

Sem vložte zadání Vaší práce.



**FAKULTA
INFORMAČNÍCH
TECHNologiÍ
ČVUT V PRAZE**

Bakalářská práce

Webová aplikace pro online web scraping

Jakub Drahoš

Katedra softwarového inženýrství
Vedoucí práce: Martin Podloucký

2. dubna 2019

Poděkování

Doplňte, máte-li komu a za co děkovat. V opačném případě úplně odstraňte tento příkaz.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 46 odst. 6 tohoto zákona tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou, a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla, a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu), licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 2. dubna 2019

.....

České vysoké učení technické v Praze

Fakulta informačních technologií

© 2019 Jakub Drahoš. Všechna práva vyhrazena.

Tato práce vznikla jako školní dílo na Českém vysokém učení technickém v Praze, Fakultě informačních technologií. Práce je chráněna právními předpisy a mezinárodními úmluvami o právu autorském a právech souvisejících s právem autorským. K jejímu užití, s výjimkou bezúplatných zákonných licencí a nad rámec oprávnění uvedených v Prohlášení na předchozí straně, je nezbytný souhlas autora.

Odkaz na tuto práci

Drahoš, Jakub. *Webová aplikace pro online web scraping*. Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2019.

Abstrakt

V několika větách shrňte obsah a přínos této práce v češtině. Po přečtení abstraktu by se čtenář měl mít čtenář dost informací pro rozhodnutí, zda chce Vaši práci číst.

Klíčová slova web scraping, extrakce dat, aplikace,

Abstract

Sem doplňte ekvivalent abstraktu Vaší práce v angličtině.

Keywords web scraping, data extraction, application, non-programmers,

Obsah

Úvod	1
1 Analýza a návrh	3
1.1 Web scraping	3
1.2 Právní aspekt	5
1.3 Analýza konkurence	11
1.4 Specifikace požadavků	17
2 Realizace	21
Závěr	23
Literatura	25
Literatura	27
A Seznam použitých zkratk	31
B Obsah přiloženého CD	33

Seznam obrázků

1.1	ParseHub	12
1.2	Octoparse	13
1.3	WebScraper	14
1.4	Dexi.io	15
1.5	Data Scraper	17

Úvod

Cíle práce

Hlavním cílem této práce je návrh a tvorba webové aplikace, která bude umožňovat uživatelům extrahovat požadovaná data z libovolné stránky v reálném čase bez jakékoli nutné znalosti programování. Při specifikaci požadavků tohoto softwaru se přihlídně k analýze stávajících řešení, jež je vedlejším cílem této práce.

Neméně důležitou součástí práce tvoří dodržení klasického vývojového cyklu softwarového projektu – analýza, design, implementace a testování.

Klíčovým aspektem aplikace je též *přehlednost a jednoduchost uživatelského rozhraní* – důraz bude kladen na intuitivní a rychlé ovládání.

Naopak v rozsahu této práce není tvorba web crawlera ani žádného jiného podobného mechanismu, jenž by systematicky procházel danou oblast webu.

Motivace

Přínos? Proč je téma důležité/aktuální? Komu tím jako prospějete? Proč tohle téma?

Členění práce

V kapitole 1 této práce je představen

Analýza a návrh

1.1 Web scraping

Web scraping (nebo také *web harvesting*, *web data extraction*) je technika získávání nejrůznějších dat z webových stránek. Nejčastěji se v tomto kontextu jedná o automatizovaný proces strojového zpracování a získávání dat, nicméně může jít i o manuální extrakci zadanou uživatelem skrze nějaký software (jako je tomu právě v našem případě). [citace z Wiki - web scraping - nebo radši ne...]

Často se také v souvislosti s pojmem web scraping používá spojení *web crawler* (nebo také *bot*, *spider*, *spiderbot*). Jedná se o automatizovaný software, který systematicky prochází danou oblast webu a během toho vytěžuje kýžená data. Jak již bylo řečeno v úvodu, touto částí web scrapingu se práce nebude zabývat.

1.1.1 Krátce z historie

Historie web scrapingu sahá k samým počátkům internetu (World Wide Web, 1989). Prvním webovým robotem, který byl vyvinut na MIT k měření velikosti webu, byl World Wide Web Wanderer (napsaný v jazyce Perl) z roku 1993. [citace z Wiki - World Wide Web Wanderer]

O něco později, v roce 2000, se ve velkém začala používat webová APIs – lidé mohli získávat čistá data přímo od serveru a scraping se tak stal o hodně jednodušším.

Dalším milníkem v historii web scrapingu je rok 2004, kdy byla vydána knihovna pro parsování HTML a XML dokumentů BeautifulSoup pro programovací jazyk Python. Ta je do dnes považována za nejsložitější a nejpokročilejší knihovnu pro web scraping.

Za zmínku stojí určitě i rok 2006, kdy je datován příchod vizuálního web scrapingu, tedy techniky, kdy uživatel skrze rozhraní aplikace označí klikáním myši, z kterých oblastí webové stránky chce extrahovat data. Tímto se otevřely

dveře web scrapingu pro všechny. [citace z <https://www.octoparse.com/blog/web-scraping-introduction>]

1.1.2 Techniky

Technik, jak z webové stránky získat data existuje mnoho, podívejme se alespoň na některé z nich:

- vyhledávání na základě textové shody – např. pomocí UNIX nástroje `grep` nebo regulárních výrazů
- HTML parsování – základní a stále ještě nejpoužívanější technika extrakce dat; informace jednoduše získáváme z HTML elementů, popř. pomocí tříd nebo id
- počítačové vidění, strojové učení, zapojení umělé inteligence – snaha napodobit způsob, jakým vidí a zpracovává webovou stránku člověk; podobný přístup zkouší např. projekt Diffbot
- vizuální web scraping – jak již bylo zmíněno výše, požadovaná data se musí ručně naklikat skrze rozhraní nějaké aplikace (značně to však usnadňuje např. hledání podobných prvků na základě prvních pár kliknutí)
- manuální vyhledávání a stahování dat (někdy nazývané také *copy-paste*)

1.1.3 Využití web scrapingu

Podob pro uplatnění scrapování dat z webu je nespočet, a to obzvlášť v dnešní době, kdy se velikost všech dat na celém internetu pohybuje v řádech Zettabajtů (1024^7 B). [citace z <https://www.nodegraph.se/big-data-facts/>]. Mezi ty hlavní patří:

- získání kontaktních informací (např. e-mail) pro marketingové účely
- indexování webových stránek (jako příklad můžeme uvést známý GoogleBot)
- data mining – proces hledání vzorců ve velkých datových setech [odkaz na Wiki]
- monitorování různých proměnných (např. sledování cen nebo hodnocení produktů)
- recyklace již někdy použitých dat za účelem vytváření „nového“ obsahu
- analýza a zpracování dat k výzkumným účelům

1.2 Právní aspekt

Podrobná právní analýza celé problematiky web scrapingu by vydala na samostatnou diplomovou práci, a tak se pokusím pouze shrnout základní body a poskytnout čtenáři alespoň náhled do této oblasti.

