

# 重大社会事件社交媒体存档平台文本集

## 文本清单

文本 1： 重大社会事件社交媒体存档平台作品介绍

文本 2： 重大社会事件社交媒体存档平台建设方案

文本 3： 重大社会事件社交媒体存档平台相关附件

文本 1:

## 重大社会事件社交媒体存档平台作品介绍



全国高校档案学专业  
大学生课外科技作品展

# 重大社会 事件社交 媒体存档 平台

团队成员：张晓宇

吴琼

贺谭涛

黄小宇

王淼

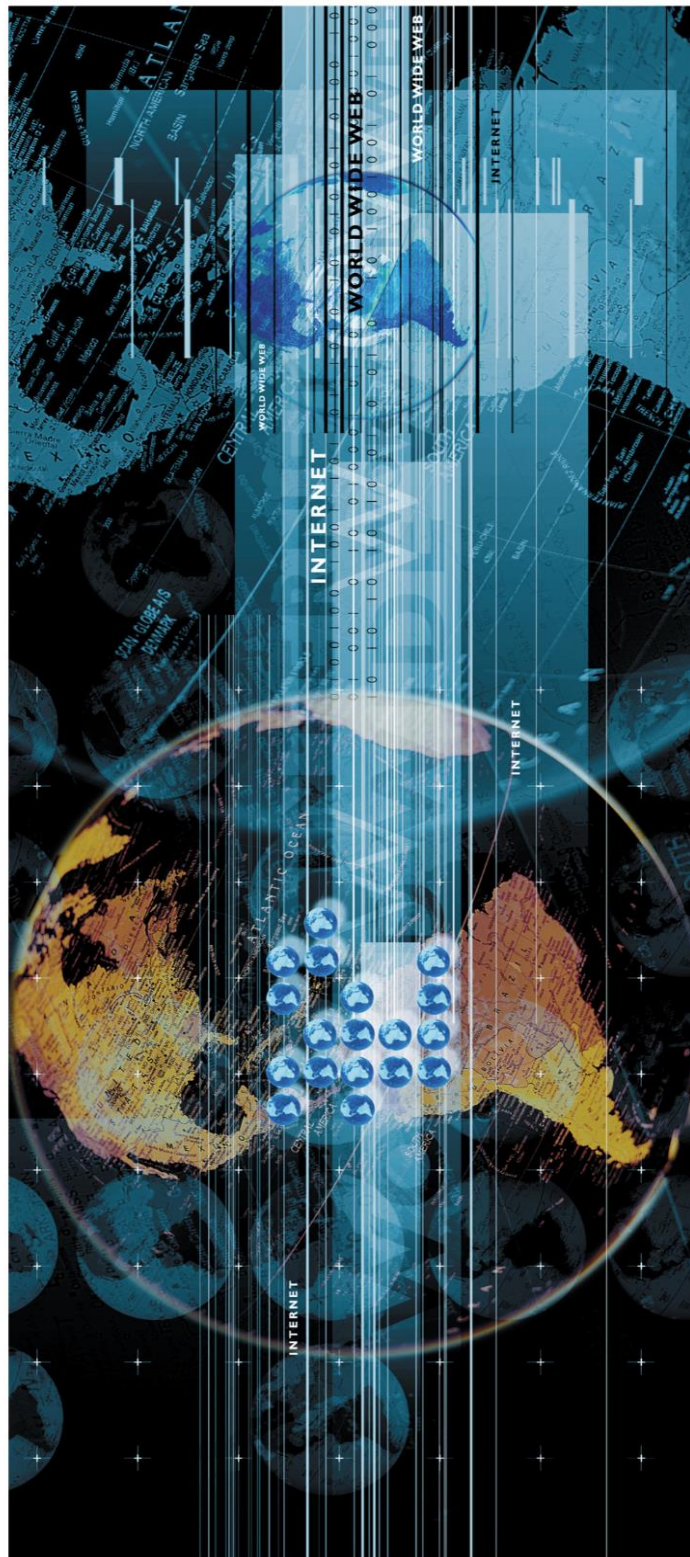
指导教师：

周文泓副研究员      (总指导)

乔健教授              (总顾问)

启动时间：2020 年 2 月

四川大学秘书档案学系



目 录

1 项目背景 ..... 1

    1.1 实践背景 ..... 1

    1.2 理论背景 ..... 2

2 项目特色及优势 ..... 5

    2.1 项目价值 ..... 8

    2.2 资源优势 ..... 9

    2.3 科技含量 ..... 9

    2.4 成果形式 ..... 9

3 作品主要内容 ..... 10

    3.1 作品基本信息 ..... 11

    3.2 资源与功能 ..... 13

    3.3 科研成果 ..... 34

4 运作模式 ..... 36

    4.1 STP 模式 ..... 36

    4.2 产品策略 ..... 37

    4.3 资金运转来源 ..... 39

5 结语 ..... 40

    5.1 现有不足 ..... 40

    5.2 未来发展方向 ..... 41

6 参考文献 ..... 43

# 1 项目背景

重大社会事件指对一个国家、地区或民族产生深远影响的事件，对塑造国家和民族的记忆与认同具有重要意义。档案的重要职能之一是保存真实证据，能最大限度保证事件的真实可靠，因而档案人应该也有义务将事件中的原始记录保存下来，由此当前世界范围内都在积极开展重大社会事件的存档项目。

从 2020 年初的新冠肺炎疫情，到即将在成都举办的 2021 年世界大学生运动会，再到档案领域备受关注的互联网信息社会化保存项目，不难发现社交媒体这个数字空间成为了展现和反映这些重大社会事件的重要组成部分。在此种现实情况下，本项目聚焦于国家、地区、行业以及档案领域的重大社会事件，以社交媒体为阵地，展开了对重大社会事件社交媒体信息存档<sup>\*</sup>的实践探索。

## 1.1 实践背景

一方面，当下世界范围内不乏着眼重大社会事件的网络存档项目，但社交媒体存档的实践甚少，且资源开发利用较为薄弱。数据类型上，仅有英国网络档案馆(UKWA)<sup>[1]</sup>、IIPC 奥运会/残奥会收藏<sup>[2]</sup>等少数项目包含社交媒体信息的归档。开发环节上，虽多数项目在归档流程构想较为完整，但信息资源的二次开发普遍较薄弱，现有实践主要有辅助研究和新形式的信息资源开发（如杜克大学的虚拟校园场景图鉴<sup>[3]</sup>）。开发对象上，多数数字记忆项目的主要对象为实体档案的数字化，原生网络信息的资源开发实践相对较少。

另一方面，面对众多的重大社会事件，当前档案工作实践主要以收集官方文件资料与媒体报道为主，缺少对社交媒体存档的考察。即使涉及到社交媒体存档，

---

<sup>\*</sup> 为简化提法，表述为重大社会事件社交媒体存档

<sup>[1]</sup> 大英图书馆.UK web archive[EB/OL]. <https://www.webarchive.org.uk/en/ukwa/index,2004-00-00>.

<sup>[2]</sup> 国际网络保存联盟内容开发工作组.IIPC 奥运会/残奥会收藏[EB/OL].<https://netpreserveblog.wordpress.com/?s=Helena+Byrne,2010-00-00>.

