

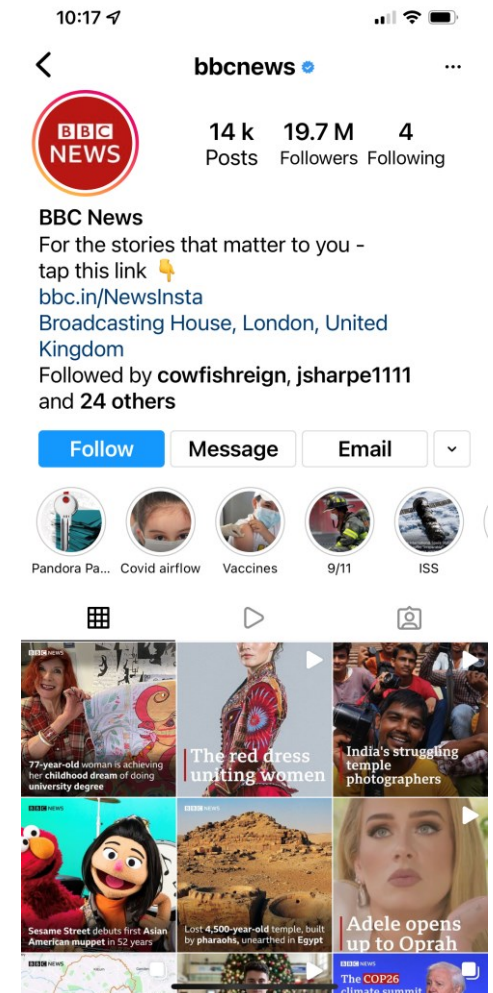
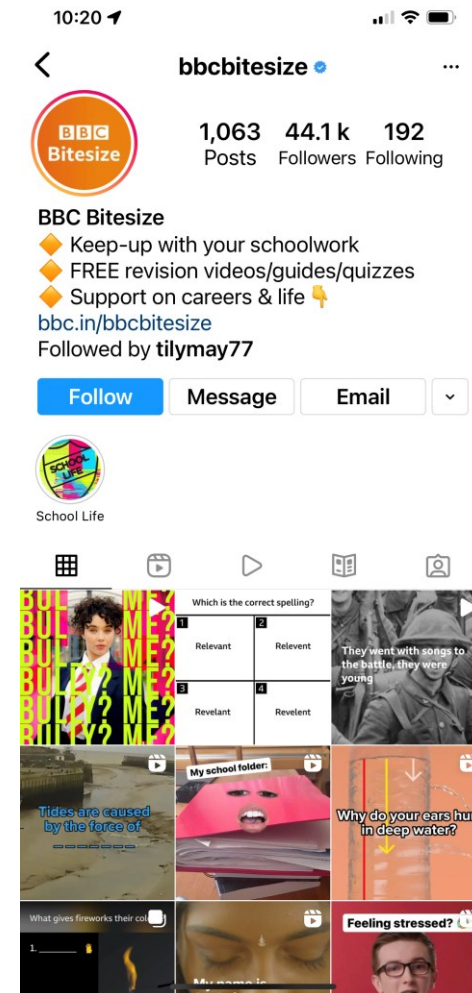
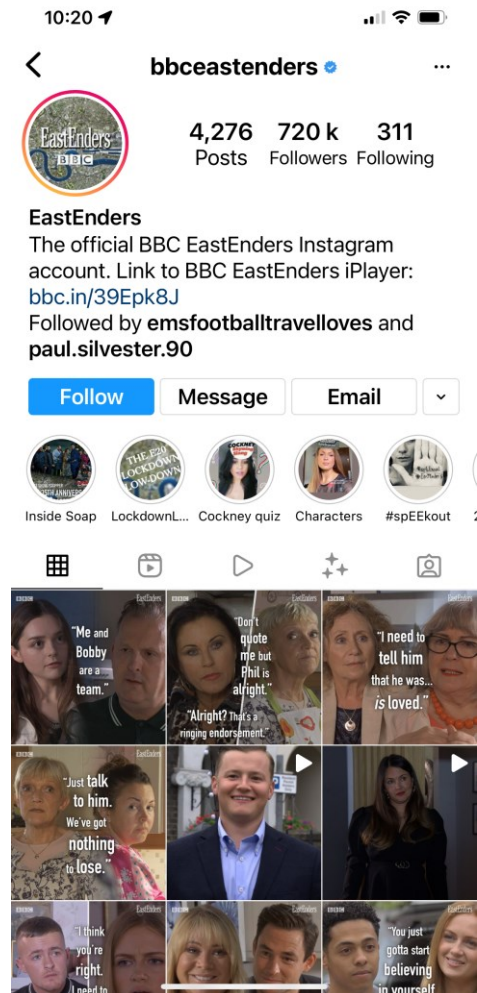
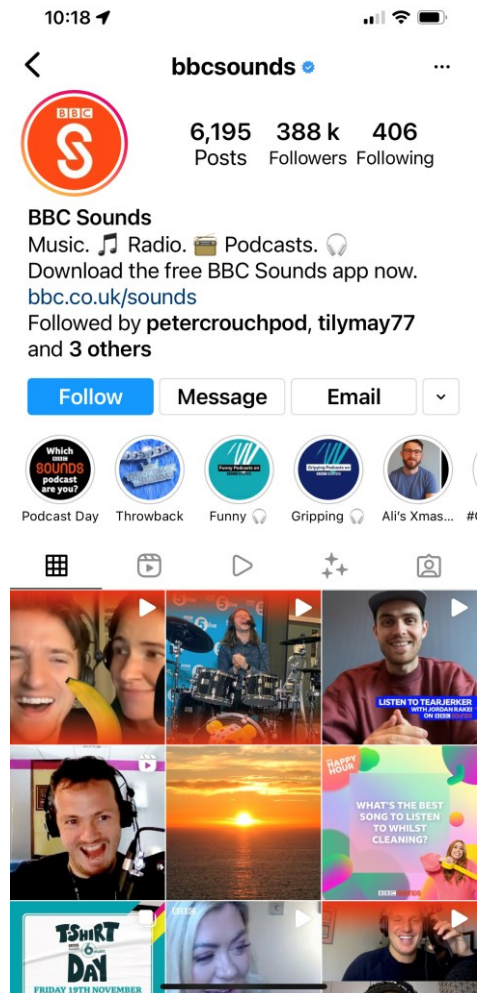
BBC ARCHIVES TECHNOLOGY AND SERVICES USING INSTALOADER TO CAPTURE BBC INSTAGRAM ACCOUNTS

