



**FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE**

Bakalářská práce

Webová aplikace pro online web scraping

Jakub Drahoš

Obor Webové a softwarové inženýrství (BI-WSI), zaměření Softwarové inženýrství

Katedra softwarového inženýrství

Vedoucí práce: Martin Podloucký

18. 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 18. 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, JavaScript, rozšíření do Chromu, právní rozbor

Abstract

Sem doplňte ekvivalent abstraktu Vaší práce v angličtině.

Keywords web scraping, data extraction, application, JavaScript, Chrome extension, legal analysis

Obsah

| | |
|--|-----------|
| Úvod | 1 |
| 1 Analýza | 3 |
| 1.1 Web scraping | 3 |
| 1.2 Právní aspekt | 5 |
| 1.3 Analýza konkurenčních nástrojů | 12 |
| 2 Návrh | 19 |
| 2.1 Specifikace požadavků | 19 |
| 2.2 Případy užití | 22 |
| 2.3 Architektura systému | 22 |
| 2.4 Návrh uživatelského rozhraní | 22 |
| 3 Realizace | 23 |
| 3.1 Diskuze možných řešení | 23 |
| 3.2 Použité technologie | 24 |
| 3.3 Představení nástroje | 25 |
| 3.4 Implementace zvoleného řešení | 29 |
| 4 Testování | 31 |
| Závěr | 33 |
| Literatura | 35 |
| A Seznam použitých zkratk | 39 |
| B Obsah příloženého CD | 41 |

Seznam obrázků

| | | |
|-----|---|----|
| 1.1 | Desktopová aplikace ParseHub[25, snímek pořídil autor] | 13 |
| 1.2 | Desktopová aplikace Octoparse[26, snímek pořídil autor] | 14 |
| 1.3 | Rozšíření WebScraper[27, snímek pořídil autor] | 15 |
| 1.4 | Webová aplikace Dexi.io[28, snímek pořídil autor] | 16 |
| 1.5 | Rozšíření Data Scraper[29, snímek pořídil autor] | 18 |
| 3.1 | Ovládací panel aplikace | 25 |
| 3.2 | Výběr řádků (tyrkysově) a sloupců (vínově) | 26 |
| 3.3 | Výběr na základě textové shody | 27 |
| 3.4 | Výběr s pomocí CSS selektoru | 27 |
| 3.5 | Navigace výběru napříč dokumentem | 28 |
| 3.6 | Tabulka obsahující náhled extrahovaných dat | 29 |

Úvod

Cíle práce

Hlavním cílem této práce je návrh a tvorba softwaru, který bude umožňovat uživatelům extrahovat požadovaná data z libovolné stránky v reálném čase bez jakékoli nutné znalosti programování. Při specifikaci požadavků tohoto nástroje se přihlídně k analýze stávajících řešení, jež je vedlejším cílem této práce. Druhým vedlejším cílem je poskytnout čtenáři úvod do právní problematiky web scrapingu a shrnout na jednom místě fakta, která máme k dispozici.

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 a především *automatizovaně* procházel danou oblast webu.

Motivace

Téma web scrapingu je v dnešní době velice aktuální a čím více dat produkujeme, tím více bude stoupat potřeba tyto informace určitým způsobem získávat a zpracovávat. Téměř kdokoli, kdo pracuje s daty dostupnými z internetu, bude nucen využít nějaký nástroj k vytěžování, aby byl vůbec schopný udržet krok s konkurencí.

Tedy důvod k vytvoření softwaru umožňující extrahovat data z webových stránek je jasný. Ač podobných nástrojů existuje několik, jejich obsluha je poměrně složitá a je nutné strávit určitý čas, než se uživatel seznámí s jejich fungováním a může je naplno využít. Právě tento aspekt se snaží aplikace vyvíjená v rámci bakalářské práce eliminovat – motivací je tak poskytnout uživatelům možnost jednoduše a rychle vytěžit požadovaná data bez zbytečného zdržování a dlouhého času stráveného seznamováním se s nástrojem.

To, čím je tato práce unikátní, je ale rozbor právního aspektu web scrapingu. V žádném případě se nejedná o hlubokou analýzu, která by byla očekávána od studenta právnické fakulty. Zároveň ale shrnuje podstatné poznatky a fakta z dané oblasti na jednom místě. Velké množství textů dostupných na internetu, jenž se týkají tohoto tématu, totiž často začíná odstavcem ve stylu „...nejsem právník a toto je jen můj názor...“, informace jsou velmi kusé a chybí jim nějaká ucelená struktura. To se práce snaží napravit a představuje tak vstupní bránu do této rozsáhlé problematiky. Zároveň může pomoci všem vývojářům, kteří tvoří software určený k web scrapingu.

Členění práce

Kapitola 1 je věnována analýze tématiky web scrapingu. První sekce shrnuje obecné informace, následuje pohled z právní strany věci a nakonec analýza stávajících řešení problému. Kapitola 2 se zaměřuje na návrh aplikace – specifikace požadavků, vymezení případů užití, architektura systému, návrh uživatelského rozhraní. Ve třetí kapitole je popsána realizace daného návrhu, výběr použitých technologií a odůvodnění rozhodnutí, která byla učiněna. Poslední kapitola je věnována testování celé aplikace. V samotném závěru se pak zabývám vyhodnocením jednotlivých cílů zadaných výše.

Poznámka

V průběhu práce se objevily okolnosti (detailně jsou diskutovány v kapitole Realizace, sekce Diskuze možných řešení), které mě donutily přehodnotit původní záměr tvorby webové aplikace na tvorbu rozšíření do internetového prohlížeče Google Chrome. To jsem bohužel nevěděl při formulaci a odevzdání zadání do systému závěrečných prací, tedy název této práce neodpovídá úplně přesně výslednému produktu. Nicméně všechny cíle zůstaly naprosto totožné.

Analýza

V této kapitole jsou představeny základní pojmy web scrapingu, jeho historie, způsoby, kterými extrakce dat z internetových stránek nejčastěji probíhá a jak se takto získaná data dají využít. Hlavní částí této kapitoly je pak právní rešerše problematiky web scrapingu a analýza již existujících aplikací poskytujících uživatelské rozhraní pro vytěžování dat z webových stránek.

1.1 Web scraping

„*Web scraping (nebo také web harvesting, web data extraction) je softwarová technika zaměřená na extrakci informací z webových stránek*“ [1, překlad autora]. 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).

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.[2]

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.[3]

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 nejsofistikovanější a nejpokročilejší knihovnu pro web scraping. Za zmínku stojí určitě i rok 2006, kdy je datován příchod vizuálního web scrapingu, tedy techniky, kdy uživatel

skrže rozhraní aplikace označí klikáním myši, z kterých oblastí webové stránky chce extrahovat data. Tímto se otevřely dveře web scrapingu pro všechny.[4]

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 [5]
- vizuální web scraping – jak již bylo zmíněno výše, požadovaná data se musí ručně naklikat skrže 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 dle [6] velikost všech dat na celém internetu pohybuje v řádech Zettabajtů (1024^7 B). Mezi ty hlavní patří:

- získání kontaktních informací (např. e-mail) pro marketingové účely
- indexování webových stránek, které však využívá hlavně web crawling (jako příklad můžeme uvést známý GoogleBot)
- data mining – proces hledání vzorců ve velkých datových setech [7]
- 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, představit hlavní právní pojmy a poskytnout čtenáři alespoň náhled do této oblasti.

