

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
FAKULTA CHEMICKÉJ A POTRAVINÁRSKEJ
TECHNOLÓGIE



ÚSTAV INFORMATIZÁCIE, AUTOMATIZÁCIE A MATEMATIKY
Oddelenie informatizácie a riadenia procesov

Webové služby pre spracovanie HTML stránok

Bakalárska práca

Vypracoval: Ján Dzurov
Vedúci práce: Ing. Michal Kvasnica, PhD.

Bratislava 2008

Touto cestou by som chcel poďakovať vedúcemu bakalárskej práce Ing. Michalovi Kvasnicovi, PhD za odborné vedenie, cenné rady i pripomienky ktoré mi poskytoval pri spracovaní bakalárskej práce.

Abstrakt

Bakalárska práca sa zakladá na využívaní služieb *dapper.net* a *Yahoo! Pipes* pre prácu s RSS. Cieľom je oboznámenie sa s formátom RSS a jeho využitie v praxi.

Teoretická časť sa zaoberá štruktúrou a významom RSS formátu na internete. V praktickej časti sa práca zaoberá tvorbou RSS prúdov a ich následným filtrovaním a zobrazovaním. Na tvorbu RSS sa využila služba stránky *dapper.net*, ktorá je tu stručne charakterizovaná. Spracovanie výstupných prúdov je uskutočnené službou *Yahoo! Pipes*. Na záver je uvedený opis rôznych funkcií tejto služby a ich demonštrovanie.

Abstract

This bachelor's thesis is based around the services *dapper.net* and *Yahoo! Pipes* for handling RSS feeds. The aim is to introduce the RSS format and to adress its practical usage.

The theoretical part deals with the structure of RSS and its meaning on the internet. The practical part deals with the creation of RSS feeds and their consecutive filtering and display. The creation of RSS feeds is realized by the *dapper.net* service which is briefly characterised here. The produced feeds are handled by the *Yahoo! Pipes* service. The various functions of this service are then charactierized and shown in practise.

Obsah

ÚVOD.....	6
1. TEORETICKÁ ČASŤ.....	7
1.1 XML.....	7
1.2 RSS.....	9
2. PRAKTICKÁ ČASŤ.....	13
2.1 Dapper.net – Opis.....	13
2.2 Práca v Dapp Factory – Vytváranie RSS feedov.....	14
2.3 Yahoo Pipes.....	19
2.3.1 Ukážka možností Yahoo! Pipes.....	20
2.3.2 Opis niektorých ďalších modulov v Yahoo Pipes.....	27
ZÁVER.....	37
LITERATÚRA.....	38

ÚVOD

Internet sa stal v posledných rokoch jedným z najdôležitejších zdrojov informácií. Užívatelia si môžu vybrať zo širokej škály webových stránok, z ktorých mnohé poskytujú často aktualizované informácie, či už sú to správy, oznamy alebo príspevky do diskusie alebo blogu. V tomto kontexte môžeme hovoriť až o prehĺtení informáciami. Mnohí ľudia trávajú veľa času prehľadávaním svojich obľúbených stránok, hľadajúc nový obsah. Takýto postup je pomalý a neefektívny, lebo prezerajú každú stránku zvlášť.

Tento nový problém si vyžaduje také riešenie, ktoré by umožnilo informovať užívateľov o novinkách resp. aktuálnych zmenách na webe. Efektívne vysporiadanie sa s týmto problémom ponúka RSS, webový formát ktorý sa stáva v súčasnosti veľmi populárnym a používa ho nielen väčšina spravodajských webov, ako sú CNN, BBC, ale ujal sa už aj na niektorých obchodných portáloch napr. *alza.cz*. Pomocou RSS prúdov sa informácie dostanú k užívateľom a prehľadne zobrazujú novinky z vybranej oblasti internetu.

Na webe ďalej existujú služby, ktoré umožňujú tieto informačné prúdy vytvárať a rôzne upravovať a sú dostupné každému. Ovládaním týchto služieb si užívateľ môže tvoriť jedinečné aplikácie na filtrovanie a vyhľadávanie obsahu na internete.

Cieľom práce je popis štruktúry a účelu RSS prúdov (feedov), ako aj vypracovanie návodu na prácu so službami na adresách *dapper.net* a *pipes.yahoo.com*.

1. TEORETICKÁ ČASŤ

1.1 XML¹

Moja práca sa zaoberá modifikovaním RSS feedov, ale pre porozumenie ako fungujú je treba najprv uviesť niektoré základné informácie o ich formáte, ktorým je XML.

Funkcia XML

XML je skratka pre Extensible Markup Language, čo v preklade znamená rozširiteľný značkovací jazyk. Tento jazyk bol vytvorený na popis dát, v kontraste z HTML, ktorý bol vytvorený na zobrazovanie dát. XML môžeme charakterizovať ako sebaopisujúci, pretože obsahuje tagy ktoré si užívateľ sám vytvorí. Slúži na uchovávanie a charakterizáciu dát spôsobom ktorý si užívateľ sám určí vytvorí.

Štruktúra XML

```
<?xml version="1.0"?>
<note>
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don't forget me this weekend!</body>
</note>
```

Na začiatku sa deklaruje verzia a potom sa uvádzajú jednotlivé elementy. Nasledujúci popis sa týka príkladu uvedeného na obrázku, t.j. správy formou listu. Základný koreňový element je <note>, ktorý má v tomto prípade štyri podelementy: <to>, <from>, <heading> a <body>. XML je citlivý na veľkosť písma, takže napr <Letter> a <letter> by boli dva rôzne elementy. Oproti HTML majú tagy v XML vždy zatvárací tag. Každý XML dokument musí obsahovať koreňový element. Všetky ostatné elementy musia byť správne vnorené do koreňového elementu. Príklad správneho vnorenia:

```
<root>
  <child>
    <subchild>
    </subchild>
  </child>
</root>
```

V XML, ako aj HTML môžu elementy mať svoje atribúty. Atribúty elementov v XML musia byť v úvodzovkách. Príklad použitia atribútov:

```
<?xml version="1.0"?>
<note date="12/11/99">
<to>Tove</to>
<from>Jani</from>
<heading>Reminder</heading>
<body>Don't forget me this weekend!</body>
</note>
```

Atribúty poskytujú dodatočné informácie, ktoré nie sú súčasťou obsahu XML dokumentu. Sú dôležité pri spracovaní týchto dokumentov a definujú sa vždy v šartovacom tagu elementu. Miesto atribútov sa dajú v XML použiť aj dodatočné elementy, tento spôsob je považovaný za lepší.

```
<?xml version="1.0"?>
<note date="12/11/99">
<to>Tove</to>
<from>Jani</from>
<heading>Reminder</heading>
<body>Don't forget me this weekend!</body>
</note>

<?xml version="1.0"?>
<note>
<date>12/11/99</date>
<to>Tove</to>
<from>Jani</from>
<heading>Reminder</heading>
<body>Don't forget me this weekend!</body>
</note>

<?xml version="1.0"?>
<note>
<date>
<day>12</day>
<month>11</month>
<year>99</year>
</date>
<to>Tove</to>
<from>Jani</from>
<heading>Reminder</heading>
<body>Don't forget me this weekend!</body>
</note>
```

V prvom dokumente je pre definovanie dátumu použitý atribút, v druhom podelement a tretí spôsob je upresnenie pomocou viacerých vnorených elementov.

1.2 RSS ²

Definícia

RSS je skratka pre “Really Simple Syndication“ (Veľmi jednoduchá syndikácia). Je to dokument vo formáte XML ktorý obsahuje často modifikované informácie z jednej alebo viacerých internetových stránok v štrukturovanej forme. Obsahuje nové články, linky, správy, zápisy do blogu a ďalšie informácie vo forme jednoduchého zoznamu.

Web Feedy

Feedmi alebo prúdmi sa nazývajú zoznamy RSS titulkov, ktoré majú spoločnú tému alebo zdroj viacerých stránok, ktoré sú štrukturované v XML formáte. Mimo RSS je najznámejší formát Atom.

Úloha

Informácie na internetových stránkach sa menia rôznym tempom, nové správy sa objavujú v nepredvídateľných časových rozmedziach. Pre užívateľov internetu, ktorý chcú dostať najnovšie informácie z viacerých internetových stránok, sa postupné prezeranie všetkých webov stáva čoraz zdĺhavejším procesom.

Posielanie emailov o novom obsahu na stránkach bolo prvotným riešením tohto problému, avšak táto metóda je neprehľadná a môže byť nesprávne interpretovaná ako spam.

RSS ponúka lepší spôsob oznamovania čerstvých informácií z internetových stránok. Správy o novinkách z viacerých zdrojov sú tak ľahko spracovateľné a výsledky sú zobrazované v organizovanej forme. RSS používa na informovanie len titulok správy alebo krátky sumár, ktorý je odkazom na článok alebo správu na dotýčnej stránke.

Použitie ³

RSS sa v počiatkoch používalo prevažne na syndikovanie, združovanie správ, ale postupom času webmasti našli širokú škálu využitia, napríklad :

- Oznamy nových produktov
- Oznamy nových pracovných príležitostí

- Správy o nových témach na fórum stránok
- Programy kín a televízie
- Oznamy meškaní v leteckej doprave

Možnosti RSS sa neustále rozširujú a okrem týchto existuje mnoho ďalších spôsobov ako túto formu komunikácie využiť.

