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/

⚡ **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
  - e.g. firefox https://support.mozilla.org/en-US/kb/feed-reader-replacements-firefox
- Some mail user agents (e.g. Thunderbird) still support RSS
  - chrome 94 recently re-enabled some RSS features
    - https://www.androidheadlines.com/2021/10/follow-sites-on-chrome-rss-powered-feed.html

RSS "item" example

From: https://www.rte.ie/news/rss/news-headlines.xml (on the day I made this slide:-)

```
<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/looked
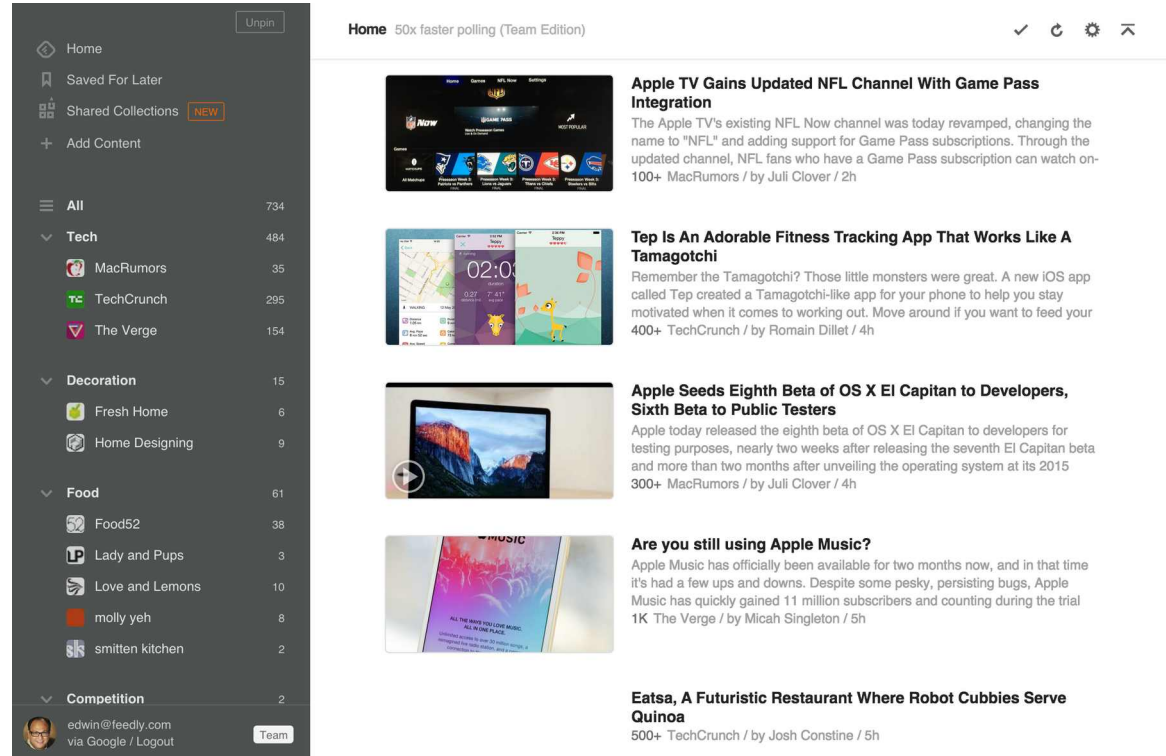
https://feedly.com/   - Note: I don't use this - It might be great, or terrible, I've no idea. (I don't like that it needs Javascript though:-) They've also gone all "AI" smart since I made this slide

Next step was dissatisfaction;-)

- 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 that we had thought (more on that later)
- In 2015 I setup jell.ie as my own server
  - jell.ie runs on a virtual private server (VPS) in a hosting site operated by a small hoster based in Amsterdam (used to be on a VPS run by a chap in Sligo, but I moved it)
  - Sometimes people call VPS' virtual machines (VM) – kinda the same thing mostly but not always
  - One server (basically a rack-mounted PC) can run multiple VMs for different customers
- Having jell.ie meant I could control more things, but what to do?