Při získávání dat z internetových stránek může nastat hned několik komplikací z právního hlediska, na které by se autor takového softwaru měl připravit. Následuje stručný souhrn informací, kterým se podrobně věnují jednotlivé části celé sekce 1.2.

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 směr 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 precedent¹, podle kterého by se soudy mohli při posuzování jednotlivých případů řídit. Proto také můžeme 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.

1.2.1 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.

¹Kontinentální právo (na rozdíl od anglosaského) nezná precedenty (tj. všeobecně závazná soudní rozhodnutí), resp. nepovažuje je právě za závazné. Na druhou stranu by soudy měly ve stejných případech postupovat stejně. Proto si zde dovolím tento výraz použít.

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.[8]

Clickwrap je oproti browse-wrap daleko lépe vymahatelný, neboť je uživatel nucen přímo vyjádřit souhlas či nesouhlas (kliknutím na tlačítko nebo 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 browsewrap 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.*“ [9, překlad autora].[10]

1.2.2 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.[11]

1.2.3 Obsah na webových stránkách a jeho možné využití

Dle [11] 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.*“ [12, § 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.*“ [13, § 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.“ [13, § 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 svém 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)². Důležité je, jestli tak činím *v souladu s běžným, očekávaným a přiměřeným použitím* – je tak nutné dát si pozor na detaily automatizovaného procházení, např. omezit počet požadavků

²Otázkou je, jestli by bylo vůbec vymahatelné, pokud by měla stránka přímo ve svých Terms of Service uvedeno, že si nepřije automatizované procházení, nehledě na to, jestli zároveň probíhá extrakce dat či nikoliv. Taková podmínka by mohla být brána jako neadekvátní a nepřiměřená.

na server, aby odpovídal běžnému použití lidským uživatelem³. Rozhodně lze ale doporučit provádět scraping webových stránek se souhlasem jejich provozovatele (poskytovatele), a to dle právního režimu dat a dané jurisdikce (situace u nás je jiná než třeba v USA).[11]

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.*...“ [13, § 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 [11] 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.*“ [13, § 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*“ [14].

„*Tato právní úprava byla do autorského zákona převzata ze Směrnice Evropského parlamentu a Rady EU 96/9/ES, o právní ochraně databází*“ [15], a tak můžeme informace čerpané z této sekce vztahovat nejen na Českou Republiku, ale na jakoukoli zemi Evropské Unie.

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*

³Tedy pokud je nějaký software schopný projít celou doménu během pár vteřin, je to náznak, že nemusí respektovat výše uvedené podmínky.

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.“[13, § 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.

“[13, § 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 jakémukoliv 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 podle nich 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.[16]

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).[17]

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.[18]

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.[18]

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 vystávají 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 pohrůzkou porušení CFAA.[19]

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 obejítí 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 [20]. hiQ mimo jiné argumentoval i tím, že sociální sítě jsou novodobá veřejná fóra (místa pro veřejné projevy, zejména politického charakteru[21]) 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.[22]

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 (údajné 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.[23]

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.[24]

1.3 Analýza konkurenčních nástrojů

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. Zaměřím se 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 odemknou až s určitým platebním plánem – tzv. freemium model) a jeden poskytuje bezplatně pouze 7denní zkušební verzi.

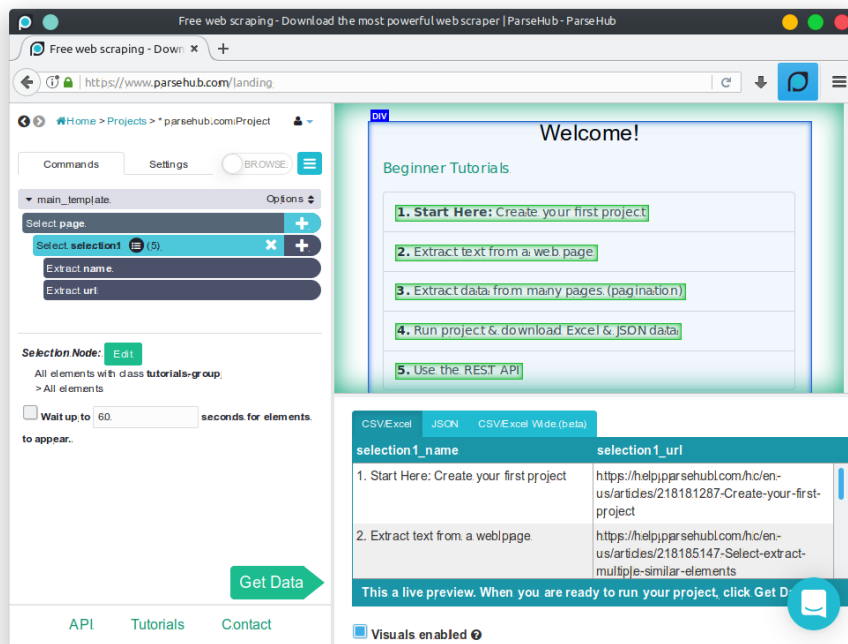
Předtím, než bude možné jednotlivé nástroje porovnávat, je nutné určit kritéria, podle kterých lze hodnotit kvalitu daného nástroje. Především půjde o jednoduchost používání, celkovou přehlednost a rychlost, se kterou se uživatel dostane k požadovaným datům. Důležitý je také způsob výběru dat, možnosti 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)



Obrázek 1.1: Desktopová aplikace ParseHub[25, snímek pořídil autor]

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

1.3.2 Octoparse

Výhody:

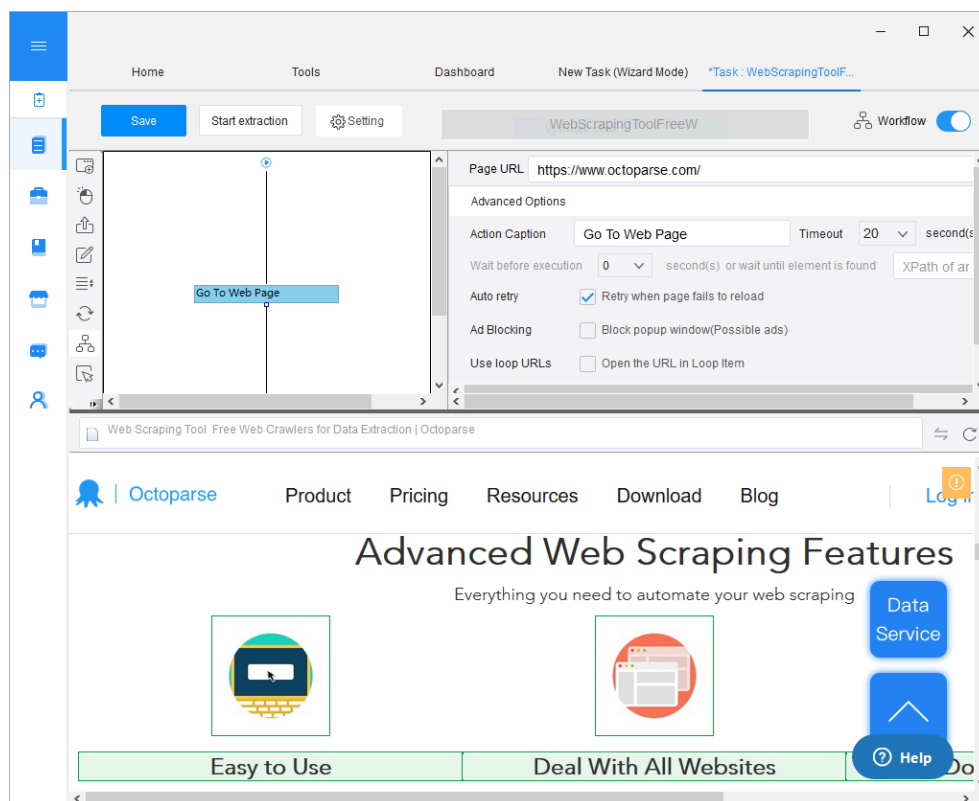
- výběr dat jak pomocí klikání (inteligentní hledání vzorců/podobností na základě prvních dvou kliknutí), tak pomocí XPath nebo regulárních výrazů
- nástroj obsahuje hotové šablony, které mohou velmi urychlit práci
- pestrá paleta možností (branch judgment, tvoření smyček apod.) – dá se vytvořit téměř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

1. ANALÝZA

- možnost řídit tasky přes API (a získávat tak data taktéž přes API); data jdou nahrát rovnou i do lokální databáze

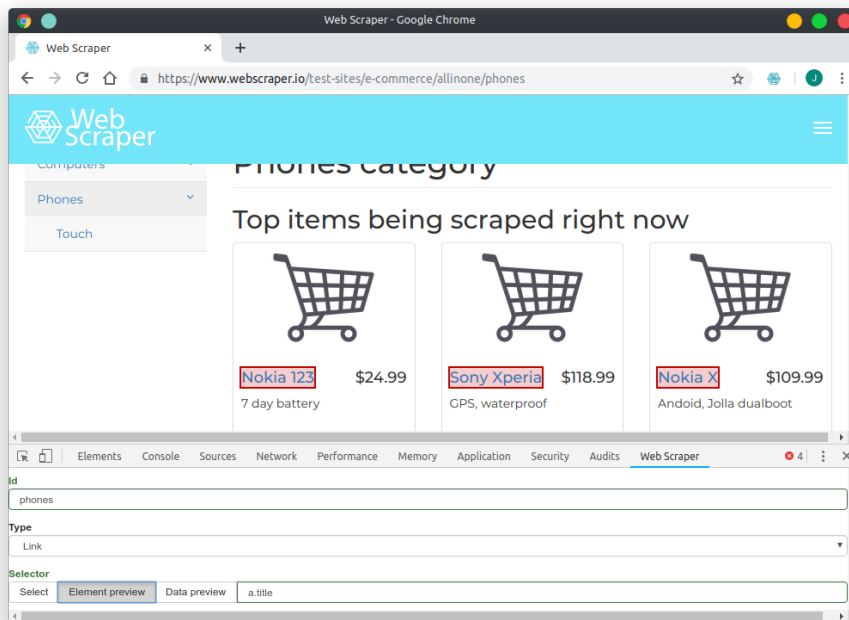
Nevýhody:

- 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: Desktopová aplikace Octoparse[26, snímek pořídil autor]

1.3.3 WebScaper



Obrázek 1.3: Rozšíření WebScaper[27, snímek pořídil autor]

Výhody:

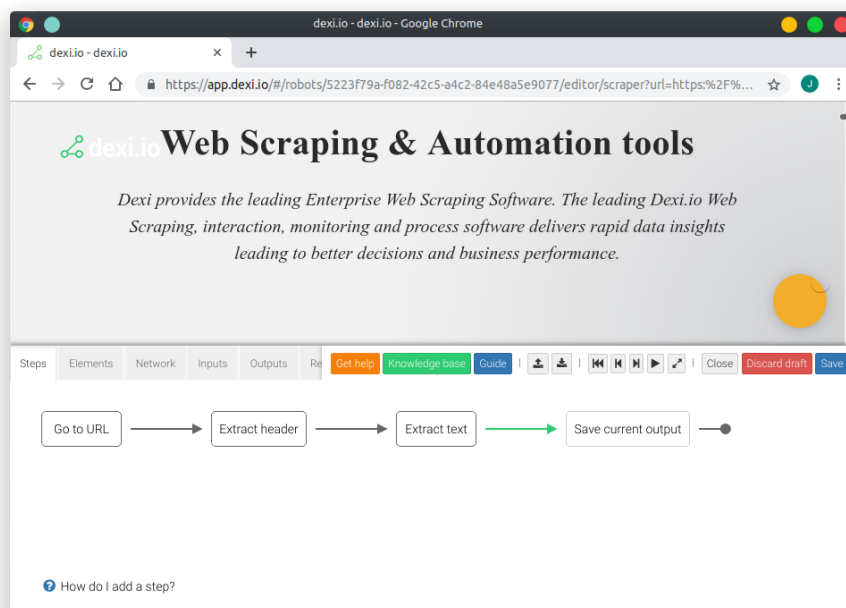
- jednoduchá instalace (jedná se pouze o rozšíření do prohlížeče Google Chrome); scrapování probíhá skrze vývojářskou konzoli
- výběr dat pomocí klikání (inteligentní hledání vzorců/podobností na základě prvních dvou kliknutí)
- tutoriály jsou formou videí – jednoduché, rychlé a naprosto postačující
- 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 a ne tak složité používání

1. ANALÝZA

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: Webová aplikace Dexi.io[28, snímek pořídil autor]

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 uživatel 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

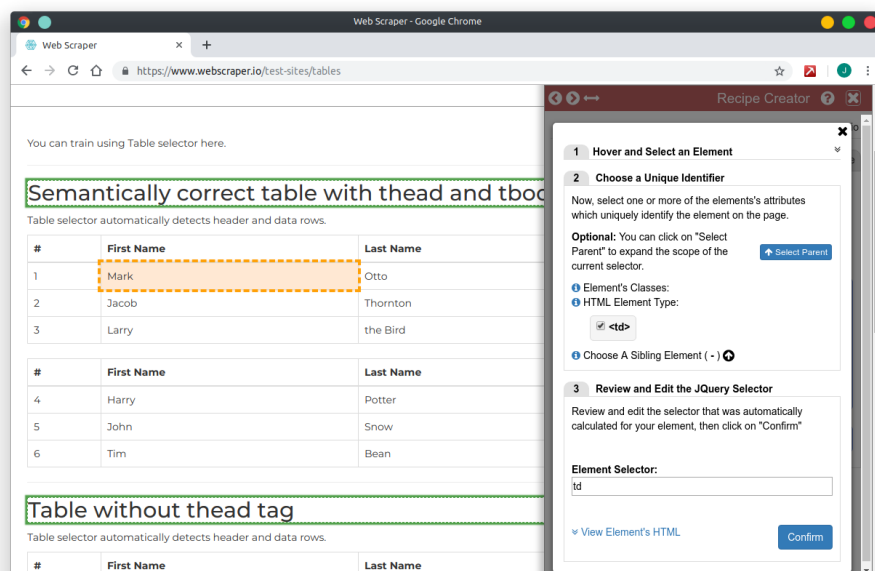
Výhody:

- 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

1. ANALÝZA



Obrázek 1.5: Rozšíření Data Scraper[29, snímek pořídil autor]

Návrh

Tato kapitola definuje funkční i nefunkční požadavky kladené na vyvíjený software, jež částečně vyplývají z předchozí analýzy stávajících řešení – problémy konkurenčních nástrojů se snaží napravit, zároveň agregují dobré vlastnosti zmíněných aplikací. Součástí kapitole je také diskuze případů užití výsledné aplikace. Poté je věnován prostor návrhu celé architektury systému a v závěru kapitoly i návrhu uživatelského rozhraní.

2.1 Specifikace požadavků

Jak jsme viděli v předchozí analýze stávajících řešení, 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 zdlouhavé 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:

2.1.1 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
- hlavní činností uživatele bude výběr elementů na jím zadané webové stránce, ze kterých bude na konci procesu vyextrahován text; tento výběr bude probíhat následujícími způsoby:
 - kliknutím myši na požadované elementy
 - na základě textové shody; zde bude mít uživatel 4 možnosti na výběr:
 1. Prvek bude vybrán, pokud jeho text začíná hledaným výrazem
 2. Prvek bude vybrán, pokud jeho text končí hledaným výrazem
 3. Prvek bude vybrán, pokud jeho text obsahuje hledaný výraz
 4. Prvek bude vybrán, pokud se jeho text přesně shoduje s hledaným výrazem
 - pomocí CSS selektorů (tedy HTML tagy, třídy, id, hodnoty atributu, různé následnosti a vše ostatní, co CSS selektory umožňují, viz přehled CSS selektorů)
 - na ovládacím panelu nalezneme i tlačítka s hotovými akcemi představující šablonu pro nejpoužívanější operace (stažení všech e-mailových adres ze stránky, všech obrázků atd.)
- pokud vybraným prvkem bude ** HTML tag, bude místo jeho textu extrahován atribut *src* (tedy zdroj obrázku)
- po kliknutí na určitý prvek s přidrženou klávesou **Ctrl/control** se program pokusí vybrat všechny podobné prvky na základě předchozí selekce (tzv. *auto-selection*)
- mezi ovládacími prvky nalezneme tlačítka **undo** a **redo**, která umožní vracet zpět provedený výběr (například v situaci, kdy nesouhlasíme s výběrem *auto-selectu*)
- všechny vybrané prvky budou barevně odlišeny, aby bylo jasné, co už je připraveno k extrakci a co ještě ne

- k dispozici bude přibližování (první potomek) a oddalování (otec) momentálního výběru pomocí ikony + a –, případně přesunutí výběru na předchozího/následujícího sourozence v DOMu (uživatel může pomocí této funkce traverzovat napříč zanořenými prvky všemi směry)
- data z vybraných elementů si bude možné kdykoli prohlédnout v tzv. *preview* módu – půjde o obyčejnou tabulku, ze které bude možné vymazat nevyhovující řádky (vymazáním řádku se odebere výběr všech relevantních prvků na stránce)
- získaná data půjdou exportovat do formátů JSON, CSV, XLS

2.1.2 Nefunkční požadavky

- půjde o webovou aplikaci běžící v internetovém prohlížeči, tedy nebude nutná žádná instalace
- celý proces bude realizován na straně klienta – bude se jednat pouze o frontend, žádný backend server nebude k dispozici
- aplikace cílí primárně na celkový zážitek uživatele – grafické rozhraní bude přehledné a co nejjednodušší, ovládání přímočaré a intuitivní
- čas, za který se uživatel dostane k požadovaným datům (tedy čas, který stráví vybíráním dat; nepočítáme čas potřebný ke stažení), bude co nejmenší
- čas nutný k samotné extrakci (od okamžiku, kdy uživatel klikne na tlačítko Download) nepřesáhne 5 vteřin

2.1.3 Nice to have požadavky

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

- uživatelské rozhraní aplikace nabídne intuitivní klávesové zkratky pro usnadnění práce – Ctrl + a a Ctrl - obstará přibližování/oddalování momentálního výběru, Ctrl n a Ctrl p vybere předchozího/následujícího sourozence vybraného prvku
- postranní panel s ovládacími prvky půjde minimalizovat (zmenšit k přilehlé hraně tak, aby zabíral co nejméně místa a nepřekážel v manipulaci s webovou stránkou) nebo přesunout na protější stranu (např. v případě, že by zakrýval nějaké prvky na stránce)

2. NÁVRH

- export dat realizovatelný i do Google Sheets, Google Drive, Dropbox
- interaktivní tutoriál, který v rychlosti představí práci s nástrojem
- na základě výběru dat uživatelem se vytvoří určitý filtr (textový řetězec), který může být ručně upraven – půjde tak o alternativu pro zkušenější uživatele, aniž by se zaneslo uživatelské rozhraní přehrší možností a celé se tak znepřehlednilo

2.2 Případy užití

...

2.3 Architektura systému

...

2.4 Návrh uživatelského rozhraní

...

Realizace

Tato část se zabývá naplněním funkčních a nefunkčních požadavků z minulé kapitoly z hlediska konkrétní implementace. Jsou diskutovány problémy a zádrhly, které nastaly při vývoji aplikace (především pak podrobně vysvětlují, proč bylo upřednostněno rozšíření do prohlížeče před webovou aplikací), stejně tak jejich možná řešení a proč bylo zvoleno zrovna dané rozhodnutí. Dále je představena výsledná aplikace, popsán způsob jejího používání a v poslední sekci nalezneme realizaci jednotlivých komponent systému.

3.1 Diskuze možných řešení

Původní záměr byl vytvořit webovou aplikaci. Důvod je jednoduchý, nikdo v dnešní době nechce cokoli stahovat a instalovat. Vzhledem k důrazu vyvíjeného nástroje kladeného na jednoduchost a rychlost používání je tak webová aplikace jasnou volbou, jelikož se jedná o nejpřímočařejší řešení a pro uživatele určitě nejpohodlnější.

Návrh tedy předpokládal jednoduchou webovou aplikaci s hlavní obrazovkou, kde by byla zobrazená uživatelem zadaná stránka a postranním panelem, který by obsahoval veškeré ovládací prvky. Jasným řešením tak byl HTML `iframe`, který reprezentuje vnořený kontext procházení (kontext procházení si můžeme představit jako jedno okno/záložku prohlížeče) – tedy umožňuje zobrazit HTML dokument uvnitř jiného HTML dokumentu.