Štruktúra ⁴

Príklad RSS dokumentu:

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<rss version="2.0">
<channel>
  <title>W3Schools Home Page</title>
  <link>http://www.w3schools.com</link>
  <description>Free web building tutorials</description>
  <item>
    <title>RSS Tutorial</title>
    <link>http://www.w3schools.com/rss</link>
    <description>New RSS tutorial on W3Schools</description>
  </item>
  <item>
    <title>XML Tutorial</title>
    <link>http://www.w3schools.com/xml</link>
    <description>New XML tutorial on W3Schools</description>
  </item>
</channel>
</rss>
```

Prvý riadok – deklarácia XML – definuje verziu XML a kódovanie znakov použitých v dokumente. V tomto prípade je dokument verzie 1.0 a používa kódovanie ISO-8859-1 (Latin-1/West European).

Nasleduje deklarácia RSS (verzia 2.0).

Najprv sa definuje element <channel>, ktorý opisuje samotný feed a má tri podelementy: <title> - definuje meno RSS feedu, alebo ináč povedané prúdu

<link> - definuje adresu feedu

<description> - tento element stručne opisuje obsah feedu

<channel> môže mať viac nepovinných podelementov, dôležité sú ešte:

<copyright> - používa sa keď je materiál licencovaný/copyrightovaný.

<image> - definuje logo feedu, ktoré sa zobrazí pri načítaní feedu. napr:

<language> - určuje jazyk v ktorom je feed napísaný

Elementy <item> obsahujú novú správu, článok, prídavok do feedu. Feed môže mať neobmedzené množstvo <item> elementov. Má nasledujúce povinné podelementy:

<title> - Určuje titulok správy

<link> - Definuje hyperlink na zdroj správy

<description> - opis správy

Niektoré nepovinné podelementy elementu <item>:

<pubDate> - Definuje dátum publikácie správy

<author> - e-mail autora správy

Čítanie RSS feedov ⁵

Existuje mnoho aplikácií na prezeranie feedov, tieto sa nazývajú *reader*, čiže čítač, alebo tiež agregátor. Browser Firefox obsahuje *RSS reader*, nové feedy sa pridávajú kliknutím na ikonu  v adresovom riadku. Feed sa tak začne odberať z dotyčného zdroja, a aktualizuje sa o nové položky pri jeho zavolaní čítačom.

Pre účely tejto práce som používal webovú službu *Google Reader*, ktorý je zadarmo, a nachádza na adrese www.google.com/reader. Pre použitie je potrebné sa zaregistrovať. Ukážka feedu stránky www.sme.sk v *Google Readeri*:



Jednotlivé elementy <item> sú v oddelených bunkách, titulky sú zvýraznené a stránka ponúka veľa možností na rôzne úpravy ako zoradenie a výpis detailov.

Na stránkach <http://www.rssreader.com/> je zadarmo na stiahnutie čítač RSS aj Atom feedov, stránka <http://www.feedreader.com/> ponúka taktiež agregátor zadarmo.

Programy na pracovanie s RSS feedmi

V súčasnosti existuje niekoľko programov, ktoré umožňujú rýchlu a jednoduchú tvorbu RSS:

Na stránkach <http://www.feedforall.com/>, <http://www.extralabs.net/feed-editor.htm> a <http://www.jitbit.com/rss-feed-creator.aspx> sa ponúkajú na stiahnutie programy za poplatok.

Na stránkach <http://www.banham.cc/feededit/index.htm>, <http://home.hetnet.nl/~bsoft/rssbuilder/index.htm>, a <http://ssrss.sourceforge.net/> sú k dispozícii na stiahnutie freeware programy.

2. PRAKTICKÁ ČASŤ

2.1 Dapper.net – Opis ⁶

Mnoho stránok v súčasnosti nemá svoj vlastný RSS feed, ale pomocou služieb na stránke www.dapper.net sa táto nevýhoda dá odstrániť. Túto funkciu som použil aj na účely tejto práce. Stránka Dapperu umožňuje extrahovať dáta z webových stránok a využiť ich pre rôzne účely v rôznych formátoch, ktoré okrem RSS zahŕňajú napr: Google Map, XML a Google Gadget.

Logged in as [johnydz](#) | [My Account and Dapps](#) | [Logout](#)



Get any content from the web

Enter a website or the content / services you want:

Choose format (RSS, XML, Google Maps...)

or [create a new Dapp](#)

Get more traffic to your site

Easily provide new means for people to access your content (such as RSS). You can use Dapper to create feeds, widgets, and APIs with your content and links.

New to Dapper?

- [Check out our demo](#)
- [Explore user generated Dapps](#)
- [Visit our forum](#) new

Developers

- [Create an API for any site](#)
- [Dapper Developer Community](#)

Business Services

- [Learn more](#)
