTu bylo dlouhou dobu Graph API, které mělo usnadnit práci s objekty (uživateli, komentáři, fotografiemi atd.). První verze označená 1.0 vznikla v roce 2010 a fungovala až do roku 2015.

Facebook si byl vědom množství osobních údajů, které bylo možné zjistit zneužitím funkcí Graph API, ale nechtěl se vzdát svých příjmů z marketingu [45]. Přestože si uživatel mohl nastavit zabezpečení částí profilu na úrovně: „pouze pro mě“, „pro přátelé“, „pro přátelé přátel“ a „pro všechny“, zacházel Facebook se všemi úrovněmi kromě „pouze pro mě“ stejně a data předával poskytovatelům aplikací třetích stran. Problémem zde nebyl samotný profil uživatele, data totiž unikala ven prostřednictvím jeho přátel, kteří mohli mít takovou aplikaci nainstalovanou.

Díky ohromnému počtu uživatelů je Facebook důležitou platformou pro komerční sektor. V roce 2012 Facebook převzal populární sociální síť Instagram, která slouží primárně pro sdílení fotografií. Z pohledu bezpečnosti Instagram nepřipouští mnoho chyb, jako tomu bylo v případě Facebooku. Pokud uživatel profil označí jako soukromý, je vidět jen titulní stránka s jednou fotografií a popisem.

Dalším produktem provozovaným pod značkou Facebook je od roku 2014 také messenger WhatsApp. Často se objevuje v žebříčcích nejpoužívanějších sociálních sítí, dle mého názoru se však o síť jako takovou nejedná. Program poskytuje služby volání, psaní zpráv a zasílání souborů, informace však nejsou

a ani nemohou být vystaveny veřejně. Částečnou náhradou je také samostatný produkt Facebook Messenger, který je kompatibilní s platformou Facebook, avšak již samostatně fungující.

Funkce Graph API, které byly v minulosti k dispozici, popisuje tabulka č. 3.

Tabulka 3 – Funkce Graph API 1.0 a 2.0 [46]

Permission Group	Permissions	Profile Items
Public profile (default)	public_profile①②	id, name, first_name, last_name, link, gender, locale, timezone, updated_time, verified
App friends	user_friends①②	bio, birthday, education, first_name, last_name, gender, interested_in, languages, location, political, relationship_status, religion, quotes, website, work,
Extended Profile Properties (xpP)*	friends_about_me①, friends_actions①, friends_activities①, friends_birthday① friends_checkins①, friends_education_history①, friends_events①, friends_games_activity①, friends_groups①, friends_hometown①, friends_interests①, friends_likes①, friends_location①, friends_notes①, friends_online_presence①, friends_photo_video_tags①, friends_photos①, friends_questions①, friends_relationship_details①, friends_relationships①, friends_religion_politics①, friends_status①, friends_subscriptions①, friends_website①, friends_work_history①	about_me, actions, activities, birthday checkins, history, events, games_activity, groups, hometown, interests, likes, location, notes, online_presence, photo_video_tags, photos, questions, relationship_details, relationships, religion_politics, status, subscriptions, website, work_history
Extended Permissions (xP)*	read_mailbox①②	inbox

Řada problémů přetrvala i ve verzi 3.0 a to až do června 2019, kdy postupnou implementací verze 3.3 byly zablokovány téměř všechny funkce, které byly zneužívány. K tomuto došlo mj. i následkem kauzy Cambridge Analytica [47].

Do hledání na stránkách Facebook (popř. přímo za URL https://www.facebook.com/) bylo do června 2019 možné vkládat řetězce, které odkrývaly jinak na první pohled neviditelný obsah. Díky tomu šlo mj.:

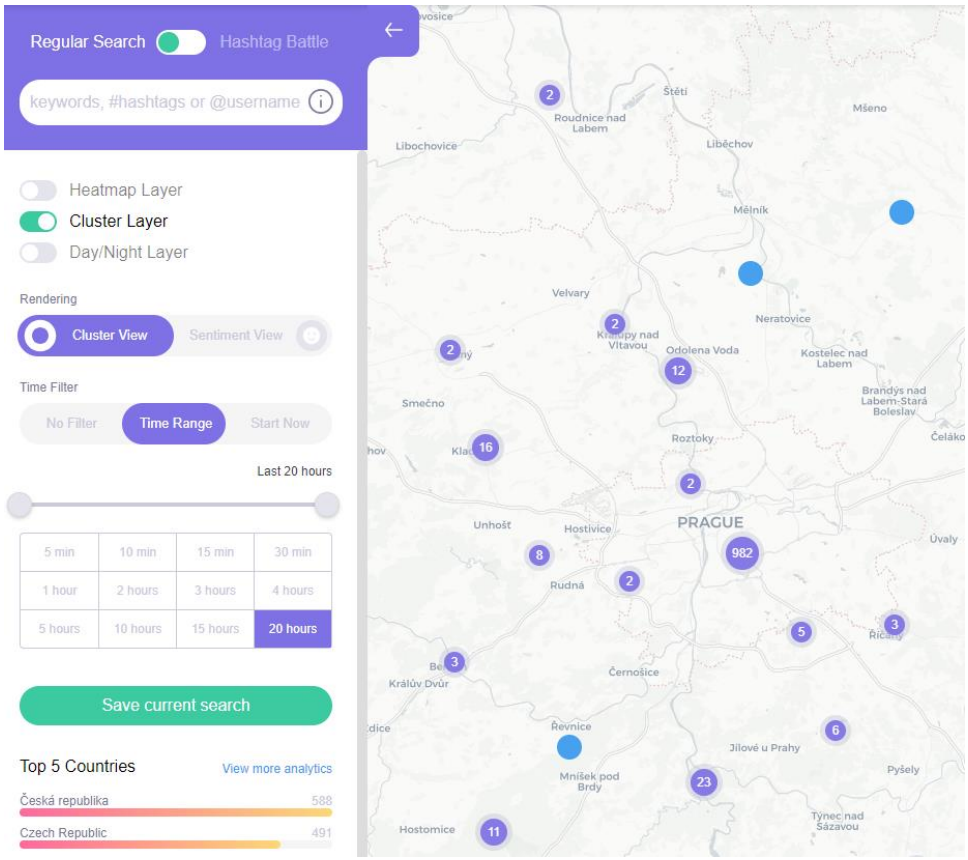
- Odkrýt seznam přátel pomocí funkce společných přátel a to použitím uzivatelA/friends?and=uzivatelB (verze 1.0), který ukázal veškeré společné přátele a později friendship/uzivatelA/uzivatelB (verze 2.0), který zobrazil počet přátel a to vč. dvou konkrétních, náhodně vybraných profilů. Dotazy bylo možné řetězit a udělat si tak představu o seznamu přátel, a to i přesto, že je uživatel měl skryté (vyjma případu nastavení seznamu přátel „pouze pro mě“).
- Zobrazit fotografie uživatele A, pokud na nich byl označen nebo se na nich vyskytoval pomocí uzivatelA/photos-tagged nebo uzivatelA/photos-of.

Existovalo mnoho dalších řetězců, které bylo možné do hledání vkládat. Bylo možné zobrazit příspěvky, kterým dal uživatel „like“, příběhy, ve kterých se objevil atd. V dnešní době lze vytěžovat už jen veřejně dostupná data.

Další, z hlediska této práce však ne tak podstatnou skutečností je, že Facebook shromažďuje i data o neregistrovaných uživatelích. Nejde o nic nelegálního, data mu posílají weby, které využívají funkce Facebook pro marketing. Vzhledem k rozšířenosti platformy jde o vysoký počet stránek a to 8,4 miliónu [48].

Twitter

Ve světě OSINT poměrně oblíbený, v ČR však nepříliš používaný, Twitter je sociální síť určená pro publikování krátkých zpráv. Každý „tweet“, jak je příspěvek na tuto síť pojmenován, s sebou může nést doplňující informace. Jednotlivé tweety v sobě mohou nést také GPS souřadnice, díky čemuž lze vidět, odkud byl příspěvek zaslán přímo na mapě, která je na obrázku č. 11.



Obrázek 11 – One Million Tweet Map [49]; snímek autora

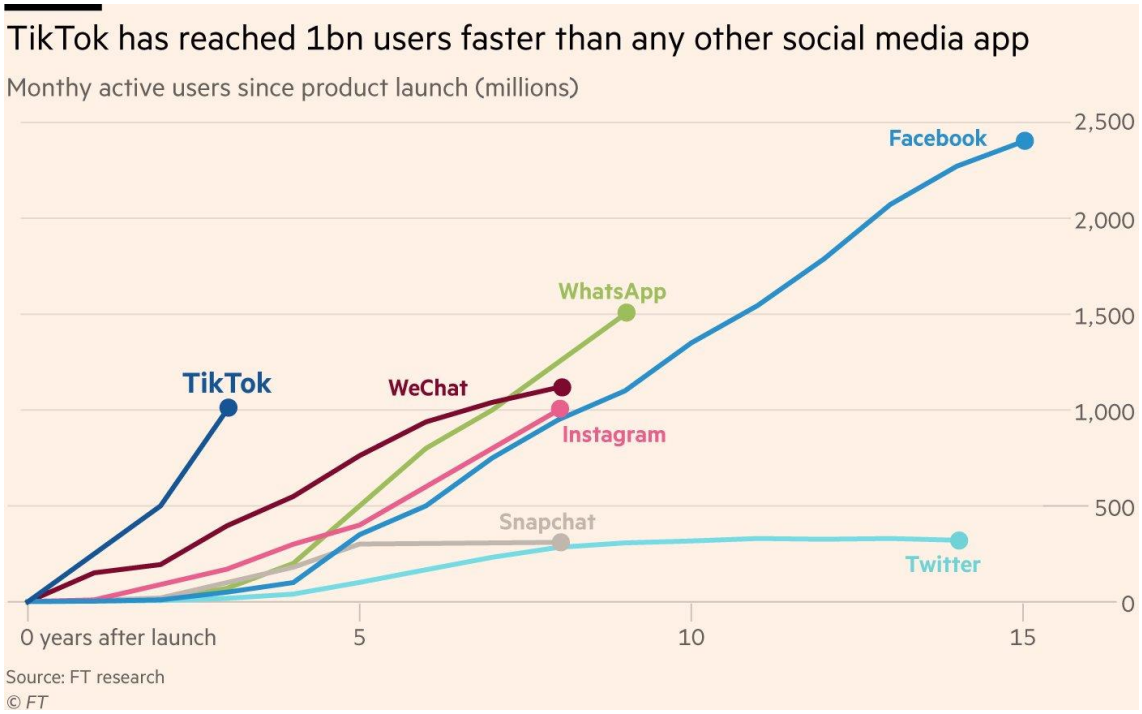
Z důvodu možnosti sdílení zpráv z různých webů Twitter sleduje uživatele obdobně jako Facebook [50].

LinkedIn

Profesní sociální síť s více než 562 milióny členů (v ČR 1, 5mil.) je výborným zdrojem informací pro personalisty, kteří hledají vhodného kandidáta vesměs na pozice vyžadující kvalifikaci nebo zkušenosti. Osoby zde publikují profesní životopisy obsahující studijní a pracovní úspěchy. Ačkoli tak činí zcela záměrně, mohou být informace zde uveřejněné zneužity v souvislosti s dalšími zdroji. Můžeme zde např. zjistit, jak vypadá osoba Jan Novák, která pracuje v konkrétní firmě a následně se pokusit hledat profil i na dalších sítích.

TikTok

Sociální síť původně známá jako Musical.ly zažívá v poslední době významný boom, jak je patrné z článku z listopadu 2019. TikTok nejrychleji rostoucí platformou, která dosáhla 1 miliardy uživatelů, jak je patrné z obrázku č. 12.



Obrázek 12 – Růst počtu uživatelů sociálních sítích v prvních letech po vzniku [51]

Tento růst je však do jisté míry dán tím, že při zrodu ostatních sociálních sítí nebyla tak velká uživatelská základna jako dnes. Navíc se po delší dobu neobjevilo nic nového, co by uživatele zaujalo. Na rozdíl od většiny dalších sítí, které byly koncipovány pro uživatele PC, je síť TikTok primárně zaměřena na uživatele mobilních telefonů. Obsahem jsou krátká, často zábavná videa a stránka, resp. aplikace, neodvádí pozornost od obsahu, který uživatel hledá.

„Jedním z problémů, který postihuje Tik Tok (ale na který narazíme např. i na serveru YouTube), patří úniky videí velmi malých dětí (do 11 let věku), která obsahují sexuálně explicitní obsah. TikTok totiž cílí především na taneční/hudební videa, ve kterých se děti pohybují - třeba poskakují na posteli, dělají různé taneční kreace apod. Na záznamu se pak často objeví i obnažené části těla dítěte, které přitahují pozornost sexuálních abuzérů. A právě tato videa mají velmi vysokou návštěvnost a jsou cíleně vyhledávána, sdílena a rozšiřována. Děti se pak stávají vyhledávaným cílem a jsou prostřednictvím Tik Toku oslovovány dospělými uživateli. Následují žádosti o zaslání dalších sexuálně explicitních fotografií apod“ [52].

Profily na této síti jsou nastaveny jako veřejné, pokud uživatel nastavení nezmění, což v souvislosti s uvedenou citací ze stránek e-bezpeci.cz může být zdrojem problémů.

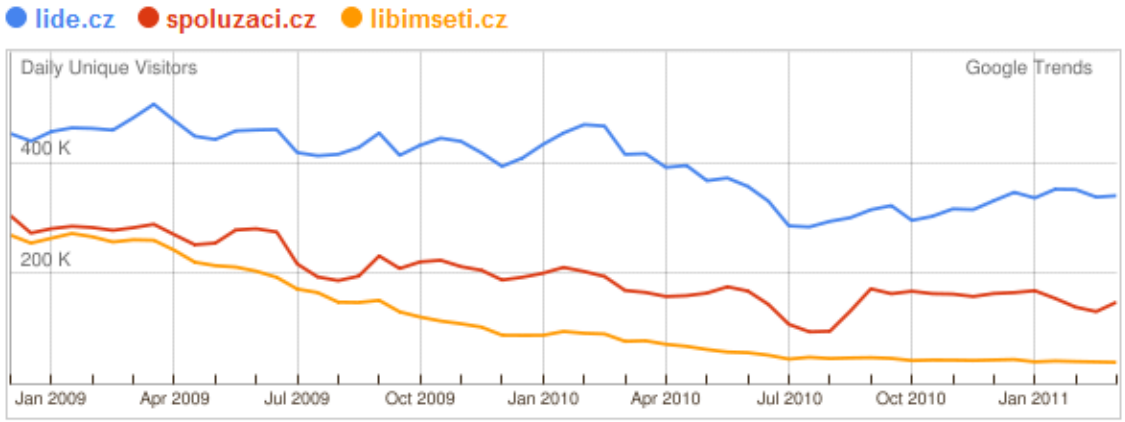
České sociální sítě

V České republice máme i přes tlak globálních sociálních sítí dlouhodobě několik vlastních sociálních sítí. Mezi oblíbené služby patřily historicky různé „chaty“, které částečně vycházely z IRC (Internet Relay Chat). Byly však přístupné z prohlížeče a bylo na nich možné vyplnit vlastní profil vč. fotografií. Uživatel se standardně připojil do místnosti, jejíž obsah mohl být někde uchováván.

Jedním z prvních byl např. XChat.cz, který vznikl na Technické univerzitě v Liberci v letech 1996/1997. Dodnes lze nalézt úryvky různých diskuzí starých i téměř 20 let. Podobných chatů postupně vznikala celá řada, nicméně službě XChat.cz se podařilo přežít dodnes a také na něm lze na některých profilech dohledat (někdy až příliš podrobné) informace. Mě osobně překvapil počet aktivních uživatelů této sítě, neboť například ve 4:30 ráno zde bylo přihlášeno 210 lidí.

V roce 2002 se k XChat.cz připojil chat Libimseti.cz, který XChat.cz postupem času sesadil a také funguje dodnes. Konkrétní počty uživatelů se mi nepodařilo dohledat. V 4:40 hod. je na Libimseti.cz online 69 lidí, a v chatu jsou aktivní 4 uživatelé, nicméně tento web vždy fungoval spíše jako internetová seznamka a podobně jako XChat.cz mají uživatelé možnost vyplnit detailně jejich profil.

Posledním zástupcem kategorie českých sociálních sítí je portál Lide.cz, který provozuje Seznam.cz. Lide.cz byl nejúspěšnější z těchto projektů a denně ho využívalo i přes 400 tisíc uživatelů. Právě sociální síť Facebook tyto lokální sítě zatlačila do ústraní. Propad viditelný na obrázku č. 13 je způsoben právě nástupem sociální sítě Facebook.



Obrázek 13 – Klesající popularita českých sociálních sítí po vzniku Facebook.com [53]

3.4.2 Inzertní portály

Vhodným zdrojem pro sběr kontaktních údajů jsou inzertní portály. Z povahy služby zde uživatelé tyto údaje publikují zcela záměrně. Výjimku v této kategorii tvoří portál Aukro.cz, který je z hlediska OSINT téměř nepoužitelný, neboť o uživateli nezveřejňuje žádné informace kromě přezdívky. Odpovědnost bere portál na sebe a v případě problému drží potřebné informace o uživatelích, které získává prostřednictvím potvrzení zaslaného poštou.

Podobná ochrana nutností kontaktu s osobou přímo prostřednictvím portálu je zavedena také na realitním portálu Bezrealitky.cz.

Mezi velmi známé inzertní portály v České republice patří služby společnosti Seznam.cz, konkrétně: Sbazar.cz, Sauto.cz a Sreality.cz. Portál Sbazar.cz se od předchozích odlišuje a to tím, že jednotlivé inzeráty jsou publikovány s uveřejněním uživatelského jména, ze kterého lze často odvodit e-mailovou adresu osoby (pokud se za jménem nenachází @, doplníme @seznam.cz). Inzeráty jsou dle tohoto jména (e-mailu) dohledatelné i pomocí vyhledávačů, které je indexují.

Dalšími inzertními portály jsou např. Bazos.cz a Hyperinzerce.cz. Inzertní portály budou mj. použity v praktické části práce pro získávání vstupních informací k uživatelům.

3.4.3 Webové archivy

Jak jsem již psal v předchozích kapitolách, lidé by na Internetu měli uveřejňovat jen nutné minimum informací. Na tuto skutečnost jsou upozorňovány dlouhá léta i jednotlivé společnosti. Přestože v některých případech společnosti data, která se týkala vnitřních záležitostí, personálu apod., ze svých webů odstranila, nemusí to automaticky znamenat, že tato data ze sítě zmizela a problém tak byl vyřešen. Realita je totiž jiná, a webová stránka může existovat na řadě dalších míst v podobě tzv. snapshotu [54].

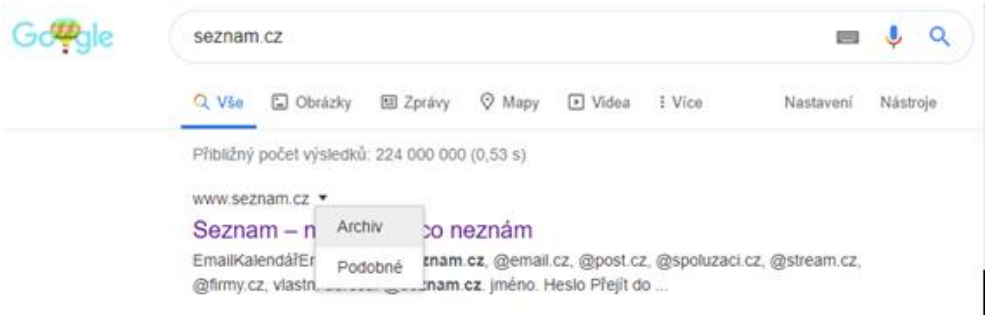
Snapshotem se rozumí otisk dat v určitém čase, byť v tomto případě nemusí být kompletní (často např. z důvodu náročnosti na diskové úložiště neobsahuje obrázky a další mediální obsah). Snímky jsou prováděny v různých intervalech, není tedy zaznamenána každá změna.

Díky archivním službám je možné zobrazit náhled stránky i po jejím smazání. Mým oblíbeným příkladem je snapshot vyhledávače Seznam.cz z roku 1996 na obrázku č. 14.

Obrázek 14 – Archivovaná podoba stránek Seznam.cz z 14. 11. 1996 [55]; snímek autora

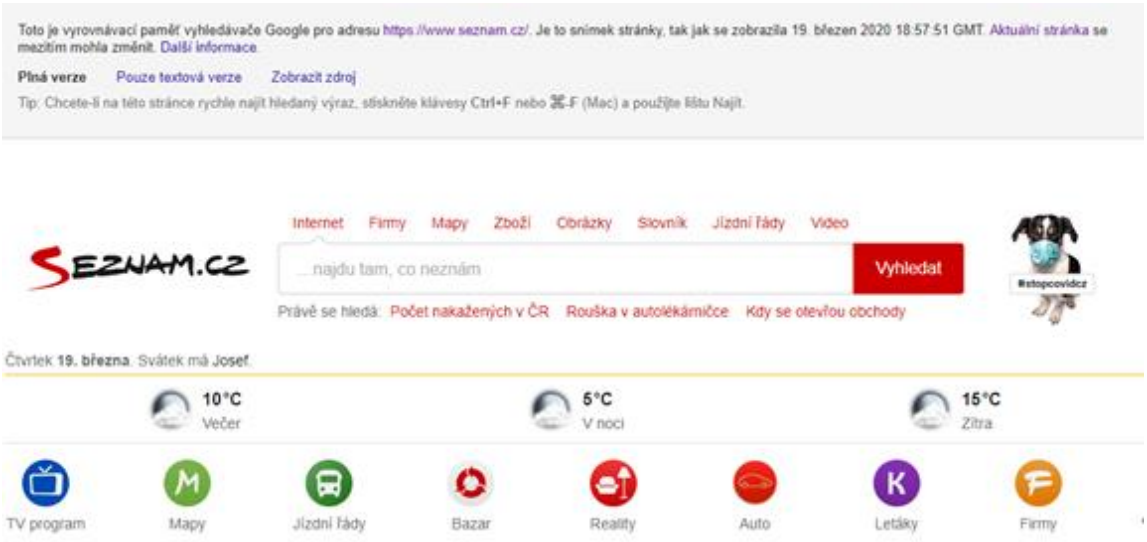
Služeb, které tuto funkci nabízí, existuje celá řada. Vůbec nejznámější a také ty, kterým se v práci budu dále věnovat, jsou stránky Archive.org (Wayback Machine), ze kterých pochází i obrázek č. 14, a Cache vyhledávače Google. Odlišností těchto služeb je, že Wayback Machine vyhotovuje snímky, které průběžně ukládá a zůstávají viditelné navždy. Funkce Cache vyhledávače Google je záležitostí zpravidla dočasnou, neboť k prohlédnutí nabízí jen jeden poslední snímek.

Cache Google je přímo součástí tohoto vyhledávače a pokud je dostupná, lze ji zobrazit kliknutím na šipku vedle odkazu, jak je patrné z obrázku č. 15.



Obrázek 15 – Možnost zobrazení stránek v Google Cache [56]; snímek autora

Po kliknutí pak budeme přesměrování na stránku uloženou v paměti Google. Obrázek č. 16 jsem vytvořil 19.3.2020 v 21:50 hod. a jak je ze sdělení Google patrné, snímek byl vytvořen v 18:57 hod. GMT, tedy v 19:57 hod. místního času. Čas obnovy snapshotu se pro jednotlivé weby liší, některé jsou aktualizovány v řádu hodin, jiné i po několika měsících.