Ač se toto zprvu zdálo jako ideální řešení, hned v úvodu jsem narazil na stěžejní implementační problém – z bezpečnostních důvodů existuje HTTP hlavička *X-Frame-Options*, která určuje, kdo může danou stránku vložit do `iframe` tagu a zobrazit ji tak v rámci své vlastní stránky (viz MDN web docs). Spousta webových stránek nastavuje tuto hlavičku tak, aby nebylo možné jejich stránky vkládat do `iframů`, čímž se brání tzv. *clickjacking* útokům (viz wikipedia). Jenže to představuje v podstatě neřešitelný problém, neboť HTTP hlavičky se v prohlížeči nedají nijak obejít a dá se předpokládat, že toto blokování bude provádět mnoho stránek.

Možnou alternativou by bylo stáhnout veškerý obsah ze zadané domény (HTML, CSS, JavaScript, všechny assety jako obrázky apod.), ten sestavit dohromady a poskytovat z vlastního serveru pouze danému uživateli. V podstatě by tak došlo k vyscrapování všech dat z cílové domény a uživatel by již pouze odfiltroval informace, o které nemá zájem. To je ale nevhodné hned z několika důvodů – jednak by byla uživateli prezentována stránka, která ve skutečnosti není tou, za kterou se vydává; mohlo by docházet k porušení copyrightu a autorských práv a v neposlední řadě by složitost takového řešení naprosto neodpovídala poměrně přímočarému úkolu výsledné aplikace.

Na problém se však lze dívat i opačně a základní myšlenku invertovat, což nás dovede k použitému řešení. Pokud není možné vložit stránku do webové aplikace, je nutné vložit aplikaci do požadované stránky. To lze elegantně vyřešit pomocí rozšíření do internetového prohlížeče Google Chrome – tzv. *Chrome-extension*.

3.2 Použité technologie

Celé řešení zadaného problému je tedy nakonec implementováno jako rozšíření do internetového prohlížeče Google Chrome (jehož popisu se věnuji níže). Z tohoto důvodu je zvoleným programovacím jazykem čistý JavaScript, resp. ECMA-Script 2018 verze 9. Dále je použit značkovací jazyk HTML k definici struktury celého ovládacího panelu a zbylých kontrolních prvků spolu s CSS jazykem určujícím styl zobrazení jednotlivých elementů. K testování aplikace byl použit JavaScriptový testovací framework Jest.

3.2.1 Rozšíření do prohlížeče Google Chrome

Chrome-extensions se skládají z několika různých komponent:

Background script je základní komponenta (můžeme si ji představit jako takový backend celého rozšíření), která se vykoná při každém spuštění prohlížeče. Zde je možné registrovat obsluhy různých událostí (např. když se otevře nová záložka v prohlížeči nebo když je dané rozšíření poprvé nainstalováno).

Browser action představuje tlačítko rozšíření umístěné v hlavním panelu nástrojů prohlížeče Google Chrome (vedle pole pro zadávání adresy). Po kliknutí na něj je vypuštěna událost mířící do background scriptu, kde se odehrává následné zpracování.

Content script je zajisté tím nejsilnějším mechanismem, které Chrome extensions nabízí. Zároveň je nejdůležitější částí tohoto ekosystému, jenž nás bude zajímat. Jedná se o kód, který je vložen do požadované stránky a spustí se v rámci jejího kontextu. To znamená, že každý takto vložený skript má

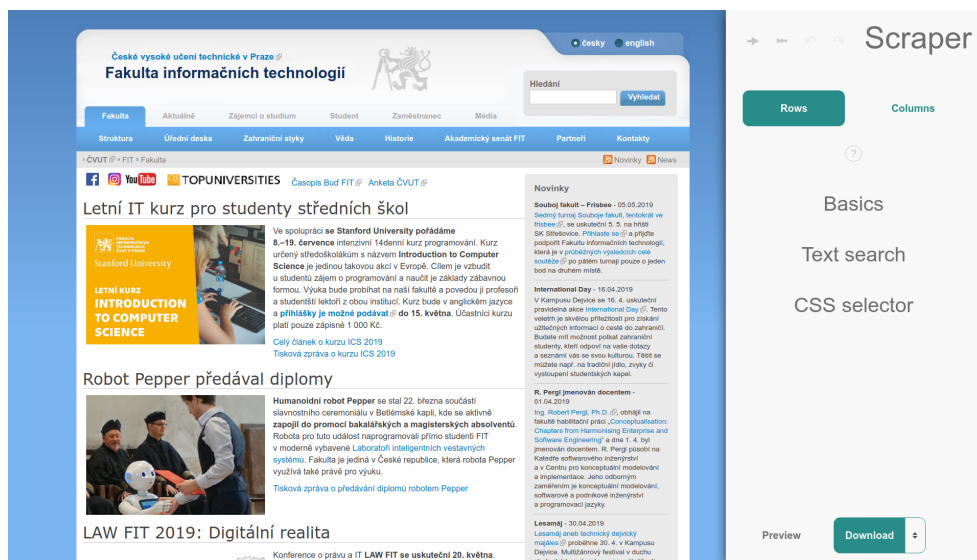
kompletní přístup k DOMu dané stránky i všem ostatním skriptům. Stane se validní součástí stránky, může *vytvářet, mazat nebo měnit prvky*, a sice do doby, než je stránka obnovena a znovu načtena.

3.3 Představení nástroje

Před používáním nástroje je nutné si uvědomit jednu velmi podstatnou věc – co je výsledek, který očekáváme na konci. Jak mají vypadat data, která plánujeme extrahovat, jakou mají mít podobu? V rámci našeho nástroje se bude ve výsledku jednat vždy o *tabulku*. Toto je nezbytné pro pochopení celého procesu.

Nyní se můžeme podívat na použití aplikace (předpokládejme, že rozšíření je již nainstalované v prohlížeči):

1. Uživatel navštíví stránku, ze které si přeje extrahovat dat.
2. Klikne na ikonu rozšíření nacházející se v pravém horním rohu prohlížeče, v hlavním panelu nástrojů.
3. Zobrazí se hlavní ovládací panel a aplikace je připravena k výběru dat, viz Obrázek 3.1.



Obrázek 3.1: Ovládací panel aplikace

Jak jsme si řekli v úvodu, výsledná data budou ve formě tabulky, tedy musíme vybrat, které elementy na stránce budou reprezentovat řádky výsledné tabulky a které budou reprezentovat sloupce, viz Obrázek 3.2. K tomu slouží dvě velká tlačítka Rows a Columns, kterými se přepíná výběr právě mezi řádky

3. REALIZACE



Obrázek 3.2: Výběr řádků (tyrkysově) a sloupců (vínově)

a sloupce. Doporučený postup je nejdřív vybrat řádky⁴ a poté uvnitř těchto větších elementů vybírat sloupce⁵.

