Používateľská príručka

[WebExtraction]

Systém pre extrakciu dát z webu

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:	Andrej Vaculčiak
Verzia číslo:	1.1
Dátum poslednej zmeny:	15.05.2017

Obsah

1 Vitajte!

V prvom rade sa Vám chceme v mene tímu WebX poďakovať za vybratie si nášho produktu, ktorého hlavnou úlohou je pomôcť Vám získať, pomerne rýchlo, jednoducho, no hlavne pravidelne a automaticky, pre Vás užitočné dáta vo veľkom množstve a z rôznych webových sídiel.

Želáme Vám čo najväčšie množstvo úspešne extrahovaných dát!

Tím WebX

2 Úvod

Táto používateľská príručka má za úlohu objasniť základné pojmy spojené s extrakciou dát, ktoré sú použité v našom systéme. Okrem toho obsahuje opis základných rozhraní a opis potrebných krokov pre úspešnú extrakciu, spolu s ukážkami obrazoviek všetkých potrebných náležitostí.

3 Základné rozhranie

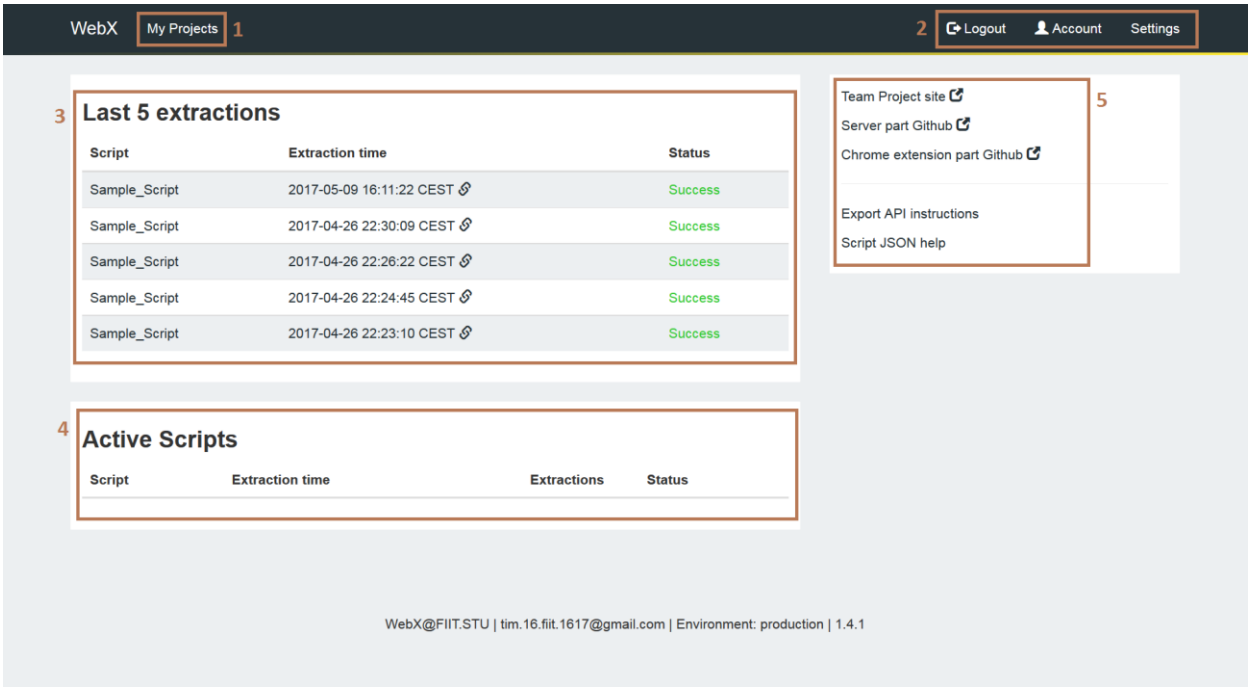
Služba sa skladá z dvoch základných častí:

1. Webové sídlo
2. Extension pre webový prehliadač

Pomocou webového sídla si používateľ spravuje svoje osobné údaje a definuje a upravuje projekty, skripty a vidí výsledky extrakcií, ktoré prebehajú alebo prebehli.

Extension pre prehliadač slúži najmä k interaktívnemu priradeniu konkrétnych polí na konkrétnych web stránkach k definovaným dátovým poliam v rámci projektu a skriptu.

3.1 Webové sídlo

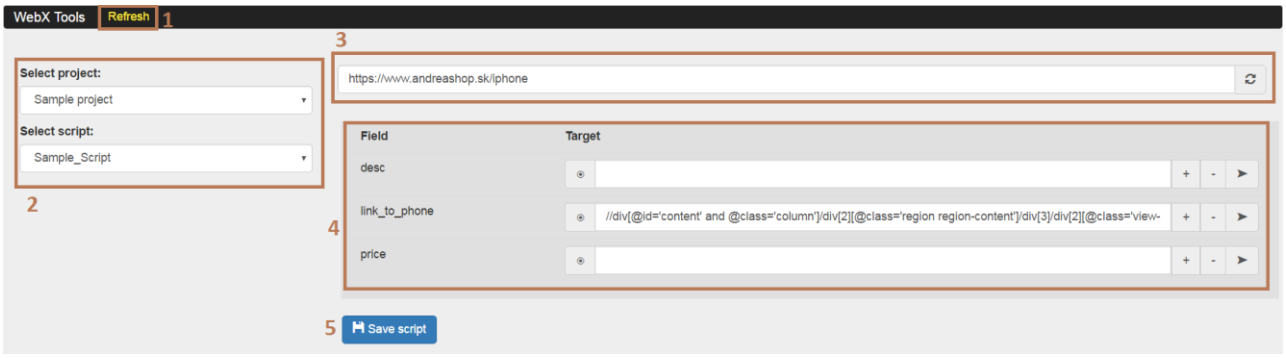


Obr. 1 – Domovská obrazovka webového sídla po prihlásení sa

Rozhranie domovskej stránky, zobrazenej na Obr. 1, pozostáva z týchto častí:

1. Tlačidlo na prechod do používateľom vytvorených projektov
2. Tlačidlá pre odhlásenie zobrazenie účtu a nastavení používateľových údajov
3. Tabuľka s prehľadom posledných ukončených 5 extrakcií
4. Tabuľka s prehľadom aktuálne bežiacich extrakcií
5. Externé odkazy pre stránku tímu, repozitár webového sídla + rozšírenia do prehliadača na Github-e a odkazy na stránky s inštrukciami a popisom API pre export dát a tvar JSONu, ktorý tvorí skript (viď kap. 5)