<sup>[3]</sup> 杜克大学.杜克大学档案馆网络档案收藏[EB/OL].<https://library.duke.edu/rubenstein/findingaids/uawebarchive/#collectionoverview,2010-00-00>.

归档对象也多限于政务型社交媒体。仅以此次新冠肺炎疫情中的归档情况为例,官方层面,各地档案馆的档案收集工作聚焦医院、疾控中心、医疗研究院所等部门的建档指导,重在疫情防控、诊疗、人员物资调配以及重点隔离区管控等工作所形成的文字、图表、照片、音像、实物的收集、建档工作,以及防控过程中形成的部署、决策、措施等相关纸质文件、电子文件、照片、录音、录像及防控过程中涌现的先进事迹和模范典型材料的收集。<sup>[4][5]</sup>民间层面上,志愿者自建的发布于 GitHub 的 nCovMemory 项目收录了疫情期间大量深度报道与非虚构作品的链接与截图。

## 1.2 理论背景

理论层面,重大社会事件存档一直以来都得到提倡<sup>[6][7]</sup>,而社交媒体存档也已非档案领域新论题。2010年,美国国会图书馆宣布与 Twitter 合作,将美国境内 IP 发布的所有推文予以存档,这一极具代表性的事件有效推动了各国国家社交媒体存档的实践与研究工作<sup>[8]</sup>,近年来我国国家层面也重点关注<sup>[9]</sup>,为保存人类活动的证据,留存社会记忆材料和信息资源做出了重要贡献<sup>[10]</sup>。随着研究的深入,来自国内外图书情报学、档案学、计算机科学等学科的学者从特定视角出发

---

<sup>[4]</sup> 浙江省档案局.浙江省档案局关于做好防控新型冠状病毒感染的肺炎疫情期间档案工作的通知[EB/OL].[http://www.zjda.gov.cn/art/2020/1/30/art\\_1378509\\_41963671.html](http://www.zjda.gov.cn/art/2020/1/30/art_1378509_41963671.html),2020-01-30.

<sup>[5]</sup> 江苏省档案局.关于面向社会征集抗击新冠肺炎疫情防控工作档案资料的公告[EB/OL].[http://www.dajis.gov.cn/art/2020/2/14/art\\_155\\_48850.html](http://www.dajis.gov.cn/art/2020/2/14/art_155_48850.html),2020-02-14.

<sup>[6]</sup> 黄新荣,曾萨.网页归档推进策略研究——基于网页归档生态系统视角[J].图书馆学研究,2018(16):63-70+16.

<sup>[7]</sup> 闫晓创.日本网络资源存档项目实践研究[J].浙江档案,2017(12):20-23.

<sup>[8]</sup> 黄新荣,高晨翔.国内外社交媒体存档研究与实践述评[J].图书情报工作,2019,63(04):122-134.

<sup>[9]</sup> 周文泓.全球社交媒体归档行动概览与展望[J].浙江档案,2016(12):16-19.

<sup>[10]</sup> 全国档案事业发展“十三五”规划纲要[J].中国档案,2016(05):14-17.

对社交媒体存档的必要性<sup>[11][12]</sup>、挑战<sup>[13][14][15]</sup>、优化策略<sup>[16][17][18][19]</sup>等得到更多探索<sup>[20]</sup>。此前的研究强调了社交媒体存档的具有证据、资产、资源、记忆等多方面价值，且在方法和技术层均有涉及。但却少有涉及事件型尤其是重大社会事件的社交媒体存档研究。基于此，如何通过框架设计、技术工具等多要素的体系来构想、开发和可持续应用重大事件社交媒体信息档案库于实践、于理论就尤为重要。

这样的理论与实践趋势同当前我国的国家发展战略规划和政策极为契合。

《全国档案事业发展“十三五”规划纲要》中提出：“要研究制定重要网页资源的采集和社交媒体文件的归档管理办法”、《国家档案局关于做好新型冠状病毒感染肺炎疫情防控期间档案工作的通知》中号召：“各级档案部门要充分利用现代通讯技术，灵活采用电话、办公网、短信、微信、移动客户端等适宜方式，做到疫情防控档案应收尽收、应归尽归。”此外，国家档案局也积极制定《重大活动和突发事件档案管理办法》，以加强重大活动和突发事件档案科学管理，确保档案齐全、完整、真实与安全。这表明在国家层面的战略规划和政策中，对社交媒体信息以及重大社会事件的归档工作的重要性得以明确。

此外，面对规模庞大的社交媒体信息，必须借助现代信息技术和计算机技术才可能全面、完整地加以归档、组织和存储，进而完整的记录、反映重大社会事件。新浪微博数据中心发布的《2018 微博用户发展报告》<sup>[21]</sup>显示，2018 年微博

<sup>[11]</sup> Archive Social. Social media archiving [EB / OL]. [2019-08-18]. <http://www.archive-social.com>.

<sup>[12]</sup> 周文泓. 互联网信息社会化保存的冷思考与热展望[J]. 图书馆论坛, 2020, 40(01): 87-95.

<sup>[13]</sup> 周文泓. 我国网络空间中档案领域的缺位审视和参与展望——基于社交媒体信息保管行动的解析[J]. 档案与建设, 2019(09): 13-17.

<sup>[14]</sup> WILEY J. The Onslaught of Big Data and the Information Governance Imperative Information Governance for Social Media [M] // ROBERT F. Smallwood. Information Governance: Concepts, Strategies, and Best Practices. New York City: Wiley, 2014: 1-22.

<sup>[15]</sup> BUSHEY J. Convergence, Connectivity, Ephemeral and Performed: New Characteristics of Digital Photographs[J]. Archives & Manuscripts, 2014 (1): 33-47.

<sup>[16]</sup> 周文泓. 社交媒体平台作为档案库的内涵及特点解析[J]. 档案学研究, 2019(02): 104-110.

<sup>[17]</sup> 王新才, 徐欣欣. 国外个人数字存档的实践经验及其启示[J]. 信息资源管理学报, 2016(4): 109-115.

<sup>[18]</sup> Helen Hockx-Yu. Preserving Social Media in the Context of Non-print Legal Deposit [EB/OL]. [2016-09-09] <http://library.ifla.org/999/>.

<sup>[19]</sup> 高晨翔, 黄新荣. 我国社交媒体文件的保管政策研究[J]. 图书馆研究, 2017(7): 47-55.

<sup>[20]</sup> 周文泓. 社交媒体信息保管模式与策略探析[J]. 档案与建设, 2018(09): 17-21.

