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
- 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;-)
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](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>
    <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>
    <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 that 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):

<table>
<thead>
<tr>
<th>Outlet</th>
<th>Web site</th>
<th>2019 State</th>
<th>RSS feed</th>
<th>2019 State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irish Times</td>
<td><a href="https://irishtimes.com/">https://irishtimes.com/</a></td>
<td>ok</td>
<td><a href="https://www.irishtimes.com/cmlink/news-1.1319%5Cok">https://www.irishtimes.com/cmlink/news-1.1319\ok</a></td>
<td></td>
</tr>
<tr>
<td>The Intercept</td>
<td><a href="https://theintercept.com/">https://theintercept.com/</a></td>
<td>ok</td>
<td><a href="https://theintercept.com/feed/?lang=en">https://theintercept.com/feed/?lang=en</a> ok</td>
<td></td>
</tr>
<tr>
<td>Guardian</td>
<td><a href="https://www.theguardian.com/">https://www.theguardian.com/</a></td>
<td>ok</td>
<td><a href="https://www.theguardian.com/world/rss">https://www.theguardian.com/world/rss</a> ok</td>
<td></td>
</tr>
<tr>
<td>Slashdot</td>
<td><a href="https://slashdot.org/">https://slashdot.org/</a></td>
<td>ok</td>
<td><a href="http://rss.slashdot.org/Slashdot/slashdotMain">http://rss.slashdot.org/Slashdot/slashdotMain</a> no TLS but nice formatting</td>
<td></td>
</tr>
<tr>
<td>El Reg</td>
<td><a href="https://www.theregister.co.uk/">https://www.theregister.co.uk/</a></td>
<td>ok</td>
<td><a href="https://www.theregister.co.uk/headlines.atom">https://www.theregister.co.uk/headlines.atom</a> probably ok</td>
<td></td>
</tr>
</tbody>
</table>
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 developer-mode
  - HOWO: 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
  - 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)

<table>
<thead>
<tr>
<th>Browser</th>
<th>Out-of-the-box chromium</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MB</td>
<td>Time (s)</td>
<td>Set-cookie</td>
<td>Hosts</td>
<td>HTTP Requests</td>
<td>Data-URLs</td>
</tr>
<tr>
<td><a href="https://www.irishtimes.com/">https://www.irishtimes.com/</a></td>
<td>3.3</td>
<td>14.43</td>
<td>147</td>
<td>44</td>
<td>253</td>
<td>7</td>
</tr>
<tr>
<td><a href="https://theintercept.com/">https://theintercept.com/</a></td>
<td>4.4</td>
<td>4.91</td>
<td>22</td>
<td>4</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td><a href="https://www.washingtonpost.com/">https://www.washingtonpost.com/</a></td>
<td>2.8</td>
<td>12.41</td>
<td>369</td>
<td>43</td>
<td>343</td>
<td>4</td>
</tr>
<tr>
<td><a href="https://www.engadget.com/">https://www.engadget.com/</a></td>
<td>4.8</td>
<td>16.42</td>
<td>2</td>
<td>19</td>
<td>341</td>
<td>1</td>
</tr>
<tr>
<td><a href="https://www.theguardian.com/">https://www.theguardian.com/</a></td>
<td>2.4</td>
<td>5.69</td>
<td>35</td>
<td>31</td>
<td>177</td>
<td>5</td>
</tr>
<tr>
<td><a href="http://www.bbc.com/">http://www.bbc.com/</a></td>
<td>1.2</td>
<td>6.16</td>
<td>19</td>
<td>26</td>
<td>193</td>
<td>1</td>
</tr>
<tr>
<td><a href="https://slashdot.org/">https://slashdot.org/</a></td>
<td>1.1</td>
<td>10.7</td>
<td>209</td>
<td>56</td>
<td>235</td>
<td>17</td>
</tr>
<tr>
<td><a href="https://www.theregister.co.uk/">https://www.theregister.co.uk/</a></td>
<td>0.8</td>
<td>5.09</td>
<td>84</td>
<td>22</td>
<td>100</td>
<td>3</td>
</tr>
<tr>
<td>Totals</td>
<td>20.8</td>
<td>75.81</td>
<td>887</td>
<td>245</td>
<td>1668</td>
<td>38</td>
</tr>
</tbody>
</table>
## 2017 Measurements (4)

<table>
<thead>
<tr>
<th>A</th>
<th></th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Browser Site</td>
<td>FF+NoScript+Ghostery+no-cookies</td>
<td>MB</td>
<td>Time (s)</td>
<td>Set-cookie</td>
<td>Hosts</td>
<td>HTTP Requests</td>
<td>Data-URLs</td>
</tr>
<tr>
<td><a href="https://www.irishtimes.com/">https://www.irishtimes.com/</a></td>
<td>1.8</td>
<td>7.28</td>
<td>0</td>
<td>4</td>
<td>97</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="https://theintercept.com/">https://theintercept.com/</a></td>
<td>3.2</td>
<td>3.44</td>
<td>0</td>
<td>3</td>
<td>11</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="https://www.washingtonpost.com/">https://www.washingtonpost.com/</a></td>
<td>2.3</td>
<td>4.12</td>
<td>0</td>
<td>3</td>
<td>23</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><a href="https://www.engadget.com/">https://www.engadget.com/</a></td>
<td>3.7</td>
<td>11.11</td>
<td>0</td>
<td>4</td>
<td>29</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="https://www.theguardian.com/">https://www.theguardian.com/</a></td>
<td>2.5</td>
<td>5.85</td>
<td>0</td>
<td>6</td>
<td>59</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.bbc.com/">http://www.bbc.com/</a></td>
<td>0.5</td>
<td>3.95</td>
<td>0</td>
<td>8</td>
<td>28</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="https://slashdot.org/">https://slashdot.org/</a></td>
<td>0.5</td>
<td>3.62</td>
<td>0</td>
<td>3</td>
<td>20</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><a href="https://www.theregister.co.uk/">https://www.theregister.co.uk/</a></td>
<td>0.5</td>
<td>2.3</td>
<td>0</td>
<td>3</td>
<td>42</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>15</td>
<td>41.67</td>
<td>0</td>
<td>34</td>
<td>309</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>
## 2017/2019 Measurements

<table>
<thead>
<tr>
<th>Jell.ie News (2017)</th>
<th>MB</th>
<th>Time (s)</th>
<th>Set-cookie</th>
<th>Hosts</th>
<th>HTTP Requests</th>
<th>Data-URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out-of-the-box chromium</td>
<td>30.7</td>
<td>17.11</td>
<td>0</td>
<td>1</td>
<td>62</td>
<td>0</td>
</tr>
<tr>
<td>FF+NoScript+Ghostery+no-cookies</td>
<td>31.8</td>
<td>16.21</td>
<td>0</td>
<td>1</td>
<td>61</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Jell.ie News (2019)</th>
<th>MB</th>
<th>Time (s)</th>
<th>Set-cookie</th>
<th>Hosts</th>
<th>HTTP Requests</th>
<th>Data-URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Out-of-the-box chromium</td>
<td>0.14</td>
<td>1.37</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>FF+NoScript+Ghostery+no-cookies</td>
<td>0.22</td>
<td>1.31</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

### Straight browser totals

| Out-of-the-box chromium        | 20.8     | 75.81     | 887        | 245   | 1668          | 38        |
| FF+NoScript+Ghostery+no-cookies| 15       | 41.67     | 0          | 34    | 309           | 3         |
The jell.ie news

• Service: https://jell.ie/news you’re welcome to use if 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?