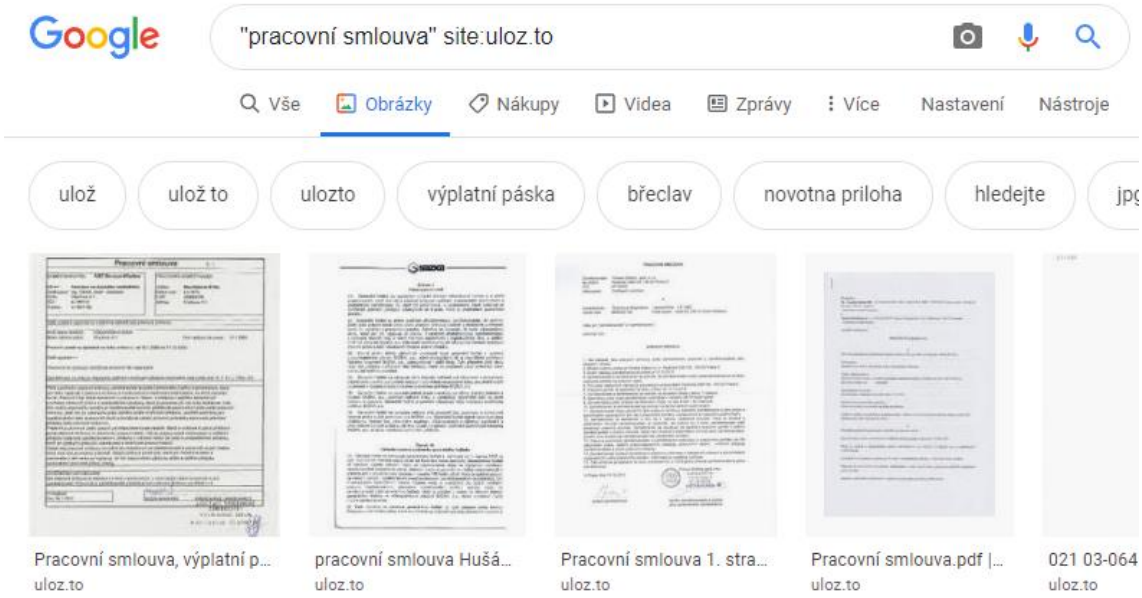


Obrázek 16 – náhled stránky z Google Cache [56]; snímek autora

Webové archivy mohou často prozradit i velmi zajímavé či citlivé informace. Způsob využití, respektive rizika zneužití budou popsána v diskuzi v kapitole 6.4.

3.4.4 Filehostingové servery

Zajímavou oblastí pro aplikaci OSINT jsou filehostingové servery. Jedná se o servery, na které může uživatel umístit téměř libovolný obsah, který bude následně přístupný buď pouze jemu (popř. okruhu uživatelů, kterým sdělí přístupové údaje), nebo celému světu. Mezi nejznámější zástupce u nás patří bezpochyby portál Ulož.to. Obsah je u těchto serverů zpravidla indexovaný vyhledávači a je tak možné jednotlivé soubory vyhledat i prostřednictvím externích služeb. Tato indexace s sebou nese i výhodu možného použití operátorů. Hledáním je možné nalézt smlouvy, mzdové výměry, formuláře pracovní neschopnosti atd. Potenciální útočník zde má k dispozici často i podpisové vzory. V případě Google jsou často zaindexovány také náhledové obrázky, jak lze vidět na obrázku č. 17.



Obrázek 17 – Příklad využití indexace Google [56]; snímek autora

Snad nejoblíbenějším obsahem, který je umisťován na Internet, jsou fotografie. Byť se v tomto případě nejedná přímo o filehostingové weby, mohu zmínit příklad některých škol, které ve snaze získat popularitu plní své weby fotografiemi žáků při různých aktivitách. Fotografie dětí ze školy v přírodě z Itálie, kde děti pózují v plavkách (obrázek č. 18), určitě udělají radost rodičům, nicméně mohou zbytečně přilákat pozornost deviantů a do otevřené části sítě Internet nepatří stejně jako dokumenty na obrázku č. 17.

Obrázek 18 – Příklad nevhodné fotografie na webu školy [57]

Takové fotografie mohou být klasifikovány jako nevhodné nejen dle názoru mého, ale také dle webu Internetembezpecne.cz na základě vyjádření k sociální síti TikTok:

„Nejpalčivější je problém sexuálních predátorů, kteří si na TikToku zvykli oslovovat velmi mladé dívky. Ty si vyhlídli na základě jejich obsahu, kde se mohly skrývat lehce sexuální momenty ve formě tance, gymnastiky, jógy či videí v plavkách“ [58].

Školy se chrání tím, že nechávají rodiče podepisovat souhlas, aby jejich dítě mohlo být při školních aktivitách fotografováno. Dříve to byla samozřejmost, a tak se lze často setkat s nepochopením nutnosti podpisu. Je však třeba vzít v potaz rozdíl mezi vytištěnou fotografií, která se nachází v albu v knihovně a digitální fotografií umístěnou na Internet.

Vina tedy neleží jen na školách, ale také na samotných rodičích, kteří si neuvědomují důsledky a sami často vystavují mnohem choulostivější fotografie. Zvlášť oblíbený je pro tyto účely portál Rajce.net. Na toto téma mj. 18. 12. 2019 vyšel na portálu Zive.cz článek s titulkem *„České Rajče je stále plné dětských nahotin. Student pomocí A.I. analyzoval miliony fotek“* [59], z čehož je zřejmé, že i přes snahu různých sdružení je problém aktuální.

„Jedním z příkladů pak může být uživatel vystupující pod pseudonymem pindulinka, která/ý začal/a s číslem 1 (pindulinka1) a momentálně zatím nese číslo 55 (pindulinka55). Pindulinka má 884 oblíbených alb (například Dětičky, Bibione, Bazén, Koupáníčko, Léto…), 587 oblíbenců a také 26 fanoušků“ [60].

Z hlediska legislativy se sice zpravidla nejedná o nic protiprávního, ale rodič by si měl před tím, než takové fotografie zveřejní do světa (či svolí k jejich pořízení a uveřejnění), uvědomit případné následky. Jakýkoli materiál, který o nás něco prozrazuje, může být zneužit a může dítě vystavit riziku (např. stalkingu).

Na školách se dnes sice o těchto rizicích vyučuje, ale samotní rodiče si případná rizika uvědomovat nemusí, proto vzniklo i několik publikací a projektů, které mají za úkol zasvětit do problematiky sociálních sítí právě i rodiče. Příkladem může být projekt Safer Internet, který vytvořilo zájmové sdružení Národní Centrum Bezpečnějšího Internetu (https://www.saferinternet.cz/), nebo Internetem Bezpečně (https://www.internetembezpecne.cz).

3.4.5 Metadata

V souvislosti se soubory na síti Internet je též nutné zmínit metadata. *„Definice metadat je jednoduchá – metadata jsou data o datech. Výstižné, leč poněkud lakonické vymezení. Proto bude zřejmější metadata charakterizovat jako data sdružená s objekty, který zbavují jejich potenciální uživatele nutnosti předběžné znalosti existence či charakteristik těchto objektů“* [61].

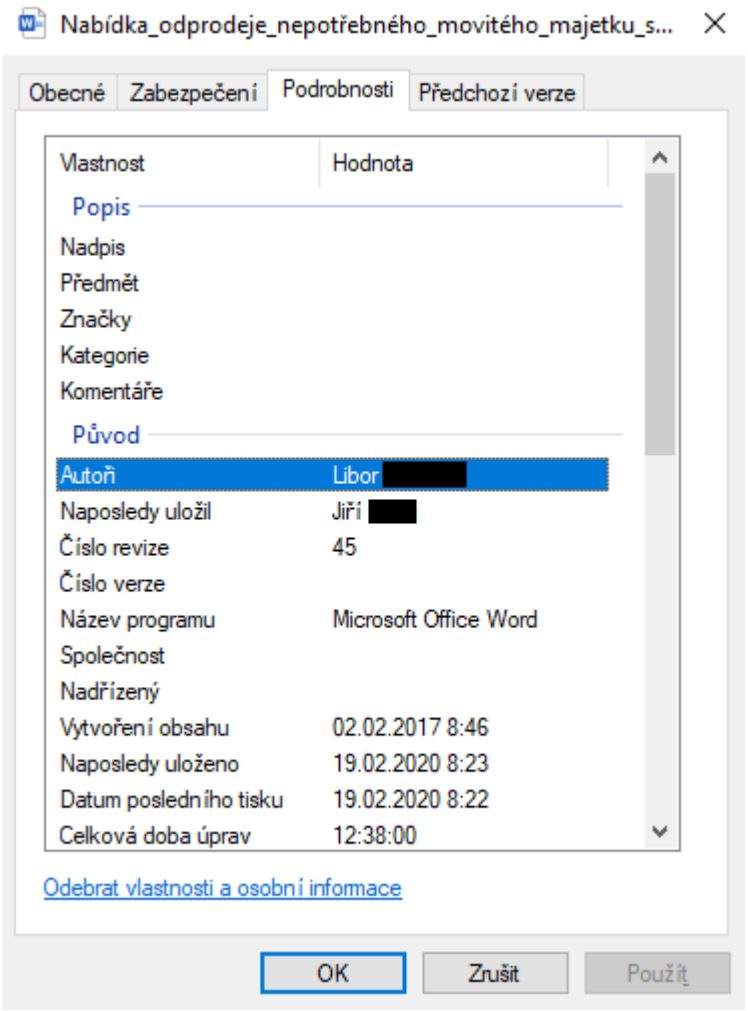
Metadata mohou být přímo součástí souboru, nebo se ukládají samostatně. Využití oddělených indexů metadat je časté např. u fotografování, ale je také využíváno operačními systémy (např. Windows Alternate Data Streams – ADS [62]) nebo systémy Data Loss Prevention (DLP) pro označování dat a bezpečnostní politiky.

Metadata mohou vznikat buď automaticky, což se typicky děje při psaní dokumentů v aplikaci Word a fotografování, nebo jsou do jednotlivých souborů přidávána ručně samotnými uživateli, jako je tomu v případě hudby (album, číslo skladby). Jak již možnost ručního vyplnění napovídá, ne vždy jsou metadata důvěryhodným zdrojem informací a je možné je podvrhnout.

Existuje několik standardů a nikde není jednoznačně určeno, co by metadata měla obsahovat. Nejčastěji se můžeme setkat s údaji:

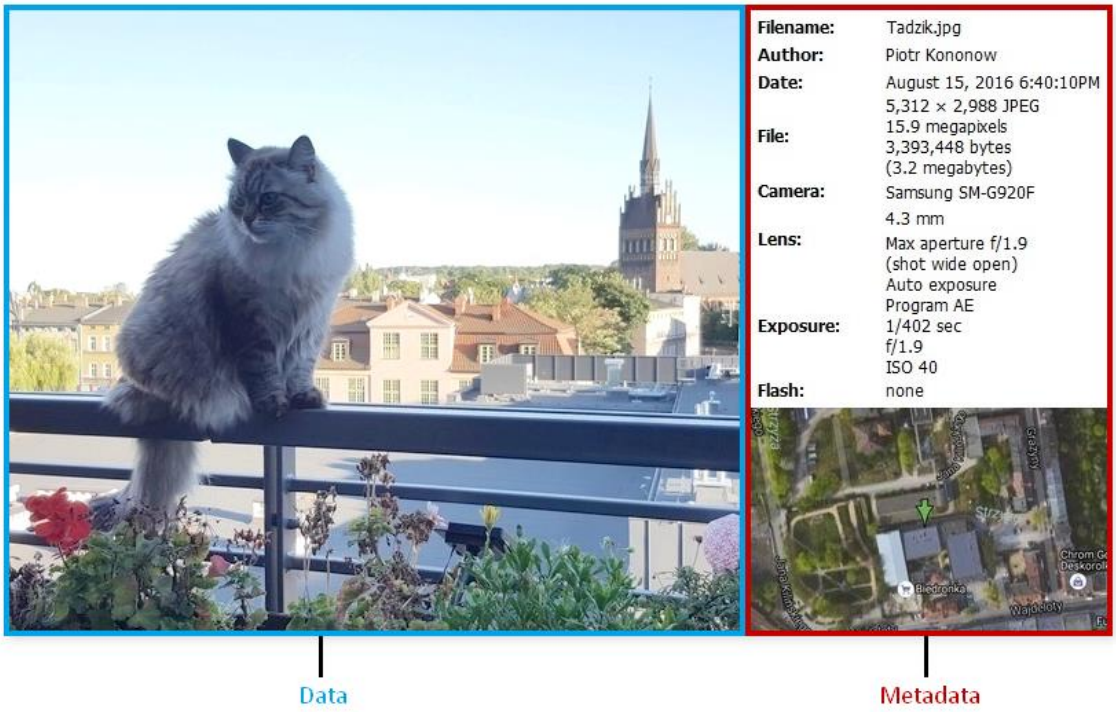
- Titulek, popis, označení.
- Kdo a kdy soubor vytvořil.
- Kdo a kdy soubor naposledy upravil.

Příkladem je náhodný soubor stažený z webu HZS ČR, jehož metadata jsou zachycena na obrázku č. 19.



Obrázek 19 – Soubor ze stránek HZS ČR; snímek autora

Kategorií, která bývá v souvislosti s OSINT často zmiňována, jsou fotografie. Ty v sobě mohou nést podrobné informace o použitém digitálním fotoaparátu (telefonu) a jeho nastavení, tzv. EXIF. Ukázka souboru je zachycena na obrázku č. 20.



Obrázek 20 – Příklad metadat u fotografie [63]

Součástí metadat v EXIF mohou být také geolokační údaje, údaje z EXIF jsou však při jejich nahrání na Internet (resp. sociální sítě) často odstraněny. I přesto, že mohou být metadata fotografií z forenzního hlediska důležité, z hlediska OSINTu je jejich využitelnost značně omezena.

Velmi důležitá metadata v sobě nesou e-mailové zprávy. Každá zpráva je dělena na hlavičku zprávy a obsah. Obsahovou část vyplňuje uživatel, zatímco hlavička je generována cestou od odesílatele k příjemci. Hlavička e-mailové zprávy může vypadat jako např. tak, jako na obrázku č. 21.

```
Return-path: <sender@senderdomain.tld>
Delivery-date: Wed, 13 Apr 2011 00:31:13 +0200
(3)Received: from mailexchanger.recipientdomain.tld([ccc.ccc.ccc.ccc])
by mailserver.recipientdomain.tld running ExIM with esmtp
id xxxxxx-xxxxxx-xxx; Wed, 13 Apr 2011 01:39:23 +0200
(2)Received: from mailserver.senderdomain.tld ([bbb.bbb.bbb.bbb] helo=mailserver.senderdomain.tld)
by mailexchanger.recipientdomain.tld with esmtp id xxxxxx-xxxxxx-xx
for recipient@recipientdomain.tld; Wed, 13 Apr 2011 01:39:23 +0200
(1)Received: from senderhostname [aaa.aaa.aaa.aaa] (helo=[senderhostname])
by mailserver.senderdomain.tld with esmtpa (Exim x.xx)
(envelope-from <sender@senderdomain.tld) id xxxxx-xxxxxx-xxxx
for recipient@recipientdomain.tld; Tue, 12 Apr 2011 20:36:08 -0100
Message-ID: <xxxxxxxx.xxxxxxxx@senderdomain.tld>
Date: Tue, 12 Apr 2011 20:36:01 -0100
X-Mailer: Mail Client
From: Sender Name <sender@senderdomain.tld>
To: Recipient Name <recipient@recipientdomain.tld>
Subject: Message Subject
```

Obrázek 21 – Příklad hlavičky e-mailové zprávy [64]

Do hlavičky se v průběhu cesty přidává kód vždy na začátek, proto je nutné číst hlavičku odspodu. Zeleně označený text zobrazuje trasu od odesílatele, červený značí odesílatele. Jednotlivé položky hlavičky pak mají následující význam:

- Subject: předmět zprávy.
- To: adresát zprávy (adresa).
- From: odesílatel zprávy (adresa).
- X-Mailer: klient, který byl použit k odeslání zprávy.
- Date: datum odeslání zprávy.
- Message-ID: ID zprávy.
- Delivery-date: datum a čas doručení zprávy.
- Return Path: adresa, na kterou se zašle upozornění, pokud e-mail nebude možné doručit.

Na filehostingových serverech se lze setkat nejen s celými e-mailovými soubory, ale i se zálohami celých schránek. Jednotlivé e-maily pak lze tímto způsobem zkoumat.

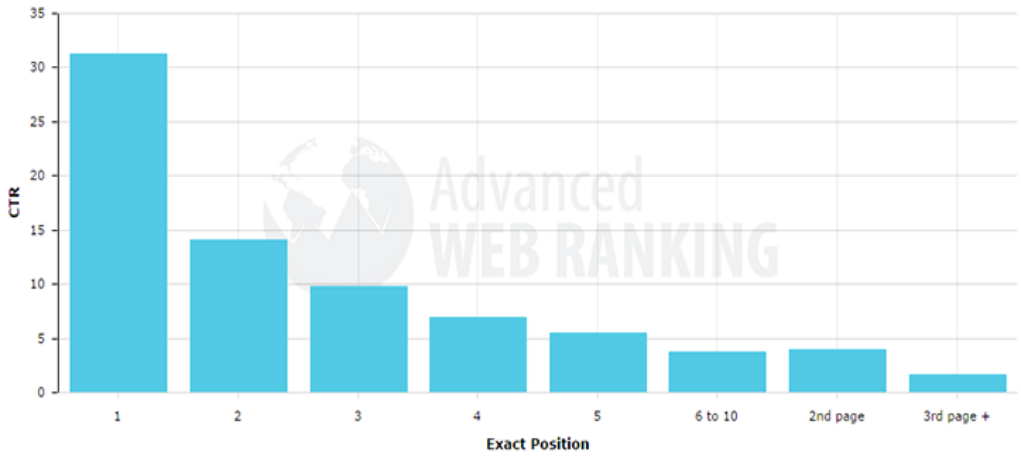
Ačkoli by souborová metadata mohla být skvělým zdrojem údajů, skutečnost bývá taková, že např. u fotografií se s nimi setkáme zpravidla jen na zdrojovém zařízení. V momentě, kdy je fotografie přeposlána prostřednictvím některého instant messengeru nebo sociální sítě, jsou metadata zpravidla odstraněna. Výjimku tvoří metadata dokumentů a fotografií umístěných na stránkách o fotografování. E-mailové hlavičky také pochopitelně v souboru exportovaném z klientů (Microsoft Outlook, Thunderbird) nebo z prohlížeče zůstávají, ačkoli některé služby jako z důvodu ochrany uživatelských údajů hlavičky maskují. V e-mailech odeslaných z webového rozhraní Gmail např. nelze dohledat IP adresu odesílatele.

3.5 Možnosti vytěžování dat

3.5.1 Manuální hledání

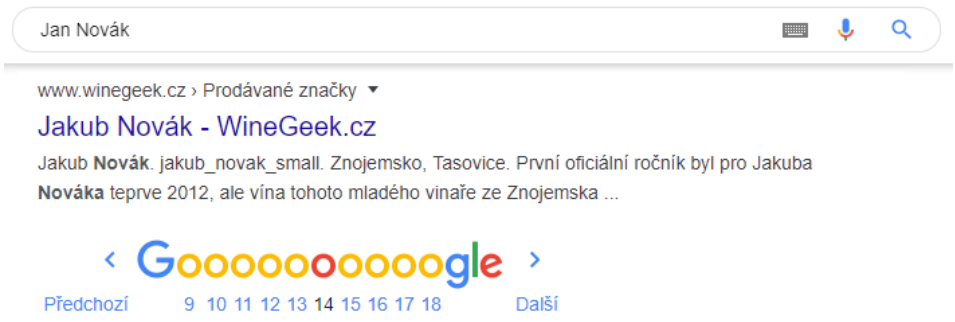
V této podkapitole se vrátím k informačnímu smogu, který byl zmíněn na začátku práce. Při hledání, ať už použijeme jakýkoli vyhledávač, se musíme smířit s tím, že výsledky budou plné irelevantního obsahu. Ten lze do jisté míry odstranit použitím operátorů, o kterých bude v této kapitole zmínka. Mimo tuto skutečnost se musíme smířit také s tím, že přední příčky vyhledávačů jsou lukrativní záležitostí, která garantuje vysokou návštěvnost webů. Tzv. optimalizaci pro vyhledávače (SEO – Search Engine Optimization) se věnují i celé společnosti.

Chování běžných uživatelů, resp. jejich prokliknutí na výsledky vyhledávání při použití vyhledávače Google, vyjadřuje obrázek č. 22.



Obrázek 22 – Počet prokliků výsledků při vyhledávání dle jednotlivých příček. ***CTR*** *– Click-through Ratio (prokliky);* ***Exact Position*** *– pozice ve vyhledávání [65]*

Odhodlaný uživatel (útočník) je schopen manuálně projít vyšší počet odkazů, řádově se však bude jednat maximálně o stovky. Otázkou pak zůstává, k čemu nám je několik miliónů výsledků, které Google najde při hledání řetězce Jan Novák. V diskuzní části bude provedena demonstrace (kapitola 6.7), ve které bude předvedeno, jak je to s výsledky ve skutečnosti. Informace o vysokém počtu výsledků by nicméně měla indikovat, že je naše hledání nedostatečně specifikované. Na 14. stránce hledání (výsledky 141–150) se navíc objevuje první náznak, že výsledky obsahují i něco, co s dotazem zcela nesouvisí, jak je patrné z obrázku č. 23.



Obrázek 23 – Irelevantní výsledek ve vyhledávání prostřednictvím Google [56]; snímek autora

Zde se dostáváme k technice, která se trochu nadneseně nazývá Google hacking. „Hacking“ v tomto případě přitom neznamená nic jiného než koncipování dotazů do vyhledávání tak, aby výsledky byly co nejpřesnější.

3.5.1.1 Google hacks

Dotazy do vyhledávače Google je možné zadávat spolu s operátory, které jsem rozdělil do dvou kategorií v tabulkách č. 4 a 5.

Tabulka 4 – Logické operátory pro vyhledávání v Google

Operátor	Význam	Příklad	Výsledek
AND (+)	každý výraz, který musí být ve výsledku (Google tento operátor používá automaticky)	jablka +hrušky	jablka a hrušky
NOT (-)	každý výrazu, který nesmí být ve výsledku	jablka –hrušky	jablka
OR (|)	alespoň jeden z výrazů musí být ve výsledku	jablka |hrušky	vše, kde jsou jablka nebo hrušky
“	všechny výrazy přesně	"jablka hrušky"	nenajde stránky, kde jsou "hrušky jablka"
*	libovolný výraz	"jablka * hrušky"	najde i "jablka švestky hrušky"
()	shlukování výrazů	"recept na (jablečný | hruškový) koláč"	najde recepty na jablečné a hruškové koláče
#..#	hledání v číselném rozmezí	"ryzlink rýnský 2014..2017"	najde všechna vína z let 2014-2017
$, €	hledání v dané měně	prague hotel $20	najde hotely v Praze s cenou $20 na noc
to, in	slouží pro převod jednotek	$20 in czk	vypíše hodnotu $20 v Kč

Tabulka 5 – Pokročilé operátory pro vyhledávání v Google

Operátor	Význam	Příklad	Výsledek
define:	vysvětlí a případně přeloží výraz do lokálního jazyka (není k dispozici v češtině)	define:apple	překlad na jablko a vysvětlení, že jde o ovoce
cache:	najde poslední uloženou verzi webu (např. při výpadku stránky nebo jejím smazání)	cache:seznam.cz	zobrazí několik minut až hodin starou verzi webu
filetype:/ext:	hledá typ souboru	jablka filetype:pdf	najde dokumenty pdf, ve kterých se vyskytují jablka
site:	hledá na konkrétní stránce	site:fbmi.cvut.cz filetype:pdf	najde pdf dokumenty na webu školy
related:	najde příbuzné weby	related:cvut.cz	najde ostatní univerzity v ČR
intitle:	hledá výraz v titulku stránky	intitle:apple	najde weby, které mají výraz apple v titulku
allintitle:	všechny výrazy musí být v tituku	allintitle:apple dictionary	najde weby, které mají oba výrazy v titulku
inurl:/allinurl:	výraz/y musí být v url	site:seznam.cz inurl:index.php	najde soubory index.php na doméně seznam.cz
intext:/allintext:	výraz musí být v textu	intext:jablka	nevyhledá výraz v url, titulku apod.
inanchor:/allinanchor:	výraz/y je v textu odkazů na stránce	inanchor:jablka site:seznam.cz	najde odkazy na stránkách seznam.cz, ve kterých je text jablka
AROUND(X)	hledá výrazy vzdálené X slov od sebe	jablka AROUND(2) švestky	výsledek může být "jablka a sušené švestky"

Pro vyhotovení tabulky byly použity vlastní zdroje a článek na webu Ahrefs.com [66].

