ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202304.00141

Web Information Archiving Architecture for Major Public Health Emergencies: Postprint of Lessons from the COVID-19 Pandemic

Authors: Zhou Wenhong, Su Yiwen, Wu Qiong, Huang Xiaoyu, Zhang Xiaoyu, Wen Lijun, He Tantao

Date: 2023-04-01T16:15:58+00:00

Abstract

[Purpose/Significance] This study conducts archival research on network information related to major public health events, aiming to explore theories and methods of network information archiving from multiple dimensions. [Method/Process] Based on a reference framework summarized from archival practices of network information for representative major social events, and combined with the specific information context of the COVID-19 pandemic, it comprehensively constructs an archival architecture for network information of major public health events. [Results/Conclusion] The fundamental content of the architecture includes: on the one hand, elaborating on the archival subjects, archival information objects, institutional and technical safeguards, archival schemes, web archives, and archival results composed of products and services for network information of major public health events; on the other hand, clarifying the overall requirements for archiving, which are manifested as implementing dynamically optimized schemes based on overall objectives, and achieving top-level design and coordination from a national perspective, characterized by whole-process control before events and whole-process collaboration among diverse social stakeholders.

Full Text

Preamble

Volume 64, Issue 15, August 2020

ChinaXiv Partner Journal

An Archival Architecture for Web-Based Information on Major Public Health Events: Insights from the COVID-19 Pandemic

ChinaRxiv 中X

Zhou Wenhong, Su Yiwen, Wu Qiong, Huang Xiaoyu, Zhang Xiaoyu, Wen Lijun, He Tantao

School of Public Administration, Sichuan University, Chengdu 610064

Abstract: *[Purpose/Significance]* This study examines the archiving of web-based information from major public health events, aiming to explore multi-dimensional theories and methods for web archiving. *[Method/Process]* Based on a reference framework derived from representative archiving practices for major social events and combined with the specific information context of the COVID-19 pandemic, this paper constructs a comprehensive archival architecture for web-based information on major public health events. *[Result/Conclusion]* The architecture comprises two fundamental components: First, it delineates the archival outcomes consisting of archiving subjects, information objects, institutional and technical safeguards, archival schemes, web archives, and associated products and services. Second, it establishes overarching requirements that emphasize dynamic optimization based on holistic objectives, pre-emptive whole-process control, and whole-process multi-stakeholder collaboration under a national-level top-down design and coordination framework.

Keywords: major public health events; major social events; COVID-19 pandemic; web-based information archiving; archival architecture

Classification Number: G270

DOI: 10.13266/j.issn.0252-3116.2020.15.023

With internet platforms hosting hundreds of millions of users, the web has become a venue for activities, dissemination, and practice across all sectors worldwide and within regions. It also serves as an integrated space for real-time information release and exchange during major social events in China, aggregating diverse perspectives from official institutions, mainstream media, stakeholder communities, and individual citizens. Beyond information management research exploring how various stakeholders should positively utilize online platforms—addressing topics such as emergency intelligence, information behavior, and post-truth dissemination—the value and handling of information recorded online after events conclude merit deeper consideration. In other words, against the backdrop of China' s explicit initiative to implement internet information preservation projects for national digital memory construction, how these vast, fragmented online records can reconstruct multi-dimensional processes of social pandemic response from historical narrative and information resource industry perspectives, and how these records can be sustained as national memory resources after dissemination, constitute important research questions. These questions can be concretized as the problem of how to archive web-based information on major public health events, using the COVID-19 pandemic as a case study.

At the practical level, institutions such as archives, libraries, and museums—entrusted with preserving national information resources and cultural heritage—

have initiated nationwide material collection campaigns. For instance, COVID-19 pandemic-related material collection is currently underway across China. However, most collection announcements fail to explicitly address the enormous scale of digital records, leaving critical questions unanswered: Should these web-based records be archived? How should they be archived? Who should archive them? What content should be archived? All lack comprehensive planning. Meanwhile, relevant experience in archiving major social events has accumulated both domestically and internationally, with libraries, archives, and global web archiving alliances participating and documenting activities across various domains, such as the European refugee crisis, Canadian federal elections, and the London Olympics. These practices demonstrate fundamental archiving concepts and methods while revealing limitations in scope definition, tools, legal and ethical risks, and resources—highlighting the compound challenges of web archiving for major social events across judicial, technical, social, cultural, and management dimensions.

At the theoretical level, discussions on archiving web-based information from major public health events have two primary orientations: First, using major social events to demonstrate the necessity and value of web archiving; second, analyzing challenges and strategies for event-based web archiving through case studies like the London Olympics. Simultaneously, research focusing on web or social media information archiving provides foundational analysis and methodological guidance for general issues, clarifying basic elements of web archiving: the necessity of archiving, compliance bases, and institutional frameworks for stakeholder rights and responsibilities; stakeholder mechanisms, including archiving actors and their division of labor; processes covering design of collection scope/strategies, capture, storage and organization, disposal, quality assurance, and utilization; and technologies, including tools and platforms required for implementation such as capture tools and retrieval algorithms. However, archiving web-based information from major public health events differs from website or account archiving centered on "subjects" in terms of information objects and methods. Moreover, as social media information constitutes a larger proportion, earlier research focusing primarily on government platforms offers limited reference. Studies on social media archiving propose few solutions at the specific "event" level for archiving major public health events and lack examples from the public health domain.

