

TEU00311

What is the Internet doing to me?
(witidtm)

Stephen Farrell
stephen.farrell@cs.tcd.ie

<https://github.com/sftcd/witidtm>
<https://down.dsg.cs.tcd.ie/witidtm>

Reading news

- We probably all have preferred sources of news
- What are yours?
- Why do you choose them?
- Do you treat all sources equally? (as a reader)
- What kind of media do you prefer (text, audio, video), for what and why?
- What kinds of publishers/aggregators are there?

How I once read tech news

- Back in pre-history (from 1980) there was usenet/netnews
 - <https://en.wikipedia.org/wiki/Usenet>
- That was (and still is!) a federated set of servers that distribute messages from named newsgroups to one another using the Network News Transfer Protocol (NNTP)
- Anyone could write a message to most groups, some were moderated
- comp.risks or “The Risks Digest” was one such newsgroup
- AFAIK, that’s where you can still see the oldest text I wrote online:
 - <https://catless.ncl.ac.uk/Risks/8.38.html>
 - That’s over there ----->
 - Obviously, a number of posters said I was stupid in followups:-)
- Amazingly the risks digest is still on the go!
 - <https://catless.ncl.ac.uk/Risks/>
 - I didn’t realise that ‘till I made this slide;-)

✂ Toshiba DOS 3.3 Backup deletes files

Fiona M Williams <fiona@euroies.ucd.ie>
Tue, 14 Mar 89 14:34:50 GMT

A colleague of mine had just started to backup the hard disk of his Toshiba 3200 using the Toshiba DOS 3.3 backup command. While backup was still looking at the root directory we had a power failure in the office. A couple of gnashes later he re-booted the T3200 only to get the message “Bad or missing command interpreter.” (This generally means that command.com has been knackered.) Also, when we looked at the backup diskette, there was nothing on it!

Having (eventually) found a Toshiba DOS 3.3 diskette we managed to have a look at the hard disk only to find that all files in the root directory *had been deleted*. (Sub-directories were ok though.) Norton’s quick un-erase came to the rescue so we managed to recover everything after about an hour.

I’d hate to think what might have happened if we’d had the power failure when backup was on its 20th diskette, rather than its first, but in any case, the moral seems to be that you should sometimes make a backup before making a backup!

Stephen Farrell, MANTIS LTD. (stephen_farrell_mantis@eurokom.ucd.ie)

How I read news a decade ago

- From ~early 2000's I had a small, slowly changing, collection of browser bookmarks for ~7-8 news sites that I'd regularly visit
- I organised those into a folder of bookmarks
- At some stage Firefox added a feature to “open all in tabs” so you could load all those pages, each on its own tab, with one click
 - That feature is still there but a bit more hidden in latest FF

FF news bookmarks circa. 2015

- I bookmarked these news sites:
 - Irish Times
 - RTE
 - /.
 - Washington post
 - BBC
 - The Intercept
 - The Guardian
 - Breakingnews.ie
- All was good but that didn't work on my phone
 - That was an issue from say ~2008 once I had a well-connected phone
 - An RSS reader did work there though

RSS Readers



- RSS = Real Simple Syndication
 - <https://en.wikipedia.org/wiki/RSS> seems accurate enough
- Designed (in 1999) to provide a simple overview of a web page (plus a link) and to be used in a “feed” so that an RSS reader can track updates as web pages change and new pages are added
- RSS is (I think) Ideal for readers of news web sites and aggregators
- Not ideal for publishers: No advertising. No analytics.
- RSS support now removed from current desktop browsers
- Some mail user agents (e.g. Thunderbird) still support RSS