Různými kombinacemi těchto dotazů a klíčových slov, lze prostřednictvím vyhledávače nalézt neuvěřitelné věci. Již několik let se tak uživatelé mohou bavit sledováním nezabezpečených IP kamer. Nejedná se o běžné webové kamery, které má většina přenosných počítačů, ale o domácí či venkovní kamerové systémy, které nejsou chráněny heslem.

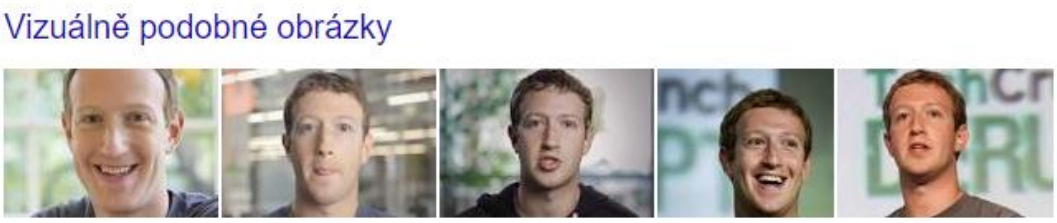
Obdobným způsobem je možné hledat nezabezpečené soubory s hesly a jiné informace, které by neměly být vidět. Sbírka řetězců, které cílí na konkrétní chyby, je k dispozici na webu https://www.exploit-db.com/google-hacking-database a je neustále aktualizována.

3.5.1.2 Hledání fotografií

Google dále nabízí možnost vyhledávání dle fotografií, jak je ukázáno na obrázku č. 24.

Obrázek 24 – Hledání fotografie v Google [56]; snímek autora

Nejenže vyhledávač často pozná, o jakou známou osobnost se jedná (výše je zakladatel sítě Facebook), ale také najde, na jakých stránkách byla fotografie dále použita, a to i v jiných velikostech. Umí také hledat vizuálně podobné obrázky, jak je vidět na obrázku č. 25.



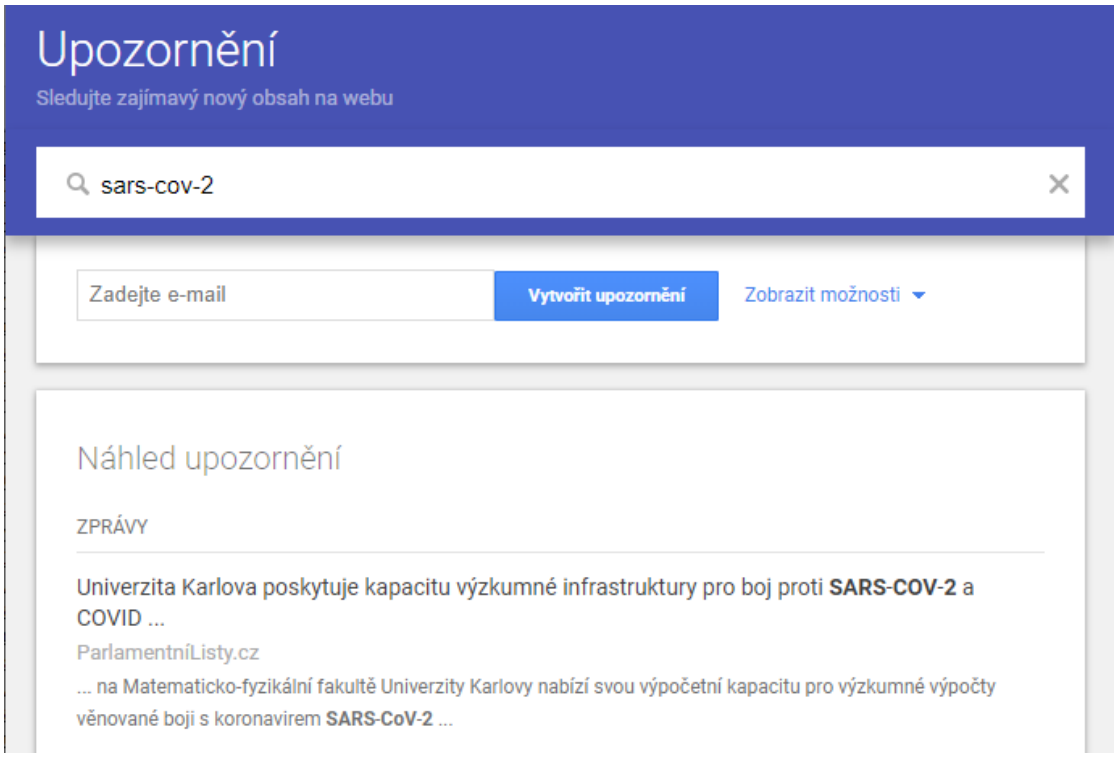
Obrázek 25 – Výsledek zobrazený k hledání dle fotografie dle fotografie známé osoby [56]; snímek autora

Google analyzuje také metadata fotografií, je proto otázkou, do jaké míry se jedná o podobu tváře a do jaké míry jde o „zkušenost" Google s touto osobou.

Podobnou službu poskytuje např. také web TinEye.com, ten však v tomto případě dohledal jen naprosto shodný obrázek bez variant v jiném rozlišení.

3.5.1.3 Google Alerts

Pokud chceme sledovat nově indexované výsledky, je za tímto účelem možné použít službu Google Alerts, která nám do e-mailové schránky bude zasílat notifikace. Nastavení služby zachycuje obrázek č. 26.



Obrázek 26 – Příklad nastavení služby Google Alerts [56]; snímek autora

3.5.2 Nástroje pro automatizovaný sběr

Pro usnadnění extrakce dat z prostředí sítě Internet je možné použít již hotové nástroje, nebo si nástroj může vyvinout každý sám dle konkrétních potřeb. V této kapitole bude pojednáno o nástrojích pro automatizovaný sběr, které lze dělit do dvou kategorií na tzv. crawlery a scrapery.

3.5.2.1 Crawler

Pod pojmem crawler si lze představit program, který navštíví web a projde celým jeho obsahem dle nastavených kritérií. Parametrem tak může být typ odkazu (zda se jedná o odkaz na další web, obrázek, video atd.), hloubka odkazu anebo určení, zda má prověřovat též externí zdroje. Účelem tohoto nástroje je stažení všech dostupných informací, vytvoření kopie webu a případné zjištění vazeb na další weby.

Příkladem pro automatizované stahování webů je program WinHTTrack, který bude použit pro demonstraci v diskuzní části. Jako crawler lze po nastavení vhodných parametrů využít také program Wget, který je součástí většiny známých linuxových operačních systémů.

3.5.2.2 Scraper

Scraper představuje nástroj, který pracuje s kódem webu a je schopen dle zadaných kritérií provést extrakci jednotlivých elementů (např. všechny nadpisy, bloky textu, ale i konkrétní řetězce). V kombinaci s crawlerem je tak možné vyhledat i nezaindexovaná data.

Scraper můžeme nastavit například tak, aby navštívil několik webů a z nich stáhl všechny e-mailové adresy. Základním nástrojem je samotný prohlížeč s konzolí. Pokud budeme chtít provést sběr e-mailových schránek, můžeme využít již zmíněný nástroj Hunter.io, do kterého budeme postupně vkládat

domény a data následně stahovat bez další interakce. Pokud nás však zajímají data o e-mailových schránkách např. z domény Seznam.cz, nebude tento nástroj možné použít. Pro takovou extrakci bychom mohli využít skriptovací jazyky, jako je například Python a spolu s ním Selenium WebDriver.

Selenium WebDriver je nástroj pro automatizaci chování prohlížeče, který je schopen ovládat i jednotlivé elementy webu, podobně jako uživatel s konzolí. Tímto způsobem lze definovat několik webů, které mají být prohlédnuty a z nich následně získat potřebné informace, tedy i již zmíněné e-mailové adresy.

Určit, která data mají být získána lze i pomocí regulárních výrazů (RegEx). Díky nim je možné hledat řetězce v určitých obvyklých tvarech a také upravovat velké množství dat. Regulární výrazy je také možné užít i v dalším zpracování dat např. v programech PSPad nebo Notepad++. Díky tomu je možné v krátkém čase formátovat text do podoby, která bude čitelná např. v MS Excel.

4 METODIKA

Výzkum bude proveden na datech k celkem 100 profilům, ke kterým budou zjišťovány údaje použitím zdrojů a metod uvedených v předchozích kapitolách. Vstupní data k subjektům budou sbírána z různých inzertních a diskuzních portálů. Způsob získání dat bude proveden buď přímo z daného webu (otevřením inzerátu, diskuzního vlákna), nebo prostřednictvím vyhledávače Google, pokud ve výsledcích budou indexována data z těchto webů.

4.1 Volba webů pro zdrojová data

Jako předmět zkoumání byly zvoleny weby Bezrealitky.cz, Sauto.cz, Hyperinzerce.cz, Sbazar.cz, Emimino.cz, Vinted.cz, Mimibazar.cz a Bazos.cz.

Weby byly rozděleny dle údajů, které se zobrazily rozkliknutím prvních 5 náhodných příspěvků či inzerátu, které byly zobrazeny po návštěvě webu. Weby Bezrealitky.cz, Sauto.cz, Hyperinzerce.cz a Bazos.cz budou primárně vytěžovány jako zdroj telefonních čísel. Weby Emimino.cz, Vinted.cz, Sbazar.cz a Mimibazar.cz budou primárně zdrojem e-mailových adres.

4.2 Prvotní vytěžování

Hledání bude prováděno dvěma způsoby dle subjektivního názoru na efektivtu. Pokud to bude možné, budou rozklikávány jednotlivé inzeráty či příspěvky zobrazené při návštěvě webu. Pokud web bude vyžadovat vstup do některé z podkategorií (diskuzní téma, oblast inzerce), budou jednotlivým kategoriím přidělena čísla od 1 do X. Následně bude použita funkce v příkazovém řádku operačního systému Windows 10 pro generování náhodných čísel, kde X znamená počet kategorií `set /a _rand=(%random%*X/32768)+1`.

Výsledné číslo bude indikovat kategorii, která bude pro vytěžení zvolena. V případě 5 po sobě jdoucích příspěvků nebo inzerátů, které neobsahují požadované informace, bude učiněn krok zpět a bude zvolena nová kategorie. Tento postup bude opakován do doby, dokud nebude získáno 10 vstupních údajů.

Druhým způsobem získávání údajů bude vkládání dotazů do vyhledávače Google konstruovaných za účelem vyhledání konkrétního identifikátoru (tedy např. přidáním slova „e-mail“, „kontakt“ nebo „volejte“). Z výsledků pak budou postupně otevírány jednotlivé výsledky, které v indexu obsahují hledaný dotaz (telefonní číslo, e-mail), a to do doby, než bude získáno 10 vstupních údajů z každého webu.

Zbylý počet 20 vzorků bude doplněn z vybrané skupiny dle potřeby a to tak, aby na vstupu bylo minimálně 50 telefonních čísel a 50 e-mailových adres a zároveň tak, aby se poměr e-mailových adres a telefonních čísel přiblížil poměru 1 ku 1.

Poté, co bude navštíven příspěvek nebo inzerát, budou z něj vypsány vstupní údaje, které budou sloužit pro další hledání.

4.3 Následné vytěžení

Po zapsání vstupních údajů bude pokračováno ve sběru informací ke konkrétnímu uživateli. Tento sběr bude prováděn ručně bez využití automatizovaných nástrojů (nebudou užity metody popsané v kapitole 3.5.2). Data nebudou sbírána z webů podmíněných registrací, ani z placených služeb. Tím bude simulováno prostředí běžného uživatele a budou tak vyhledány jen výsledky, které jsou viditelné kýmkoli.

Ke každému druhu vstupního údaje bude provedena shodná série dotazů do vyhledávače Google. Při dohledávání budou dle potřeby používány operátory

uvedené v kapitole 3.5.1.1 a to zejména ty, které specifikující vyhledávání na konkrétních doménách nebo hledání celého výrazu (site:, ""). Při nalezení informací, které se budou jevit jako relevantní, bude navštíven konkrétní web a budou z něj vypsány údaje, které souvisí s daným uživatelem.

Budou prohlíženy jen webové stránky, které již v hledání signalizují přítomnost hledaného obsahu a nemusí tedy být prověřovány veškeré výsledky. To platí zejména v případech diskuzních fór a podobně.

Dle potřeby bude v průběhu výzkumu používána služba Archive.org a Google Cache (kapitola 3.4.3), a to zejména v případě potřeby zobrazení nedávno smazaných dat, která se stále nachází v indexu Google nebo při nalezení webové stránky se vztahem k uživateli, která neposkytuje kompletní informaci (např. firemní stránky, ze kterých byla odstraněna fotografie osoby nebo její identifikátor).

Dle jména bude vyhledán profil na sociální síti Facebook. V případě méně rozšířených jmen (do 5 výsledků při hledání ve tvaru „jméno příjmení“ na síti Facebook.com) bude předpokládáno, že uživatelem je hledaná osoba, pokud zároveň korespondují další údaje jako zdánlivý věk, lokalita. V případě obvyklých jmen bude proveden pokus o přiřazení dle fotografie, pokud byla zjištěna. Obsah jednotlivých profilů nebude blíže zkoumán, nebude prověřováno okolí uživatele.

Pokud bude nalezena fotografie uživatele, bude pomocí Google vyhledán její výskyt na dalších stránkách (metoda popsána v kapitole 3.5.1.2). Dále bude zjišťováno bydliště a adresa pracoviště nebo sídlo firmy (bude porovnáno s katastrem nemovitostí).

Vzhledem k časové náročnosti profilace a počtu uživatelů, které bude výzkum zahrnovat budou použity pouze metody, které mají simulovat vyhledávání potenciálních obětí. Metody zaměřené na cílení jednotlivých uživatelů (např.

zjišťování okolí subjektu, viz kapitola 3.4.1) by obnášely mnohem větší časovou dotaci i prostor. Výsledná data by byla několikanásobně větší a jejich analýza by byla výrazně složitější.

4.4 Způsob zadávání dotazů

Hledání uživatelů bude primárně probíhat dle vstupních identifikátorů, kterými je telefonní číslo, e-mailová adresa, jméno a případně též identifikátory firmy (IČO, DIČ, sídlo).

Telefonní číslo bude zadáváno ve tvarech „123456789“ a „123 456 789“. Použití uvození umožní vyhledávat konkrétní tvar, ve kterém je mezera platným znakem.

E-maily budou vyhledávány ve tvaru „uživatel@doména.cz“. V případě negativních výsledků bude použito též samotné uživatelské jméno. Pokud bude zjevné, že se jedná o totožnou osobu, budou data použita stejně, jako by byl zjištěn e-mail. V případě pochybností (obvyklé uživatelské jméno) budou zjištěné údaje ignorovány. Pro e-mailové adresy, bude dále proveden dotaz na stránku haveibeenpwned.com. Zde bude zjištěno, jestli byl tento účet kompromitován v některém z úniků.

Jména budou hledána ve tvaru „jméno příjmení“ a „příjmení jméno“. V případě, kdy bude zjištěna také adresa bydliště či zaměstnání, bude k tomuto vyhledávání přidáno „DIČ“. Pokud bude adresa korespondovat, bude předpokládáno, že se jedná o totožnou osobu a bude k ní tento identifikátor přiřazen.

4.5 Vyhodnocení

Při vyhodnocení bude rozlišován projev vůle osoby sdělit kontaktní informace. Tato vůle bude vyjádřena jako množství vstupních údajů, které budou

zjištěny. Údaje zjištěné dalším hledáním, které se podaří k subjektu přiřadit, budou klasifikovány jako údaje, které subjekt sděluje nevědomě. To platí pro údaje vložené třetí stranou i pro údaje vložené uživatelem na jiný portál.

Kategorizace vychází z myšlenky, že v případě, kdy např. uživatel na jednom portálu sdílí pouze e-mailovou adresu a již neuvádí své telefonní číslo, není pro něho žádoucí, aby jeho telefonní číslo bylo veřejně publikováno.

Projev vůle nebude součástí analýzy MCDA. Účelem analýzy je demonstrovat možnost útočníka a nezáleží tedy na tom, zda oběť uveřejnila informace sama, nebo tak učinila třetí strana.

Získaná data budou zavedena do tabulky, ve které bude vyhodnocena míra rizika ve vztahu k hrozbám uvedeným v kapitole 3.3. Výsledky budou kategorizovány do skupin:

Bez rizika

- Byl zjištěn jen e-mail nebo telefonní číslo, další informace se nepodařilo zjistit.
- Na uživatele je obtížné cílit plošné útoky, připadá v úvahu jen náhodný spam.

Nízké riziko

- Bylo zjištěno jméno osoby nebo druhý identifikátor (jméno/e-mail).
- Tato kategorie značí, že uživatel může být obtěžován spamovými a phishingovými e-maily, které mohou být i částečně personalizovány (např. oslovení křestním jménem na začátku e-mailu), nebo obtěžován náhodnými telefonickými prodejci apod.

Střední riziko

- Bylo zjištěno jméno osoby nebo druhý identifikátor, a to buď spolu s profilem na sociální síti, fotografií, bydlištěm nebo pracovištěm.

- Uživatel je vystaven vyššímu riziku v prostředí sítě, jeho identifikátor je provázán s jeho soukromím ve větší míře, než je nutné, např. může být snazším cílem pro stalkery nebo se stát obětí sofistikovanějších útoků pomocí sociálního inženýrství.

Vysoké riziko

- Zjištěný e-mail byl součástí některého z úniků (riziko kompromitovaného hesla), nebo bylo zjištěno datum narození či rodné číslo.
- Uživatel by měl zkontrolovat data, která jsou o něm zjistitelná, hrozí prolomení zabezpečení jeho účtů nebo jiné zneužití osobních údajů.

4.6 MCDA analýza

Závěrečná MCDA analýza bude vypracována z pohledu útočníka a bude simulovat možnou volbu útoku, provedenou na základě zjištěných informací.

Pro hrozby bude individuálně stanovena důležitost relevantních údajů. Důležitost bude vyjadřovat potřebu znát konkrétní údaj pro možné provedení každého z útoků a bude se tedy pro každý útok lišit. Důležitost jednotlivých hodnot bude stanovena na stupnici od 0 do 5, kde 0 bude znamenat nevýznamný údaj a 5 podstatný údaj. Hodnoty budou sečteny, přičemž pro každý typ útoku je stanoven jiný maximální počet dosažených bodů. Následně bude u každého profilu procentuálně vyjádřeno množství získaných bodů pro každý z útoků.

Dalším krokem bude individuální stanovení hranice proveditelnosti pro každou z hrozeb, podle které bude získaný počet bodu klasifikován do jedné z kategorií – snadno proveditelná, proveditelná, obtížně proveditelná a neproveditelná, čímž bude provedena normalizace hodnot, aby bylo možné výsledky pro jednotlivé útoky navzájem porovnat.

Posledním krokem analýzy bude úprava hodnot koeficientem priority, který bude vyjadřovat preferenci útočníka vykonat složitější typ útoku, pokud k němu bude mít dostatek údajů.

Výsledkem analýzy bude volba jedné varianty útoku a nebude odrážet vůli oběti uveřejnit informace.

5 VÝSLEDKY

5.1 Shrnutí výzkumu

Při výzkumu byly použitím metod OSINT shromážděny údaje ke sto uživatelům sítě Internet. Vstupními údaji bylo v 47 případech telefonní číslo, v 36 e-mailová adresa a v 17 případech byly na vstupu oba údaje.

Dle metodiky práce byly použity portály, z nichž byla získána vstupní data a následným vytěžením byly zjištěny údaje,

Zjištěné údaje shrnuje tabulka č. 6.

Tabulka 6 – Shrnutí zjištěných údajů

	e-mail	telefonní číslo	únik hesla	jméno	datum narození	adresa	fotografie	sociální sítě
vstupní údaj	53	64	0	25	2	4	0	0
dohledáno	17	7	41	25	14	28	24	20
celkem	70	71	41	50	16	32	24	20

Úspěšnost hledání dalších identifikátorů osoby byla dle vstupních identifikátorů následující:

- Z 36 e-mailových adres bylo zjištěno 7 telefonních čísel a 9 jmen.
- Ze 47 telefonních čísel bylo zjištěno 17 e-mailových adres a 13 jmen.
- V 17 případech byly známé od počátku oba identifikátory a jméno bylo zjištěno u 3 osob.

Z toho vyplývá, že pokud byla vstupním údajem e-mailová adresa, bylo telefonní číslo zjištěno pouze v 19,4 % případů, pokud bylo na vstupu telefonní číslo, byl e-mail zjištěn v 36,2 % případů.

Ačkoli je e-mailová adresa určena pro komunikaci v síti Internet, bylo snazší dohledat spojitost s určitou osobou prostřednictvím telefonního čísla, tedy identifikátoru fungujícím v „reálném“ světě.

K 50 celkově zkoumaným jménům bylo 8krát zjištěno datum narození a 6krát DIČ odpovídající rodnému číslu (2krát uvedeno již na vstupu).

Kategorizaci výsledků dle rizik uvedených v metodice výzkumu zachycuje tabulka č. 7.

Tabulka 7 – Kategorizace rizikovosti informací

Kategorizace zjištěných informací								
bez rizika	nízké riziko		střední riziko			vysoké riziko		
bez dalších informací	jméno	druhý identifikátor	profil, fotografie	bydliště, pracoviště	obě kritéria	RČ/datum narození	únik	obě kritéria
25	20	41	25	31	15	16	41	10
Kategorizace jednotlivých zkoumaných profilů								
bez rizika	nízké riziko		střední riziko			vysoké riziko		
25	13		15			47		

Překvapivým výsledkem byl počet e-mailových adres, ke kterým je vázán některý z úniků informací. Úniky uživatelských údajů byly zjištěny pro 41 e-mailových adres z celkových 70 shromážděných, což odpovídá 58,6 %. Díky tomu se většina profilů dostala do kategorie vysokého rizika.

Jako nejkritičtější informace, které bylo výzkumem možné zjistit, byly stanoveny data narození a rodná čísla osob, spolu s úniky hesel uživatelských účtů. Celkem o 45 uživatelích (vyjma 2, kteří DIČ sdělují sami) sděluje citlivé informace, třetí strana. V případech DIČ lze tvrdit, že se jedná spíše o chybu v nastavení systému, který primárně neměl dopustit, aby rodné číslo bylo zároveň veřejným identifikátorem. Dle zjištěných informací by tyto identifikátory měly být v blízké budoucnosti nahrazeny [67].

Profily na sociálních sítích byly nalezeny celkem u 20 uživatelů. Na tento výsledek měly vliv 2 jevy.

Ačkoli jsem původně zamýšlel diplomovou práci primárně zaměřit na oblast sociálních sítí (zejména Facebook), došlo v průběhu schvalování zadání k vydání Facebook Graph API v3.3, která řadu funkcí, jež jsem chtěl pro výzkum využít, zneplatnila. Toto mělo jednoznačně velký dopad na úspěšnost při hledání uživatelů. Na druhou stranu se podařilo dohledat celkem 12 profilů díky skutečnosti, že již samotné jméno (případně ve spojení s bydlištěm či zdánlivým věkem) vedlo k individuální identifikaci osoby na sociální síti.

5.2 MCDA analýza výsledků

Pro provedení MCDA analýzy byly zvoleny vybrané hrozby představené v kapitole 3.3. Informacím, které byly předmětem výzkumu byla pro jednotlivé hrozby přiřazena důležitost na stupnici od 0 do 5, kdy 0 nepředstavuje žádné riziko a informace s hodnotou 5 je pro hrozbu stěžejní informací.

Tato analýza má simulovat volbu potenciálního útočníka založenou na informacích, které se podařilo metodami OSINT zjistit a měla by interpretovat šanci útočníka na úspěšné provedení, pokud by potřebné informace získal. Jejím výstupem bude volba útočníka uchýlit se k vybranému druhu útoku.

Nejprve byly údaje obodovány dle důležitosti pro jednotlivé typy útoků (0 = nepodstatná informace, 5 = významná informace). Bodování bylo stanoveno na základě užitné hodnoty informací pro spáchání vybraných trestných činů. Toto bodování bylo zaneseno do tabulky č. 8.

