Building a National Archive of Government Social Media

Claire Newing

The National Archives, UK, claire.newing@nationalarchives.gov.uk

Jake Bickford

The National Archives, UK, jake.bickford@nationalarchives.gov.uk

Archiving social media content is a vital but difficult task for The National Archives, UK. The complexities of archiving social media content are well known, but despite this we now have an archive of several million posts from seven platforms which is growing by the day, much of which is freely accessible to the public. In this poster we will show how we tackled the task of capturing and preserving the rapidly growing and increasingly important social media output of UK central government organisations. We will describe the tools we used and how we adapted, and continue to adapt, our processes in the face of technical and policy challenges.

Additional Keywords and Phrases: web archiving, social media archiving, born-digital collections, UK Government

CCS CONCEPTS • Digital libraries and archives • Social media • E-Government

1 CONTEXT

The National Archives (UK) is the official archive and publisher for the UK Government. The Public Records Act (1958) empowers us to collect and preserve the public record in any format or medium, and make it as widely available as possible. Starting in 2004 (though with backfilled content dating from 1996), the UK Government Web Archive (UKGWA) began collecting government websites for permanent preservation via remote web crawling, and today holds some nine billion digital objects in a freely available online archive. The emergence of social media posed new challenges to the completeness of this online public record, from both a technical perspective (since social media platforms are considerably more complex than traditional websites) and in terms of the policy and ethical challenges presented by the medium (as they contain material that forms part of the public record published on platforms that are outside government ownership, and that is frequently presented alongside non-government content). It was also recognised that content published on social media was even more fragile than other forms of digital record, being by its nature ephemeral and subject to deletion or rapid changes to the host platform. This poster tells the story of The National Archives and our partners' innovative work to preserve this vital but vulnerable component of the public record, from our initial experiments to the development of production processes. It will highlight our recent efforts to navigate the increasingly restrictive environment for the preservation of social media, and explore how researchers can make use of our growing archive.

2 HISTORY OF THE SOCIAL MEDIA ARCHIVE

The UKGWA began systematically capturing social media in 2014. Initial experiments focused on archiving Twitter using the same remote crawling system as for websites. This proved technically complex given the highly interactive and

dynamic nature of social media sites, while this capture method made separating out in-scope government content from other material challenging. It was therefore decided to focus on only capturing content itself (for example, the text of a tweet, along with any associated media and metadata) rather than the look and feel of specific platforms. This mitigated the risk of capturing content that was outside our legal remit and opened alternative approaches to capture via platform APIs. It also maintained our existing model of archiving public content without requiring engagement from account holders. Working with our then partners, Internet Memory Foundation, we began regular harvesting of Twitter and YouTube. We also developed a process for resolving and crawling embedded links in tweets. From 2017, working with a new supplier, MirrorWeb, we increased capture frequency to a daily check of the APIs, increased the number of channels captured, and began providing access to Tweets via a new playback platform. In 2019, in part prompted by Flickr's announcement that it would limit free account holders to 1000 images, we began capturing Flickr content, backfilling over a decade's worth of images from government accounts. At this time, in response to user research, we also introduced what we believe to be the first full-text search across any social media archive.

This approach served us well for YouTube, Twitter and Flickr, and during this period, we accrued and published some 2.5 million Tweets, 126,000 YouTube videos, and 262,000 Flickr images. But archiving other platforms such as Meta products and LinkedIn via API proved impossible. We were, however, able to implement a limited in-house capture process for Instagram using Instaloader. This precedent proved helpful when changes to Twitter's API rules made our model impossible to sustain from mid-2023. In the run-up to the 2024 General Election, we began in-house capture of selected Twitter and Facebook accounts using similar client-side tools. However, these tools are less reliable than using official APIs and at time of writing we are once again unable to capture Meta social media via this method. While we continue to accrue YouTube and Flickr content according to our established processes, our Twitter archive has also become more fragmentary and uneven following the API changes (down from around 850 accounts to around 100 regularly captured). We have therefore had to re-assess our preference for a remote capture approach and concluded that for high-priority accounts (for example Public Inquiries) we need to engage with account holders and ask them to use the platform's data portability features to extract social media content for deposit. We have piloted this approach with Twitter and have begun receiving our first accessions from account holders. Content captured in-house or submitted by account holders is currently stored securely while we work to add it to the public archive.

3 FUTURE WORK

We continue to work to adapt our operating model to the more challenging social media archiving environment that confronts us, and we will shortly be publishing our first user-deposited Twitter content in the main archive. We will then focus on integrating content that has been captured in-house, focusing on Twitter, Instagram and Facebook. We have carried out experimental captures of TikTok and BlueSky, and we will continue to monitor these platforms, with a view to eventually archiving them as available resources permit. We are also working with our partners to better integrate social media content with the wider web archive and have recently implemented a script that links embedded YouTube videos in the web archive to the corresponding video in the social media archive, allowing users to click through and view the video. This remains work in progress, and more sophisticated forms of linkage between the collections are possible future developments. Finally, we believe that the social media archive is an underused research resource, and are keen to support researchers from all disciplines who would like to explore the official record of government social media.