1.2.1 Odpovědnost v případě protiprávního jednání

Ve chvíli, kdy se uživatel dopustí protiprávního jednání při využívání určitého nástroje, odpovědnost jednoznačně spočívá na něm, nikoliv na tvůrci aplikace. Zároveň je třeba upozornit, že autor softwaru odpovídá v situaci, kdy je software primárně určen k protiprávnímu jednání. Lze tedy jen doporučit, aby byl uživatel prokazatelně seznámen s odpovědností za užívání softwaru v souladu s právem.[1]

1.2.2 Základní pojmy

Terms of Service (někdy také *Terms of Use*, *Terms and Conditions*) je soubor pravidel sepsaný provozovatelem služby a říká, jak se uživatel smí chovat při užívání dané služby (v kontextu této práce se jedná o webové stránky) a co naopak dělat nesmí.

Browsewrap je jeden ze způsobů dohody mezi dvěma stranami kontraktu. Není nutná žádná přímá interakce s uživatelem, ať už jde o souhlas či nesouhlas. Místo toho je na webové stránce (nejčastěji v dolní části) umístěna krátká zpráva informující, že pouhým procházením daného webu souhlasí s podmínkami používání (Terms of Service). Ty jsou umístěny na samostatné stránce, na kterou vede odkaz, jenž je součástí této zprávy. U tohoto způsobu je těžké posoudit, zdali je tu jasný projev vůle, a tak je jeho vymahatelnost sporná a liší se případ od případu.[2]

Clickwrap je oproti browse-wrap daleko lépe vymahatelný, neboť je uživatel nucen přímo vyjádřit souhlas či nesouhlas (kliknutím na tlačítko, zaškrtnutím políčka) se všemi uvedenými podmínkami, a to *před* použitím dané služby. Tím je zde jasně určen projev vůle. Podmínky jsou stejně jako v případě browse-wrap často umístěny na samostatné stránce a je k nim uveden pouze odkaz, i když někdy je k dispozici i celé jejich znění. Jedná se o tzv. *ber nebo nech být* smlouvu – „*Ber nebo nech být smlouva, také nazývána adhezní smlouva, říká, že smluvní podmínky nemůžou být vyjednávány.*“ [3, překlad autora].[4]

CFAA

1.2.3 Obsah na webových stránkách a jeho možné využití

Dle [1] může být obsah chráněn zejména jako:

- projev osobní povahy
 - zde lze připomenout z občanského zákoníku – „*Nikdo nesmí zasáhnout do soukromí jiného, nemá-li k tomu zákonný důvod. Zejména nelze bez svolení člověka narušit jeho soukromé prostory, sledovat jeho soukromý život nebo pořizovat o tom zvukový nebo obrazový záznam, využívat takové či jiné záznamy pořízené o soukromém životě člověka třetí osobou, nebo takové záznamy o jeho soukromém životě šířit. Ve stejném rozsahu jsou chráněny i soukromé písemnosti osobní povahy.*“ [5, § 86]
 - do této kategorie můžou spadat třeba i komentáře uživatelů na internetovém fóru
- autorské dílo (včetně databáze, viz níže); z autorského zákona lze zmínit:
 - volné užití – „*Za užití díla podle tohoto zákona se nepovažuje užití pro osobní potřebu fyzické osoby, jehož účelem není dosažení přímého nebo nepřímého hospodářského nebo obchodního prospěchu, nestanoví-li tento zákon jinak.*“ [6, § 30 odst. 1]
 - citaci – „*Do práva autorského nezasahuje ten, kdo*
 - a) *užije v odůvodněné míře výňatky ze zveřejněných děl jiných autorů ve svém díle,*
 - b) *užije výňatky z díla nebo drobná celá díla pro účely kritiky nebo recenze vztahující se k takovému dílu, vědecké či odborné tvorby a takové užití bude v souladu s poctivými zvyklostmi a v rozsahu vyžadovaném konkrétním účelem,*
 - c) *užije dílo při vyučování pro ilustrační účel nebo při vědeckém výzkumu, jejichž účelem není dosažení přímého nebo nepřímého hospodářského nebo obchodního prospěchu, a nepřesáhne rozsah odpovídající sledovanému účelu;**vždy je však nutno uvést, je-li to možné, jméno autora, nejde-li o dílo anonymní, nebo jméno osoby, pod jejímž jménem se dílo uvádí na veřejnost, a dále název díla a pramen.*“ [6, § 31 odst. 1]

- osobní údaje (čl. 2 GDPR)

Je tedy možné shrnout, že použití pro osobní účely (resp. domácí činnosti) je v zásadě neomezené. Za vyzdvižení však stojí věta z odstavce o volném užití: „... jehož účelem není dosažení přímého nebo nepřímého hospodářského nebo obchodního prospěchu ...“ – když použiji získaná data na mém osobním blogu, kde ale mám určitou formu výděлку třeba v podobě reklamy, mohu se již dopouštět protiprávního jednání.

Důležité je dát si velký pozor také při zpracování a využití osobních údajů, které upravuje čl. 2 GDPR – toto téma by samo vydalo na několik desítek

stránek, a tak se jím v této práci nebudu zabývat a je zde zmíněno jen pro úplnost.

V neposlední řadě je třeba poznamenat, že při vytěžování webu není podstatné, jakým způsobem k získání obsahu došlo (zdali prostřednictvím automatizovaného nebo manuálního postupu). Důležité je, jestli tak činím v souladu běžným, očekávaným a přiměřeným použitím. (Otázkou je, jestli by bylo vůbec vymahatelné, pokud by měla stránka přímo ve svých Terms of Service uvedeno, že si nepřeje automatizované procházení, nehledě na to, jestli zároveň probíhá extrakce dat či nikoliv. Taková podmínka by mohla být brána jako neadekvátní a nepřiměřená.) Rozhodně lze ale doporučit provádět scraping webových stránek se souhlasem jejich provozovatele (poskytovatele).[1]

1.2.4 Obsah chráněný autorským zákonem

Autorské právo chrání na internetu různý obsah, zejména budou chráněny různé články, obrázky, videa atd. Vždy však bude muset naplňovat znaky autorského díla, tj. bude muset být „... *jedinečným výsledkem tvůrčí činnosti autora a být vyjádřen v jakékoli objektivně vnímatelné podobě včetně podoby elektronické, trvale nebo dočasně, bez ohledu na jeho rozsah, účel nebo význam.*...“ [6, § 2 odst. 1].

