Univerzita Karlova v Praze

Filozofická fakulta

Ústav informačních studií a knihovnictví

Studijní program: Informační studia a knihovnictví

Studijní obor: Informační studia a knihovnictví

Bakalářská práce

Viktor Dobrovolný

Použití information scrapingu pro tvorbu výukových simulací

The use of information scraping for the development of educational simulations

Praha, 2012

Vedoucí práce: Mgr. Vít Šisler, PhD.

Rád bych poděkoval vedoucímu této bakalářské práce Mgr. Vítu Šislerovi, PhD. za trpělivost a ochotu řešit se mnou všechny problémy, na které jsem při jejím psaní narazil. Také bych chtěl poděkovat Mgr. Cyrilu Bromovi, PhD. za odpovědi na mé otázky ohledně výukových simulací a podnětné připomínky.

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 29. července 2013

.....

podpis studenta

Abstrakt

Předložená bakalářská práce se zabývá možnostmi využití metody web scrapingu při tvorbě výukových simulací.
V úvodu první části se autor věnuje rozdílům mezi pojmy information scraping, web scraping a screen scraping a vytvořením definice po zbytek práce používaného pojmu web scraping. Dále se věnuje historii a kontextu této metody zejména v americkém prostředí a na jednoduchém příkladu ukazuje, jak může vypadat web scraper.
Rozebírá i příklady použití web scrapingu, na které je možné na internetu narazit a popisuje konkrétní projekty, které metody využívají. Závěrem první části se věnuje složité právní situaci, která kolem web scrapingu panuje a zmiňuje etické problémy, na které je možné v souvislosti s používání scrapingu narazit.
V druhé části se krátce věnuje výukovým simulacím a prozkoumává možnosti využití web scrapingu pro tvorbu a aktualizaci výukových simulací, včetně typických znaků datové struktury simulací, které jsou pro použití s web scrapingem vhodné.
V závěru je představen návrh příkladu výukové simulace z oblasti energetiky, která by mohla z použití scraperů ve fázi tvorby i ve fázi aktualizací těžit.

Klíčová slova: information scraping, web scraping, screen scraping, data mining, výuková simulace, edutainment

Abstract

This bachelor thesis deals with possibilities of the use of web scraping method in the development of educational simulations.
Author notes differences between the terms information scraing, web scraping and screen scraping and defines term web scraping, which is used for the remander of the thesis. He introduces history nad context of this method especially in the american background and shows simple example of a web scraper.
Examples of the use of web scraping that can be seen on the Internet are described and analyses particular cases of projects that make use of this method. In the end of the first part the thesis deals with complicated juridical situation around web scraping and notes ethical difficulties related to the use of scraping.
Author then briefly deals with educational simulations and explores the possibilities of the use of web scraping in thedevelopment and updating of simulations including typical structure of data used in simulations that are suitable to be developed with help of scrapers. In the final part a draft of educational simulation about energetics, which could profit by the use of scrapers both in the process of development and in the process of updating, is introduced.

Keywords: information scraping, web scraping, screen scraping, data mining, educational simulation, edutainment

Obsah

1 Úvod

Tato bakalářská práce se zabývá metodou information scrapingu, která je často využívaná a je nedílnou součástí některých notoricky známých webových služeb, jako jsou například webové vyhledávače, a možnostmi implementace této metody do procesu tvorby výukových simulací.

K výběru tohoto tématu mě vedl osobní zájem a úvaha, že metoda, která umožňuje automatické získávání velkých objemů informací by mohla být komplementární k potřebám simulací, které často vyžadují množství dat, aby mohly vytvářet model komplexního systému. Některé výukové simulace pracují s daty, která by měla být co nejaktuálnější a za tím účelem je třeba každoročně systém upravovat. S použitím scraperů by mohlo být možné aktualizace dat alespoň částečně automatizovat a tím celý proces zjednodušit.

V úvodních částech práce se budu věnovat samotné metodě information scrapingu, její historii a kontextu. Pokusím se představit, kde se metoda používá, a na konkrétních případech její užití ukázat. Rád bych též zmínil základní etické a právní specifika, které se scrapery souvisí. Poté se budu věnovat výukovým simulacím a tomu, jestli je možné při jejich tvorbě scrapingu využít. Nakonec bych rád zpracoval krátký návrh toho, jak by mohla vypadat výuková simulace z prostředí energetiky, která by využívala potenciálu information scraperů.

Cílem práce je analyzovat možnosti využití scrapingu při tvorbě a aktualizaci výukových simulací a navrhnout příklad konkrétní simulace, která by s touto metodou pracovala.

Veškeré bibliografické záznamy jsou zpracovány dle citační normy ISO 690:2010.

2 Information scraping

2.1 Definice

V souvislosti se scrapingem se můžeme setkat s několika termíny, které jsou běžně používány - obecné termíny information scraping a data scraping a specializované termíny screen scraping a web scraping. Pravděpodobně nejčastěji používaným je screen scraping, který existuje již od dob prvních grafických uživatelských rozhraní a v souvislosti s webem se používá, byť v poněkud odlišném významu, od jeho vzniku v 90. letech. Pro účely této práce bude pravděpodobně nejdůležitější termín web scraping, který zužuje téma pro použití s daty zobrazovanými ve webových prohlížečích.

Není snadné se dobrat k jednoznačné definici těchto termínů, protože je možné najít data z různých zdrojů, které se zcela neshodují. To samozřejmě do jisté míry souvisí s tím, které zdroje na information scraping nahlíží z jakého pohledu (tedy jestli z pohledu programátora, správce serveru, novináře,...), ale i mezi podobně zaměřenými definicemi je možné nalézt rozdíly. Proto bych chtěl v následující části ukázat některé z definic, které jsou dostupné na internetu a v některých výkladových slovnících, a na jejich základě postavit jednoznačnou definici, podle které budu s termínem v rámci této práce dále nakládat.

Na internetu poměrně rozšířená definice pochází z encyklopedie Computer Desktop Encyclopedia od firmy The Computer Language Company Inc., kterou používají jako zdroj termínů počítačové a informační vědy například webové stránky pcmag.com nebo thefreedictionary.com.

Screen scraping

Získávání dat viditelných na obrazovce manuálně, pomocí příkazu uložit, nebo s pomocí programu. Webové stránky jsou neustále cílem scrapingu za účelem ukládání užitečných dat pro budoucí použití. Má-li být scraping automatizovaný, program musí být napsán tak, aby rozpoznal konkrétní data. Viz scraping a screen scraper.[1]

1 Vlastní překlad z anglického originálu: „screen scraping(redirected from data scraping) - Acquiring data displayed on screen by capturing the text manually with the copy command or via software. Web pages are constantly being screen scraped in order to save meaningful data for later use. In order to perform scraping automatically, software must be used that is written to recognize specific data. See scraping and screen scraper.“

Screen scraper

(1) Označován též jako frontware, může být programem, který přidává grafické uživatelské prostředí (graphical user interface - GUI) do textových systémů mainframů nebo starších mini-počítačových aplikací. Screen scraper je spuštěn na počítači uživatele a připojuje se k systému pomocí sítě nebo emulátoru terminálu. Viz terminal emulation.

(2) Program, který automaticky získává data ze zobrazeních určených k uživatelské interakci bez příčinění uživatele. Tento význam byl primárně používán v době textových terminálů.

(3) Program, který automaticky získává data z HTML stránek, nebo jiných dokumentů, které jsou normálně používány pro interakci s uživatelem.[2]

Scraping

(1)Získávání dat z výstupu určeného pro obrazovku nebo tiskárnu, na místo použití původních souborů či databází. Například webové stránky formátované pomocí jazyku HTML jsou často cílem scrapingu. Scraping je způsob jak získat data bez přístupu k formátům a databázím v kterých jsou normálně uloženy. Pomocí webových stráneki a vyhledávačů může být na uživatelově obrazovce zobrazeno a staženo obrovské množství informací. Viz screen scraping.

(2)Získávání e-mailových adres nebo dalších dat z rozsáhlých webových stránek nebo vyhledávačů. Boti jsou použiti ke sbírání dat, která pak mohou být přeorganizována a prezentována speciálním způsobem, který má přilákat návštěvníky za účelem většího výdělku na reklamách. Data mohou být též použita pro zločinné účely, jako prodej informací spammerům a cyberzločincům.[3]

---

2 Vlastní překlad z anglického originálu: „screen scraper

(1) Also called "frontware," it is software that adds a graphical user interface (GUI) to character-based mainframe and earlier minicomputer applications. The screen scraper runs in the user's computer and connects to the mainframe or mini via the network or terminal emulation. See terminal emulation.

(2) Software that automatically extracts data from interactive screens without user intervention. This usage of the term was primarily in the days of character-based terminals.

(3) Software that automatically extracts data from HTML pages or other documents that are normally viewed interactively by the user (see screen scraping).“

3 Vlastní překlad z anglického originálu: „scraping

(1) Extracting data from output intended for the screen or printer rather than from original files or databases. For example, Web pages formatted in HTML are often scraped. Scraping is a way to obtain data without having access to the formats and databases the data are naturally stored in. With Web sites and search engines, an enormous amount of data can be brought to the user's screen and captured. See screen scraping.

(2) Extracting e-mail addresses or other data from a large Web site or search engine. Bots are used to cull the data, which may be reorganized and presented in a unique manner that attracts visitors, the purpose of which is to make money on ads. The data might also be used for nefarious reasons, such as selling the information to spammers and cybercriminals.“

Webový slovník dictionary.com má databázi hesel aktualizovanou až do roku 2010, ale termíny data, information ani web scraping nezná a z termínu screen scraping automaticky přesměruje na termín screen scraper, který obsahuje definici z roku 1995, jež se vztahuje k použití screen scrapingu pro využití společně se systémy z doby před vznikem webu za účelem využití grafických rozhraní.

Server wisegeek.com, který si dal za cíl zodpovídat časté dotazy na internetu, obsahuje článek v odpovědi na otázku “Co je screen scraper.” Článek označuje HTML scraping a web scraping jako synonyma screen scrapingu a definuje je jako program, který získává textová data z výstupu jiného programu, odstraňuje formátování, (vytvořené například jazykem HTML) a umožňuje další úpravy (například indexaci stránek, jako to dělají web crawlery.) Dále se článek zmiňuje o použití scrapingu v podnikání jakožto zdroje kompetetivních informací a zabývá se sociálními problémy, které mohou vznikat při nezodpovědném používání této metody.

Stránka webopedia.com neobsahuje termíny information či data scraping, ale rozlišuje screen scraping, jakožto způsob zobrazování dat (zejména) ze starších systémů (tzv. middleware) a web scraping, který definuje takto:

Web scraping je aplikací, která zpracovává kód HTML webové stránky za účely jako například konverze stránek do jiného formátu(i.e. z HTML do WML). Skripty a aplikace zabívající se web scrapingem simulují uživatele, který stránku zobrazuje pomocí prohlížeče. S těmito skripty je možné se připojit k webu a vyžádat si konkrétní stránku stejně, jako by to udělal prohlížeč. Webový server zašle zpět stránku, která pak může být dále upravována, či z ní mohou být extrahovány specifické informace.[4]

4 Vlastní překlad z anglického originálu: „Web Scraping refers to an application that processes the HTML of a Web page to extract data for manipulation such as converting the Web page to another format (i.e. HTML to WML). Web Scraping scripts and applications will simulate a person viewing a Web site with a browser. With these scripts you can connect to a Web page and request a page, exactly as a browser would do. The Web server will send back the page which you can then manipulate or extract specific information from.“

Na server about.com nejsou dostupné termíny information, data či screen scraping, ale poskytuje definici termínu web scraping a uvádí, že web scraping nejčastěji používají “spam” stránky a zloději webového obsahu, bez jakéhokoli udání zdroje těchto informací - navíc poskytuje odkaz na stránky o obraně proti web scrapingu, které přímo říkají, že web scraping se rovná krádeži. Proti důvěryhodnosti těchto stránek mluví fakt, že nabízejí objednávku nerůznějších řešení jak se vypořádat s nežádoucími scrapery, od knih, přes počítačové programy až po komplexní řešení od profesionálních firem. Na druhou stranu, ačkoli je text zjevně zaujatý, velmi dobře vystihuje náhled na information scraping, se kterým se je na internetu často možné setkat.

Server techopedia.com termíny data a information scraping neobsahuje a screen scraping staví na rovinu web scrapingu. Navíc zde nalezneme termín content scraping, který je popsán jako ilegální proces kopírování dat chráněných autorskými právy a jejich opětovnou publikaci na jiné webové sídlo(bez udání pravého autora) za účelem zvýšení publicity a výdělku pomocí reklam.