Therefore, this study summarizes representative web archiving practices for major social events to develop a reference framework for public health events, while examining the archiving structure for China' s major public health events through the lens of the COVID-19 pandemic' s information context.

2. A Reference Framework for Web-Based Information Archiving of Major Social Events: Based on Representative Cases

Through search engine queries, surveys of archives and library websites, and the International Internet Preservation Consortium (IIPC) website, we identified representative web archiving projects in the United States, Canada, the United Kingdom, Germany, Denmark, Ukraine, China, Singapore, Malaysia, Australia, and other countries. These projects have created digital memory resource repositories for major social events across political, cultural, sports, and economic domains, covering web-based information generated by official organizations, mainstream institutions, social groups, and individuals. From these achievements and challenges, we derive a reference framework for web archiving of major social events.

To develop this summary reference framework, our methodology proceeds as follows: First, we convert online research findings on project backgrounds, content, and outcomes into specific project descriptions. Second, drawing on archival science' s principle of provenance to clarify organic relationships among information from the same event, we classify and extract project content according to web archiving fundamentals—subjects, objects, methods, and processes (including capture, integration, preservation, and publication)—identifying five core elements and their sub-elements: archiving subject, web-based information object, archiving scheme, archiving outcome, and archiving safeguards. Third, by integrating content across projects under different elements, we specify each element' s manifestation in major social event contexts and summarize framework essentials.

2.1 Representative Case Categories

Current web archiving projects with available outcomes and documented practices for major social events reveal several representative approaches based on the scale of archiving subjects:

2.1.1 International Collaboration-Led Archiving These practices emphasize obtaining shared methodologies, resources, and tools through international cooperation, forming global capacity to archive transnational major events. The IIPC Content Development Group (IIPC CDG) exemplifies this approach: institutions from 45 countries and regions, including libraries and archives, collaborate to build publicly accessible, cross-national, multilingual, multi-perspective web archive collections on major international social events. IIPC CDG' s collaborative collection themes include the Olympics, Paralympics, European refugee crisis, and climate change—events with international impact. Targeted web information forms are diverse, encompassing websites, articles, news reports, blogs, and social media platforms like Facebook and Twitter. The final collections cover web content from multiple countries and languages;

for instance, the 2016 Rio Olympics collection encompassed 125 countries and recorded 34 languages. The project employs public nomination and citizen decision-making for collection and preservation, with IIPC member institutions and the public contributing recommended website lists after open web meetings. The integration process includes quality control and organizational description through crawler tool reporting and manual review, with detailed metadata cataloging. The resulting major social event web archives enter subsequent research programs—for example, IIPC CDG extracted gender data from Olympic web archives to study gender distribution among participants. Throughout this process, IIPC CDG primarily utilizes third-party collaborative development tools and platform services, such as the URL nomination tool developed by the University of North Texas for its 2010-2014 Olympics collection, with publication and preservation relying on Internet Archive' s Archive-It platform services.

2.1.2 National-Level Archiving These practices adopt a national perspective to capture, preserve, and provide access to web-based information on major social events that generate national public opinion and possess research value, emphasizing leadership by national memory institutions with legal deposit obligations. The UK Web Archive (UKWA) project by the British Library and the UK Government Web Archive (UKGWA) project by The National Archives exemplify this approach in their archiving of the 2012 London Olympics. The UK constructed web-based information repositories on the Olympics for two reasons: first, to document the grandeur of the 2012 London Olympics from a national perspective and construct Olympic national memory for British citizens; second, to preserve memory resources about the Olympics for people worldwide. Captured web information covers Olympic history, competition reports, typical cases, and exemplary stories, providing comprehensive resources for the public. Operationally, UKWA's Olympic archiving first determined capture themes and scope, then integrated and published them directly under established themes. The projects proceeded smoothly due to institutional safeguards regarding stakeholder authority, resource acquisition, and access. Technically, both currently use open-source tools and enterprise services to capture Olympics-related content, though archiving technologies for interactive content like reposts remain under development.

2.1.3 Regional or Institutional Archiving These practices archive web-based information on events significantly impacting regions or institutions, led primarily by corresponding memory institutions or departments. Based on regional perspectives, institutional functional activities, and user perspectives, these practices select events for archiving across broad platforms. The Bibliothèque et Archives nationales du Québec (BAnQ) web archiving project serves as an example: BAnQ considers web archiving part of its mission to preserve Quebec' s published heritage materials, using web archives to support the retention of Quebec' s historical and contemporary social contexts. Archiving objects include web information reflecting specific historical periods in Quebec society,

including major event-related content such as ecological accidents, social elections, and commemorative activities. BAnQ primarily employs manual capture for event-based web content collection, adopting different capture frequencies based on event intensity and information dissemination scope: "weekly capture" for periodic events like political elections and "daily capture"for sudden incidents like ecological accidents during initial outbreak phases. Captured information undergoes strict screening and integration across value, copyright, and quality dimensions before cloud storage, with final publication as lists on the BAnQ portal providing thematic and URL search functionality. For technical and institutional safeguards, the project uses internationally recognized open-source software (Heritrix, OpenWayback) for archiving and replay, while BAnQ, as Canada' s legal deposit institution, possesses collection authority though web archiving publication still requires permission from information producers.

