PRESERVING ONLINE JOURNALISTIC CONTENT IN DISRUPTIVE TIMES

The case of collection.news

Lok Hei Lui
University of Toronto
Canada
kenlh.lui@mail.utoronto.ca
0000-0001-5077-1530

Abstract – Journalistic content is a crucial part of history, yet its longevity always remains uncertain without proper curation and preservation. This is true in particular when it comes to journalistic content under authoritarian regime contexts, where freedom of the press and information freedom are usually in vain. The article explores the case of collection.news, a community initiative that crawled, disseminated and hosted the journalistic content of Apple Daily, a pro-democracy media outlet that was forcibly shut down by the authority in Hong Kong. By discussing the key events and tools used by collection.news initiative, the three distinctive features of it, namely exigency, decentralization, and anonymity, are highlighted. Finally, suggestions to the digital preservation field for supporting these community initiatives in authoritarian regimes will be given.

Keywords – collection.news, archives-at-risk, authoritarianism, community archives, Hong Kong

Conference Topics – Digital accessibility, inclusion and diversity

I. INTRODUCTION

There is a common saying that "journalism is the first draft of history". News content is one of the important records that document the events happening around the world, and also has become an indispensable part of many people's lives. That said, freedom of the press is not guaranteed in many parts of the world, especially for people living in authoritarian regimes. Critical journalism platforms operating under such regimes are always being targeted by the authorities since true journalism, which involves exposing government wrongdoings, could be a threat to these regimes. When these media platforms are cracked down by the regime, the associated news content, if not well preserved by third parties, is usually vanished.

This article will discuss the case of collection.news. The project is a community initiative that preserved the web content of Apple, a now-defunct pro-democracy media platform based in Hong Kong. The research methods will be outlined in the next section. After that, the key events, tools used, and approaches adopted of the initiative will be illustrated and then the analysis follows. Lastly, a conclusion will be drawn and suggestions for the digital preservation field will be provided.

II. RESEARCH METHODS

The paper adopts a qualitative approach in this study by analyzing primary and secondary sources. These sources include forum posts, collection.news website and its GitHub repo documentation. Drawing upon the analysis, the author will further discuss the tools used, coordination and distinctive features of the preservation project, make analysis, and give suggestions.

III. BACKGROUND OF APPLE DAILY

Apple Daily was a prominent pro-democracy media outlet before its forced closure in June 2021.

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

https://hdl.handle.net/2142/121103

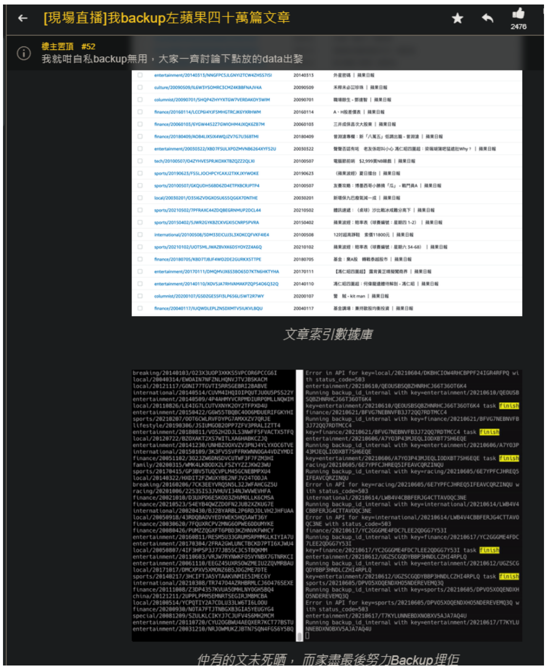
In 2019, Hong Kong experienced the largest-scale pro-democracy movement, the Anti-Extradition Bill Movement, in the territory. In response to the political unrest in Hong Kong, the Chinese government promulgated the controversial Hong Kong National Security Law (NSL) on July 1, 2020, the 22nd anniversary of the handover of Hong Kong's sovereignty from the United Kingdom to the People's Republic of China.

Pro-democracy media platforms were deemed as one of the high-risk groups being targeted by the authority under NSL [1]. Two months after the NSL came into effect, Jimmy Lai, the founder of Apple Daily, was arrested by the National Security Department of the Hong Kong Police Force on suspicion of "collusion with foreign forces". On the same day, the police also searched the headquarters of Apple Daily. Despite the arrest, Apple Daily kept its business as usual afterward.