- [Service Level Agreements](#)

Services That Use Dapper

**Semantify**
(beta)
Add semantics to your site

**MashupAds**
Content is the best ad.

**Facebook App Maker**
Create Facebook Applications

**DapperFox**
Firefox extension

Featured Partners

**Yedda**

**Swivel**

**WorkLight**

2.2 Práca v Dapper Factory – Vytváranie RSS feedov

Prvý krok je zadanie zdroja ktorý chceme využiť, tj. URL stránky, alebo existujúceho feedu.

Dapper Dapp Factory Welcome to the Dapp Factory [having problems?](#)

Welcome to the Dapp Factory

Where is the content you want to use?

☐ In a website ☒ In an existing RSS Feed

Enter the URL of the website:

[Advanced](#)

Choose a format (You can always change this later)

☐ Dapp XML

☒ **RSS feed**

☐ Google Gadget

☐ Netvibes Module

☐ Google Map

☐ iCalendar

☐ Image Loop

☒ New version (Beta) ☐ Old version [Next Step](#)

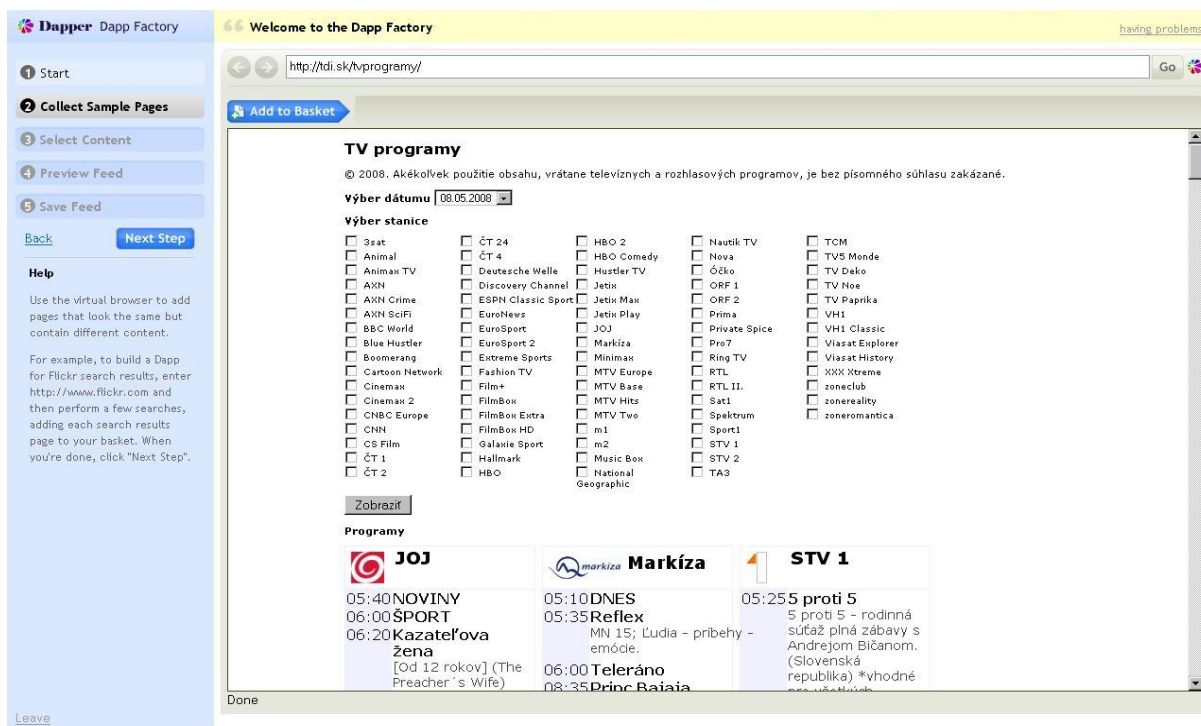
Not sure what will happen next? See the "Create a Dapp" demo.

If you are using an ad blocking extension, consider disabling it as it may cause problems.

Na začiatku si tiež môžeme zvoliť formát výstupu, ten ale môžeme neskôr zmeniť.

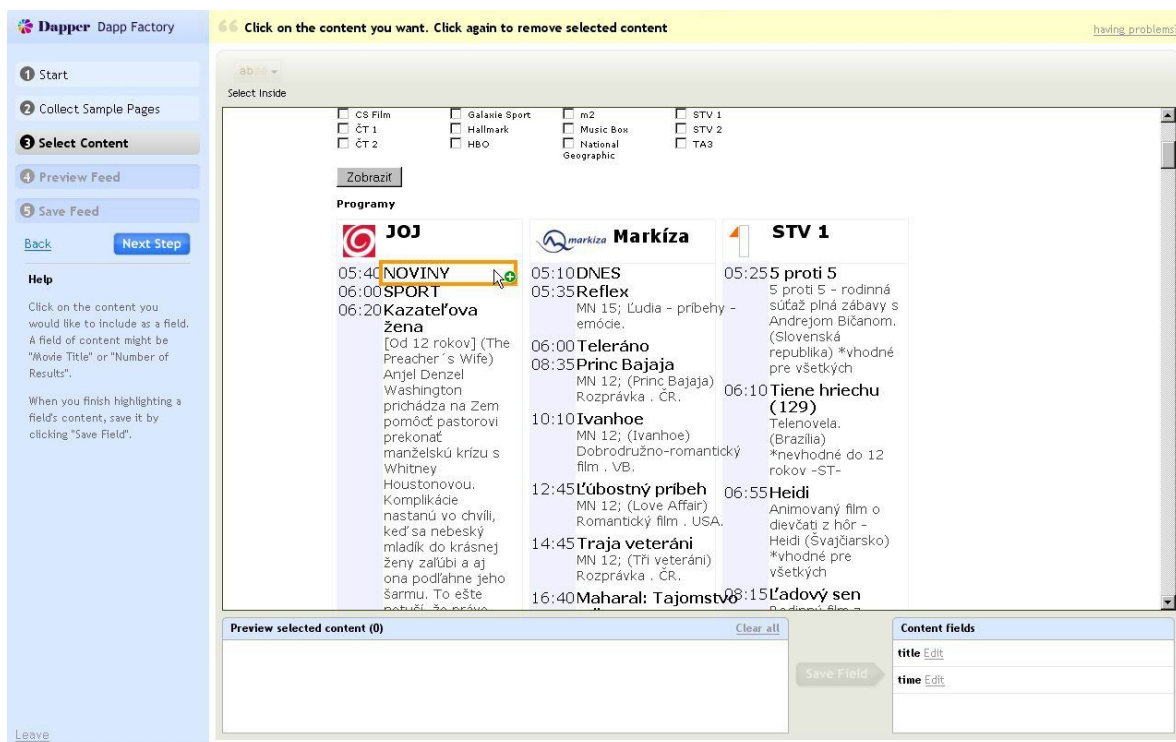
K dispozícii je taktiež video dokumentujúce jednoduchú prácu s aplikáciou.

Po potvrdení zadaných údajov stlačením "Next Step" sa presunieme do ďalšieho kroku.



V hlavnom okne sa zobrazí takzvaný “virtual browser”, ktorý zobrazuje zadanú stránku. Na to, aby sme s touto stránkou mohli pracovať, musíme ju pridať do košíka pomocou “Add to Basket“. Je umožnené pridať viac stránok, a to buď zadaním URL, alebo navigáciou vo zvolenej stránke.

Takto zvolené stránky by mali mať podobnú štruktúru prvkov, akými sú napr. zoznamy a tabuľky. Napomáha to Dapperu pri výbere obsahu stránky, ktorá nasleduje v ďalšom kroku. Je odporúčané, aby sa vybrali aspoň dve stránky.



V tomto kroku je užívateľ vyzvaný aby si vybral tie časti stránky, s ktorými chce pracovať. Výber sa koná na pôvodnej stránke, ktorú sme zadali na začiatku.

Kliknutím na akúkoľvek časť stránky sa nám táto časť označí. Označené časti sa odstraňujú opätovným kliknutím. Je veľmi výhodné pracovať so stránkami, ktoré sú prehľadné a dobre naformátované vo forme tabuliek, lebo *Dapper* spoznáva isté znaky označených dielov a označuje aj niektoré ďalšie ktoré sú podobné. Týmto spôsobom sa dajú kliknutím na jeden prvok - "Noviny" - označiť všetky názvy programov na stránke.

Chcel som vytvoriť feed pre jednu TV stanicu, napr. STV1, takže musím z výberu vylúčiť časti stránky v stĺpcoch pre ostatné stanice. *Dapper* spoznáva, že všetky údaje má brať z jedného stĺpca. Keď máme všetky položky ktoré potrebujeme, môžeme náš výber uložiť.

Po kliknutí na „Save Field“ a pomenovaní, sa údaje zgrupujú do jedného zoznamu dát. Podobným spôsobom som označil aj vysielačie časy v prvom stĺpci.

1 Start

2 Collect Sample Pages

3 Select Content

4 **Preview Feed**

5 Save Feed

[Back](#) [Next Step](#)

Help

When two or more fields are related to one another, they can be grouped together.

For example, if you created "Theater Name" and "Theater Address" fields, you can group them together into a "Theater" group. The resulting Dapp will have one group for every movie theater found.

[Leave](#)

66 Your group has been saved [having problems?](#)

Extracted Content:

TV item

time 05:25

title 5 proti 5

TV item

time 06:10

title Tiene hriechu (129)

TV item

time 06:55

title Heidi

TV item

time 08:15

title Ľadový sen

TV item

time 09:40

title Clifford

TV item

Content Fields [Select all](#)

☐ title

☐ time

Groups

TV item [Edit](#)

[Save Group](#)

Vytvorí sa zoznam prvkov, ktorý naznačuje ako budú vyzerat' prvky výsledného RSS feedu. Prednastavené prvky obsahujú všetky položky ktoré sme v minulom kroku vytvorili. *Dapper* priradí ku každej položke "title" jednu položku z množiny "time", tak ako boli usporiadané na pôvodnej stránke. Keby som si označil aj opis programu, aj ten by sa pridal k týmto dvom položkám a spoločne by tvorily jeden RSS záznam s tromi elementami.

- 1 Start
- 2 Collect Sample Pages
- 3 Select Content
- 4 Preview Feed
