Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
Ilkovičova 2 , 842 16 Bratislava

Extrakcia dát z webu

[WebExtraction]

Dokumentácia k inžinierskemu dielu

Tím:	č. 16, WebX
Vedúci tímu:	Ivan Srba
Členovia tímu:	Ján Brechtl, Tomáš Juhaniak, Martin Kalužník, Rastislav Krchňavý, Michal Kren, Martin Lacek, Andrej Vaculčiak
Akademický rok:	2016/2017
Autor:	Ján Brechtl, Tomáš Juhaniak, Martin Kalužník, Rastislav Krchňavý, Michal Kren, Martin Lacek, Andrej Vaculčiak
Verzia číslo:	1.5
Dátum poslednej zmeny:	15.05.2017

1 Úvod

V dnešnom modernom svete plnom digitálnych technológií nepredstavuje až taký problém prístupnosť informácií, ako ich nekonzistentné uchovávanie, resp. ich reprezentácia. Poznáme rôzne druhy uchovávania informácií, či už pomocou textových alebo iných typov súborov. Keď z nich však chceme jednotlivé informácie získavať, nehovoriac o ich uchovávaní na jednom mieste, nastáva dosť ošemetná situácia, ktorá sa vyznačuje časovo náročným zberom dát z rôznych zdrojov a ich následná úprava do jednotného formátu, z ktorým sa následnej pracuje ďalej.

To isté platí aj pre webové služby. Aj keď aj v tejto oblasti existujú určité štandardy, stále to nie je len jeden, a teda je potrebné vedieť pracovať s viacerými alternatívami reprezentácie.

Tento problém otvára priestor pre realizáciu systému, ktorý bude vykonávať zber dát z rôznych webových serverov uchovávajúcich informácie v rôznej podobe. Jednou zo základných a veľmi užitočných vlastností by mala byť možnosť opakovaného zberu dát, nakoľko je dobré udržiavať dáta neustále aktuálne a kontrolovať ich manuálne je dosť nepraktické a pri obrovskom množstve aj nemožné.

Obsahom tohto dokumentu je celkový pohľad na systém, spolu s objasnením dekompozície na moduly. Ku každému modulu sú samostatne (v samostatných dokumentoch) uvedené informácie o analýze, návrhu a následnej implementácií, pričom sa nesmie vynechať testovanie, ktoré je ťažiskom pri tvorbe jednotlivých modulov.

2 Globálne ciele pre ZS

Keďže ide o tímový projekt, na ktorom spolupracujú ľudia, ktorí predtým spolu nepracovali, celý zimný semester, no najmä prvá polovica sa vyznačuje inicializáciou činností, dohadovaním a konfiguráciou mnohých komponentov a podporných nástrojov, no v neposlednom rade rozdelením si zodpovedností a začatím implementácie prvých verzií vybraných modulov systému.

Cieľom projektu je vytvoriť funkčný systém na pravidelnú extrakciu vybraných dát z vybraných webových stránok.

Používateľ sa má možnosť zaregistrovať a následne prihlásiť na server, ktorý mu poskytuje možnosť vytvorenia projektov, pričom v jednom projekte môže mať uložených viacero skriptov, každý pre rôznu stránku, ktoré definujú, aké dáta sa budú extrahovať.

Okrem definovania extrahovaných dát systém umožní nastaviť dátum spustenia automatickej extrakcie a pravidelnosť jej spúšťania v rôznych intervaloch (minúty, hodiny, atď.).

Samozrejmosťou je prehľadný výpis používateľových projektov, skriptov, aj samotných extrakcií spolu s časom ich spustenia a vykonania.

Aby však používateľ mohol tieto extrakcie definovať, potrebuje na vybranej stránke nejakým spôsobom vybrať, ktoré dáta chce extrahovať. Túto funkcionalitu mu poskytne rozšírenie do prehliadača, pomocou ktorého sa prihlási na svoj účet a môže si dané skripty podrobnejšie definovať.

3 Globálne ciele pre LS

Pre letný semester sa tak isto stanovili určité ciele. Základným cieľom je dokončiť implementáciu základného prototypu, ktorý odovzdáme potencionálnym používateľom na testovanie.

Čo sa týka funkcionality, je potrebné (oproti verzií zo zimného semestra) dopracovať parsovanie údajov, výber elementov do prijateľnej podoby a rozšíriť možnosti spracovania elementov po ich získaní zo stránky (formátovanie a pod.).