Web scraping

Web scraping je termín pro různé metody používané pro sbírání informací v prostředí Internetu. Obecně je toho dosaženo pomocí programu, který simuluje prohlížení webu lidským uživatelem za účelem získání konkrétních informací z různých webových stránek. Uživatelé programů utilizujících web scraping mohou vyhledávat konkrétní data k prodeji dalším uživatelům nebo k použití za účelem zvýšení publicity nějakých stránek.[5]

Na stránkách techterms.com se nachází pouze termín “scraping”, který ovšem určuje web scraping jako synonymum. Tato definice scraping určuje jako proces získávání většího množství informací z webu, včetně přímého stažení kompletních webových stránek. Připouští též možnost automatizace procesu pomocí k tomu určených programů a nebo vytvoření opakovaných přístupů pomocí botů.

---

5 Vlastní překlad z anglického originálu: „Web Scraping
Web scraping is a term for various methods used to collect information from across the Internet. Generally, this is done with software that simulates human Web surfing to collect specified bits of information from different websites. Those who use web scraping programs may be looking to collect certain data to sell to other users, or to to use for promotional purposes on a website.“

Nakonec na stránce wikipedia.org je data scraping uváděn jako mateřský termín termínům screen scraping a web scraping. Data scraping podle wikipedie je získávání informací z výstupu nějakého systému automatizovaným způsobem. Bohužel v popisu se termíny web a screen scraper několikrát zaměňují, takže není jejich definice jednoznačná (wikipedia.org, heslo data scraping 2013). Wikipedie, ale ve svém popisu termínu web scraping poskytuje popis nejrůznějších dílčích metod web scrapingu a také jako jedna z mála definic na internetu dostupných popisuje web crawlery a jak s web scrapingem souvisí (Wikipedia.org, heslo web sraping, 2013)

Na konci tohoto průzkumu definic je zjevné, že termín information scraping jsem nezvolil zcela vhodně a že lepší bude použít jeden z konkrétnějších termínů. Ačkoli termín screen scraping je častější a v mnoha zdrojích je s web scrapingem zaměňovaný, rozhodl jsem se pro účely této práce používat termín web scraping, protože se jedná o termín, který je novější, konkrétnější a nehrozí jeho spojení se staršími použitími screen scrapingu (například jako prostředníka mezi originálním textovým systémem a grafickým uživatelským rozhraním.)

Všechny výše uvedené definice se shodnou v tom, že scraping je metoda získávání dat, která jsou určená pro zobrazení uživatelem. Tato vlastnost je navíc pro uživatele scraperu velmi užitečná, protože nevyžaduje přístup přímo ke zdroji dat (jako jsou databáze a nejrůznější interní tabulky), který často není dostupný, nebo je dostupný pouze za určitých podmínek, a stává se tak často často důvodem proč se scrapingem vůbec začít (Alba 2008).

Vzhledem k tomu, že pro účely této práce budu používat termín web scraping, je vhodné, aby se v definici vyskytlo, že se jedná o metodu používanou v prostředí internetu.

Dalším důležitým bodem definice je, že se jedná o metodu automatizovanou. Na tom se sice neshodnou všechny výše uvedené definice, ale myslím si, že uvažovat manuální metody, jako prosté kopírování, by bylo pro účely této práce kontraproduktivní.

Většina těchto definicí se shodne na tom, že součástí scrapingu je též nějaká bazální úprava získaných dat (odstranění HTML kódu, vybrání konkrétních tabulek,...), která je podstatná pro další využití. S tím do jisté míry souvisí i to, že by výsledná data měla být uložena ve formátu, který bude pro další zpracování vhodný. Na základě těchto bodů jsem vypracoval definici, kterou budu po zbytek této práce používat.

Web scraping je metoda automatizovaného získávání dat z výstupu určeného pro zobrazení uživatelem v internetovém prohlížeči. Data jsou automaticky upravována a uložena ve formátu vhodném pro další použití.

2.2 Screen scraping

Termín screen scraping, se v prostředí internetu s web scrapingem často zaměňuje (viz. definice a určení synonym na jednotlivých stránkách, které jsem pro tvorbu definice použil), ale jako termín existuje mnohem déle. Sledování kořenů screen scrapingu je podle mého názoru velmi důležité, protože může pomoci pochopit kontroverzi, která kolem scraperů existuje a účel, za kterým původně vznikly, který nakonec není tak odlišný od důvodů pro užívání scrapingu v dnešní době.

Historie termínu vychází ze zdrojů od autorů žijících v USA a sleduje tedy vývoj ve Spojených státech - samozřejmě v jiných zemích se vývoj mohl více či méně lišit na základě různých politických a sociálních faktorů. Bohužel české zdroje, které by se problematice nějakým způsobem věnovaly, prakticky neexistují, přestože web scraping je i na českém webu již zjevně a delší dobu využíván. Osobně si ale myslím, že použití v českém prostředí bylo inspirováno právě úspěchy agregačních webů ve Spojených státech.

Pojem screen scraping byl původně používán v 90. letech zejména jako metoda, kterou se data z “legacy”(česky “odkaz” nebo “dědictví”, vztahuje se na systémy, které jsou zastaralé a neumožňují důležité funkce, které jsou považované za

standart nebo za předpoklad pro nějakou rozšířenou službu - v tomto případě prohlížení webu) systémů převáděla do grafického uživatelského rozhraní (net.dictionary 1997, Booker 1997). Cílem původních scraperů tedy bylo překládat chování uživatele v grafickém rozhraní do vstupu, který mohl být pochopitelný pro jinak textově orientovaný systém.

Ke konci roku 1999 se začaly na internetu objevovat agregační weby, které umožňovaly uživatelům po zadání svých přihlašovacích údajů, využívat služby všech bankovních účtů, které měli, nehledě na to, že byly od různých, na sobě zcela nezávislých bank. V jejich souvislosti se screen scraping (Weisul 2000, Rektwa 2000) stal velmi diskutovaným a kontroverzním tématem. V tomto momentě už bychom mohli mluvit o termínu web scraping, protože se jednalo o automatizované programy, které bankovním systémům předložili přihlašovací údaje uživatele, získali data o jejich službách a zůstatcích a uložily je na serveru agregátoru, který je poté zobrazoval v uživatelském rozhraní. Termín web scraping je nicméně novější, proto se ve všech zdrojích z této doby setkáme s termínem screen scraping nebo pouze “agregátor”.

Koncem roku 1999 a začátkem roku 2000 se většina bank ve Spojených státech nějakým způsobem vyjádřila k agregátorům (Financial Service Online 2000), ať již spoluprácí s některou z agregačních společností, jako např. yoodle.com (Power 2000), nebo zaujetím negativního postoje, varujíce své zákazníky před scrapingem a bezpečnostními riziky s tím spojenými. V některých případech tyto negativní postoje vyústily až v žalobu, ačkoli žádná z firem poskytujících služby spojené se screen scrapingem nebyla v této době shledána vinnou. Typickým příkladem může být v tomto případě banka First Union, která firmu agregující uživatelské účty zažalovala, ale do tří měsícu měla vypracovaná pravidla pro spolupráci se scrapery (The Disclosure 2001), žalobu stáhla a začala vyvíjet vlastní agregátor (Toonkel 2000a).

Přestože byla v té době idea agregačních webů nová, začátkem roku 2001 se již odhadovali počty uživatelů agregátorů na stovky tisíc (The Disclosure 2001).

Bankám nezbývalo než se s touto novou koncepcí smířit (Toonkel 2000b), protože zákazníci si ji žádali a neexistoval způsob, jak jim zabránit, aby informace o svých účtech nesdíleli s firmami, které agregátory provozovali. Scraping totiž nebylo snadné rozeznat od aktivity lidského uživatele, a snahy technologicky zamezit scrapingu pomocí průběžných změn v architektuře uživatelského rozhraní finálně vedly jen k další nevoli uživatelů, kteří si tak museli na časté změny zvykat.

Pro společnosti, které s agregátory přišly, byla situace, ve které o ně byl čím dál tím větší zájem nejen ze strany koncových uživatelů, ale také ze strany jednotlivých bank, výzvou. Mnoho z nich přešlo od poskytování služeb uživatelům na vývoj agregátorů pro jednotlivé banky (Hackett 2000). Některé takové agregátory místo dat ze screen scrapingu přijímaly data přímo ze systémů bank za pomoci interní API (The Disclosure 2001).

Od roku 2002 se objevují další a další agregátory, které už se nevěnují jen bankovním systémům, ale umožňují srovnávat ceny zboží, doby dodání, atp. Tyto weby hledají nová data, která lze scrapovat a dále využívat. Tento vývoj již podruhé během pár let nutí firmy aby se přizpůsobily, když staví vedle sebe možnosti a ceny jednotlivých služeb a výrazně tak zvyšuje informovanost klientely (Madnick, Siegel 2002). Scraping se tak stává běžnou součástí internetového ekosystému.

2.3 Web Scraping

.Web scraper je program, který naplňuje nějakou konkrétní informační potřebu a to buď jednorázově a nebo během delšího časového úseku. Jako zdroj používá webové stránky zpravidla napsané pomocí jazyku HTML. Za účelem získání konkrétních informací používá funkci, kterou nabízejí nejrůznější knihovny - mohu zmínit třeba lxml, Beautiful Soup, jTidy, html Parser, nebo HTML::Parser pro různé programovací jazyky, jakými jsou například Python, Ruby, PHP, Perl, Java, atp. Tato funkce se nejčastěji označuje jako parsing resp. html parsing(Wikipedia, lxml.de, htmlparser.sourceforge.net, whatwg.org, foo.be 2013).

Podle slovníku The American Heritage® Dictionary of the English Language

(2000) je parsing synonymem pro syntaktickou analýzu a v prostředí počítačové vědy má následující význam:

“To analyze or separate (input, for example) into more easily processed components.”

Podrobit analýze nebo rozdělit (na příklad data ze vstupu) na snáze upravovatelné komponenty.

V případě jazyka html to znamená rozebrání zdrojového kódu na jednotlivé tagy (v případě párových tagů přiřazení koncových tagů) a vytvoření stromu (tzv. Parse Tree), který umožní další úpravy (Wikipedia, Webopedia.com 2013). Součástí parsingu je i vyhledávání ve zdrojovém kódu, které je pro web scraping v jakékoli jeho formě velmi zásadní. Ať už je cílem scrapingu pouze odstranit tagy a naformátovat celý text nějakým konkrétním způsobem, abychom získali dobře čitelný a snadno šířitelný dokument, indexování informací (jak to dělají web crawlery), nebo výběr určitého typu informací, je vyhledávání něčím, co musí být programátorem pečlivě zváženo a vhodně navrženo.

Stejně jako v případě jakékoli jiné informační metody (ať již je jejím cílem prosté získávání informací či tvorba komplexního informačního systému) je i v případě použití web scrapingu velmi důležité mít jasně definovanou informační potřebu (Kendal 1991, Řepa 1999, Polák 2003). Od toho se nutně odvíjí celková podoba scraperu - jaké informace získává, kolik a jakých zdrojů využívá a v jakém formátu je ukládá.

2.3.1 Příklad tvorby web scraperu

Myslím si, že je vhodné proces tvorby web scraperu ukázat na konkrétním příkladu. Aby bylo použití názorné, mělo by se jednat o větší množství dat, jejichž získání, pokud by mělo probíhat manuálně, by bylo časově náročné. Proto jsem si jako cíl scraperu v tomto příkladu vybral zjištění průměrných denních teplot naměřených v Praze v průběhu celého roku tak, aby data byla co možná

nejaktuálnější. Využití pro takové informace může být různé (počítání spotřeby vytápění, plánování dovolené podle průměrných teplot za poslední roky, vytváření teplotních srovnání, atp.), ale pro účely tohoto příkladu nebudu výstupní data přizpůsobovat jejich finálnímu využití, místo toho je nechám ve formátu, díky kterému bude možné s daty dále snadno manipulovat.

Prvním krokem bude nalezení vhodných zdrojů, ze kterých by scraper mohl čerpat. V praxi je často možné se setkat s tím, že je jasně dané, který informační zdroj bude cílem scraperu (Merill 2013, Ornstein 2013), zejména v případě že se podobné informace na jiných místech nenacházejí - pak je samozřejmě možné tento krok přeskočit. V tomto případě, však se jedná o informace běžně a veřejně dostupné a je tedy možné si vybírat cíle z mnoha stránek, které meteorologická data poskytují. Volba scrapované stránky bude mít velký dopad na podobu kódu a na úplnost a spolehlivost získaných dat, takže je užitečné zprvu vybrat více možností a poté z nich vyčlenit vhodného kandidáta. Pro účely tohoto příkladu jsem nalezl čtyři meteorologické weby, které poskytují data o průměrných denních teplotách: wunderground.com, pr-asv.chmi.cz (webové stránky Českého hydrometeorologického ústavu), in-pocasi.cz a pocasi.divoch.net. Podobných webů se patrně na internetu nachází více, na druhou všechny výše uvedené na první pohled vypadají jako vhodné cíle pro scraping, takže širší vyhledávání by mohlo být kontraproduktivní. V případě, že by žádná z vyhledaných stránek vhodná nebyla, je samozřejmě možné vyhledávání dodatečně rozšířit.