- 5 **Save Feed**

[Back](#)
[Next Step](#)

Save Your Dapp

Save your Dapp so that you and others can use it.

Dapp Name: **TV STV1 version 1 (XHTML)**

Description:

Tags:
Enter one or more tags, separated by commas.

Input Variables: **None** [Edit](#)
optional [help](#)

Advanced: ☐ This is a test Dapp [help](#)
optional
☐ Only include full groups
☒ **Make private**

After saving, you'll be taken to your new Dapp

[Save](#)

[Leave](#)

Teraz je potrebné náš výtvar pomenovať, opísať a stanoviť, kto k tomu bude mať prístup. Po uložení si ešte môžeme niektoré veci upraviť.

Logged in as **johnydz** | [My Account and Dapps](#) | [Logout](#)

[Get Content from the web](#)
[Content Owners](#)
[Developers](#)

TV STV1 version 1 (XHTML)

Dapp creator options: [Edit Dapp](#) [Make public](#) [Delete Dapp](#)

Is this a Dapp for your site? [Expose semantics to search engines](#)

Data Mapping

Coming Soon...
"Screenshot"
This Dapp is in the queue and will have its pretty picture taken shortly
Screenshot 1 month ago, <http://tdi.sk/stvprogramy/>

Dapp Description

Program stanice STV1

Terms of Use

Content Owner: [tdi.sk](#)

Dapp Creator

[see all Dapps by this user](#)

Direct Permission Required

[<](#) Choose another format

Create RSS Feed

Each Item in an RSS feed has a standard template(Title, Link, Summary, and Date). Choose the fields you want and preview the RSS feed below.

Choose Content

Item Title:

Item Text: ☐ title ☒ time

Date Published:

☐ Fix Dates (experimental)

Set Input

[Update Input](#)

RSS <http://www.dapper.net/transform.php?dappName=TVSTV1vers>

Preview

5 proti 5
05:25

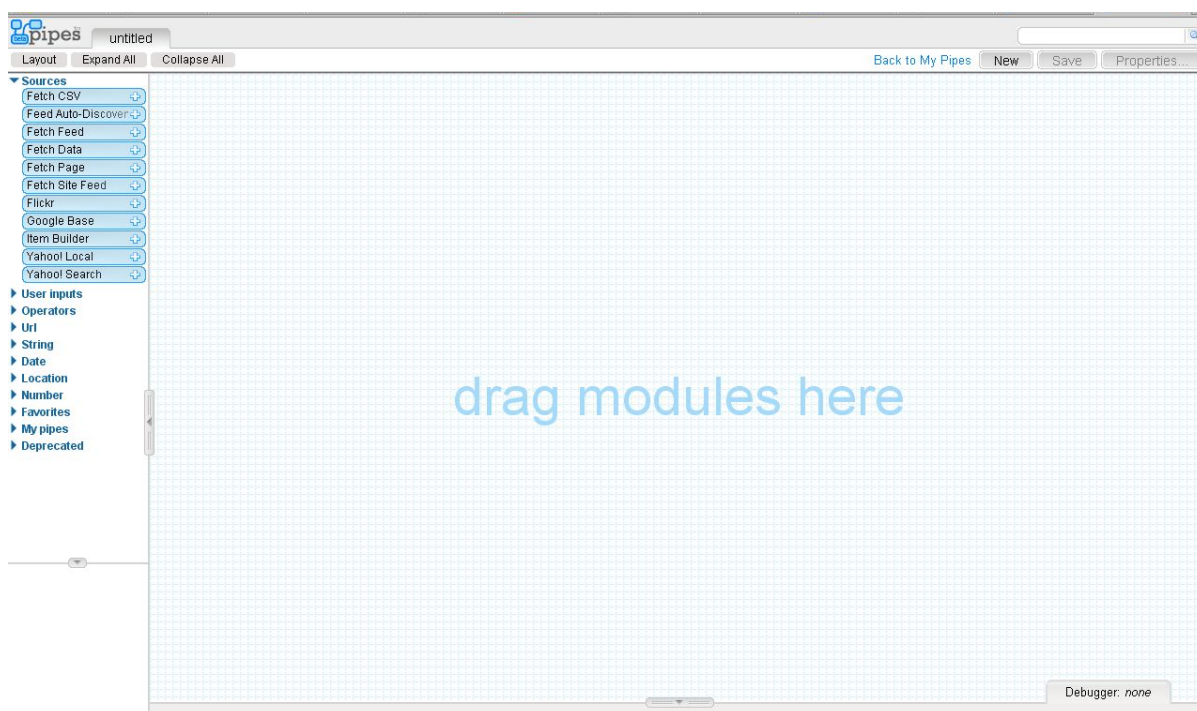
Tiene hriechu (129)
06:10

V tomto kroku si môžeme zmeniť formát, v ktorom údaje uložíme. Moje položky sú vhodné pre RSS feed, takže som si ho zvolil z menu.

Pre dokončenie práce si musíme zvoliť, čo máme priradiť do potrebných položiek RSS feedu. Pre tento formát je povinný element “title”, a pre moje účely sa hodí aby tento prvok tvorili mená programov, ktoré som si zgrupoval do jednej množiny. Element “description” som nastavil tak, aby obsahoval vysielacie časy. Prvok “Date Published” som nechal prázdny, pretože som nemal dostupné informácie zo stránky a pre môj feed by takéto informácie ani neboli užitočné. Dátum, kedy bola dotyčná položka RSS feedu vytvorená je dôležitý pre feedy s charakterom správy.

Výsledkom celého procesu práce v *Dapp Factory* je link, ktorý obsahuje náš feed, a ktorý sa dá otvoriť každým RSS prehliadačom, a je priebežne aktualizovaný.

2.3 Yahoo! Pipes ⁷



Yahoo! Pipes je produktom firmy Yahoo! ktorý umožňuje zlučovať, filtrovať a ďalšími spôsobmi manipulovať s obsahom internetových stránok.

Názov *Pipes* korešponduje s Unixovým znakom „|“ ktorý vykonáva príkaz na predošlom výstupe. Tento princíp sa uskutočňuje aj na tejto stránke, pomocou

prehľadného grafického rozhrania, ktoré je tvorené blokmi príkazov spojených dohromady pomocou pipeov, v preklade rúr.

Pre tvorbu vlastných „rúr“ je potrebná registrácia na stránkach yahoo.com. Užívateľ svoje „rúry“ tvorí presúvaním blokov do pracovnej časti metódou „drag & drop“ a ich vzájomným spájaním. V nasledujúcej časti opíšem spracovanie feedov TV programov.

2.3.1 Ukážka možností Yahoo Pipes

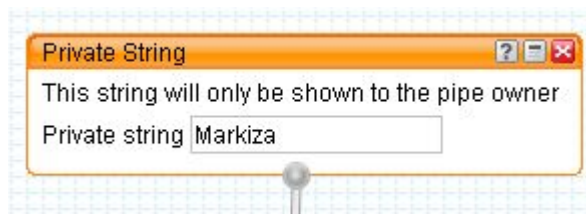
Filtrovanie RSS feedov pre niektoré TV stanice. Cieľom je umožniť užívateľom hľadať programy tak, aby si mohli vybrať stanicu a vysielač čas.

Fetch Feed



Tento blok obsahuje odkaz na feed s ktorým budeme ďalej pracovať. Tlačidlo „+“ umožňuje načítanie viacerých zdrojov, ktoré sa majú zlúčiť. V našom prípade obsahuje jeden RSS feed ktorý som vytvoril pomocou služby *dapper* zo stránky <http://tdi.sk/tvprogramy/>, konkrétne zo stĺpca obsahujúceho TV program stanice Markíza. Keby som blok fetch feed pripojil priamo na výstup z pipu, po spustení (Run Pipe) by sa zobrazil dotýčny feed. Pre účely tejto práce som použil viacero týchto blokov, s odkazmi aj na feedy programov STV1 a JOJ.

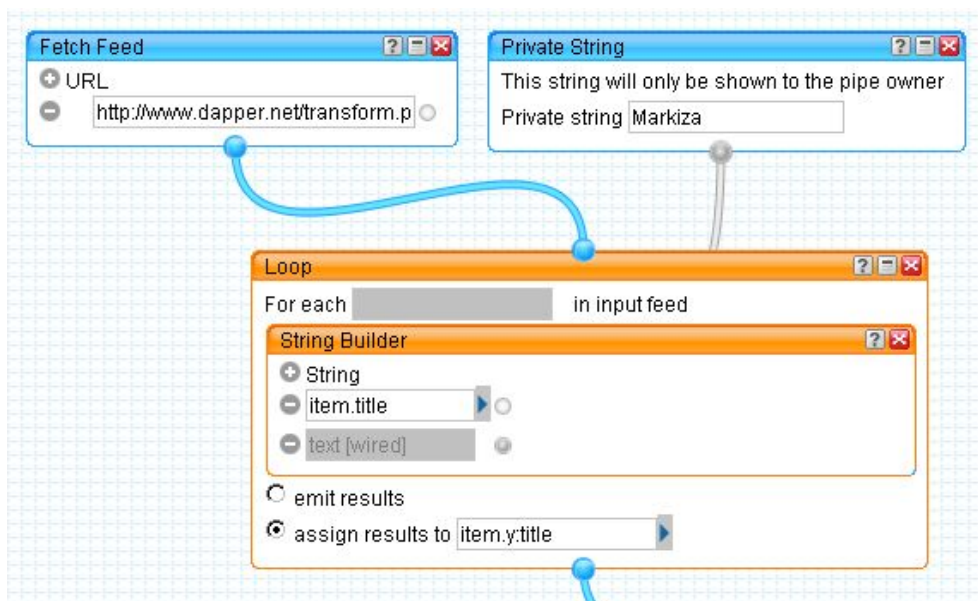
Private String



Pomocou tohoto bloku definuje tvorca pipu textový reťazec, ktorý sa dá pripojiť na vstup iného bloku ktorý to umožňuje. Pre ďalšie spracovanie mojich zdrojov som