3.2 *Extension pre Google Chrome*



Obr. 2 – Základné rozhranie rozšírenia do prehliadača Google Chrome

Rozšírenie pre prehliadač Chrome má tieto základné časti:

1. Tlačidlo pre aktualizáciu dát v rozšírení (v prípade, že sme uskutočnili úpravu projektu/skriptu vo webovej časti)
2. Dropdown polia pre výber konkrétneho projektu a skriptu
3. Pole pre URL stránky z ktorej chceme dáta extrahovať
4. Zoznam dátových polí a k nim prislúchajúce xPath-y + tlačidlá pre operácie s nimi
5. Tlačidlo na uloženie aktuálneho stavu skriptu

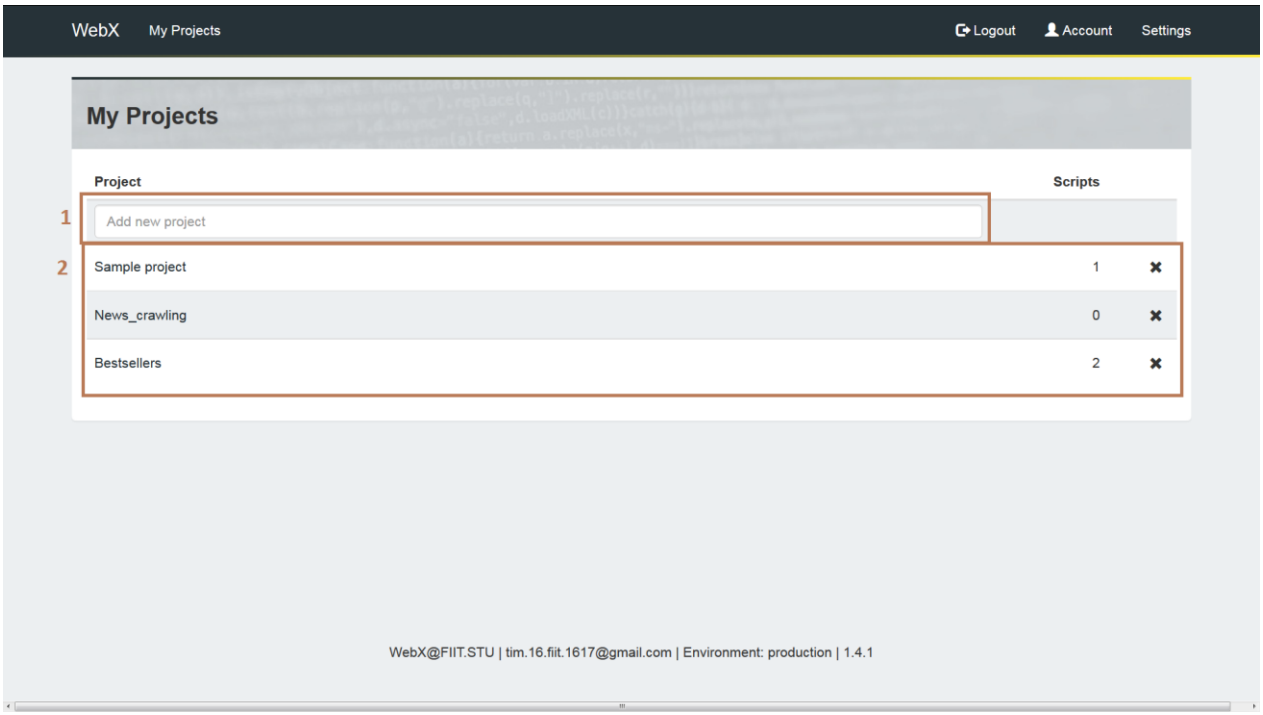
4 Projekt

Projekt sa v kontexte nášho systému chápe ako doména, v ktorej chceme získavať dáta. Môžeme chcieť napríklad extrahovať informácie z rôznych e-shopov o notebookoch alebo z viacerých spravodajských portálov získavať rôzne titulky článkov. Ako ďalší príkald môže byť viacero stránok s počasím, ktoré chceme zbierať a vytvoriť si napríklad podklady (datasety), na ktorých postavíme náš systém a pod.

4.1 Vytvorenie projektu

Ak sa chceme pracovať s projektami, stačí, keď na úvodnej stránke klikneme na tlačidlo "My Projects" (viď. Obr. 1).

Zobrazí sa nasledujúca obrazovka:

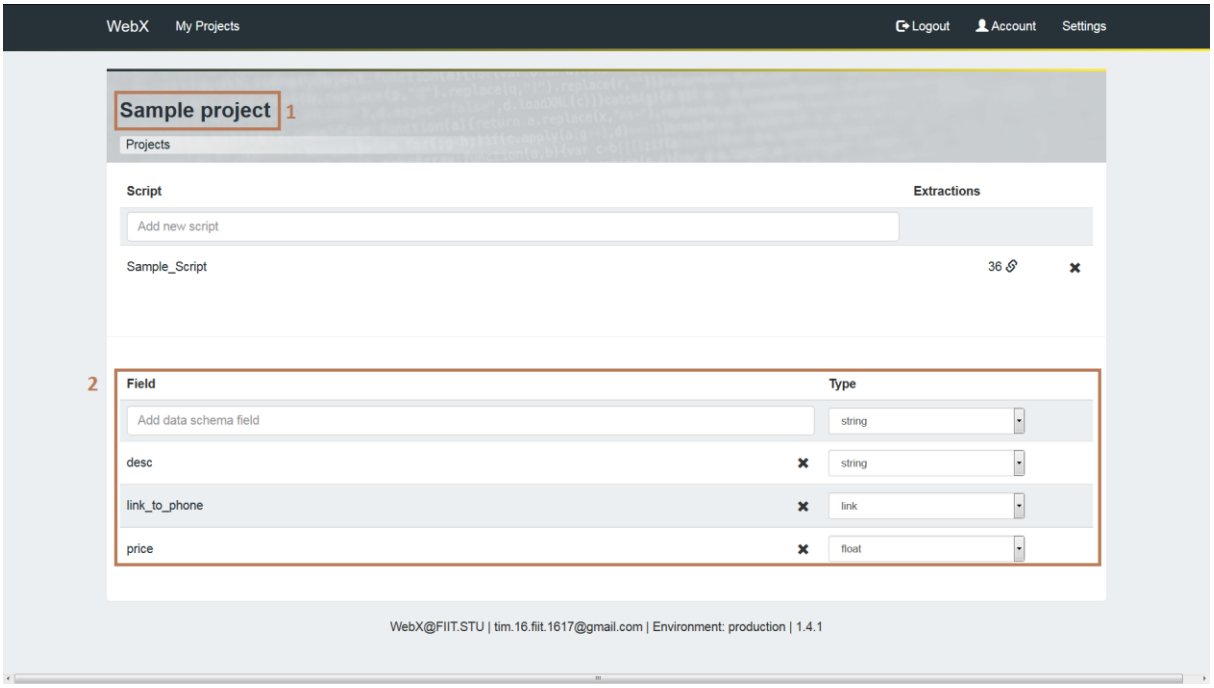


Obr. 3 – Obrazovka pre prácu s projektami

Na tejto obrazovke môžeme vytvoriť nový projekt (1) alebo vidíme prehľad už vytvorených projektov (2).