<sup>[21]</sup> 新浪微博数据中心. 2018 微博用户发展报告[EB/OL]. [2020-04-15]. <https://data.weibo.com/report/repo>

日均文字发布量达 1.3 亿，日均图片发布量超 1.2 亿。以新冠肺炎疫情为例，从 1 月 20 日到 2 月 4 日，国务院部委和各级人民政府已发布超过 55 万条疫情相关微博。卫健系统凝聚 1000 余个账号，已发布超 6 万多条疫情信息。此外，包括一线医务人员、医院、政府机构、媒体、热心网友在内，社会各界已经发布了 19.5 万条求助微博。鉴于社交媒体平台的不稳定性，如果不对有价值的社交媒体信息进行合理保存，宝贵的公共历史与记忆资源就有可能消失。<sup>[22]</sup>

因而，为了充分记录并反映在众多重大社会事件的大背景下社交媒体中的各方参与主体及其社会活动，项目以宏观鉴定理论和职能分析<sup>[23]</sup>、来源原则为指导，综合运用 Web 应用程序的自动化测试工具（Selenium）、数据库、超文本标记语言（HTML）、层叠样式表（CSS）和 JavaScript 等多种语言、工具和技术，参考国内外重大社会事件网络存档和社交媒体存档项目，开发了“重大社会事件社交媒体存档平台”，以期提供社交媒体档案存取资源库与存档协作空间，并在实践基础上形成系列理论与实践成果。其中，宏观鉴定和职能分析用以确定归档范围和归档优先级，来源原则用于指导归档信息的组织和存储过程。Selenium 用于信息捕获、数据库用于信息存储，HTML、CSS 和 JavaScript 用于网站建设。

---

rtDetail?id=433,

<sup>[22]</sup> 蒋冠,冯湘君.国家综合档案馆在社交媒体信息存档中的角色定位研究[J].档案学研究,2018(01):10-13.

<sup>[23]</sup> 特里·库克,李音.宏观鉴定与职能分析[J].中国档案,2012(01):51-53.

## 2 项目特色及优势

基于调研结果，项目收集并整理了当前的社交媒体存档实践与数字记忆项目，以及往届“全国高校档案学专业大学生课外科技作品竞赛”中入围的信息资源开发类作品。这些已有案例为本项目的构建提供了良好的参照和创新基础，如现有社交媒体存档项目的存档流程与存档技术的参考，以及往届竞赛作品的成果形式借鉴等。

通过对当前的社交媒体存档实践、数字记忆项目这些与本作品定位相关的实践项目的梳理，在不同类型的实践中选择了其中代表性项目，从主体、方法、资源、技术、利用等维度进行了分析，以下为分析结果（见表 1—表 2）。

| 比较维度<br>项目 |              | 国际存档组织          | 地区文献机构  | 高校文献机构  | 个人志愿者       | 本项目优势                                            |
|------------|--------------|-----------------|---------|---------|-------------|--------------------------------------------------|
|            |              | IIPC CDG        | UKGWA   | 杜克大学    | nCov-Memory |                                                  |
| 构建主体       |              | 国际网络保存联盟内容开发工作组 | 英国国家档案馆 | 杜克大学档案馆 | Github 个人用户 | 专业属性、与创造力兼具的年轻团队                                 |
| 构建方法       | 重大社会事件维度     | √               | √       | √       | √           | 已有项目虽都涉及了重大事件与社交媒体两种方法，但单纯利用社交媒体存档构建重大社会事件的项目较少。 |
|            | 社交媒体维度       | √               | √       | √       | √           |                                                  |
|            | 重大社会事件社交媒体存档 |                 | √       |         | √           |                                                  |

|      |      |                            |              |      |                                   |
|------|------|----------------------------|--------------|------|-----------------------------------|
| 技术框架 | 暂未说明 | 只说明采集工具为<br><b>Hertrix</b> | 暂未说明         | 暂未说明 | 统一的技术框架将更好地保障实践开展                 |
| 利用深度 | 暂无   | 暂无                         | 虚拟校园<br>场景年鉴 | 暂无   | 本项目对存档资源进行更好的利用既发挥了资源价值又增强项目的吸引力。 |

表 1 社交媒体存档实践项目分析表

|                  | 国家<br>记忆项目 | 地区<br>记忆项目                               | 重大事件记<br>忆项目       | 本项目优势                                                                             |
|------------------|------------|------------------------------------------|--------------------|-----------------------------------------------------------------------------------|
|                  | 中国记忆项目实验网站 | 记忆四川                                     | 2019nCov<br>Memory |                                                                                   |
| 开<br>发<br>主<br>体 | 国家图书馆      | 四川省新闻<br>网传媒集团<br>四川省非物<br>质文化遗产<br>保护中心 | Github<br>个人用户     | 当前进行数字记忆项目构建的多为图书机构以及非遗保护中心一类的官方机构，档案机构的参与深度与广度不及前两者。本作品能够补充从档案视角出发构建的重大社会事件记忆项目。 |

|        |                               |                         |      |                                                                                                                 |
|--------|-------------------------------|-------------------------|------|-----------------------------------------------------------------------------------------------------------------|
| 资源对象   | 专题文献、采集口述史料、影像史料等新类型文献与传统纸质文献 | 民间文学、传统音乐、舞蹈、曲艺等非物质文化遗产 | 网络信息 | 当前的数字记忆类项目所采用的资源多为人物、非遗项目、传统文献、传统艺术品的专题，且资源形式大多为文字、图片、视频等传统形式，采用原生网络信息进行整合与开发的项目少之又少。本项目则聚焦于原生网络信息，并将其整合、开发与利用。 |
| 成果形式   | 网站                            | 网站                      | 网站   | 网站是目前应用最广使用频率最高的开发成果形式，利用网站能够更好地展现已有资源和开发成果，并且可以提供信息、资源互动的渠道。                                                   |
| 开发利用形式 | 专题资源汇编                        | 专题资源汇编（包含视频、音频与文字等）     | 暂无   | 多为专题资源汇编，通过文字与音频视频结合的方式对某类活动进行介绍，缺乏开发的深度与趣味性。本项目则对开发利用模式进行创新，提高开发利用形式的互动性和创新性。                                  |

表 2 数字记忆实践项目分析表

同时，项目也对历届“全国高校档案学专业大学生课外科技作品竞赛”中入围的信息资源开发类作品进行了梳理，并从研究对象、成果形式和所用技术方面进行了分析，得到如下表格（见表 3）。

|      | 往届同类参赛作品                | 本项目优势                                                             |
|------|-------------------------|-------------------------------------------------------------------|
| 研究对象 | 多为历史档案、少数民族档案等传统形态的档案资料 | 本作品采集的是新浪微博平台上重大社会事件的社交媒体资源，这是档案作品展里第一次以原生网络信息为研究对象而进行采集和开发利用的项目。 |
| 成果形式 | 资料汇编、网站、APP、资源开发设计方案等   | 本作品具体成果包含档案网站等技术实践类成果，也包含存档全过程框架等理论研究成果，形式多样。                     |
| 所用技术 | 数字人文技术、数据库、网页设计等        | 本作品在构建过程中使用了包括技术爬虫技术、数据库、网站建设技术、网页设计等多种具体技术，科技含量较高。               |

表 3 历届信息资源开发类参赛作品分析表

综合上述表格的分析内容，总结出了本项目的特色与优势所在：

## 2.1 项目价值

在社会价值方面，本项目在对重大社会事件的社交媒体信息进行采集时，遵循“多种类型、多种主题、多种声音”的原则，存档的事件包括国家、地区、行业不同层次不同类型的重大社会事件，存档的用户汇集了政府机构、社会组织、新闻媒体和个人用户等多元主体。希望能构建多层次事件存档、多元主体参与、多元话语权共存的重大社会事件记忆。