DECEMBER 2021

BBC Social Media Archiving - rationale

- The BBC has over 1000 social media accounts across Instagram, Facebook, Twitter, YouTube and TikTok
- The Royal Charter requires representative distributed BBC content to be kept
- Provides a corporate record of BBC promotion
- It's an historical and cultural record of the BBC
- Content can be repurposed

Examples of BBC Instagram Accounts



Why Instaloader

- The BBC has hundreds of Instagram Accounts. We have solutions for Twitter and YouTube, but needed something for Instagram
- We originally heard about Instaloader at a DPC event in November 2019
- We downloaded from GitHub and set it up at the start of 2020 to capture 30 closing BBC accounts
- If Instagram changes something that breaks the capture process, the open-source community is likely to release a fix Instaloader quicker than an engineer could.
- Being able to use this tool to automate capture greatly reduces the amount of manual hours required to capture content.
- By having a consistent file naming scheme, we can later refer to the assets which we can later use to make the asset searchable and browsable.
- The assets are currently not available to search yet

Scale of capture after 2 years

System Summary

November Report	Accounts	Posts	Media Files	Images	Videos	Size
Total	184	400,473	525,781	383,565	142,216	1.6TB
Change Since Last Report	0	4,852	8,113	5,628	2,485	31.2GB

The BBC consistently publishes over 2,000 videos and 4,000 images on Instagram each month

Using the tool

- Requires someone with system engineering and code experience to deploy
- The tool provides some common functions through the command line interface
- It can be used to download entire profiles
- It can be run from a standalone desktop/laptop provided it has Python installed and has sufficient storage to hold the assets captured. We run our version from a local server.
- We have to have a number of 'dummy' accounts to be able to sign-in to Instagram to archive it
- Currently we upload the assets to AWS S3 and periodically store the assets on an LTO tape. However, our script requires the files to be stored locally so that it is able to determine what posts have already been captured.