RSS “item” example

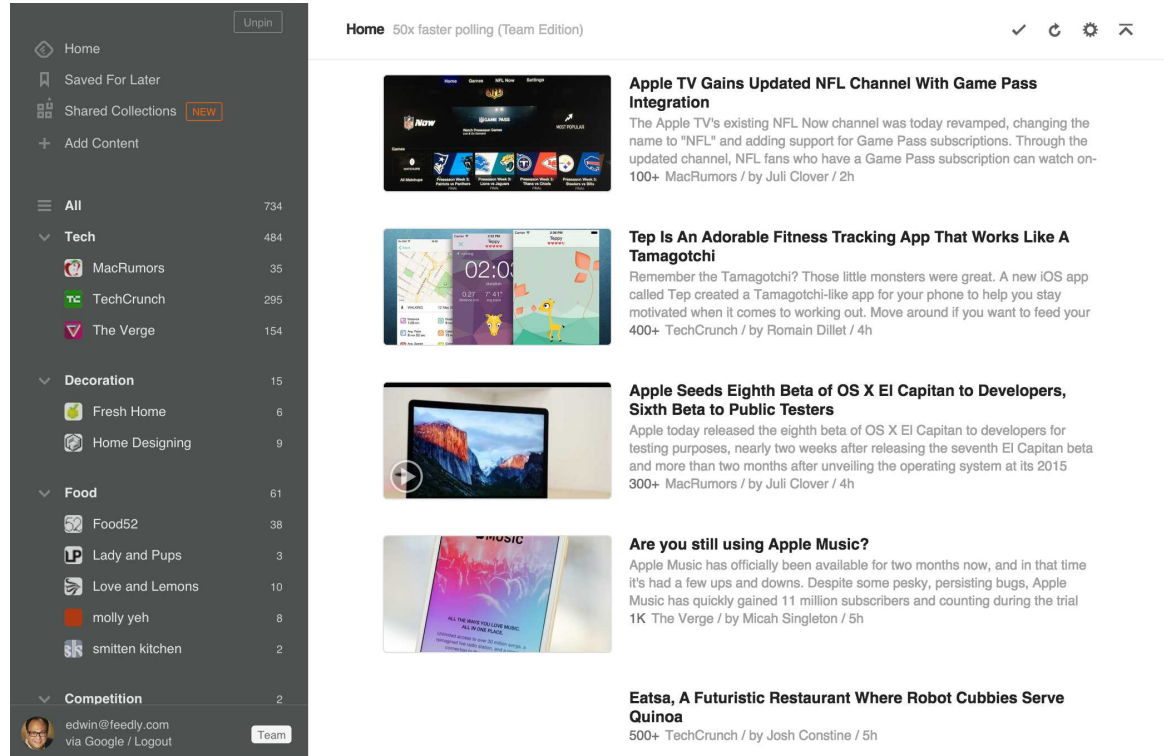
From: <https://www.rte.ie/news/rss/news-headlines.xml>

```
<item>
  <title>Taoiseach awaits outcome of 'volatile' week in UK</title>
  <link>http://www.rte.ie/news/brexit/2019/0902/1073287-brexit-eu/</link>
  <description>
    Taoiseach Leo Varadkar has said he may meet British Prime Minister Boris Johnson next week and says he will of course listen
    to alternatives to the backstop.
  </description>
  <pubDate>Mon, 02 Sep 2019 11:50:21 +0000</pubDate>
  <guid>http://www.rte.ie/news/brexit/2019/0902/1073287-brexit-eu/</guid>
  <category>Brexit</category>
  <media:content url="https://img.rasset.ie/00129ec1-800.jpg" width="800" type="image/jpeg" height="450"/>
</item>
```

The syntax used there is XML (eXtensible Markup Language)

- XML is (sort-of) a superset of HTML
- An RSS “feed” is just a set of items plus some wrapping stuff
- An RSS reader has URLs like the above configured and then renders the items

How an RSS reader looks



<https://feedly.com/> - Note: I don't use this - It might be great, or terrible, I've no idea. (I don't like that it needs Javascript though:-)

Next step was dissatisfaction;-)

- So, for quite a while, (maybe 2010-2016), I was fine with browser bookmarks and an RSS reader on my phone
 - More-or-less the same sources though not identical
- Eventually though, the amount of web crap (ads, JS, “accept cookie”) got annoying
 - And Snowden’s news told us lots more was being surveilled than we had thought (more on that later)
- About 2017 time I setup jell.ie as my own server
- Which meant I could control more things, but what to do?
 - First: measure

2017 Measurements (1)

- 8 web sites measured
- Just re-checked those URLs (20190908):

	A	B	C	D	E
1	Outlet	Web site	2019 State	RSS feed	2019 State
2	Irish Times	https://irishtimes.com/	ok	https://www.irishtimes.com/cmlink/news-1.1319	ok
3	The Intercept	https://theintercept.com/	ok	https://theintercept.com/feed/?lang=en	ok
4	Washington Post	https://www.washingtonpost.com/	GDPR'd	http://feeds.washingtonpost.com/rss/rss_blogp	Broken TLS + wall of text!
5	Engadget	https://www.engadget.com/	GDPR'd	https://www.engadget.com/rss.xml	ok
6	Guardian	https://www.theguardian.com/	ok	https://www.theguardian.com/world/rss	ok
7	BBC	http://www.bbc.com/	ok	https://feeds.bbc.co.uk/news/rss.xml?edition=ir	wall of text
8	Slashdot	https://slashdot.org/	ok	http://rss.slashdot.org/Slashdot/slashdotMain	no TLS but nice formatting
9	El Reg	https://www.theregister.co.uk/	ok	https://www.theregister.co.uk/headlines.atom	probably ok