2.1.4 Community-Level Archiving These practices select events important for community development, collecting members' event-related web records with unified formal standards but rich, inclusive community and individual characteristics in content. The "2020 COVID-19 Pandemic Individual Stories" project on GitHub exemplifies this approach, with individual GitHub users as archiving subjects. In terms of objectives, this practice represents ordinary Chinese citizens' personal life record backups during the pandemic. The project selects Douban platform weblogs as archiving objects, with most records from Wuhan, Hubei, reflecting real lives of ordinary Wuhan citizens during the anti-pandemic period and constructing precious community network memory. The project adopts a collaborative citizen strategy, with participants capturing diary text and images from Douban diary pages according to uniform formats and procedures, providing basic descriptions when merging files with cataloging elements including republication authorization, author, source, and publication time. Archived content is publicly released on the GitHub platform.

2.2 Framework Discovery and Analysis

Based on inductive analysis of existing major social event web archiving projects, a relatively universal action framework emerges (see Figure 1 [Figure 1: see original paper]), with specific components as follows:

2.2.1 Framework Content: Archival Elements and Their Relationships Practical experience reveals that a relatively complete web archiving framework for major social events includes these elements and functions:

(1) **Archiving Subject**: Stakeholders in major social event information, primarily comprising initiators and collaborators. Initiators serve as dominant actors throughout the entire process from advocacy, initiation, coordination, and design to implementation. In practice, initiators are mainly of two types: traditional memory institutions (archives, libraries, museums) as professional information custodians with supporting resources;

and socially conscious third parties such as information generators, disseminators, users, and social organizations or individuals providing resources and technical support who benefit from participation. As archiving subjects, initiators and collaborators jointly identify major social events and their web-based information, deepen understanding of information formation, value, and characteristics, form basic cognition of archiving needs and actions, establish organizational mechanisms, and clarify archiving objectives and schemes.

(2) **Web-Based Information Object**: Relevant records of major social events on web platforms. These constitute both the source of archived information and the target of archiving actions, documenting events according to their nature, scope, and depth. They influence basic information circumstances including quantity, format, and content, reflecting different values and characteristics that affect determinations of archiving necessity, extent, required resources, and specific schemes. More importantly, comprehensive investigation helps 梳理 and refine the main threads and dimensions of major social events, providing a foundation for understanding information and overall relationships from contextual layers.

(3) **Archiving Safeguards**: Systems and technologies that guide and implement archiving schemes throughout the process. Institutionally, this encompasses rule sets guiding major social event web archiving, helping formulate archiving schemes. Given that major social event web information often features cross-contextual, multi-producer characteristics with difficult-to-delineate information boundaries and rights/responsibilities, comprehensive archiving guidance is essential. Institutional functions manifest in two aspects: first, providing compliance bases for archiving necessity and risk avoidance, such as information ownership, usage rights, privacy, and intellectual property; second, offering action guidance by implementing overall models, processes, and activity details from theoretical, methodological, principle, and normative perspectives. For example, virtually all archiving projects establish deletion mechanisms allowing users to contact administrators to remove web archives containing privacy-violating content.

The other safeguard is **technology**, a critical enabler for implementing archiving schemes. Web archiving from scheme to implementation urgently requires supporting technology applications, from capture to publication and even development and utilization. Archiving subjects can independently develop technology, seek "outsourced" support from specialized fields, or adapt shared open-source technologies.

(4) **Archiving Scheme**: The overall configuration implementing capture, integration, preservation, and publication processes. Archiving scheme design is a procedural activity encompassing confirmation of archiving scope, capture frequency and methods, information preservation formats, quality standards, and publication conditions. Notably, integration is the pro-

cess of organizing web information, including organization, description, appraisal, and other specific steps. Given web information' s open characteristics, maintaining compliant "online visibility" is also essential archival content. Archiving schemes should be dynamically adjusted based on implementation effectiveness and feedback, which in turn drives institutional improvement and technical optimization.

(5) **Archiving Outcomes**: First, **web archives**—the direct results of archiving scheme implementation and integrated information resources retaining major social event memory and evidence. Their content and form directly reflect archiving scheme effectiveness. Second, **web archive products and services**. Beyond providing raw materials of major social event web information, archiving subjects can gradually develop higher "information density" products and services through deep development.

2.2.2 Framework Essentials To connect elements into a complete major social event web archiving framework and ensure its effective application, basic strategies and methods must permeate all aspects:

(1) **Participation and Collaboration as Fundamental Organizational Principles**: First, strong leadership is essential for archiving major social event web information, requiring several conditions: archiving consciousness and "volunteerism" —recognizing the value of major social event web information and willingness to act (e.g., the UK established the UK Web Archive in 2004 to address potential "digital black holes," aiming to collect, access, and preserve academically and culturally significant UK domain web resources); archiving capacity—the ability to coordinate and execute archiving from professional schemes with comprehensive management, technical, and humanities expertise; archiving resources—the capacity to allocate corresponding talent, funding, and infrastructure; and archiving responsibility—institutions, groups, or individuals with legal or socially recognized mandates to preserve social memory resources can achieve advocacy and action coordination. Second, major social event web information often forms across platforms, groups, formats, and even nations, with complex dissemination paths in social and content networks, extensive stakeholder coverage for information rights and responsibilities, and compound professional demands requiring multi-stakeholder participation. For example, IIPC' s content nomination mechanism balances CDG leadership with broad public participation, leveraging social forces to obtain extensive information sources while ensuring overall coordination for social memory construction. Similarly, expanding tool applications from independent development to more open-source or partially paid tools alleviates technical pressure on archiving institutions.