Taková kritéria však může splňovat i obsah, který by se na první pohled vůbec nemusel zdát chráněný autorským zákonem – jako příklad může posloužit celkové rozvržení stránky (neboli *layout*), které ponese určitý prvek originality a bude na první pohled asociovatelné s danou webovou stránkou.

Naopak výše zmíněnou definici určitě nesplňují různá počítačem generovaná data, tedy například logy chráněné autorským zákonem nebudou.

1.2.5 Web scraping a zvláštní práva pořizovatele databáze

Dle [1] žádný zvláštní zákon věnující se výslovně vytěžování webových stránek neexistuje. Z naší právní úpravy je tomu nejbližší úprava zvláštního práva pořizovatele databáze (hlava III autorského zákona), která definuje, co to je databáze – „*Databází je pro účely tohoto zákona soubor nezávislých děl, údajů nebo jiných prvků, systematicky nebo metodicky uspořádaných a individuálně přístupných elektronickými nebo jinými prostředky, bez ohledu na formu jejich vyjádření.*“ [6, § 88]. Za nezávislé se v tomto kontextu považují prvky, „*které lze od sebe oddělit, aniž by tím byl dotčen jejich informační, literární, umělecký, hudební nebo jiný obsah*“ [7].

Dále jsou upraveny některé způsoby užití databáze v souladu s autorským zákonem:

- Omezení zvláštního práva pořizovatele databáze – „*Do práva pořizovatele databáze, která byla zpřístupněna jakýmkoli způsobem veřejnosti, nezasahuje oprávněný uživatel, který vytěžuje nebo zužitkovává kvalitativně nebo kvantitativně nepodstatné části obsahu databáze nebo její*

části, a to k jakémukoli účelu, za podmínky, že tento uživatel databázi užívá běžně a přiměřeně, nikoli systematicky či opakovaně, a bez újmy oprávněných zájmů pořizovatele databáze, a že nezpůsobuje újmu autorovi ani nositeli práv souvisejících s právem autorským k dílům nebo jiným předmětům ochrany obsaženým v databázi.“ [6, § 91]

- Bezúplatné zákonné licence – „Do práva pořizovatele jím zpřístupněné databáze též nezasahuje oprávněný uživatel, který vytěžuje nebo zužitkovává podstatnou část obsahu databáze

- a) pro svou osobní potřebu; ustanovení § 30 odst. 3 zůstává nedotčeno,
- b) pro účely vědecké nebo vyučovací, uvede-li pramen, v rozsahu odůvodněném sledovaným nevýdělečným účelem, a
- c) pro účely veřejné bezpečnosti nebo správního či soudního řízení.

“[6, § 92]

Tedy pokud využívám databázi čistě pro osobní potřebu či pro vědecké nebo vyučovací účely, kdy uvedu zdroj, je vše v pořádku. Taktéž pokud využívám pouze nepodstatnou část obsahu databáze, a pokud tak dělám v souladu s běžným, očekávaným a přiměřeným použitím, nikoliv systematicky a opakovaně, je v pořádku využití dokonce k jakémukoli účelu. Důležité je zde ale spojení *oprávněný uživatel* – v každém případě musím mít k datům autorizovaný přístup.

1.2.6 Důležité soudní případy v Evropě

Rozbor jednotlivých soudních případů je velice důležitý, neboť právě tím vzniká precedent, podle něhož se mohou soudy řídit při posuzování nových žalob a soudních rozepří. Stejně tak se o ně mohou opírat žalobci i obhájci a pro širokou veřejnost může být více předvídatelné, jak by se obdobné problémy mohly řešit v budoucnu.

Mezi nejvýznamnější rozhodnutí na půdě Soudního dvora Evropské unie patří:

C-444/02 kdy společnost Fixtures Marketing Limited (dále jen "Fixtures") žalovala společnost Organismos Prognostikon Agonon Podosfairou AE (dále jen "OPAP") kvůli opakovanému vytěžování rozpisů ligových soutěží ve fotbale v Anglii (které vytvářela, sestavovala a zveřejňovala společnost Fixtures) a jejich následnému užití na webových stránkách společnosti OPAP.[8]

Výsledkem jednání bylo, že i *rozpis fotbalových utkání je databází* ve smyslu čl. 1 odst. 2 směrnice 96/9 a jako takový může být předmětem ochrany, kdy pořizovatel databáze má právo zabránit vytěžování nebo zužitkování dat. (ačkoliv u společnosti Fixtures nebyl prokázán podstatný vklad, jenž by mohl odůvodnit poskytnutí ochrany).[9]

C-30/14 kdy společnost Ryanair Ltd. žalovala společnost PR Aviation BV kvůli automatizovanému sběru dat o cenách, letech a letových řádech, jež byly volně přístupné spotřebitelům z webových stránek Ryanair Ltd. K přístupu musel uživatel potvrdit souhlas (viz clickwrap) se všeobecnými podmínkami, ve kterých byl mimo jiné výslovně zakázán způsob, jakým data využívala PR Aviation BV. Tyto data pak společnost PR Aviation BV používala ke srovnávání cen na svém vlastním portálu.[10]

Soudní dvůr rozhodl ve prospěch Ryanair – *jestliže se autor databáze, která není chráněna autorským zákonem nebo právem pořizovatele, rozhodne poskytnout souhlas s jejím využitím, nic mu nebrání ve stanovení smluvních podmínek, jež by omezily používání databáze ze strany třetích osob, aniž by přitom bylo dotčeno použitelné vnitrostátní právo*.[10]

1.2.7 Důležité soudní případy ve Spojených státech

Úprava v USA je postavena na jiné právní úpravě, a závěry tak nejsou automaticky přenositelné do našeho práva (což ovšem nevylučuje možnost použití těchto rozhodnutí pro účely argumentace). Představíme si dva nejdůležitější případy v právní historii web scrapingu, se kterými mimo jiné vyvstávají otázky jako „Komu doopravdy patří data uživatelů na internetu?“ a „Kdo má rozhodovat o přístupu k veřejným datům, jestli o tom vůbec má někdo rozhodovat?“.