Všechna uvedená webová sídla poskytují požadované informace (tedy průměrné denní teploty v Praze). Na serverech wunderground.com a pocasi.divoch.net je zdrojem dat meteorologická stanice v pražské Ruzyni. Web in-pocasi.cz zdroj neuvádí. Na stránkách ČHMÚ (Český hydrometeoroligický ústav) nalezneme data z dvou pražských meteorologických stanic (Praha-Ruzyně a Praha-Libuš), navíc, vzhledem k tomu, že se jedná o oficiální web těchto stanic a je zaštiťován ústavem zřízeným státem, je možné že informace na ostatních stránkách pocházejí právě z tohoto webu. Web pr-asv.chmi.cz se tak jeví jako nejlepší řešení, protože poskytuje údaje z dvou různých stanic ve vybrané lokalitě a protože se data, která poskytuje, zdají jako nejspolehlivější.

Druhým krokem je rozbor zdrojového kódu (Nguyen 2010). K tomu je možné použít funkci tzv. “web inspector”, kterou má v nějaké podobě většina používaných webových prohlížečů a která umožňuje zobrazit přímo zdrojový kód elementu, který je momentálně označen (např. v google chrome kliknutím pravého tlačítka a výběrem možnosti “Zkontrolovat prvek”). To přináší možnost poměrně snadno získat přehled o zdrojovém kódu a zjeména o tom, kde v kódu se nachází potřebná data (Nguyen 2011). V tomto případě je po prohlédnutí kódu jasné, že jednotlivé teploty jsou na stránce uváděny v párovém tagu s třídou (class) označenou jako “vgtext”. Teploty v získané z pražských meteorologických stanic jsou tedy ve zdrojovém kódu zapsány takto:

“ -0.3

-1.2 ”

Pravděpodobně nejsnažší způsob jak vybrat potřebná data, je pomocí třídy tagu (Davies 2012). Ruzyně a Libuš jsou 27. a 28. výskytem tagu s třídou “vgtext” na stránce. Cílem scraperu bude tedy lokalizovat tyto dva tagy, získat data, která jsou v nich zapsaná a uložit je v nějakém použitelném formátu. Také bude nutné, aby scraper tuto činnost prováděl každý den, protože se jedná o data aktuální.

Třetím krokem je výběr vhodných nástrojů pro tvorbu scraperu. To je do jisté míry otázkou osboních preferencí, protože možností je opravdu mnoho. Já jsem použil nástroje dostupné na webu scraperwiki.com, které umožňují ukládání dat do databáze scraperwiki a automatizované periodické spuštění scraperu. Příjemným bonusem je volba ze tří různých programovacích jazyků a tedy Pythonu, PHP a Ruby a předinstalované knihovny pro parsing a ukládání dat.

Čtvrtým a posledním krokem je vytvoření samotného programu. V tomto případě

byl program vytvářen v prostředí scraperwiki za pomocí jazyka Python a knihovny lxml. Dále uvádím jeho zdrojový kód, spolu s poznámkami.

```
import scraperwiki

#načtení knihovny scraperwiki, která umožňuje načítání zdroje pro scraping, ukládání dat do přednastavených tabulek a export do souborů s příponou csv

import lxml.html

#načtení knihovny lxml, která obsahuje nástroje pro parsing

html = scraperwiki.scrape("http://pr-asv.chmi.cz/synopy-map/pocasiin.php?ukazatel=prumtep&pozadi=mapareg&graf=ano")

root = lxml.html.fromstring(html)

#deklarování proměnné root, jako zdrojového kódu stránek ČHMÚ

data = {
        'Datum' : root.cssselect("div.vgtext")[0].text,
        'Praha_Libus' : root.cssselect("div.vgtext")[26].text,
        'Praha_Ruzyne' : root.cssselect("div.vgtext")[27].text
    }

#deklarování proměnné data, která bude zdrojem pro import do tabulek scraperwiki

#příkaz cssselect prohledává proměnou root vybere 1., 27. a 28. výskyt tagu <div> s třídou "vgtext" a jejich obsah uloží do jednotlivých proměnných, první výskyt tagu je datum, které je velmi vhodným unikátním klíčem, protože zaručí, že nebudou dva záznamy z jednoho dne a že budou teploty časově ukotveny

scraperwiki.sqlite.save(unique_keys=['Datum'], data=data)

#příkaz, který uloží tabulku v proměnné data do databáze sraperu na scraperwiki a utřídí jí podle unikátního klíče „Datum“
```

Ačkoli je tento scraper teoreticky plně funkční a díky nastavení na scraperwiki se spouští jednou denně a získává tak aktuální průměrné teploty z obou

meteorologických stanic v Praze, je třeba si uvědomit, že pro použití v konkrétním projektu by bylo vhodné udělat několik úprav.

Zjevným problémem scraperu v tomtu příkladu je totiž jeho naivita. Scraper sice funguje dobře, ale jen v případě, že veškerá data jsou vždy dostupná na stránkách ČHMÚ a že script, který spouští scraper na scraperwiki za všech okolností funguje dokonale. Za tři měsíce, po které jsem nechal program běžet (nechával jsem získávat data ze všech meteorologických stanic na pr-asv.chmi.cz aby bylo vidět, kde konkrétně se případný problém vyskytuje), jsem se setkal s občasnými výpadky scraperwiki (během prvních 30-ti dní se scraper spustil 26-krát) a s poměrně běžnými výpadky některých stanic - nikdy se sice nejednalo o pražské stanice, ale vzhledem ke stavbě scraperu založené na pořadí teplotních dat v záznamu, to na výsledné data mělo vliv.

V době, ve které scraper fungoval, chyběla opakovaně data na stanicích v Přerově a v Čáslavi. Vzhledem k tomu, že jsou obě v seznamu před sledovanými pražskými stanicemi způsobuje jejich výpadek nesprávnost získaných dat. Nejsnažším řešením by bylo nechat scraper zkontrolovat počet uváděných teplot a pokud by se ukázalo, že některé stanice chybí, upravit pořadí stanic, ze kterých jsou data získávána. Na druhou stranu se jedná o velmi neohrabané řešení, které nebude fungovat, pokud se vyskytne nějaká neočekávaná chyba (například pokud by vypadla stanice, která je v seznamu za těmi pražskými). Vhodnější by v tomto případě bylo zaměřit se na jiná data, která poskytuje zdrojový kód stránek ČHMÚ a zlepšit metodu, podle které scraper konkrétní teploty vyhledává. V tomto případě se nabízí vlastnost “title” tagu , ve které jsou uváděny názvy meteorologických stanic podle místa, kde se nacházejí. Stejně by se dala využít i vlastnost “style”, ve které je uváděna jejich absolutní poloha na mapě.

Vyřešit nespolehlivost scraperwiki je náročnější. Samozřejmě by bylo možné ji zkrátka nepoužívat a najít způsob jak scraper periodicky spouštět jiným způsobem. Na druhou stranu model, ve kterém je nutné aby program běžel bez výjimky každý den, je problematický, protože je snadno rozhozen jakoukoli

neočekávanou událostí. Zjevné řešení by bylo nechat scraper prozkoumat teploty pár dní do minulosti a ujistit se, že všechna data jsou v pořádku. Bohužel stránky ČHMÚ nemají archiv průměrných denních teplot. Ostatní stránky, které jsem zmiňoval na začátku tohoto příkladu naštěstí archivy mají, takže je možné vytvořit scraper, který by jednou týdně získával teplotní data za celý měsíc (například v případě serveru wunderground.com se data za celý měsíc zobrazují na jedné stránce, takže dopad na vytížení serveru bude prakticky nulový). Takový scraper sice nebude poskytovat data, která by byla denně aktuální, ale bude mnohem spolehlivější - samozřejmě není problém oba koncepty spojit a získat tak data, která budou i spolehlivá i aktuální (ať již použitím dvou scraperů nebo vytvořením jednoho komplexnějšího).

2.4 Použití web scrapingu v praxi

2.4.1 Web crawlery

Jedním z nejviditelnějších příkladů použití web scrapingu v praxi jsou tzv. web crawlery. Ty jsou speficifkým typem web scraperu, které procházejí webové stránky a indexují data, která nacházejí a ukládají je pro další použití (Kaefer 2013). Mohou být použity pro automatizované vyhledání dat na konkrétní téma, pro zmapování nějakého internetového trendu nebo v extrémním případě pro systematickou indexaci webových stránek, jak je tomu u webových vyhledávačů, jakým je například Google web search (Newth 2013, Google 2013). Web crawlery vyhledávačů procházejí jednotlivé webové stránky skrze odkazy, které se na nich nachází a automatizovaným procesem z nich vytváří index, který je zdrojem dat pro výsledky hledání zobrazované uživatelům. Dalším odvětvím, které crawlery využívá, je lingvistika, kde umožňují vyhledání slov a slovních spojení typických pro webové konverzace a články a usnaďňují tak například tvorbu jazykových korpusů (Kaefer 2013). Mohou být též použity pro získání většího množství dat na určité téma například pro účely výzkumu nějakého trendu.

Web crawlery pracují z pravidla s velkými objemy dat (např. komentáře k novinovým článkům vybraneho tématu a jazyku v případě lingvistiky, všechny dostupné webové stránky v případě vyhledávačů,...). Velkou výhodou web

crawlerů je, že jim není potřeba zadávat konkrétní strukturu, ve které mají hledat, ani konkrétní webové sídlo jako cíl (Google 2013). Problémem naopak je, že získaná data nejsou zcela spolehlivá, protože ne každá webová stránka je pro crawlery snadno přístupná a nestandartně formátované stránky mohou přinášet klamavé výsledky (Google 2013). Další nevýhodou je fakt, že web crawlery se nejsou schopny dostat do hlubokého webu a přináší tak výsledky pouze z webu povrchového (Najork 2013). Výstup získaný crawlery je velmi vhodný pro případy kdy je třeba získat větší množství dat, která nemusí být stoprocentně relevantní, ale v případě, že je vyžadována přesnost a záruka pravdivosti obsažených informací, je třeba výsledky dále kontrolovat.

2.4.2 Mediální a novinářská činnost

Web scraping je používán také v rámci novinářské činnosti. Podle novináře Paula Bradshawa (2012), je pro novináře velkou výhodou, když dokáže mít informace rychleji než ostatní a má-li jich více. Web scraping jako metoda má potenciál v mnoha případech tuto výhodu poskytnout. Díky nejrůznějším programům, které umožňují provádět scraping pomocí grafického uživatelského prostředí není třeba, aby měl novinář programátorské dovednosti (The data journalism handbook 2013). Takové programy mohou ušetřit mnoho kopírování a umožnit získat data, když nejsou dostupná pomocí API nebo nebo vhodného formátu (např. jako excelová tabulka). Na druhou stranu se znalostí kódu je možné upravovat a získávat data mnohem flexibilněji (Nguyen 2012). Ve světě existuje celá řada mediálních projektů a článků, které využívají možností web scrapingu.

Velmi zajímavý projekt je například webová aplikace Dollars for Docs od firmy ProPublica (Merrill 2013). Ta obsahuje data o doktorech v USA, kteří přijímali peníze od farmaceutických společností za propagaci konkrétních léků mezi lety 2009-2012. Tato data byly firmy donuceny zveřejnit americkou vládou a často tak udělaly na svých stránkách v nejrůznějších formátech, které nejsou vhodné pro kopírování a v některých případech byl problém byť i s jejich stažením (Ornstein 2013). Aplikace Dollars for Docs všechna tato data pomocí web scrapingu shromáždila a udělala z nich přehlednou databázi, ve které je možné snadno

vyhledávat a roztřídit data podle států, společností i jednotlivých doktorů. Na základě tohoto projektu vzniklo také několik článků v lokálních novinách jednotlivých států USA (Nguyen 2010).

Dalším příkladným projektem je automaticky updatovaná stránka Mug Shots, která je spravovaná zaměstnanci z novin Tampa Bay Times. Web scraper stahuje informace o všech policií zadržených lidech ve čtyřech krajích, kterými se noviny zabývají. Informace získává ze stránek tamních policejních stanic a spolu se jmény a základními údaji je shrnují na jedinou stránku. Kromě seznamu stránka poskytuje i různé statistiky a vyhledávání v databázi (Mug Shots 2013).

Web scraping může novinářům poskytnout plně automatizovaný přísun dat ze zdrojů, které používají jednotné formátování (například výše uvedené data o zadržených osobách) a velmi usnadnit práci s informacemi z rozsáhlých a/nebo nevhodně formátovaných zdrojů. Novináři často používají programy, které jim pomohou se získanými informacemi dále pracovat a upravovat je. Vhodně použitý scraper s pomocí vhodného nástroje (např. OpenRefine) může data stáhnout upravit, setřídit a do jisté míry i ověřit jejich úplnost (Bradshaw 2012, OpenRefine 2013, Šverák 2010).