(2) **Deepening Cognition of "Limited Completeness" in Archiving Outcomes**: Archives created from major social event web information represent only a portion of the whole. Some information has limited value

or constitutes noise, while some is unsuitable for inclusion in publicly accessible resource collections. Furthermore, technical and resource constraints make comprehensive archiving impossible. As Duke University Archives' collection policy notes, certain social media platforms are difficult or impossible to collect due to service structures, and continuous collection of all materials cannot be guaranteed as social media services constantly evolve. Therefore, interim archiving results do not equal final archiving scheme objectives, and each archiving scheme component requires meticulous configuration. For example, capture must determine which information has higher priority based on value or difficulty within the archiving scope, as does integration and publication—information captured in the same batch may have inconsistent integration progress and publication timing due to various reasons. In summary, archiving results must tolerate "completeness" limitations. For instance, the UK Government Web Archive (UKGWA) project explicitly states that due to technical limitations, currently archivable social media information excludes Facebook platforms and interactive content (reposts, comments, etc.) across all platforms, making current event depictions incomplete. Future decisions on whether and how to conduct web archiving for major social events will consider government needs and resource realities.

(3) **Clarifying Archiving as a Nonlinear Continuous Process**: The recording breadth and depth of major social event web information pose challenges from resources to specific actions across social, cultural, technical, management, and judicial dimensions, making linear, one-time archiving impossible. Instead, archiving requires dynamic adjustment and improvement with actions before, during, and after events. **Pre-event archiving** focuses on establishing routine mechanisms for major social event web information, providing universal framework foundations for targeted schemes. Projects like BAnQ, Australia' s Pandora Archive, and UKWA have established relatively complete routine web archiving frameworks as guiding documents. **During-event archiving** ensures timeliness, addressing web information' s vulnerability to loss and distortion, requiring archiving to address increasingly complex information associations in network environments. **Post-event archiving** targets information requiring long-term or tracking capture, with subsequent processing, management, and publication also being time-consuming. For example, the Library of Congress' s web archiving of Brazil' s 2010 presidential election ran from September 2010 to January 2011, capturing pre-election warm-up messages and post-election public reactions and follow-up reports to more completely present actual conditions. Additionally, archiving scheme design and implementation during events reveal deficiencies, even in information source discovery requiring post-event improvement.

(4) **Parallel Institutional and Technical Improvement**: The workload and complexity of major social event web archiving are fully demonstrated in practice, with no current archiving outcome serving as a complete bench-

mark. First, information ownership issues require legal perspectives, while complex information association analysis requires archival science to further expand connotations of "context" and "organic relationships." These must be translated to institutional settings to ensure scientific archiving schemes. Second, facing massive and complex archiving activities, broad public participation alone cannot meet workload demands, and information quality review and value determination from mass participation also entail substantial work. For example, current COVID-19 pandemic crowdsourcing projects on GitHub have limited effectiveness, partly because information authenticity verification alone consumes substantial time. Therefore, comprehensive technology application in archiving has become an indispensable strategy. For instance, IIPC recommends 84 available tools from capture to publication.

(5) **Event-Centered Archiving Units that Grasp "Event" Characteristics**: Compared with archiving practices centered on creating subjects, major social event web archiving emphasizes the "event" dimension from information formation understanding to archiving action. This requires information object investigation to extend beyond information producers, starting from the event itself to associate relevant subjects, scenarios, activities, and evolutionary contexts. In essence, the question is which aspects and progressions of events should be recorded, with different recording subjects representing only one archival dimension. For example, although some web archiving results display different events, limited event perspectives during archiving lead to scattered, poorly associated information and insufficient event reconstruction. The Webarchiv project' s final 成果 is limited to publicly accessible URL lists arranged alphabetically, lacking deeper organization and failing to vividly present more event facets.

3. The Web-Based Information Context of Major Public Health Events in China: The COVID-19 Case

Major public health events possess enormous social influence due to their relevance to human survival needs. Their broad attention base, amplified by today' s rich web platforms, creates massive and complex information dissemination phenomena. Using the COVID-19 pandemic that gradually gained attention in late 2019 as an example, several basic characteristics emerge:

(1) **Web-based information on major public health events like COVID-19 primarily forms on two platform types**: First, self-built websites, mainly from institutional actors such as party and government organs, news media, and other large-scale social organizations. Party and government websites at all levels and systems, along with various news platforms, have established pandemic information columns covering daily outbreak notifications, prevention and control news, announcements, pandemic knowledge 普及, and medical worker profiles. This authoritatively released information has strong reference value. CCTV, Xinhua, and

People' s Daily have all produced special reports and even live broadcasts. Second, third-party platforms supported by network service providers such as Weibo, WeChat, Toutiao, Douyin, Kuaishou, Douban, and Zhihu. Based on social media characteristics, broad user groups have created vast and diverse pandemic-related web information. In terms of subjects, these include both official accounts from organizations at all levels and numerous individual accounts closely related to the pandemic, spanning medical personnel, volunteers, and ordinary citizens. Content-wise, web information created by these broad subjects reflects all pandemic aspects. For example, Sina Weibo—the most widely used social media platform during the pandemic—shows pandemic-related topics generating billions of views and tens of millions of discussions: "Novel Coronavirus Pneumonia Super Topic," "Pneumonia Patient Assistance Super Topic," "Quarantine Diary," "Frontline Anti-Pandemic," "Wuhan Jiayou!" and daily hot search topics. A dedicated "Wuhan Diary" section on the discovery page records various life aspects during the pandemic.