However, less than a year later, on June 17, 2021, the National Security Police arrested other management of Apple Daily and searched the headquarters again. The authorities also froze Apple Daily's assets, which eventually led to the media's cessation of operations. On June 23, the board of Apple Daily announced that the company would terminate its operations no later than June 26, and its digital platform would be shut down by midnight June 24.

IV. THE EMERGENCE OF COLLECTION.NEWS

Following the forced shutdown of Apple Daily on June 24, 2021, a netizen "五大素球缺汁不可" (user id: #355204) created a thread on LIHKG forum, which is a Reddit-like forum based in Hong Kong, announcing that they had web-crawled over four hundred thousand articles from Apple Daily's website [2] (Fig. 1). In addition, the original poster expressed their wish to index the content afterward for web hosting and invited other forum users to contribute ideas on how to distribute and host the content. Some forum users suggested in the thread that there should be a frontend website for hosting the news article content with search functionality, while others proposed some distribution methods/platforms such as InterPlanetary File System (IPFS), BitTorrent, and GitLab for disseminating the news content data.



[現場直播]我backup左蘋果四十萬篇文章

2476

樓主置頂 #52
我就咁自私backup無用，大家一齊討論下點放的data出嚟

文章索引數據庫

仲有的文未死晒， 而家盡最後努力Backup埋佢

Figure 1 Snapshot of the inaugural thread discussing the crawling of Apple Daily web content on the LIHKG Forum

Two days later, the original poster created another thread [3] and included a GitHub repository link [4] to the initiative's documentation. On the GitHub repository, the author outlines the initiative's position and aims. Below is a translated version:

- The initiative primarily aims to back up the textual content of Apple Daily as I strongly believe in the power of words.
- The initiative aims to index the content and host an SEO-friendly website for people to search for old articles.
- Revealing the data is meant to promote brainstorming and encourage us to think about how we can utilize the data. The initiative does not intend to conceal the data. In fact, most people will not extensively browse the data after it has been backed up.
- Revealing the data can achieve the goal of decentralization. Even if someone who possesses the data gets into trouble later on, others can still continue.
- The initiative does not intend to crawl all of Apple Daily's content, such as images, videos, Instagram accounts, YouTube channels, Telegram channels, Facebook accounts, etc. I am aware that someone else is working on this.

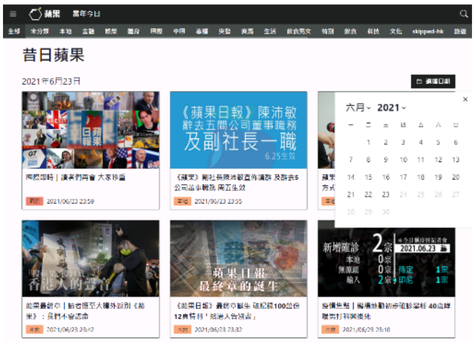
The GitHub repository also provided a tutorial on how the end-users could download a copy of the media content through the Resilio Sync download tool in the forum post. The author explained the adoption of Resilio Sync: because of its decentralized, high-speed P2P sharing and flexibility in the modification of source files features.

V. COLLECTION.NEWS FRONTEND ACCESS AND ITS FUNCTIONALITY

Less than a month after the original post, on July 21, 2021, the same user created another thread on the LIHKG forum. They mentioned that after some effort throughout the weeks, they crawled more than 2.2 million articles from Apple Daily's website and hosted a website[1] for frontend access to news articles. The original poster also mentioned that the aim of the website is hoping an essential part of Hong Kong history would not be faded out because of the closure of Apple Daily.

Fig. 2 is the landing page of Apple Daily's content on collection.news website. By clicking on the boxes, users will be directed to the corresponding article. Akin to the layout of Apple daily's original website, the top grey bar lists different categories of articles. The date selection menu, represented by the middle black box (選擇日期), allows users to sort articles based on their publication dates.

In the top right-hand corner, users can access the website's search function. The indexing service is provided by Google (Fig. 3). This feature allows users to search articles by keywords. Mentioned in the FAQ section of collection.news website, using Google's indexing service is based on financial considerations and mitigating the risk of experiencing cyber attacks on the indexing server by hackers.