potreboval priradiť mená jednotlivých TV staníc k ich feedom. V tomto prípade som zadal reťazec “Markíza”. Ďalšie dva bloky obsahujú texty “STV1” a “JOJ”.

Loop



Blok “Loop” (Slučka) umožňuje vykonať operácie pre každú položku vstupného feedu. Do vnútra bloku sa pridá buď blok s príkazmi alebo hotový pipe. Príkazy sa vykonávajú pre celý feed alebo pre istú jeho časť, ktorá spĺňa kritériá vnoreného prvku. Vstup do tohto bloku je v tomto prípade blok „Fetch Feed“ s linkom na feed stanice Markíza. Ďalšie dva bloky „Loop“ majú identickú vnútornú časť a privádza sa do nich feed zvyšných dvoch staníc.

Výstup sa dá buď vyviesť alebo definovať ako prvok feedu. Druhá možnosť prepíše pôvodné hodnoty. Výstup som priradil do prvku „y.item.title.“

String Builder



Umožňuje zostaviť reťazec z viacerých prvkov. Môže byť zadaný manuálne alebo pripojený z iného bloku obsahujúceho reťazec.

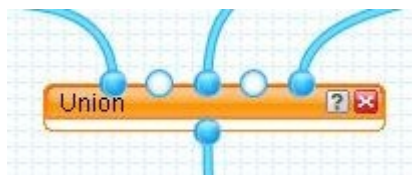
Pridaním nového reťazca tlačidlom “+” sa ten pripojí na koniec predošlého. Pre moje účely som celý blok „String Builder“ vnoril do bloku „Loop“. Týmto spôsobom sa ako reťazec dá zvoliť aj prvok feedu ktorý je vstupom do bloku „Loop“.

Prvú časť môjho reťazca som si zvolil ako prvok „item.title“, čiže titulok prvku, ktorý je pre tento feed totožný s menom programu. Druhú časť reťazca som pripojil z toho bloku „private string“, ktorý obsahuje názov stanice pre daný feed. Pre tento prípad je to reťazec “Markíza”. Príklad výsledného reťazca: “MenoProgramuMarkiza”.

Výsledok bloku Loop

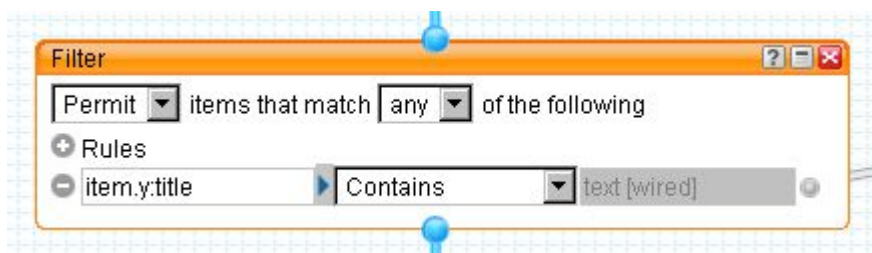
Výstup z tejto časti pipu je pôvodný feed s korigovaným prvkom „y.item.title“. Tieto jeho prvky majú od teraz hodnoty “MenoProgramuMarkiza”. Tento krok bol dôležitý v tom, že umožnil označiť ku ktorému feedu, tj. ku ktorej stanici jednotlivé položky patria.

Union



Tento blok slúži na spojenie až piatich vstupov do jedného výstupu. Výsledkom je jeden spoločný feed. V mojom pipe som spojil výstupy zo všetkých troch blokov „Loop“.

Filter

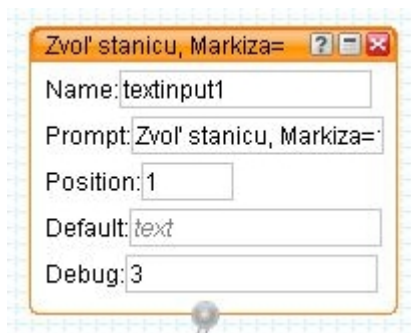


Ďalej som chcel umožniť aby užívateľ mohol istým spôsobom meniť zobrazený výsledok pipu. Blok „Filter“ mu umožňuje aby sa zobrazovali alebo naopak vyradili určité členy feedu.

Prvý selektor v bloku dáva na výber, či sa majú členy ktoré spĺňajú nižšie uvedené kritéria zakázať alebo povoliť (Block/Permit). Druhé políčko stanovuje či sa majú filtrovať členy ktoré spĺňajú všetky alebo hociktoré z požiadavok (all/any).

Pravidlá (Rules) sa pridávajú stlačením “+”. Pravidlá sa vzťahujú na prvky feedu, ako je napr. „item.title“ alebo „item.description“. Musí sa určiť čo má tento prvok spĺňať. Tento konkrétny filter obsahuje pravidlo na výber stanice, preto som si ako definujúci prvok zvolil „item.y:title“, ktorý som si zmenil v predošlej časti. Podmienkou bolo aby sa prepúšťali len tie časti feedov, ktoré mali v prvku item.description meno stanice ktorú si zvolí užívateľ.

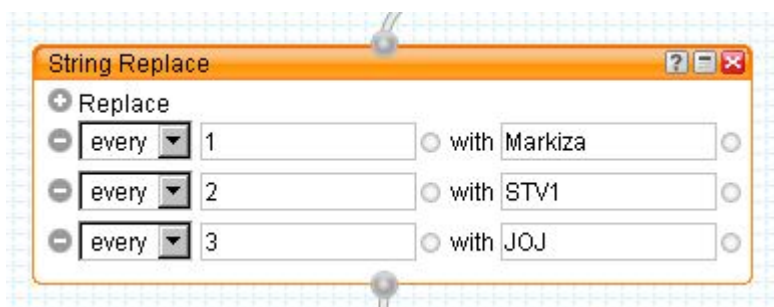
Text Input



Pretože môj cieľ je predovšetkým umožniť užívateľom aby mohli zadávať požiadavky, musím do pipu zakomponovať užívateľský vstup.

Najvhodnejším je v tomto prípade blok „Text Input“. Ten má viacero polí, z ktorých prvý definuje meno vstupu. Zvolil som si predvolenú hodnotu “textinput1”. Druhé pole “Prompt“ slúži na komunikáciu s užívateľom, vyzve ho na zadanie textu. Zadal som “Zvol’ stanicu, Markiza=1, STV1=2, JOJ=3”. Ďalej blok potrebuje poradové číslo vstupu, podľa ktorého ich zobrazí po spustení aplikácie. Zadal som “1”, pretože tento vstup je pre užívateľa prvoradý. Políčko “Default” určuje pre aký vstup sa má vygenerovať výsledok keď sa pipe prvýkrát spustí. Nechal som ho prázdny, pre túto prácu sa nehodí. “Debug” hovorí pipu aby si jeho pri tvorbe generoval výstupy pre zadaný text za účelom kontroly práce.

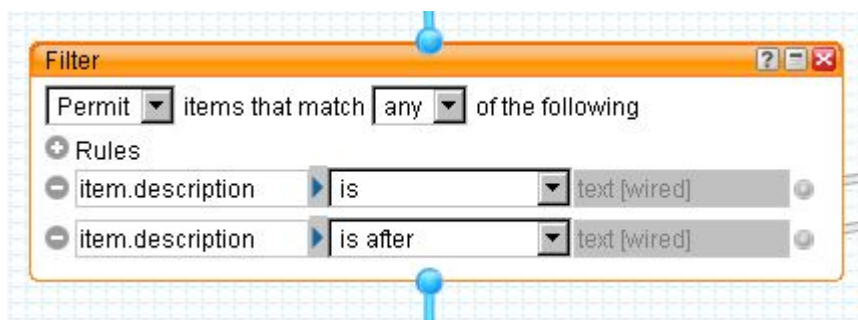
String Replace