(2) **High proportion of original records**: The pandemic has imperceptibly accelerated paperless operations across sectors, with web space' s cross-temporal and diversified attributes fully realized. From government websites to social media to online office tools, different functions are performed, making pandemic web information not only format-diverse but also with original record content comprising important components. For example, medical personnel recorded first-hand anti-pandemic actions through video, images, and text on social media like Douyin, Kuaishou, and Weibo. Remote work became an important mode for pandemic decision-making and action arrangements, with platforms like DingTalk and Tencent Meeting providing live replay functions, enabling complete process recording of pandemic-related official activities in web space. These important "business" records are not converted from other forms or migrated from other facilities but are directly generated and recorded on web platforms.

(3) **Enhanced "co-creator" characteristics across subjects**: From the pandemic' s social impact perspective, co-creators number in the hundreds of millions at the macro level—unprecedented in both generator scale and duration. From infected individuals and medical personnel to decision-makers, grassroots actors, cross-sector collaborators, and the public, all contributed different aspects of effort, forming different records around the pandemic on web space due to business needs, information disclosure, and public participation factors, also creating information from respective perspectives within the same activities. For example, Sina Weibo' s COVID-19 super topic served as an important channel for potential infections seeking rescue, with over 1,000 help posts and tens to thousands of reposts or comments under each post collectively forming partial records of assistance and rescue.

(4) **Content mapping social conditions beyond the pandemic itself**:

As a national coordinated crisis response in China, from central to local levels, from decision-making to grassroots, from health departments to all sectors, and from official to public spheres, comprehensive action formed ubiquitous pandemic-related web information. This information extends beyond specific health department business and activities to include all social activities for pandemic prevention and response. Thus, viewing the information holistically reveals a recording system interwoven around various activities and involved subjects. Content covers diverse pandemic-related aspects across different times and spaces.

(5) **Concurrent value and challenges**: In terms of value, as a crucial pandemic response battlefield, web space has generated large-scale, broad-scope, diverse-content information with high evidentiary, memory, resource, and asset value. It can contribute to accumulating disease control experience, optimizing health systems, building emergency management systems, demonstrating and advancing China's governance modernization, telling China' s story, and promoting national digital memory construction. The system' s enormous scale—over 100 hot search topics and billion-level discussion volumes on Sina Weibo alone—also provides a testbed for archival socialization practices across platforms, subjects, formats, and content.

In terms of challenges, where should massive web information obtain supporting archival resources? Which subjects should and can archive this information? How are information rights and obligations identified and delineated? How are information noise, low-value information, duplicate information, or illegal account postings identified? How are associations formed around pandemic activities and practical subjects fully revealed? How are archiving schemes designed, archiving scope set, information value determined, and archiving action priorities established? What legal and ethical risks exist and how should they be addressed? These challenges require first resolving a fundamental question: What is the web archiving architecture for major public health security events? That is, what constructions exist in the archiving rule system, what is missing, and how can cross-platform, cross-subject, cross-format, cross-content web information have a relatively unified archiving framework? This determines which elements corresponding archiving actions should address and provides a basic discussion scope for solving these problems.

4. A Web Archiving Architecture for Major Public Health Events in China

How should web-based information on major public health events be archived in China? First, China has certain policy and practical foundations for web archiving. As an IIPC member, the National Library of China has long-standing exploration experience, leading internet information social preservation projects for national digital memory. On April 22, the National Library launched a "War Against Epidemic" memory repository covering digital information. Un-

der the "Guidelines for Web Archiving of Government Websites" issued by the National Archives Administration in 2019, local archives and website organizers are undertaking archiving actions. Additionally, major social event web archiving practices worldwide in Canada, the UK, and other countries—though with limited publicly available public health event outcomes—offer reference value within the major social event category for political, sports, and social disaster events. Meanwhile, the COVID-19 pandemic provides a specific information context. Therefore, based on China' s national conditions and the previous major social event web archiving framework, and focusing on resolving current limitations in basic schemes, we should first confirm the elements and overall requirements in a public health event web archiving architecture.

4.1 Basic Settings for Archival Elements

4.1.1 Archiving Subjects First, archives, libraries, and museums currently entrusted with safeguarding information assets should play leading roles. These institutions across China are currently collecting various pandemic materials from society, encouraging participation at different levels, which should extend to web-based information objects. The National Library' s "War Against Epidemic"memory repository project explicitly proposes incorporating web information into capture scope and advocates collaboration among memory institutions and social forces at all levels. Thus, under current decentralized actions, COVID-19 web archiving can become a space for libraries, archives, and museums to establish a nationwide networked system for co-constructing and sharing archival resources. Second, from government websites to various third-party platforms, network service providers—especially social media service providers that primarily share information through APIs—should become important archiving assistants and resource providers. Official websites of emergency management and health departments, as well as news media and social media platforms playing crucial dissemination roles, should be incorporated into collaborative systems through social advocacy and official coordination. Third, third-party social organizations, particularly academic groups, technical organizations, and enterprises providing archiving infrastructure, can become important supplements. Finally, the public should become a continuous "supply" force, both as recorders and recorded subjects with certain rights to decide information retention and deletion, and as collaborative participants whose collective wisdom, resources, and "workforce" can achieve what single parties cannot.