在专业价值方面，本项目从档案专业与档案视角出发对重大社会事件社交媒

体存档做出实践尝试，积极探索社交媒体存档的实践框架，从而在当前的实践领域中增添档案视角，贡献档案人的力量。此外，作为档案专业的大学生，应积极把握本学科理论与实践的前沿热点。在当前数字转型的理论背景下、在当前社交媒体资源海量产生的现实前提下，对社交媒体作为档案资源进行保存进行相关的理论学习与实践探索是与本学科、本专业的发展要求高度契合的。

## 2.2 资源优势

本项目采集的是新浪微博平台上的新冠肺炎疫情、2021 成都世界大学生运动会、互联网信息社会化保存项目等各个层级、各个行业的社交媒体信息资源，这是档案作品展里第一次以原生网络信息为研究对象而进行采集和开发利用的项目，甚至在我国档案馆的实践里也未曾出现，具有高度的创新性。

## 2.3 科技含量

本项目在构建过程中采用了多个领域的具体技术，其中包括：资源捕获时所用的爬虫技术、前期设计中所用的网页设计技术、网站构建时所用的数据库以及网站建设技术，后续在对产品进行开发的过程中还将涉及到动画制作、3D 建模、传播学理论与知识的学习等，作品本身与其构建过程具有较高的科技价值与含量。

## 2.4 成果形式

本项目包含的成果既有档案网站及其后台资源库以及开发产品与服务等技术实践类成果，也包含项目实证研究报告、存档全过程框架等理论性研究成果，形式多样，能够为后续同类型研究提供理论参考与实践对照。

### 3 作品主要内容

重大社会事件社交媒体存档平台自开发以来，形成了阶段性成果，在不同的场景中得到关注和讨论，持续拓展科学研究和实践应用空间。

在档案学界，中国人民大学冯惠玲教授在其发表的文章《数字记忆：文化记忆的数字宫殿》<sup>[24]</sup>中对该项目有所提及，并将其归属为“重大事件记忆项目”，肯定了该项目在参与构建数字记忆中发挥的积极作用。在主题为“赋能知识社会建构社媒记忆”的2020国际档案周前沿论坛暨第五届档案社交媒体圆桌会议中，与会学者从构建社交媒体记忆的角度介绍了该项目的具体情况。此外，在涵盖图书情报与档案管理、公共管理、法学、管理科学与工程等学科的“公共管理学院跨学科应急信息管理论坛”上，与会学者基于本项目提出从应急建档角度将重大社会事件社交媒体存档纳入应急管理体系。

在档案实践领域，通过与成都市档案馆积极交流与沟通，该作品对2021成都大运会事件的社交媒体存档已获得成都市档案馆的认可。项目团队将作为2021成都大运会社交媒体存档工作的承接团队，参与2021成都大运会的档案管理工作。一方面，项目实践中存档的大运会相关的社交媒体信息将作为传统档案的有力补充，成为大运会档案馆藏的一部分；另一方面，项目团队将与成都市档案馆协同探讨，共同形成“重大社会事件社交媒体存档”的全过程方法与具体规范，以期能够在相关标准规范或指南等方面有所贡献。

在社交媒体上，四川大学未来档案实验室微信公众号发布推文“星火之力筑抗疫记忆：社交媒体信息存档在路上”，宣布启动面向“新冠肺炎”疫情的社交媒体存档项目。本项目作为该项目的具体实践之一，依托该项目宣传，亦在社交媒体平台获得了广泛关注。

---

<sup>[24]</sup> 冯惠玲. 数字记忆:文化记忆的数字宫殿[J]. 中国图书馆学报,2020,v.46;No.247(03):4-16.

## 3.1 作品基本信息

### 3.1.1 作品定位

一个开放、参与、互动的资源空间。用户可在该平台浏览、检索、获取重大社会事件社交媒体存档资源，并加以利用。同时，通过编研室、“故事树”等核心功能，用户可以参与社交媒体存档过程，推动构建多元叙事空间。

### 3.1.2 用户群体

（1）首先，公众在网站中可以获取、浏览到重大社会事件的社交媒体信息，并且可以参与归档过程，对构建国家记忆和民族记忆有重要意义。

（2）其次，专家学者可以从中获取重大社会事件的原始信息和数据，从而为其学术研究提供支撑。

（3）此外，处于档案共同体的人可以在“档案领域”专题中寻找和构建档案共同体的专属记忆，且网站本身也可以作为档案学者研究社交媒体存档的实践案例，从而促进社交媒体存档的理论探索。

（4）最后，从国家这一宏观层面来看，社会各界积极参与重大社会事件社交媒体存档，对讲述中国故事、构建数字中国记忆都有重大意义。

### 3.1.3 开发原因

（1）以档案之力讲述中国故事。该作品通过面向重大社会事件的社交媒体信息进行归档，尽量客观、全面的保护、记录这些对人们产生重要影响的社会事件，以档案之力讲述各式各样的中国故事。

（2）以社交媒体建构记忆权力。疫情中的医护人员、成都大运会中的志愿者、70 周年阅兵中的青年学生，这些普通群众的社交媒体信息都是该作品的归档对象。对社会普通成员的关注，有利于推动普通群众的记忆话语权的建构。

(3) 以交互共享激发档案意识。通过用户分享亲身经历或真实事件，该作品让公民主动参与信息归档、分享与著录，可增加档案工作的影响力，促进公民了解档案事业。

(4) 以数字技术拓展档案空间。随着“互联网+”的逐步深入，网络空间与数字技术已然进入档案领域，该作品利用数字技术，能更加高效地归档、整理和存储信息，实现社交媒体存档。从而积极响应《全国档案事业发展“十三五”规划纲要》等国家战略规划，探索重大社会事件社交媒体存档路径，为数字转型贡献力量。

### 3.1.4 开发团队

| 姓名  | 专业年级                          | 分工        |
|-----|-------------------------------|-----------|
| 张晓宇 | 四川大学公共管理学院<br>2017 级档案学专业     | 统筹项目 报告撰写 |
| 吴 琼 | 四川大学公共管理学院<br>2018 级档案学专业     | 报告撰写 美工设计 |
| 贺谭涛 | 四川大学公共管理学院<br>2018 级档案学专业     | 报告撰写 方案设计 |
| 黄小宇 | 四川大学公共管理学院<br>2018 级档案学专业     | 报告撰写 开发设计 |
| 王 淼 | 四川大学计算机学院 2018<br>级计算机科学与技术专业 | 技术实现      |

表 4 团队成员分工表

## 3.2 资源与功能

### 3.2.1 资源建设

资源建设是本项目平台构建的关键一环，本平台的资源建设通过平台自身搭建、用户上传贡献、平台与用户互动过程中形成并采集等多种方式，形成了一个由多方参与、共建共享的资源集合。

