April 11, 2014 4:31 pm

How to preserve the web's past for the future

By Hannah Kuchler

Staff working at the Internet Archive in San Francisco

When a group of investigative reporters was raided last month in Crimea by masked gunmen, they made an unlikely call: to a group of archivists 6,000 miles away in San Francisco. After years of working to expose corruption in Ukraine, the Crimean Center for Investigative Journalism was concerned its reports could be taken off the internet in a moment.

The place they turned to was the Internet Archive, a not-for-profit digital library dedicated to preserving the internet's past for the use of future historians. Within 10 minutes of the call, the archive had started to store the group's web pages. By entering the web address to their online tools, they managed to freeze the pages in time and ensure they collected a new record each day.

Brewster Kahle, the Internet Archive's 53-year old founder, recalls the phone conversation he had with an employee that day. "They said, 'This is just too important to go down.'"

The recent conflict in Ukraine is just the latest occasion when websites that could prove to be vital historical resources have come under serious threat. The Internet Archive hosts collections created by more than a thousand librarians. An archive of websites from the Arab uprising protests that began in 2011 includes an Egyptian site that memorialised victims of violence, and images of protests from Flickr and YouTube. Another, which curated information on last year's bombings at the Boston marathon, includes amateur videos of the explosion and blogs displaying tweets from the immediate aftermath, accusing everyone from "Muslims" to "Koreans" of perpetrating the attack.

The archive's tools have also been used to collect resources that might provide insights into rarely documented areas, such as the home cooking featured on a collection of Mexican food blogs or a middle school in Iowa that has archived its learning materials for the future. Preserved in this way, says Kahle, the internet is a "blessing", an opportunity to transform the study of history with a previously unheard of depth of data from our increasingly online lives. Instead of relying on official reports, easily destroyed private letters and occasional oral history projects, historians could have the potential to study the lives of ordinary people, or niche interests, in as

much detail as they have traditionally traced the tales of the most powerful.

"It is a golden age for librarians, historians and scholars and it is the sweep of digital tools in the humanities that make it so," he says. "In the past, if you wanted to study the evolution of language for a PhD or the roles of women in different eras, you had to do all the grunt work with references and citations all done by hand. Now it can be done by machine at an astonishing rate."

Kahle is a serial tech entrepreneur, a member of the "Internet Hall of Fame" for his work developing WAIS, a publishing system that predated the world wide web. WAIS was sold to AOL in 1995 and his next company, Alexa Internet, which provides web traffic data, was sold to Amazon in 1999. The Internet Archive, which he founded in 1996, is a non-profit organisation, roughly half funded by libraries, who pay for services, with the rest from donations.

Tucked in a quiet corner of San Francisco, away from the techies perpetually chasing the next big thing, the archive is located beneath the imposing arches of a former church. Flashing servers are stacked 10ft high, like old tomes, each blue blink a signal that someone somewhere is trying to reach a web page frozen in time in its archive.

Sitting in a rocking chair in the basement, Kahle explains that one of the biggest drivers behind the idea was his fear that culture and history would be lost to future generations if they were not preserved online. "The web is locked in the perpetual present. It is what people want you to see right now and that's not good enough – that's not how you run a society or open culture," he says. "The best of the web is already not online."

The potential is vast. But the pitfalls are significant too. Not only could it change the way history is told but there are wider questions about who has the rights to guard the web's past and, inevitably in these post-Snowden leak times, what the availability of this data means for individual privacy. So what is the best way to make history from the internet ?

. . .

The Internet Archive is one of just a handful of institutions, including parts of the British Library and the Library of Congress, trying to ensure that what is online now is saved for the future. It does this by capturing more than 1bn web pages a week, though it doesn't try to archive every page of every website – on the fast-moving web the average page is changed every 100 days – or any social media. This snapshot of the web has been taken every two months since 1996 and the gateway to the archive, the "Wayback Machine", is one of the most popular sites online.

The scale of the task is huge. The Library of Congress has a deal with Twitter for all the tweets ever – sent at a rate of about 500m a day – but no legal requirement to receive copies of US websites. Last year Ed Vaizey, the British culture minister, granted the British Library the right of "legal deposit" for ".uk" websites, about 5m sites and 1bn pages. At first it will only harvest every site once a year – with a couple of hundred, such as the BBC, more often – and it still has to scour by hand to find others, such as ".com" sites with predominantly British content.

Niels Brügger, director of the centre for internet studies at Aarhus University in Denmark, recalls his frustration at the way the object of his study used to disappear before his eyes. Now, using the Danish national web archive, which takes a snapshot of all ".dk" websites four times a year, he can track how the internet as a whole is developing in his country, from the different types of websites to the balance between text and images.

He is surprised at how few historians make use of the internet as a source but expects that to change rapidly in five or 10 years as a new generation of scholars better understands its potential and acquires the tools for rigorous data analysis, which are required to study such an ocean of information. "It really is an astonishing new source for historians," he says. "It gives us a great opportunity to study the daily life of people. It is as if we had a tape recorder on the marketplace in the Middle Ages."