2.4.3 Agregační weby

Agregační weby byly v zásadě prvním využitím web scrapingu, které dosáhlo velké pozornosti veřejnosti (viz kapitola screen scraping). Za posledních patnáct let prošly agregátory velkým vývojem a jsou dnes běžnou součástí internetu (Beatie 2010). Ne každý agregační web je však výsledkem práce scraperů, velká část agregátorů preferuje stahování informací z k tomu určených kanálů a využívá tak funkce, které mnoho webových stránek poskytuje - API, RSS, a XML feedů. Web scraping však zůstává jedinou možností agregace v případě stránek, které takové kanály neposkytují a možnou alternativou, pokud nejsou považovány za dostatečně spolehlivé (Alba 2008, Hartley 2012).

Typickým agregačním webem je služba Google News, která agreguje zprávy z různých webů podle preference a jazyka uživatelů a to plně automatizovanou formou (Co jsou Google zprávy 2013). Protože jsou Google News postaveny na automatickém sběru dat, je možné poměrně snadno utilizovat nápady na specializované agregační služby, jako například agregátor novinek o volbách a jednotlivých kandidátech (Čížek 2013). Web Mug Shots zmiňovaný výše, by také mohl být považován za agregační web, protože stahuje a dává dohromady data ze čtyř různých míst. Z českého prostředí bych mohl uvést například webové stránky heureka.cz, které agregují zboží z různých e-shopů za pomocí jejich API, stránku mixo.cz, která agreguje portály webových slev a jejich agregátory za pomocí XML feedů, nebo stránku webtrhy.cz, která dává dohromady data o jednotlivých webtrhových portálech na základě RSS feedů. Bohužel žádného agregačního webu v českém prostředí, který by přímo využíval metody web scrapingu si nejsem vědom.

Dalším příkladem agregačních webů jsou stránky, které jsou na internetu známe pod pojmy “scraper site” nebo “spam site” (Scraper site v News Agregator 2010, About.com 2011). Jedná se o web, který pomocí scraperů stahuje plné texty článků z novin, časopisů a blogů, integruje je na jednom místě společně s větším množstvím internetových reklam za účelem výdělku. Často takové stránky neuvádí své zdroje a autory jednotlivých článků, v každém případě však texty shromažďují bez jejich svolení a odebírají tak pozornost, a s ní spojené výdělky autorům a webům, na kterých byli články původně publikovány. Tento typ agregačních webů je jedním z důvodů, proč je web scraping v některých zdrojích považován za metodu přímo spojenou s krádeží obsahu, nebo metodu, která je sama o sobě nelegální.

Podskupinou agregačních webů jsou tzv. “mashupy”. To jsou webové stránky nebo aplikace, které kombinují data a/nebo funkce z dvou a více zdrojů za účelem vytvoření nové služby (Peenikal 2009). Od klasického agregátoru je tedy odlišuje to, že kromě dat kombinují i funkcionalitu různých aplikací. Dobrým příkladem by mohla být například stránka, která by získávala data z databáze pražských restaurací, zobrazovala je pomocí nějaké mapové aplikace (třeba Google Maps) a

při výběru restaurace zobrazila guestbook, kam by návštěvníci mohli psát své dojmy. Mashupy jsou běžnou součástí internetu a jsou často používány novináři či jednotlivými uživateli. Podle výzkumu v roce 2007, se ale pouze 21 procent dotázaných firem vyjádřilo, že by o použití mashupů uvažovalo (Clarkin, Holmes 2007). Zájem o mashupy pro firemní použití však vzrůstá a objevují se články s velmi přesnými návody jak vytvořit mashup pro firmu a jaká pozitiva to s sebou může nést (Crupy, Warner 2009).

2.4.4 Další informační potřeby

Se všemi výše uvedenými příklady použití web scrapingu se již v nějaké formě průměrný uživatel internetu setkal, nebo alespoň setkat mohl, protože se jedná o příklady, jejichž výsledky jsou veřejně dostupné. To samozřejmě není pravidlem a v praxi se objevují i případy použití pro interní potřeby firem, nebo jednotlivých uživatelů. V případě jednotlivých uživatelů jsou příklady běžně dostupné ať již na stránkách, které web scraping nějakým způsobem usnadňují (jako např. ScraperWiki), nebo na stránkách, které pouze poskytují fóra komunitám, které scrapery vytvářejí (např. StackOverflow). Jako příklad, který naplňuje prostou uživatelskou informační potřebu mohu použít scraper, jehož autor chtěl vědět, zda-li články uveřejňované v rubrice Coment is Free časopisu The Guardian pod tématem Atheism jsou frekventovanější v závislosti na měsíci, respektive na křesťanských svátcích s daty spojenými (Hawker 2013). Za tímto účelem vytvořil jednoduchý scraper, který sbírá počty článků v dáné rubrice a tématu a zobrazuje je v grafu, který ukazuje procentuální zastoupení článků v jednotlivých letech a pro každý rok ještě v rámci jednotlivých měsíců. Bez použití web scraperu by podobná analýza mohla trvat poměrně dlouho, nicméně s pomocí ScraperWiki celý program zabral cca 120 řádků kódu.

Získání informací o tom, jak scraping interně využívají firmy je výrazně náročnější, protože firmy se o svých interních metodách zpravidla příliš nešíří. Infomace je však možné čerpat ze stránek, které nabízejí tvorbu scraperů pro firmy na zakázku nebo z článků, které se reflektují nějaký konkrétní příklad, kdy bylo použití scraperu zviditelněno - například, když o něj byl veden soudní proces. Jednou

možností jak firmy mohou web scraping využít je získávání cen zboží a/nebo služeb konkurence a určení obchodní strategie na jejich základě (Use of Web "scraper" did not violate CFAA in absence of notice of site restrictions 2003). Manuálně taková práce může trvat velmi dlouho a vyžaduje procházení celé nabídky. S použitím web scrapingu je možné data porovnat zcela automatizovaně a poté pouze zkontrolovat výsledky. Další možností použití scrapingu pro interní firemní účely je migrace dat do nových formátů (Scraperwiki 2013). V tomto případě není možné mluvit o web scrapingu, ale o data scrapingu, protože data nemusí být nutně ve formě webových stránek, ale též ve formě klasických dokumentů. Zejména starší dokumenty ve formátu pdf jsou vhodným kandidátem pro scrapery, protože manuálně je velmi nesnadné s nimi dále pracovat. Naopak při použití scrapingu může jeden program snadno převést velké množství dokumentů, pokud jsou podobně strukturované.

2.5 Web scrapery a právo

Vzhledem k tomu, že web scraping jako metoda není nikde oficiálně popsán, nemohou existovat ani zákony, které by ho přímo upravovaly. Na druhou stranu, při používání scraperů a zejména při používání dat scrapery získaných, je možné narazit na konflikt se zákonem, zejména se zákony o autorských právech a zákony o informačních a informatických zločinech (Sokol, Smejkal 2009). V České Republice můžeme mluvit o zákonech č. 121/2000 Sb. (zákon o autorských právech) a č.40/2009 Sb. (trestní zákoník, zejména Část II. - Hlava V: Trestné činy proti majetku, §230-232), který vychází z Úmluvy o počítačové kriminalitě, schválené Výborem ministrů Rady Evropy 8. 11. 2001. V USA se jedná zejména o Copyright Act (United States Code, hlava 17, sekce 101) a o Computer Fraud and Abuse Act (United States Code, hlava 18, sekce 1030).

V českém prostředí, ačkoli to možná souvisí s nízkou informovaností ohledně scrapingu, je nesnadné najít konkrétní doporučení, či případy, které by se scrapingem přímo souvisely, naproti tomu v USA je takovýchto příkladů velké množství, proto se dále budu věnovat zejména situaci ve Spojených státech.

Nejzjevnější způsob jakým se může scraper stát předmětem soudního řízení je

porušení smlouvy (v originále breach of contract). V tomto případě se jedná zejména o porušení podmínek použití konkrétních stránek, či webového portálu. Pokud je někde na stránkách viditelný odkaz na podmínky použití, které stanovují, že činnost, kterou scraper provádí je zakázaná, musí to uživatel scraperu respektovat. V opačném případě může dojít k zaslání výzvy k ukončení protiprávní činnosti (v originále cease and desist letter) a případně k žalobě (Zabriskie 2009). Jak důležité jsou podmínky použtí pro vymáhání těchto zákonů je velmi dobře vidět na příkladu, kdy soud rozhodl, že živatel scraperu nemůže být odsouzen podle CFAA, protože na stránkách chyběly podmínky použití a uživatel tedy nemohl bez vší pochybnosti vědět, že jeho činnost překračuje jeho autoritu (Use of Web "scraper" did not violate CFAA in absence of notice of site restrictions, 2003).

Další možnost právního konfliktu je v případě Computer Fraud and Abuse Act (dále jen CFAA). Jedná se o část trestního práva, které se zabývá podvody a zneužitími v rámci počítačových sytémů. Podle znění, ale případná žaloba na základě CFAA vyžaduje, aby svou činností scraper: za a) přistupoval k datům na serveru zcela neoprávněně nebo překračujíc svou autoritu (tedy autoritu uživatele, který scraper používá) a za b) způsobil svou činností serveru škody (popsané jako citelné zpomalení či výpadek činnosti serveru) nebo finanční ztráty (alespoň v hodnotě 5000 dolarů). Kvůli těmto podmínkám je právní situace poměrně nejistá (Zabriskie 2009), protože soudy se neshodnou na přesné definici oprávněného přístupu a škody, které scraper může způsobit se většinou nedají snadno vyčíslit. Jako příklad mohu uvést případ zažalované firmy BoardFirst (Southwest Airlines Co. v. BoardFirst, L.L.C. 2007), která používala scraper, aby svým zákazníkům umožnila za poplatek získat preferovaná sedadla v letadlech Southwest Airlines. Soud firmu neshledal vinnou z porušení CFAA, protože Southwest Airlines nebyly schopné prokázat finanční ztrátu (protože služba BoardFirst de facto fungovala jako reklama, která na stránky aerolinií přitahovala nové zákazníky), navíc soud konstatoval, že použití systému rezervace letenek k rezervaci letenek nelze považovat za neoprávněné použití nehledě na využitou technologii. Firma Boardfirst však byla uznána vinnou pro porušení smlouvy, na základě podmínek použití rezervačního systému Southwest Airlines, které jasně stanovili, že je portál určen pouze pro osobní účely a později dokonce přímo zakazovali použití automatizovaných prostředků pro rezervace. Zástupci BoardFirst navíc byli

nadevší pochybnost obeznámeni s podmínkami portálu, protože v průběhu existence firmy dostali dvě výzvy k ukončení protiprávní činnosti.

Dalším příkladem by mohl být výrazně medializovaný případ A. Auernheimera, který zjistil, že databáze s interními daty o uživatelích firmy AT&T nevyžaduje žádné přístupové údaje kromě fyzického kódu jejich iPadu, a pomocí programu, který generoval fyzické kódy a scrapoval k nim příslušící údaje, získal maily více než 100.000 uživatelů (Blaze 2012, Auernheimer 2012). Ty poté publikoval, údajně za účelem upozornění na chyby v databázi. Byl zažalován na základě CFAA a odsouzen na 41 měsíců vězení (Zetter 2013). Soud rozhodl, že stahování v takovém rozsahu nebylo nutné, a že na základě jeho komunikace na IRC jeho pravým motivem bylo zviditelnit sebe a zdiskreditovat AT&T.

Copyright Act (dále jen CA) je další z věcí, na které si uživatel web scrapingu musí dát pozor. Jedná se o soubor zákonů, které upravují autorské právo a podmínky, za kterých se uplatňuje. V případě web scrapingu se jedná zejména o různé novinové a odborné články, blogy a zdrojové kódy. Číselná data a prostá fakta bez expresivního textu pomocí CA chráněna nejsou. Za normálních okolností samotné kopírování nelegální není, ale v případě, že se v podmínkách používání stránky píše, že je zakázané využívání automatizovaných metod a/nebo komerční použití získaných informací a uživatel to nerespektuje, může se stát že soud rozhodne, že se nejedná “fair use” a uživatel se tak dopustí porušení autorského zákona (Zabriskie 2009). Jako příklad je možné uvést proces s A. Swartzem který byl obviněn ze stahování velkého množství odborných článků z placené databáze firmy JSTOR (Kravets 2012). Program, který k tomu použil technicky vzato nebyl scraper, protože stahoval dokumenty ve formátu .pdf a nijak je dále neupravoval, ale zákonný princip to nijak neovlivňuje. Byl zažalován státním zástupcem a souzen za 13 různých přečinů spadajících zejména pod CA a CFAA. Trest se mohl vyšplhat až na 50 let vězení a pokutu milion dolarů.