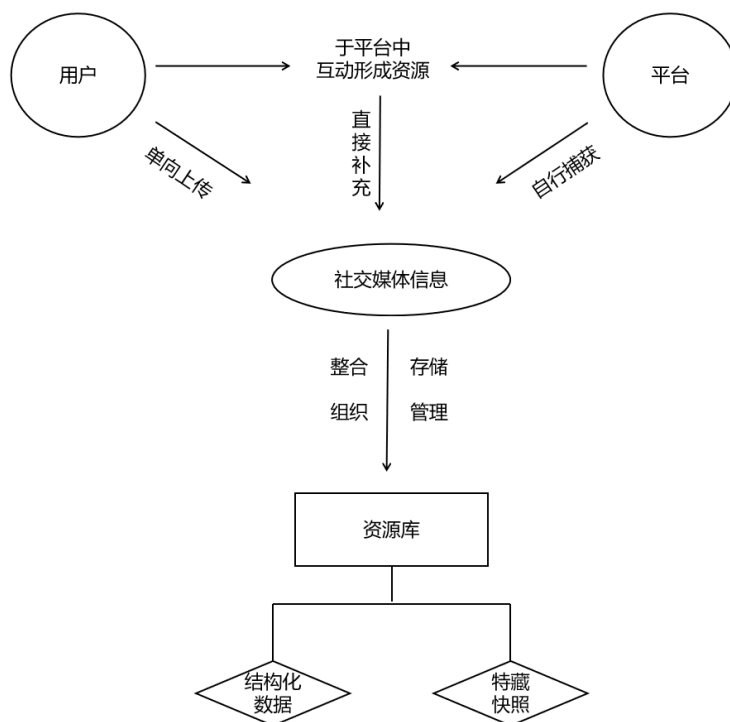


图 1 资源建设示意图

#### 3.2.1.1 平台：事件识别与资源搭建

在平台构建初期，事件的选择与资源的构建主要依赖于平台自身的识别与获取，这一阶段平台的资源搭建可以主要从内容和形式两方面展开说明。

##### （1）内容上：持续的重大社会事件识别

本项目平台致力于面向社会发展进程持续识别与归档重大社会事件并在此基础上提供利用，并鼓励社会各界参与社交媒体存档过程，构建多元叙事空间。

目前平台对于社会重大事件的选取涵盖国家、地区以及行业三个主要方面。

此外，项目还专门针对档案领域的重大社会事件开辟存档板块，希望对档案专业热点进行保存与记录。目前已入选归档范围的事件列表如下（见表 5）。

| 层级   |      | 事件           | 备注    |
|------|------|--------------|-------|
| 国家层面 |      | 新冠肺炎疫情       | 基本完成  |
|      |      | 国庆 70 周年阅兵   |       |
|      |      | 民法典草案通过      |       |
| 地区层面 |      | 2021 成都大运会   | 正在进行中 |
|      |      | 上海率先实行垃圾分类   |       |
|      |      | 2020 乌镇戏剧节   |       |
| 行业层面 | 教育行业 | 高考改革         | 正在进行中 |
|      | 文体行业 | 中国电竞——IG 夺冠  |       |
|      | 信息行业 | 5G           |       |
| 档案领域 |      | 2020 国际档案日   |       |
|      |      | 《档案法》修订      | 正在进行中 |
|      |      | 互联网信息社会化保存项目 |       |

表 5 事件存档范围表

## （2）形式上：多形式的资源构建

除了覆盖多层级、多领域的事件集，平台在资源形式方面的建设同样丰富。当前平台资源库以数据库储存的结构化数据为主要内容，但由于归档工具与呈现技术的局限，此类信息在对社交媒体信息原始面貌与状态的呈现效果欠佳。因此，本平台在以数据库储存的结构化数据为主要内容的基础上，利用网络快照技术对部分事件的社交媒体信息进行存档作为特藏快照作为补充，可以更加真实地还原社交媒体信息原貌并对资源库进行形式和内容上的扩充。

**①数据库储存的结构化数据。**以数据库储存的结构化信息具有存储结构清晰、后台管理便捷、方便检索和开发利用等特点，故平台各社会事件均要以结构化数据的形式归档原始社交媒体信息。但限于技术和合规限制，每个事件主要就 10-50 个同事件关联最为密切的社交媒体账户进行内容存档，如截至目前，项目就

2021 成都大运会事件的 28 个微博用户的近 900 条微博进行了前期存档，且将随着 2021 年赛事的正式启动进行主体内容的追踪与存档。此外，项目正在对《档案法》修订事件的 12 个微博用户以及高考改革事件的 13 个微博用户的相关社交媒体信息进行存档。

**②网页快照。**网页快照相较于数据库储存的结构化信息来说，具有易于捕获、对社交媒体信息还原度高等特点。但由于其内含的信息要素无法拆解与重组，给资源的结构化存储管理与开发利用带来问题，故将其作为当前技术限制下的资源补充形式。当前，平台对新冠肺炎疫情选取了 200 个最具影响力的微博账户的近 20000 条微博进行存档，每条快照下还归档了将近 15 条评论，若每条评论算一条信息，则有将近 300000 条信息。

### 3.2.1.2 用户：线索与具体资源贡献

若要实现平台长期、可持续的资源构建，就要将用户发展成为平台资源的重要来源，通过广泛的用户群体实现资源的多渠道、多维度补充。当前平台建设中对于用户主动的资源贡献主要涉及以下两方面的内容：

**（1）用户对于资源线索的提供。**用户对于资源线索的提供也可以划分为两种类型种，一是用户对于重大事件的提名，用户可以在平台上提交自己认为具有较高存档意义和价值的事件并阐述自己的理由，平台通过后台审核后，将选择其中有价值的事件由平台进行存档。另一方面是对某条社交媒体信息的具体位置进行线索提供。由于平台限制或用户自身技能掌握情况存在差异，部分社交媒体资源存在无法直接获取并上传的情况，在此情况下用户可以向平台提供线索，由平台进行归档。这两种方式下用户不直接提供具体资源，而是对应存档的事件、具体资源进行题名及位置的提供，故称其为资源线索的提供。

**（2）用户对于具体资源的上传。**与提供资源线索相对应，用户还可以直接向平台提供自己前期归档或存储的社交媒体资源。为此平台开辟了专门的资源贡

献板块，用于收集用户贡献的源链接、图片、视频、文本等具体资源，通过这种方式一方面可以增强用户与平台的互动，另一方面也是对平台资源的补充与完善。更重要的是，形成用户上传资源的互动机制、培养用户的参与积极性对于平台的后续可持续发展推广具有重要意义与价值。

### 3.2.1.3 互动过程中：特殊的资源获取方式

本平台的资源构建不仅局限在平台自身的资源搭建与用户的单向贡献上传，还可以通过平台开放的开发利用功能对某一专题的社交媒体档案资源进行直接采集。

例如，通过平台开发的“故事树”工具，用户可以通过讲述某一重大社会事件的专题故事进行资源贡献。平台可以利用此工具直接批量收集关于某一事件的专题资源对原有资源集进行补充完善。

此外，利用平台的“快闪聊天室”的回忆沙龙活动还可以通过发布主题邀请用户交流而形成关于某一事件的“线上口述档案”。此类特殊形式的资源在平台后期开发利用过程中具有独特作用，并且可以为研究人员提供更加丰富多样的研究资源。