hiQ v. LinkedIn kdy analytická společnost hiQ automatizovaně extrahovala veřejně dostupná (bez nutnosti registrace účtu) data z profilů uživatelů sítě LinkedIn. Ta společnosti hiQ zaslala tzv. *cease and desist letter* a požadovala okamžité ukončení činnosti ze strany hiQ pod pohrůžkou porušení CFAA.[11]

K dispozici je zatím pouze předběžné opatření ve prospěch hiQ, které říká, že *nelze zabránit přístupu k veřejně dostupným informacím* – k porušení CFAA (jako neoprávněný přístup) by došlo pouze v případě obehnutí systému autentizace. Ze strany společnosti LinkedIn padl argument, že automatizovaný přístup k veřejným datům není to samé jako normální „osobní“ užití, stejně jako je odlišné dlouhodobé sledování osoby přes GPS zařízení od letmého potkávání [12]. hiQ mimo jiné argumentoval i tím, že sociální sítě jsou novodobá veřejná fóra (místa pro veřejné projevy, zejména politického charakteru[13]) a rozhodování o udělení přístupu k nim (ať už ze strany státu či soukromé společnosti) je porušením svobody slova a vyjadřování, a tím i samotné ústavy státu Kalifornie.[14]

Facebook v. Power Venture (Vachani) kdy společnost Power Venture (CEO Steven Suraj Vachani), jež umožňovala agregovat různé sociální sítě a používat všechny jejich klíčové funkce z jednoho místa, byla žalována sociální sítí Facebook kvůli neoprávněnému vytěžování profilů a zasílání e-mailů uživatelům, které úmyslně vypadaly jako odeslané společností Facebook (údajně

porušení CFAA, CAN-SPAM a California Penal Code, sekce 502). Facebook zaslal Power Venture *cease and desist letter* a snažil se zamezit přístupu k datům pomocí blokování IP adres. Power Venture přesto pokračoval ve své činnosti.[15]

Soud rozhodl ve prospěch společnosti Facebook – jednání Power Venture bylo v rozporu se zákonem. Jako hlavní argument posloužil fakt, že Power Venture pokračoval ve svém jednání i po explicitním odejmutí oprávnění přístupu k datům společnosti Facebook. Zde je ale vhodné uvést, že Power Venture *přistupoval k profilům uživatelů pouze s jejich vlastním souhlasem*, ač bez souhlasu Facebooku.[16]

1.2.8 Shrnutí

V předchozích sekcích jsme si představili hlavní právní pojmy týkající se problematiky web scrapingu a nyní je na čase shrnout naše poznatky. Při získávání dat z internetových stránek může nastat hned několik komplikací z právního hlediska, na které by se autor takového softwaru měl připravit.

Obsah může být chráněn autorským zákonem, pokud nabývá určitých rysů – zejména se musí jednat o výsledek tvůrčí činnosti autora. Zároveň pod tuto oblast mohou spadat věci jako je obyčejná databáze, způsob, jakým jsou určitá data rozvrstvena a uspořádána na stránce nebo třeba rozpis fotbalových utkání. Další kategorie, do kterých data mohou spadat, jsou osobní údaje a projevy osobní povahy. Při jejich zpracování je nutné postupovat přesně podle stanovených pravidel v příslušných zákonech a právních úpravách. A i když data nejsou chráněna žádným zvláštním zákonem, stále je nutné řídit se smluvními podmínkami, které se mohou vztahovat na jakýkoliv obsah.

Tím nejkritičtějším místem každého web scrapingu je ale způsob využití samotných dat. Ve směs je možné říci, že pokud používám data čistě pro svoji osobní/domácí potřebu, pro vědecké nebo pedagogické účely (kde má ale každé užití své podstatné náležitosti) a bez účelu dosažení hospodářského prospěchu, s největší pravděpodobností se nedopustím žádného protiprávního jednání. Vždy je ale tím nejbezpečnějším řešením kontaktovat provozovatele dané stránky a na detailech se dohodnout.

Nutno také podotknout, že celá tato oblast je relativně nová a zatím neexistuje jednotný právní precedent, podle kterého by se soudy mohli při posuzování jednotlivých případů řídit. Proto také člověk může nabýt pocitu, že i když se jedná o často velmi podobné případy, výsledky soudních sporů jsou diametrálně odlišné. To se ale může změnit s případem *hiQ v. LinkedIn*, který je pravděpodobně tím největším milníkem v právní historii web scrapingu – pokud se výsledek ještě zvrátí ve prospěch společnosti LinkedIn, mohlo by to znamenat velké omezení otevřeného přístupu k informacím pro všechny.

1.3 Analýza konkurence

První skupinou, na kterou můžeme při hledání na internetu narazit, jsou společnosti, které nabízejí zákazníkům kompletní péči v rámci extrakce dat. Cílí především na velké korporace, jimž postaví scrapovací nástroj přesně na míru, který poté také hostují a spravují. Zákazník tedy dostane data a o nic víc se již nemusí starat. Jako příklad lze jmenovat třeba ContentGrabber, Mozenda a další.

Pro tuto práci mnohem relevantnější kategorií je konkurenční nabídka nástrojů poskytujících uživatelům rozhraní k web scrapingu. Budu se zaměřovat pouze na takové nástroje, které nevyžadují jakoukoli znalost programování – tedy žádné knihovny, API a nástroje pro budování vlastních scraperů.

Mezi ty největší představitele patří ParseHub, Octoparse, WebScaper, Data Scraper a Dexi.io. Čtyři ze zmíněných nástrojů jsou volně dostupné (které mají však velmi omezenou funkcionalitu a pokročilejší operace se odepknou až s určitým platebním plánem – tzv. freemium model) a jeden poskytuje bezplatně pouze 7denní zkušební verzi.

Předtím, než bude možné jednotlivé nástroje porovnávat, je nutné určit kritéria, podle kterých lze hodnotit kvalitu daného nástroje. Především půjde o jednoduchost používání, celkovou přehlednost a rychlost, se kterou se uživatel dostane k požadovaným datům. Důležitý je také způsob výběru dat, možnosti exportu získaných dat, jak aplikace sama dokáže uživatele seznámit s používáním a také, v jaké formě se nástroj vůbec používá a čím se od ostatních odlišuje (ať už v pozitivním či negativním smyslu).