Tabulka 8 – Klasifikace důležitosti informací z pohledu pachatele

	E-mail	Tel.č.	Jméno	RČ	Únik hesla	Firma	Bydliště	Sociální síť	Fotografie
Spam, šíření virů	5	2	1	0	3	2	0	2	0
Scam (podvody)	4	2	3	4	4	3	1	3	1
Phishing	5	3	3	2	5	3	1	4	0
Pronásledování	4	4	4	3	4	2	4	5	4
Spear phishing	5	4	5	5	5	4	3	4	1

Údaje, které byly v průběhu výzkumu shromážděny, pak byly pro každou z hrozeb ohodnoceny dle této tabulky. U každé hrozby byl porovnán získaný počet bodů s maximálním možným počtem a byl procentuálně vyjádřen.

Dle tabulky č. 9 byla vyhodnocena proveditelnost jednotlivých útoků pro každý profil. Proveditelnost byla vyjádřena číselně, kdy 0 = nemožné a 1 = snadné provedení.

Tabulka 9 – Stanovení proveditelnosti hrozeb

Podmínky proveditelnosti	Spam	Scam	Phishing	Stalking	Spear phishing	Bodové ohodnocení
Lze provést snadno	>50 %	>60 %	>70 %	>70 %	>80 %	1
Lze provést	30–50 %	40–60 %	50–70 %	50–70 %	70–80 %	0,75
Lze provést s obtížemi	<30 %	<40 %	<50 %	<50 %	<70 %	0,5
Nelze provést	0–20 %	0–20 %	0–30 %	0–30 %	0–50 %	0

Pro jednotlivé skutky byly následně stanoveny koeficienty priority jednotlivých hrozeb. Hodnota 1 znamená maximální zájem útočníka provést daný typ útoku a hodnota 0,5 vyjadřuje zájem minimální. Tyto hodnoty jsou vyjádřeny v tabulce č. 10.

Tabulka 10 – Stanovení priority hrozeb

Priority	Spam	Scam	Phishing	Stalking	Spear phishing
Koeficient	0,5	0,6	0,75	0,8	1

Pro shromážděná data byly s využitím uvedených tabulek provedeny výpočty, na jejichž základě byla sestavena výsledná tabulka č.11, která značí možnost a zájem pachatele uchýlit se k danému jednání.

Tabulka 11 – Výsledek MCDA analýzy pro zkoumané profily

Bez rizika	Spam	Scam	Phishing	Stalking	Spear phishing
22	26	8	12	17	15

V porovnání s předchozí tabulkou rizik, která kategorizovala jednotlivá zjištěná data dle jejich kritičnosti, hodnotí tato analýza veškerá získaná data jako celek. Pokud tedy byl uživatel zařazen do kategorie hrozby spear phishingového útoku, znamená to, že množství informací, které se o něm podařilo z otevřených zdrojů zjistit, je skutečně kritické. Na rozdíl od předchozího hodnocení nepostačuje, že se jeho heslo objevilo v některém z úniků, ačkoli samozřejmě taková informace sama o sobě závažná je.

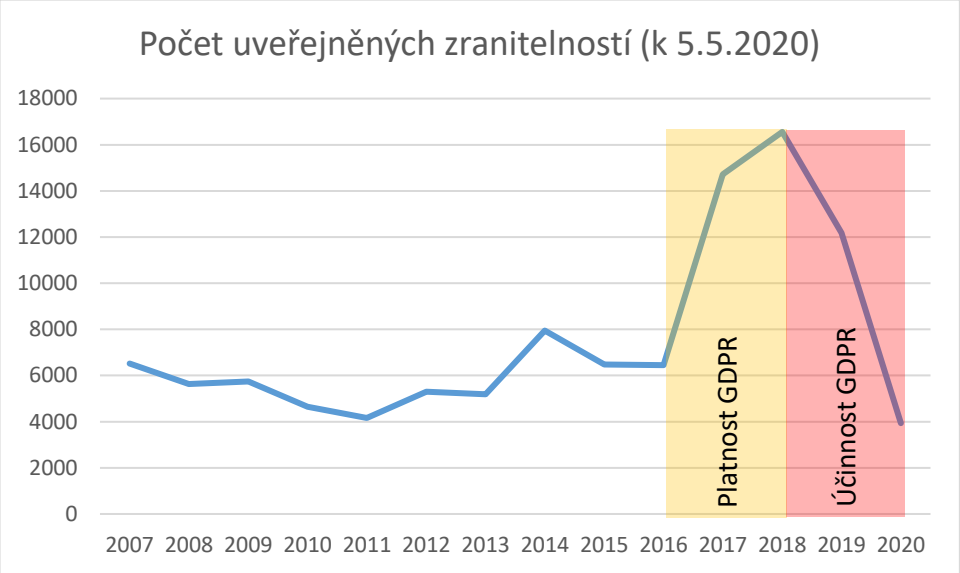
Uživatelé zařazení do této kategorie o sobě na síti Internet doslova poskytují kompletní obraz. Zpravidla je veřejně známo jejich jméno, telefonní číslo, e-mailová adresa, bydliště, často byl jejich účet kompromitován, je známo též jejich datum narození a případně i profil na sociální síti.

Přestože tento výzkum nebyl koncipován tak, aby zkoumal okolí uživatele, bylo během relativně krátké doby u 15 uživatelů zjištěno takové množství informací, že by na jejich základě mohl útočník skutečně uživatele ohrozit, pokud by k tomu měl motiv.

5.3 Vyhodnocení hypotéz

5.3.1 Hypotéza 1: Neexistuje korelace mezi počtem úniků uživatelských dat a směrnicí GDPR.

Tato hypotéza vychází z myšlenky, že počet informačních systémů ve světě neustále roste. Nařízení GDPR zapříčinilo, že je kybernetické bezpečnosti věnována větší pozornost. To má za následek také zvýšenou aktivitu v oblasti penetračního testování a tím pádem nárůst počtu nalezených zranitelností v systémech. Pro tento účel byla z webu cvedetails.com stažena evidence nalezených zranitelností, jejichž počet byl zanesen do grafu na obrázku č. 27.



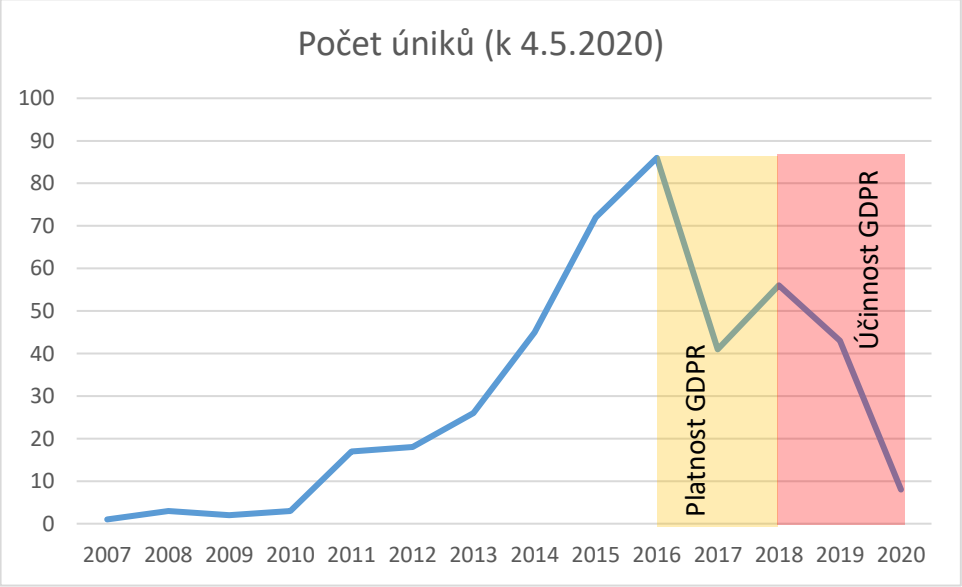
Obrázek 27 – Počet zjištěných zranitelností v jednotlivých letech, zdrojem dat je web cvedetails.com

Pokud budeme předpokládat, že počet uveřejněných zranitelností souvisí s počtem incidentů, lze při pozitivním pohledu na věc tvrdit, že uveřejňování chyb zmenšuje prostor pro jejich exploitaci. Při negativním pohledu lze soudit, že zranitelnosti mohou být do doby jejich záplatování zneužity útočníky. Před uveřejněním zranitelnosti bývá zpravidla období, kdy mají správci jednotlivých systému čas provést potřebná opatření. Pokud však opatření

zanedbají, je defacto formou evidence chyb útočníkům předáván postup, jakým lze systém napadnout.

Ten, kdo chybu přímo zjistí, má na výběr mezi zneužitím zranitelnosti, nebo jejím ohlášením. K ohlášení pak bývá finančně motivován ze strany společnosti, v jejímž produktu chybu objevil, prostřednictvím tzv. bug bounty programů (programy pro vyplácení odměn za oznámené chyby). Zvýšený počet nalezených chyb v posledních letech pravděpodobně odráží také tuto skutečnost.

Z webu haveibeenpwned.com, který byl použit pro zjištění možné kompromitace uživatelských účtů, byla extrahována data o počtu úniků v jednotlivých letech. Ze seznamu úniků byly odfiltrovány kompilace, které pouze duplikují již uveřejněná data. Jednalo se celkem o 19 záznamů z celkových 440. Výsledný graf je prezentován na obrázku č. 28.



Obrázek 28– Počet uveřejněných úniků v jednotlivých letech, zdrojem dat je web haveibeenpwned.com

Pokud pominu, že některé úniky stále nemusely být uveřejněny, je z grafu patrná dlouhodobě stoupající tendence. V roce 2016, kdy vešlo v platnost

nařízení GDPR, nastal zlom a nelze tak vyloučit, že nařízení může mít souvislost se sníženým počtu úniků.

Oba grafy tedy podporují teorii, že nařízení GDPR mohlo mít vliv na snížení počtu úniků a **hypotéza se tedy nepotvrdila**.

5.3.2 Hypotéza 2: Občané ČR se v souvislosti s ochranou osobních údajů v síti Internet chovají zodpovědně.

Údaje, které byly pro výzkum zjištěny, byly ve většině případů vloženy do textu inzerátu samotným uživatelem, jednotlivé portály by je samy o sobě neuveřejňovaly. Inzerenti takto obchází nutnost kontaktu prostřednictvím služby. Přímý kontakt se pravděpodobně jeví jako uživatelsky přívětivější. Pro jednotlivé portály byl proveden stručný rozbor:

- **Bezrealitky.cz**

  Uživatelská data šla dohledat jen tehdy, pokud uživatel obešel standardní cestu a vepsal své údaje přímo do textu inzerátu, za jiných okolností by k nim nebyl přístup.

- **Sauto.cz**

  Zobrazuje e-mail, pokud byl vložen, telefonní číslo uživatele skrývá a je zobrazeno až po kliknutí. Telefonní číslo je však vidět i bez kliknutí ve zdrojovém kódu webu a je indexováno vyhledávači.

- **Hyperinzerce.cz**

  Bez dalších zabezpečujících prvků, nicméně není nutné vyplňovat údaje, které budou viditelné. Je tedy na vůli inzerujících, zda tyto uveřejní.

- **Sbazar.cz**

  Stejně jako Sauto.cz implementuje skrývání telefonního čísla nedostatečně. E-mail lze ve většině případů odvodit z URL (sbazar.cz/email).

- **Emimino.cz**

  Identifikátory z tohoto webu byly sbírány pouze využitím vyhledávače Google, získané údaje byly z diskuze (uveřejněné uživateli) a nebyly v profilu.

- **Vinted.cz**

  Portál sám o sobě neuveřejňuje žádné informace, data byla sbírána přímo z inzerátů a popř. též dle uživatelských jmen, která byla ve tvaru jmeno@e-mail.cz.

- **Mimibazar.cz**

  Portál také neuveřejňuje žádné osobní údaje. Při zkoumání však byla největší úspěšnost právě zde, neboť uživatelé ve svých profilech uvádí nad rámec běžných identifikátorů také informace o bydlišti, bankovních účtech a DIČ (RČ).

- **Bazos.cz**

  Telefonní číslo inzerujícího je skryto a může ho zobrazit pouze ověřená osoba po zadání svého telefonního čísla. Ve zdrojovém kódu není vidět, není tudíž viditelné ani vyhledávači. Získané údaje tedy byly vloženy uživatelem nestandardně.

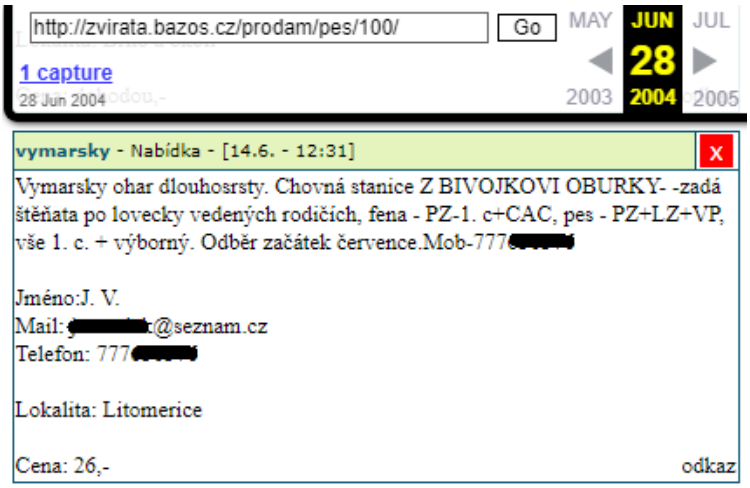
V případě Bazos.cz byla nabídnuta možnost zobrazení kontaktu na osobu až po uvedení telefonního čísla, na které je následně zaslán ověřovací kód. Nad rámec výzkumu jsem se pokusil o vložení čísla a použil kontakt, na který je

možné nechat veřejně zasílat zprávy (web https://receive-smss.com/). Toto číslo bylo odmítnuto, což je pozitivní zjištění.

Vzhledem k tomu, že na žádném z portálu není viditelný úmysl údaje o uživatelích publikovat, lze soudit, že je tak činěno cíleně s vědomím potřeby chránit uživatele. Pochopitelně by tak portál mohl činit také proto, aby měl lepší přehled o počtu odpovědí, které standardně musí být zasílány prostřednictvím dané služby. Každopádně je efekt z hlediska OSINT analýzy kladný.

Z hlediska ochrany osobních údajů ve vztahu k OSINT lze na základě provedeného výzkumu tvrdit, že někteří uživatelé nejen ignorují problematiku ochrany osobních údajů, ale jejich jednáním doslova sabotují snahu o ochranu ze strany zkoumaných portálů. Tato aktivní účast uživatelů na uveřejňování údajů má také další negativní dopad.

Jak je vidět na obrázku č. 29, uživatelská data, která jsou vložena do těla inzerátu mohou být zaindexována archivem, ve kterém potom budou viditelná nejspíše na věky (v případě tohoto konkrétního inzerátu je tomu již bezmála 16 let).



Obrázek 29 – Inzerát na webu Bazoš.cz z 14. 6. 2004 s kontaktními údaji [55]; snímek autora (upravený)

V roce 2004 pochopitelně nebyla ochrana uživatelských dat takovým tématem, jakým je dnes a Bazoš.cz tehdy neposkytoval svým uživatelům ochranu formou zprostředkování kontaktu. To však nic nemění na skutečnosti, že pokud uživatelé tuto ochranu obchází, vystavují se riziku, že jejich údaje budou zvěčněny podobně, jako bylo zachyceno na obrázku č. 29.

Hypotéza tedy byla vyvrácena. Tato skutečnost mne dále vedla k myšlence prověřit vztah GDPR k webovým archivům a zařadit pro toto téma samostatnou diskuzní kapitolu 6.1.

6 DISKUZE

6.1 GDPR vs. webové archivy

Na základě zjištění uvedeného v hypotéze č. 2 bylo provedeno zkoumání webu Archive.org se zaměřením na české portály, u nichž je předpoklad výskytu osobních údajů.

Archive.org a další podobné služby sehrávají v této problematice zvláštní roli. Pokud uživatel zjistí, že se v aktivní části webu vyskytují osobní údaje, které mají vazbu k jeho osobě, je oprávněn využít svého práva být zapomenut (kapitola 3.2.5) a zaslat provozovateli příslušného webu žádost, aby data odstranil. Aplikace práva být zapomenut je však v případě archivních služeb výrazně ztížená až nemožná. Ačkoli je prostřednictvím zmíněného institutu možné zajistit, že označené osobní údaje z dané webové stránky zmizí, mohla být data zkopírována na mnoho dalších míst, o čemž uživatel rozhodně nebyl a logicky ani nemohl být informován.

Pro provádění archivace lze v prostředí sítě Internet uplatnit následující výjimky:

- Z článku 14 odst. 5 písm. b) nařízení GDPR vyplývá, že není třeba dotčené osoby informovat, pokud poskytnutí takových informací není možné nebo by vyžadovalo nepřiměřené úsilí, což platí zejména v případě zpracování pro účely archivace ve veřejném zájmu.
- Dle článku 17 odst. 3 písm. d) nařízení GDPR se neuplatní právo být zapomenut u archivace ve veřejném zájmu, pokud je pravděpodobné, že by aplikace tohoto práva znemožnila nebo vážně ohrozila splnění cílů uvedeného zpracování.

Archivy přístupné online, včetně webu Archive.org těchto výjimek pochopitelně využívají, což však neznamená, že právo být zapomenut nelze aplikovat vůbec. V souvislosti s tím, že osoba není o zpracování jejích údajů

informována, je však nemožné se práva domáhat v plném rozsahu. Osoba by tak musela činit např. požadavkem na odstranění či pseudonymizování všech čísel v archivu, která by mohla být jejím rodným číslem, což by teoreticky mohlo způsobit poškození dalších nesouvisejících dat.

Předpokladem je však to, že si je osoba vědoma skutečnosti, že se zde její údaje nachází. Zjištění však nelze provést nijak snadno, jako v případě aktivní části sítě, protože možnost vyhledávání v archivu je značně omezena. Plošná aplikace cenzury by pak patrně poškodila mnoho záznamů, a navíc by nemohla být zcela efektivní. Konkrétně web Archive.org je navíc spravován neziskovou organizací Internet Archive se sídlem v San Francisku v USA, přičemž GDPR je legislativní normou Evropské unie. Podobně jako v případě společnosti Google by tak pravděpodobně došlo k soudním sporům (viz kapitola 3.2.5).

Podchytit veškeré podobné situace může být i přes snahu zpracovatele nemožné a ačkoli je v prostředí sítě Internet snaha o ochranu uživatelských dat zjevná, je řešení historických dat problematické.

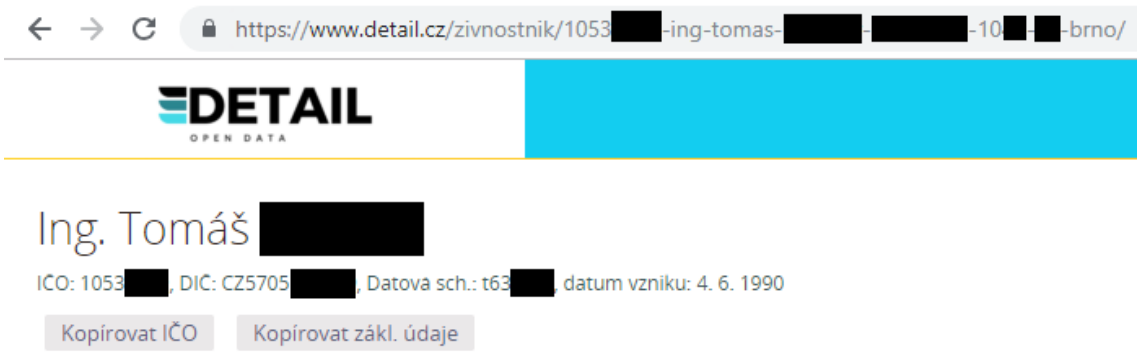
Pro potvrzení této skutečnosti byl vykonstruován dotaz ve vyhledávání na webu Archive.org k webu Peníze.cz. Tento web byl mj. využit i při výzkumu, a to díky skutečnosti, že se zde mohou nacházet jména a adresy osob spolu IČO a DIČ. Dotaz do služby Archive.org pak přinesl výsledky viditelné na obrázku č. 30.

http://rejstrik.penize.cz:80/dph/cz731130[redacted]-roman-[redacted]	text/html	Jun 1, 2016	Jul 4, 2016
http://rejstrik.penize.cz:80/dph/cz735129[redacted]-erika-[redacted]	text/html	May 5, 2016	Jun 9, 2016
http://rejstrik.penize.cz:80/dph/cz750527[redacted]-ing-ondrej-[redacted]a	text/html	Feb 26, 2017	Feb 26, 2017
http://rejstrik.penize.cz:80/dph/cz760914[redacted]-judr-robert-[redacted]	text/html	Jan 29, 2016	Mar 4, 2016
http://rejstrik.penize.cz:80/dph/cz800601[redacted]-martin-[redacted]	text/html	Mar 28, 2014	Mar 28, 2014
http://rejstrik.penize.cz:80/dph/cz801010[redacted]-bohdan-[redacted]nyy	text/html	Sep 17, 2017	Aug 18, 2019
http://rejstrik.penize.cz:80/dph/cz810720[redacted]-mgr-lukas-[redacted]ek	text/html	Oct 6, 2016	Oct 6, 2016
http://rejstrik.penize.cz:80/dph/cz860731[redacted]-radek-[redacted]	text/html	Jul 31, 2017	Jul 31, 2017
http://rejstrik.penize.cz:80/dph/cz870925[redacted]-lubomir-[redacted]	text/html	Sep 15, 2017	Sep 15, 2017
http://rejstrik.penize.cz:80/dph/cz880809[redacted]-trong-quyet-[redacted]	text/html	Mar 4, 2016	May 5, 2016

Obrázek 30 – DIČ v URL webů ve službě Archive.org [55]; snímek autora (upravený)

Již samotné výsledky obsahují osobní údaje a není tak ani třeba jednotlivé odkazy navštívit. Situace je pochopitelně zapříčiněna jednak tím, že web Peníze.cz v URL přímo tyto údaje uvádí, ale také samotným systémem České republiky, který rodná čísla využívá i pro další účely, jak již bylo uvedeno v kapitole 5.1.

V souběhu s existencí archivních služeb tvoří tyto okolnosti prostředí, které je z hlediska ochrany dat nepřehledné. Web Peníze.cz pochopitelně není jediným, informace o podnikajících fyzických osobách jsou dostupné např. také na webu Detail.cz, kde je v URL uvedeno IČO, jméno a adresa, zatímco DIČ je uvedeno „pouze“ v obsahu, jak je patrné z obrázku č. 31.



Obrázek 31 – Informace podnikající fyzické osoby viditelné na webu Detail.cz [68]; snímek autora (upravený)

Informace zde viditelné však nelze dohledat pomocí Archive.org tak snadno jako v prvním případě, neboť z URL není patrné právě DIČ. V případě, kdy by k došlo k nápravě a DIČ by již rodná čísla neobsahovala, bude stejně v archivu možné na stránky Peníze.cz i Detail.cz dál nahlížet.

Z tohoto hlediska je nutné zmínit, že hlavní registr DPH, který je ve správě Ministerstva financí ČR (aplikace Adisreg), neumožňuje indexování dotazů. Byť aplikace nepůsobí na první pohled tak moderně jako zmíněné dva weby Peníze.cz a Detail.cz, je ve skutečnosti promyšlenější. Web funguje na principu dotazů, které jsou dynamicky generovány do obsahu stránky a výsledky tak není možné indexovat. Podobně funguje také systém Evropské unie Vies.

Ačkoli tedy je z oficiálních zdrojů možné DIČ zjistit, je do budoucna v případě potřeby poměrně snadné problém vyřešit. To samé však nelze tvrdit v případě dvojice „moderních webů“, které již ve spolupráci s archivy zapříčinily nekontrolovatelné rozšíření těchto identifikátorů.