## 3.2.2 资源存储

两类不同形式的社交媒体信息，在存储格式和存储方式上均有所不同。

对于结构化信息，项目利用 **Selenium** 模块应用程序进行捕获，随后自动录入 **MySQL** 数据库中，对于一条微博的描述字段包含用户 ID、用户昵称、用户性别、用户所在地、用户生日、用户简介、职业信息、教育信息、微博正文、微博 URL、微博原图、转发带图、视频 URL、微博发布地址、微博发布时间、微博发布终端、点赞数、转发数、评论数。部分字段如下（见图 2）。

```
mysql> desc user;
```

| Field           | Type         | Null | Key | Default | Extra |
|-----------------|--------------|------|-----|---------|-------|
| id              | varchar(20)  | NO   | PRI | NULL    |       |
| nickname        | varchar(30)  | YES  |     | NULL    |       |
| gender          | varchar(10)  | YES  |     | NULL    |       |
| location        | varchar(200) | YES  |     | NULL    |       |
| birthday        | varchar(40)  | YES  |     | NULL    |       |
| description     | varchar(140) | YES  |     | NULL    |       |
| verified_reason | varchar(140) | YES  |     | NULL    |       |
| talent          | varchar(200) | YES  |     | NULL    |       |
| education       | varchar(200) | YES  |     | NULL    |       |
| work            | varchar(200) | YES  |     | NULL    |       |
| weibo_num       | int(11)      | YES  |     | NULL    |       |
| following       | int(11)      | YES  |     | NULL    |       |
| followers       | int(11)      | YES  |     | NULL    |       |

13 rows in set (0.00 sec)

```
mysql>
```

图 2 结构化数据存储结构图（节选）

对于网页快照，项目同样利用 Selenium 模块应用程序进行捕获，再通过数据过滤环节进行数据清洗，最后根据来源原则，将其根据不同来源即用户类别进行归档和排序，数据格式为 MTHML。部分结果如下（见图 3）

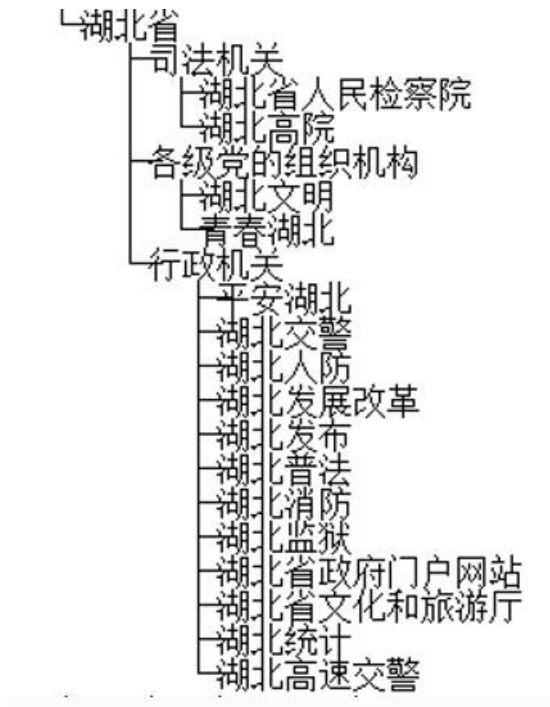


图 3 网页快照用户归档图（节选）

### 3.2.3 平台功能

该网站分为核心功能、常规功能和辅助信息三大模块。除网站资源浏览、检索获取等常规功能服务外，开放、创新、社会化的开发利用功能是平台的重点功能。一方面，平台不定期推出内容或形式创新的旨在构建数字中国记忆的在线展览，以社交媒体档案面向不同事件的记忆要点丰富展览厅板块。另一方面，通过编研室、用户贡献、故事树和回忆沙龙等核心功能强化重大社会事件存档的社会参与，努力构建多元叙事空间，从而形成国家记忆、地区记忆以及集体记忆。此外，辅助信息模块中，用户可以了解平台规范、政策以及该项目的具体情况。

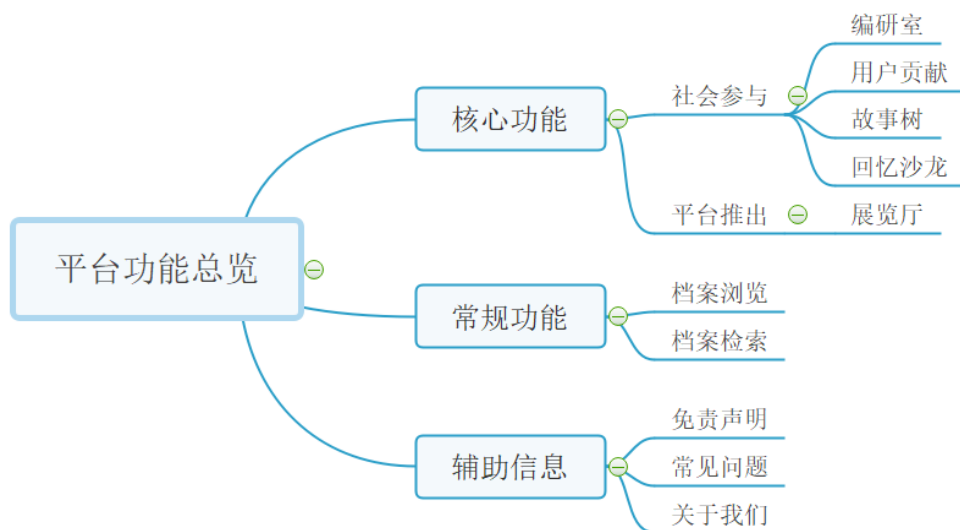


图4 平台功能总览

#### 3.2.3.1 核心功能

##### 3.2.3.1.1 档案展览

| 功能需求                 | 功能介绍                 | 创新性与价值                                                                                        |
|----------------------|----------------------|-----------------------------------------------------------------------------------------------|
| 用户想查看档案展览或对某一具体专题感兴趣 | 为用户提供有组织、有价值的专题性档案展览 | <p><b>展览内容有组织、有价值</b></p> <p>本平台在熟悉所拥有资源的基础上开展了对他人感兴趣的档案展览主题的调查询问，并结合其他档案展览案例对档案的内容展开挖掘、关</p> |

|  |  |                                                                                                                                      |
|--|--|--------------------------------------------------------------------------------------------------------------------------------------|
|  |  | <p>联和组织，形成了有趣、有组织、有价值的档案专题展览。</p> <p><b>展览资源可拓展</b></p> <p>作为一个可持续性平台，展览资源也是可持续拓展的。所收纳的丰富资源支持平台拓展各类专题，而不断更新的资源又保证每个专题的具体内容可不断更新。</p> |
|--|--|--------------------------------------------------------------------------------------------------------------------------------------|