4.1.2 Web-Based Information Objects Comprehensive investigation and 梳理 are required. Full-scale surveys should grasp web information formation platforms, accounts, quantities, and growth patterns. By cross-referencing COVID-19 with information, we should identify which information aspects correspond to pandemic-related activities, conducting association 梳理 from subject, platform, and content perspectives to form an information mapping 脉络 oriented to the pandemic as a major social event. For example, from the perspective of social activities behind information, analysis should unfold from health

and emergency response domains, with auxiliary peripheral domains mapped according to proximity to the pandemic core. Proximity assessments should be conducted across decision-making, business, and functional support activities in various domains. Meanwhile, given the pandemic' s uncertain timeline, archiving actions often commence after preliminary event 脉络梳理. To avoid archiving difficulties from information accumulation during event development, archiving subjects should track fragmented themes and capture information during the event to ensure complete background description, facilitating later-stage or post-event organization.

4.1.3 Archiving Safeguards **Institutionally**, China currently has foundational systems from laws to standards, regulations to departmental rules, including the Cybersecurity Law, NPC Standing Committee Decision on Strengthening Network Information Protection, Provisions on Ecological Governance of Network Information Content, and GB/T 33994-2017 Information and Documentation—WARC Format. However, improvements are needed regarding information rights confirmation, archiving compliance, long-term preservation, and information leakage prevention. More specific guidelines require dominant institutions like archives, libraries, and museums to provide relatively unified standards, addressing archiving terminology, methods, scope, processes, and multi-stakeholder collaboration mechanisms.

Technically, support can be obtained from three sources: First, adapt and improve open-source shared technologies—IIPC shares diverse full-process supporting tools for selection based on characteristics, and technical open-source platforms like GitHub also offer reference tools and potential collaborators for common problems. Second, utilize paid archiving tools like Archive-It provided by Internet Archive to countries and regions. Third, independently develop or jointly develop with other social forces. Technology should not only serve all archiving process activities such as automatic information identification, authenticity appraisal, capture, organization, and description, but also build networked platforms for multi-stakeholder collaboration that enable interconnection, mutual trust, and communication to present social participation outcomes holistically.

4.1.4 Archiving Schemes Archiving schemes must ensure complete configuration and standardized integration across information investigation, capture, integration, preservation, and publication, with executable specific standards and bases such as archiving scope, capture frequency and methods, information preservation formats, quality standards, and publication conditions—drawing reference from existing web archiving practices. For public health events like COVID-19, special attention should focus on: First, **archiving scope setting**. Facing the pandemic' s extensive, deep, and complex records, comprehensive archiving is unrealistic, requiring balancing of diverse formation subjects, platforms, formats, and content behind information. Actors must conduct detailed classification, value determination, and action priority setting, potentially in-

vestigating existing projects or negotiating shared captures. Second, archiving must not only identify which web information objects fall within archival scope but also clarify which information should be excluded due to quality, value, or stakeholder rights protection (e.g., privacy) and receive appropriate disposal—pandemic-related personal privacy information should be effectively protected. Third, archiving schemes must be detailed to ensure complete mapping of technical requirement lists. Fourth, schemes must allow flexibility—pandemic development is uncertain, and random long-term archiving should ensure adjustment possibilities, with optimization necessary as archiving experience accumulates.

4.1.5 Archiving Outcomes The COVID-19 web archive repository is the direct archiving outcome, positioned as: comprehensive, open, and mappable. First, it should comprehensively present the full action picture across all levels, domains, and platforms throughout the pandemic process. Second, openness is essential for web archiving—online access should be provided as open data standards where technically feasible and compliant with stakeholder rights protection. Third, the repository should map the pandemic, providing not just archived information but multidimensional classification results enabling different 脉络 presentations of COVID-19. In developing web archive products and services, the COVID-19 repository should build corresponding association platforms—not just resource preservation and basic access channels, but also functioning as a "memorial hall" and "laboratory," providing processed outcomes to users while collaborating with them to deeply mine information assets into products serving diverse social needs, maximizing social value.

4.2 Overall Archiving Requirements: Strengthening Top-Level Design from a National Perspective

Web-based information from public health events like COVID-19 holds major value for the nation, with archiving complexity requiring coordinated planning from the national strategic level across resources, organizational structure, and action guidelines. However, beyond scattered social archiving actions like Huazhong University of Science and Technology's "War Against Epidemic" Digital Museum and technically specialized individuals' projects on GitHub documenting ordinary Chinese lives during the pandemic, national-level projects proposing larger-scale, multi-perspective, richer content are limited to the National Library's China "War Against Epidemic" Memory Repository launched on April 22, 2020. This project explicitly includes new technologies and social management forms like digital life, online work, and distance learning in its capture scope. However, compared with IIPC's 5,000 archivable web resources with online retrieval and website access permission, China's actions show time lags and limited social participation. Meanwhile, implementing archiving elements across contexts, subjects, time-space, and objects as holistic actions requires effective coordination mechanisms. Thus, national-level top-level design and spatio-temporal coordination demonstrate necessity through:

(1) **Implementing Dynamic Optimization Schemes Based on Holistic Objectives**: Given COVID-19 web information' s extensive scope, recording depth, and complex associations, archiving requires both detailed sub-objectives for different participants and strategic direction toward overall national digital memory construction goals. Whether sub-objectives or overarching goals, archiving schemes cannot achieve matching schemes instantly—they must set process-oriented schemes adapted to different stages, archiving conditions, and requirements for different participants, implementing archiving actions progressively. This means scheme completeness is relative to specific time-space contexts, not absolute, and allows "compromised" deficiencies constrained by objective conditions and capabilities. For example, currently it is difficult to write different format data into .md files while maintaining complete correspondence with original records, so "distributed" capture prioritizing element completeness should be permitted.

