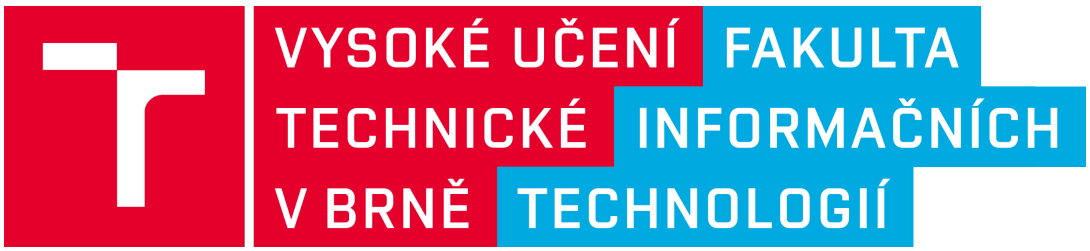
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ



Síťové aplikace a správa sítí

Čtečka novinek ve formátu Atom s podporou TLS

19. listopadu 2018

Jan Vávra (xvavra20)

Obsah

1 Úvod

Tato dokumentace se zabývá implementací a použití aplikace *feedreader*, která umí zpracovat XML feedy druhu Atom a RSS zadané URL adresou nebo souborem s URL adresami a vypsat jejich informace na standartní výstup. Aplikace podporuje šifrované nebo nešifrované stránky s feedy.

Atom a RSS jsou dokumenty XML formátu, které mají specifikcé strukturování elementů. [1] Obecně obsahují položky, které jsou dále specifikovány svými elementy (nadpis, autor, čas úpravy, apod.) Tento formát lze použít pro předávání zpráv v jednotné struktuře, kterou si uživatelé mohou stahovat a zobrazit ve čtečkách, které ho interpretují na logicky sttrukturovaný text. (webový prohlížeč, RSS feed reader).

2 Specifikace feedů

Tento dokument se zaměřuje na 3 verze feedů: RSS, RSS 1.0, která byla specifikována společností Netscape (dnes AOL), a RSS 2.0, která je specifikována RSS Advisory Board a dále Atom, který byl specifikován v RFC4287.

2.1 Atom

Každý Atom feed začíná elementem <*feed*> a atributem xmlns (namespace), ve kterém je adresa http://www.w3.org/2005/Atom. Uvnitř tohoto elementu se nachází elementy s metadaty o celém feedu a jednotlivé položky specifikované elementem <*entry*> [1][2]

Povinné elementy metadat jsou:

- <*id*> – obsahuje jedinečné URI, které identifikuje feed
- <*title*> – nadpis celého feedu
- <*updated*> – obsahuje čas poslední aktualizace feedu

Povinné elementy pro <*entry*> jsou:

- <*id*> – obsahuje jedinečné URI, které identifikuje položku
- <*title*> – nadpis pro položku
- <*updated*> – obsahuje čas poslední aktualizace položky

2.2 RSS 1.0

RSS 1.0 začíná elementem <*RDF*>, s atributy namespace, kde musí být http://www.w3.org/1999/02/22-rdf-syntax-ns# (zpětná kompabilita s RSS 0.9) a http://purl.org/rss/1.0/ (specifikující fromát RSS 1.0). Informace o celém feedu je obsažena v elementu <*channel*>, který obsahuje povinný atribut *about*, ve kterém najdeme url k feedu.[3]

Povinné elementy pro <*channel*> jsou:

- <*title*> – nadpis celého feedu
- <*link*> – URL, obsahující běžně domovskou/rodičovskou stránku feedu
- <*description*> – popis feedu
- <*items*> – seznam jednotlivých položek ve feedu, každá položka obsahuje svoje url v atributu *resource*

Jednotlivé položky začínají elementem <*item*>, který obsahuje atribut *about*, který má URL adresu na článek (stejná url se musí vyskytovat i v atributu *resource* předchozího seznamu položek)

Povinné elementy pro <*item*> jsou:

- <*title*> – nadpis pro položku
- <*link*> – URL, vstahující se na článek položky

2.3 RSS 2.0

RSS 2.0 začíná elementem <*rss*> s povinným atributem *version* (pro rss 2.0 bude roven 2.0). Element dále obshahuje jeden element <*channel*>, ve kterém se nachází popis kanálu (hlavička) a jednotlivé položky. [6] [7]
Povinné elementy pro hlavičku jsou:

- <*title*> – nadpis celého feedu
- <*link*> – URL, obsahující běžně domovskou/rodičovskou stránku feedu
- <*description*> – popis feedu

Položka feedu je v elementu <*item*> Ten musí obsahovat alespoň jednu z těchto položek:

- <*title*> – nadpis pro položku
- <*description*> – popis položky

2.4 Dublin Core

Feedy lze doplnit o namespace atributy, které rozšiřují počet elementů, které může čtečka využít. Jedním takovým rozšířením je Dublin Core.
Vybrané elementy z DublinCore. [4] [5]

- <*title*> – nadpis pro položku/feed
- <*creator*> – autor položky/feedu
- <*date*> – datum spjaté s událostí v položce
- <*source*> – odkaz na informace, odkud autor položky čerpal
- <*modified*> – čas poslední aktualizace

3 Implementace

3.1 Knihovny

Aplikace používá 2 speciální knihovny, které nejsou součástí C standardů: Openssl a Libxml2. Openssl je používána pro šifrované připojení a pro nachystání certifikačních souborů. Libxml2 je použito pro rozparsování xml souboru a jeho následné vypsání na standartní výstup. Minimální verze knihoven jsou: pro Openssl 1.0.2k, pro Libxml2 2.9.1. Tyto knihovny aktuálně běží na serveru merlin, verze na serveru eva jsou kompaktibilní. Doporučené verze jsou: pro Openssl 1.1.0h, pro Libxml2 2.9.4.

3.2 Zpracování argumentů

Seznam vstupních argumentů:

- *URL* – URL, ze kterého chceme feed číst
- *-f feedfile* – přepínač -f se specifikací cesty k souboru s URL odkazující na feedy
- *-c certfile* – přepínač -c se specifikací cesty k souboru s certifikátem, který má použít při připojení na URL
- *-C certaddr* – přepínač -C se specifikací cesty k adresáři, který obsahuje
- *-T* – přepínač -T, který zajistí vypsání času aktualizace položky, pokud je dostupný
- *-a* – přepínač -T, který zajistí vypsání autora položky, pokud je dostupný
- *-u* – přepínač -T, který zajistí vypsání URL položky, pokud je dostupný

Aplikace zpracovává argumenty zleva doprava. Alespoň jeden z parametrů *URL* a *-f feedfile* musí být na vstupu, je možné zadat oba najednou. V souboru feedfile by měli být jednotlivé adresy být oddělené buď znakem konce řádku (\n) a nebo mezerou (" "). V souboru se můžou vykytovat řádkové komentáře uvozené znakem mřížky (#), kde vše za tímto znakem je ignorováno do znaku konce řádku.

Všechny ostatní parametry jsou volitelné. Přepínače lze psát jednotlivě (*./feedreader -f file.txt -a -u*) nebo dohromady (*./feedreader -fau file*). Při zadání přepínačů dohromady, parametry, které potřebují specifikaci souboru/adresáře, si berou další argumenty v pořadí, jakém jsou zadané na vstupu (pro vstup *./feedreader -fcC file1 file2 addr1* bude přiřazeno přepínači -f file1, -c file2 a -C addr1).

Při zadání špatných argumentů se vypíše nápověda. Nápovědu lze zavolat explicitně přepínačem *-h*, který vypíše nápovědu a ukončí program.

3.3 Získání feedu

Jednotlivé feedy jsou uloženy ve frontě a zpracovány způsobem FIFO. URL se rozdělí na doménové jméno(a případný port) a cestu k souboru na serveru (pro adresu http://www.fit.vutbr.cz/news/news-rss.php rozdělení je http://www.fit.vutbr.cz a /news/news-rss.php). Podle přípony URL se rozhodne jestli jde o šifrované či nešifrované připojení (http nebo https).

Poté se pokusí poslat požadavek na server. Pokud byl úspěšný, aplikace si přečte hlavičku a zjistí jestli přijatá zpráva je OK (HTTP1.1 200 OK) a jestli přijal správný typ zprávy (Content-Type: */xml). Pokud je vše v pořádku načte zprávu do bufferu a předá parseru. Je možné, že v hlavičce najde řádek *Transfer-Encoding: chunked*, poté po načtení celé zprávy do bufferu se zbaví chunků ve zprávě. [8]

Při použití https připojení, je možné použít vlastní certifikáty specifikované přepínači *-c* a *-C*. Je ovšem potřeba předem certifikáty poslat aplikaci c_rehash pro vytvoření symbolických odkazů na certifikáty.