蘋果 當年今日

昔日蘋果

2021年6月23日

選擇日期

六月 2021

《蘋果日報》陳沛敏辭去五間公司董事職務及副社長一職 6.25生效

2021/06/23 23:59

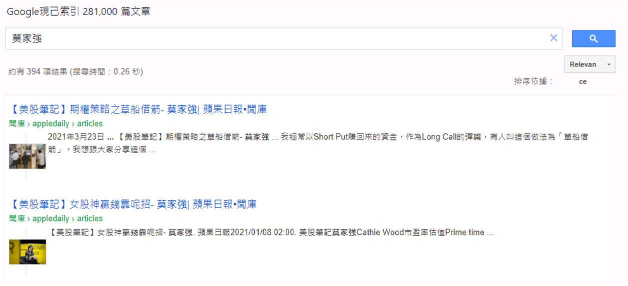
2021/06/23 23:55

2021/06/23 23:42

2021/06/23 23:32

2021/06/23 23:10

Figure 2 Screen capture displaying the landing page of Apple Daily's content on collection.news



Google現已索引 281,000 篇文章

莫家強

約有 394 項結果 (搜尋時間：0.26 秒)

排序依據： Relevance

【美股筆記】期權策略之草船借箭- 莫家強| 蘋果日報•聞庫

聞庫 › appledaily › articles

2021年3月23日 ... 【美股筆記】期權策略之草船借箭- 莫家強 ... 我經常以Short Put賺回來的資金，作為Long Call的彈藥，有人叫這個做法為「草船借箭」，我想跟大家分享這個 ...

【美股筆記】女股神贏錢靠呢招- 莫家強| 蘋果日報•聞庫

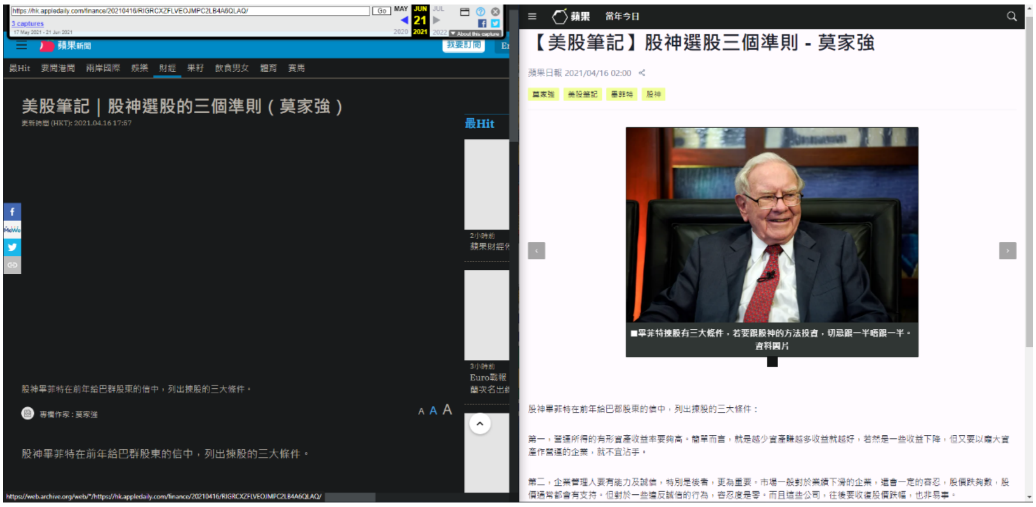
聞庫 › appledaily › articles

【美股筆記】女股神贏錢靠呢招- 莫家強. 蘋果日報2021/01/08 02:00. 美股筆記莫家強Cathie Wood市盈率估值Prime time ...

Figure 3 Screen capture showcasing the search functionality on collection.news

VI. STRENGTHS AND LIMITATIONS OF COLLECTION.NEWS

The Internet Archive is another important platform for archiving Apple Daily's web content. However, the two platforms serve different functionalities, and have their pros and cons. The left and right sides are screenshots of the same article from the Internet Archive's Wayback Machine and collection.news respectively (Fig. 4). In comparison, the most significant advantage of collection.news is its ability to showcase attached photos, an essential



https://hk.appledaily.com/finance/20210416/RIGRCXZFLVEOJMPC2LB4A6QLAQ/

3 captures

17 May 2021 - 21 Jun 2021

美股筆記｜股神選股的三個準則（莫家強）

更新時間 (HKT): 2021.04.16 17:57

股神畢菲特在前年給巴郡股東的信中，列出揀股的三大條件。

專欄作家：莫家強

股神畢菲特在前年給巴郡股東的信中，列出揀股的三大條件。

最Hit

2小時前 蘋果財經

3小時前 Euro戰報

https://web.archive.org/web/*/https://hk.appledaily.com/finance/20210416/RIGRCXZFLVEOJMPC2LB4A6QLAQ/