(2) **Achieving Pre-emptive Whole-Process Control**: Web information forms and spreads rapidly through complex networks. To prevent disorderly accumulation and random distortion or loss after dissemination, comprehensive deployment planning should begin as early as possible. With information continuing to grow while the pandemic remains unresolved, archiving frameworks and schemes must be formulated simultaneously. For this pandemic' s relatively lagging web archiving, experiences should be promptly summarized and incorporated into national memory institutions' business upgrades and institutional construction, building pre-emptive frameworks and general contingency plans for public health and major social event web archiving. The COVID-19 pandemic' s more complex recording subjects, quantities, content diversity, and platform multiplicity require maximally compact action despite lagging starts, potentially through pilot projects that expand into overall layout and panoramic action.

(3) **Whole-Process Reliance on Multi-Stakeholder Collaboration**: Web information complexity necessitates multi-stakeholder collaboration for systematic and large-scale archiving, including: memory institutions (archives, libraries, museums) safeguarding national information assets; core record-generating subjects like health and emergency management departments, hospitals and frontline medical staff, communities, and volunteers; third-party providers of technology and resources like media, academic groups, enterprises, and non-profits; web platform service providers with pandemic themes who have established information preservation mechanisms; and, under China' s institutional advantages, maximally encouraged public participation to transform this universally participated pandemic event into socially co-constructed, reconstructive archiving practice.

References

[1] Royal Library of Denmark. Netarkivet.dk [EB/OL]. [2020-02-20]. http://netarkivet.dk/in-english/.

[2] LAC. GCWA [EB/OL]. [2020-02-20]. http://webarchive.bac-lac.gc.ca/#c.

[3] THE BRITISH LIBRARY. UK Webarchive [EB/OL]. [2020-02-20]. https://www.webarchive.org.uk/en/ukwa/index.

[4] Ma Ningning, Qu Yunpeng, Xie Tian. Research and Enlightenment on Major Web Resource Collection Projects in Europe [J]. Library and Information Service, 2013, 57(12): 10-15.

[5] Qiu Zhuangli, Xu Dongling. Research on Influencing Factors of Archival Web Information Selection Strategy [J]. Archives Science Study, 2011(3): 63-66.

[6] Zhou Yi. On Web Archiving Right and Its Generation [J]. Journal of Library Science in China, 2011, 37(1): 102-108.

[7] Zhao Junling, Du Guofang. Analysis of Copyright Law Impact on Web Information Resource Preservation [J]. Modern Information, 2005(5): 72-74.

[8] VLASSEENROOT E, CHAMBERS S, DIPRETORPO E, et al. Web archives as a data resource for digital scholars [J]. International journal of digital humanities, 2019, 1(1): 85-111.

[9] Zhou Wenhong, Chen Yi, Zhang Yujie, et al. Case Analysis and Enlightenment of Web Archiving at The National Archives of the United Kingdom [J]. Archives Management, 2018(4): 4-7, 74.

[10] Zhao Junling, Lu Zhenbo. Analysis of Responsibility System for Web Information Preservation [J]. Journal of Academic Libraries, 2006(2): 94-97, 88.

[11] Wu Shuona, Huang Xinrong. Research on Development of Web Archiving Lifecycle Model [J]. Digital Library Forum, 2018(10): 41-45.

[12] Chen Qingwen. Collection Strategy and Method for Long-term Preservation of Web Information Resources [J]. Information Research, 2006(12): 47-48.

[13] Wang Wenling, Qu Yunpeng. Theoretical Framework Research on Web Archiving Data Quality Assurance Strategy [J]. Knowledge Management Forum, 2018, 3(2): 106-115.

[14] Ma Ningning, Qu Yunpeng. Research and Suggestions on Chinese and Foreign Web Resource Collection Information Service Methods [J]. Library and Information Service, 2014, 58(10): 85-89, 116.

[15] Tang Guangqian. Research on Automatic Acquisition and Identification Archiving Technology for Web Resources [J]. Modern Library and Information Technology, 2003(5): 57-61.

[16] GOSSEN G, RISSE T, DEMIDOVA E. Towards extracting event-centric collections from web archives [J]. International journal on digital libraries, 2018: 1-15.

[17] TAMIMENT LIBRARY OF NEW YORK UNIVERSITY. Special collections [EB/OL]. [2020-02-20]. https://specialcollections.library.nyu.edu/search/?f%5Brepository_sim%5D%5B%

[18] UNIVERSITY OF TEXAS AT AUSTIN LIBRARY. Webarchive [EB/OL]. [2020-02-20]. http://legacy.lib.utexas.edu/hr.