3.4 Výpis informací

Zpráva je předána parseru, který jí rozbije na stromovou strukturu. Pokud se v xml vyskytne chyba, pokusí se parser zotavit z chyb. Pokud je xml rozparsováno úspěšně, je vrácen kořenový uzel, ve kterém je název elementu, podle kterého aplikace určí druh feedu, podle kterého prochází stromovou strukturu.

Aplikace se v tomto stavu chová "žravě". Tím je myšleno, že se nedívá na povinné elementy nebo namespace atributy, ale snaží se najít všechny elementy napříč všemi druhy feedů (Atom, RSS 1.0, RSS 2.0), přičemž bere první údaj na který narazí (až na vyjímky viz. 3.4.1). Důvodem tohoto chování je možnost specifikovat namespace atributy, které specifikují možnost použití elementů z jiných feedů (např. Atom element <updated> v RSS feedu)

V první fázi se snaží najít elementy *title* a *item/entry*. Při nalezení *title* se vypíše na standartní výstup nadpis feedu. Při nenalezení *title* se vypíše nadpis "*** Bez názvu ***". Při nalezení *item/entry* přejde do funkce pro zpracování položky, kde se snaží najít elementy ve kterých se nachází údaje o nadpisu, autorovi, poslední aktualizaci, a URL položky. Při nenalezení nadpisu položky se vypíše "Bez názvu", při nenalezení doplňujících informací se vypíše "chybí *(doplňující informace)*". Položky jsou ukládány do fronty, pro případ, kdyby autor feedu vložil nadpis feedu někde jinde než před načtením prvního elementu *item/entry*. Pokud aplikace nenajde ani napis a zároveň ani položku, prohlásí feed na chybný a vrátí chybu.

3.4.1 Podporované elementy

- nadpis – <*title*>
- autor – <*dc:creator*>, <*author*>
- čas aktualizace – <*updated*>, <*pubDate*>, <*dc:modified*>
- URL – <*link*>, <*guid*>

Poznámka k elementům:

<*dc:creator*>, <*dc:modified*> – Tyto elementy jsou spjaty s rozšířením DublinCore. Aplikace podporuje tyto 2 elementy + <*dc:title*>, vzhledem k tomu, že obsahují informace, které hledáme (autor, čas aktualizace, nadpis) a je častým výskytem ve feedech.

<*autor*> – autor se vyskytuje jak ve Atom feedech tak RSS 2.0. Odlišují se syntaxi. Atom uvnitř elementu <author> obsahuje element <*name*> a někdy i <*email*>. Aplikace se snaží najít oba tyto elementy a dát je za sebe (Autor: *name email*). Při nenalezení elementu <autor> uvnitř entry, chová se aplikace následovně: "Pokud <entry> neobsahuje elementy <author>, poté se elementy <author> hledají pod elementem <source>. Pokud nejsou uvnitř <entry> žádné elementy <author>, použije se <author> celého feedu"(RFC4287, sekce 4.2.1)[1]. RSS 2.0 má uvitř elementu autor jméno autora.

<*link*>, <*guid*> – RSS 2.0 má potencionální 2 URL. Aplikace se rozhodne na základě atributu *isPermaLink*. Pokud je tento atribut nastaven na true vezme se URL z elementu <*guid*> [7]. Jinak načte URL z <*link*>, pokud se uvitř elementu <item> nachází.

<*link*> – u Atom feedu se může specifikovat vztahový (*rel*) atribut. Výzozí hodnota, pokud není specifikován, je *alternate*. Hodnoty atributu *rel*, které aplikaci zajímají, jsou *alternate* a *via* (prioritně se bere alternate), vzhledem k tomu že poskytují odkaz na stránku, který je popisován v <entry>. Hodnoty *related* a *enclouser* aplikace ignoruje, protože na tyto linky se článek pouze odkazuje. Hodnota *self*, je taky ignorována, protože to je odkaz na aktuální feed [1].