表 6 档案检展览功能表



图 5 展览厅页面

## 3.2.3.1.2 编研室

| 具体功能 | 功能需求                          | 功能介绍                      | 创新性与价值                                                                                                          |
|------|-------------------------------|---------------------------|-----------------------------------------------------------------------------------------------------------------|
| 在线编研 | 用户想要在线使用网站所拥有的档案资源进行编研        | 可以自由选取所需要的档案资料；可以在线编辑批注资源 | <p><b>层次完善，内容丰富的档案资源</b></p> <p>本产品可直接向用户提供平台中的所有档案资源，为编者形成疫情相关专题的编撰提供选材上的便利。</p>                               |
| 导出保存 | 用户想要将自己编研的成果保存至本地             | 编辑结束后可以下载导出到本地            | <p><b>可在线使用的，便捷的编辑工具</b></p> <p>在线编辑工具基本满足了一般的编纂需求，实现了从选材到编辑的一站式服务。</p>                                         |
| 互动分享 | 用户想要将自己的编研成果分享到平台或想要查看他人的编研成果 | 可以在线上传自己的编纂作品；可浏览已有的编撰作品。 | <p><b>线上的互动交流功能</b></p> <p>互动交流功能，满足了用户学习交流、自我展示的需求，也为作品提供了丰富资源的途径。</p>                                         |
|      |                               |                           | <p><b>功能顺应并推动档案事业发展</b></p> <p>随着《全国档案信息化实施纲要》的发布实行，数字档案工作日渐发展。而数字档案编研工作也已经摆在档案编研人员的案头。本功能有利于推动线上数字编研工作和探索。</p> |

表 7 线上编研室功能表

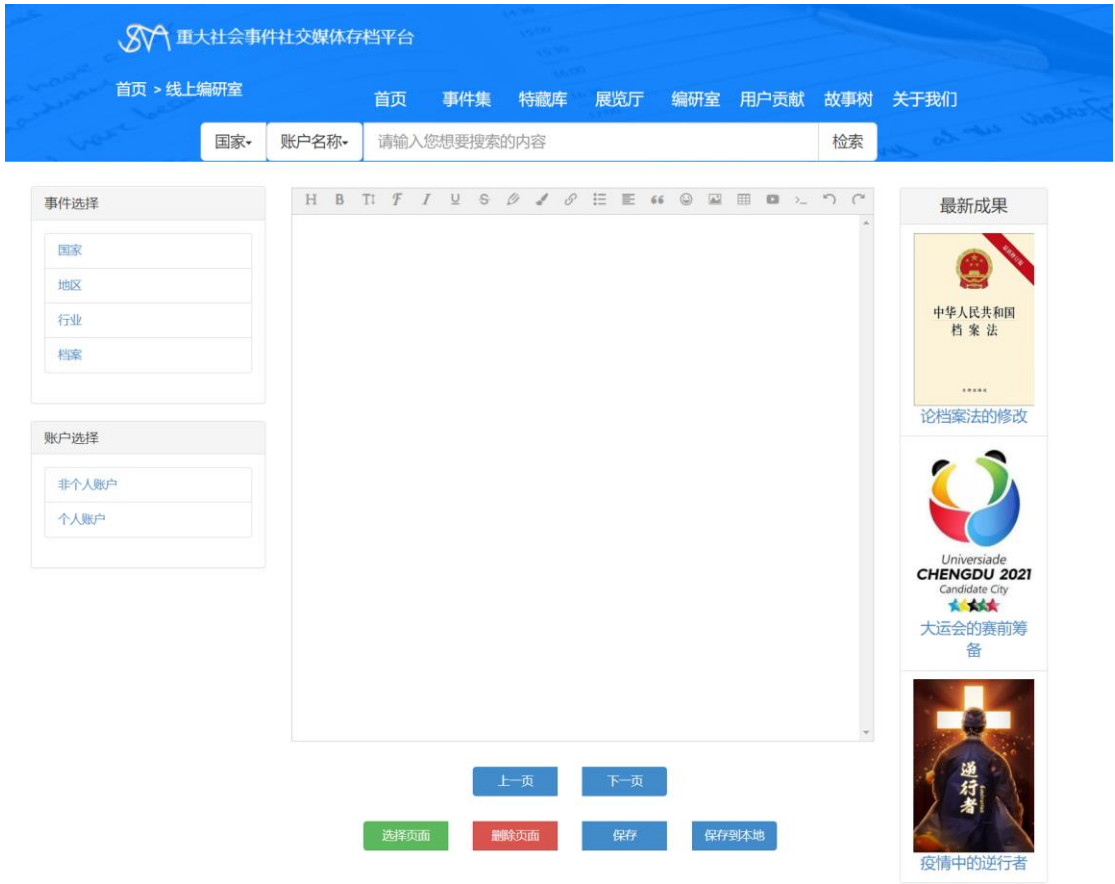


图 6 线上编研室页面

3.2.3.1.3 用户贡献

| 具体功能        | 功能需求                        | 功能介绍                                               | 创新性与价值                                                                                                                                          |
|-------------|-----------------------------|----------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|
| 社交媒体信息资源上传  | 用户希望分享自己拥有的某一特定事件的社交媒体信息    | 提供某一事件的社交媒体信息或对事件、专题进行提名，经平台审核后前者可补充资源库，后者可纳入归档范围。 | <b>形式贡献</b><br>支持社交媒体信息资源的上传，同时也鼓励用户提供归档事件或专题                                                                                                   |
| 事件提名及相应资源上传 | 用户对尚未归档的事件或专题有浓厚兴趣，希望平台对此归档 |                                                    | <b>多格式上传</b><br>支持文本、图片、以及快照等多种形式的资源上传<br><b>参考规范标准</b><br>都柏林核心元素集具有简易性、通用性、等特点，电子文件归档与电子档案管理规范是我国电子档案管理的行业标准。借鉴上述标准规范资源上传有利于一定程度上保证档案真实性和完整性。 |

表 8 用户贡献功能表

资源的上传需要明确作品主题、信息形式、源链接、封面图片、资源所有者、创建时间、信息描述、信息类型，并附上传的具体资源。信息形式包括文本、图片、视频和其他，而源链接考虑到用户的信息素养差异，设定为选择性填写内容。

事件提名需要提供事件或专题的统一资源定位器(URL)，并说明提名理由。此外，用户可以选择上传相应事件的社交媒体信息资源，作为该事件归档资源的有力补充。具体页面展示如下（见图 7—图 8）

SA 重大社会事件社交媒体存档平台

首页事件集特藏库展览厅编研室用户贡献故事树关于我们

首页 > 用户贡献 > 资源贡献

国家- 账户名称- 请输入您想要搜索的内容 检索

来源贡献 资源上传

作品主题

请输入主题

信息形式

☐ 网页链接 ☐ 文本 ☐ 图片 ☐ 视频 ☐ 其他

网页链接

请输入网页链接

封面图片

选择图片

目前可支持png jpg pdf 等格式

资源所有者

请输入资源所有者

创建时间

请选择具体日期

信息描述

请输入信息描述

信息类型

☐ 社交媒体档案 ☐ 信息处理作品 (单选)

上传附件

网页链接 图片 文档 截屏

H B T I U S

请确保您上传的信息由您本人发布或已获得发布者授权

上传

免责声明

联系我们: scu\_archivelab@163.com

常见问题

关注我们: 未来档案实验室(微信公众号)