The not-for-profit group's founder Brewster Kahle

At the University of Leicester, Ruth Page, a lecturer in linguistics, has already made sources such as Wikipedia central to her work. She studies how entries in the online encyclopedia are edited as a particular event unfolds; for example, the Meredith Kercher murder case in Italy. The site's Italian-language version cites different sources and gives more prominence than the English-language version to material presenting Amanda Knox as guilty.

Page believes that, in future, historians will have to transform the way they work. “I’m an empiricist so I like data. It is like being let loose in a very large sweet shop,” she says. “[But] the days of the lone scholar are gone, in my personal opinion we really need to embrace creative ways to work collaboratively.”

According to Philip Howard, a professor of public policy at the Central European University in Budapest, another consequence of historians having access to greater amounts of data might be to make important research more affordable. Howard used social media to study the Arab uprising, finding that an online civil society with connections to outside observers helped protesters outmanoeuvre authoritarian regimes. But he also found that the internet cut the cost of his painstaking research.

During the protests, he could track where people were using geolocation tags and read “lots of little snippets of diary”. By contrast, he says, “to do something like that 20 years ago, we would need hundreds of thousands of dollars to send in surveyors and a survey instrument 30 questions long. That would be the traditional way to study a movement – it was very expensive.”

For these academics, the benefits are clear. But there are concerns too, chiefly that the internet companies who own much of this data – and whose primary business model tends to be selling it to data-hungry advertisers – have too much control over it.

Two months after Margaret Thatcher’s death in April 2013, Ruth Page decided she would like to study the Twitter reactions. By then, however, she could not archive it contemporaneously and so would have had to pay Twitter’s private data partners for the tweets. “It is completely commercially driven,” she says. “It is really frustrating for researchers who don’t have the money to pay.”

Brügger acknowledges a conflict between companies who see “data as a commodity” and historians who see it as source material. “Companies owning much of the data don’t have a long-term perspective. Cultural heritage institutions have to preserve it for eternity, which is a weird thing for a company to do in a way.”

Twitter has introduced a pilot programme to allow researchers access to their public and historical data. Research institutions must submit a proposal to Twitter, which will select a “small number” to receive free data.

Issues surrounding privacy also play an important part. Some companies have privacy policies, which prevent them from making identifiable information available, while researchers say they struggle to get information from Facebook, where people often restrict what they post to friends only. The Internet Archive removes personal material if people request it. Kahle says it is a “trade-off” the archive has made to allow it to put pages in the Wayback Machine immediately. “Often it is personal things, like a blog about a marriage that someone would like to forget,” he says.

From the historians’ perspective at least, little thought seems to have been given to whether the right to privacy should end with a person’s death or a set period of time afterwards, a guideline such as the 20-year rule for the release of government documents in the UK. Brügger’s belief is that it is best to archive everything and leave it to the scholars to consider ethical considerations for each research project. “If we don’t archive it, it is gone two years later,” he says.

. . .

For some, the answer to these problems lies in giving people – and groups of people – the ability to preserve their own online material in a kind of individual time capsule.

Phil Libin is chief executive of Evernote, a service for organising information across multiple devices, which is thinking about how to keep people’s personal archives after they die. He says everything would be private by default but there should be some guarantee that the information goes where the user wants it to, even decades after they stop paying for the service. “The data will be around for a hundred years and you can control who has access to it and when – should it be their children, grandchildren or 50 years from now,” he says.

Google’s Cultural Institute has also created a tool for people to make galleries of their lives on their own websites. Amit Sood, who heads the institute, says Google did not want to be a “digital curator” but, instead, wanted to allow everyone from museums and art galleries to individuals to do it themselves. “I think it will come down to a marriage of the professional and the amateur. If the content is really benefiting a large group of people, you will always find people to preserve this stuff,” he says.

Fast-moving technology companies are not the obvious guardians of the past. As they dash for the next big thing, be it virtual reality or internet-connected homes, even a year can seem

like a long time in Silicon Valley. In San Francisco, however, a new museum hopes to redress the balance a little. The Long Now Foundation, an organisation founded in 1996 to promote long-term thinking, wants to create a space to encourage people to stop and think about how the decisions they make now will affect the next 10,000 years.

Laura Welcher, the foundation's director of operations, says for years they have feared a "digital dark age" where resources kept only online disappear. Initially, the project looked at ways to help people constantly migrate their files to ensure, for example, old Microsoft Word documents were still readable in the newer versions. Then, they got much more ambitious, building a new version of the Rosetta Stone, a silicon disc inscribed with thousands of pages documenting human languages.

"We were very purposeful about creating a future artefact . . . intentional migration of information into the future is much harder digitally," she says.

Yet she, too, is optimistic about the opportunity to create and preserve your own space online. "I think keeping a story of an individual or of a gender or of a cultural group is more egalitarian because access to archiving your stuff is easier," she says. "It is a very new thing to have your voice out there like never before."

Hannah Kuchler is the FT's San Francisco correspondent