[19] DUKE UNIVERSITY ARCHIVES. Uawebarchive [EB/OL]. [2020-02-20]. https://library.duke.edu/rubenstein/findingaids/uawebarchive/#collectionoverview.

[20] NATIONAL LIBRARY OF THE CZECH REPUBLIC. Webarchiv [EB/OL]. [2020-02-20]. https://www.webarchiv.cz/en/.

[21] NATIONAL ELECTRONIC ARCHIVES OF UKRAINE. Collections [EB/OL]. [2020-02-20]. https://tsdea.archives.gov.ua/.

[22] DEUTSCHE NATIONALBIBLIOTHEK. Webarchive [EB/OL]. [2020-02-20]. https://www.dnb.de/EN/Professionell/Sammeln/Sammlung_Websites/sammlung_websites_node.htm#

[23] NATIONAL LIBRARY OF NEW ZEALAND. New Zealand webarchive [EB/OL]. [2020-02-20]. https://natlib.govt.nz/collections/a-z/new-zealand-web-archive?search%5Bpath%5D=items&search%5Btext%5D=web+archive.

[24] INTERNATIONAL INTERNET PRESERVATION CONSORTIUM. CDG collections [EB/OL]. [2020-02-20]. https://netpreserveblog.wordpress.com/?s=Helena+Byrne.

[25] JIAYI LIUJIAYI. 2020 nCov individual archives [EB/OL]. [2020-02-20]. https://github.com/jiayiliujiayi/2020nCov_individual_archives.

[26] NATIONAL ARCHIVES. Webarchive [EB/OL]. [2020-02-20]. http://www.nationalarchives.gov.uk/webarc

[27] BANQ. Services [EB/OL]. [2020-02-20]. https://www.banq.qc.ca/services/bibliotheque_nationale/depot_

[28] NATIONAL LIBRARY OF AUSTRALIA. Pandora overview [EB/OL]. [2020-02-20]. http://pandora.nla.gov.au/overview.html.

[29] LIBRARY OF CONGRESS. Web-archiving [EB/OL]. [2020-02-20]. https://www.loc.gov/programs/web-archiving/about-this-program/.

[30] Feng Huiling, He Jiasun. The Essence of Fonds Theory—Second Exploration of Fonds Theory [J]. Archives Science Bulletin, 1988(5): 8-11.

[31] Cai Na. Research on Major Event Archives Management Mechanism [J]. Archives Science Bulletin, 2012(3): 65-67.

[32] National Health Commission of the People' s Republic of China. Epidemic Notification [EB/OL]. [2020-02-20]. http://www.nhc.gov.cn/xcs/xxgzbd/gzbd_index.shtml.

[33] Xinhua Media Creative Workshop. Epidemic Data [EB/OL]. [2020-02-20]. http://fms.news.cn/swf/2020_sjxw/3_12_worldYQ/index.html?v=0.8829852148310229.

[34] People's Daily. Epidemic Reports [EB/OL]. [2020-02-20]. http://society.people.com.cn/GB/369130/431577

[35] Sina Weibo. #COVID-19 Pandemic# [EB/OL]. [2020-02-20]. https://s.weibo.com/weibo?q=%23%E6%96

[36] National Library of China. Announcement on Launching China "War Against Epidemic" Memory Repository Project and Collecting Anti-COVID-19 Theme Resources [EB/OL]. [2020-02-20]. http://www.nlc.cn/dsb_zx/zyjyqk/.

[37] DA/T 80-2019, Guidelines for Web Archiving of Government Websites [S]. Beijing: National Archives Administration, 2019.

[38] INTERNET ARCHIVE. Archive-it [EB/OL]. [2020-02-20]. https://www.archive-it.org/.

[39] Huazhong University of Science and Technology Press. Wuhan "War Against Epidemic"Digital Museum [EB/OL]. [2020-02-20]. https://loftehuati.lofter.com/tag/%E6%88%98%E6

Author Contributions:

Zhou Wenhong: Conceived research questions and framework, wrote and revised manuscript;

Su Yiwen: Identified and analyzed cases, wrote manuscript;

Wu Qiong: Identified cases, revised manuscript;

Huang Xiaoyu: Identified cases, revised manuscript;

Zhang Xiaoyu: Identified cases, revised manuscript;

Wen Lijun: Identified literature;

He Tantao: Identified literature.

Abstract: *[Purpose/significance]* Research on archiving web-based information from major public health events aims to explore multi-dimensional theories and methods for web archiving. *[Method/process]* Based on a reference framework derived from representative archiving practices for major social events and combined with the specific information context of the COVID-19 pandemic, this paper constructs a comprehensive archival architecture for web-based information on major public health events. *[Result/conclusion]* The architecture comprises two fundamental components: First, it delineates the archival outcomes consisting of archiving subjects, information objects, institutional and technical safeguards, archival schemes, web archives, and associated products and services. Second, it establishes overarching requirements that emphasize dynamic optimization based on holistic objectives, pre-emptive whole-process control, and whole-process multi-stakeholder collaboration under a national-level top-down design and coordination framework.

Keywords: major public health events; major social events; COVID-19 pandemic; web-based information archiving; archival architecture

Note: Figure translations are in progress. See original paper for figures.

ChinaRxiv 中X

Source: ChinaXiv —Machine translation. Verify with original.