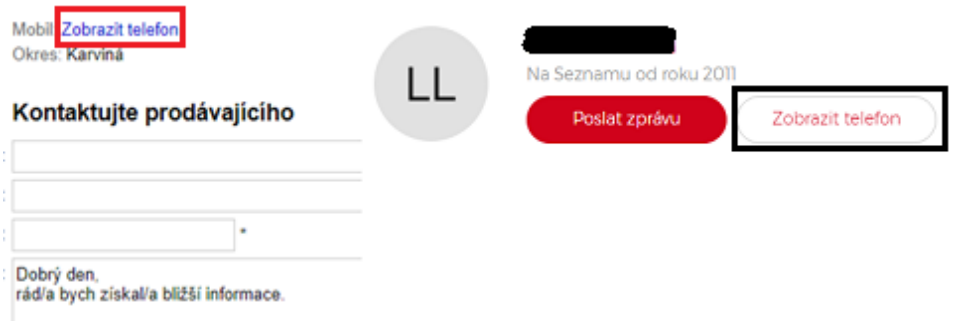
Vzhledem k tomu, že uživatel zpravidla není schopen zjistit ani dohledat, kde jsou jeho osobní údaje v rámci sítě Internet zpracovávány a má tak omezenou možnost domáhat se svého práva být zapomenut, je toto v kolizi s původním záměrem nařízení GDPR, který byl citován v kapitole 3.2.1. Cílem nařízení bylo přizpůsobit právní rámec ochrany osobních údajů dnešní době a posílení práv subjektů.

Problematika archivů však nebyla dle mého názoru upravena dostatečně. Uznávám, že nebylo snadné postihnout právě oblast, která se v mé práci jeví jako velmi problematická. Lokální problémy tohoto typu však měly být ošetřeny. Primárně tedy mělo nejpozději s účinností ZoZOÚ dojít k eliminaci DIČ obsahujících rodná čísla a také měla být podchycena problematika uveřejňování údajů subjekty, kterým dotčená osoba nedala se zpracováním souhlas.

V souvislosti s tím se navíc domnívám, že pokud již třetí strana zpracovává evidenci DIČ, ze kterých jsou snadno odvoditelná rodná čísla, stává se automaticky správcem osobních údajů, což pravděpodobně nikdo nepopírá. Dle mého názoru se však v tomto případě rozhodně nelze odvolávat na výjimku, že ke zpracování dochází ve veřejném zájmu. Ve veřejném zájmu je již tato evidence zpracovávána ze strany státu, který data v případě zájmu poskytuje. Navíc na rozdíl od služby Archive.org, která se odvolává na výjimku ve veřejném zájmu a je provozována neziskovou organizací, jsou weby Peníze.cz a Detail.cz komerčními záležitostmi.

6.2 Ochrana uživatelských dat na inzertních portálech

V této kapitole se vrátím k technikám a poznatkům učiněným v souvislosti s prováděním případových studií. Z tohoto hlediska bude prvním bodem diskuze ochrana uživatelských dat na portálech, které byly zahrnuty do výzkumu. Nedostatky v této oblasti byly objeveny pouze na službách společnosti Seznam.cz a to konkrétně na portálech Sauto.cz a Sbazar.cz. Na obou službách Seznam.cz zavedl dynamicky zobrazovaný obsah zachycený na obrázku č. 32.

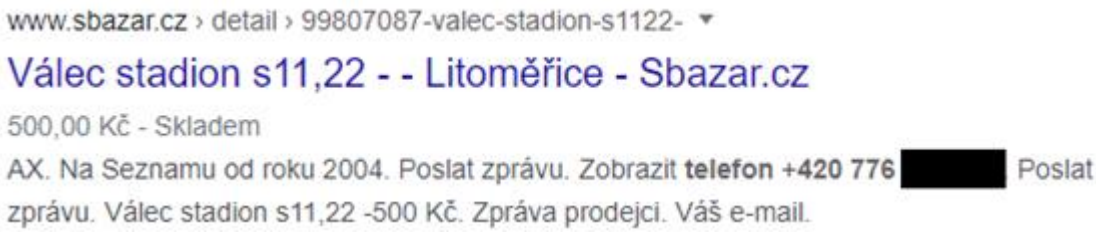


Obrázek 32 – Skrytí kontaktních údajů. ***Vlevo*** *– Sauto.cz [69];* ***vpravo*** *– Sbazar.cz [70]; snímky autora (upravené)*

Účelem dynamického obsahu je skrytí dat, která by mohla být snadno zneužita. Přestože je u jednotlivých inzerátů možné snadno a automatizovaně odkaz rozkliknout a telefonní číslo zobrazit.

Ochrana dynamickým obsahem může být efektivní a zabránit indexaci obsahu za předpokladu, že je správně implementována. Na webech Sauto.cz a Sbazar.cz je však obsah skrytý pouze pomocí html. Ačkoli standardně uživatel musí kliknout, aby číslo odkryl, číslo se po celou dobu nachází ve zdrojovém kódu stránky. Tudíž je indexováno vyhledávači a metoda tak v případě takové implementace postrádá smysl. Jak bylo zmíněno v kapitole 3.4.2 Seznam.cz provozuje také portál Sreality.cz. Na tomto webu je dynamický obsah vytvářen pomocí javascriptu a telefonní číslo není vidět ani ve zdrojovém kódu. Zde je nutné zmínit, že ve výzkumu nebyla použita žádná data z webu Sreality.cz a to nikoli z důvodu, že zde žádná z osob neinzerovala, ale proto, že data nejsou indexována, tudíž je nebylo možné nijak dohledat.

Funkci v Sreality.cz jsem blíže nezkoumal, ale svůj účel prokazatelně splnila, což však nelze tvrdit v případě webů Sbazar.cz a Sauto.cz, kde je kontakt viditelný přímo z vyhledávače Google. Příklad hledání čísla, které je „skryto“ na stránce Sbazar.cz je vidět na obrázku č. 33.



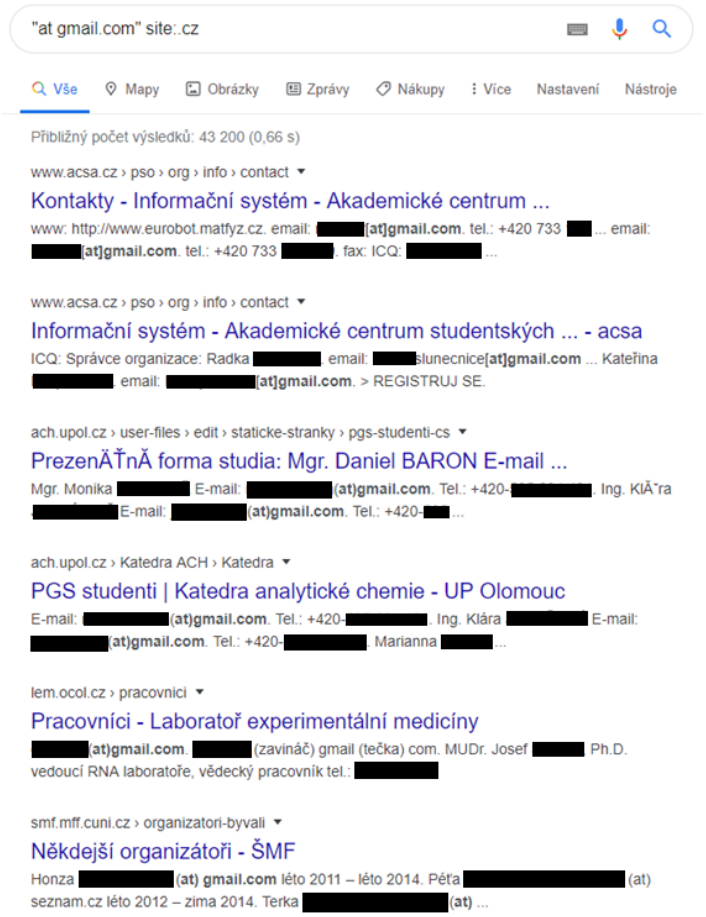
Obrázek 33 – Výsledek hledání telefonního čísla v Google – ukázka následku špatné implementace skrývání údajů [56]; snímek autora (upravený)

Proč jsou na těchto službách implementovány funkce, které neplní svou roli mi není jasné. Seznam by mohl využít stejnou metodu, kterou využívá na portálu Sreality.cz, čímž by problém vyřešil.

Díky uvedeným skutečnostem tak je možné použít portály Sbazar.cz a Sauto.cz jako zdroj informací, a navíc je možné i použití metod automatizovaného sběru dat přímo z výsledků vyhledávání (kapitola 3.5.2).

Pro srovnání uvedu další 2 známé metody pro ochranu před sběrem.

Vůbec nejstarším způsobem ochrany e-mailových adres je nahrazování znaku „@“ slovem „at“, a to i v dalších podobách jako [at], (at) nebo nahrazování „.“ za slovo „dot“. Jedná se o nedostatečnou metodu, o čemž vypovídá výsledek hledání na obrázku č. 34.



Obrázek 34 – Výsledek hledání e-mailových adres při využití „at“ místo „@“ [56]; snímek autora (upravený)

Hledání slova „at“ před názvem domény „gmail.com“ bez problému vyhledá jakýkoli zápis, ať jsou zde použity závorky či nikoli. Hledáním konkrétního jména by pochopitelně odkaz viditelný byl, což nejspíše není na škodu v případě, že o tom dotčené osoby ví. Metoda k zamezení vyhledání e-mailové adresy však naprosto selhala.

Jako nejvhodnější způsob, který zamezí vyhledávačům indexovat uživatelská data, se mi jevila konverze z textu na obrázek, čehož lze dosáhnout např. v jazyce PHP pomocí funkce imagestring z knihovny GD (https://www.php.net/manual/en/function.imagettftext.php 21.3.).

Tato metoda je účinnější než využití dynamického obsahu, neboť ani strojovou interakcí na samotném webu nelze data vytěžit. Řetězec, který má být převeden, se pochopitelně nesmí nacházet ve viditelném kódu webové stránky v plain textu (v čitelné podobě). To by v případě PHP hrozit nemělo, neboť skript běží na straně serveru a uživateli se předkládá jen výsledek. Na podobném principu znečitelnění pro stroje funguje také většina metod „captcha“, které slouží k vynucení uživatelské interakce. Znemožnění strojové interakce je přímo jejich cílem.

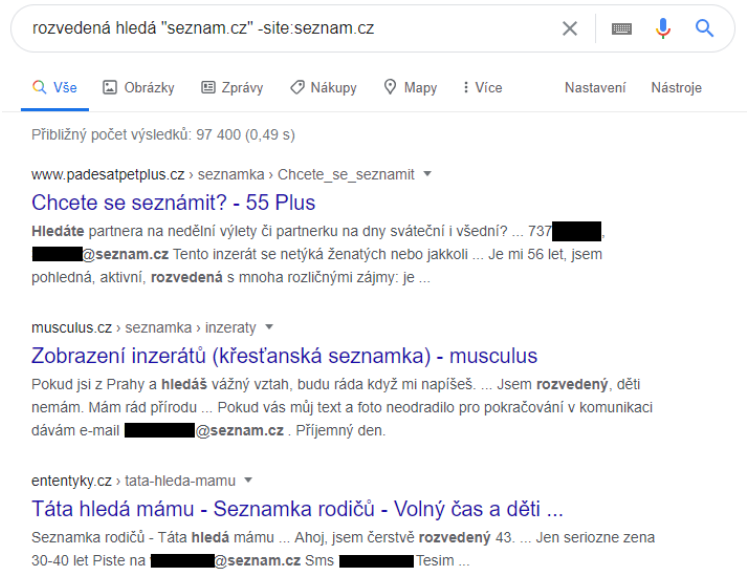
Tato metoda však zároveň vyloučí možnost strojového předčítání webu handicapovaným osobám. Ačkoli je tedy z hlediska ochrany konverze textu do obrázku účinnější, lze pro praktické využití doporučit generování spíše generování dynamického obsahu způsobem, který využívá web Sreality.cz.

6.3 Využití technik pro hromadný sběr dat

Metody uvedené v kapitole 6.2 mají především zamezit možnosti strojového sběru dat. Proč není sběr žádoucí a způsob jakým může probíhat budu demonstrovat v této kapitole. Využiji zde scénáře, kdy chce útočník získat e-maily osob, které by mohly být vhodným cílem pro tzv. romance scam.

Romance scam je druhem nigerijských podvodů (kapitola 3.3.1.2), kde útočník láká oběť k seznámení, přičemž často užívá fotografie modelek, vojenských důstojníků, doktorů apod. Útočník s obětí udržuje kontakt a ve vhodné době požádá oběť o finanční pomoc.

Výzkum provedený prof. Whitty jako nejčastější oběti označil ženy ve věku mezi 35 až 57 let, a to pravděpodobně z důvodu, že se v této kategorii vyskytuje vyšší počet osob, které mají zájem o seznámení [71]. Na základě toho jsem jako vhodná klíčová slova zvolil „rozvedená hledá“. Aby byly nalezeny e-mailové adresy, bylo do hledání přidáno „seznam.cz“ a vyloučeno hledání v doméně seznam.cz (kapitola 3.5.1.1). Výsledek hledání je viditelný na obrázku č. 35.

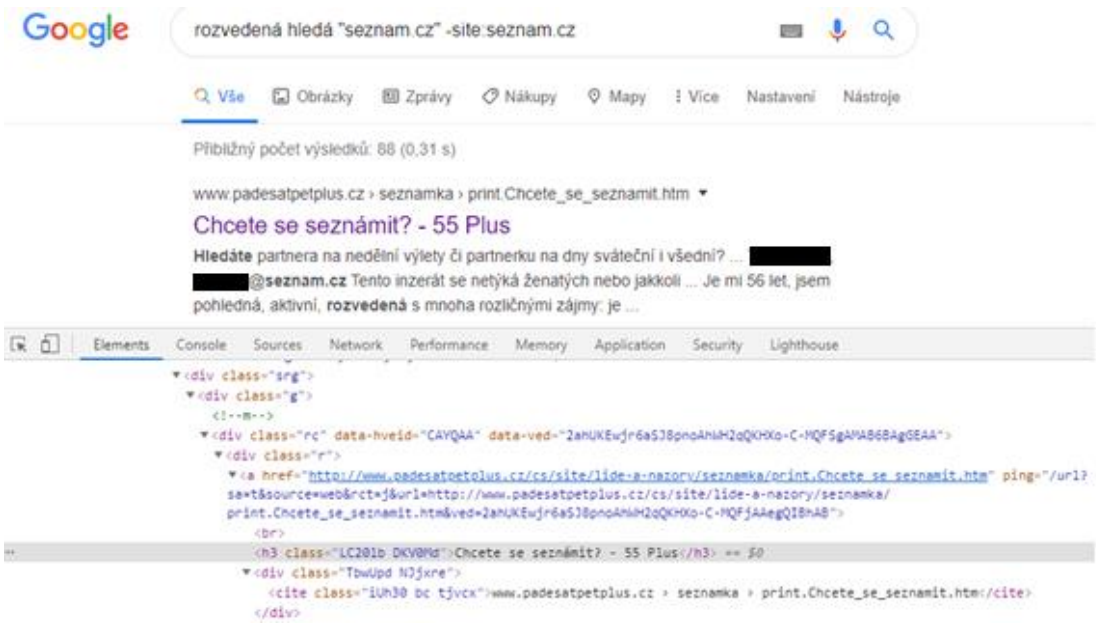


Obrázek 35 – Příklad vyhledávání specifického okruhu uživatelů [56]; snímek autora (upravený)

Výsledky hledání se pro účely sběru jeví jako relevantní, přičemž útočník výsledky hledání nemusí ručně prohlížet, postačí, když uvedené adresy vloží do nástroje, který za něj jednotlivé weby navštíví, vyhledá v nich veškeré e-mailové adresy a tyto uloží. V nastavení vyhledávače Google lze zvolit možnost zobrazení až 100 výsledků, čímž docílíme toho, že bude možné z jedné stránky extrahovat více výsledků a postup nebude nutné opakovat tolikrát.

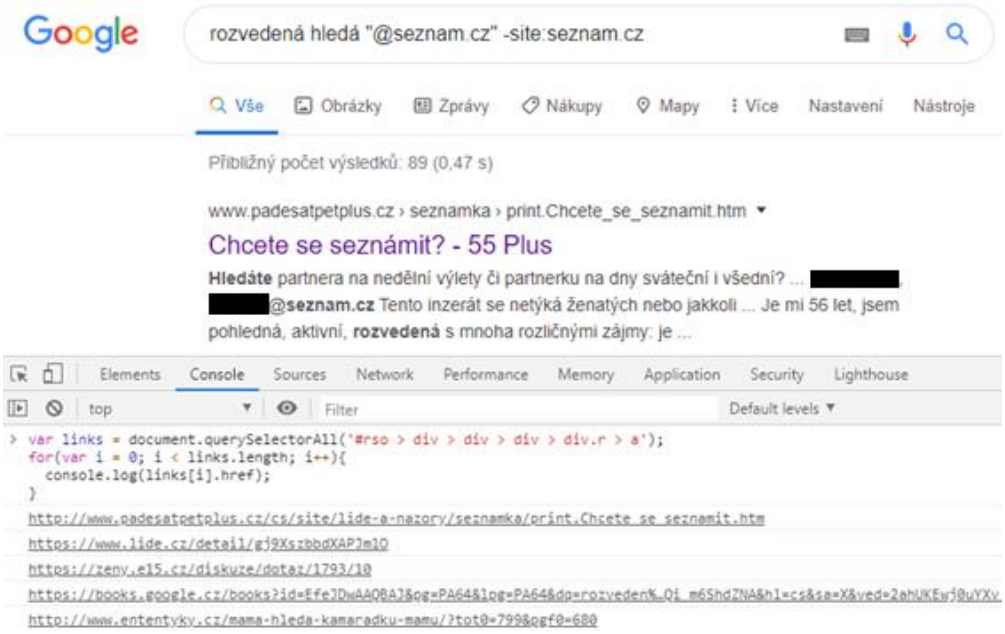
Následně použijeme vývojářské rozhraní prohlížeče, ve kterém budou provedeny 2 kroky:

1) Zjištění elementu, který obsahuje URL v kolonce „Elements“ (obr. 36).



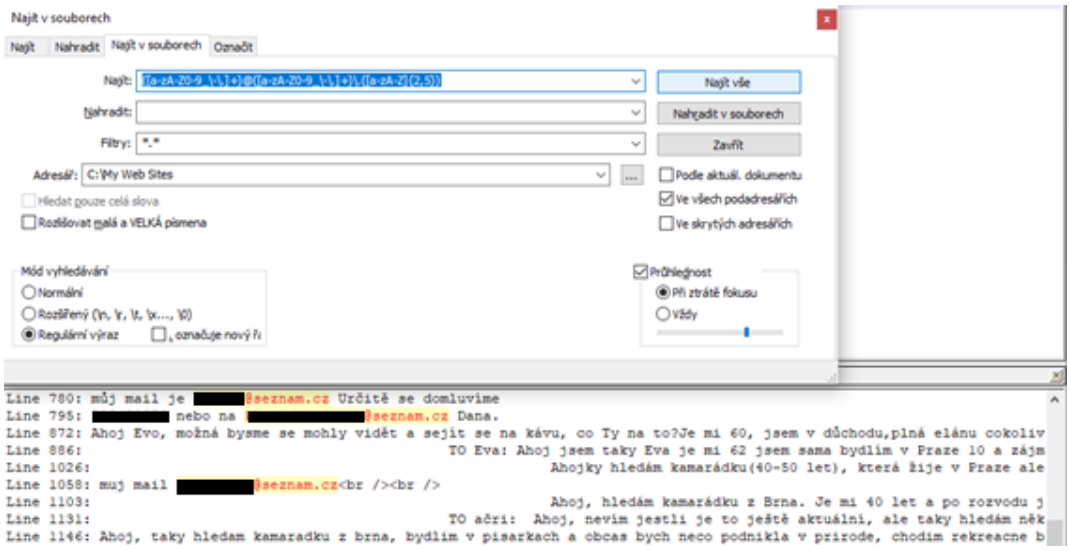
Obrázek 36 – Zjištění elementu obsahujícího požadované informace [56]; snímek autora (upravený)

2) Použití javascriptu pro extrakci všech URL z výsledků hledání (obr. 37).



Obrázek 37 – Použití konzole k extrakci výsledků hledání [56]; snímek autora (upravený)

Přestože by bylo efektivnější dále využít např. nástroj ScrapeBox, který by jednotlivé weby prozkoumal a vypsal e-mailové adresy, které se v nich nachází, budu dále postupovat manuálně, abych předvedl, jak nástroj pracuje. Adresy URL z konzole lze následně uložit do textového souboru a předat do dalšího nástroje, který stáhne obsah všech webových stránek (např. již zmíněný WinHTTrack). Z obsahu všech stažených webů pak lze extrahovat e-mailové adresy např. pomocí programů s podporou hledání pomocí regulárních výrazů (PSPad, Notepad++ atd.), jak je znázorněno na obrázku č. 38.



Obrázek 38 – Použití RegEx pro hledání e-mailových adres v nástroji Notepad++

Jak je z výše uvedeného obrázku zřejmé, objevují se zde texty nasvědčující tomu, že osoby mohou být vhodnými cíli pro romance scam. V případě, že by útočník takto postupoval, měl by zcela jistě vyšší pravděpodobnost úspěchu než použitím náhodného seznamu e-mailových adres.

Metoda nebyla v praktické části využita. Postup měl pouze demonstrovat možnosti extrakce dat a poukázat na skutečnost, že zdroje, které jsou indexovatelné, je možné snadno vytěžit.

6.4 Využití archivů

V této kapitole budu pokračovat v rozboru technik sběru dat, přičemž se zaměřím na již uvedené archivy. V první části se budu věnovat nástroji Google Cache a poté zmíním kazuistiku využití Archive.org.