Do konkrétneho projektu sa dostaneme kliknutím na jeho názov.

4.2 *Úprava projektu a definícia dátových polí*



Obr. 4 – Obrazovka pre úpravu projektu a manažment dátových polí pre projekt

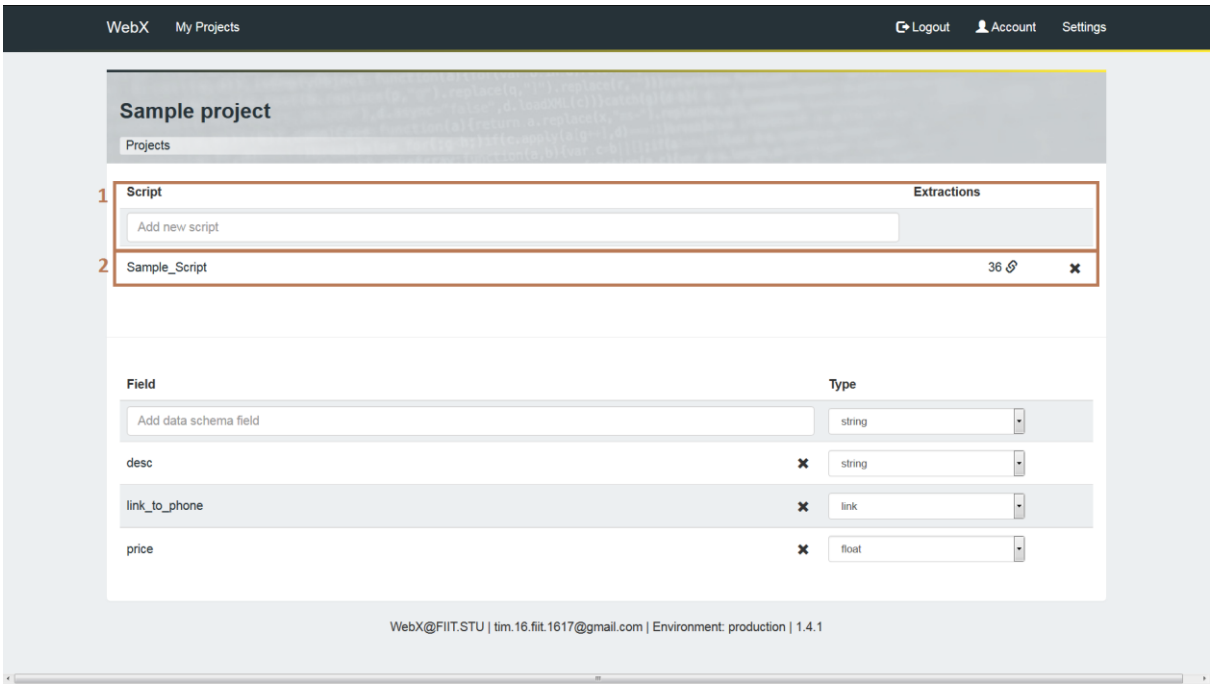
Po prechode na obrazovku úprav (Obr. 4), môžeme:

1. Upraviť názov projektu (kliknutím na názov)
2. Pridať/ upraviť existujúce dátové pole spolu s jeho typom

5 Skript

Každý skript predstavuje jednu web stránku v rámci danej domény, z ktorej chceme extrahovať dáta. Ak napríklad extrahujem titulky novinových článkov, vytvorím si jeden skript pre BBC News a jeden pre Al Jazeera.

5.1 Vytvorenie skriptu



Obr. 5 – Vytvorenie skriptu v rámci projektu

Vytvorenie nového skriptu prebieha na obrazovke správy projektu (pole 1 na Obr. 5).

Okrem toho tu vidíme aj zoznam už vytvorených skriptov (2), spolu s počtom vykonaných extrakcií.

5.2 Úprava a možnosti skriptu

Do úprav novo vytvoreného skriptu sa dostaneme kliknutím na jeho názov (pole 2 na Obr. 5).

Na obrazovke skriptu (Obr. 6) máme k dispozícii tieto možnosti:

1. Upraviť názov skriptu (po kliknutí na názov)
2. Upraviť JSON skriptu (v podstate teda samotný skript)

Pri úpravách skriptu je potrebné dodržiavanie určitých zásad a je potrebné poznať jeho štruktúru. Základný pomocník pre skript nájdeme na adrese:

http://team16-16.studenti.fiit.stuba.sk/webx/json_help

3. Prehľad informácií z poslednej vykonanej extrakcie (pokiaľ nejaká prebehla) spolu s jednoduchou štatistikou prázdnych polí.

Po kliknutí na text s dátumom poslednej extrakcie sa zobrazia extrahované dáta v tejto extrakcií.

4. Nastavenie úrovne log záznamov (viac v kap. 5.2.1)
5. Uložiť úpravy JSONu a zmenu úrovne log záznamov
6. Zobraziť zoznam extrakcií pomocou API (“API List Extractions“)