Tento blok slúži na zmenenie určitých reťazcov. Pridaním podmienky sa ich môže meniť viac. Pri tomto príkaze sa musí určiť rozsah zmien. Z menu som vybral “every” aby sa všetky výskyty reťazca zmenili na mnoho určenú hodnotu. Pri predošlom bloku som užívateľa vyzval aby zadal číslo reprezentujúce očakávanú stanicu. V tomto kroku som tento vstup zaviedol do aplikácie tým, že som premenoval všetky výskyty čísel 1, 2 a 3 na korešpondujúce mená staníc.

Výstup tohto bloku som zaviedol do Filtra ktorý som si vopred vytvoril.

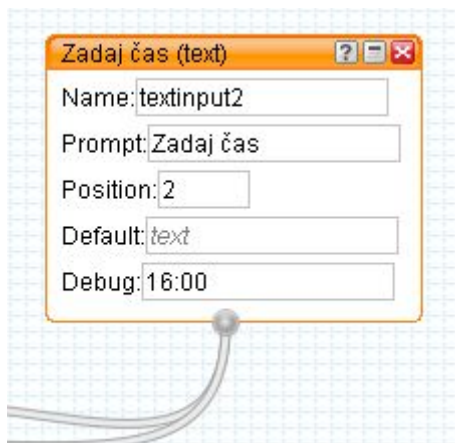
Filter



Ďalší “Filter” je potrebný na to aby sa výsledky ďalej zredukovali na určité časové rozmedzie. Čas vysielania relácie sa nachádza v prvku feedu s názvom “item.description”, takže som ho zvolil ako predmet filtrovania.

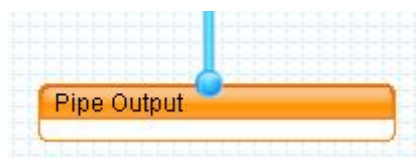
Zvolil som si dve podmienky, a to “is” a “is after”, čo značí, že chcem povoliť len tie časti feedu ktoré majú parameter času vysielania rovný alebo väčší než zadaní čas. Pre dosiahnutie žiadaného výsledku som prikázal filtru aby pre podmienky “item.description is” a “item.description is after” povolil tie údaje ktoré sa zhodujú z hociktorou týchto kritérií. Teraz sa mi pri zadaní času zobrazí všetko čo sa začína presne v ten čas alebo potom.

Text Input



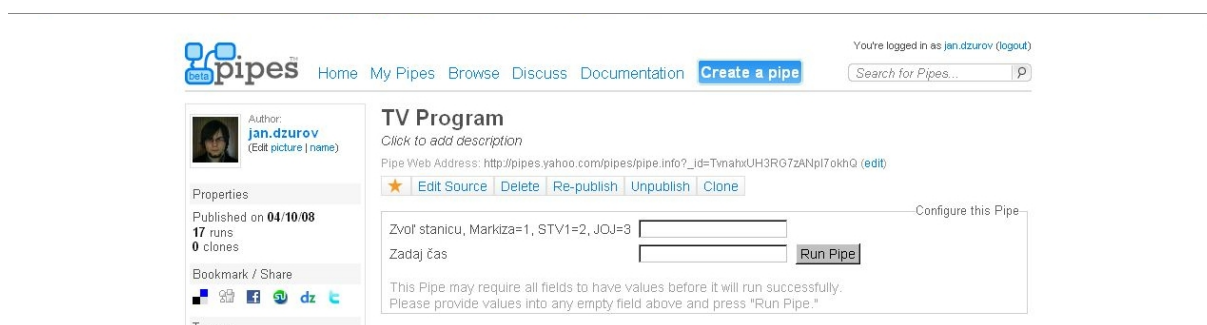
Na to aby užívateľ mohol tento údaj zvoliť, musel som do pipu priviesť ďalší užívateľský vstup. “Prompt” v tomto prípade obsahuje “Zadaj čas”, tento vstup je pri zobrazení na druhom mieste, preto je “Position” na hodnote “2”. Nastavil som taktiež “Debug” aby som mohol vidieť nejaké výsledky už pri práci. Výstup z tohto bloku vstupuje do oboch pravidiel v predošlom filtri.

Pipe Output



Práca sa v *Yahoo! Pipes* ukončuje vždy týmto blokom, ktorý spustí všetky bloky v aplikácii v tom poradí, v ktorom boli graficky spojené. „Debug“ pri tomto bloku dáva taký istý výsledok ako spustenie celého pipu.

Výsledok po spustení



Po stlačení “Run Pipe” v editovacom okne sa aplikácia spustí. Výsledkom sú v tomto prípade políčka na zadávanie užívateľských premenných. Po zadaní údajov podľa požiadaviek výzvy sa vygeneruje výsledný feed.

The screenshot shows the Yahoo! Pipes web interface. At the top, there's a navigation bar with links like Home, My Pipes, Browse, Discuss, Documentation, and a 'Create a pipe' button. A user is logged in as 'jan.dzurov'. The main content area displays a pipe titled 'TV Program' by 'jan.dzurov'. It shows the pipe's address and options to edit, delete, re-publish, unpublish, or clone. Below this is a configuration section with input fields for 'Zvof stanicu, Markiza=1, STV1=2, JOJ=3' (set to 3) and 'Zadaj čas' (set to 14:00), with a 'Run Pipe' button. To the left of the main area is a sidebar with pipe properties (published on 04/10/08, 17 runs, 0 clones), tags (TV, schedule), sources (dapper.net), and modules (fetch, filter, textinput, union, strconcat, strreplace, loop, privatestring). At the bottom of the sidebar is a visual diagram of the pipe's data flow. The main area also features a 'List' button and a feed of 10 items, including 'ZloĀtiny podĀľa Mary Higgins Clarkovej: Kde si, ocko?' at 14:30, 'Auto na Āšteku' at 16:20, 'PanelĀk (56)' at 18:20, 'KRIMI NOVINY' at 19:00, 'NOVINY' at 19:30, 'Ā PORT' at 19:55, 'PanelĀk (57)' at 20:00, and 'C. S. I. : KriminĀlka Miami (20)' at 20:45.

V hornej časti stránky sú k dispozícii nástroje na ďalšiu modifikáciu. Náš *pipe* je možné publikovať, čo znamená pridanie do databázy *pipov* ku ktorým bude mať prístup každý návštevník stránok *Yahoo! Pipes*. *Pipe* je taktiež možné “klonovať”, čo v praxi znamená že si *pipe* môžeme ďalej editovať podľa našich potrieb.

Panel na ľavej strane uvádza detaily o *pipe* a jeho majiteľovi. Je tu možné pridať tkzv. tagy, ktoré opisujú vlastnosti aplikácie. Tagy sú používané pri vyhľadávaní v databáze *pipov*. Ďalej je tu stručný prehľad použitých modulov a schéma *pipe*.

Poznámka: Diakritka sa nezobrazuje správne, dalo by sa tomu predísť výmenou znakov s diakritikou za znaky bez nej, táto požiadavka však nebola súčasťou zadania práce.

2.3.2 Opis niektorých ďalších modulov v Yahoo Pipes

Nasledujúce bloky som nepoužil v mojom predchádzajúcom *Pipe*, ale pre užívateľov poskytujú mnoho možností ako svoj *Pipe* vylepšiť alebo rozšíriť.