4 Příklady spuštění a výpisů

4.1 Spuštění s URL

```
xvavra20@merlin: ~/ISA$ ./feedreader http://www.fit.vutbr.cz/news/news-rss.php
*** FIT VUT v Brně ***
1.9.2018 Erasmus+ výjezdy a zahraniční stáže/praxe
14.9.2018 - 4.2.2019 Rozdělení do přednáškových skupin BIA a BIB na ZS 2018/2019
27.9-21.11.2018 Mimořádné stipendium pro 500 studentů 1. ročníků BC s nejlepší maturitou
31.10.2018 Rezevace seminárních místností v knihovně FIT
2.11.2018 Údržba knihovního fondu v knihovně FIT
2.11.2018 Školení z vyhlášky č. 50 (§5)
15.11.2018 ERASMUS+ 2019/2020 - BESEDA
14.12.2018 - 1.2.2019 Dny otevřených dveří na FIT
2.1-1.2.2019 Přehled vybraných zkoušek po ZIMNÍM semestru
```

4.2 Spuštění s URL a dodatečnými informacemi

```
xvavra20@merlin: ~/ISA$ ./feedreader https://tools.ietf.org/dailydose/dailydose_atom.xml -a -T -u
*** The Daily Dose of IETF ***
The Daily Dose of IETF - Issue 3235 - 2018-11-16
Autor: chybí autor
Aktualizace: 2018-11-16T05:00:15Z
URL: https://tools.ietf.org/dailydose/3235.html

The Daily Dose of IETF - Issue 3234 - 2018-11-15
Autor: chybí autor
Aktualizace: 2018-11-15T05:00:15Z
URL: https://tools.ietf.org/dailydose/3234.html

The Daily Dose of IETF - Issue 3233 - 2018-11-14
Autor: chybí autor
Aktualizace: 2018-11-14T05:00:18Z
URL: https://tools.ietf.org/dailydose/3233.html

The Daily Dose of IETF - Issue 3232 - 2018-11-13
Autor: chybí autor
Aktualizace: 2018-11-13T05:00:17Z
URL: https://tools.ietf.org/dailydose/3232.html
```

4.3 Spuštění se souborem s https URL adresami a specifikací certifikátu

```
xvavra20@merlin: ~/ISA$ ./feedreader -f https.txt -c /etc/ssl/certs/ca-bundle.crt
*** The Register - Hardware ***
IDC: Big data biz worth $16.9 BILLION by 2015
Boffins build blood-swimming medical microbot
Dell mothership hovers over backup startup, beams it aboard
Dell Latitude E6220 12.5in Core i7 notebook

*** xkcd.com ***
Evaluating Tech Things
Indirect Detection
Trig Identities
Wishlist

*** What If? ***
Earth-Moon Fire Pole
Electrofishing for Whales
Toaster vs. Freezer
Coast-to-Coast Coasting
Hide the Atmosphere
```

4.4 Spuštění se souborem s https URL adresami a specifikací nevalidního certifikátu

```
xvavra20@merlin: ~/ISA$ ./feedreader -f https.txt -c /dev/null
Chyba: nepodařilo se ověřit platnost certifikátu serveru www.theregister.co.uk

Chyba: nepodařilo se ověřit platnost certifikátu serveru xkcd.com

Chyba: nepodařilo se ověřit platnost certifikátu serveru what-if.xkcd.com
```

5 Literatura

[1] Nottingham, M., Ed., and R. Sayre, Ed.,*The Atom Syndication Format*, RFC 4287 [online]., Prosinec 2005,[cit. 2018-11-19]. , DOI 10.17487/RFC4287. Dostupné z: https://tools.ietf.org/html/rfc4287

[2] *FEED Validator Introduction to Atom* [online]. [cit. 2018-10-27]. Dostupné z: https://tools.ietf.org/html/rfc4287

[3] *RDF Site Summary (RSS) 1.0* [online]. [cit. 2018-10-27]. Dostupné z: http://web.resource.org/rss/1.0/spec

[4] *RDF Site Summary 1.0 Modules: Dublin Core* [online]. [cit. 2018-10-27]. Dostupné z: http://web.resource.org/rss/1.0/modules/dc/

[5] *Dublin Core* [online]. [cit. 2018-10-27]. Dostupné z: https://feedforall.com/dublin-core.htm

[6] *RSS 2.0 Specification* [online]. [cit. 2018-10-27]. Dostupné z: http://www.rssboard.org/rss-specification

[7] *FEED Validator RSS 2.0 specification* [online]. [cit. 2018-11-19]. Dostupné z: https://validator.w3.org/feed/docs/rss2.html

[8] *Transfer-Encoding* [online]. [cit. 2018-10-27]. Dostupné z: https://developer.mozilla.org/en-US/docs/Web/HTTP/Headers/Transfer-Encoding