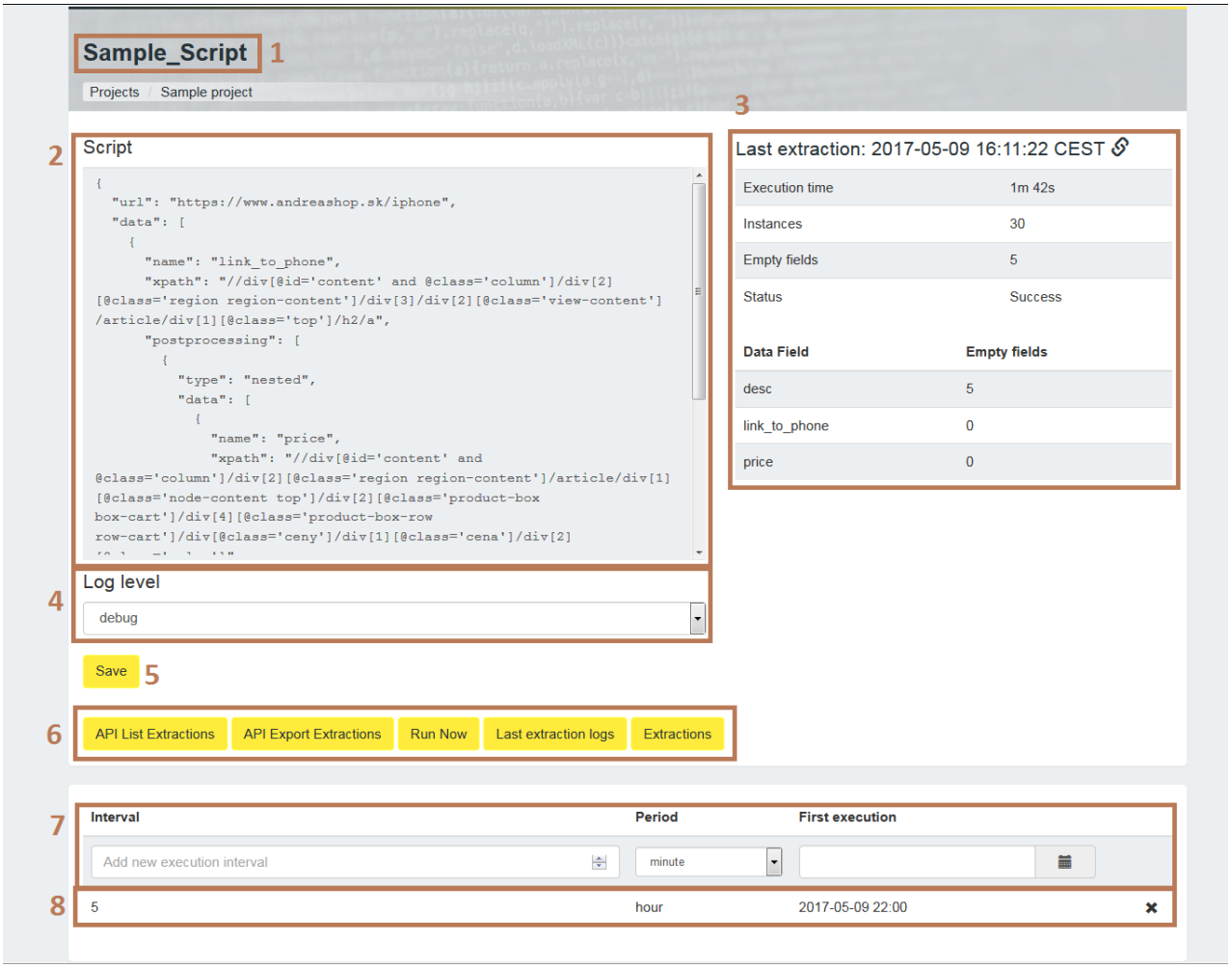
Zobraziť dáta posledných 10 extrakcií (“API Export Extractions“)

Spustiť extrakciu (“Run Now“)

Zobraziť log záznamy poslednej extrakcie (“Last extraction logs“)

Zobraziť prehľad extrakcií pre daný skript (“Extractions“)

7. Definovať nový interval vykonávania extrakcie (viac v kap. 5.2.2)
8. Upraviť už existujúci interval spúšťania extrakcie



Obr. 6 – Obrazovka pre úpravy skriptu

5.2.1 Úrovne log záznamov

Ku každej extrakcií je možné vytvárať log záznamy rôznej úrovne. Predvolená hodnota je „Error“.

Označenie úrovne	Popis
Debug	Najpodrobnejšia úroveň. Uchováva podrobné informácie o stave extrakcie.
Warning	Menej podrobná. Uchováva najmä neočakávané udalosti a chyby počas extrakcie
Error	Najmenej podrobná. Uchováva len chybové hlásenia extrakcie.

5.2.2 Nastavenie pravidelnosti spúšťania

Pre každý skript je možné nastaviť jeden alebo viac intervalov spúšťania. Pri definovaní tohto intervalu (pole 7 na Obr. 6) sa nastavujú nasledovné parametre:

- Interval – koľkokrát uplynie perióda (Period), kým sa daná extrakcia znova spustí
- Period – časové kvantum pre jeden Interval (minúta, hodina, deň, mesiac)
- First Execution – dátum a presný čas prvej extrakcie

Príklad:

Chceme daný skript vykonávať každých 5 hodín. Hodnoty budú nasledovné:

Interval = 5, Period = hour

Ak chceme vykonávať skript každých 15 minút, bude situácia nasledovná:

Interval = 15, Period = minute

6 Extrakcie

Extrakcia je vo svojej podstate samotná operácia, počas ktorej sa podľa JSONu v skripte extrahujú konkrétne dáta určené xPathom pomocou rozšírenia do prehliadača Chrome. Spúšťa sa pravidelne, automaticky, podľa používateľom definovaného intervalu (kap. 5.2.2) alebo manuálne po kliknutí na tlačidlo na obrazovke pre prehľad a úpravy konkrétneho skriptu (pole 6 na Obr. 6).