Truncate



Blok “Truncate” limituje počet prvkov vo výslednom feede. Počet môžeme zadať manuálne alebo spojením s iným blokom.

Tail



Blok “Tail” má opačný účinok ako blok “Truncate”, čiže jeho výsledkom je posledných niekoľko prvkov feedu, toto číslo sa zadáva manuálne alebo napojením iného bloku.

Blok Yahoo Search



Pomocou tohoto bloku sa vyhľadáva zadaný reťazec na stránke www.yahoo.com a výsledok sa zobrazí vo forme RSS feedu. Umožňuje spresniť hľadanie na konkrétny URL, pomocou pridania „Site restriction“. Výsledkom budú odkazy ktoré Yahoo nájde na dotyčnej stránke.



Google Base

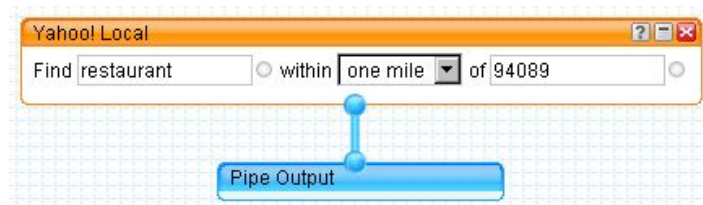


„Google Base“ je služba firmy Google na adrese <http://base.google.com/> ktorá umožňuje užívateľom propagovať svoje informácie v rôznych formátoch, vrátane PDF (.pdf), Microsoft Excel (.xls), Text (.txt), HTML (.html), Rich Text Format (.rtf), Word Perfect (.wpd), ASCII, Unicode a XML.

Využitie aplikácie je v oblasti propagácie firiem, produktov, spoločenských akcií ale aj publikácií a osobných profilov. Informácie si môže užívateľ organizovať do elementov. Blok “Google Base” v *Yahoo! Pipes* vyhľadáva v databáze *Google Base* a uľahčuje tak vyhľadávanie v tomto prostredí. Ponuka oblastí hľadania zahŕňa najpopulárnejšie oblasti využitia *Google Base*.

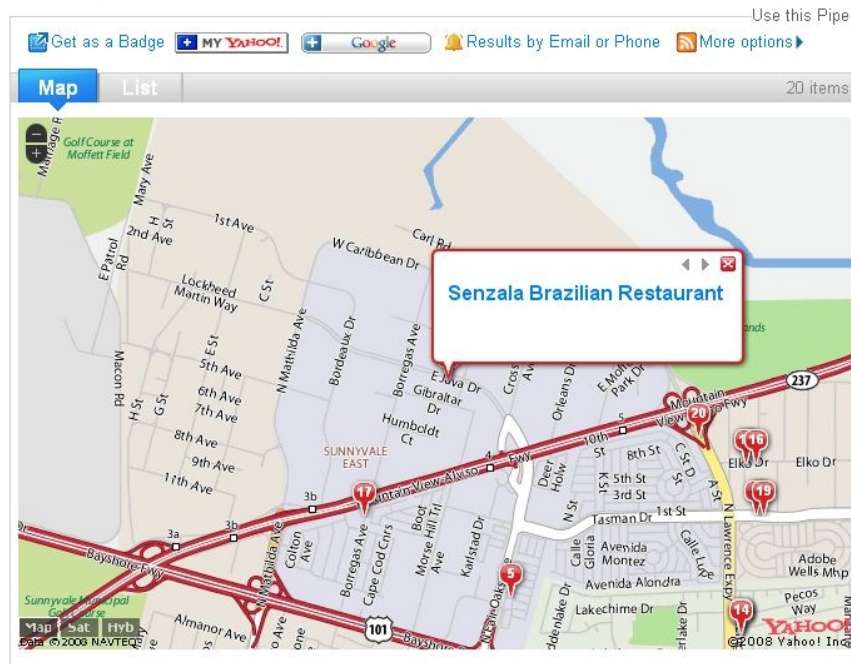
Zadáva sa kľúčové slová a lokácia. Ďalej sa určuje aj vzdialenosť od miesta určenia v míľach. Miesto sa zadáva vo forme adresy alebo ZIP kódu, ktorý je používaný v USA.

Yahoo Local

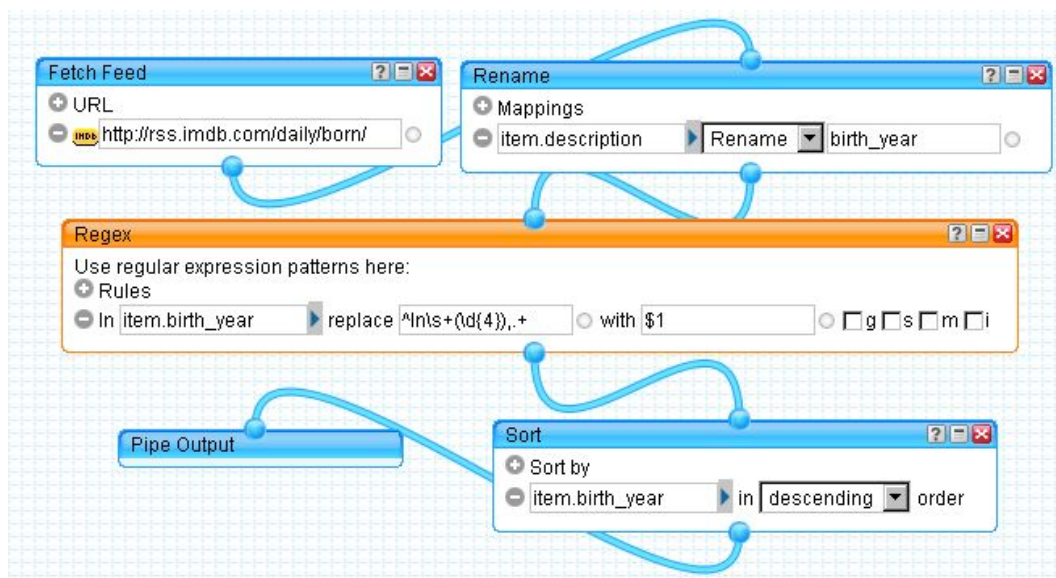


Yahoo Local je samostatná služba na stránke <http://local.yahoo.com/> ktorá ponúka vyhľadávanie rôznych služieb a inštitúcií v špecifickej lokácii. Blok “Yahoo Local” volá túto službu a zobrazuje na výstupe mapu okolia. Akceptujú sa ZIP kódy

mesta a štátu v USA. Hľadané body sú na mape zvýraznené a v okne sa zobrazuje meno miesta, firmy, atď’.



Regex



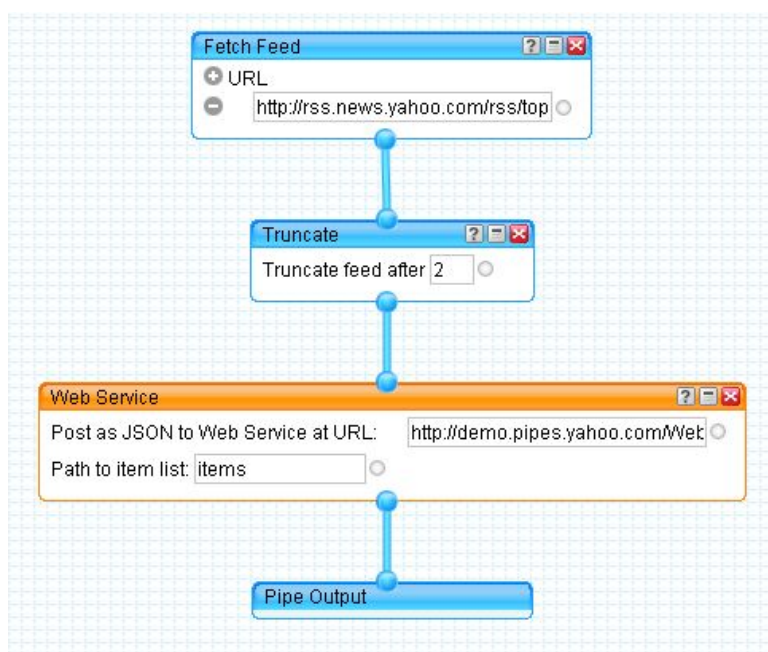
Tento modul používa “regular expressions”, (regex – v preklade pravidelné výrazy), na zmenu obsahu elementov vo feede.

Feed v tomto príklade má v elementoch <description> definované meno osoby, rok a miesto narodenia vo formáte: “In [rok], [osoba] was born on this date in [miesto]”.

Pre zprehľadnenie si tieto elementy premenujeme blokom Rename na <birthdate>. Pole replace v bloku Regex určuje segment z obsahu elementu, ktorý sa má zmeniť. V tomto prípade sa má spracovať celý reťazec. V druhom poli si stanovujeme nový reťazec ktorý nahradí predchádzajúci. \$1 je referenciou na vranaté zátvorky v regexe. Tam sa nachádza rok narodenia, a všetko ostatné sa vymaže.

Ďalej je použitý modul „Sort“ na zorganizovanie feedu podľa novovytvorených elementov.

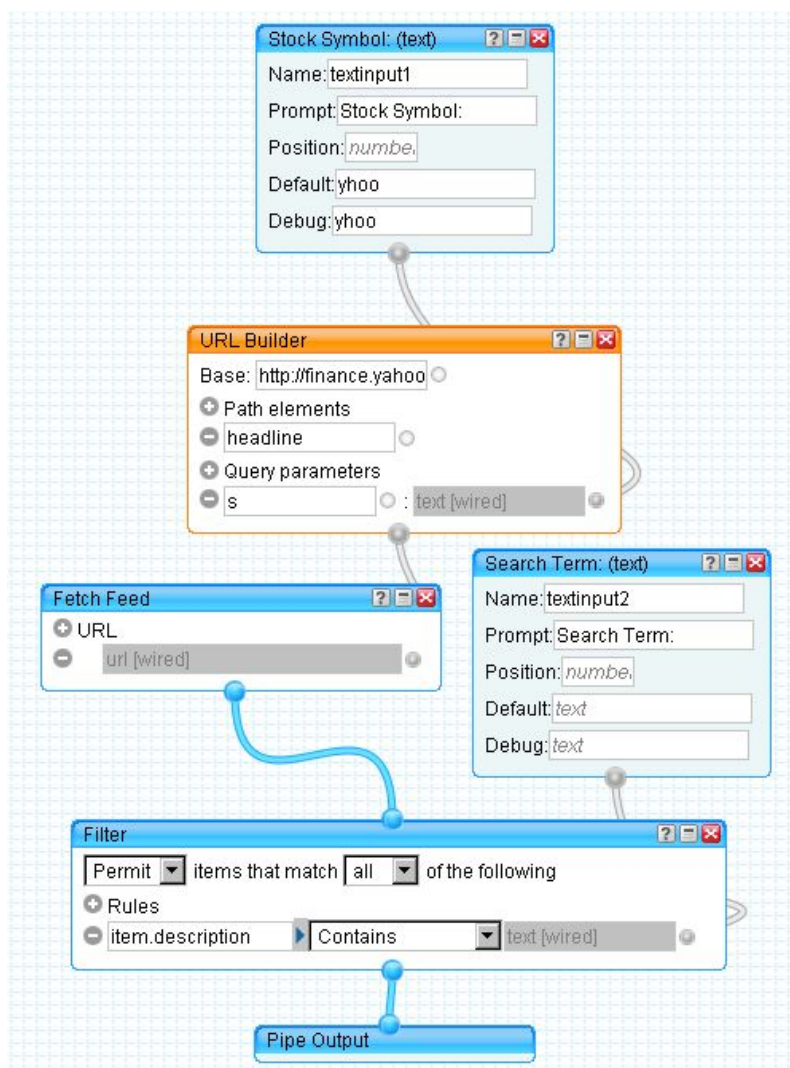
Web Service



Blok „Web Service“ umožňuje poslať feed alebo výstup z *Pipu* web službám. Informácie sú spracované do formátu JSON (JavaScript Object Notation) a tvoria vstup do skriptu služby, ktorej adresu zadáme. Výstup z *Pipu* je výstup zo skriptu a mal by byť v RSS alebo JSON formáte aby sa s ním dalo ďalej pracovať.

Tento blok poskytuje možnosť spracovať údaje takým spôsobom akým *Pipes* nedokáže, a je vhodný pre vývojárov.

URL Builder

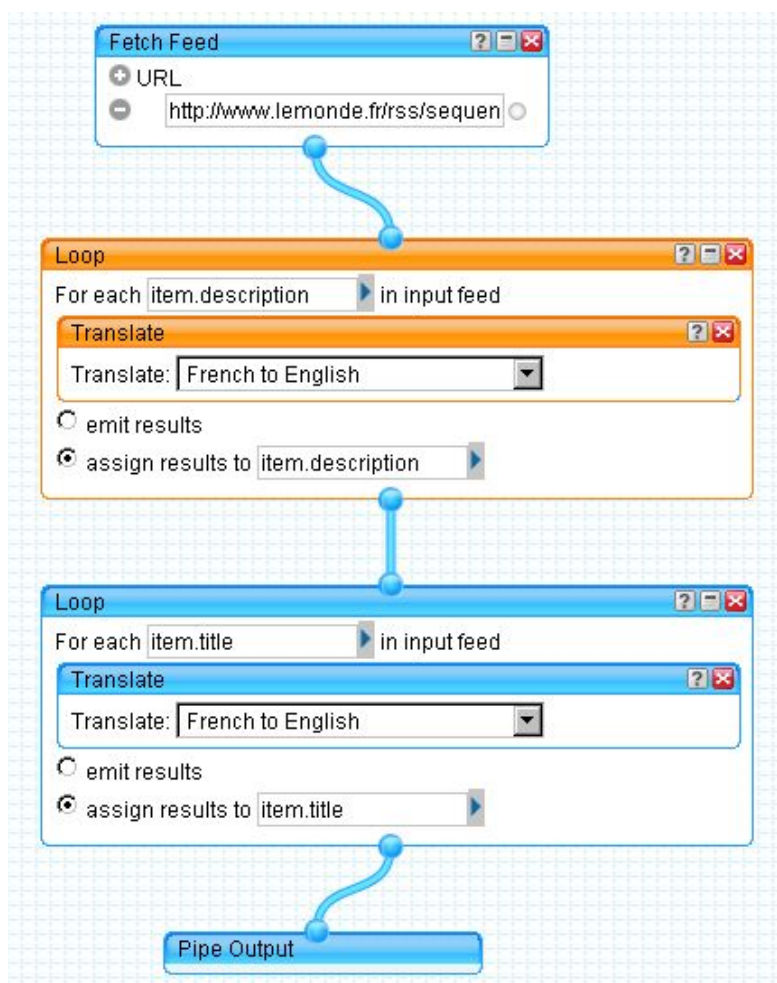


Zdroje na internete sú definované ich URL. Blok „URL builder“ umožňuje ich ľahšie zadávanie. URL majú tri hlavné časti: prvá definuje server, tu je to finance.yahoo.com, druhá cestu (path) a tretie dopyt po parametroch (query).

Základná stránka obsahuje RSS feedy na firmy podľa ich skratky. Parameter “s” definuje ktorý feed bude zobrazený.

Napr. na stránke <http://finance.yahoo.com/rss/headline?s=yhoo> sa nachádza feed pre finančné novinky firmy Yahoo. Pipe na obrázku používadva užívateľské vstupy. Prvý je na zadanie skratky firmy, ktorá sa zavedie do „URL builder“, ako parameter pre query. „URL builder“ načíta stránku ktorá je následne napojená na „Fetch Feed“, čím sa načíta RSS feed v nej. Ďalší vstup slúži na filtrovanie tak, aby vo výslednom feede boli len také prvky ktoré užívateľ chce zobrazit’.

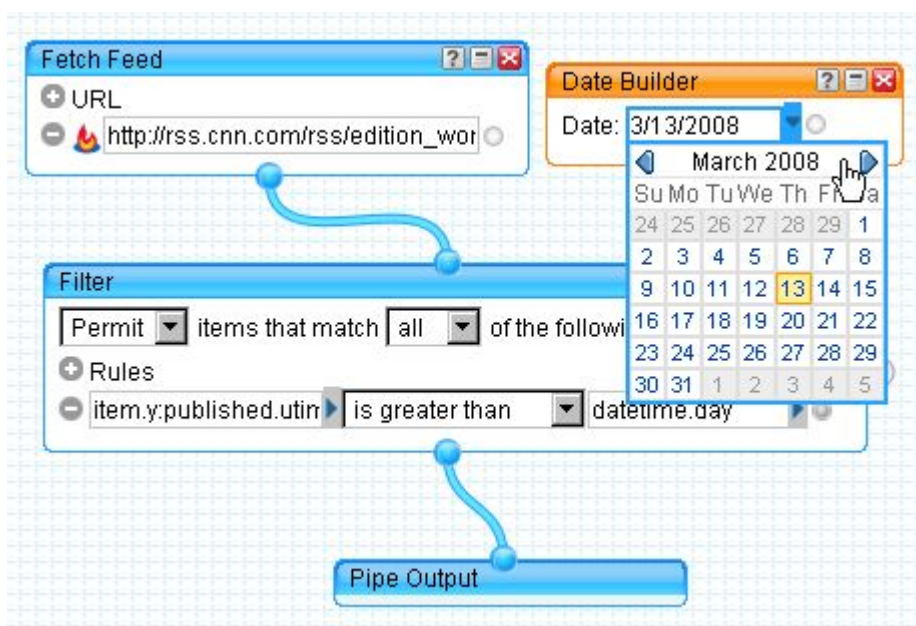
Translate