Maintaining the tool

- Adding and removing accounts requires some manual intervention from the engineering team.
- Due to issues with regards to Instagram blocking requests from web crawlers, our script may stop running, the engineering team are sent an email alert when this happens so that they can investigate.
- Usually the fix is to restart the script after a period of cooldown. If our 'dummy' account has been blocked, then we roll over to a new 'dummy' account
- In our case there is server maintenance to consider. However, as this server is used for other applications it is already maintained by the engineering team.

Shortcomings of the tool

- We ran into a fair number of issues with requests being blocked by Instagram. We found the built-in rate controller was not enough to mitigate requests being blocked by Instagram. We got around this by adding preventative measures to our code such as cycling through our own 'dummy' accounts.
- Instaloader may not work with all cloud platforms. When we tried to run the Instagram archiver on AWS, Instagram was able to identify the IP as an Amazon IP address and can deny requests from it.
- Instaloader is a great tool but there is uncertainty on how well the application will work in the future.
- There is no inbuilt interface to search and browse the captured data.

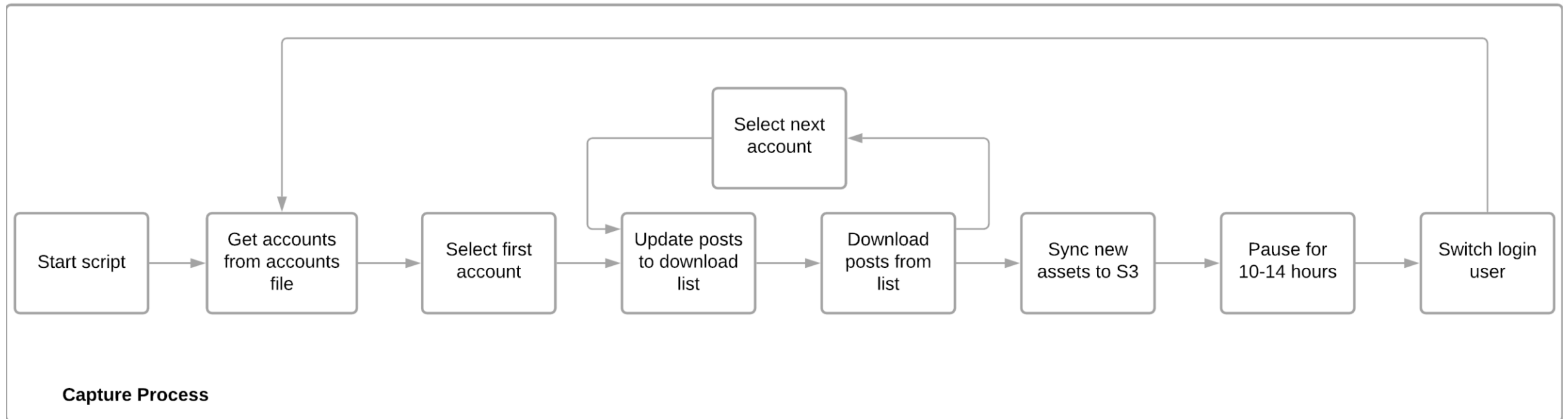
Example Script Output

```
2021-11-28 22:10:15 - INFO - Fast update download.
2021-11-28 22:10:21 - INFO - 1/120 : 5livesport - Downloading 4 posts
5livesport 1/4 : CW03rydolF9
5livesport 2/4 : CW01ybAIP3M
5livesport 3/4 : CW0xmLcoQDG
5livesport 4/4 : CW0ehgUoRRk
2021-11-28 22:11:49 - INFO - 2/120 : bbc - Downloading 2 posts
bbc 1/2 : CW1C0TBiPiH
bbc 2/2 : CW0Z1zutSV4
2021-11-28 22:13:18 - INFO - 3/120 : bbc_archive - Downloading 1 posts
bbc_archive 1/1 : CW0uOn_qzXR
2021-11-28 22:14:29 - INFO - 4/120 : bbc_culture - Archive already up to date, skip download
2021-11-28 22:15:35 - INFO - 5/120 : bbc_hardtalk - Archive already up to date, skip download
2021-11-28 22:16:41 - INFO - 6/120 : bbc_midlands - Downloading 2 posts
bbc_midlands 1/2 : CS6CNB9KTOh
Download failed, skipping: CS6CNB9KTOh

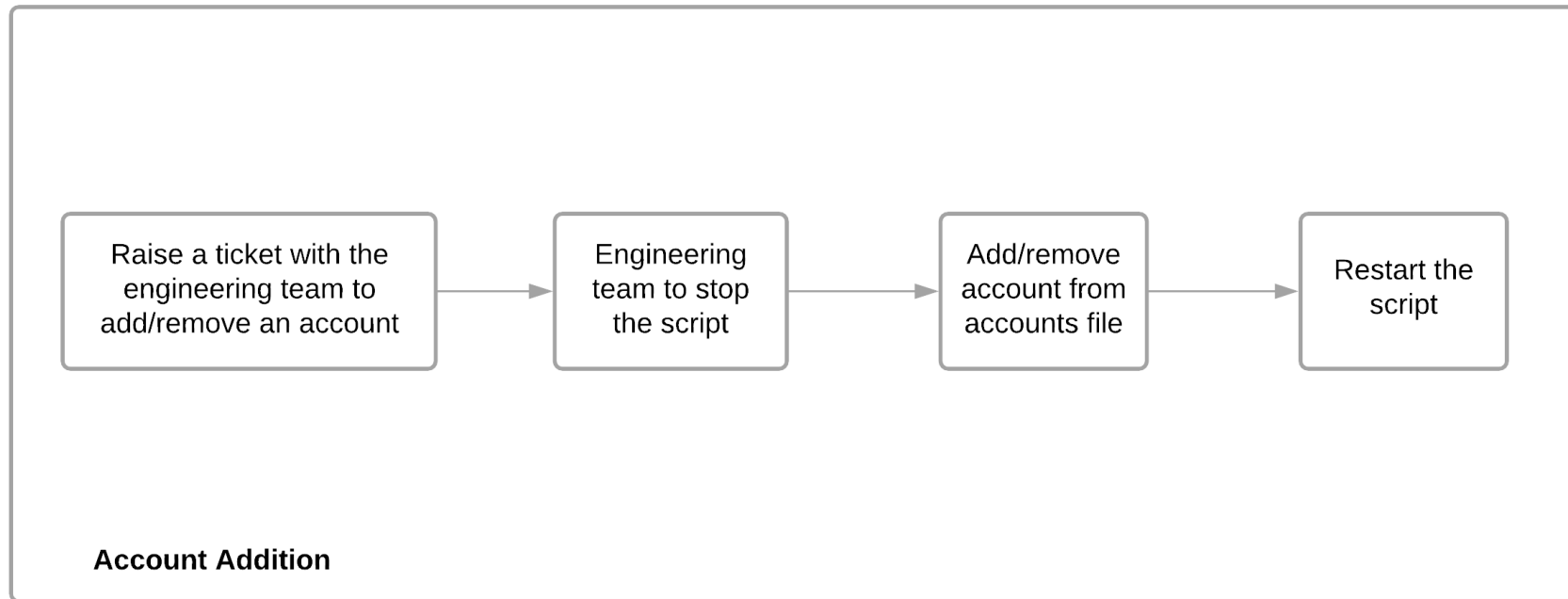
Fetching Post metadata failed.

bbc_midlands 2/2 : CW1C3e1IX1s
2021-11-28 22:17:50 - INFO - 7/120 : bbc_reel - Archive already up to date, skip download
2021-11-28 22:18:53 - INFO - 8/120 : bbc_travel - Downloading 1 posts
bbc_travel 1/1 : BUvUuV3FLl_
download_pic(): HTTP error code 410. [retrying; skip with ^C]
download_pic(): HTTP error code 410. [retrying; skip with ^C]
Download failed, skipping: BUvUuV3FLl_
```