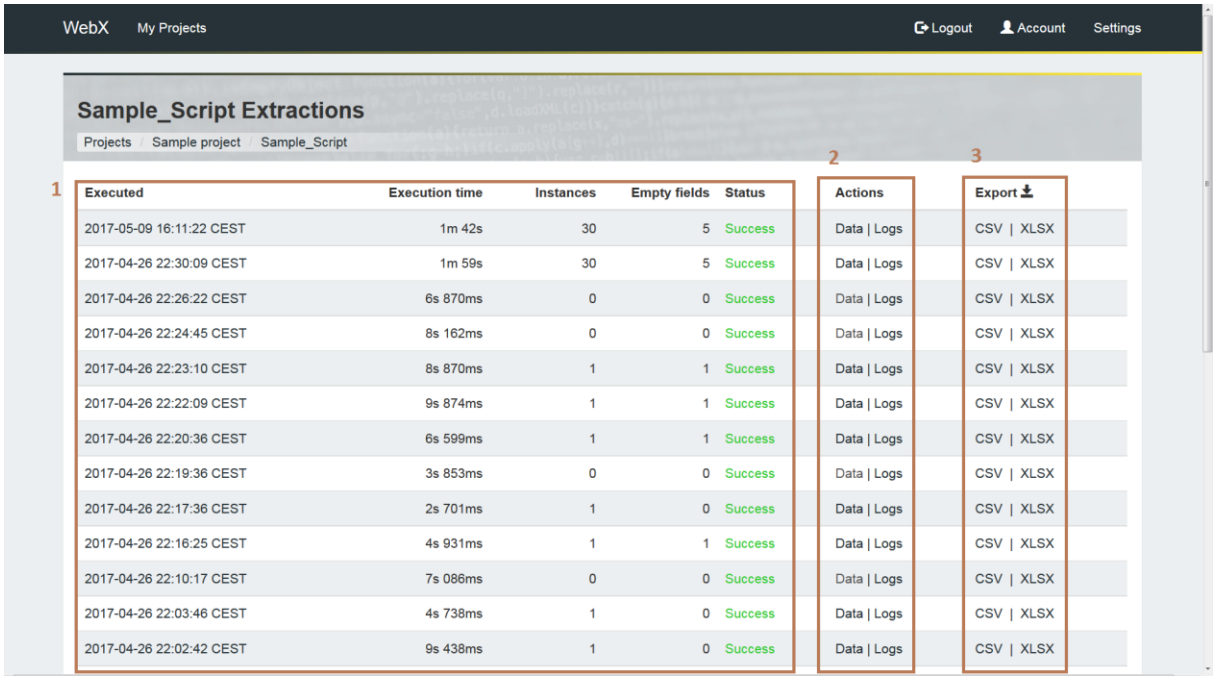
Každá extrakcia má svoje špecifické hodnoty atribútov. Medzi atribúty zaraďujeme:

- Čas vykonávania – koľko daná extrakcia trvala
- Status – konečný stav extrakcie (Success, Running, Fail)

Okrem toho má aj odvodené atribúty:

- Počet prázdnych polí – ak sa extrahuje prázdna hodnota
- Počet inštancií – počet objektov v rámci extrakcie

6.1 Zobrazenie extrakcií pre skript



Obr. 7 – Prehľad extrakcií pre konkrétny skript

Do zobrazenia všetkých extrakcií pre konkrétny skript sa dostaneme tlačidlom „Extractions“ (pole 6 na Obr. 6). Nájdeme tam prehľad základných informácií o každej extrakcií (pole 1) spolu s dostupnými akciami (pole 2).

6.2 Akcie nad dátami

Pre každú extrakciu vieme buď zobraziť konkrétne extrahované dáta alebo zobraziť log záznamy pre konkrétnu extrakciu (pole 2 na Obr. 7).

6.2.1 Zobrazenie log záznamov

Pri zobrazovaní log záznamov je možné tieto záznamy filtrovať podľa úrovne (viď kap. 5.2.1)

6.2.2 Export dát

Všetky dáta, ktoré získame pomocou extrakcie sú uložené na serveri, z ktorého si ich vieme stiahnuť buď pomocou API alebo v štandardizovanej forme CSV, prípadne XLSX.

6.2.3 API

Export dát pomocou API je možný buď priamo pomocou tlačidiel (pole 6 na Obr. 6) alebo priamo dopytom, pričom je ale potrebné vedieť tento dopyt korektne odoslať. Pomôcku nájdeme na adrese: http://team16-16.studenti.fiit.stuba.sk/webx/export-api-instructions

6.2.4 Formát CSV/XLSX

Pokiaľ chceme dáta stiahnuť v prehľadnej forme, môžeme použiť tlačidlá na export vo formáte CSV alebo XLSX. Tie sú dostupné na viacerých miestach.

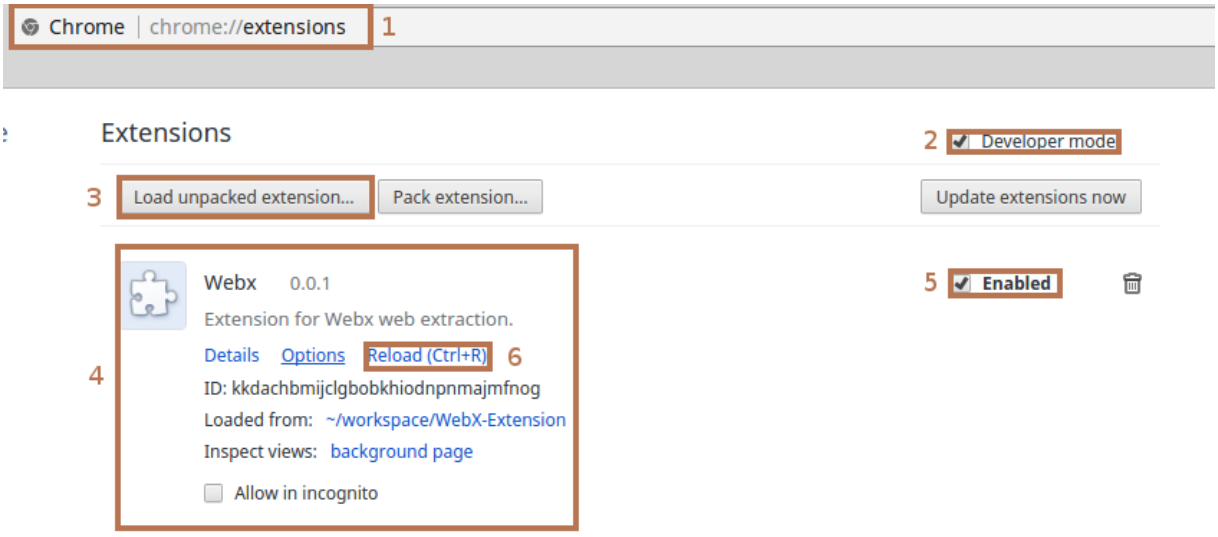
7 Extension

Extension je rozšírenie do Chrome prehliadača, pomocou ktorého sa vytvára skript (vo formáte JSON) pre projekt. Na základe skriptu sú následne sťahované dáta zo stránky. Najprv je potrebné si stiahnuť extension z repozitáru, ktorý sa nachádza na adrese: https://github.com/michal55/WebX-Extension

7.1 Pridanie do prehliadača

Na to, aby bolo možné používať rozšírenie v Chrome prehliadači, je potrebné najprv toto rozšírenie pridať do Chrome prehliadača. Na obrázku 8 sú naznačené nasledovné kroky:

1. V novom okne najprv načítať link chrome://extensions/.
2. Povoliť developer mode.
3. Načítanie samotného extensionu z priečinku, v ktorom je uložené na disku.
4. Po správnom načítaní sa rozšírenie zobrazí na obrazovke.
5. Možnosť deaktivovať alebo aktivovať rozšírenie.
6. Možnosť opakovane načítať rozšírenie v prípade, že nastali nejaké zmeny v rozšírení.