Blok "Translate" umožňuje preklad reťazcov z jedného jazyka do druhého. Ponúkané jazyky sú: angličtina, francúzština, japončina, holandčina, nemčina, gréčtina, taliančina, portugalčina, čínština, ruština, kórejščina, a španielčina.

Na príklade je ukážka tohto bloku v kombinácii s blokom „Loop“. Takto možno použiť „Translate“ na každý žiadaný element v rámci feedu. Tento blok môže byť taktiež užitočný pre preklad užívateľského vstupu.

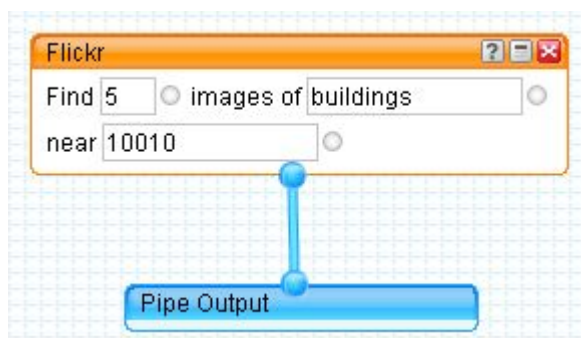
Date Builder



Jeden z najdôležitejších elementov vo feedoch zvyčajene býva dátum, kedy bol určitý prvok feedu publikovaný. “Date Builder” umožňuje vybrať dátum pomocou kalendára, ktorý sa objaví po stlačení šípky, alebo manuálnym zadánym.

Blok spoznáva väčšinu dátumových formátov, napr: “07/04/2007”, “16:32 EST”, ale aj “tomorrow”, čiže zajtra, alebo “3 days ago”, tj. pred troma dňami. Využíva sa v spojení s blokom „Filter“ ako kritérium na zobrazenie prvkov feedu. „Item.y:published.utime“ zadáva publikovaný čas, „datetime.day“ označuje deň publikácie. Takisto sa môže voliť rok, mesiac, hodina, minúta a sekunda.

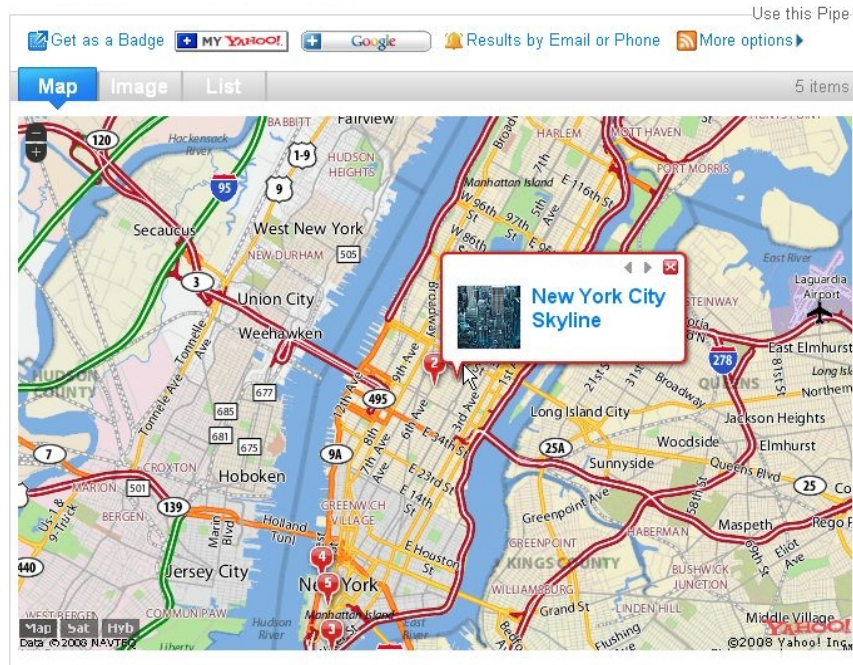
Flickr



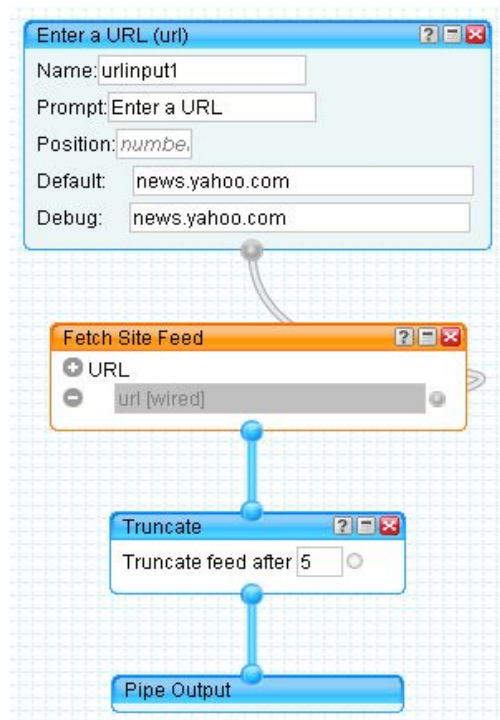
Flickr (www.flickr.com) je webová stránka na zdieľanie fotiek. Fotky načítané na stránku môžu mať takzvané “geotagy”, ktoré určujú polohu na mape. *Pipes* môže

blokom „Flickr“ vyhľadať takéto obrázky zadáním klúčového slova, počtu obrázkov ktoré sa majú zobrazit', a miestom kde sa majú fotky hľadať.

Tu tiež platí že zadávame ZIP kódy mesta v USA. Výsledkom bude buď mapa so značkami ktoré reprezentujú fotky, zoznam obrázkov alebo slideshow obrázkov.



Fetch Site Feed



“Fetch Site Feed” je modul ktorý nájde RSS feed v zadanom URL. Keď je na stránke viac feedov, nájde a zobrazí vždy prvý. Tento blok je užitočný keď chceme zistiť či existuje RSS feed pre danú stránku.

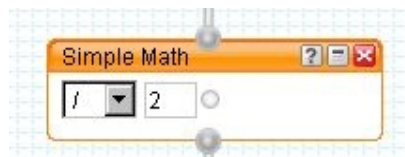
V hore uvedenom príklade je užívateľ vyzvaný na zadanie adresy a výsledok je prvých 5 prvkov z feedu ktorý je obsiahnutý v dotyčnej URL.

Split



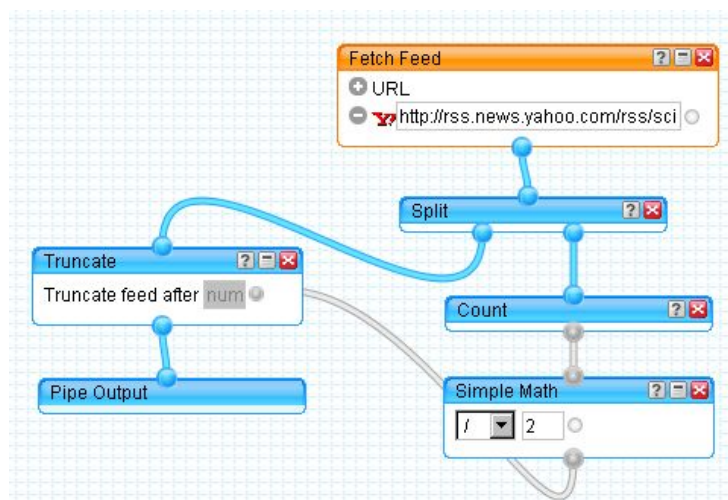
Blok “Split” rozdelí feed na dva identické feedy.

Simple Math



Tento blok umožňuje vykonávať základné algebraické operácie, ako násobenie, delenie, umocňovanie, sčítavanie a odčítavanie.

Count



“Count” je modul ktorý spočíta prvky vo feede a emituje toto číslo ako výsledok. V príklade je demonštrované jeho použitie na zníženie zobrazovaných prvkov na polovicu. Najprv pomocou bloku “Split” rozdelíme feed na dva totožné, aby sme s nimi mohli pracovať oddelene. Feed napravo prejde cez blok “Count”, čím sa získa počet prvkov v ňom. Potom sa tento počet rozdelí na polovicu napojením na blok “Simple Math”. Výsledok je polovica počtu prvkov vo feede. Ten sa načíta do bloku “Truncate” spojeného s druhým feedom vystupujúcim z bloku “Split”. Tým sa dosiahne skrátenie výsledku na polovicu.

ZÁVER

Cieľom tejto práce bolo demonštrovať tvorbu a využitie RSS prúdov pre špecifické účely. RSS sú v súčasnosti populárnym spôsobom šírenia informácií na internete a ich využitie zefektívňuje hľadanie na webe.

V práci som sa zaoberal tvorbou prúdov službou *dapper.net*, ktorá rýchlo a jednoducho umožňuje transformovať obsah stránok do formy RSS prúdov. Pre dodatočné spracovanie vytvorených prúdov som použil *Yahoo! Pipes*. Pomocou tejto aplikácie som zjednotil a prefiltroval obsah RSS prúdu tak, aby som zobrazil len tie prvky ktoré som hľadal. Aj ďalšie možnosti *Yahoo! Pipes*, ktoré som nepoužil v príklade, som tiež podrobne opísal.

RSS sa stal veľmi populárnym pre firmy, ktoré propagujú svoju činnosť na internete. Tak sa z formátu, ktorý bol pôvodne vytvorený len na šírenie správ, stáva trvalá súčasť internetu. V budúcnosti možno očakávať jeho ďalší vývoj a rozšírenie využitia.

Literatúra

- 1 - http://www.xmlfiles.com/xml/xml_intro.asp
- 2 - <http://rss.softwaregarden.com/aboutrss.html>
- 3 - <http://www.feedforall.com/rss-uses.htm>
- 4- http://www.w3schools.com/rss/rss_syntax.asp
- 5 - http://www.w3schools.com/rss/rss_readers.asp
- 6 - <http://www.dapper.net/>
- 7 - <http://pipes.yahoo.com/pipes/>