2017 Measurements (2)

- As we saw already 'shift-ctrl-l': let's you see what's up as a page loads
- Another thing you can do in developer-mode is save a "har" file – an HTTP archive file that stores details of most of the interactions you can see in developer-mode
 - HOWO: in browser, shift-ctrl-l; choose network tab; reload page; right-click, select "save as har" (FF) or "save as har with content" (chromium/opera)
 - HAR files are large JSON things
 - Careful to clear browser cache before measuring!
- (With a bit of scripting) that allowed me to count attempts to set cookies and how many hosts were being contacted for each page load
- One could automate all that, but I didn't, I tested 2 configurations:
 - Out-of-the-box chromium
 - FF+NoScript+Ghostery+cookies-off

2017 Measurements (3)

	A	B	C	D	E	F	G
1	Browser	Out-of-the-box chromium					
2	Site	MB	Time (s)	Set-cookie	Hosts	HTTP Requests	Data-URLs
3	https://www.irishtimes.com/	3.3	14.43	147	44	253	7
4	https://theintercept.com/	4.4	4.91	22	4	26	0
5	https://www.washingtonpost.com/	2.8	12.41	369	43	343	4
6	https://www.engadget.com/	4.8	16.42	2	19	341	1
7	https://www.theguardian.com/	2.4	5.69	35	31	177	5
8	http://www.bbc.com/	1.2	6.16	19	26	193	1
9	https://slashdot.org/	1.1	10.7	209	56	235	17
10	https://www.theregister.co.uk/	0.8	5.09	84	22	100	3
11	Totals	20.8	75.81	887	245	1668	38
12							

2017 Measurements (4)

	A	B	C	D	E	F	G	
13	Browser	FF+NoScript+Ghostery+no-cookies						
14	Site	MB	Time (s)	Set-cookie	Hosts	HTTP Requests	Data-URLs	
15	https://www.irishtimes.com/	1.8	7.28	0	4	97	0	
16	https://theintercept.com/	3.2	3.44	0	3	11	0	
17	https://www.washingtonpost.com/	2.3	4.12	0	3	23	3	
18	https://www.engadget.com/	3.7	11.11	0	4	29	0	
19	https://www.theguardian.com/	2.5	5.85	0	6	59	0	
20	http://www.bbc.com/	0.5	3.95	0	8	28	0	
21	https://slashdot.org/	0.5	3.62	0	3	20	0	
22	https://www.theregister.co.uk/	0.5	2.3	0	3	42	0	
23	Totals	15	41.67	0	34	309	3	
24								

2017/2019 Measurements

	A	B	C	D	E	F	G
25	Jell.ie News (2017)	MB	Time (s)	Set-cookie	Hosts	HTTP Requests	Data-URLs
26	Out-of-the-box chromium	30.7	17.11	0	1	62	0
27	FF+NoScript+Ghostery+no-cookies	31.8	16.21	0	1	61	0
28							
29	Jell.ie News (2019)	MB	Time (s)	Set-cookie	Hosts	HTTP Requests	Data-URLs
30	Out-of-the-box chromium	0.14	1.37	0	1	5	0
31	FF+NoScript+Ghostery+no-cookies	0.22	1.31	0	1	5	0
32	Straight browser totals						
33	Out-of-the-box chromium	20.8	75.81	887	245	1668	38
34	FF+NoScript+Ghostery+no-cookies	15	41.67	0	34	309	3
35							

The jell.ie news

- Service: <https://jell.ie/news> you're welcome to use if you want, even better if you re-produce based on your own choices!
- Text description of why and what:
 - <https://down.dsg.cs.tcd.ie/witidtm/refs/jell.ie-news.html>
 - Includes measurement details given above
- We'll walk through it...

Jell.ie news == VPS+Apache2+simplepie

- Jell.ie is a VM/VPS I rent from a hoster running Ubuntu 16.04 with Apache2
 - That's totally run of the mill
- Given I have a web server and a set of news sources I'd like to read, why not have the web server contact the various sources, cache the content and then display that to me?
- Checked who'd done this before and found simplepie, a PHP script for just that
 - <https://github.com/simplepie/simplepie/>
 - Tweaked that to better do what I want, mostly forget what I changed
- Default for simplepie was to include links to images on the source site, but that'd mean my IP connecting to the source, so I turned on caching of images on my VPS as well
 - But then I turned of images entirely they don't add much

Example of what control means

- I got fed up reading news stories featuring the name of the current US president
- So on each page load, that string is replaced by a randomly selected Dutch name:-)
 - <https://randomuser.me/api/?nat=nl>
 - Dutch names are funnier somehow

What's the difference in privacy footprint between the jell.ie news and the status quo ante?