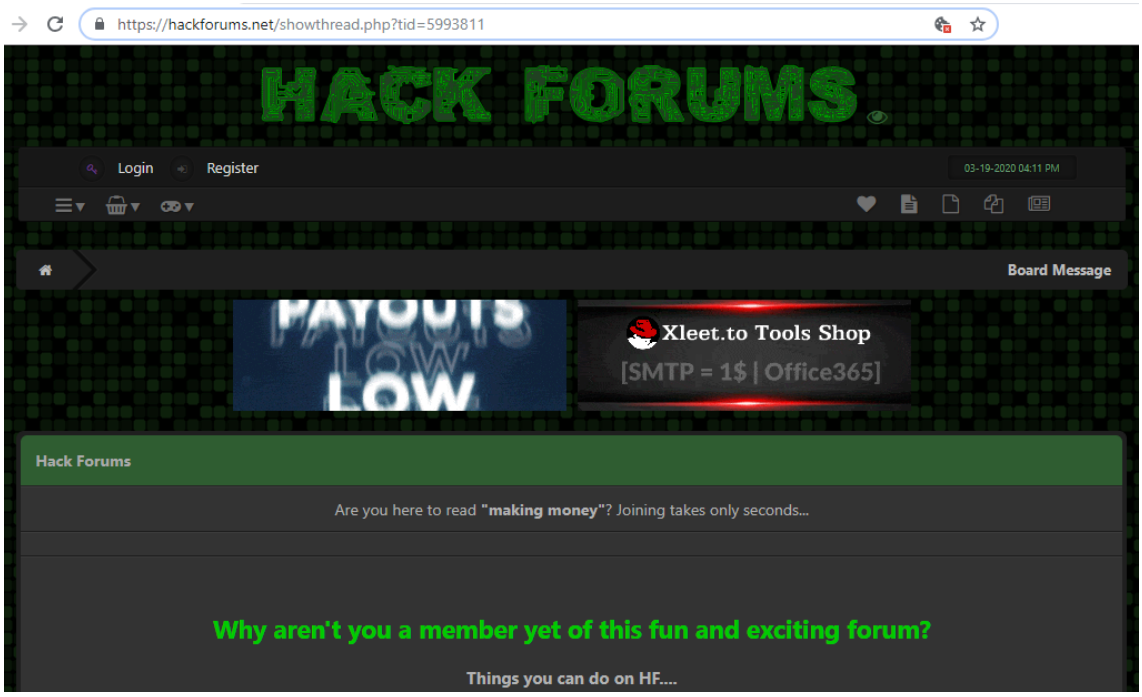
6.4.1 Google Cache

Ke Google Cache si nejprve dovolím citovat z diplomové práce Vondrušky, která je dle mého názoru poměrně zdařilou a komplexní kompilací nástrojů pro OSINT. Vondruška zde k problematice Google Cache uvádí:

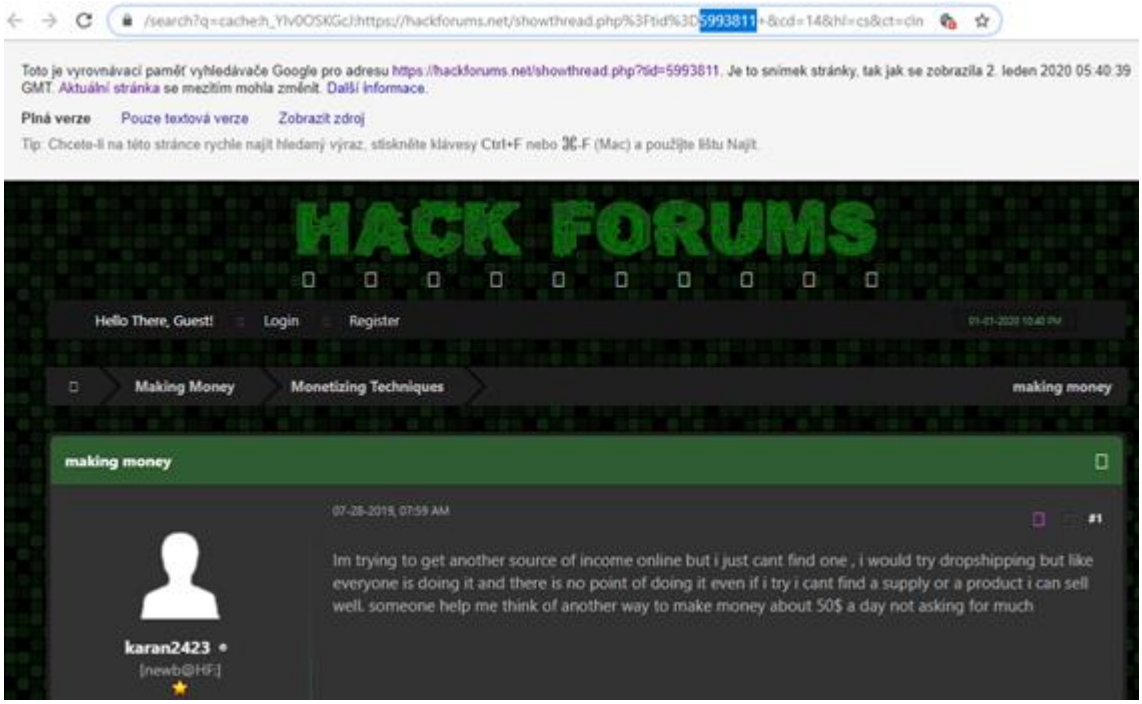
„Indexace webových stránek není možná v případě existence dynamického obsahu nebo v případě nutnosti autorizovaného přístupu ke stránce, kdy uživatel zadává své uživatelské jméno a heslo (například diskuzní fóra). Přesto jsou k dispozici i výsledky zobrazující obsah některých diskuzních fór (či jiných stránek vyžadujících autentizaci) a to díky možnosti zobrazení výsledků z vyrovnávací paměti vyhledávače" [72].

Citace obsahuje tvrzení o dynamickém obsahu, které částečně potvrzuje zjištění z kapitoly 6.2. Vondruška však nespecifikuje, že dynamický obsah musí být generován jinak, než je tomu např. u služeb Sauto.cz a Sbazar.cz. Dále zde Vondruška zmiňuje poměrně zajímavou skutečnost, že Google Cache je schopna zobrazit jinak nedostupný obsah. Nechává však tento paradox nevysvětlený. Osobně jsem se s tímto několikrát setkal také a sám jsem záměrně hledal stránky pomocí operátoru „cache:", abych mohl nahlížet na obsah různých diskusních fór bez nutnosti (často i placené) registrace. Tato funkce byla veřejně odhalena na webu hackforums.net, přičemž i tato stránka je v některých případech k tomuto „hacku" náchylná. Konkrétní příspěvek se v cache nenachází, nicméně je možné demonstrovat techniku na přístupu např. do vlákna s ID 5993811.

Na obrázku č. 39 je zachycen pokus o přístup do vlákna standardní cestou a na obrázku č. 40 je vlákno zpřístupněno pomocí Google Cache.



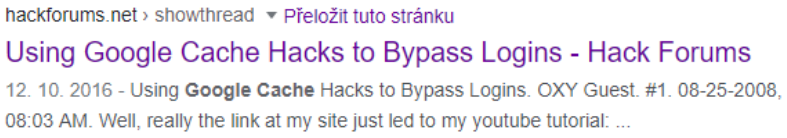
Obrázek 39 – Standardní přístup na uzamčené fórum [73]; snímek autora



Obrázek 40 – Obsah jinak uzamčeného fóra při využití Google Cache [56]; snímek autora

Jak je patrné z obrázku č. 40, je při využití odkazu na cache z vyhledávače Google zobrazen obsah vlákna pro účet „Guest“. Google Cache obsahuje snapshot surového kódu HTML, který Googlebot obdržel od serveru, na němž je obsah publikován. Tento HTML kód je pak interpretován prohlížečem uživatele [74].

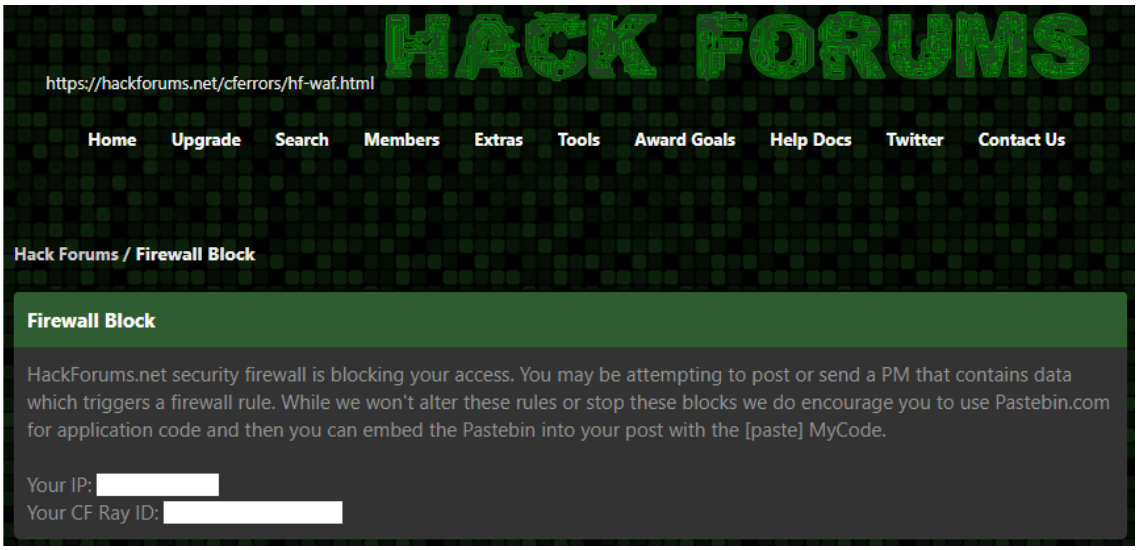
Záhada má tedy poměrně jednoduché vysvětlení. Administrátor na svůj web dobrovolně vpouští crawlery a ty nechává indexovat obsah webu, aniž by crawlerům sdělil, že si nepřeje obsah archivovat. V tomto případě se konkrétně jedná o crawler Googlebot. Ten je aktivním návštěvníkem webu, kterému je pro indexaci přiděleno oprávnění vstupu (Guest). Pokud administrátor nechce, aby byl web tímto způsobem archivován, je jednoduchým řešením zakomponování kódu `<meta name=”robots” content=”noarchive” />` do stránek, které nemají být uloženy do cache, přičemž tyto budou nadále indexovány. Při indexování se však tímto způsobem nevyhne skutečnosti, že bude část obsahu součástí indexu, jak je zřejmé z obrázku č. 41.



Obrázek 41 – Příklad indexu prozrazujícího část obsahu jinak uzamčeného fóra [56]; snímek autora

Ačkoli šipka vedle odkazu na přeložení stránky často obsahuje možnost zobrazení cache, byla v tomto případě pod šipkou skryta možnost „zobrazit podobné“, která využívá vyhledávacího operátoru „related:“. Obsah na obrázku č. 41 uvedený pod odkazem tedy pochází z indexace, nikoli z cache. Pokud byl bot instruován, aby cache neprovedl, je tento náhled to jediné, co lze zjistit. Je tedy čistě na administrátorovi webu, zda botům cachování (případně též indexaci) umožní. Případné indexaci webu lze zabránit přidáním kódu `<meta name=”robots” content=”noindex”>`.

Pokud bychom chtěli simulovat chování Googlebota a získat obsah podobným způsobem, nepůjde to. Pokusu o simulaci přístupu crawleru jsem docílil podvržením user-agent pomocí rozšíření User-Agent Switcher v prohlížeči Chromium. User-agent slouží k poskytnutí informací o prohlížeči návštěvníka. Rozšíření jsem nakonfiguroval tak, aby prohlížeč vystupoval jako „`Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)`“. Po návštěvě webu se pak zobrazila stránka zachycená na obrázku č. 42.



Obrázek 42 – Pokus o přístup s podvržením user-agent [73]; snímek autora

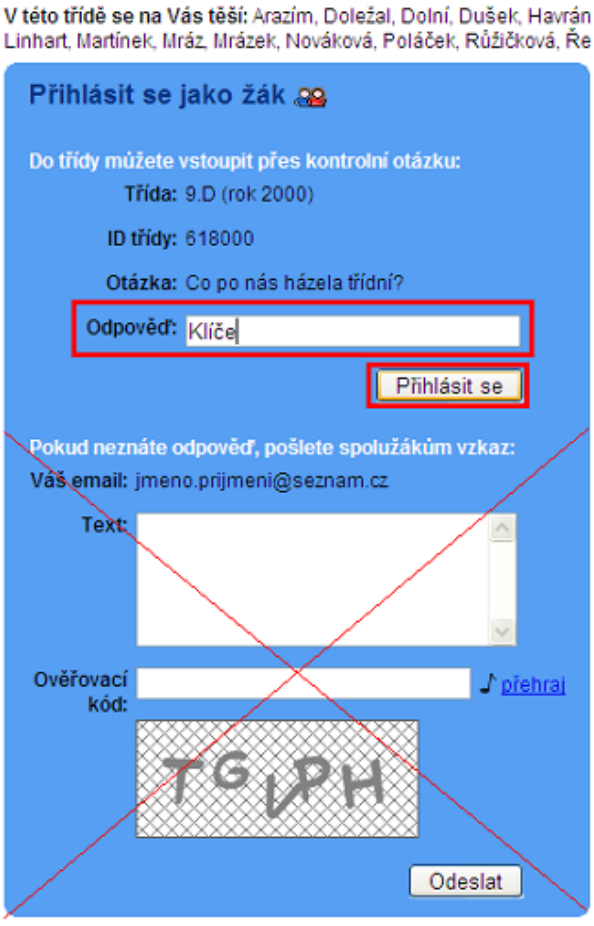
Z textu vyplývá, že pro přístup Googlebota je definováno pravidlo ve firewallu. Sdělení je dále irelevantní, nicméně ze zprávy vyplývá, že je přístup bota omezen na IP adresy, které užívá Google, a to právě z důvodu, aby nemohlo dojít k zneužití. Pokud by měl bot vyšší oprávnění než neregistrovaný návštěvník (jako v tomto případě), mohl by se uživatel za bota touto metodou vydávat a zpřístupnit tak skrytý obsah.

Pokud tedy Googlebot navštíví danou stránku, je obsah webu indexován a cachován, pokud administrátor neurčí jinak.

6.4.2 Archive.org

V této kapitole budu demonstrovat konkrétní možnost zneužití popsaných archivů. Pro tento účel využiju již neexistující web Spoluzaci.cz. Z hlediska ochrany osobních údajů byl web Spoluzaci.cz špatně koncipován a následující příklad poslouží i jako ukázka toho, co vše může být špatně.

Ačkoli se mi ani s pomocí Archive.org nepodařilo vzkřísit potřebnou stránku, pomohl mi sám Seznam.cz, na kterém je do dnešního dne k dispozici nápověda k tomuto již ukončenému webu. Z nápovědy byl pořízen následující snímek, který je v obrázku č. 43.



Obrázek 43 – Snímek z nápovědy Seznam.cz ke službě Spolužáci.cz [75]

Prvním problémem je, že již přihlašovací okno zobrazuje příjmení spolužáků. Veškeré spolužáky však bylo možné zobrazit už při rozkliknutí každé třídy v určité škole.

Nejedná se sice o až tak závažnou informaci, že by nesměla být publikována, nicméně i zjištění spolužáků případného cíle by mohlo pomoci. Těžko bychom např. na síti Facebook hledali někoho jménem Jan Novák. Pokud však víme, s kým Novák chodil na základní školu, můžeme se pokusit dohledat profily spolužáků. Třeba se Novák objeví v přátelích některého bývalého spolužáka.

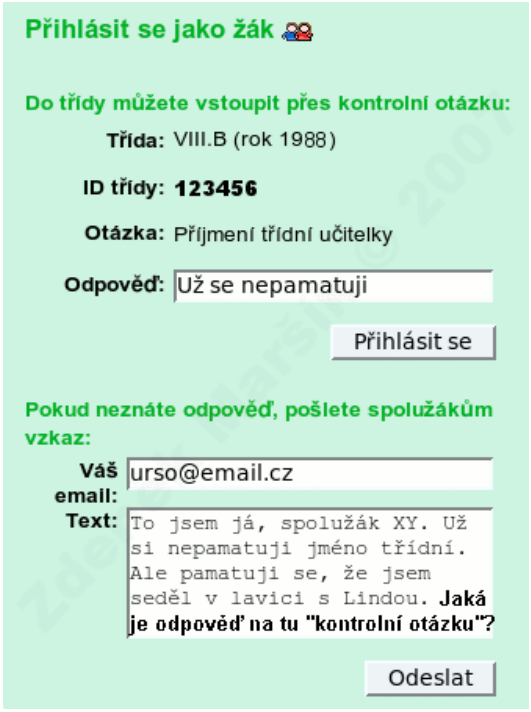
Druhý, mnohem závažnější problém, nám na snímku názorně demonstruje i sám Seznam.cz. Odpověď na otázku *„Co po nás učitelka házela?“* by jistě uhodl každý. Bylo by jen s ohledem na diakritiku a velká a malá písmena třeba vyzkoušet několik možností zápisu slova „klíče“ a „křídu“. O síle hesla bylo pojednáno mj. v kapitole 3.3.1.5. Z principu se mělo jednat o heslo, které by měl znát každý spolužák, nicméně by nemělo být snadno odvoditelné nikým dalším (např. dedukcí přezdívky z příjmení).

Pochopitelně by se uvedeným postupem útočník dopustil přinejmenším přečinu neoprávněného přístupu k počítačovému systému dle § 230 odst. 1 trestního zákoníku.

Z dalšího webu pak pochází snímek, po kterém bude následovat text publikovaný s vědomím, že je služba vypnuta a nemůže se tak jednat o návod k trestné činnosti. Jde tedy pouze o demonstraci možných rizik a možnost zneužití archivů a metod OSINT.

Přihlášení do třídy bylo možné po přihlášení na Seznam.cz a vyplnění odpovědi na otázku zvolenou správcem třídy (uvedená otázka „Co po nás házela třídní?“).

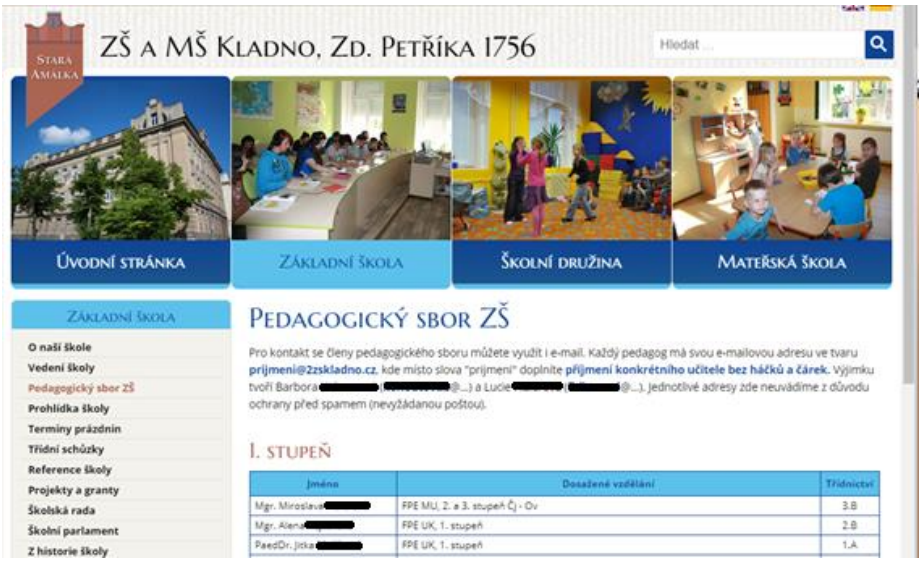
Nejčastějším heslem pro vstup však bylo příjmení třídní učitelky. To lze ostatně vidět i na obrázku č. 44.



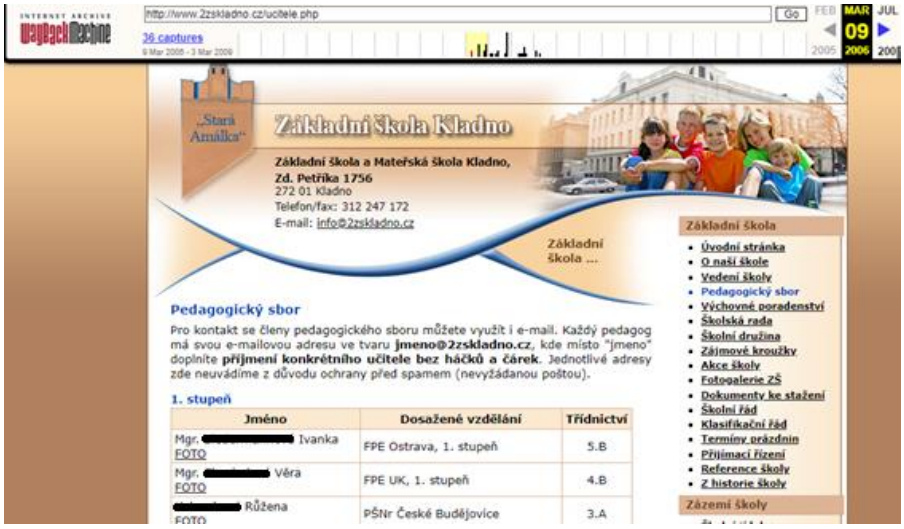
Obrázek 44 – Přihlašovací formulář do třídy na webu Spolužáci.cz [76]

Slabinou tohoto hesla je, že pokud škola uveřejnila seznam vyučujících, bylo možné zjistit heslo pro vstup do třídy. Pro případ, že by útočník potřeboval zjistit vstup do některé ze starších tříd, bylo možné využít právě službu Archive.org. Přestože hledat informace by zrovna v případě roku 1988 bylo výrazně obtížnější, mohli bychom vyzkoušet celý učitelský sbor. Pochopitelně by v roce 2020 byla poměrně malá šance, že zde učitelka – „hledané heslo“ stále učí. Mohli bychom se však pokusit využít archivu a zobrazit stránku co nejstarší.

Pro názornost jsem zadal do vyhledávače Google dotaz „zs kladno“, první odkaz mne zavedl na web 2zskladno.cz, který patří ZŠ a MŠ Kladno, Zd. Petříka 1756. Seznam třídních učitelů se zde vyskytuje, jak je zřejmé z obrázku č. 45, na obrázku č. 46 je pak zachycen snapshot webu z roku 2006.



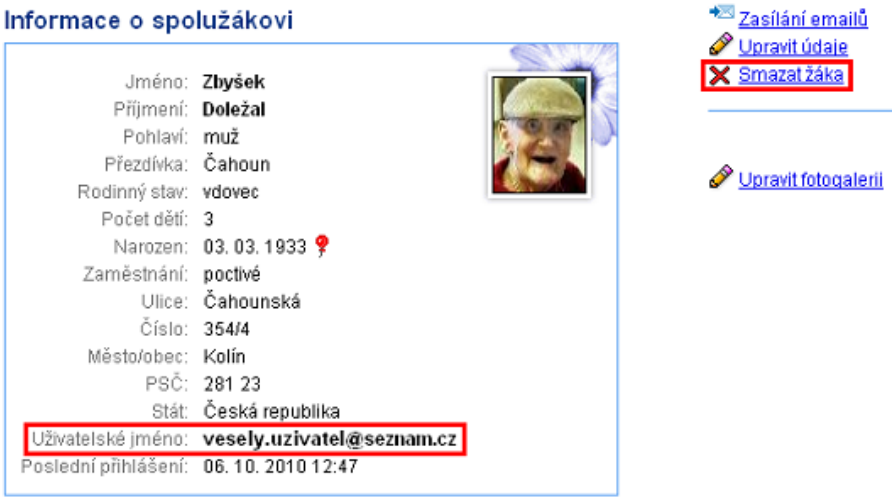
Obrázek 45 – Pedagogický sbor ZŠ a MŠ Kladno, Zd. Petříka 1756 v roce 2020 [77]; snímek autora



Obrázek 46 – Pedagogický sbor ZŠ a MŠ Kladno, Zd. Petříka 1756 v roce 2006 [55]; snímek autora

Pokud by tedy útočník v minulosti chtěl získat přístup do některé z tříd této základní školy zabezpečené heslem „jméno třídní učitelky“, mohl by tak díky archivu a škole, která uveřejnila seznam učitelů, učinit poměrně snadno.

Důvodem pro přístup by přitom mohla být potřeba zjistit údaje o některém ze studentů, které jsou viditelné na obrázku č. 47.



Obrázek 47 – Informace o spolužákovi na webu Spolužáci.cz [75]

Mimo údaje viditelné na předchozím obrázku zde mohlo být také telefonní číslo nebo číslo služby ICQ (I Seek You – jeden z prvních instant messengerů), která byla v době největší slávy této sociální sítě často využívána. Vážnější problém byla skutečnost, že údaje o sobě nemusel zadat sám žák, ale mohl tak učinit správce třídy. Ve třídě se dále mohly nacházet fotografie.

V tomto případě si byl Seznam.cz těchto závažných chyb ve svém systému vědom a rozhodl se službu ukončit. Oceňuji také upřímnost jejich vyjádření k portálu Spolužáci.cz: *„Od 25. května začne platit Obecné nařízení o ochraně osobních údajů. Abychom mohli změny, které z něj vyplývají, implementovat na Spolužácích, museli bychom službu od základů přepsat. Toto řešení se ukázalo jako neefektivní a finančně příliš náročné, proto jsme dospěli k rozhodnutí službu ukončit a úsilí i peníze věnovat do rozvoje jiných služeb“* [78].

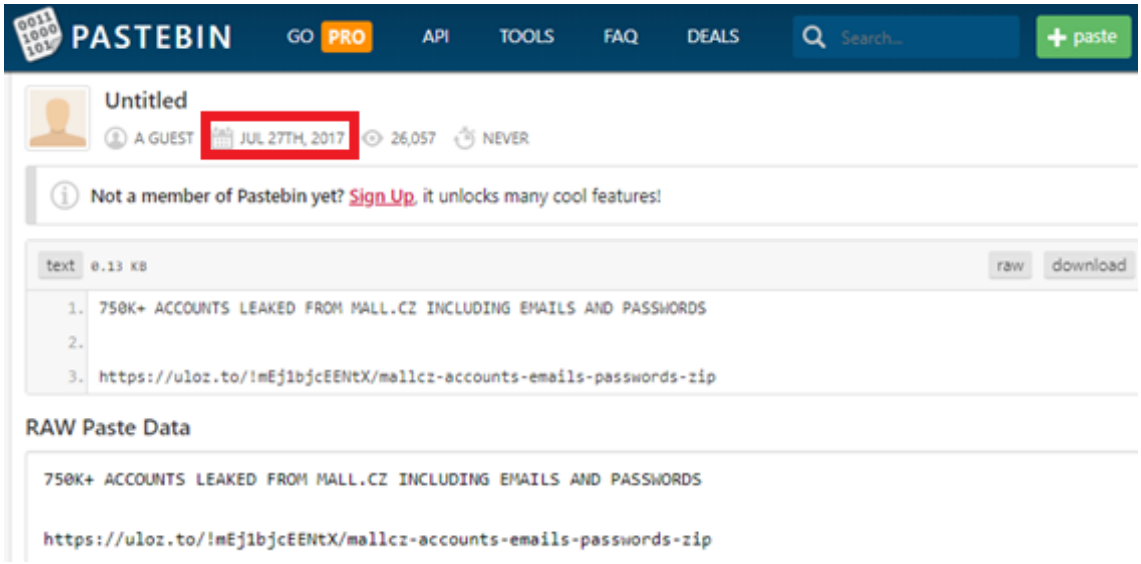
Jak je patrné z výše uvedených demonstrací, jsou nástroje Google Cache a Archive.org velmi užitečné a v OSINT mají široké spektrum využití.