蘋果 當年今日

【美股筆記】股神選股三個準則 - 莫家強

蘋果日報 2021/04/16 02:00

莫家強 美股筆記 畢菲特 股神

■畢菲特揀股有三大條件，若要跟股神的方法投資，切忌跟一半唔跟一半。資料圖片

股神畢菲特在前年給巴郡股東的信中，列出揀股的三大條件：

第一，營運所得的有形資產收益率要夠高。簡單而言，就是越少資產賺越多收益就越好，若然是一些收益下降，但又要以龐大資產作營運的企業，就不宜沾手。

第二，企業管理人要有能力及誠信，特別是後者，更為重要。市場一般對於業績下滑的企業，還會一定的容忍，股價跌夠數，股價通常都會有支持。但對於一些違反誠信的行為，容忍度是零。而且這些公司，往後要收復股價跌幅，也非易事。

Figure 4 Screen capture comparing Wayback Machine and collection.news platforms for the same article

[1] https://collection.news/

component of online news articles, whereas Wayback Machine was unable to crawl the picture for this article and many other instances.

However, one major problem for the collection.news platform is that the content might not be up-to-date and may affect the data integrity of the content. As shown in the timestamp, Wayback Machine successfully crawled a more recent version of the webpage (2021.04.16 17:57), whereas the content hosted on collection.news was from an earlier version (21.04.16 02:00), which was 15 hours earlier. This discrepancy is most likely due to the limitation of web crawling from a legacy source before the complete shutdown of Apple Daily's web server.

Also, despite collection.news being a newly-built website for hosting the archived news content of Apple Daily with search functionality, it is unable to preserve the user interface and layout of Apple Daily's website, unlike Wayback Machine does.

VII. Analysis of the preservation process

Table 1 Summary of tools used by collection.news categorized by usage

Frontend Access	collection.news
Content Distribution	Resilio Sync, IPFS
Announcement/Coordination	LIHKG forum
Documentation	GitHub, GitLab

Table 1 summarizes the tools and platforms adopted by collection.news project. In contrast to traditional institutional approaches to implement preservation projects with long-term planned, structured and centralized features, the whole collection.news digital preservation project was an autonomous, decentralized and anonymous digital preservation movement initiated by passionate netizens. The three distinctive characteristics of this community-led project are exigency, decentralization and anonymity, respectively.

Exigency is one notable characteristic of this project. The Apple Daily web content was an archives-at-risk with only a small window of time to plan and execute the preservation process. From Apple Daily being searched by the National Security Police on June 17,2021, to the time that Apple Daily eventually ceased operation by midnight June 24, 2021, there was less than a week of time. This tight timeframe posed challenges to the preservation project facilitators, since they would have to work under intense pressure and grasp the golden period before the complete shutdown of service to crawl the data as much as they could. This urgency also meant that the preservation plan was likely to be incomplete and rough, potentially leading to critical data loss.

Decentralization is also another distinctive feature. Most digital preservation projects, due to financial, management and staffing considerations, are usually managed by GLAM (Galleries, Libraries, Archives, and Museums) institutions with a centralized operational approach, whereas the community-led collection.news project was operated in a decentralized way:

1. For preservation storage, there was no centralized data repository or platform for long-term preservation. Instead, the project publicly disseminated the news content data to end-users and relied on every single end-user for long-term preservation. This practice was entirely different from most institutional centralized approaches.
2. For data dissemination, the collection.news initiative made use of peer-to-peer protocol tools such as Resilio Sync and IPFS to disseminate the news content. The main advantage is the decentralized feature that could disseminate data with multiple users simultaneously while avoiding download speed bottlenecks. Another benefit of using peer-to-peer protocols is to prevent government internet censorship or denial-of-service attacks on a single hosting platform.
3. Adopting GitHub and GitLab as the platforms for documentation was also a decentralized approach. These open-source project platforms enable open collaboration and backup of content from every user without restrictions. This can ensure further access to the documentation. Also, similar to the case of Mainland China internet users, hosting documentation and organizing community archives on GitHub could be a way to circumvent Chinese government internet censorship [5].

Another distinctive characteristic of this project is its emphasis on anonymity. While most digital preservation projects were organized by identifiable institutions or organizations, collection.news initiative was largely operated under the radar. The user name of the original poster's account on the LIHKG forum was a pseudonym. The GitHub repo was also owned by a brand new, designated account with no prior history. In addition, the initiative never publicly recruited volunteers nor openly organized crowdfunding campaign for funding. The organizational and operational details, such as funding, the number of facilitators and the decision-making model, remain concealed. This was, as mentioned in the FAQ on collection.news, intended to reduce the potential political risks.

