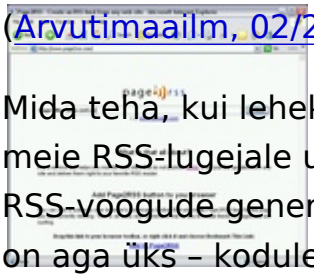
Teeme ise RSSi

18 aastat tagasi - 27.02.2007 Autor: AM



(Arvutimaailm, 02/2007)

Mida teha, kui lehekülg, mille uuendamisega tahame olla kindlalt kursis, ei paku meie RSS-lugejale uudistevoogu? Aga sel juhul teeme sellele lehele RSS-voo ise. RSS-voogude genereerijaid veebilehtedele leiab lausa mitu. Nende tööpõhimõte on aga üks – kodulehel, kus puudub RSS-uudisvoo väljund, analüüsitakse HTML-i ja eemaldatakse sealt kõik mittetekstiline info. See pannakse omakorda ümber XML-kujule ja kui selles tekstilises osas toimub muutusi, ongi uus artikkel uudisvoos sündinud.

Selline lihtne mootor töötab mõnede veebilehtedega väga hästi, mõnedega aga üldse mitte. Näiteks kui lehel uueneb pidevalt midagi, siis hakkab RSS-i genereerija spämmima tähtsusetute tehniliste muutuste uudistega. Kui aga koduleht püsib kaua ühesugune ja seda uuendatakse harva, saab RSS-i genereerijaga vajaliku teate operatiivselt otse oma uudisvoogude lugejasse, ilma et kodulehte peaks uudisteootuses kogu aeg vaatamas käima.

Kasutada saab selleks näiteks teenust Page2RSS ehk „leheküljest RSSiks” (www.page2rss.com). Nimi iseloomustab seda teenust täpselt: Google´i laadsele lehele tuleb sisestada koduleht, mis vaatamata RSSi ajastule pole viitsinud omal seda lisavõimalust teha ja hetke pärast genereeritakse sellel lehel toimuvatest muutustest juba RSS-voog.

Kuna robot, mis RSSi jaoks vajalikku infot veebilehelt kogub, ei hiilga just suure intelligentsiga, siis sellist kunstlikult loodud RSSi saab pruukida vaid teatud tüüpi veebilehtede puhul. Esiteks peab see leht olema piisavalt staatiline ehk muutusi peab seal toimuma küllalt harva. Teiseks peab koduleht olema ka üsna lihtne, ilma freimide ja tihti muutuvate tehniliste osadeta (näiteks kasutajate sisselogimisinfo vms). Kuid tavaliselt ongi staatilised, harva muutuvad lihtsad lehed need, kuhu ei viitsita ka RSSi teha. Selliste suhteliselt lihtsate lehtede jaoks Page2RSS ongi.
Veidi veebis ringi kolades selgus, et näiteks Eesti päevalehtedest Page2RSS ei suuda RSS-voogu adekvaatselt genereerida. See-eest saab ta väga hästi hakkama CNNi uudistevoo loomisega.

Mis on RSS?
RSS on uudistevoo formaat, mille abil saab edastada tihedalt muutuvate veebilehekülgede uuendusi, näiteks blogides või uudistelehekülgedel olevaid uusi sissekandeid. RSSi loetakse spetsiaalsete uudisteagregaatoritega, mis käivad soovitud lehekülgedel kontrollimas, kas on tekkinud uusi sissekandeid. (Wikipedia)

Kuid ega Page2RSS pole ainuke seda liiki teenus, samamoodi teeb linkidest uudisvoogusid ka Web2RSS teenus (http://www.baekdal.com/web2rss/). Sellegi teenuse puhul saadetakse RSS-voona teade viimastest muutustest lehel. Erinevalt Page2RSSist püsib see teenus stabiilsem ja leiab ka rohkem uudiseid lehelt. Kuid kõik sõltub jällegi konkreetsest leheküljest. Mugav on sellise teenusega tekitada omale RSS voog näiteks mõnest otsingutulemusest. Niipea, kui otsitava märksõna kohta mõni otsimootor miskit uut leiab, saab sellest ka teate RSS-lugejasse.

Web2RSS-i suurepärane omadus on ka otsida muutusi HTMLi seest ja muu lehekülg tähelepanu alt välja jätta. Nii võib jälgida vaid seda osa leheküljest, kus tõesti on olulised muutused ja kõik muu kõrvale jätta. Näiteks võib lisavõimaluste lehel Match lahtrisse HTML-i jupi, mille muutust vaja kindlasti kontrollida ja exclude lahtrisse jupi, mida ignoreerida. Tekst "<table id="myygitulemused">.+?</table> näiteks lahtris „match” tähendab, et RSS-voogu tulevad uudised vaid siis, kui toimub mingi muutus lehel olevas „myygitulemused” nimelises tabelis.

Kaido Einama

- Lahendused