<WRWeb 15.3 Reading File ****>

A worldwide effort to stop the web losing its memory

Mic Moroney
The Irish Times

Photo: Jane Kelly / Shutterstock.com

Not all human life is yet online, but we're getting there. The ever-ballooning web, currently estimated at five billion indexed webpages alone, is a vast surf of expression and knowledge, constantly evolving in structure and complexity. Yet it's all highly ephemeral: the average website lasts only 100 days before being altered or deleted.

Since the web's infancy, an evangelistic global community of archivists, librarians and researchers has grown up, often connected through cross-membership of bodies such as the Digital Preservation Coalition or the Internet Memory Foundation, all united by the imperative to preserve the web's molten resources for future use.

From eastern Europe to the huge Canadian state archive, institutions and small labs conduct daily web crawls using automated bots. Some conduct emergency operations: the Archive Team stepped in when Yahoo snuffed out GeoCities in 2009, while others are preparing to rescue Vine, after Twitter announced last month that it would suspend the video-sharing service, appalling its 200 million-odd active users.

Perhaps the grandaddy of the project is Brewster Kahle, founder and digital librarian of the San Francisco-based nonprofit Internet Archive, which celebrated its 20th anniversary on October 26th.

Kahle, a pioneer of AI and parallel supercomputing, began bulk-capturing the web in 1996 and storing it on the online Wayback Machine (which now carries more than 273 billion often-defunct webpages), and is still harvesting 250 million webpages a week.

Speaking from San Francisco, he enthusiastically reiterates his ethos of "making all human knowledge available online for free, now", and of eventually being able to replay the web from its earliest days.

Meanwhile, he gets on with projects such as restoring a million links on Wikipedia (all part of "fixing the web"), while, with Firefox, the Internet Archive is test-piloting an add-on which diverts users from 404 pages to the Wayback Machine if dead links can be traced. The archive also recently posted up what it claims is an entire corpus of Balinese-language material.

Even in the "big data" age, the analytic tools do not yet exist to truly mine the web's vast, heterogeneous trove of information, and even the Wayback Machine is not yet fully text-searchable. So web data is currently being stashed and "futureproofed" up on servers across the world.

Open-source applications

Internet Memory Foundation director Julien Masanès, a web archivist since 2000 at the Bibliothèque Nationale de France, points to developing open-source applications such as Flink, Hadoop and HBase, the latter now being developed by Irishman Michael Stack and adopted by Facebook to underpin their mail and other high-traffic apps.

Masanès preaches urgent collaboration between the world's "memory institutions" to preserve the web's resources, not just as a rich data stream for future social historians, economists, technologists, linguists and so on, but as something that one day, by allowing live tracking of mass opinion, might even help predict large-scale social events.

At the National Library of Ireland (NLI), such persuasive enthusiasm is incarnate in its director, Dr Sandra Collins, who, after only a year in the post, has brought renewed urgency to the issue. A maths-physics PhD and a coder with telecom software patents to her

name, she helped set up the infrastructure behind the fledgling Digital Repository Ireland in 2011.

Collins now promotes the NLI's online digital resources, this year inviting the public to select websites that best record Ireland in 2016 or commemorate 1916. The NLI regularly commissions the Internet Memory Foundation to archive "thematic" trawls, using the Internet Archive's open-source Heritrix webcrawler.

The NLI's web archive now hosts regularly captured snapshots of many Irish sites, reconstituted with the original look and feel, and with working hyperlinks.

Back in 2007, the NLI commissioned the Internet Archive to conduct a full trawl of the .ie domain, only to realise they could not make it available without written permission from thousands of site owners and rights holders. This "ginormous ball of data", says Collins, still sits on the library's server unprocessed in WARC (Web ARChive) format, although the NLI hopes eventually to make it available to readers, perhaps on standalone terminals.

Under the legal deposit clause of the Copyright and Related Rights Act 2000, a copy of every Irish paper book or journal must be deposited with the NLI, Trinity College Dublin and the six other universities. Yet there is no provision for the mandatory deposit of Irish ebooks or web-published material (or for allowing such material to be harvested).

The NLI has campaigned for an amendment to allow for such digital legal deposit, which would allow the NLI to exercise its full statutory obligations.

Committee

In 2011, the then government appointed a copyright review committee, chaired by barrister Eoin O'Dell. Two years later it released its considered 180-page *Modernising Copyright* report. This duly gathered dust until Minister of Jobs, Enterprise and Innovation Mary Mitchell O'Connor's press release of August 4th last, announcing an upcoming Bill and talking of digital legal deposit "on a voluntary basis".

With no provision for open "pre-legislative scrutiny", the proposed Bill is now being drafted, while O'Dell, a colourful and instructive blogger, is left "reading the runes" of the press release and musing on the "notable omissions".

Some see the department's interest in exploring US "fair use" common law as an exceedingly ill fit with more restrictive European and even Irish copyright law. They point to the long saga of the Google Books case. Google's vast copyright grab in scanning all books without permission was eventually deemed "fair use" (under snippet-displaying circumstances) by the US Supreme Court.

Yet even with some form of digital legal deposit, the sheer scale of the NLI's web-archiving ambitions makes the mind boggle, particularly considering its current resources of just 11 people in its digital collections department. Although the NLI saw a slight funding recovery this year, it has suffered a 40 per cent funding drop, and a 26 per cent reduction in staff, since 2008.

There are countless other challenges: how to capture the "deep web", with its password- or paywall-protected sites, or those carrying bot-blocking software, most of which are "respected" by web-harvesting bots; or the problems involved in capturing rapidly updated social-media sites, bristling with embedded media.

All eyes are on the US Library of Congress, which, in 2010, in association with Twitter, vowed to archive and preserve every tweet ever posted. Six years on, there is no sign of a properly searchable archive, with tweets daily dumped unsorted (more than half a billion of them a day) on to the library's servers. However, Kahle notes the appointment by President Barack Obama of Carla Hayden, who is dedicated to digitisation and universal access, as head librarian.

There are other legal issues surrounding an NLI web archive: commercial sensitivity, libel, laws against hate speech, and so on, which may require, in Collins's words, some "curation".

Might this mean censorship of the web record, even for researchers? Imponderables fan off in every direction. But while public libraries, in these austerity times, have seen resources cut to the bone, Kahle tells *The Irish Times*: "I have yet to see any legislation which

says we do not want libraries any more. So, until I see libraries being banned and stopped and destroyed, I am going to continue doing my job.”