The Digital Millennium Copyright Act (United States Code, hlava 17, sekce 1201; dále jen DMCA) může v některých případech s použitím scraperů také kolidovat. Tato sekce zákoníku říká, že je trestné obcházet jakékoli ochrany technického charakteru při přístupu k autorským dílům. Fakticky je tento zákon je velmi zřídka

používán proti běžným uživatelům scraperů, protože obcházet technické ochrany serveru vyžaduje vědomý zásah do programu scraperu. Poněkud speciálním případem v této kategorii jsou tzv. CAPTCHA (pozn.: zkratka pro Completely Automated Public Turing test to tell Computers and Humans Apart, jedná se o obrázky na kterých je text, který musí uživatel zadat, aby dokázal, že je člověk a webové stránky ho pustili dál). Některé scrapery zejména v podnikatelské praxi se snaží tyto systémy obcházet, což bohužel v zákonech Spojených států nechává další právní nejistotu, protože soudy nejsou jednotné ve svých rozhodnutích. Zjevným příkladem obcházení CAPTCHA je případ, ve kterém firma Ticketmaster, žalovala přeprodejní společnost Wiseguys za automatizované nakupování nejlepších lístků na významné koncerty a jejich následný prodej koncovým uživatelům (Zetter 2010). Na soudní proces nakonec nedošlo, protože se majitelé firmy doznali. Rozsudek byl poměrně mírný, ačkoli podle zákonů, jejichž porušení jim bylo přiznáno mohli dostat až pět let vězení, vyvázli s podmínkou a povinností částečné náhrady škod (Sisario 2011).

Jak je vidět, zákonů, které mohou scraping ovlivnit je poměrně velké množství, na druhou stranu, běžného uživatele ve většině případů nemohou ohrozit, protože aby vstoupily v platnost je třeba aby scraping probíhal ve větším měřítku, nebo působil vyčíslitelné škody. V českém prostředí je velmi těžké odhadnout, jak by podobné procesy probíhaly, protože právní systém ČR je velmi odlišný od toho amerického, a stejně tak celé pojetí ochrany duševního vlastnictví je v obou státech jen stěží srovnatelné. Nicméně se dá očekávat, že dříve nebo později se podobné případy objeví i v ČR s tím, jak se scrapery rozšíří více do informační praxe firem.

2.6 Sociálně zodpovědný web scraping

Jak vyplývá z předchozího textu, je web scraping uživateli internetu často vnímán s nelibostí a nahlížen jako neetický. Podle mého názoru, je to z velké části proto, že je o něm poměrně malé povědomí a nejviditelnější jsou ty příklady, které jsou nějakým způsobem negativní. Web scraper je velmi účinný nástroj, který může přinést mnoho užitku, a rozšířit možnosti získávání informací z internetových

zdrojů, jak z hlediska obsahu, tak z hlediska formátu. Je o všem důležité si uvědomit, že pokud je využíván neuvědoměle nebo bezohledně, může také napáchat mnoho škod.

Při rešerši jsem narazil jen na dva články, které se v nějaké míře věnovali tomu jak scraper vytvořit zodpovědně, přestože návodů na tvorbu jednodušších scraperů na internetu rozhodně není nedostatek. Občas se dokonce můžeme setkat s návody, které etické pravidla zcela ignorují a učí například jak obcházet nejrůznější limity paralelních dotazů za pomoci změny IP adresy (Hartley 2013). Z příspěvků v některých internetových fórech navíc vyplývá, že co se v rámci scrapingu považuje za etické není zcela jasné ani profesionálům, kteří se vytvářením scraperů živí (Scraper site v News Agregator 2010).

První bych rád zmínil náhled na etiku scrapingu ze stran autorů, kteří se jí do jisté míry zajímali a metody, které navrhují aby autoři scraperů dodržovali, pokud se neetickému použití chtějí vyhnout. Poté představím vybrané případy, ve kterých autoři scraperů, podle mého názoru, etickou stránku věci nepromysleli dostatečně, či jednali přímo způsobem, který by za neetický mohl být považován.

V dokumentu Data manipulation for science and industry je etika web scrapingu pojata velmi umírněně s tím, že tvůrce scraperu by měl písemně požádat o svolení majitele webových stránek, pokud jejich obsah chce získávat automatizovanou metodou (Summet 2010). Autor textu argumentuje tím, že je mnoho důvodů, proč by mohl být obsah stránek určen výhradně pro lidské uživatele - například pokud majitel spoléhá na financování pomocí reklam obsažených na stránce, pokud stránky odkazují na nějaký pay-per-view obsah (jako je tomu například u portálu heureka.cz, kde obchodníci platí za každý uživatelský přístup z portálu do svého e-shopu), nebo pokud jsou data na stránce nějakým způsobem licencována. Ve všech těchto případech by unáhlené použití scraperu mohlo mít neblahé následky spojené s finanční ztrátou správce webu.

Naproti tomu článek An introduction to compassionate screen scraping představuje jediné, ale zato poměrně zajímavé pravidlo zodpovědného scraperu - scraper by neměl být rozeznatelný od běžného (byť vysoce aktivního) uživatele (Larson 2008). Podle autora lze toto pravidlo prakticky aplikovat s použitím tří dílčích prostředků - používání cache (samozřejmě tam kde je to možné), odmlky

mezi dotazy na server a nakonec získávání pouze potřebných dat (je třeba se zejména vyvarovat stahování redundantních dat jen proto, že se na serveru nalézají).

Podle mého názoru jsou oba tyto přístupy validní ukázkou etického používání web scrapingu, ačkoli samozřejmě není možné je univerzálně aplikovat. První přístup může být poněkud neohrabaný a zejména při používání informací, které jsou přímo určené k veřejnému šíření a nejsou upravované autorskými právy, pak může být snaha kontaktovat správce stránek zbytečně složitá či dokonce zcela bezúčelná. Při získávání informací z více serverů může tento postup též neúměrně prodlužovat dobu, která je na potřeba na tvorbu scraperu, a to do té míry, že by mohlo být výhodnější data získat prostým kopírováním. Druhý přístup zase může způsobit problémy například v případech, kdy cílové stránky používají pay-per-view obsah a reklamy, protože scraper zpravidla reklamy ignoruje - naopak obsah na principu pay-per-view může považovat za relevantní data a jeho činnost na webových stránkách tak bude pro provozovatele čistě ztrátová.

Myslím si, že nejdůležitější je si tvorbu scraperu předem dobře promyslet, uvědomit si jaké situace z činnosti scraperu na stránkách mohou vzniknout a snažit se předejít všem negativním vlivům, které by mohl provozovateli způsobit. Tam, kde je použití scraperů limitováno podmínkami použití nebo v rámci souboru robot.txt, je možné, že se jedná o krok ze strany provozovatele, který má být obranou proti automatizovaným metodám, které jsou použity bezohledně a zpomalují pak činnost serveru. V takových případech je dobré provozovatele kontaktovat s žádostí o výjimku a s návrhem jakým způsobem by scraper data získával, aby provoz stránek žádným způsobem neomezoval. Tvůrce scraperu by měl vždy přemýšlet nad dopadem svých programů a jednotlivé případy posuzovat individuálně.

Etické problémy spojené s neopatrným používáním web scrapingu jsou bohužel poměrně časté. Jedním z poslední doby mediálně známým problémem bylo použití scraperu na webových stránkách patientslikeme.com (Angwin 2010). PatientsLikeMe je diskuzní server, na kterém jednotliví pacienti mají vlastní profil, který informuje o jejich chorobách a lécích, které užívají a je tak příležitostí sdílet zkušenosti s ostatními lidmi postiženými stejnými chorobami. 7. května 2010

software, který server využívá na vyhledávání podezřelé aktivity účtů zaregistroval účet, u kterého se následně ukázalo, že byl vytvořen za účelem sběru dat a že s pomocí web scrapingu stahuje osobní informace pacientů pro účely firmy Nielsen Co., která mimo jiné dodává data z prostředí internetu velkým farmaceutickým společnostem. Naštěstí byl celý vyřešen během dvou týdnů pomocí žádosti o ukončení protiprávní činnosti, na kterou firma Nielsen reagovala kladně a s omluvou. Akce byla dokonce následována odstoupením šéfa Nielsen Co. a prohlášením jeho nástupce, který scraping PatientsLikeMe odsoudil a vyjádřil se, že se firma rozhodla podobné praktiky nadále neužívat. Podle mého názoru, se zde tvůrci scraperu dopustili bezohlednosti, jak proti jednotlivým uživatelům, tak proti majitelům serveru. Server sám prodává anonymizované informace o svých uživatelích, takže kopírování těchto dat jinou firmou by je připravovalo o možnost tyto informace efektivně prodávat. Samotní uživatelé pak byli vystaveni tomu, že jejich osobní data byla získána třetí stranou a někteří ztratili v důvěru, v to, že by server dokázal jejich data patřičně ochránit, a někteří dokonce služby severu přestali využívat. Podle mého názoru je obecně nevhodné scrapovat jakýkoli obsah hlubokého webu bez svolení majitele takového obsahu.

Dalším typickým problémem, který bohužel vzniká příliš často je porušení autorských práv s použitím scraperů. Velmi zjevným zástupcem tohoto problému jsou výše zmiňované scraper/spam weby (About.com 2011). Kopírování cizího obsahu za účelem výdělku je nejen velmi neetické, ale též proti zákonné. Podobný problém nastal též v případě Andrewa Auernheimera, který stahoval osobní data uživatelů Ipadů v databázi firmy AT&T pomocí web scraperu, byť údajně za účelem dokázání bezpečnostní chyby (Zetter 2013).

Je důležité si uvědomit, že ne veškerá data dostupná na internetu mohou být scrapována bez následků a porušování autorských práv, a že ne všechna data, která mohou být stažena eticky a legálně mohou být použita libovolným způsobem. Zejména při komerčním využití dat je třeba mít jistotu, že se jejich získáváním a publikováním autor nedopouští něčeho neetického.

3 Tvorba výukových simulací

3.1 Definice

Termín výuková simulace je poměrně často používán a jeho význam je různými autory poměrně přesně vysvětlován, přesto bych chtěl nejprve zmínit pár krátkých definic termínu simulace, který je obecnější a usnadní pochopit hlavní ideu, která se za výukovými simulacemi skrývá. Podle Encyclopedia of Computer Science (Smith 1998) je simulace „navrhnutí modelu reálného, nebo imaginárního systému a pokusy, které jsou s takovým modelem prováděny.[6]“

Podle Webster's College Dictionary (2010), který je využíván jako zdroj například webovými stránkami thefreedictionary.com, pak definice termínu simulace zní: „reprezentace chování či charakteristika jednoho systému pomocí užití systému jiného, zejména za použití počítače[7]“

6 Vlastní překlad z anglického originálu: „designing a model of a real or imagined system and conducting experiments with that model“

7 Vlastní překlad z anglického originálu: „the representation of the behavior or characteristics of one system through the use of another system, esp. using a computer.“

Takové definice jsou bohužel pro účely této práce příliš široké, jelikož mohou zahrnovat nejrůznější typy simulací - od vědeckých modelů až po komerční počítačové hry. Obecně je však možné říci, že simulace je modelem, který usnadňuje uchopení nějakého složitějšího systému - vývoj výukové simulace, je tedy logickým vyústěním tohoto principu, při kterém se její autor snaží využít tuto vlastnost, za účelem výuky.

Podle několika zdrojů jsou hlavními rozdíly mezi klasickou simulací a počítačovou hrou možnost počítačové hry přímo nekopírovat reálný systém (Oblinger 2006) a fakt, že simulace není hráčsky hodnocena – neexistuje stav „výhry“ (Klopfer 2009). V případě výukových simulací to ovšem nemusí být úplně pravda, protože výuková simulace často přebírá prvky komerčních her za účelem zvýšení vnitřní motivace studentů (Clayton 2010), mezi které typicky patří právě hodnocení v podobě jasného cíle, jehož splnění se rovná „výhře“. Navíc, vzhledem k tomu, že cílem výukové simulace není vytvořit systém, který bude dokonale funkční, ale naučit konkrétní znalosti a/nebo dovednosti, nemusí být pravda ani to, že výuková simulace dokonale kopíruje reálný systém, jak ukazuji i u jednotlivých případových studií, kterým se věnuji v další kapitole.

Na základě toho formuluji definici termínu výuková simulace pro účely této práce:

„Výuková simulace je počítačový program, který se snaží simulovat abstraktní model určitého systému, nebo jeho části, a předložit ho tak, aby uživateli usnadnil jeho pochopení. K tomuto cíli se výukové simulace snaží používat model stimulující u uživatelů vnitřní motivace, podobně jak tomu bývá u komerčních počítačových her.“

3.2 Proč používat simulace?

Toto téma je ve skutečnosti velmi široké a existuje na něj velké množství názorů. Já se v rámci práce nebudu tímto problémem zabývat příliš podrobně a zůstanu u

několika základních ideí, které stojí v pozadí vzniku výukových simulací.