Pojďme se tedy na některé nástroje podívat blíže:

1.3.1 ParseHub

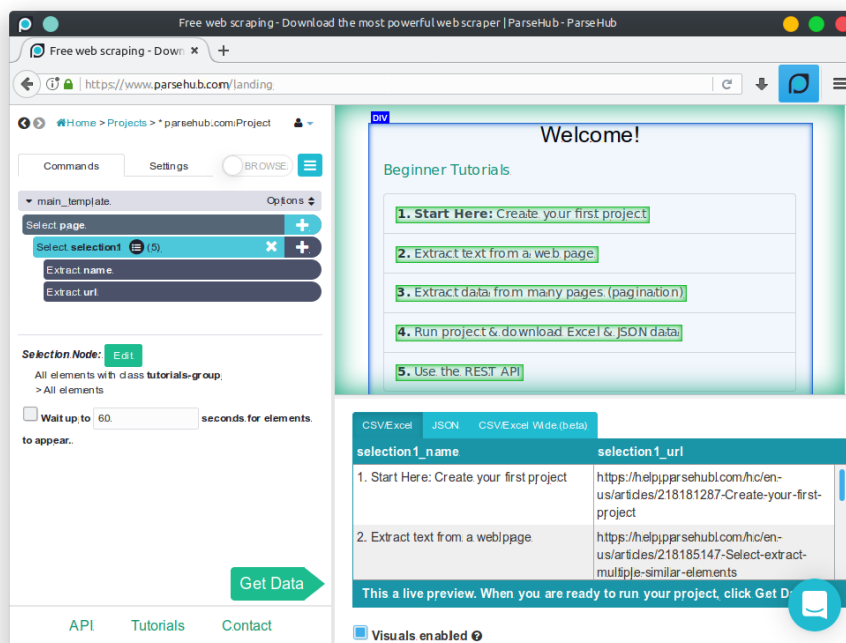
Výhody:

- výběr dat jak pomocí klikání (inteligentní hledání vzorců/podobností na základě prvních dvou kliknutí), tak pomocí XPath, regulárních výrazů nebo CSS selektorů
- aplikace obsahuje interaktivní tutoriál, který na jednoduchých příkladech ukáže, jak s nástrojem zacházet
- možnost získání dat různými formami - přes API, jako CSV/XLS, do GoogleSheets nebo do Tableau
- různé módy kliknutí (výběr, relativní výběr, kliknutí), zooming in/out na HTML elementy – když se uživatel netrefí (nebo ani trefit nemůže) přesně na požadovaný prvek, lze na něj lehce přejít pomocí této funkce
- automatická rotace IP adresy (tedy nedochází k blokování ze strany serveru)

1. ANALÝZA A NÁVRH

Nevýhody:

- nutnost stažení aplikace (ale je zde podpora pro Windows, Linux i Mac)
- aplikace je celkově těžkopádná, nemá moc přívětivé uživatelské rozhraní, ovládání působí nepřehledně a přehlcně – na uživatele se vyvalí hodně informací a možností najednou



Obrázek 1.1: ParseHub

1.3.2 Octoparse

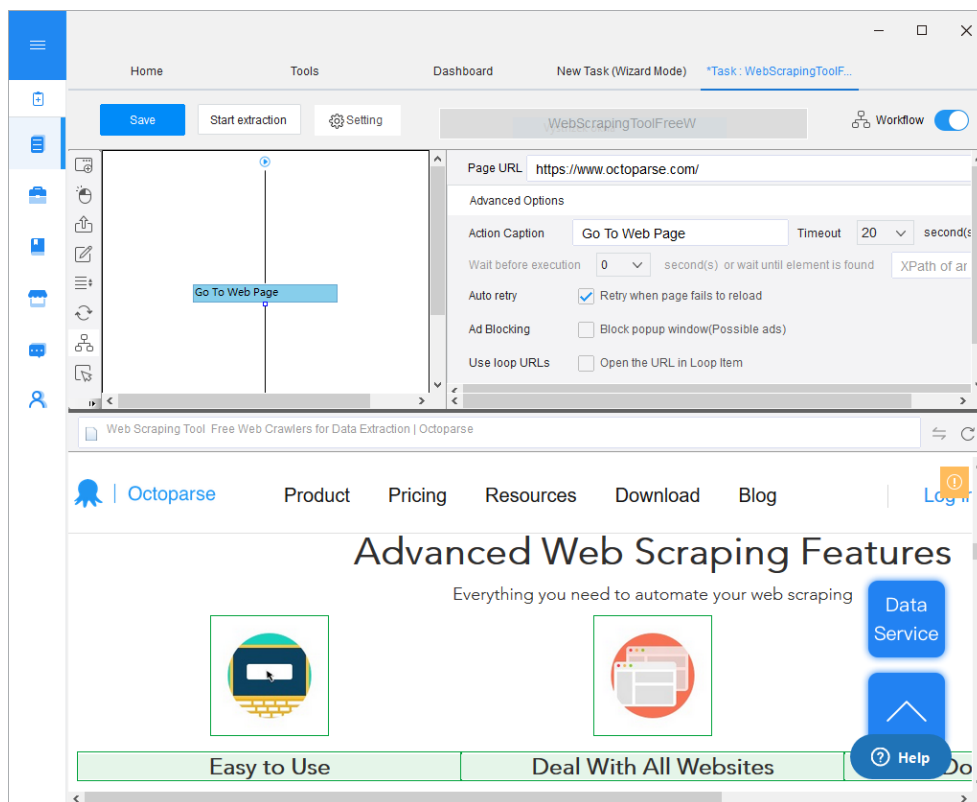
Výhody:

- výběr dat jak pomocí klikání (inteligentní hledání vzorců/podobností na základě prvních dvou kliknutí), tak pomocí XPath nebo regulárních výrazů
- nástroj obsahuje hotové šablony, které mohou velmi urychlit práci
- pestrá paleta možností (branch judgment, tvoření smyček apod.) – dá se vytvořit téměř jakákoli logika procházení webu a extrakce dat

- lehký způsob, jak scrapování automatizovat
- možnost řídit tasky přes API (a získávat tak data taktéž přes API); data jdou nahrát rovnou i do lokální databáze