Dle zadání v kapitole Návrh, sekce Specifikace požadavků, probíhá selekce třemi způsoby – ruční označování prvků, hledání na základě textové shody a výběr pomocí CSS selektorů. Tyto tři druhy výběru jsou stejné jak pro výběr řádků, tak pro výběr sloupců. Jakýkoli výběr lze vzít zpět nebo následně provést znovu pomocí tlačítek Undo a Redo v horní části nástroje. Následuje popis každého ze způsobů výběru:

Ruční označování probíhá klikáním myši na požadované elementy a zapíná se přepínačem nacházejícím se na kartě Basics. Pokud uživatel při vybírání podrží klávesu Ctrl, aplikace se pokusí vybrat všechny podobné prvky na základě předchozího kliknutí. Kliknutím na již označený element se výběr zruší.

Největší část výběrů bude představovat právě tento způsob.

Textová shoda se nalézá na kartě Text search, která se skládá ze tří samostatných formulářů. Uživatel má možnost vybrat všechny elementy na stránce, jejichž text bude:

- začíná daným výrazem,
- končí daným výrazem,

⁴To budou nejčastěji různé „kontejnery“, které sdružují data jedné entity dohromady – jako příklad lze uvést kartu produktu v přehledu produktů e-shopu.

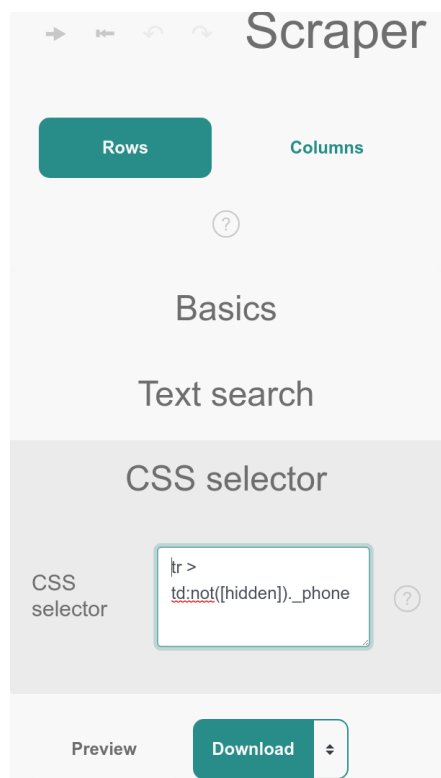
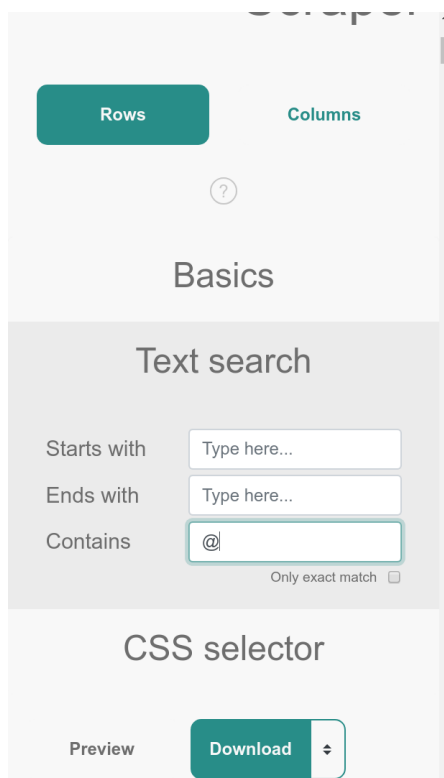
⁵To může být například cena nebo jméno daného produktu.

c) obsahuje daný výraz nebo se mu přímo rovná.

Tento způsob se osvědčí například v případě, kdy chceme vybrat všechny e-mailové adresy na stránce – stačí hledat prvky, které obsahují zavináč.

CSS selektory najdeme na kartě s názvem **CSS selectors** a jedná se o jednoduché textové pole, které přijímá libovolný CSS selektor. Můžeme tedy pomocí něj vybírat na základě HTML tagů (`jmenoTagu`), tříd (`.jmenoTridy`), atributů (`[atribut=hodnota]`), různé následnosti (`otec > syn + naslednik`) a zkrátka vše, co CSS selektory umí, viz přehled selektorů.

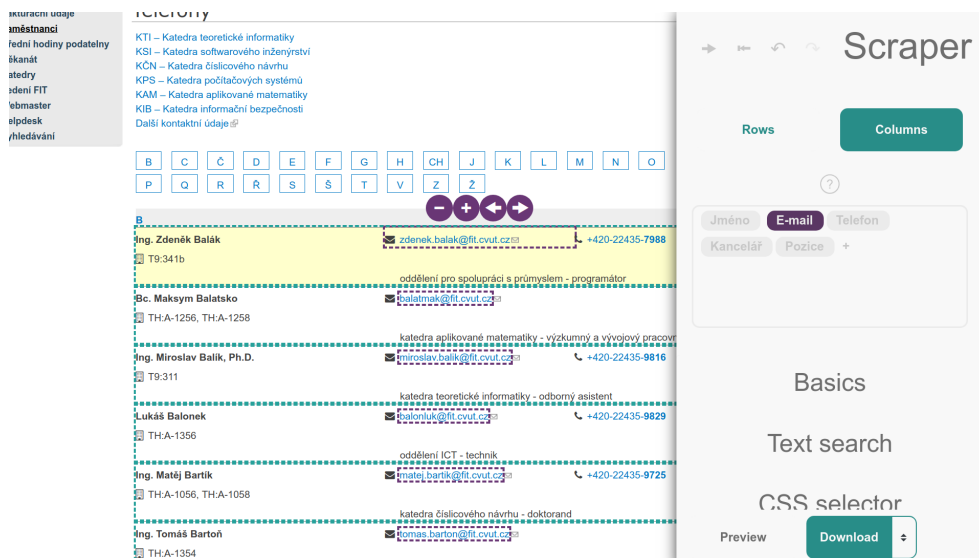
Toto je jistě nejsilnější z uvedených způsobů, jelikož s ním jde vybrat libovolná skupina prvků, avšak předpokládá alespoň základní znalost HTML, CSS a především struktury stránky. Je tedy spíš pro pokročilejší uživatele.



Obrázek 3.3: Výběr na základě textové shody Obrázek 3.4: Výběr s pomocí CSS selektoru

3. REALIZACE

Může se stát, že požadovaný element nejde označit ručně. Pro tyto případy je každý vybraný prvek opatřen čtveřicí tlačítek, která se objeví, pokud uživatel najede kurzorem na daný element, viz Obrázek 3.5. Tlačítko + posune označení na prvního syna prvku, – na otce, ← na předchozího sourozence a tlačítko → na následujícího sourozence. Tímto způsobem může uživatel traverzovat napříč celým DOMem a vybrat tak libovolný prvek.