  - First: measure, then see what I want to control

2017 Measurements (1)

- 8 web sites measured that had corresponding RSS feeds

- Re-checked those URLs on 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_blogpo	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.bbci.co.uk/news/rss.xml?edition=in	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	probablly ok

2017 Measurements (2)

- As we saw already 'shift-ctrl-I': 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 a browser's developer-mode
  - HOWTO: in browser, shift-ctrl-I; 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
    - JSON is JavaScript Obect Notation - a way to represent structured data that's currently fashionable
  - Be 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

2017 Firefox setup

- Turn off cookies in preferences, except for an allowlist of sites I want to let use cookies (about 10)
  - That means only those 10 sites can set cookies that FF will re-send to the site
- NoScript is an add-on that turns off Javascript for all but a set of allowlisted web sites
  - FF won't execute Javascript code embedded into a non-allowlisted web page
  - Default config allows "well known" things like gmail (maybe 100 sites) but I turn all those off and only allow those I want/need (also about 10, about 8 of which are also allowed cookies)
- Ghostery is an ad/tracker blocker add-on
  - Has a blocklist of sites that are trackers (about 3000! and growing!), default is to allow a bunch of things (same as NoScript) but I turn 'em all off
  - HTTP requests to those trackers that would otherwise happen automatically (e.g. 1x1 pixel images) are intercepted and not sent
- "HTTPS Everywhere" add-on
  - Has an allow-list of domains so that whenever the domain in an HTTP schemed request is on the allowlist the URL is automatically changed to an HTTPS schemed URL (i.e. it automates turning on encryption); Uninstalled when broke at some stage;
  - Modern browsers have that as a built-in setting now, and it is, or will become, the default
  - I'm currently still running this but EFF (the authors) have declared victory and will deprecate it soon
    - https://www.eff.org/deeplinks/2021/09/https-actually-everywhere

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							

Frequently leaving breadcrumbs at 245 hosts, many of whom are purely there for tracking, would be horrendous, and even 34 hosts seems ickky...

...so I made https://jell.ie/news/

jell.ie News

Read at: 2019-09-21T13:04:46+01:00

How the farmers are being talked down

The latest moves by some of the protesting farmers to abandon the blockades seem to have been agreed upon after local negotiations with factory owners and political representatives in some areas.

Source: News Headlines | 21 Sep 2019 | 12:50 pm IST

New Zealand 23-13 South Africa: Defending champions hold off spirited Springboks

Defending champions New Zealand produce a clinical display to hold off a spirited South Africa in their World Cup Pool B opener.

Source: BBC News - Home | 21 Sep 2019 | 12:38 pm IST

Trump's week of dithering over Iran makes America look weak and foolish

The president's misconceived Middle East policy has been laid bare, and few allies will rush to the rescue

They must be laughing their socks off in Tehran. The days following last weekend's attacks on Saudi oil

Example of what control means

- I got fed up seeing headlines featuring the name of a certain US president
- Each time I generate a page, a bit of PHP code I added to my web page replaces that string with a randomly selected Dutch name:-)
    - https://randomuser.me/api/?nat=nl is a site that gives random user information (names, d-o-b, logins, emails) designed to look credible – it's used for evasion (creating phoney accounts) but more legitimately for software testing
    - It supports about 16 nationalities but Dutch names were funnier somehow
- I don't actually call the API each time any more (that'd be another breadcrumb!) but have downloaded a snapshot of their Dutch names and randomly pick first and last names from that locally
    - https://github.com/RandomAPI/Randomuser.me-Node

jell.ie News

Read at: 2019-09-21T13:06:11+01:00 (US Pres==Sharelle Van Den Nieuwendijk)

How the farmers are being talked down

The latest moves by some of the protesting farmers to abandon the blockades seem to have been agreed upon after local negotiations with factory owners and political representatives in some areas.