6.5 Úniky hesel

Faktor, který v provedeném výzkumu mnoho uživatelů zařadil do kategorie vysokého rizika, byl únik hesla k účtu spojenému se zjištěnou e-mailovou adresou. V této kapitole proto provedu rozbor úniku dat z Mall.cz, který postihl více než 700 tisíc uživatelů převážně z České republiky a o kterém postižená společnost informovala veřejnost 27. srpna 2017 [79].

6.5.1 Únik dat z Mall.cz

Mall.cz sice poměrně vhodně informoval uživatele prostřednictvím e-mailu a okamžitě provedl reset hesel, což je dobrý předpoklad pro řešení situace. Uživatel tak může zareagovat a učinit příslušná opatření, aby nedošlo ke kompromitaci jeho dalších účtů. Nicméně Mall.cz uživatele vyrozuměl až celý měsíc po uveřejnění databáze s uniklými hesly, jak je patrné z obrázku č. 48.

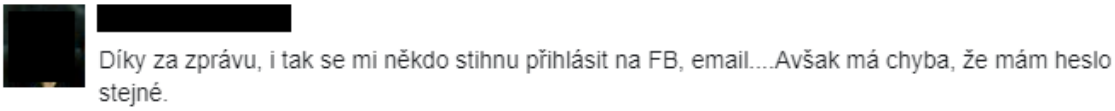


Obrázek 48 – Příspěvek na webu Pastebin.com, kde se objevil odkaz ke stažení databáze Mall.cz [80]; snímek autora (upravený)

Pomocí Archive.org se k tomuto webu podařilo zjistit, že ke dni 28. 8. 2017 již odkaz ke stažení vidělo 448 lidí, jak je patrné z obrázku č. 49.

Obrázek 49 – Archivovaný příspěvek na webu Pastebin.com z 28. 8. 2017 [55]; snímek autora (upravený)

Pozdní informování pak mohlo mít za následek další škody. O tom hovoří např. uživatel v komentáři na stránce s oznámením Mall.cz, jehož příspěvek je na obrázku č. 50.



Obrázek 50 – Příspěvek k vyjádření o úniku Mall.cz na jejich webu [79]

K síle hesel jsem se vyjádřil v teoretické části v kapitole 3.3.1.5. Je tedy částečně chybou uživatele, pokud lze jeho heslo použít i na dalších službách. Pochopitelně nejhorší situace nastává tehdy, pokud je heslo zároveň použito pro přístup k samotnému e-mailovému účtu. Útočník je pak schopen resetovat hesla do veškerých služeb, o kterých zjistí, že mají vazbu k účtu a nemají další stupeň ochrany (např. dvou-fázové ověření pomocí SMS).

K samotnému incidentu vytvořil Mall.cz webovou stránku s vyjádřením, ve kterém uvádí: „*Nedávno jsme zaznamenali pokus o narušení bezpečnosti, který se dotkl starší databáze uživatelských účtů, jež neměly dostatečně silné heslo*“ [79].

Toto tvrzení si dovolím zpochybnit, neboť uniklá data neobsahovala jen hesla, ale také další údaje (telefonní čísla a jména) [81]. V tomto případě již tedy rozhodně není vina na samotných uživatelích a jejich slabých heslech a Mall.cz

tedy lže. Leda by Mall.cz spojením „jež neměly dostatečně silné heslo“ mínil své vlastní databáze, nebo mluvil o jiném případu.

Mall.cz dále svádí problém na mechanismus hashování MD5:

„Od listopadu 2012 jsme bezpečnost hesel zajišťovali hashovací metodou SHA1 + unikátní solí a od října 2016 chráníme přístupové údaje jednou z nejsilnějších hashovacích metod bcrypt. Do roku 2012 byly údaje hashovány metodou MD5, která dnes již není považována za bezpečnou“ [79].

Je pravda, že MD5 skutečně není považována za bezpečnou, nicméně bezpečná dnes není už ani SHA1. Prolomení samotné hashe SHA1 pomocí hrubé síly je přibližně 3krát náročnější než prolomení MD5. Dodatečnou ochranu zde poskytuje až použití soli (salt). Skutečnost, že Mall.cz držel hesla zahashovaná bez salt je z hlediska bezpečnosti faux pas.

Jak uvádí Špaček: *„Za 45 minut jsem měl vylámáno 165 tisíc hesel, tedy cca 43 %. Za 12 hodin jsem cracknul skoro všechna, zůstalo mi 935 nevylámaných hesel“* [36].

Rozluštění databáze v takto krátkém čase bylo možné právě díky absenci salt. Salt se v kryptografii rozumí několik náhodných znaků, které jsou použity při hashování jednosměrnou funkcí, čímž je eliminována možnost zpětného rozluštění hash pomocí slovníkových útoků nebo rainbow-tables (prekalkulovaných tabulek s výsledky hash).

V souvislosti s tímto incidentem hrozí do budoucna Mall.cz problémy se žalobami. První případ úspěšné žaloby se již objevil a byť jsou o osobě, která žalobu podala, údaje veřejně známé a zjistitelné, bylo žalující straně dáno za pravdu [81].

Michal Špaček se případem zabýval podrobněji a zjistil, že únik byl pravděpodobně ještě rozsáhlejší. Společnost Mall.cz tedy pravděpodobně byla schopna reagovat jen na základě uniklého souboru s hesly a sama neví, jaké škody útočník napáchal [36].

6.5.2 Možnost prolomení hesla

Samotná skutečnost, že došlo ke kompromitaci databáze s uživatelskými účty ještě nemusí nutně znamenat, že je útočník schopen data o uživatelích využít. Důležitou roli hraje již zmíněná salt. Zásadní je však také použití bezpečné šifrovací funkce. V případech funkcí SHA1 a dalších totiž nejde o funkce určené pro ukládání hesel, byly navrženy pro rychlý běh, což je v případě hesel naopak nežádoucí. Běžnému uživateli při přihlášení nevadí, když musí např. 5 vteřin čekat.

Mall.cz tedy po zkušenostech správně přikročil k použití funkce bcrypt určené pro uchovávání hesel. To pochopitelně obnáší vyšší náklady na výpočetní výkon při každém přihlášení, nicméně se jedná o seriózní řešení a je rozhodně vhodnější než stále migrovat data do databází zabezpečených novější hashovací funkcí.

Výkon grafických karet, které se pro prolamování používají, roste každým rokem a v případě použití rychlých hashovacích funkcí je tak jen otázka času, kdy bude nutné opět přecházet na novější variantu. Ochranu uživatelských dat není vhodné podceňovat, a naopak bych se přikláněl k volbě vyšší úrovně zabezpečení dat i za cenu vyšších provozních nákladů.

6.6 Využití Google hacking

Metody Google hacking byly stěžejní pro hledání informací v rámci výzkumné části této práce. Jedná se však o snadno zneužitelnou pomůcku, která při specifikování určitých typů dotazů umí nalézt např. špatně zabezpečené soubory s hesly, databáze, kamerové systémy atd. Přehled řetězců, které vedou k podobným výsledkům, je k dispozici na webu https://www.exploit-db.com/google-hacking-database.

Vzhledem k tomu, že jsem se setkal s názorem, že se využitím Google hackingu může člověk dopustit trestného činu, rozhodl jsem se provést rozbor situace.

Ke Google hackingu se vyjadřuje Vondruška na str. 35, kde uvádí: *„Zneužití Google hackingu by mohlo vést k protiprávnímu jednání dle následujících paragrafů Trestního zákoníku“* [71, s. 35], přičemž dále navazuje citace Kratochvíla ze dne 12. 3. 2013, který v článku v magazínu Chip z roku 2009 uvedl:

„Jednání pachatele trestného činu podle § 257a TrZ spočívá v získání přístupu k nosiči informací a zároveň: v neoprávněném užití informací (§ 257a odst. 1a); ve zničení, poškození nebo učinění informací neupotřebitelnými (§ 257a odst. 1b); v zásahu do technického nebo programového vybavení počítače (§ 257a odst. 1c)“ [82].

Kratochvíl v článku uvádí ustanovení ze zákona č. 140/1961 Sb., trestního zákona, který byl v té době platný. Vondruška však psal práci v roce 2013, kdy již platil nový zák. č. 40/2009 Sb., trestní zákoník, který nabyl účinnosti 1. ledna 2010. Ačkoli Vondruška zákon pojmenovává názvem „trestní zákoník“, což je název právě zákona č. 40/2009 Sb., cituje dále znění původního trestního zákona. Uvedené jednání je v novém trestním zákoníku kodifikováno jako § 230 Neoprávněný přístup k počítačovému systému a nosiči informací.

Samotné využití Google hackingu nelze v žádném případě kvalifikovat jako trestný čin a jako trestné nelze považovat ani následné zpřístupnění dat, která ze své povahy mohou být citlivá. Dle Šámalova komentáře k § 230 trestního zákoníku:

„Nově je trestný samotný neoprávněný přístup k počítačovému systému nebo k jeho části. Tento čin bývá též označován anglickým termínem „hacking“ (osoba, která se jej dopustí, se pak označuje jako „hacker“). Podmínkou trestnosti však je, že pachatel překonal bezpečnostní opatření“ [83, s. 2086].

Pokud tedy nebylo nutné překonat bezpečnostní opatření, nemohla být naplněna skutková podstata tohoto trestného činu. Data, která se nachází v otevřené části sítě Internet a k jejichž zpřístupnění není požadováno oprávnění, lze považovat za data veřejná.

Uvedené jednání je trestné až za předpokladu, že získávání výsledků na dotazy, směřuje k budoucímu trestnému činu. Jmenovitě se však jedná pouze porušení tajemství dopravovaných zpráv dle § 182 trestního zákoníku, nebo již zmíněný neoprávněný přístup k počítačovému systému dle § 230 trestního zákoníku, a to dle následujícího ustanovení:

§ 231 Opatření a přechovávání přístupového zařízení a hesla k počítačovému systému a jiných takových dat

(1) ***Kdo v úmyslu spáchat trestný čin porušení tajemství dopravovaných zpráv podle § 182 odst. 1 písm. b), c) nebo trestný čin neoprávněného přístupu k počítačovému systému a nosiči informací podle § 230 odst. 1, 2 vyrobí, uvede do oběhu, doveze, vyveze, proveze, nabízí, zprostředkuje, prodá nebo jinak zpřístupní, sobě nebo jinému opatří nebo přechovává***

a) zařízení nebo jeho součást, postup, nástroj nebo jakýkoli jiný prostředek, včetně počítačového programu, vytvořený nebo přizpůsobený k neoprávněnému přístupu do sítě elektronických komunikací, k počítačovému systému nebo k jeho části, nebo

b) ***počítačové heslo, přístupový kód, data, postup nebo jakýkoli jiný podobný prostředek, pomocí něhož lze získat přístup k počítačovému systému nebo jeho části****,*

bude potrestán odnětím svobody až na jeden rok, propadnutím věci nebo jiné majetkové hodnoty nebo zákazem činnosti.

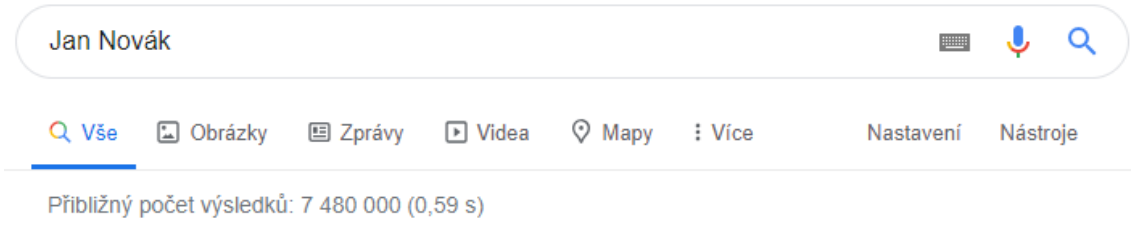
Trestné by tedy bylo až jednání, kdy by pachatel použil pro zpřístupnění dat, byť i výchozí heslo systému, který takovým způsobem objevil (typicky účet admin a heslo admin). V roce 2009 skutečně mohl být protiprávní už samotný přístup k nosiči informací, pokud byla data neoprávněně užita, poškozena nebo pozměněna, popř. bylo pozměněno technické nebo programové vybavení, avšak za předpokladu, že byl čin spáchán s úmyslem způsobit jinému škodu nebo jinou újmu nebo získat sobě nebo jinému neoprávněný prospěch.

6.7 Vyhledávání v Google

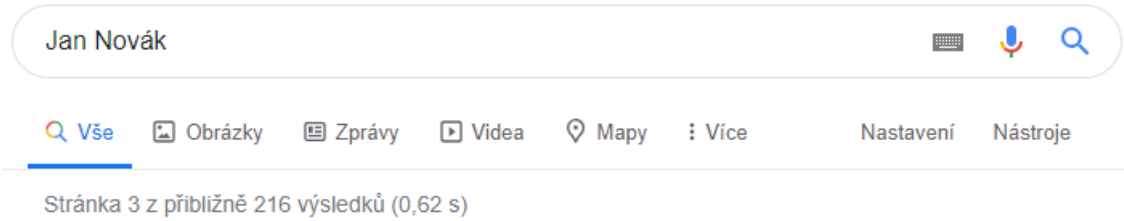
V této kapitole navážu na problematiku hledání v Google. Konkrétně zde zmíním problém, se kterým jsem se setkal v průběhu výzkumné části. Při zadávání dotazů do Google jsem zaznamenal, že se jednotlivé výsledky hledání přesouvaly a ztrácely. Nejednalo se přitom o to, že by výsledky byly v průběhu odstraněny, ale o fakt, že se výsledky vyhledávání Google v různých podmínkách měnily. Problém jsem se rozhodl replikovat a provést jeho rozbor.

Pro demonstraci byly použity prohlížeče Iridium (dále prohlížeč č. 1, jedná se o derivaci prohlížeče Chromium) a Chromium (dále prohlížeč č. 2). Oba prohlížeče jsem nastavil, aby zobrazovaly 100 výsledků na jedné stránce.

Výsledek z prohlížeče č. 1 je zachycen na obrázku č. 51, následná změna výsledků je viditelná na obrázku č. 52.

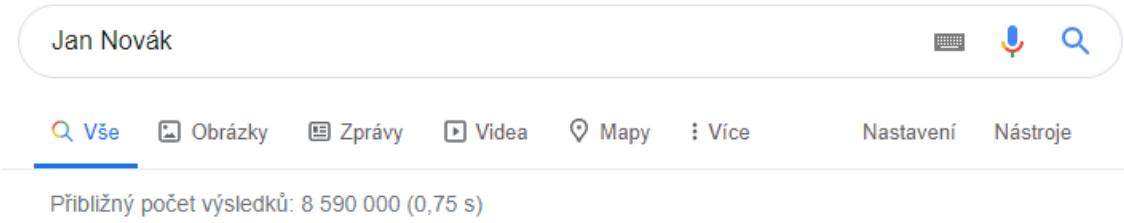


Obrázek 51 – Prohlížeč č.1 – původní počet výsledků vyhledávání v Google [56]; snímek autora

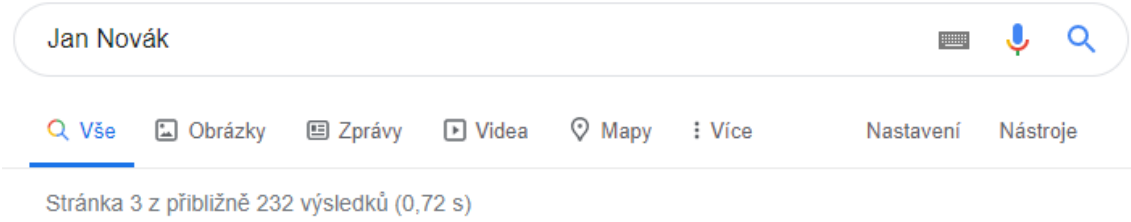


Obrázek 52 – Prohlížeč č.1 – změna počtu výsledků vyhledávání v Google [56]; snímek autora

Google neuvádí zcela korektní a pravdivé informace, neboť původně oznámil, že našel 7,48 miliónu výsledků, ale reálně zobrazil jen 216 a ve stejnou chvíli navíc v druhém prohlížeči uvádí jiné hodnoty, jak je viditelné na obrázcích č. 53 a 54.



Obrázek 53 – Prohlížeč č.2 – původní počet výsledků vyhledávání v Google [56]; snímek autora



Obrázek 54 – Prohlížeč č.1 – změna počtu výsledků vyhledávání v Google [56]; snímek autora

Počet výsledků se měnil i v průběhu psaní této kapitoly. Rozhodl jsem se proto výsledky zapsat a porovnat. S použitím techniky pro extrakci odkazů (kapitola 6.3) jsem provedl komparaci prohlížečů a dospěl k výsledkům uvedeným v tabulce č. 12.

Tabulka 12 – Rozdílné počty výsledků při vyhledávání v Google dvěma různými prohlížeči

	Prohlížeč 1	Prohlížeč 2	Celkem
Počet výsledků celkem	224	242	466
Po korelaci duplicit	224	238	462
Unikátních výsledků	23	37	60

Z tabulky je zřejmé, že ani jeden z prohlížečů neposkytl kompletní přehled, a to s poměrně vysokou odchylkou. Celkově bylo nalezeno 261 unikátních odkazů. V případě prohlížeče č. 1 nebylo zobrazeno 37 výsledků hledání, což znamená, že Google v tomto případě zamlčel cca 14,2 %, prohlížeč č. 2 nezobrazil 23 výsledků což odpovídá 8,8 % odkazů. Spolehlivě oba prohlížeče dohledaly 201 odkazů, ale 60 odkazů bylo zobrazeno jen v jednom případě, což odpovídá 29,9 %.

Google tedy ve stejném čase generuje výsledky dynamicky a nepředkládá kompletní výsledky, což je nutné při vytěžování dat pomocí tohoto vyhledávače vést v patrnosti.

Vzhledem k zjištěnému je tedy při vytěžování otevřených zdrojů na místě nespoléhat jen na výsledky z Google a případně využít dalších vyhledávačů (např. Duckduckgo.com). Alternativou je provedení opakovaných dotazů z různých prostředí. Situace může být částečně ovlivněna snahou Google o personalizaci výsledků hledání. Prohlížeč č. 1 užívám každodenně, prohlížeč č. 2 neužívám takřka vůbec.

7 ZÁVĚR

Ačkoli bych se v diplomové práci chtěl jednotlivým oblastem věnovat podrobněji, bylo nutné obsáhnout co nejširší spektrum problémů. Snažil jsem práci pojmout prakticky a provést rozbor prostředí, které je známé nejen odborné veřejnosti, ale také laikům. Mnoho lidí pocítilo na vlastní kůži nepříjemnosti, které může v prostředí Internetu způsobit lhostejnost nebo nedbalost při uveřejňování dat, a věřím, že s jejich názory bude má práce v mnohých ohledech korespondovat. Těm ostatním by práce mohla ukázat nový úhel pohledu a motivovat je k větší míře obezřetnosti a rozvaze při sdílení svého soukromí. Z jednotlivých demonstrací technik a praktických rozborů mohou čerpat také pracovníci bezpečnostních sborů, které by má práce mohla inspirovat.

Abych mohl práci napsat, nestačilo využít jen dosavadní zkušenosti. Naopak jsem musel prostudovat další materiály související s problematikou OSINT, k čemuž bych se pravděpodobně jinak nedostal. Bylo nutné jednotlivé dílčí metody zkoumat podrobněji a testovat je, nejen je běžně užívat.

Na základě poznatků, které z práce vyplynuly je možné v budoucnu pokračovat v dalším výzkumu prostředí. Jako největší problém se jeví možnost zneužívání archivů, jejichž existence je z hlediska ochrany osobních údajů velice problematická. Vhodné by bylo též zaměřit pozornost i na další portály, které by mohly s osobními údaji zacházet způsobem, který byl popsán v kapitole 6.1.

Výsledky provedeného výzkumu k internetovým archivům a komerčním webům jasně signalizují nedostatky, které dle mého názoru mohou kolidovat s legislativní úpravou a záměrem GDPR. Nedbalost a nezájem zamyslet se nad možnými důsledky uveřejňování osobních údajů, může v této oblasti způsobit škody, které již nikdy nepůjdou napravit. Hlavní problém přitom dle mého názoru pramení z toho, že si dotčené subjekty neuvědomují potenciální rizika zneužití jejich webů právě metodami OSINTu. Při nakládání s citlivými údaji by však zejména tato rizika měla být brána v potaz.

Jako zajímavost na závěr mohu uvést, že v rámci praktické části byl mj. nalezen inzerát, který mne zavedl na diskuzní fórum doktoronline.cz. Jeden z uživatelů zde nabízel prodej léku Tramal obsahujícího opioid tramadol. Lék je vázaný na lékařský předpis a jeho distribuce je tudíž nelegální. Zjištěné informace proto byly předány Policii ČR (k e-mailové adrese se podařilo dohledat telefonní číslo a pravděpodobně také jméno osoby, tudíž by neměl být problém věc objasnit). Případů prodeje léků na předpis prostřednictvím sítě Internet jsem zjistil celou řadu. Namátkově jsem prohlédl i další prodejce a dle očekávání zjistil, že většina z nich využívá zahraniční e-mailové služby známé pro jejich vyšší míru anonymity (Protonmail, Tutanota). Tito jistě využívají i dalších pseudonymizačních služeb jako jsou VPN (Virtual Private Network) nebo Tor (The Onion Routing) a jejich odhalení je tak značně ztížené.