Koncom semestra (cca 10. týždeň), prebehne publikácia produktu, pričom je podmienkou, že produkt má mať charakter MVP (Minimum Viable Product). Podrobnosti sú uvedené v príslušnej podkapitole.

Počnúc publikáciou sa produkt podrobí testovaniu, a to aj z pohľadu funkcionality, aj z pohľadu použiteľnosti rozhrania.

3.1 MVP

Po konzultovaní s PO (Product Owner), sa stanovili požiadavky na vyvíjaný systém, ktoré musia byť splnené, aby mohlo prebehnúť nasadenie do prevádzky a testovanie použiteľnosti. Aj keď celý vývoj prebieha v spolupráci s PO, ktorý definuje, akú funkcionalitu

postupne požaduje, je potrebné mať definované, aké vlastnosti a funkcie má mať produkt, aby bolo možné ho nasadiť do testovacej, ale v podstate reálnej prevádzky.

MVP pre náš systém bolo definované nasledovnou funkcionalitou:

- User management
  - Registrácia a prihlasovanie používateľov
  - Administrácia účtu
- Project and script management
  - CRUD (Create, Read, Update, Delete) pre projekty a skripty
  - Nastavenie pravidelnosti spúšťania skriptu
  - Nastavenie dátumu a času prvého spustenia
- Extension
  - Výber jedného xPathu (graficky)
  - Výber viacerých xPathov (graficky)
- API
  - API pre komunikáciu s Extension
  - API pre export extrahovaných dát
- Crawler
  - Základná funkcionalita (extrakcia dát zo špecifikovanej URL a xPathu)
  - Postprocessing
    - Vnorený xPath
    - Restrict
    - Možnosť výberu extrahovanej hodnoty (napr. text vs. atribút elementu)
    - Kontrola dátových typov
    - Logovanie extrakcie
    - Spúšťanie extrakcie pomocou plánovača (scheduler)
    - Spustenie extrakcie manuálne
- Dáta
  - Ukladanie do DB (Extrahované dáta, logy)
  - Zobrazovanie výsledkov extrakcií
  - Zobrazovanie logov

Produkt je dostupný na webe:
http://team16-16.studenti.fiit.stuba.sk/webx/login

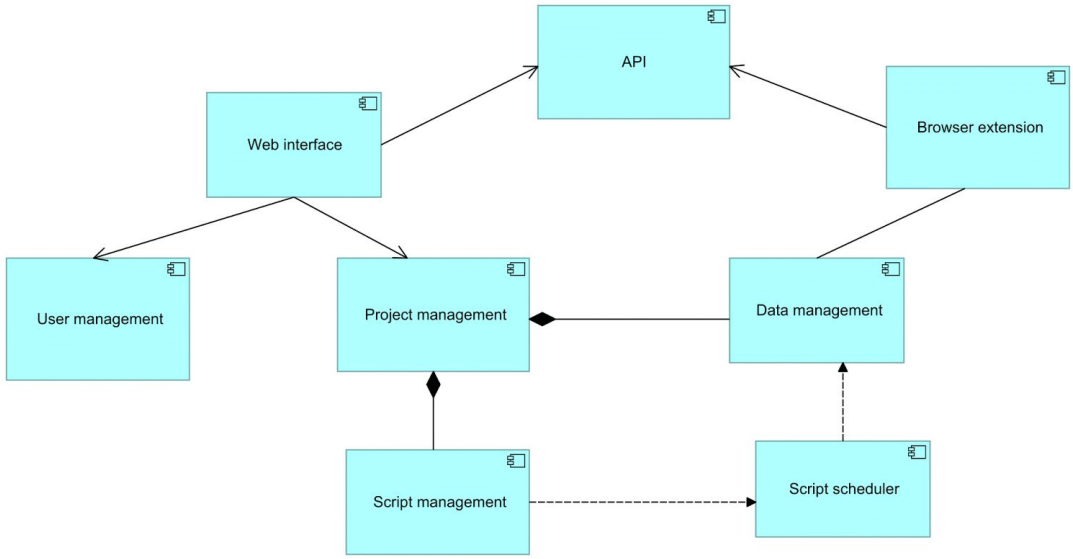
3.2 Testovanie

Testovanie MVP prebiha z dvoch pohľadov. Prvú predstavuje prípadová štúdia, ktorá je realizovaná v spolupráci s organizáciou Slovensko.digital a druhá predstavuje len testovanie UX z pohľadu bežného používateľa, nakoľko prípadová štúdia nevyžaduje používateľské rozhranie ako také.

UX testovanie je opísané v samostatnom dokumente.