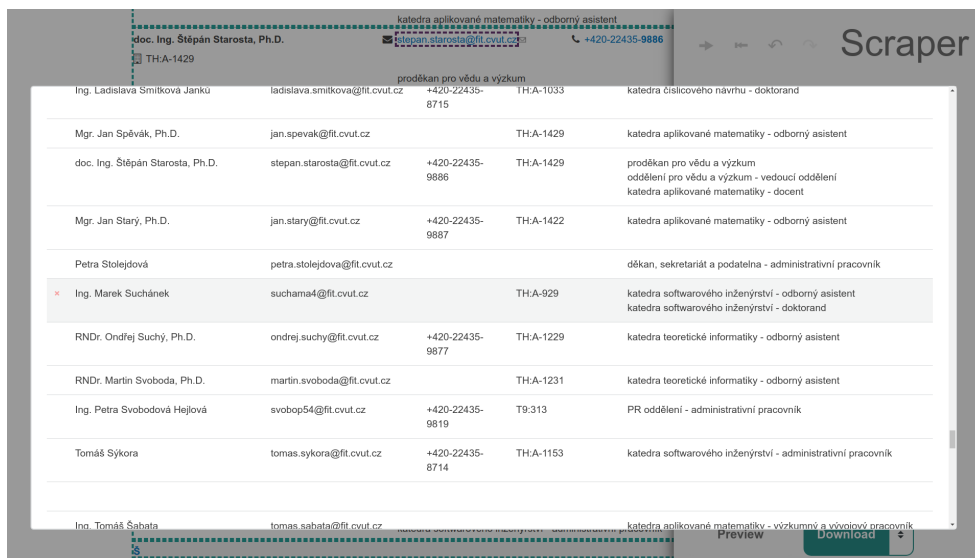
Obrázek 3.5: Navigace výběru napříč dokumentem

Poté, co jsme s výběrem hotovi, je možné data prohlédnout v tzv. *Preview módu*. Zobrazí se tabulka obsahující námi zadané řádky a sloupce⁶. V tuto chvíli je možné zkontrolovat veškerá extrahovaná data a případně vyřadit ta, která nevyhovují našim potřebám pomocí křížku na levé straně každého řádku (ten se objeví až po najetí kurzorem na daný řádek), viz Obrázek 3.6. Vyřazení záznamu způsobí zrušení výběru elementů, které obsahují příslušná data.

Na závěr stačí pomocí rozbalovacího výběru zvolit formát, do kterého chceme data exportovat (zatím je na výběr CSV a JSON) a kliknout na tlačítko Download.

⁶Nutno podotknout, že prvky označené jako sloupce, které se *nenacházejí* uvnitř prvku označeného jako řádek, nebudou zahrnuty do výsledku.

3.4. Implementace zvoleného řešení



| Ing. Ladislava Smítková Janku | ladislava.smilkova@fit.cvut.cz | +420-22435-8715 | TH:A-1033 | katedra číslicového návrhu - doktorand |
|----------------------------------|--------------------------------|-----------------|-----------|---|
| Mgr. Jan Spěvák, Ph.D. | jan.spevak@fit.cvut.cz | | TH:A-1429 | katedra aplikované matematiky - odborný asistent |
| doc. Ing. Štěpán Starosta, Ph.D. | stepan.starosta@fit.cvut.cz | +420-22435-9886 | TH:A-1429 | proděkan pro vědu a výzkum oddělení pro vědu a výzkum - vedoucí oddělení katedra aplikované matematiky - docent |
| Mgr. Jan Starý, Ph.D. | jan.starý@fit.cvut.cz | +420-22435-9887 | TH:A-1422 | katedra aplikované matematiky - odborný asistent |
| Petra Stolejšová | petra.stolejsova@fit.cvut.cz | | | děkan, sekretariát a podatelna - administrativní pracovník |
| Ing. Marek Suchánek | suchama4@fit.cvut.cz | | TH:A-929 | katedra softwarového inženýrství - odborný asistent katedra softwarového inženýrství - doktorand |
| RNDr. Ondřej Suchý, Ph.D. | ondrej.suchy@fit.cvut.cz | +420-22435-9877 | TH:A-1229 | katedra teoretické informatiky - odborný asistent |
| RNDr. Martin Svoboda, Ph.D. | martin.svoboda@fit.cvut.cz | | TH:A-1231 | katedra teoretické informatiky - odborný asistent |
| Ing. Petra Svobodová Hejlová | svobop54@fit.cvut.cz | +420-22435-9819 | T9:313 | PR oddělení - administrativní pracovník |
| Tomáš Sýkora | tomas.sykora@fit.cvut.cz | +420-22435-8714 | TH:A-1153 | katedra softwarového inženýrství - administrativní pracovník |
| Ing. Tomáš Šabala | tomas.sabala@fit.cvut.cz | | | katedra aplikované matematiky - výzkumní a vývojový pracovník |

Obrázek 3.6: Tabulka obsahující náhled extrahovaných dat

3.4 Implementace zvoleného řešení

3.4.1 Inicializace a spuštění

Jak již bylo řečeno v úvodu této kapitoly, použité řešení spočívá ve vložení ovládacího panelu do libovolné stránky. To umožňuje právě komponenta content script, jež byla zmíněna výše. Spuštění aplikace vypadá následovně:

1. Po instalaci rozšíření do prohlížeče Google Chrome (a každém jeho spuštění) je vykonán kód, který je obsažen v background scriptu. Zde se nachází pouze obsluha vyzývající content script k zobrazení nebo schování ovládacího panelu.
2. Když je rozšíření aktivní, do *každé* načtené stránky je vložen content script, jenž poslouchá zprávy od background scriptu.
3. Kliknutím na ikonu rozšíření v hlavním panelu nástrojů je vyvolána událost, na kterou reaguje background script – aktivnímu oknu/záložce zašle zprávu, aby byl otevřen ovládací panel.
4. Tu odchytlí content script, který na dané stránce poslouchá a vloží do těla stránky nový iframe, do něhož načte HTML dokument, který představuje samotný ovládací panel. Při dalších žádostech je iframe pouze skryt/zobrazen.
5. Vhodným nastavením příslušných CSS vlastností je iframe umístěn na boční straně prohlížené stránky a nabývá formy ovládacího panelu.

Obrázek 3.1 ilustruje, jak ovládací panel vypadá po vložení do stránky.

3.4.2 Výběr dat

...

3.4.3 Komunikace mezi komponentami

...

3.4.4 Extrakce dat

...

3.4.5 Ostatní

...

Testování

...

Závěr

V bakalářské práci jsem se zabýval analýzou web scrapingu a především pak právní problematikou, jež představuje první ze dvou vedlejších cílů. Uvedl jsem základní pojmy a zasadil web scraping do kontextu českého právního systému. Také jsem představil nejvýznamnější soudní případy zabývající se právě vytěžováním dat z internetových stránek, díky kterým jsme se mohli přesvědčit, že zatím neexistuje jednotné stanovisko při určování, co je v rozporu s právem a co ne. V nejbližší budoucnosti však můžeme být svědky formování a zrodu jednotného precedentu, dle něhož se soudy mohou řídit. Užitek bakalářské práce pak spatřuji hlavně v tomto právním souhrnu, který může posloužit jako dobrý úvod do hlubšího pátrání.

Poté jsem se věnoval druhému z vedlejších cílů, a sice řešerši stávajících řešení. Nástrojů poskytujících uživatelské rozhraní pro vytěžování dat z webových stránek existuje nespočet, avšak téměř všechny se snaží nabídnout co nejvíce funkcionalit a možností na úkor uživatelského zážitku a jednoduchosti používání. Přesně tomu se snaží vyhnout software vyvinutý v rámci této práce, čímž přecházím k hlavnímu cíli.