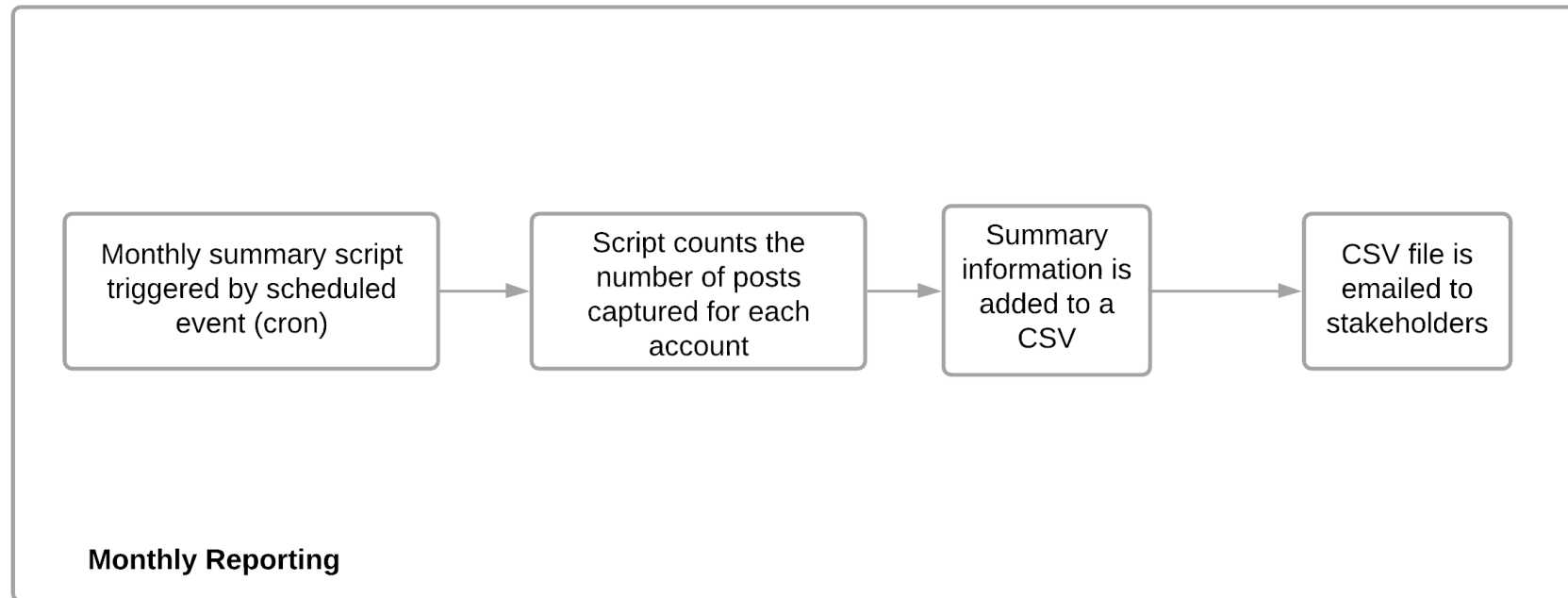
Workflow – Capture Process















Workflow – Add Accounts



Workflow – Monthly Reporting



File storage

<input type="checkbox"/> Name	Date modified	Type	Size
 2016-05-08_18-33-42_(BFKbJURDHPz)	13/08/2020 09:24	JSON File	17 KB
 2016-05-08_20-34-07_(BFKOtGQDHrv)	13/08/2020 09:24	JPG File	63 KB
 2016-05-08_20-34-07_(BFKOtGQDHrv)	13/08/2020 09:24	JSON File	20 KB
 2016-05-11_14-46-47_(BFRVV1jDHmN)	13/08/2020 09:24	JPG File	80 KB
 2016-05-11_14-46-47_(BFRVV1jDHmN)	13/08/2020 09:24	JSON File	24 KB
 2016-05-16_14-42-04_(BFEMxovjHvC)	13/08/2020 09:24	JPG File	52 KB
 2016-05-16_14-42-04_(BFEMxovjHvC)	13/08/2020 09:24	JSON File	23 KB
 2016-05-17_20-01-23_(BFhWHPHjHqP)	13/08/2020 09:24	JPG File	39 KB
 2016-05-17_20-01-23_(BFhWHPHjHqP)	13/08/2020 09:24	JSON File	20 KB
 2016-05-17_20-06-07_(BFhWp5WjHrk)	13/08/2020 09:24	JPG File	93 KB
 2016-05-17_20-06-07_(BFhWp5WjHrk)	13/08/2020 09:24	JSON File	22 KB
 2016-05-19_07-16-05_(BFIIH17DHIv)	13/08/2020 09:24	JPG File	45 KB

We need to build an search interface or add assets to a pre-existing digital asset management system to be indexed for searchability