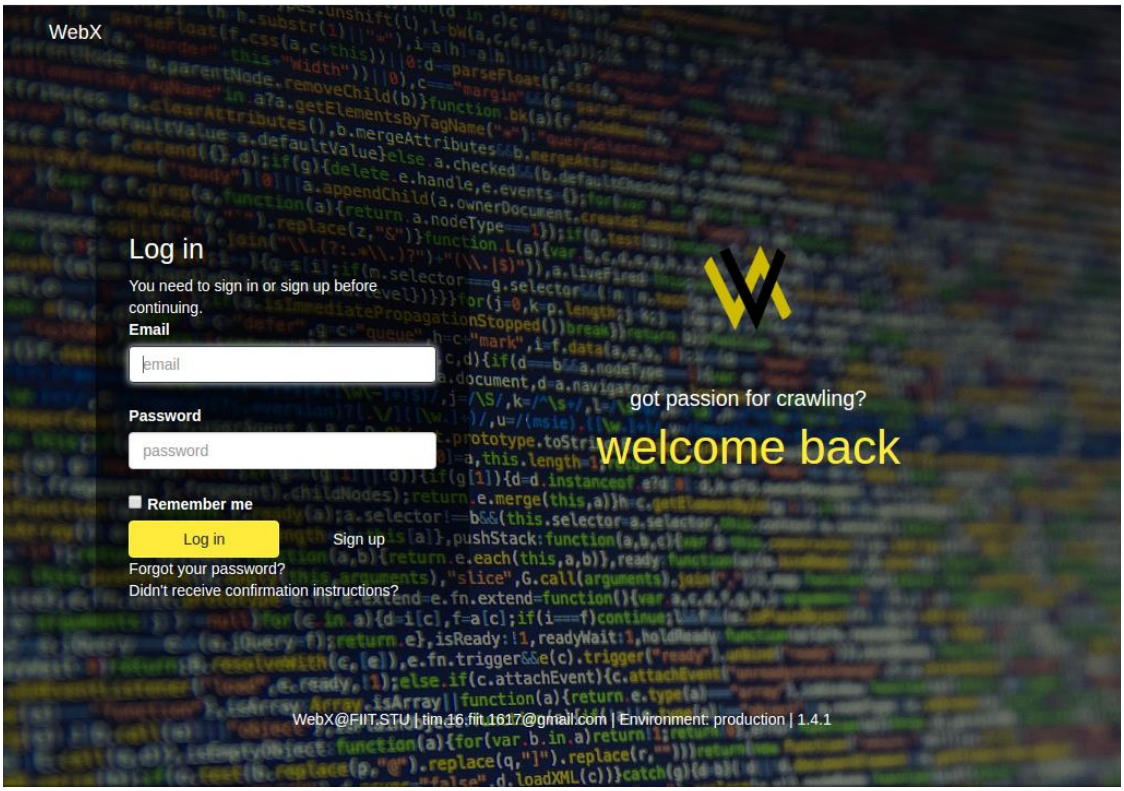
Obr. 8 – Pridanie extensionu do Chrome prehliadača

7.2 Výber elementov

Po načítaní extensionu do Chrome prehliadača je možne ho začať používať. Najprv je potrebné otvoriť developer tools (Vývojárske nástroje), štandardne je to za pomoci klávesy F12. Obrazovka na obrázku 9 zobrazuje extension pred prihlásením:

1. Extension medzi možnosťami vývojárskych nástrojov.
2. Tlačidlo pre prihlásenie. Po stlačení sa zobrazí obrazovka pre prihlásenie sa do webovej aplikácie - obrázok 10.
3. Tlačidlo pre znovu načítanie extensionu.