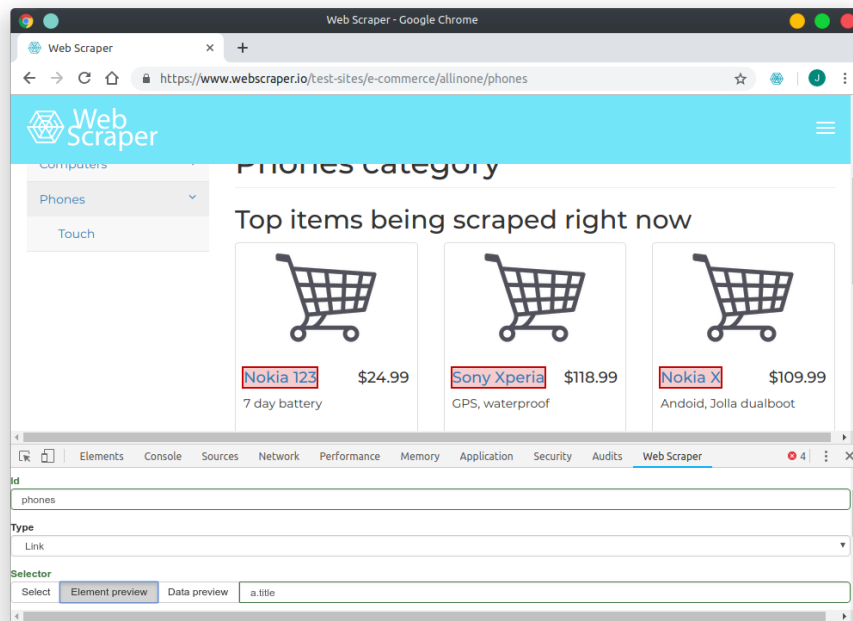
Nevýhody:

- nutnost stažení aplikace (která je navíc pouze pro Windows)
- těžkopádné a pomalé ovládání, neintuitivní rozhraní
- tutoriál je v podstatě nic neříkající
- připravených šablon je jenom pár a jsou velmi konkrétní



Obrázek 1.2: Octoparse

1. ANALÝZA A NÁVRH



Obrázek 1.3: WebScaper

1.3.3 WebScaper

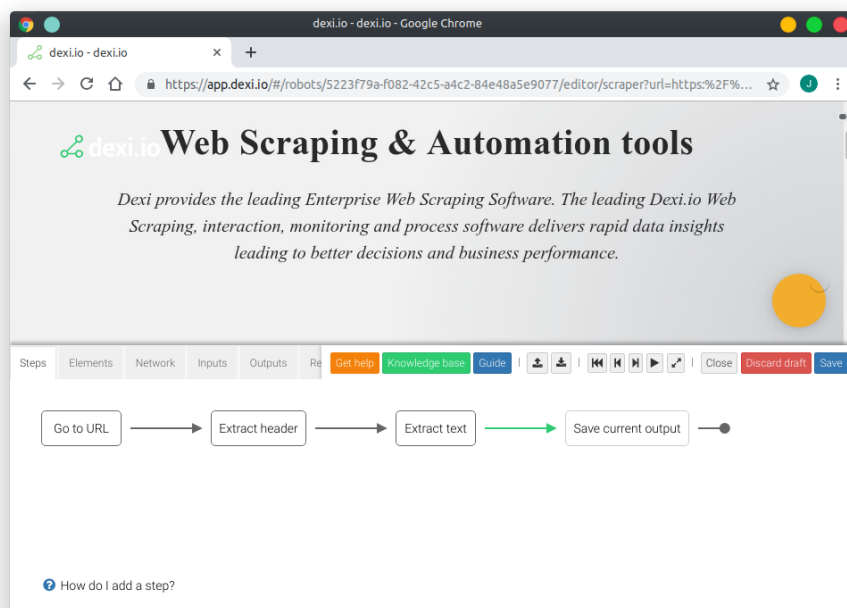
Výhody:

- jednoduchá instalace (jedná se pouze o rozšíření do prohlížeče Google Chrome); scrapování probíhá skrze vývojářskou konzoli
- výběr dat pomocí klikání (inteligentní hledání vzorců/podobností na základě prvních dvou kliknutí)
- tutoriály jsou formou videí – jednoduché, rychlé a naprosto postačující
- různé typy elementů, které vybíráme (text, odkaz, scroll down), takže lze celkem snadno projít celou doménu
- možnost získání dat různými formami – přes API, jako CSV/XLS nebo do Dropboxu
- klávesové zkratky při výběru elementů velmi usnadňují práci
- možnost využít jejich cloud k automatizaci celého procesu
- oproti konkurenci nabízí přehledné rozhraní, rychlé a jednoduché používání

Nevýhody:

- nutnost používat Google Chrome, což pro některé uživatele může být překážka
- nelze vyhledávat podle klíčových slov ani podle HTML nebo CSS, tudíž všechno se musí manuálně naklikat

1.3.4 Dexi.io



Obrázek 1.4: Dexi.io

Výhody:

- bez nutnosti stahování aplikace – vše se ovládá přes webové rozhraní
- výběr dat jak pomocí klikání (inteligentní hledání vzorců/podobností na základě prvních dvou kliknutí), tak pomocí HTML, CSS nebo textové shody
- mnoho návodů dostupných na stránkách, interaktivní rádce přímo při scrapování
- všechny možné druhy kliknutí, takže lze lehce projít celou doménu

- možnost exportovat data do CSV, JSON, XLS, získat přes API, poslat do Google Drive, Google Sheets nebo Amazon S3
- různé módy aplikace – scraping, crawler, pipes (skládání menších scrape botů) a autobot (extrahování z více stránek najednou se stejným rozložením); možnost takto automatizovat celý proces.
- nápomocné jsou různé addony (např. na obcházení Captchy)

Nevýhody:

- široká nabídka možností, a tak chvíli trvá, než se člověk zorientuje
- placený nástroj, zadarmo je dostupná pouze týdenní zkušební verze
- úvodní tutoriál je velmi strohý a žádné velké seznámení s nástrojem se nekoná

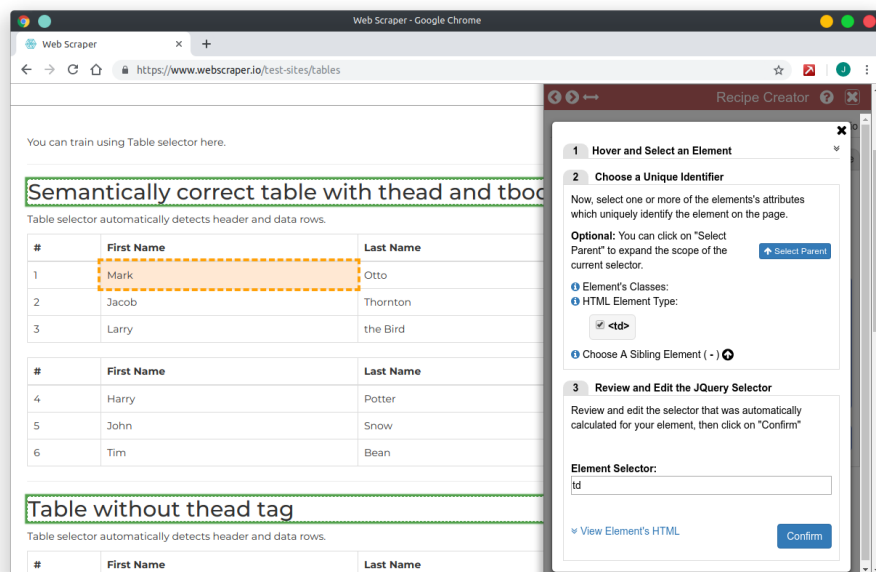
1.3.5 Data Scraper

Výhody:

- jednoduchá instalace (jedná se pouze o rozšíření do prohlížeče Google Chrome).
- velmi jednoduché ovládání a přehledné rozhraní
- výběr dat probíhá pomocí klikání
- klikáním se utváří JQuery selektor, který si uživatel může podle svého upravit a doladit tak drobné detaily, jež by jinak nutně zahltily uživatelské rozhraní (tedy je možné vyhledávat i podle HTML tagů, id, CSS selektorů – zkrátka vše, co umí klasické JQuery)
- různé druhy kliknutí
- možnost spustit na stránce libovolný JavaScriptový kód v rámci scrapování

Nevýhody:

- nutnost používat Google Chrome, což pro některé uživatele může být překážka
- oproti ostatním nástrojům se může zdát velmi chudý na různé funkce



Obrázek 1.5: Data Scraper

1.4 Specifikace požadavků

Jak jsme viděli v předchozí analýze konkurenčních nástrojů, největšími neduhy, které se prolínají napříč valnou většinou aplikací, jsou *těžkopádné uživatelské rozhraní*, *neintuitivní ovládání* a *rychlost* (nebo spíš pomalost), se kterou se uživatel dostane k požadovaným datům. Pro aplikaci, již se tato práce zabývá, bude klíčové se výše zmíněným nedostatkům vyhnout a nabídnout jejich přesný opak. Také jsme se přesvědčili, že nejpříjemnější cestou je celou aplikaci ovládat přes webové rozhraní *bez nutnosti stahování a instalace*.

Na druhou stranu se lze u konkurence i inspirovat. Za vyzdvižení stojí určitě *různé druhy výběru dat* – *klikání* přímo na stránce spolu s inteligentním hledáním podobných prvků jistě tvoří mocný mechanismus. Avšak je potřeba zajistit i ostatní způsoby výběru (jako je např. *textová shoda*, *HTML tagy*, *CSS selektory*) pro případ, kdy je pouhé klikání zdoluhavé či nevyhovující. Rovněž široký výběr způsobů exportu dat, intuitivní klávesové zkratky a zooming in/out na prvky může uživatelům zpříjemnit práci s nástrojem.

Neméně důležitou vlastností aplikace je také schopnost sebe sama kvalitně, ale svižně představit, *seznámit uživatele s používáním* a poskytnout mu alespoň pro začátek určité vodítko. Pro většinu ovládacích prvků by však mělo platit to stejné, co platí pro správný kód – měly by být tzv. *self-explanatory*. Tedy každému by mělo být na první pohled jasné, co který element dělá.

Pojďme si nyní všechny požadavky shrnout do několika bodů a rozdělit

na funkční a nefunkční:

1.4.1 Funkční požadavky

- uživatelské rozhraní se skládá z hlavní pracovní plochy, kde se bude nacházet uživatelem zadaná stránka a z postranního panelu, obsahující všechny ovládací prvky
- postranní panel skryje tlačítka, formuláře a ostatní elementy k ovládání aplikace do několika záložek – tímto se na uživatele nevyvalí velké kvantum informací a možností najednou; podle potřeby si každý rozbalí tu možnost, kterou potřebuje
- výběr dat bude probíhat těmito způsoby:
 - kliknutím myši na požadované elementy (na základě předchozích kliknutí se program pokusí označit všechny podobné prvky, výběr však půjde uživatelem zrušit)
 - na základě textové shody (uživatel jednoduše zadá text, jenž má být obsažen v extrahovaných datech)
 - pomocí HTML tagů (např. image, header, article), které se budou psát do textového pole
 - pomocí CSS selektorů (třídy, id, hodnota atributu, různé následnosti); k tomu poslouží formulář umožňující vše přehledně zadat
 - na ovládacím panelu nalezneme i tlačítka s hotovými akcemi představující šablonu pro nejpoužívanější operace (stažení všech obrázků ze stránky, všech emailových adres atd.)
- po kliknutí na určitý prvek se tento barevně označí; taktéž všechny již vybrané prvky budou barevně odlišeny, aby bylo jasné, co už je připraveno k extrakci a co ještě ne
- k dispozici bude přibližování/oddalování momentálního výběru pomocí ikony + a – (uživatel klikne na daný element a pomocí této funkce může traverzovat napříč zanořenými prvky oběma směry)
- na základě výběru dat uživatelem se vytvoří určitý filtr (textový řetězec), který může být ručně upraven – půjde tak o alternativu pro zkušenější uživatele, aniž bychom zanesli uživatelské rozhraní přehrší možností a celé ho tak znepřehlednili
- získaná data půjdou exportovat do formátů JSON, CSV, XLS, pokud se bude jednat o text; v případě obrázků, videí nebo zvukových souborů poskytne aplikace výstup v zabaleném archivu ZIP

1.4.2 Nefunkční požadavky

- půjde o webovou aplikaci běžící v internetovém prohlížeči, tedy nebude nutná žádná instalace
- program se bude skládat ze dvou částí:
 - *frontend* – kód, který poběží u klienta v prohlížeči; představuje celé uživatelské rozhraní aplikace
 - *backend* – kód, který poběží na serveru; bude naslouchat požadavkům a zpracovávat je; zde bude probíhat samotná extrakce dat
- aplikace cílí primárně na celkový zážitek uživatele – grafické rozhraní bude přehledné a co nejjednodušší, ovládání intuitivní
- čas, za který se uživatel dostane k požadovaným datům (tedy čas, který stráví vybíráním dat; nepočítáme čas potřebný ke stažení), bude co nejmenší

1.4.3 Nice to have požadavky

V předchozích dvou sekcích jsme si shrnuli, jaké požadavky by naše aplikace v každém případě měla splňovat a bez nichž by neměla vůbec být uvedena k dispozici uživatelům. Pak tu máme ale také požadavky, které rozhodně zlepšují celkovou kvalitu a pocit z nástroje samotného, avšak nejsou již pro nás vitální a pokud by se jejich implementace nepovedla, aplikace bude stále plně funkční a připravená k použití. Patří sem:

- uživatelské rozhraní aplikace nabídne intuitivní klávesové zkratky pro usnadnění práce – klikání s přidržanou klávesou **Ctrl** bude automaticky vybírat všechny podobné elementy; **Ctrl+** a **Ctrl-** obstará přibližování/oddalování momentálního výběru; ...
- export dat realizovatelný i do Google Sheets, Google Drive, Dropbox
- interaktivní tutoriál, který v rychlosti představí práci s nástrojem

Realizace

Závěr

Literatura

Literatura

- [1] VAŠÍČEK, Libor. [osobní sdělení]. Sídlo advokátní kanceláře Legal Partners, s.r.o., Záhořanského 1944/4, Praha 2. 29. března 2019.
- [2] KUNKEL, R. G. Recent developments in shrinkwrap, clickwrap and browsewrap licenses in the United States. *MurUEJL* [online]. 2002, roč. 9, č. 3 [cit. 1. 4. 2019]. Dostupné z <http://www5.austlii.edu.au/au/journals/MurUEJL/2002/34.html>.
- [3] What Is a Take It or Leave It Contract?. In: *UpCounsel* [online]. © 2019 UpCounsel, Inc [cit. 1. 4. 2019]. Dostupné z: <https://www.upcounsel.com/take-it-or-leave-it-contract>.
- [4] OBAR, J. A.; OELDORF-HIRSCH, A. The Clickwrap: A Political Economic Mechanism for Manufacturing Consent on Social Media. *Social Media + Society* [online]. 2018, roč. 4, č. 3 [cit. 1. 4. 2019]. DOI: 10.1177/2056305118784770. ISSN 2056-3051. Dostupné z <https://journals.sagepub.com/doi/10.1177/2056305118784770>
- [5] ČESKO. Zákon č. 89/2012 Sb, občanský zákoník. In: *Zákony pro lidi.cz* [online]. © AION CS 2010-2019 [cit. 30. 3. 2019]. Dostupné z: <https://www.zakonyprolidi.cz/cs/2012-89>.
- [6] ČESKO. Zákon č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon). In: *Zákony pro lidi.cz* [online]. © AION CS 2010-2019 [cit. 30. 3. 2019]. Dostupné z: <https://www.zakonyprolidi.cz/cs/2000-121>.
- [7] POLČÁK, Radim. *Právo informačních technologií*. Praha: Wolters Kluwer, 2018. 656 stran. ISBN 978-80-7598-045-8.
- [8] Věc C-444/02, stanovisko generální advokátky ze dne 8 června 2004 ve věci Fixtures Marketing Ltd proti Organismos prognostikon agonon podofairou AE (OPAP). In: *EUR-Lex* [online]. © European Union, 1998-

- 2019 [cit. 30. 3. 2019]. Dostupné z: <https://eur-lex.europa.eu/legal-content/CS/TXT/?uri=CELEX:62002CC0444>.
- [9] Věc C-444/02, rozsudek Soudního dvora (velkého senátu) ze dne 9. listopadu 2004 ve věci *Fixtures Marketing Ltd v. Organismos prognostikon agonon podosfairou AE (OPAP)*. In: *EUR-Lex* [online]. © European Union, 1998-2019 [cit. 30. 3. 2019]. Dostupné z: <https://eur-lex.europa.eu/legal-content/CS/TXT/?uri=CELEX:62002CJ0444>.
- [10] Věc C-30/14, rozsudek Soudního dvora (druhého senátu) ze dne 15. ledna 2015 ve věci *Ryanair Ltd v. PR Aviation BV*. In: *EUR-Lex* [online]. © European Union, 1998-2019 [cit. 30. 3. 2019]. Dostupné z: <https://eur-lex.europa.eu/legal-content/CS/TXT/?uri=CELEX:62014CJ0030>.
- [11] Case 17-cv-03301-EMC, order granting plaintiff's motion for preliminary injunction, filed 08/14/17. In: *The United States District Court for the Northern District of California* [online]. United States District Court [cit. 30. 3. 2019]. Dostupné z: <https://www.cand.uscourts.gov/filelibrary/3170/C-17-3301-hiQ-v.-LinkedIn-Order-Docket-No.-63.pdf>.
- [12] WILLIAMS, Jamie. 'Scraping' Is Just Automated Access, and Everyone Does It. In: *Electronic Frontier Foundation* [online]. Electronic Frontier Foundation [cit. 1. 4. 2019]. Dostupné z: <https://www.eff.org/deeplinks/2018/04/scraping-just-automated-access-and-everyone-does-it>
- [13] Forums, In: *Legal Information Institute* [online]. Cornell Law School [cit. 30. 3. 2019]. Dostupné z: <https://www.law.cornell.edu/wex/forums>.
- [14] NARULA, Prayag. LinkedIn Vs. hiQ Ruling Casts A Long Shadow Over The Tech Industry. In: *Forbes* [online]. © 2019 Forbes Media LLC [cit. 30. 3. 2019]. Dostupné z: <https://www.forbes.com/sites/forbestechcouncil/2017/09/20/linkedin-vs-hiQ-ruling-casts-a-long-shadow-over-the-tech-industry/#73e4bada5e6c>.
- [15] Case 13-17154, *Facebook v. Vachani*. In: *United States Court of Appeals for the Ninth Circuit* [online]. [cit. 31. 3. 2019]. Dostupné z: <https://cdn.ca9.uscourts.gov/datastore/opinions/2016/07/12/13-17102.pdf>.
- [16] KERR, Orin. 9th Circuit: It's a federal crime to visit a website after being told not to visit it. In: *The Washington Post* [online]. © 1996-2019 The Washington Post [cit. 31. 3. 2019]. Dostupné z: <https://www.washingtonpost.com/news/voikh-conspiracy/wp/2016/07/12/9th-circuit-its-a-federal-crime-to-visit->

a-website-after-being-told-not-to-visit-it/?noredirect=
on&utm_term=.ee6e01250364.

Seznam použitých zkratek

HTML HyperText Markup Language

XML eXtensible Markup Language

API Application Programming Interface

MIT Massachusetts Institute of Technology

CSS Cascading Style Sheets

CSV Comma-Separated Values

IP Internet Protocol

JSON JavaScript Object Notation

XLS formát souboru používaný aplikací Microsoft Excel

CFAA Computer Fraud and Abuse Act

CAN-SPAM Controlling the Assault of Non-Solicited Pornography and Marketing Act

Obsah přiloženého CD

	readme.txt	stručný popis obsahu CD
	exe	adresář se spustitelnou formou implementace
	src	
	impl.....	zdrojové kódy implementace
	thesis	zdrojová forma práce ve formátu L ^A T _E X
	text	text práce
	thesis.pdf	text práce ve formátu PDF
	thesis.ps	text práce ve formátu PS