4 Celkový pohľad na systém

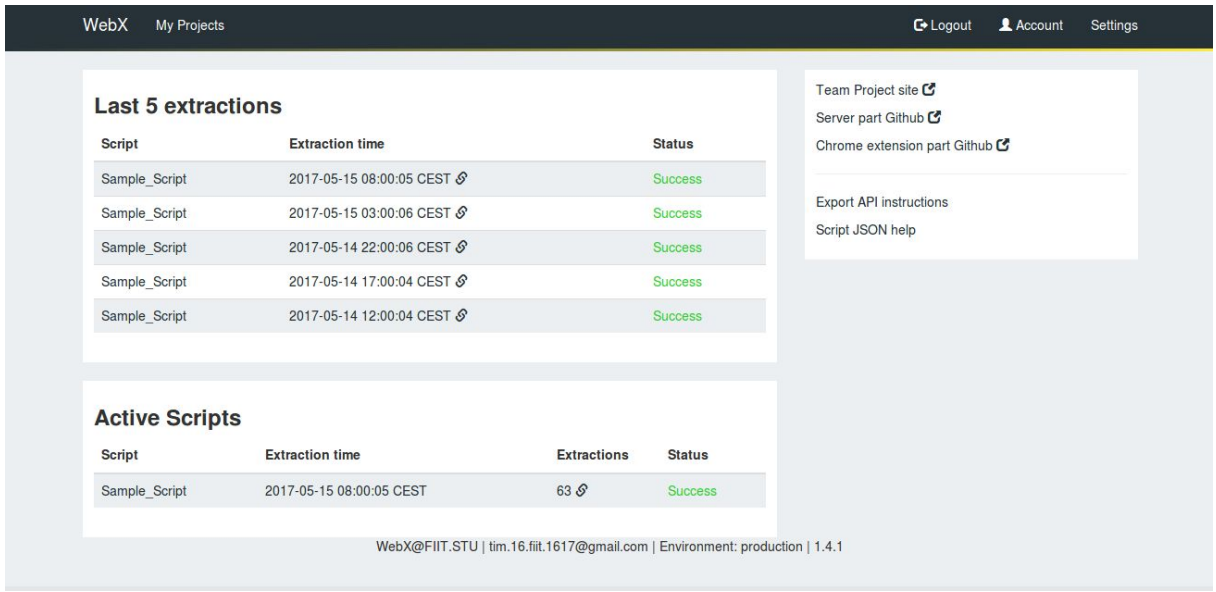
Podľa informácií od zákazníka, resp. vlastníka produktu je hlavnou funkciou systému automatizované zbieranie preddefinovaných dát z vybraných webových stránok. Používateľ si pomocou grafického rozhrania (webová stránka) môže spravovať vlastné projekty, skripty a dátové schémy, podľa ktorých sa vykonáva extrakcia. Anotácia dát na požadovanej stránke je realizované pomocou rozšírenia do webového prehliadača.



Obr. 1 - Zobrazenie modulov systému a ich vzťahy

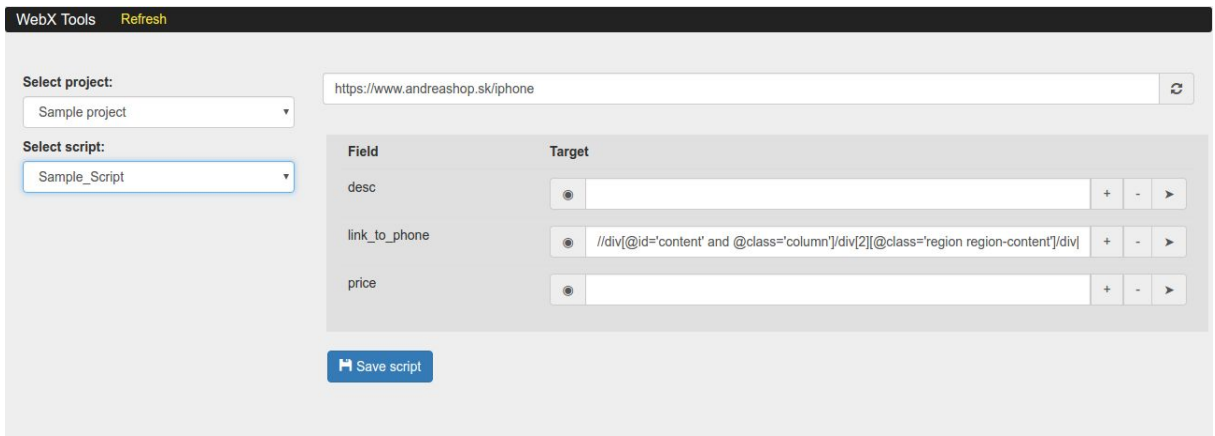
Systém ako taký je tvorený dvoma základnými časťami. Dekompozícia celého systému na časti je znázornená na obr. 1.