Obr. 9 – Prihlásenie do extensionu



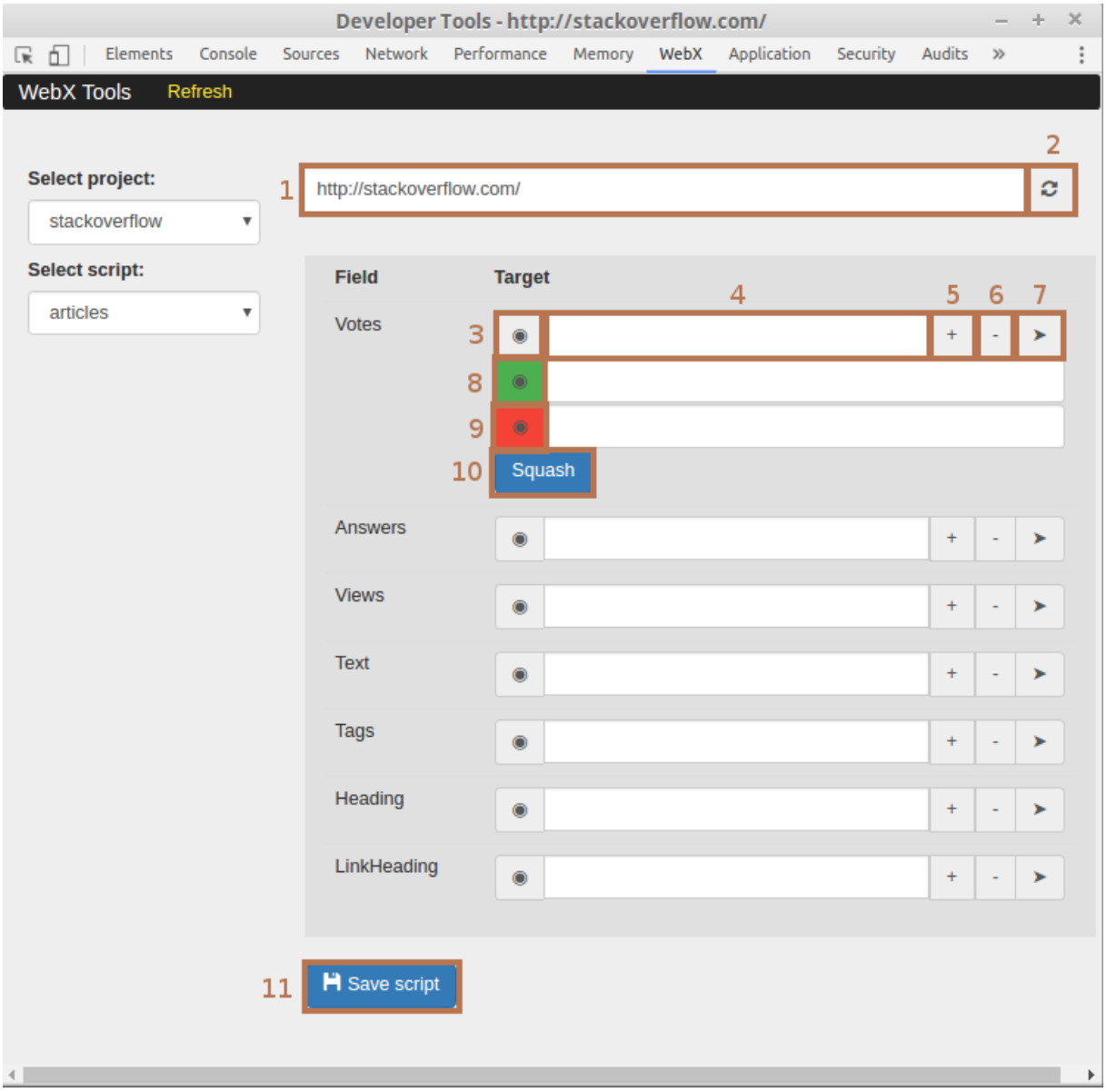
Obr. 10 – Prihlasovacie okno

Po prihlásení sa na obrazovke vľavo zobrazí možnosť zvoliť projekt. Projekt musí už existovať vo webovej aplikácii pred tým ako je možné ho zvoliť. Po zvolení projektu sa zobrazí možnosť zvoliť konkrétny skript. Skript musí taktiež existovať vo webovej aplikácii pre zvolený projekt (v prípade vytvorenia projektu alebo skriptu po prihlásení sa do extension je možné použiť tlačidlo „Refresh“). Po zvolení projektu aj skriptu sa načítajú dátové polia (field) a k ním príslušné akcie. Na obrázku 11 je obrazovka s načítanými dátovými poľami.

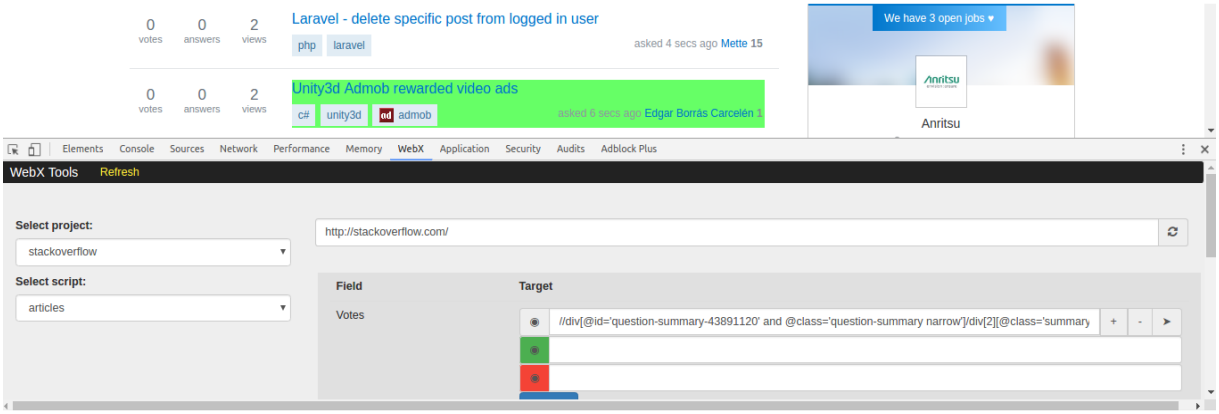
1. Link stránky na ktorej skript začína (url).
2. Tlačidlo pre načítanie adresy stránky aktuálneho okna.
3. Tlačidlo na výber elementu zo stránky. Po stlačení je možné na aktuálnej stránke vybrať element. Jednotlivé elementy sa zvýrazňujú tak, ako je to zobrazené na

obrázku 12. Po kliknutí na konkrétny element sa načíta príslušný xPath, ktorým je daný element opísaný.

4. Pole do ktorého sa načíta xPath. V prípade potreby je možné tento xPath upravovať. Po kliknutí do poľa sa zvýrazní element, ktorý xPath opisuje.
5. Tlačidlo pre pridanie ďalšieho poľa pre xPath ku konkrétnemu dátovému poľu. Používa sa keď chceme zadefinovať xPath, ktorý opisuje pole elementov. Používa sa v kombinácii s tlačidlom z bodu 10.
6. Tlačidlo pre pridanie ďalšieho poľa pre xPath ku konkrétnemu dátovému poľu. Používa sa keď chceme zadefinovať xPath, ktorý opisuje pole elementov avšak chceme vylúčiť konkrétny element z tohto poľa. Používa sa v kombinácii s tlačidlom z bodu 10.
7. Tlačidlo pre pridanie ďalšieho spracovania elementov. Toto je opísane v časti 7.3
8. Tlačidlo s rovnakou funkcionalitou ako v bode 3.
9. Tlačidlo s rovnakou funkcionalitou ako v bode 3.
10. Tlačidlo na zjednotenie xPathov pre konkrétne dátové pole. Po stlačení sa všetky xPathy spoja do jedného, ktorý všeobecne opisuje všetky elementy, ktoré boli zadefinované jednotlivými xPathmi.
11. Tlačidlo na uloženie skriptu.