Source: News Headlines | 21 Sep 2019 | 12:50 pm IST

New Zealand 23-13 South Africa: Defending champions hold off spirited Springboks

Defending champions New Zealand produce a clinical display to hold off a spirited South Africa in their World Cup Pool B opener.

Source: BBC News - Home | 21 Sep 2019 | 12:38 pm IST

Sharelle Van Den Nieuwendijk 's week of dithering over Iran makes America look weak and foolish

The president's misconceived Middle East policy has been laid bare, and few allies will rush to the rescue

They must be laughing their socks off in Tehran. The days following last weekend's attacks on Saudi oil

The jell.ie news

- Service is https://jell.ie/news and you're welcome to read it if you want
    - Even better if you make your own service based on your own choices!
- Text description of why and what:
    - https://down.dsg.cs.tcd.ie/witidtm/refs/jell.ie-news.html
    - https://github.com/sftcd/witidtm/blob/master/refs/jell.ie-news.md
        - Includes measurement details given above
- Does what I want, on laptops, tablets and phones
    - We'll walk through it...

Jell.ie news == VPS+Apache2+simplepie

- Jell.ie is a VM/VPS I rent from a hoster running Ubuntu 22.04.1 LTS, with Apache2
  - Totally run of the mill, cost is <1 coffee/week
  - Learning HOWTO cost is not too high, mostly reading
- Given I have a web server and a set of news sources I'd like to read, why not have my web server contact the various news sources, cache the content and then display that to me without me leaving breadcrumbs all over?
- Checked who'd done what before and found simplepie, a PHP script for just that
  - The VPS (not my browser) updates the RSS feeds and caches them
  - https://github.com/simplepie/simplepie/
  - Tweaked that to better do what I want, mostly forget what I changed

LAMP Server

- LAMP server is usually a VM/VPS with Linux Apache MySql and PHP
  - https://en.wikipedia.org/wiki/LAMP_%28software_bundle%29
  - That or similar are commonly setup with nearly one click at a hoster
- PHP is a recursive acronym for PHP: Hypertext Preprocessor (geeks, eh;-)
  - https://www.php.net/
  - PHP is a simple scripting language that runs inside the web server in order to generate tailored HTML pages
- Javascript (in contrast) is a scripting language such that the web server gives the browser the (possibly complex) script which is then run by the browser
- Many purist programmers hate bits or all of both PHP and Javascript
  - Others don't:-)

Jell.ie news == VPS+Apache2+simplepie

- Instead of my browser talking to many web sites and trackers, it talks to my VPS – jell.ie

- Simplepie PHP code, running inside the https://jell.ie web server acts as an RSS client pulling the RSS feeds and turning those into ok-to-nice HTML

- So my browser doesn't interact with the source web sites, nor trackers

- Default for simplepie includes links to feed-item images on the source site, but that'd mean my IP connecting to the image source, so I turned on caching of images and mapping of img URLs as well

- But then I turned off images entirely as they don't add much

Jell.ie news == VPS+Apache2+simplepie

- Simplepie has a cache, (I set it to 1 hour) so that each load of the page does not result in a new fetch from the source sites – use that so I don't expose timing/presence information to the web sites

- I have a "cron" job that also runs on the jell.ie VPS that refreshes that cache every hour
  - Last I looked (2019), bingbot was readng the jell.ie news about every 30 mins

- So the source sites just see the VPS once per hour and never see timing related to my browser accessing the news (nor yours!)

- Until I click through to read the detail of a story when it's back to "normal" web access

- One recent (~Aug 2022) change: the US NPR web site publish nice text-only versions of articles (https://text.npr.org/) that don't require cookie click-throughs, so I now munge npr.org URLs to link direct to the text-only version of the article

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							

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?