Při hraní počítačových her člověk přijme velké množství informací a velkou část z nich se naučí aktivně používat. U komerčních her to bývají nejrůznější údaje o fiktivních světech, znalosti o nejrůznějších strategiích a postupech částí konkrétních systémů a samozřejmě znalosti o ovládacích prvcích. Mnoho mladých lidí tráví hraním her velké množství času a počítačová hra, jakožto nové médium, se pro ně stává intuitivním způsobem jak přijímat informace (Gee 2005).

Výukové simulace se snaží využít potenciál počítačových her k instrukci studentů. Navazují na dlouholetou snahu o vytvoření výukové hry, ale na rozdíl od projektů, které vznikaly v začátcích počítačové éry, se nesnaží využít médium pouze jako vnější motivaci, ale jako komplexní prostředí, které umožňuje simulaci abstraktního problému, který by za jiných okolností, tedy bez využití simulace, nebyl snadno uchopitelným a/nebo by neumožňoval experiment bez následků (Gredler 2004). Typickými příklady jsou simulátory, které při své výuce používají piloti, vojáci, lékaři a jaderní inženýři.

3.3 Tvorba výukových simulací

3.3.1 Výukový záměr

Při tvorbě výukových simulací je vhodné mít jasný a předem daný výukový záměr, který je esenciální pro případnou evaluaci a zejména pro zapojení simulace do praxe, ať již se jedná o běžnou školní výuku nebo vzdělávání v rámci podniku či instituce (Aldritch 2009).

3.3.2 Iterační proces

Tvorba výukové simulace by měla být do určité míry iteračním procesem (Salen 2003) - je tomu tak u mnoha komerčních počítačových her (beta verze, patche, enhanced verze, datadisky,...) U výukových simulací je fáze pozměňování ještě mnohem důležitější a proces navrhování složitější. Je třeba myslet nejen na to,

aby simulace byla zábavná, ale též na to, aby byla opravdu výuková, aby učila to, co si klade za cíl a aby její herní mechanismy nepřekážely výukovému procesu.

3.3.3 Dostupnost

Simulace musí být dostupná pro všechny studenty, kterým bude nabídnuta - pokud má být šířeji použita, nesmí příliš zvýhodňovat jednu skupinu studentů. To znamená zejména to, že výuková simulace musí být dostatečně nenáročná na pochopení, aby příliš nezvýhodnila studenty, kteří mají s prací na počítači, potažmo s hraním her, větší zkušenosti. Pokud má být ovládání složitější, je třeba ho dostatečně vysvětlit a neočekávat, že všichni studenti budou schopni do simulace intuitivně vstoupit. Je třeba si dávat pozor na to, že i nevhodně zvolený rámec, do kterého bude simulace zasazena, může mít výrazně negativní dopad na motivaci některých studentů (Malone 1972).

3.3.4 Herní systém

Ačkoli se simulace snaží přiblížit realitu, a co možná nejpřesvědčivěji simulovat konkrétní problematiku, je třeba udělat ústupky tak, aby byla „hratelná“. Základní ovládací prvky, které jsou nutné k obsluze simulace, by měly být dostatečně pochopitelné a intuitivní, aby se pozornost koncentrovala na výuku konkrétní problematiky a studenti nebyli přehlcováni informacemi, které nejsou podstatné pro pochopení problému. Zároveň by studenti měli být stále motivováni k dalšímu prozkoumávání simulace (Aldritch 2009).

3.3.5 Platforma

Pří výběru platformy, na které výuková simulace bude fungovat, je důležité myslet na kompatibilitu se systémy používanými ve školách. Simulace by proto měla být schopna fungovat v rámci systému MS Windows, neměla by být příliš náročná a její případná instalace by měla být co nejintuitivnější.

3.4 Případové studie

3.4.1 Evropa 2045[8]

Simulace Evropa 2045 vznikla v rámci projektu „Podpora rozšíření a využití informačních technologií ve výuce společenských věd“ financovaného Evropským sociálním fondem, státním rozpočtem České republiky a Hlavním městem Prahou. Simulaci vyvíjí nezisková organizace Generation Europe ve spolupráci s odborníky z Univerzity Karlovy (MFF UK, UISK FF UK), Asociace pro mezinárodní otázky, CIANT a Gymnázia Sázavská.

Simulace má dvě základní roviny, které na sobě nejsou vzájemně zcela závislé. První rovina je ekonomická - umožňuje studentům procházet rozpočet státu a sledovat, jak úpravy investic ovlivní ukazatele dostupné v simulaci. Studenti si tak například mohou vyzkoušet, že jakákoli změna politiky sociálních dávek bude mít velmi výrazný dopad na státní pokladnu a pravděpodobně i na spokojenost obyvatel. Druhá rovina se zabývá samotnou politikou uvnitř EU a agendami, které se studenti snaží prosadit pomocí hlasování. Simulace hodnotí studenty výlučně za to, jak se jim daří prosazovat agendy, které zapadají do jejich projektu, který je jakýmsi ideálním obrazem Evropy pro určitou politickou frakci.

V současné době simulace probíhá v rámci výuky mnoha českých středních škol a její obsah je upravován alespoň jednou ročně. Mezi každoroční úpravy patří aktualizace jednotlivých agend, případné přidávání nových agend společně s tím, jak vznikají pro EU důležitá témata politických diskuzí a úpravy ekonomických a statistických dat jednotlivých zemí EU tak, aby byla aktuální pro poslední proběhlý rok.

Použití web scrapingu by mohlo usnadnit každoroční úpravy, protože statistická a ekonomická data, která jsou potřebná pro simulaci, jsou dostupná v nejrůznějších ekonomických databázích a statistických ročenkách v prakticky neměnném formátu. Scraping by zde ale mohl vyřešit pouze problém s číselnými daty, protože údaje o nových agendách a tématech podstatných pro EU nejsou snadno kvantifikovatelné a bylo by náročné hledat vhodné a spolehlivé zdroje. Existuje i

8 Dostupné z: http://www.evropa2045.cz

teoretická možnost vytvořit program, který by na základě diskutovanosti nových témat vybíral vhodné kandidáty pro nové agendy, ale obávám se, že by jeho výroba byla poměrně náročná, a navíc by získané informace stěží dosahovaly kvality, kterou může nabídnout např. konzultace s odborníkem, který situaci v EU sleduje.

3.4.2 CellCraft[9]

Hlavním designérem simulace CellCraft je Anthony Pecorella a na programování se podílel Lars Doucet, který je v poslední době známý jako vývojář nezávislého herního studia Level Up Labs. Cílem simulace je informovat studenta o vnitřním fungování buňky a o jednotlivých organelách. Simulace probíhá v reálném čase a jednotlivé procesy, jako například tvorba enzymů, jsou velmi komplexní a blízké realitě, aniž by studenta příliš zatěžovaly složitým ovládáním. Je též možné si kdykoli prohlédnout encyklopedii, která obsahuje velké množství hesel a poskytuje poměrně komplexní popis všeho, co se v simulaci vyskytuje.

Ohledně simulace CellCraft byla v době jejího dokončení vedena kontroverzní debata, ve které někteří odpůrci tvrdili, že ji není vhodné používat ve školách, protože podvědomě učí teorii kreacionismu a přímo odporuje teorii evoluce. Z toho, co se dá v článcích odpůrců vyčíst, je hlavním problémem fakt, že jednotlivé organely se v buňce jednoduše objeví, místo toho, aby byl simulován proces evoluce. Bohužel, nikdo nenavrhuje žádné konkrétní změny a já osobně si nedokáži představit, jak by bylo možné takový proces simulovat tak, aby byl zachován smysl celkového produktu. Podle mého je v pořádku, že simulace se nesnaží učit věci, které nejsou jejím výukovým záměrem a pokud má být uvedena do škol, je úkolem vyučujícího, aby případné nejasnosti doplnil, či uvedl na pravou míru.

V dnešní době již CellCraft není ze strany svých tvůrců nijak podporován, nevznikají nové verze a ani domovská stránka projektu již neexistuje. Naštěstí je simulace volně šiřitelná a je k dispozici na mnoha stránkách, které se zabývají

9 Dostupné z: http://www.sciencegeek.net/Biology/CellCraft/CellCraft.html

výukou (jako sciencegeek.net) nebo které nabízejí volně šiřitelné hry vytvořené ve Flashi (jako kongregate.com).

CellCraft je výborným příkladem toho, že rozhodně ne všechny výukové simulace mohou těžit z metody information scrapingu. Neprobíhají žádné úpravy, takže by bylo zbytečné používat scraper pro účely usnadnění aktualizací. Navíc i při samotné tvorbě by sice bylo možné použít scraper např. pro získání názvů jednotlivých organel, ale poté by stejně bylo nutné všechna data překontrolovat a vytvořit vztahy a systémové vlastnosti jednotlivých prvků. Navíc i encyklopedické záznamy by bylo třeba tvořit nezávisle, protože vyžadují spolupráci odborníků z oboru biologie. Použití scraperů by se tak stalo spíše kontraproduktivním.

3.4.3 Super Energy Apocalypse[10]

Super Energy Apocalypse je další projekt, na kterém se podílel Lars Douchet. Nazývat jej však výukovou simuací by bylo velmi nadsazené - jedná se flashovou hru z postapokalyptického prostředí, kde je cílem hráče vytvářet nejrůznější stacionární zbraně, aby se ubránil hordám agresivních mimozemšťanů. Co je na hře velmi zajímavé, a proč má vůbec smysl se o ní zmiňovat v rámci této práce, je ekonomický model, který je použit v herním systému. Kromě základních surovin jako je uhlí, ropa a železo, které se ve hře nacházejí, je pro provoz obraných budov potřebná také elektrická energie, která je produkována v nejrůznějších typech elektráren. Jednotlivé výkony, spotřeby a případné rizika elektráren jsou založena na reálných datech, byť pro účely hratelnosti byly upraveny např. poměry velikostí jednotlivých budov (zejména sluneční a větrné elektrárny ve hře použité by vyžadovaly výrazně větší plochu, než jakou zabírají oproti např. jaderné elektrárně.)

V současné době není hra, podobně jako CellCraft, přímo podporována tvůrci - nevznikají tedy nové verze a hra nemá žádnou domovskou stránku.

10 Dostupné z: http://www.kongregate.com/games/larsiusprime/super-energy-apocalypse-recycled

Web scraping by zde bylo možné použít k průběžné aktualizaci dat, čímž by se hra stala více simulací. Takový zásah by mohl být použit k zajímavým možnostem a úpravám, kdy by se například výkony elektráren mohly měnit podle aktuálních možností na reálném trhu. Bohužel by hrozilo, že se hra, bez dohledu tvůrců, stane nevyváženu a některé elektrárny se stanou ve hře zcela nepoužitelnými, protože ekonomický model simulace samozřejmě nepočítá se všemi možnými reálnými faktory.

3.5 Použití web scrapingu při tvorbě výukových simulací

3.5.1 Typy výukových simulací

Výukové simulace můžeme rozdělovat na mnoho různých typů na základě nejrůznějších hledisek, od pojetí času (dělení na simulace v reálném čase a simulace tahové), přes komplexitu vstupů a výstupů, až po způsoby, jakými jsou uchyceny v pozadí a příběhu. Pro účely této práce bohužel tato běžně používaná rozdělení nestačí. Z hlediska možností použití web scrapingu je zajímavá zejména datová struktura simulace a její potřeba aktualizací.

Data, na kterých je simulovaný systém postaven můžeme dělit na taková, která jsou výrazně propojená a vázaná velkým množstvím pravidel - vhodný příklad jsou organely v simulaci CellCraft (tedy cca 20 jednotek, které mají své specifikace a jsou na sobě nejrůznějším způsobem závislé) - a na data, kterých je sice relativně velké množství, ale mají minimální vzájemné vztahy - v extrémním případě se může jednat o prostou tabulku. Příkladem takových neprovázaných dat jsou statistické a ekonomické informace v simulaci Evropa 2045, kde se nachází 24 států, z nichž každý má zhruba 10 údajů o velikosti, rozpočtu, počtu obyvatel,... Podle struktury dat tedy můžeme výukové simulace rozdělit na provázané (závisí více na vztazích jednotek) a obsáhlé (závisí spíše na větším počtu jednotek) - samozřejmě žádná simulace nebude spadat výlučně do jedné kategorie, přesto může být podobné kategorizování velmi dobrým indikátorem, jestli má smysl snažit se web scraping do tvorby simulace zapojit.