Obr. 11 – Dátové polia v extension

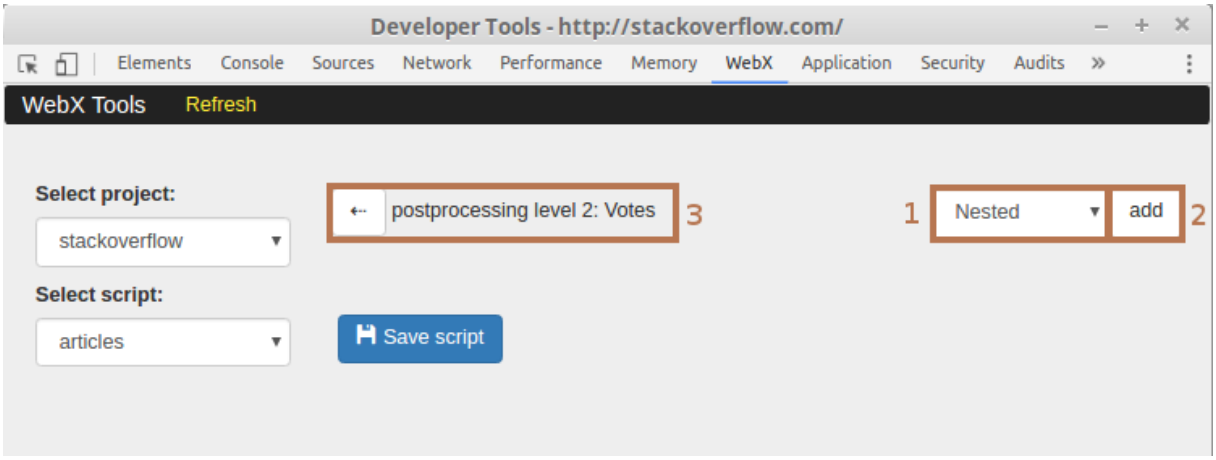


Obr. 12 – Zvýrazňovanie elementov na stránke

7.3 *Postprocessing extrahovaných dát*

Postprocessing je ďalšie spracovanie dátového poľa po extrakcii. V rozšírení je možné ho pridať cez tlačidlo opísane v časti 7.2 bodu 7. Po kliknutí na tlačidlo sa zobrazí obrazovka z obrázku 13. Na tejto obrazovke pribudli tri nové možnosti:

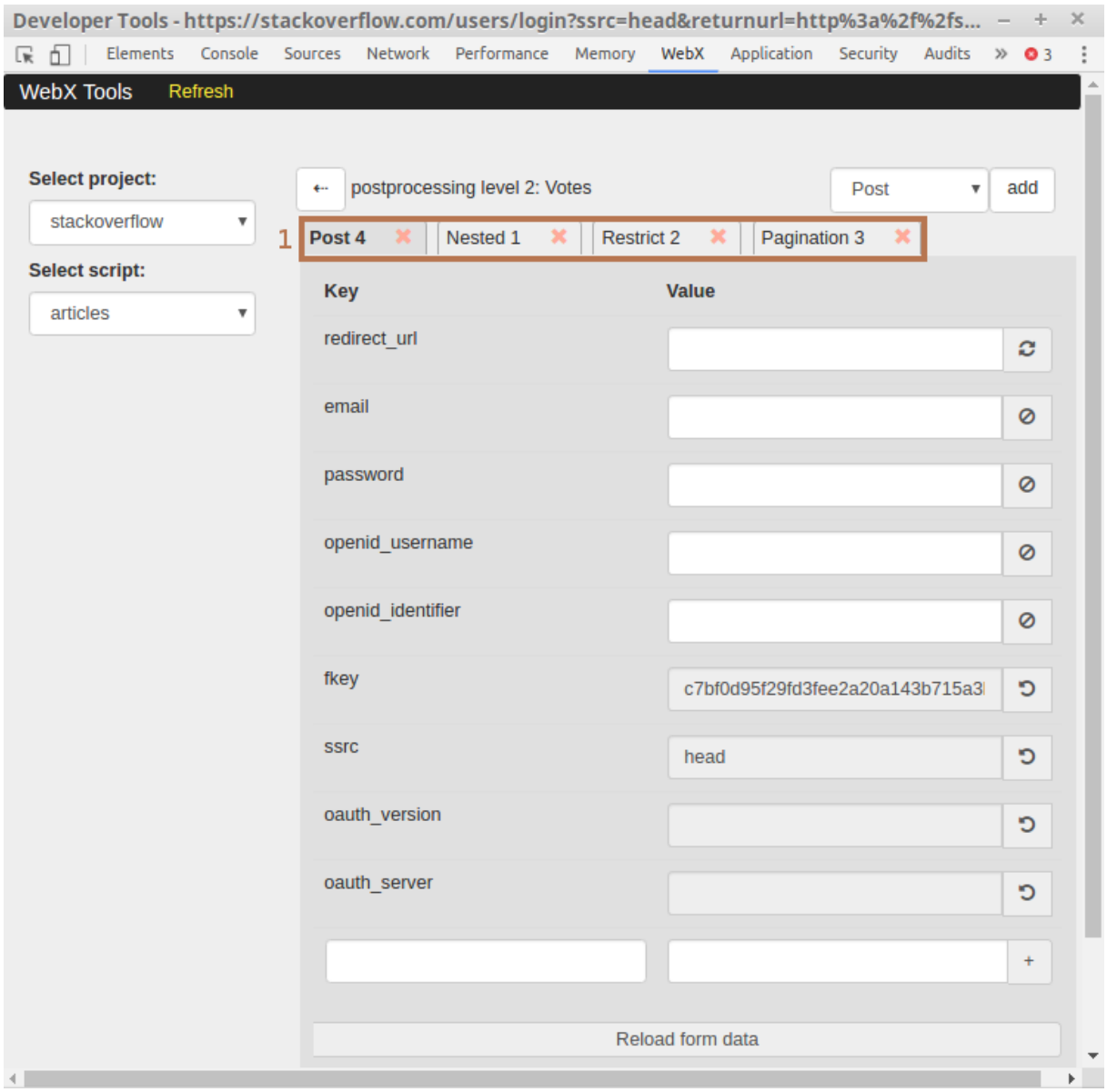
1. Výber z možností ďalšieho spracovania.
2. Pridanie konkrétneho ďalšieho spracovania.
3. Návrat na predošlú obrazovku.



Obr. 13 – Postprocessing základná obrazovka

Pre každé dátové pole je možné vybrať viacero spracovaní. Tieto sa vykonávajú v poradí, v akom boli uložené. Na obrázku 14 je obrazovka s viacerými spracovaniami. Každé spracovanie má svoje číslo a spracovania je možné medzi sebou ľubovoľne vymieňať potiahnutím myšou, alebo ich vymazať. Poradie spracovaní v skripte je postupne zľava doprava. Pridať je možné nasledovné spracovania:

Názov spracovania	**Funkcionalita**
Nested	Vnorenie skriptu. Používa sa keď zadefinujeme xpath, ktorý ukazuje na link, na ktorom chceme extrahovať ďalšie dáta.
Restrict	Obmedzenie na extrahovanie dát v rámci určitého elementu.
Pagination	Skript prechádza stránkovanie v prípade že je to potrebné. Používa sa na stránkach, pri ktorých sú dáta stránkované.
Attribute	Extrahovanie určitého atribútu z html prvku (elementu).
Post	Slúži na prihlásenie, ak nie je možné extrahovať dáta bez prihlásenia.
Filter	Slúži na filtrovanie dát podľa dátumu, extrahovaného v rámci dát, napríklad keď chceme dáta, ktoré boli pridané predošlý deň. (iba pre dátový typ dátum)



Obr. 14 – Príklad postprocessingov