British Library sets out to archive the Web (Update 3)

April 4 2013, by Jill Lawless

In this photo taken Wednesday, April 3, 2013, people work on their computers as old books kept at the British Library are displayed, in London. Capturing the unruly, ever-changing Internet is like trying to pin down a raging river. But the British Library is going to try. For centuries the library has preserved a copy of every book, pamphlet, magazine and newspaper published in Britain. Starting Saturday, April 6, 2013, it will also be bound to record every website, e-book and blog, in a bid to preserve the nation's "digital memory." (AP Photo/Lefteris Pitarakis)

Capturing the unruly, ever-changing Internet is like trying to pin down a raging river. But the British Library is going to try.

For centuries the library has kept a copy of every book, pamphlet, magazine and newspaper published in Britain. Starting Saturday, it will also be bound to record every British website, e-book, online newsletter and blog in a bid to preserve the nation's "digital memory."

As if that's not a big enough task, the library also has to make this digital archive available to future researchers—come time, tide or technological change.

The library says the work is urgent. Ever since people began switching from paper and ink to computers and mobile phones, material that would fascinate future historians has been disappearing into a digital black hole. The library says firsthand accounts of everything from the 2005 London transit bombings to Britain's 2010 election campaign have already vanished.

"Stuff out there on the Web is ephemeral," said Lucie Burgess, the library's head of content strategy. "The average life of a web page is only 75 days, because websites change, the contents get taken down.

"If we don't capture this material, a critical piece of the jigsaw puzzle of our understanding of the 21st century will be lost."

The library is publicizing its new project by showcasing just a sliver of its content - 100 websites, selected to give a snapshot of British online life in 2013 and help people grasp the scope of what the new digital archive will hold.

They range from parenting resource Mumsnet to online bazaar Amazon Marketplace to a blog kept by a 9-year-old girl about her school lunches.

Like reference collections around the world, the British Library has been attempting to archive the Web for years in a piecemeal way and has collected about 10,000 sites. Until now, though, it has had to get permission from website owners before taking a snapshot of their pages.

That began to change with a law passed in 2003, but it has taken a decade of legislative and technological preparation for the library to be ready to begin a vast trawl of all sites ending with the suffix .uk.

In this photo taken Wednesday, April 3, 2013, people work on their computers at the British Library in London. Capturing the unruly, ever-changing Internet is like trying to pin down a raging river. But the British Library is going to try. For centuries the library has preserved a copy of every book, pamphlet, magazine and newspaper published in Britain. Starting Saturday, April 6, 2013, it will also be bound to record every website, e-book and blog, in a bid to preserve the nation's "digital memory." (AP Photo/Lefteris Pitarakis)

An automated web harvester will scan and record 4.8 million sites, a total of 1 billion web pages. Most will be captured once a year, but hundreds of thousands of fast-changing sites such as those of newspapers and magazines will be archived as often as once a day.

The library plans to make the content publicly available by the end of this year.

"We'll be collecting in a single year what it took 300 years for us to collect in our newspaper archive," which holds 750 million pages of newsprint, Burgess said.

And it is just the start. Librarians hope to expand the collection to include sites published in other countries with significant British content, as well as Twitter streams and other social media feeds from prominent Britons.

The archive will be preserved at the London institution and at five other British and Irish "legal deposit libraries" - the national libraries of Wales and Scotland, as well as university libraries at Oxford, Cambridge and Trinity College, Dublin.

This is not the biggest attempt to archive the digital universe. The nonprofit, San Francisco-based Internet Archive - which developed the Web-crawling technology the library is using - has collected 240 billion pages since 1996 on its Wayback Machine at archive.org.

The Library of Congress in Washington, D.C., preserves American digital content such as e-books and e-journals and archives online content in collections built around themes and events, but does not routinely save all websites.

"The Library of Congress is committed to saving e-books, but it is not

committed to saving what everybody is saying about e-books on the Web," said technology historian Edward Tenner.

Britain is one of the first countries to commit in law to capturing its entire digital domain.

The challenge is not just saving the material, but preserving it. The British Library, which has a collection of 150 million items as much as 3,000 years old, says it wants researchers in future centuries to have access to the content. But anticipating changing technology can be tricky - some years ago it was suggested the library's vast collection should be saved to CD-ROM.

To ensure the collection doesn't decay, there will be multiple self-replicating copies on servers around the country, and staff will transfer files into updated formats as technology evolves.

Tenner says keeping up with technology is only one challenge the project faces. Another is the inherently unstable nature of the Web. Information constantly mutates, and search engines' algorithms can change results and prices in an instant - as anyone who has booked airline tickets online knows.

"It is trying to capture an unstable, dynamic process in a fixed way, which is all a librarian can hope to do, but it is missing one of the most positive and negative aspects of the web," Tenner said.

"Librarians want things as fixed as possible, so people know where something is, people know the content of something. The problem is, the goals of the library profession and the structure of information have been diverging."

British Library spokesman Ben Sanderson acknowledged that this is new

territory for an institution more used to documents written on parchment, paper and the fine calfskin known as vellum.

"Vellum - you don't need an operating system to read that," he said.

Citation: British Library sets out to archive the Web (Update 3) (2013, April 4) retrieved 30 April 2024 from https://phys.org/news/2013-04-british-library-archive-web.html