Dalším důležitým hlediskem je potřeba data aktualizovat. Výukové simulace můžeme rozdělit na stálé - takové, které simulují systém, který je založen na konkrétní teorii či souboru dat, které pravděpodobně zůstanou delší dobu neměnné. Ideálním příkladem je opět simulace CellCraft, jejíž systém je založen na základě fungování buňky, které je poměrně detailně prozkoumané a nehrozí, že by byl simulovaný systém výrazně ovlivněn novými poznatky. Druhým typem jsou simulace obnovované, jejichž data je třeba průběžně upravovat nebo doplňovat, aby zůstaly aktuálními (statistická data jednotlivých zemí v Evropě 2045).

Pro použití s metodou information scrapingu jsou nejvhodnější simulace obsáhlé a obnovované. V případě obnovovaných dokonce může být scraping vhodný i pro samotné aktualizace - aniž by byl použit při tvorbě simulace.

3.5.2 Vlastnosti získávaných dat

Při zavádění web scrapingu do tvorby výukových simulací je také velmi důležité si uvědomit, že proto, aby mohl být scraping použit, je potřeba mít dostupná vhodná data. Naštěstí, formát, ve kterém jsou data dostupná, není natolik důležitý, protože scraper dokáže pracovat téměř s čímkoli, co dokáže v nějaké podobě číst engine, který data zobrazuje koncovému uživateli. Důležitými se tak stávají jiné vlastnosti získávaných informací.

Zřejmě nejdůležitější je samotná dostupnost. Je třeba si položit otázku, jestli data, která se pro simulaci hodí, jsou určená pro použití s automatizovanými metodami získávání informací, popř. jestli je možné pro konkrétní scraper získat výjimku (samozřejmě, pokud potřebná data nejsou dostupná vůbec, bude nemožné vytvořit simulaci, nehledě na to, zda-li bude použita metoda web scrapingu.)

Další důležitou vlastností je spolehlivost. Pokud mají být data použita ve výuce, je nemyslitelné, aby nebyla pravdivá, a naopak je velmi vhodné, aby je bylo možné zaštítit jménem nějaké konkrétní organizace, která za jejich správnost alespoň

teoreticky zodpovídá.

Podstatná je též úplnost získávaných dat. Tato vlastnost je nutná pro automatické fungování scraperu, protože scraper většinou nedokáže získávat informace z alternativních zdrojů. Samozřejmě, pokud je dostupných více neúplných zdrojů, mohou dát dohromady jeden úplný, pak ovšem bude těžší scraper vytvořit tak, aby byl zcela spolehlivý. I v případě, že úplná data nejsou dostupná, může být použití scraperu vhodné a usnadnit manuální dohledávání.

Nakonec bych zmínil stálost formátu dat. Přestože samotný formát není příliš podstatný, ačkoli samozřejmě jsou formáty, se kterými se pracuje lépe, pokud bude potřeba simulaci aktualizovat, je velmi vhodné vybrat si zdroj, u kterého je menší riziko, že bude změněn jeho formát, nebo zdrojový kód. U simulací, které by měly být často aktualizovány, může být opakovaná změna formátu dat dokonce důvodem, proč radši simulaci vůbec nevytvářet.

Pro tyto účely jsou velmi vhodná data z veřejné správy, která jsou dobře dostupná, poměrně spolehlivá a formátem neměnná.

3.5.3 Návrh výukové simulace s použitím web scrapingu

Poslední věc, které bych se rád v rámci této práce věnoval, je příklad, jak by mohla vypadat výuková simulace, která by využívala metody web scrapingu pro získání dat, na základě kterých by byla utvořena a dat potřebných pro její aktualizace. Jak již bylo řečeno v zadání práce, návrh simulace bude z oblasti energetiky.

Jejím výukovým záměrem je seznámit studenty s aktuálním stavem energetiky z hlediska ekonomického a ukázat rozdíly mezi tradičním pojetím energetiky, kdy ekonomický model stojí na velkých, výkonných a centralizovaných elektrárnách a pojetím alternativních energetických zdrojů, které jsou zpravidla decentralizované a stojí na ochotě soukromých subjektů a jaké podmínky jsou potřeba pro částečné

smazání těchto rozdílů, jak se tomu stalo např. v některých severských zemí, nebo do jisté míry v Německu.

Sekundárním výukovým záměrem je naučit studenty pracovat s ekonomickými rozhodnutími a volit vhodné krátkodobé cíle tak, aby co nejlépe naplnili ty dlouhodobé.

Simulovaný systém

Základem systému bude samotný stav cen energetiky. Podstatné budou momentální reálné ceny, které budou aktualizovány každý měsíc podle současného stavu v ČR. Další důležitou součástí systému bude analýza složení ceny elektřiny pro jednotlivé typy elektráren, podstatné bude zejména to, kolik procent finální ceny je třeba použít na palivo, kolik na samotnou stavbu elektrárny (rozpočítáno na její očekávanou životnost), kolik na případné další výdaje (zbavování se odpadů, fond oprav,...) a kolik bude činit finální čistý zisk firmy. V některých případech bude naopak třeba započítat případné dotace, závislé na vyhláškách EU a ČR. Poslední součástí systému budou data, která budou celou simulaci dokreslovat: ceny jednotlivých paliv, ceny staveb elektráren, ceny a vlastnosti pozemků. Systém se bude snažit kopírovat události v reálném světě a bude umožňovat odehrát historii energetiky od roku 1990 (pokud by měla jít dále do historie, bylo by třeba najít další zdroje dat, protože starší data o cenách energií u nás nejsou tak snadno dostupná) do současnosti (respektive do chvíle poslední aktualizace simulace.)

Hratelnost

Student bude postaven do role člověka, který plánuje budování elektráren na určitém území. Jeho cílem bude vyprodukovat určité množství energie, tedy zásobovat určité množství podniků/domácností, s tím, že ani příliš velká, ani příliš malá produkce nebude výhodná (je třeba vyrovnávat nabídku vůči poptávce.) Na základě informací, které bude mít o době stavby elektráren, jejich životnosti, produkci energie, očekávaných nákladech na palivo, atp. bude jeho cílem naplňovat poptávku a při té příležitosti vykázat co nejvyšší zisk.

Ve svém nejsnažším pojetí bude simulace fungovat jako nástroj, který bude

možné využít pro podporu výuky, s tím, že bude vyžadovat spolupráci instruktora. Jediné informace, které bude simulace poskytovat studentovi navíc, budou měsíční zprávy o změnách cen, průběžné zprávy o zavádění nových technologií a nových vlastnostech elektráren a příležitostné novinky, které budou mít vztah k důležitým událostem na poli energetiky (jako např. změna podmínek pro získávání dotací na solární energii.) Možné využití simulace by pak bylo, kdyby instruktor po vysvětlení nějaké části látky vyzval studenty, aby si doma spustili simulaci, odehráli dobu mezi např. lety 2003-2008, a na další hodinu si připravili krátký referát, co se stalo zajímavého, jaké změny se odehrály a proč se tomu tak stalo.

Myslím si, že aby byla simulace užitečnější v praxi, bylo by vhodné, aby byla lépe využitelná ve třídě, a tedy umožňovala alespoň do jisté míry účast více studentů, nebo aby sama plnila funkci instruktora a byla tudíž praktičtější pro samostudium. Bohužel, účast více studentů, zejména blížil by se jejich počet třiceti, jak tomu často ve třídách bývá, vidím jako těžko zaveditelnou do podobného systému, už jen proto, že by bylo těžké pro ekonomicky méně zběhlé studenty konkurovat těm schopnějším, což by mohlo vést ke ztrátě motivace. Naopak si myslím, že přidání instruktážních prvků by mohlo této simulaci velmi prospět, a to nejen z hlediska množství předaných znalostí, ale i z hlediska lepší hratelnosti.

Instruktážní prvky by bylo možné přidat v podobě jednotlivých úkolů, které by se odehrávaly v systémem určeném časovém rozpětí, na daném místě a s danými cenami paliv/pozemků. Tak by mohly být předvídány podmínky a celý úkol by mohl ukazovat výhody a nevýhody jednotlivých typů elektráren a zároveň postavit studenty před herně zajímavé úlohy, ve kterých by museli aktivně využívat své vědomosti.

Ovládání a uživatelské rozhraní

Simulace bude ovládána pomocí myši s klávesovými zkratkami pro jednotlivé typy elektráren. Na hlavní obrazovce bude seznam všech elektráren, které současný uživatel vlastní, spolu s jejich základními údaji (kolik energie vyrábí, jaký činí výdělek, údaje o pozemku, k jakému roku byla elektrárna renovována, celkové provozní náklady ode dne stavby a celkový výdělek od počátku provozu a

případné další údaje jako produkovaný odpad či znečištění) a základními ovládacími prvky (možnost elektrárnu nahradit jinou, renovovat (čímž se uvedou v platnost případné technologie, které výrobu upravují,) dočasně uzavřít (například pokud je zrovna drahé palivo) či prodat.)

Kromě toho se ve spodní části obrazovky bude nacházet menu se základním nastavením (to se bude patrně lišit podle softwaru použitého pro tvorbu simulace; za použití nástroje Adobe Flash by se mohlo jednat o nastavení kvality zobrazení a ovládání zvuků), s přístupem do encyklopedie a s možností návratu do hlavního menu. Encyklopedie bude obsahovat informace o každém typu elektráren, případné důležité technologie související s renovací (tedy ty, které budou dostatečně podstatné, jako např. nová generace jaderné elektrárny, která by byla schopna zpracovávat palivo efektivněji, a tudíž by dokázala částečně zpracovat i některé sloučeniny, které do té doby byly součástí jaderného odpadu), údaje o současných cenách (podle roku, ve kterém se simulace nachází) a základní informace o ovládání a systému simulace.

Posledním ovládacím prvkem bude ovládání času simulace v pravém horním rohu, které bude umožňovat čas zrychlit, zpomalit a dočasně zastavit.

Použitá data

Pro většinu dat z oblasti energetiky by byly vhodným zdrojem stránky Energetického regulačního úřadu, které obsahují data o cenách, spotřebě a dalších důležitých veličinách, které s elektrickou energií v ČR souvisí a to ve formátu HTML, který je již delší dobu neměnný a ani v budoucnosti zcela pravděpodobně žádné zásadní změny nechystají. Data jsou tak zaštítěná státní organizací, která do jisté míry odpovídá za jejich pravost, jsou neměnná, takže je možné používat jeden scraper delší dobu a nevšiml jsem si jakékoli neúplnosti.

International Energy Agency vydává publikaci Projected costs of generating electricity. Jejím cílem je ukázat efektivitu nejrůznějších metod vyrábění elektrického proudu a mimo jiné obsahuje informace o nových technologiích, u kterých se očekává, že budou v následujících pěti letech implementovány do reálné produkce a jaký je jejich očekávaný dopad. Nejnovější, tedy sedmý, díl této

publikace vyšel roku 2010 a je možné ho použít jako relevantní a aktuální zdroj. V jednotlivých publikacích jsou drobné formální změny, kvůli kterým by bylo potřeba buď vyvinout scraper, který by byl do jisté míry schopen prohledávat obsah, nebo pro každou publikaci udělat separátní scraper v rámci aktualizace. Vzhledem k tomu, že taková aktualizace by byla třeba cca jednou za pět let, nevidím v manuální variantě problém, naopak manuální aktualizace by v tomto případě umožňovala i tvorbu nových záznamů do encyklopedie a zavedení drobných úprav do simulovaného systému, pokud by to bylo potřeba.

Data o cenách pozemků by bylo možno získávat ze stránek jednotlivých městských úřadů. Větší problém by byl s jejich vlastnostmi - základní vlastnosti pozemků, které jsou pro energetiku důležité, tedy např. slunečnost, průměrná síla větru, etc. je sice možné získat ze stránek českého meteorologického ústavu, ale pravděpodobně bude nutné jejich ceny a vlastnosti nějakým způsobem upravit tak, aby studentům sice ukazovaly, že není možné postavit elektrárnu kdekoli, ale zároveň aby se nestaly zbytečnou překážkou pro hratelnost.

Překážky

Před další fází iteračního procesu tvorby této simulace, která by obsahovala detailnější design dokument a následně první verzi simulace, by bylo třeba vyřešit několik dílčích problémů.

Bylo by vhodné udělat průzkum trhu a zájmu ze strany škol, aby se ukázalo, jestli zadaný výukový záměr má smysl a případně jej upravit.

V současném návrhu nikde není popsán přesný způsob, jakým bude simulovaný systém fungovat a na bázi jakého ekonomického modelu. Před samotným začátkem tvorby první verze simulace je samozřejmě taková informace potřebná, protože model bude hlavním prvkem, kolem kterého bude utvářená hratelnost i případné specifické prvky ovládání a instruktáže. Před tím, než by se takový model dal přesně definovat, by byla zapotřebí konzultace s odborníky z oborů energetiky a ekonomie, aby bylo možné vytvořit návrh tak, že bude co možná

nejpřesnější, ale zároveň pochopitelný a hratelný.