V rané fázi vývoje nastaly neočekávané komplikace. Implementace návrhu jakožto webové aplikace najednou nepřípadala v úvahu kvůli HTTP hlavičce omezující možnost zabudovat libovolný HTML dokument uvnitř své vlastní stránky. Musel jsem tedy upustit od svého původního záměru a najít lepší řešení. Tím je implementace v podobě rozšíření do internetového prohlížeče, které je dle mého názoru elegantním řešením daného problému.

Jak již bylo několikrát řečeno, přínos samotné aplikace spočívá především v její jednoduchosti. Z toho mohou těžit uživatelé, kteří se nezabývají programováním nebo tvorbou webových stránek. Využije ji tak kdokoli, kdo potřebuje jednorázově získat data z webové stránky, jež obsahuje velké množství dat pohromadě na jedno místě. Z důvodu chybějící automatizace a crawlingu je naopak nevhodná k pravidelnému získávání dat (jako je například dlouhodobé sledování cen produktů) či ke zpracování stránek, kde se jednotlivá data nacházejí rozptýlená po celé doméně.

Hlavní i vedlejší cíle tak považuji za úspěšně splněné. Zároveň se zde otevírá obrovský prostor pro potenciální navazující práce. Na tomto základu by mohla vzniknout opravdu hluboká a plnohodnotná právní analýza celé domény. Taktéž nástroj samotný je připraven pro další vylepšení, ať už by se jednalo o dokončení všech definovaných požadavků či přidání klíčové funkcionality, která by diametrálně změnila jeho konkurenceschopnost – web crawlingu nebo možnosti automatizace.

Literatura

- [1] VARGIU, Eloisa; URRU, Mirko. Exploiting web scraping in a collaborative filtering-based approach to web advertising. *Artificial Intelligence Research* [online]. 2013, roč. 2, č. 1 [cit. 13. 4. 2019]. DOI: 10.5430/air.v2n1p44. ISSN 1927-6974. Dostupné z: <http://www.sciedu.ca/journal/index.php/air/article/view/1390/1115>.
- [2] KOBAYASHI, Mei; TAKEDA, Koichi. Information Retrieval on the Web. *ACM Computing Surveys* [online]. 2000, roč. 32, č. 2 [cit. 13. 4. 2019]. DOI: 10.1145/358923.358934. ISSN 0360-0300. Dostupné z: <https://dl.acm.org/citation.cfm?doid=358923.358934>.
- [3] World Wide Web Wanderer. In: *History-Computer* [online] [cit. 13. 4. 2019]. Dostupné z: <https://history-computer.com/Internet/Conquering/Wanderer.html>.
- [4] Web Scraping: How It All Started And Will Be. In: *Octoparse's blog* [online]. © 2019 Octopus Data Inc. [cit. 13. 4. 2019]. Dostupné z: <https://www.octoparse.com/blog/web-scraping-introduction>.
- [5] ROUSH, Wade. Diffbot Is Using Computer Vision to Reinvent the Semantic Web. In: *Xconomy* [online]. © 2007–2019, Xconomy [cit. 13. 4. 2019]. Dostupné z: <https://xconomy.com/san-francisco/2012/07/25/diffbot-is-using-computer-vision-to-reinvent-the-semantic-web/>.
- [6] WASSÉN, Olivia. Big Data facts – How much data is out there? In: *NodeGraph* [online]. © 2019 NodeGraph [cit. 13. 4. 2019]. Dostupné z: <https://www.nodegraph.se/big-data-facts/>.
- [7] CLIFTON, Christopher. Data mining. In: *Encyclopædia Britannica* [online]. ©2019 Encyclopædia Britannica, Inc. [cit. 13. 4. 2019]. Dostupné z: <https://www.britannica.com/technology/data-mining>.

- [8] 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>.
- [9] 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>.
- [10] OBAR, Jonathan A.; OELDORF-HIRSCH, Anne. 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>.
- [11] VAŠÍČEK, Libor. [osobní sdělení]. Advokát specializující se na ICT právo a autorské právo. Sídlo advokátní kanceláře Legal Partners, s. r. o., Záhořanského 1944/4, Praha 2. 29. března 2019.
- [12] 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>.
- [13] 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>.
- [14] POLČÁK, Radim. *Právo informačních technologií*. Praha: Wolters Kluwer, 2018. Právní monografie (Wolters Kluwer ČR). 656 stran. ISBN 978-80-7598-045-8.
- [15] SMEJKAL, Ladislav. Právní ochrana databází v novém autorském zákoně. *Ikaros* [online]. 2001, roč. 5, č. 3 [cit. 13. 4. 2019]. urn:nbn:cz:ik-10688. ISSN 1212-5075. Dostupné z: <http://ikaros.cz/node/10688>.
- [16] Věc C-444/02, stanovisko generální advokátky ze dne 8. června 2004 ve věci Fixtures Marketing Ltd v. 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>.
- [17] Věc C-444/02, rozsudek Soudního dvora (velkého senátu) ze dne 9. listopadu 2004 ve věci Fixtures Marketing Ltd v. 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:62002CJ0444>.

-
- [18] 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>.
- [19] 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>.
- [20] 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>
- [21] Forums. In: *Legal Information Institute* [online]. Cornell Law School [cit. 30. 3. 2019]. Dostupné z: <https://www.law.cornell.edu/wex/forums>.
- [22] 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>.
- [23] 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>.
- [24] 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/volokh-conspiracy/wp/2016/07/12/9th-circuit-its-a-federal-crime-to-visit-a-website-after-being-told-not-to-visit-it>.
- [25] PARSEHUB. *ParseHub. Version 54.0.1* [software]. 2015 [cit. 13. 4. 2019]. Dostupné z: <https://www.parsehub.com/quickstart>.
- [26] OCTOPUS DATA INC. *Octoparse. Version 7.1.2* [software]. 2018 [cit. 13. 4. 2019]. Dostupné z: <https://www.octoparse.com/download>.
- [27] WEBSCRAPER. *WebScraper. Version 0.3.8.9* [software]. 2016 [cit. 13. 4. 2019]. Dostupné z: <https://chrome.google.com/webstore/detail/web-scraper/jnhgnonknehpejjnehehlklplmbmhn>.

LITERATURA

- [28] DEXI APS. *Dexi.io* [software] [cit. 13. 4. 2019]. Dostupné z: <https://app.dexi.io/>.
- [29] SOFTWARE INNOVATION LAB LLC. *Data scraper. Version 3.299.84* [software]. 2015 [cit. 13. 4. 2019]. Dostupné z: <https://chrome.google.com/webstore/detail/data-scraper-easy-web-scr/nndknepjnldbdbepjfgmncbggmopgden>.

Seznam použitých zkratk

API Application Programming Interface

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

CFAA Computer Fraud and Abuse Act

CSS Cascading Style Sheets

CSV Comma-Separated Values

DOM Document Object Model

HTML HyperText Markup Language

IP Internet Protocol

JSON JavaScript Object Notation

MIT Massachusetts Institute of Technology

XLS formát souboru používaný aplikací Microsoft Excel

XML eXtensible Markup Language

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 |