VIII. Conclusion

This article introduces the case of collection.news, an autonomous and anonymized community initiative for preserving the online content of a Hong Kong-based newspaper platform, Apple Daily. The article then overviews the preservation process and approaches adopted by collection.news, by highlighting the key events and tools used. In the later part, this article points out three distinctive features of the whole community initiative compared to conventional digital preservation projects, namely exigency, decentralization, and anonymity. This case study should be helpful for readers to understand community-led digital preservation activism issues under authoritarian regime contexts.

In recent years, the world has been experiencing serious global democratic backsliding. With the expansion of authoritarianism, unfortunately, there might be a growing trend of more cases like the sudden collapse of Apple Daily. With reference to the distinctive features of collection.news discussed in the previous section, the digital preservation field could take specific actions to support these community initiatives:

1. Exigency: Authoritarian governments' crackdown on their targets is always unexpected. It is critical to plan ahead to collect and preserve the records and data before they vanish. While there is relatively sophisticated development for research data management cycles and digital preservation lifecycles, such as the DCC's Curation Lifecycle [6] and DPC's preservation Lifecycle [7], our field should consider developing standalone lifecycle frameworks for community digital preservation projects. These frameworks could help civil society actors, especially those in authoritarian regimes, to plan in advance and avoid abrupt crackdowns that leave little time for preservation efforts, just like in the case of collection.news.
2. Decentralization: Decentralization is an effective way to rapidly and widely disseminate censored data while mitigating political risks. However, without central management, the longevity and integrity of these digital assets remain uncontrollable and uncertain. To address this challenge, institutions from the free world could provide storage and techniques for parties to relocate their endangered digital materials. One example is Safe Havens for Archives at Risk Initiative [8], which is dedicated to providing support to organizations or individuals that need to deposit their records documenting human rights violations in reliable repositories.
3. Anonymity: Anonymity is crucial when it comes to conducting archiving initiatives in authoritarian regimes, as it ensures the safety of the initiative's facilitators. To support the facilitators of these community initiatives in circumventing state surveillance, more tutorials and technical assistance should be provided to teach them how to use encryption platforms and tools, such as Tor [9] and Session [10], for communication and operation without being detected by state authorities.

ACKNOWLEDGEMENTS

The author wishes to thank the guidance provided by Steve Marks, and, in particular, Jess Whyte from the University of Toronto Libraries along the way. Also, the author wants to salute the unsung heroes who took the initiative to preserve such precious records of Hong Kong’s history.

REFERENCES

[1] Amnesty International, “Hong Kong: Targeting of pro-democracy newspaper is threat to press freedom,” *Amnesty International*, Aug. 10, 2020. https://www.amnesty.org/en/latest/news/2020/08/hong-kong-targeting-of-pro-democracy-newspaper-is-threat-to-press-freedom/ (accessed Jun. 24, 2023).

[2] “[現場直播]我 backup 左蘋果四十萬篇文章 | LIHKG,” *LIHKG 討論區*, Jun. 24, 2021. https://lihkg.com/thread/2588517/page/1 (accessed Mar. 10, 2023).

[3] “話說我之前 backup 左蘋果四十萬篇文章想大家幫手 download | LIHKG,” *LIHKG 討論區*, Jun. 26, 2021. https://lihkg.com/thread/2591598/page/1 (accessed Mar. 10, 2023).

[4] appledailyarchive, “collection-news/appledaily-archive-directory: 蘋果日報文字備份目錄,” *GitHub*, Nov. 29, 2021. https://github.com/collection-news/appledaily-archive-directory (accessed Mar. 10, 2023).

[5] A. Acker and L. Flamm, “COVID-19 Community Archives and the Platformization of Digital Cultural Memory,” presented at the Hawaii International Conference on System Sciences, Jan. 2021. doi: 10.24251/HICSS.2021.312.

[6] Digital Curation Centre, “Curation Lifecycle Model,” *Digital Curation Centre*. https://www.dcc.ac.uk/guidance/curation-lifecycle-model (accessed Mar. 10, 2023).

[7] Digital Preservation Coalition, “Preservation Lifecycle,” *Digital Preservation Coalition*. https://www.dpconline.org/digipres/tags/preservation-lifecycle (accessed Mar. 10, 2023).

[8] Safe Havens for Archives at Risk Initiative, “Safe Havens for Archives at Risk Initiative,” *Safe Havens for Archives at Risk Initiative*. https://safehavensforarchives.org/en/about-the-initiative/ (accessed Mar. 10, 2023).

[9] The Tor Project, “The Tor Project | Privacy & Freedom Online.” https://torproject.org (accessed Mar. 10, 2023).

[10] Session, “Session | Send Messages, Not Metadata. | Private Messenger,” *Session*. https://getsession.org/ (accessed Mar. 10, 2023).