Na základě ujasnění prvků ovládání, uživatelského rozhraní a modelovaného systému (popřípadě též softwaru použitého pro tvorbu simulace), by také měl být vytvořen první grafický návrh, který by ukazoval alespoň rámcové zobrazení, jak by vypadalo pro koncového uživatele.

4 Závěr

Cílem práce bylo analyzovat možnosti využití scrapingu při tvorbě a aktualizaci výukových simulací a navrhnout příklad konkrétní simulace, která by s touto metodou pracovala.

Na začátku práce byla představena metoda information scrapingu a na základě definic z různých zdrojů, byl jako termín využívaný pro účely této práce určen web scraping, s tím, že je užší a lépe vystihuje konkrétní potřeby této práce.

Velkou část práce jsem se věnoval praktickým problémům, se kterými se uživatel musí potýkat, pokud web scraping využívá. Popsal jsem současnou právní a etickou situaci a poukázal na problémy s tím související.

V druhé části jsem se věnoval výukovým simulacím a pokusil jsem se definovat jaká simulace musí být, aby mola těžit z výhod information scrapingu. Z bakalářské práce vyplývá, že web scraping rozhodně není použitelný pro tvorbu jakékoli simulace, ale že může být velmi nápomocný u simulací, které pracují s velkým množstvím nepříliš provázaných informací, a to jak při jejich tvorbě, tak později při případných aktualizacích.

V závěru své práce jsem navrhl příklad simulace, která by využila potenciál scrapingu a to zejména díky tomu, že pracuje s daty, která se aktualizují každý měsíc a manuální aktualizace by tak byly prakticky nemyslitelné.

Myslím si, že scraping rozhodně nabízí další možnosti při tvorbě výukových simulací a pokud se použije vhodně může ušetřit mnoho manuální práce.

5 Seznam použité literatury

ALBA, Alfredo - BHAGWAN, Varun - GRANDISON, Tyrone. Accessing the Deep Web: When Good Ideas Go Bad. *OOPSLA Companion* [online]. 2008. s. 815-818.

ALDRITCH, Clark. A field guide to educational simulations. New York: ASTD, 2009.

ANGWIN, Julia - STECKFLOW, Steve. 'Scrapers' Dig Deep for Data on Web. *The Wall Street Journal*. 2010-10-11. Dostupné z: http://online.wsj.com/article/SB10001424052748703358504575544381288117888.html

AUERNHEIMER, Andrew. Forget Disclosure — Hackers Should Keep Security Holes to Themselves. *Wired.com* [online]. 2012-11-29 [cit. 2013-03-11]. Dostupné z: http://www.wired.com/opinion/2012/11/hacking-choice-and-disclosure

ČÍŽEK, Jakub. Google má zpravodajský agregátor k prezidentským volbám. *zive.cz* [online]. 2013 [cit. 2013-03-19]. Dostupné z: http://www.zive.cz/bleskovky/google-ma-zpravodajsky-agregator-k-prezidentskym-volbam/sc-4-a-167085/default.aspx

BLAZE, Matt. AT&T iPad Hacker's Real Crime Was Embarrassing the Wrong People. Wired.com [online]. 2012-11-27 [cit. 2013-03-11]. Dostupné z: http://www.wired.com/opinion/2012/11/att-ipad-hacker-when-embarassment-becomes-a-crime/

BOOKER, Ellis. IBM puts a Web face on legacy systems. *Computing Canada.* 1997-06-23, 23(13), s. 26.

BRADSHAW, Paul. Two reasons why every journalist should know about scraping. journalism.co.uk [online]. 2012 [cit. 2013-03-19]. Dostupné z: http://www.journalism.co.uk/news-commentary/-two-reasons-why-every-journalist-should-know-about-scraping-/s6/a550001

CLARKIN, Larry - HOLMES, Josh. Enterprise Mashups. *The Architecture Journal.*

Microsoft. 2007 [cit. 2013-03-19]. Dostupné z: http://msdn.microsoft.com/en-us/architecture/bb906060.aspx

ČÍŽEK, Jakub. Google má zpravodajský agregátor k prezidentským volbám. *zive.cz* [online]. 2013 [cit. 2013-03-19]. Dostupné z: http://www.zive.cz/bleskovky/google-ma-zpravodajsky-agregator-k-prezidentskym-volbam/sc-4-a-167085/default.aspx

CRUPY, John - WARNER, Chris. Enterprise Mashups : The New Face of Your SOA : Bringing value to the enterprise. *SOA World*. 2009. [cit. 2013-03-19]. http://soa.sys-con.com/node/719917

DAVIES, Tim. Foraging for Data with Scraper Wiki. In: *Open Data Cook Book* [online]. 2012-10-24 [cit. 2013-10-03]. Dostupné z: http://opendatacookbook.net/wiki/recipe/foraging_for_data_with_scraper_wiki

GEE, James P. *Why Are Video Games Good For Learning?* Wisconsin: AADLC, 2006.

Google. Adding a site to Google. 2013 [cit. 2013-03-11]. Dostupné z: http://support.google.com/webmasters/bin/answer.py?hl=en&answer=34397

HACKETT, John. Domesticating Account Aggregators. *Bank Technology News* 13. 2000-08-10. s. 40-46.

HARTLEY, Brody. I Don't Need No Stinking API: Web Scraping For Fun and Profit. Personal blog, 2012-08-12 [cit. 2013-03-11]. Dostupné z: http://blog.hartleybrody.com/web-scraping

HAWKER, Mark. Oh my Ateism. *Scraperwiki.com*. 2013 [cit. 2013-03-19]. Dostupné z: https://scraperwiki.com/views/oh_my_atheism_1

KAEFER, Heather. What Is a Web Crawler? *Wisegeek* [online]. 2013-02-18 [cit. 2013-03-19]. Dostupné z: http://www.wisegeek.org/what-is-a-web-crawler.htm

KRAVETS, David. Feds Charge Activist with 13 Felonies for Rogue Downloading of Academic Articles. *Wired.com* [online]. 2012-09-18 [cit. 2013-03-11]. Dostupné z: http://www.wired.com/threatlevel/2012/09/aaron-swartz-felony

LARSON, William. *An Introduction to Compassionate Screen Scraping*. 2008-10-8 [cit. 2013-03-10]. Dostupné z: http://lethain.com/an-introduction-to-compassionate-screenscraping

MADNICK, Stuart - SIEGEL, Michael. Seizing the opportunity: Exploiting web aggregation. *MIS Quarterly.* 2002-03-10, 1 (1), s. 35-46.

MALONE, Thomas W. What makes things fun to learn? : heuristics for designing instructional computer games. In *Proceedings of the 3rd ACM SIGSMALL symposium*. California : [s.n.], 1980. s. 162 - 169.

MERRILL, Jeremy B. - ORNSTEIN, Charles - WEBER, Tracy - WEI, Sisi - NGUYEN, Dan. Dollars for Docs : How Industry Dollars Reach Your Doctors. *ProPublica.org* [online]. 2013-04-11 [cit. 2013-03-19]. Dostupné z: http://projects.propublica.org/docdollars

Mug Shots. *Tampa Bay Times*. 2013 [cit. 2013-03-19]. Dostupné z: http://www.tampabay.com/mugshots/

NAJORK, Marc. Web Crawler Architecture. [cit. 2013-03-19]. Dostupné z: http://research.microsoft.com/pubs/102936/eds-webcrawlerarchitecture.pdf

NEWTH, Alex. What Is Automatic Indexing? *Wisegeek* [online]. c2013 [cit. 2013-03-19]. Dostupné z: http://www.wisegeek.com/what-is-automatic-indexing.htm

NGUYEN, Dan. Scraping for Journalism : A guide for collecting data. Chapter 4: Scraping data from html. *Propublica.org* [online]. 2010-12-30 [cit. 2013-03-11]. Dostupné z: http://www.propublica.org/nerds/item/scraping-websites

NGUYEN, Dan. *The Bastards Book of Ruby* [online]. 2011-03-12 [cit. 2013-03-11]. Dostupné z: http://ruby.bastardsbook.com/chapters/web-scraping

NGUYEN, Dan. A defense of web-scraping as a vital tool for journalists, danwin.com [online]. 2012 [cit. 2013-03-19]. Dostupné z: http://danwin.com/2012/06/a-defense-of-web-scraping-as-a-vital-tool-for-journalists

OBLINGER, Diana. Simulations, Games and Learning. 2006-05 [cit. 2013-03-11]. Dostupné z: http://mobilelearningcourse.pbworks.com/f/Games%2Band%2BLearning%2BELI3004.pdf

ORNSTEIN, Charles - WEBER, Tracy - Nguyen, Dan. About the Dollars for Docs Data. ProPublica.org [online]. 2013-04-11 [cit. 2013-03-19]. Dostupné z: http://www.propublica.org/article/about-our-pharma-data

PEENIKAL, Sunilkumar. Mashups and Enterprise. MphasiS, white paper, 2009 [cit. 2013-03-19]. Dostupné z: http://www.mphasis.com/pdfs/Mashups_and_the_Enterprise.pdf

POWER, Carol. While Others Quail at 'Screen Scraping,' FleetBoston Will Embrace It on New Site. *American Banker*. 2000-02-01. s. 1.

Random House Kernerman Webster's College Dictionary. Tel Aviv: K Dictionaries Ltd., 2010. Dostupné z: http://www.thefreedictionary.com/simulation

RETKWA, Rosalyn. Data sweepers. *Registered Rep.* 2000-09-24, 24 (9), s. 69-76.

Screen scrapers: Gotta love 'em...or hate 'em. *Financial Service Online.* 2000-03-12, 5 (4), s. 6.

Scraper site v News Agregator. *Webproworld.com* [online]. Forum discussion. 2010-01-24 2010-01-26 [cit. 2013-03-19]. Dostupné z: http://www.webproworld.com/webmaster-forum/threads/95399-Scraper-Site-v-News-Agregator

SISARIO, Ben. Probation, Not Prison, for Scalpers. *New York Times*. 2011-06-09 [cit. 2013-03-11]. Dostupné z: http://www.nytimes.com/2011/06/10/business/media/10wiseguys.html?_r=0

SMITH, Roger D. Simulation article. 1998. In: *Encyclopedia of Computer Science*. New York: Grove's Dictionaries, 2000.

SOKOL, Tomáš - SMEJKAL, Vladimír. Postih počítačové kriminality podle nového trestního zákona. Právní rádce. *ihned.cz* [online]. 2009 [cit. 2013-03-19]. Dostupné z: http://pravniradce.ihned.cz/c1-37865090-postih-pocitacove-kriminality-podle-noveho-trestniho-zakona

SUMMET, Jay. Reading the web. In: Data manipulation for science and industry. Verze 1.0.1. 2010 [cit. 2013-03-19]. Dostupné z: http://www.summet.com/dmsi/html/readingTheWeb.html

ŠVERÁK, Petr. Google uvolnil open source aplikaci Refine 2.0 pro třídění dat. *Computer world*. 2010 [cit. 2013-03-19]. Dostupné z: http://computerworld.cz/software/google-uvolnil-open-source-aplikaci-refine-2-0-pro-trideni-dat-8073

The American Heritage Dictionary of the English Language. Fourth Edition. Boston: Houghton Mifflin Company, 2000.

TOONKEL, Jessica. 2000a. First Union Yields to Screen Scraping; Customer Demand Seen Key Motive. *American Banker.* 2000-04-13. s. 1.

TOONKEL, Jessica. 2000b. Banks Stop Whining, Learn to Love Aggregation. *American Banker.* 2000-09-08. s. 3A.

Use of Web "scraper" did not violate CFAA in absence of notice of site restrictions. *Computer and Internet Lawyer.* 2003-04-20. 20 (4), s. 39.

Web aggregation--an overview. *The Disclosure.* 2001-03-18, 18 (3), s. 3.

WEISUL, Kimberly. Screen Scraping Makes Web Comeback. *Interactive week*. 2000-04-17, s. 24.

ZABRISKIE, John F. Bots, Scrapers, and Other Unwanted Visitors to Your Web Site: Can You Keep Them Out? *The Computer & Internet Lawyer.* 2009-07-07, 26 (7), s. 5-11.

ZETTER, Kim. 2010a. Wiseguys Indicted in $25 Million Online Ticket Ring. *Wired.com* [online]. 2010-03-01 [cit. 2013-03-11]. Dostupné z: http://www.wired.com/threatlevel/2010/03/wiseguys-indicted

ZETTER, Kim. 2010b. Is Breaking CAPTCHA a Crime? Wired.com [online]. 2010-07-07 [cit. 2013-03-11]. Dostupné z: http://www.wired.com/threatlevel/2010/07/ticketmaster/

ZETTER, Kim. AT&T Hacker 'Weev' Sentenced to 3.5 Years in Prison. *Wired.com* [online]. 2013-03-18 [cit. 2013-03-11]. Dostupné z: http://www.wired.com/threatlevel/2013/03/att-hacker-gets-3-years/