Web musings deserve their place in history

Michael Skapinker

Hearing that the Nationalbibliothek might be interested in his web musings, a blogger called Night Watchman thought the German national library should mind its own business.

Robert Basic, another blogger, was thrilled, however. "My parents are never going to believe I'm going to be catalogued by the German national library," he wrote.

As my colleague Gerrit Wiesmann reported last week, the German internet was buzzing with news that the library was going to force every blogger and website to submit material, or face a fine of €10,000 (£8,300, $12,700). This was an exaggeration. The library was saving a selected number of websites and the fines were there only as a last resort.

The fuss did demonstrate how much trouble national libraries are having deciding what to store from the internet. Some would say "nothing", on the grounds that libraries are there to store books. But national libraries already contain much more than books, a lot of it useful and much of it now online.

In the US, the Library of Congress has a free digital library called American Memory with 10m documents, including the papers of the early presidents, civil war photographs, films and cartoons.

On the British Library website, you can listen to recordings of Alexander Fleming talking about penicillin or a somewhat expressionless William Butler Yeats reciting: "I will arise and go now, and go to Innisfree."

There are also large numbers of ephemera on the British Library site, such as an 1878 advertisement ("Cadbury's Cocoa is Absolutely Pure; many foreign cocoas sold as 'Pure' are adulterated") and an 1885 ticket for dancing lessons at West Kensington Hall.

Archives like this are popular. When Europeana, the new European Union digital library containing more than 2m books, recordings, paintings and films went live last month, it promptly collapsed under the impact of 10m hits an hour.

These digital archives are also helpful to researchers, providing ready examples of culture, high and low, from earlier ages.

Anyone studying our time will want to know what was on the internet, the transforming technology of our day. Won't future generations simply be able to look it up on the internet themselves? Possibly not. Internet pages die. Colin Webb, an official at the National Library of Australia, said much of the web content from the 2000 Sydney Olympics disappeared as quickly as the athletes.

The British Library forecasts that by 2020 the majority of research will be published in digital form only

Technology changes too: generations from now, people will almost certainly be doing their shopping and arranging their weekends on devices unimaginable to us.

We probably won't even have to wait that long. Last year, a Financial Times article listed once leading-edge technologies from which researchers now struggle to extract information, such as five-hole paper computer tape and laser discs readable only by a mid-1980s computer. Compact discs are still around but, as the FT article observed, they "turn to dust far more quickly than paper".

Libraries will have to turn internet records into whatever technology people are using. But how much should they keep in the meantime?

Storage of books is challenging enough. The British Library, which receives a copy of every publication produced in the UK and Ireland, has to build 12km of new shelves every year to accommodate it all.

The web is far more voluminous. The Library of Congress has 32m books. It has already captured web content equivalent to 55m books.

There are ambitious web archiving initiatives, such as the US-based Internet Archive and several international inter-library efforts. But as Michael Day of Bath University noted in a useful paper: "No single organisation can realistically hope to collect the entire web."

So what should libraries keep? Online academic journals, clearly. The British Library forecasts that by 2020 the majority of research will be published in digital form only.

What of political blogs, online shopping lists, amateur YouTube performances? "We don't have any predefined criteria," Rory McLeod, the British Library's digital preservation manager, said. The library captures large slices of the web, but what is important is to retain the internet's changing technological specifications so that future researchers can reconstruct them.

A wide array of internet content is surely worth holding on to. Today's archeologists are as interested in ancient cooking implements as they are in pottery shards. Tomorrow's will be the same.

Much on the internet is unreliable or plain wrong. In this the web differs from books, where the publishing process offers a rough guarantee of quality.

But the presence of books has not stopped debate about the proper interpretation of the past. At least the web will give tomorrow's historians something to argue about.

michael.skapinker@ft.com