Prvou je webová aplikácia poskytujúca prístup k účtu používateľa, s možnosťou správy projektov a skriptov na zber dát. Okrem toho je súčasťou aj služba v podobe "crawler-a", ktorý vykonáva samotné, používateľom definované skripty, a teda zbiera už konkrétne dáta. Tak isto je v tejto aplikácií možné spravovať svoj vlastný účet (napr. zmeniť heslo).



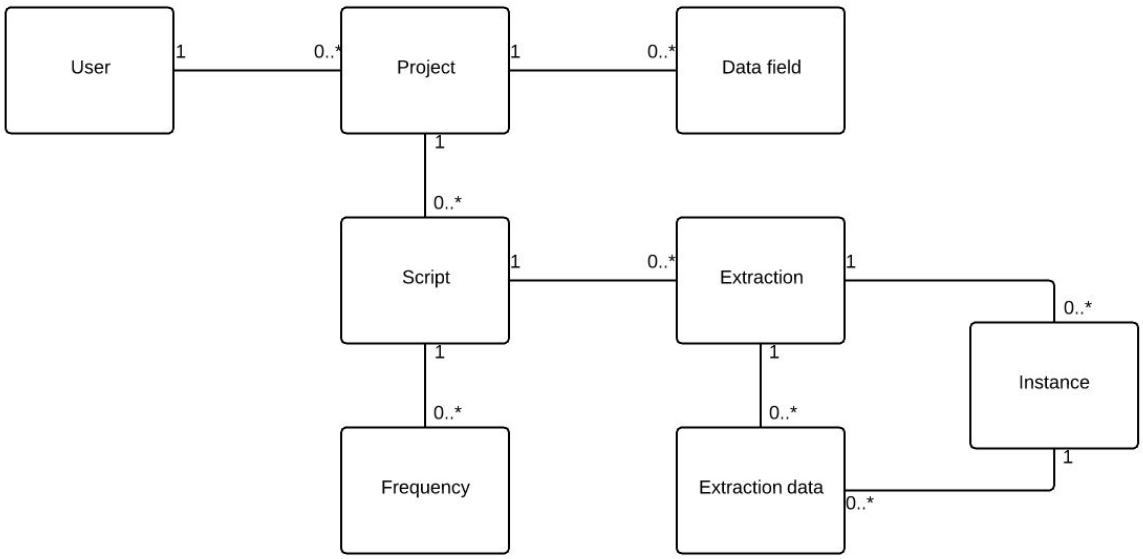
Obr. 2 - Webová aplikácia

Druhá časť je už spomínané rozšírenie do prehliadača, pomocou ktorého používateľ (po úspešnom prihlásení sa) jednoducho definuje, aké dáta a z ktorej stránky sa majú extrahovať.



Obr. 3 - Extension do prehliadača Google Chrome

Dátovým model našej aplikácie obsahuje tabuľku pre používateľov (User). Používateľovi patrí viacero projektov (Project), pričom každému projektu patrí viacero dátových polí (Data Field) a skriptov (Script). Samotnému skriptu prislúcha viacero frekvencí spustenia (Frequencies) a extrakcií (Extraction). Každá extrakcia má viacero inštancií (Instances) a každá inštancia viacero dát extrakcie (Extraction Datum).



Obr. 4 - Dátová schéma

5 Moduly systému

Pre lepšie pochopenie systému, pomerne veľkú komplexitu problému a možnosť určitej systematickej implementácie boli definované 4 základné moduly, ktoré logicky vychádzajú z diagramu v predchádzajúcej kapitole. Moduly prechádzajú postupnou implementáciou a každý z nich je podrobnejšie opísaný v samostatnom dokumente, ktorý sa venuje analýze, návrhu, implementácií a testovaniu konkrétneho modulu.

Systém teda pozostáva z týchto modulov:

1. User management - zaoberá sa používateľom a správou jeho účtu
2. Project management - venuje sa používateľom vytvoreným projektom a skriptom
3. Browser extension - rieši všetko, čo sa týka rozšírenia do prehliadača Google Chrome
4. Extraction management - obsahuje informácie o riešení procesu extrakcie a spracovaní dát