8 SEZNAM POUŽITÝCH ZKRATEK

ADS – Alternate Data Stream (funkce souborového systému NTFS – New Technology File System, která umožňuje ukládání metadat souborů)

A. I. – Artificial intelligence (umělá inteligence)

API – Application Programming Interface (rozhraní pro programování aplikací)

CSV – Comma-separated value (čárkami oddělené hodnoty, formát dat, lze otevřít použitím MS Excel)

CTR – Click Through Rate (míra prokliku, poměr mezi počtem zobrazení a kliknutí)

CVV/CVC – Card Verification Value/Card Verification Code (ochranný prvek platební karty)

DIČ – Daňové identifikační číslo

DLP – Data Loss Prevention (systém pro ochranu dat)

EXIF – Exchangeable Image File Format (specifikace pro formát metadat fotografií)

GD – Grafická knihovna jazyka PHP

GDPR – General Data Protection Regulation (Obecné nařízení o ochraně osobních údajů)

GMT – Greenwich Mean Time (greenwichský střední čas, staré označení časových pásem)

GPS – Global Positioning Systém (globální družicový navigační systém)

HTML – Hypertext Markup Language (značkovací jazyk pro tvorbu webových stránek)

HZS ČR – Hasičský záchranný sbor České republiky

ICQ – I Seek You (program pro instant messaging)

IČO – Identifikační číslo osoby

IP – Internet Protocol (protokol pro komunikaci v počítačových sítích)

IRC – Internet Relay Chat (otevřený protokol pro textovou komunikaci)

MCDA – Multiple-Criteria Decision Analysis (multikriteriální rozhodovací analýza)

MD5 – Message-Digest (hashovací funkce libovolného vstupu vytvářející výstup fixní délky)

OSINT – Open Source Intelligence (zpravodajství z otevřených zdrojů)

PHP – Hypertext Preprocessor (dříve Personal Home Page, jazyk pro programování dynamických webových stránek a aplikací)

REGEX – Regular Expression (regulární výraz)

SEO – Search Engine Optimization (optimalizace pro vyhledávače)

SHA1 – Secure Hash Algorithm 1 (hashovací funkce libovolného vstupu vytvářející výstup fixní délky)

SSL/TLS – Secure Sockets Layer/ Transport Layer Security (kryptografické protokoly pro bezpečnou komunikaci v rámci sítě)

TOR – The Onion Routing (protokol pro komunikaci a pseudonymizaci připojení v rámci sítě Internet)

URL – Uniform Resource Locator (jednotná adresa zdroje, označuje doménu, umístění zdroje a protokol k přístupu)

VK – zkratka ruské sociální sítě VKontakte

VPN – Virtual Private Network (možnost bezpečné komunikace mezi počítači v síti Internet, které se chovají obdobně, jako by byly propojeny lokálně – přístup přes VPN tak mj. umožňuje vystupovat do Internetu prostřednictvím jiného PC)

ZoEK – Zákon č. 127/2005 Sb., o elektronických komunikacích

ZoZOÚ – Zákon č. 110/2019 Sb., o zpracování osobních údajů

9 SEZNAM POUŽITÉ LITERATURY

1. AKHGAR, Babak, Petra Saskia BAYERL a Fraser SAMPSON. *Open Source Intelligence Investigation: From Strategy to Implementation*. 1. New York: Springer International Publishing, 2017. ISBN 9783319476711.
2. COLQUHOUN, Cameron. A Brief History of Open Source Intelligence. *Bellingcat: the home of online investigations* [online]. Amsterdam: Bellingcat, 2016, July 14, 2016 [cit. 2020-01-28]. Dostupné z: https://www.bellingcat.com/resources/articles/2016/07/14/a-brief-history-of-open-source-intelligence/
3. SCHULTZ, Jeff. How Much Data is Created on the Internet Each Day? *Micro Focus* [online]. 2019, 08.06.2019 [cit. 2020-04-25]. Dostupné z: https://blog.microfocus.com/how-much-data-is-created-on-the-Internet-each-day/
4. ČERNOHLÁVKOVÁ, Kateřina. *Informační hygiena*. Brno, 2006. Dostupné také z: http://is.muni.cz/th/109574/ff_b/. Bakalářská práce. Masarykova univerzita, Filozofická fakulta.
5. *Joint Publication 2-0: Joint Intelligence* [online]. Suffolk (Virginia): Joint Chiefs of Staff, 2013. Dostupné také z: https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp2_0.pdf
6. ROZSYPAL, Michael a Michael ERHART. Českou dezinformační jedničkou je Aeronet. Věří mu zejména starší lidé, říká novinář. In: *Český rozhlas* [online]. Praha: Český rozhlas, 18. prosinec 2017 [cit. 2020-03-11]. Dostupné z: https://plus.rozhlas.cz/ceskou-dezinformacni-jednickou-je-aeronet-veri-mu-zejmena-starsi-lide-rika-6504360
7. STEPHENS, Stefanie. Learn about Evaluating Sources: CRAP Test. In: *CCConline Library* [online]. [2016] [cit. 2020-03-11]. Dostupné z: https://ccconline.libguides.com/c.php?g=242130&p=2185475

8. RIEGER, Kristy. Week of 3/07: Using the C.R.A.P. Test. In: *Marquette Library: Media Center* [online]. Chicago, Illinois, Mar 6, 2016 [updated Mar 11, 2016] [cit. 2020-03-18]. Dostupné z: https://sites.google.com/site/msriegerreads/coursework/7th-year-2/weekof307-usingthecraptest
9. GRANVILLE, Kevin. Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens. In: *The New York Times* [online]. New York, New York: The New York Times, 19. března 2018 [cit. 2020-02-15]. ISSN 0362-4331. Dostupné z: https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html
10. Co je GDPR. *Ministerstvo vnitra České republiky* [online]. Praha: Ministerstvo vnitra České republiky [cit. 2020-03-19]. Dostupné z: https://www.mvcr.cz/gdpr/clanek/co-je-gdpr.aspx
11. MIHULKOVÁ, Jitka a Martin KORNEL. Co je, co není a co bude osobní údaj podle GDPR. *Frank Bold advokáti* [online]. 10. 2. 2018 [cit. 2020-03-19]. Dostupné z: https://www.fbadvokati.cz/cs/clanky/541-co-je-co-neni-a-co-bude-osobni-udaj-podle-gdpr
12. Citlivé osobní údaje. In: *GDPR.cz: Obecné nařízení o ochraně osobních údajů prakticky* [online]. Praha [cit. 2020-03-19]. Dostupné z: https://www.gdpr.cz/gdpr/heslo/citlive-osobni-udaje/
13. Rozsudek Soudního dvora z 13. 5. 2014 Google Spain SL, Google Inc. proti Agencia Española de Protección de Datos (AEPD), Mario Costeja González, C-131/12 [2014] ECLI:EU:C:2014:317.
14. KOLOUCH, Jan. *CyberCrime*. Praha: CZ.NIC, z.s.p.o., 2016. CZ.NIC. ISBN 978-80-88168-15-7.
15. ŠVEC, Michal. Tzv. „právo být zapomenut“ a jeho uplatnění v praxi. In: *Epravo.cz* [online]. Praha: epravo.cz, 10. 6. 2015 [cit. 2020-03-19]. ISSN 1213-

189X. Dostupné z: https://www.epravo.cz/top/clanky/tzv-pravo-byt-zapomenut-a-jeho-uplatneni-v-praxi-98087.html

16. Odstranění obsahu na základě ochrany soukromí uživatelů z EU. *Google* [online]. [cit. 2020-03-19]. Dostupné z: https://www.google.com/webmasters/tools/legal-removal-request?complaint_type=rtbf
17. DOSKOČILOVÁ, Veronika. Jedete na dovolenou a nikdo není doma? Na sociální sítě nic nesdílejte. *Měšec.cz* [online]. Praha: Internet Info, 23. 8. 2018 [cit. 2020-03-20]. ISSN 1213-4414. Dostupné z: https://www.mesec.cz/aktuality/jedete-na-dovolenou-a-nikdo-neni-doma-na-socialni-site-nic-nesdilejte/
18. BLAKEMORE, Eve. How do spammers get my email address? In: *Microsoft Security Blog* [online]. Redmond (Washington): Microsoft, August 6, 2010 [cit. 2020-03-11]. Dostupné z: https://www.microsoft.com/security/blog/2010/08/06/how-do-spammers-get-my-email-address/
19. BRINKS, Melissa. The 'Nigerian Prince' Scam Is Actually 200 Years Old. *Ranker* [online]. Los Angeles (California): Ranker.com, February 5, 2019 [cit. 2020-05-11]. Dostupné z: https://www.ranker.com/list/history-of-nigerian-prince-trick/melissa-brinks
20. Nigerian scams. *ACCC: Australian Competition and Consumer Commission* [online]. Canberra: Australian Competition and Consumer Commission [cit. 2020-02-14]. Dostupné z: https://www.scamwatch.gov.au/types-of-scams/unexpected-money/nigerian-scams
21. LEONHARDT, Megan. 'Nigerian prince' email scams still rake in over $700,000 a year—here's how to protect yourself. *CNBC: make it* [online]. Englewood Cliffs (New Jersey): CNBC, Apr 18 2019 [cit. 2020-02-14]. Dostupné z: https://www.cnbc.com/2019/04/18/nigerian-prince-scams-still-rake-in-over-700000-dollars-a-year.html

22. Pwned websites. *';--have i been pwned?* [online]. [cit. 2020-02-14]. Dostupné z: https://haveibeenpwned.com/PwnedWebsites
23. HERN, Alex. WannaCry, Petya, NotPetya: how ransomware hit the big time in 2017. *The Guardian* [online]. London: The Guardian, 30 Dec 2017 [cit. 2020-02-14]. ISSN 1756-3224. Dostupné z: https://www.theguardian.com/technology/2017/dec/30/wannacry-petya-notpetya-ransomware
24. *Hunter* [online]. [cit. 2020-04-12]. Dostupné z: https://hunter.io/
25. Policejní virus dál straší Čechy. In: *Novinky.cz* [online]. Praha: Borgis, 4. 11. 2013 [cit. 2020-02-14]. Dostupné z: https://www.novinky.cz/Internet-a-pc/clanek/policejni-virus-dal-strasi-cechy-208535
26. JERIE, Ladislav. Hlášení o zablokování počítače policií jsou falešná, jedná se o vir. In: *Benešovský deník.cz* [online]. Praha: VLTAVA LABE MEDIA, 2. 1. 2014 [cit. 2020-02-14]. Dostupné z: https://benesovsky.denik.cz/zlociny-a-soudy/laj-hlaseni-o-zablokovani-pocitace-policii-jsou-falesna-jedna-se-o-vir-20140102.html
27. The State of Ransomware in the US: Report and Statistics 2019. *Emsisoft.com* [online]. New Zealand: Emsisoft, December 12th, 2019 [cit. 2020-02-14]. Dostupné z: https://blog.emsisoft.com/en/34822/the-state-of-ransomware-in-the-us-report-and-statistics-2019/
28. MAGDOŇOVÁ, Jana. Na nemocnici v Benešově útočil ruský virus Ryuk. Jermanová odmítá, že by někdo požadoval výkupné. *IRozhlas.cz* [online]. Benešov: Český rozhlas, 14. 1. 2020 [cit. 2020-02-14]. Dostupné z: https://www.irozhlas.cz/zpravy-domov/nemocnice-benesov-kyberneticky-utok-ransomware-vykupne-ochrana-osobnich-udaju_2001140615_cha
29. RYUK Ransomware Information. In: *Trend Micro* [online]. Tokyo: Trend Micro, 12 Dec 2019 [cit. 2020-02-14]. Dostupné z:

https://success.trendmicro.com/solution/1123892-ryuk-ransomware-information

30. CORFIELD, Gareth. Don't pay off Ryuk ransomware, warn infoseccers: Its creators borked the decryptor. *The Register* [online]. London: Situation Publishing, 10 Dec 2019 [cit. 2020-02-14]. Dostupné z: https://www.theregister.co.uk/2019/12/10/ryuk_decryptor_broken_latest_strain/
31. KOŽÍŠEK, Martin a Václav PÍSECKÝ. *Bezpečně n@ internetu: průvodce chováním ve světě online*. Praha: Grada Publishing, 2016. ISBN 978-80-247-5595-3.
32. *W3schools.com* [online]. [cit. 2020-05-15]. Dostupné z: https://www.w3schools.com/html/tryit.asp?filename=tryhtml_intro
33. ČERMÁK, Miroslav. Phishing: přichází nová generace phishingu. *Clever And Smart* [online]. 18. 12. 2011 [cit. 2020-02-14]. Dostupné z: https://www.cleverandsmart.cz/phishing-prichazi-nova-generace-phishingu/
34. Československá obchodní banka – certifikát. *ČSOB* [online]. [cit. 2020-05-15]. Dostupné z: https://www.csob.cz/
35. Pozor na phishing a podvodné e-maily. *mBank* [online]. Praha: mBank, 03-06-2014 [cit. 2020-05-11]. Dostupné z: https://www.mbank.cz/blog/post,505,pozor-na-phishing-a-podvodne-e-maily.html
36. ŠPAČEK, Michal. *Crackování hesel z úniku Mall.cz* [online]. In: . 2. ledna 2018 [cit. 2020-04-12]. Dostupné z: https://www.michalspacek.cz/crackovani-hesel-z-uniku-mall.cz
37. KELLEY, Diana a Seema KATHURIA. Spear phishing campaigns—they're sharper than you think. *Microsoft Security Blog* [online]. Redmond (Washington): Microsoft, December 2, 2019 [cit. 2020-05-11]. Dostupné z:

https://www.microsoft.com/security/blog/2019/12/02/spear-phishing-campaigns-sharper-than-you-think/

38. ČÍRTKOVÁ, Ludmila. *Moderní psychologie pro právníky: [domácí násilí, stalking, predikce násilí]*. Praha: Grada, 2008. Psyché (Grada). ISBN 978-80-247-2207-8. 9788024768861.
39. ČERNÁ, Alena. *Kyberšikana: průvodce novým fenoménem*. Praha: Grada, 2013. Psyché (Grada). ISBN 978-80-210-6374-7.
40. LOSEKOOT, Michelle a Eliška VYHNÁNKOVÁ. *Jak na sítě: ovládněte čtyři principy úspěchu na sociálních sítích*. Brno: Jan Melvil Publishing, 2019. Žádná velká věda. ISBN 978-80-7555-084-2.
41. RUTLEDGE, Patrice-Anne. *The Truth about Profiting from Social Networking*. London: Pearson Education, 2008. ISBN 9780789737885.
42. GLADWELL, Malcolm. *Bod zlomu: o malých příčinách s velkými následky*. 4. vydání, první v BizBooks. Brno: BizBooks, 2015. ISBN 978-80-265-0404-7.
43. BHAGAT, Smriti, Moira BURKE, Carlos DIUK, Ismail ONUR FILIZ a Sergey EDUNOV. Three and a half degrees of separation. In: *Facebook Research* [online]. February 4, 2016 [cit. 2020-02-17]. Dostupné z: https://research.fb.com/blog/2016/02/three-and-a-half-degrees-of-separation/
44. KREBS, Valdis. Social Network Analysis: An Introduction. *Orgnet* [online]. [cit. 2020-02-17]. Dostupné z: http://www.orgnet.com/sna.html
45. ALBRIGHT, Jonathan. The Graph API: Key Points in the Facebook and Cambridge Analytica Debacle. *Tow Center* [online]. A Medium Corporation, Mar 21, 2018 [cit. 2020-02-17]. Dostupné z: https://medium.com/tow-center/the-graph-api-key-points-in-the-facebook-and-cambridge-analytica-debacle-b69fe692d747
46. SYMEONIDIS, Iraklis, Pagona TSORMPATZOUDI a Bart PRENEEL. *Collateral damage of Facebook Apps: an enhanced privacy scoring model*. 2015. Dostupné také z: https://eprint.iacr.org/2015/456.pdf

47. SHU, Catherine. Changes to Facebook Graph Search leaves online investigators in a lurch. *TechCrunch* [online]. Verizon Media, June 11, 2019 [cit. 2020-02-18]. Dostupné z: https://techcrunch.com/2019/06/10/changes-to-facebook-graph-search-leaves-online-investigators-in-a-lurch
48. MARTINEAU, Paris. Facebook is tracking you on over 8.4 million websites. *The Outline: THE FUTURE* [online]. BDG Media, MAY—18—2018 [cit. 2020-02-18]. Dostupné z: https://theoutline.com/post/4578/facebook-is-tracking-you-on-over-8-million-websites?zd=1&zi=lvue4n6j
49. *One Million Tweet Map* [online]. [cit. 2020-02-18]. Dostupné z: https://onemilliontweetmap.com
50. BROWN, Joshua E. Study: On Facebook And Twitter Your Privacy Is At Risk—Even If You Don't Have An Account. In: *The University of Vermont* [online]. Burlington (Vermont): The University of Vermont, 01-21-2019 [cit. 2020-02-18]. Dostupné z: https://www.uvm.edu/uvmnews/news/study-facebook-and-twitter-your-privacy-risk-even-if-you-dont-have-account
51. NICOLAOU, Anna. How to become TikTok famous. *Financial Times* [online]. London: The Financial Times, NOVEMBER 8 2019 [cit. 2020-02-18]. Dostupné z: https://www.ft.com/content/dd7234e8-fcb9-11e9-98fd-4d6c20050229
52. KOPECKÝ, Kamil. Problém zvaný Tik Tok. *E-Bezpečí* [online]. Olomouc, 4. leden 2019 [cit. 2020-02-19]. ISSN 2571-1679. Dostupné z: https://www.e-bezpeci.cz/index.php/rizikove-jevy-spojene-s-online-komunikaci/socialni-site/1403-problem-zvany-tik-tok
53. DOČEKAL, Daniel. Vzestup a pád českých sociálních sítí. *POOH.CZ* [online]. 02/04/2011 [cit. 2020-04-11]. Dostupné z: https://pooh.cz/2011/04/02/vzestup-a-pad-ceskych-socialnich-siti/
54. ORIYANO, Sean-Philip a Michael GREGG. *Hacker techniques, Tools, and Incident Handling*. Sudbury, MA: Jones & Bartlett Publishers, 2010. ISBN 978-0-7637-9183-4.

55. *Archive.org: Wayback Machine* [online]. 14. 11. 1996 [cit. 2020-03-19]. Dostupné z: https://web.archive.org/
56. *Google* [online]. [cit. 2020-05-15]. Dostupné z: https://www.google.com
57. Ozdravný pobyt v Itálii 2019. *Základní škola Praha - Radotín* [online]. Praha, 2019 [cit. 2020-03-19]. Dostupné z: https://www.skola-radotin.cz/cs/skolni-druzina/akce-skolni-druziny/ozdravny-pobyt-v-italii-2019.html
58. TIKTOK: CO BYSTE MĚLI VĚDĚT O NEJRYCHLEJI ROSTOUCÍ SOCIÁLNÍ SÍTI DNEŠKA. *Internetem bezpečně* [online]. Karlovy Vary: you connected, 23.10.2019 [cit. 2020-03-19]. ISSN 2571-3736. Dostupné z: https://www.internetembezpecne.cz/tiktok-co-byste-meli-vedet-o-nejrychleji-rostouci-socialni-siti-dneska/
59. ČÍŽEK, Jakub a Filip ŠEDIVÝ. České Rajče je stále plné dětských nahotin. Student pomocí A.I. analyzoval miliony fotek. In: *Živě.cz* [online]. Praha: CZECH NEWS CENTER, 18. prosince 2019 [cit. 2020-03-19]. ISSN 1212-8554. Dostupné z: https://www.zive.cz/clanky/ceske-rajce-je-stale-plne-detskych-nahotin-student-pomoci-ai-analyzoval-miliony-fotek/sc-3-a-201672/default.aspx
60. VOKROUHLÍKOVÁ, Kateřina. RAJČE = PEDOFILŮ RÁJ! *Internetem bezpečně* [online]. Karlovy Vary: you connected, 18. 11. 2019 [cit. 2020-03-29]. ISSN 2571-3736. Dostupné z: https://www.internetembezpecne.cz/rajce-pedofilu-raj/
61. SKLENÁK, Vilém. *Data, informace, znalosti a Internet*. Praha: C.H. Beck, 2001. C.H. Beck pro praxi. ISBN 80-717-9409-0.
62. ABRAMS, Lawrence. Windows Alternate Data Streams. *BLEEPINGCOMPUTER* [online]. Melville (New York): Bleeping Computer, February 17, 2004 [cit. 2020-03-29]. Dostupné z: https://www.bleepingcomputer.com/tutorials/windows-alternate-data-streams/

63. KONONOW, Piotr. What is Metadata (with examples). *Dataedo* [online]. Gdańsk: Dataedo, 2018-09-16 [cit. 2020-03-29]. Dostupné z: https://dataedo.com/kb/data-glossary/what-is-metadata/
64. Email Header: How to Read and Analyze the Email Header Fields and Information about SPF, DKIM, SpamAssassin. *Arclab* [online]. Wiesent: Arclab Software [cit. 2020-04-01]. Dostupné z: https://www.arclab.com/en/kb/email/how-to-read-and-analyze-the-email-header-fields-spf-dkim.html
65. PETRESCU, Philip. Google Organic Click-Through Rates in 2014. *Moz* [online]. Seattle (Washington): Moz, October 1st, 2014 [cit. 2020-04-03]. Dostupné z: https://moz.com/blog/google-organic-click-through-rates-in-2014/
66. HARDWICK, Joshua. Google Search Operators: The Complete List (42 Advanced Operators). *Ahrefs blog* [online]. Singapore: Ahrefs, December 24, 2019 [cit. 2020-04-03]. Dostupné z: https://ahrefs.com/blog/google-advanced-search-operators/
67. KUČERA, Petr. Rodné číslo už nemá být povinnou součástí DIČ. *Peníze.cz: Největší web o osobních financích* [online]. Partners media, 27. 5. 2019 [cit. 2020-04-20]. ISSN 1213-2217. Dostupné z: https://www.penize.cz/podnikani/406614-rodne-cislo-uz-nema-byt-povinnou-soucasti-dic/
68. *Detail open data* [online]. [cit. 2020-05-15]. Dostupné z: https://www.detail.cz/
69. *Sauto.cz* [online]. [cit. 2020-04-29]. Dostupné z: https://www.sauto.cz/
70. *Sbazar.cz* [online]. [cit. 2020-04-29]. Dostupné z: https://www.sbazar.cz/
71. WHITTY, Monica T. Do You Love Me? Psychological Characteristics of Romance Scam Victims. In: *Warwick Research Archive Portal* [online]. Coventry, Velká Británie: The University of Warwick, 2017, 28 Jun 2017, s. 1-5 [cit. 2020-05-14]. DOI: 10.1089/cyber.2016.0729. Dostupné z:

http://wrap.warwick.ac.uk/89709/2/WRAP-do-you-love-psychological-romance-scam-Whitty-2017.pdf

72. VONDRUŠKA, Petr. *Metody a nástroje OSINT*. Praha, 2013. Diplomová práce. Bankovní institut vysoká škola Praha. Vedoucí práce Ing. Vladimír Beneš.
73. *Hack Forums* [online]. [cit. 2020-04-15]. Dostupné z: https://www.hackforums.net/
74. CIESLAK, Maria. Why Google Cache lies to you and what to do about it (if anything). *SearchEngineLand.com* [online]. Redding (Connecticut): Search Engine Land, October 15, 2018 [cit. 2020-04-07]. Dostupné z: https://searchengineland.com/why-google-cache-lies-to-you-and-what-to-do-306343
75. Jak změnit uživatelské jméno. In: *Seznam.cz Nápověda* [online]. Praha: Seznam.cz [cit. 2020-04-12]. Dostupné z: https://napoveda.seznam.cz/cz/jak-zmenit-uzivatelske-jmeno/
76. MARŠÍK, Zdeněk. *Jak se zadat jako nový spolužák na serveru spolužáci* [online]. [cit. 2020-04-12]. Dostupné z: http://kooperace.czechian.net/novyspol/novyspol.html
77. *Pedagogický sbor ZŠ* [online]. ZŠ a MŠ Kladno, Zd. Petříka 1756 [cit. 2020-04-12]. Dostupné z: https://www.2zskladno.cz/zs/ucitele.php
78. KAPUCIÁNOVÁ, Aneta. Služba Spolužáci.cz na konci srpna skončí. *SBLOG* [online]. Seznam.cz, 4. 4. 2018 [cit. 2020-04-12]. Dostupné z: https://blog.seznam.cz/2018/04/sluzba-spoluzaci-cz-na-konci-srpna-skonci/
79. Q&A - Vše, co jste chtěli vědět o bezpečnosti na MALL.cz. *MALL.CZ BLOG* [online]. Internet Mall, 27. srpna 2017 [cit. 2020-04-14]. Dostupné z: https://blog.mall.cz/o-nas/q-a-vse-co-jste-chteli-vedet-o-bezpecnosti-na-mall-cz-451.html
80. *Pastebin* [online]. [cit. 2020-04-12]. Dostupné z: https://www.pastebin.com/

81. SLÍŽEK, David. Unikátní rozsudek: Mall.cz musí zaplatit uživateli odškodnění za únik hesla. *Lupa.cz: server o českém Internetu* [online]. Internet Info, 23. 10. 2019 [cit. 2020-04-12]. ISSN 1213-0702. Dostupné z: https://www.lupa.cz/clanky/unikatni-rozsudek-mall-cz-musi-zaplatit-uzivateli-odskodneni-za-unik-hesla/
82. KRATOCHVÍL, Petr. Google hacking: cíl zaměřen. *Chip: Magazín informačních technologií* [online]. Praha: Vogel Publishing, 2009, (06), 102-105 [cit. 2020-05-14]. ISSN 1210-0684. Dostupné z: https://www.chip.cz/soubory/dokumenty/332acaee340c790375e1151169626eca.pdf
83. ŠÁMAL, Pavel. *Trestní zákoník: komentář*. Praha: C.H. Beck, 2010. Velké komentáře. ISBN 978-80-7400-109-3.

10 SEZNAM POUŽITÝCH OBRÁZKŮ

11 SEZNAM POUŽITÝCH TABULEK