图 7 资源上传页面

SA 重大社会事件社交媒体存档平台

首页事件集特藏库展览厅编研室用户贡献故事树关于我们

首页 > 用户贡献 > 事件提名

国家- 账户名称- 请输入您想要搜索的内容 检索

提名事件

请输入事件名称

事件类别

☐ 国家重大社会事件 ☐ 地区重大事件

☐ 行业重大社会事件

☐ 档案领域重大事件

提名原因

请输入提名原因

网页链接

请输入网页URL

其他附件

图片 文档 截屏 压缩包

若您已有关于此事件的其他资源收集, 请在此处一并提交上传

请在下方留下您的联系方式, 方便我们在资源获取过程中与您取得联系。

联系电话

请输入联系电话

电子邮件

请输入邮件地址

请确保您上传的信息由您本人发布或已获得发布者授权

上传

图 8 事件提名页面

## 3.2.3.1.4 “故事树”

| 具体功能 | 功能需求                 | 功能介绍                                 | 创新性与价值                                                                                                                                                                                                                                                                                                                                                                                     |
|------|----------------------|--------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 种植   | 用户希望以具象化的方式呈现某一事件或专题 | 通过用户提名或平台指定的方式，确定“种植”的重大社会事件或专题      | <p><b>反映用户需求</b></p> <p>在事件或专题的选择上，用户可以通过提名的方式进行推荐，平台将对此做出选择，最大限度的满足用户需求。</p> <p><b>多元主体共建</b></p> <p>目前大多数档案网站的建设、开发主体单一，不能反映事实上存在的多元主体。该板块一改传统网站的不足，主张多主体参与故事树的种植、培育等过程，从而实现对事件或专题的全景式叙事。</p> <p><b>具象化的表现形式</b></p> <p>采取“故事树”这一具象化的表现形式，可以有效的吸引用户参与“故事树”培育过程，同时也能有效降低操作的复杂度。</p> <p><b>可生长的树形结构</b></p> <p>采用树形结构除了有利于对事件进行划分和细化外，也有易于扩展的特性。所以用户容易在原有的“故事树”中接入新的分支或节点，而不影响原有资源的结构。</p> |
| 添枝   | 用户希望更全面、细致地描述事件      | 用户可以自主发挥，为这一事件“添枝”，形成具有个人特色的二级类目或子专题 |                                                                                                                                                                                                                                                                                                                                                                                            |
| 加叶   | 用户丰富“故事树”中的社交媒体存档资源  | 用户可以在“故事树”上添加社交媒体档案，增加“故事树”的树叶总量     |                                                                                                                                                                                                                                                                                                                                                                                            |

表 9 “故事树”功能表

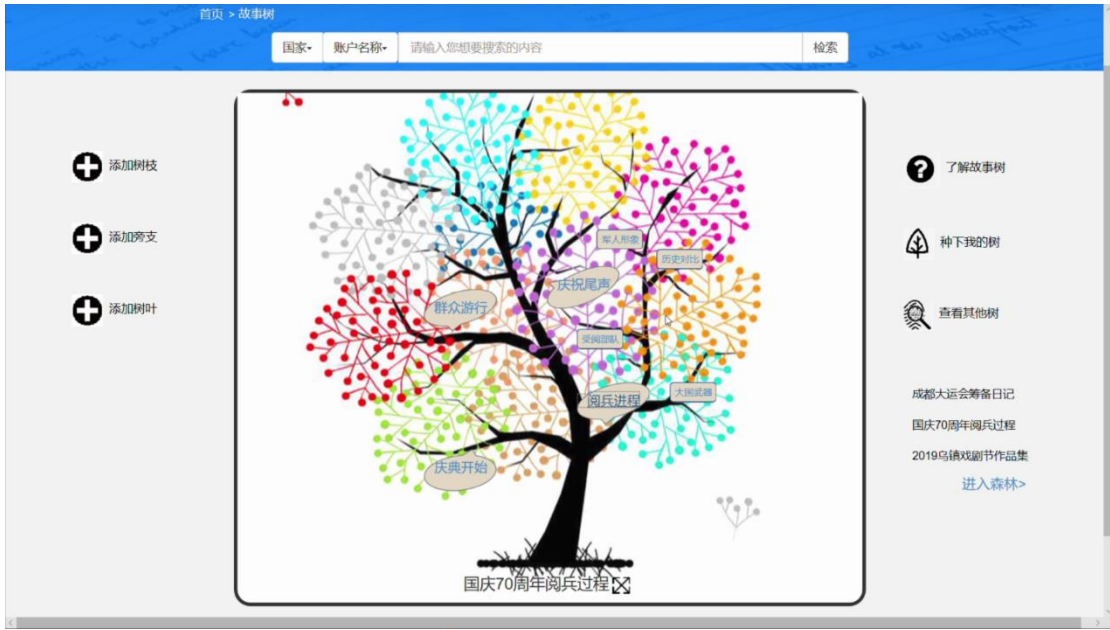


图 9 “故事树”页面

3.2.3.1.5 “回忆沙龙”

| 具体功能   | 功能需求                 | 功能介绍                     | 创新性与价值                                                                                                                                                               |
|--------|----------------------|--------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 发布讨论专题 | 用户希望平台内存在关于某一事件的回忆空间 | 平台不定期发布关于某一事件的讨论专题       | <b>打造回忆空间</b><br>每个人都有自己的故事和经历，这些故事和经历都有一定的事件背景。平台通过提供这一事件背景，让用户在该空间中回忆、分享自己的独特经历<br><b>创新资源收集方式</b><br>用户在讨论交流的过程中形成的文字材料可以成为该事件或专题口述档案的原始资源，而用户分享的社交媒体信息资源也是对该事件或专 |
| 分享自身想法 | 用户有分享、表达自身观点和想法的需求   | 在该板块分享自身看法与观点，并与其他用户即时交流 |                                                                                                                                                                      |

|        |                     |                         |                                                                                                |
|--------|---------------------|-------------------------|------------------------------------------------------------------------------------------------|
| 贡献专题资源 | 用户希望能够提供社交媒体资源供大家讨论 | 在聊天的过程中，即时分享自己拥有的社交媒体信息 | 题资源建设的有效补充。<br><b>即时性和灵活性兼具</b><br>讨论的即时性能够调动用户的积极性，从而提高资源获取的可能性。专题发布的灵活性可降低运营成本，并参考社会热点不定期发布。 |
|--------|---------------------|-------------------------|------------------------------------------------------------------------------------------------|

表 10 “回忆沙龙”功能表



图 10“会议沙龙”页面

3.2.3.2 常规功能

3.2.3.2.1 档案浏览

| 具体功能       | 功能需求                      | 功能介绍                       | 创新性与价值                                                                                                                                                                                                                                      |
|------------|---------------------------|----------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 事件集结构化数据浏览 | 用户想要浏览关于某一事件的社交媒体信息       | 为用户提供可浏览的每一事件相关社交媒体档案      | <p><b>以原始形式为基础呈现社交媒体档案</b></p> <p>档案的特性是原始记录性，因而在展现档案时也应应以原始数据和真实呈现形式为根本。本平台将所搜集的资源以原始或接近原始呈现方式的形式展现所有内容，以保证真实性。</p> <p><b>以事件维度组织杂乱的社交媒体信息</b></p> <p>社交媒体平台上的信息丰富却杂乱无章，难以进行集中了解和研究。本平台将所有信息以筛选搜集后以事件维度划分，在不同事件下展示不同档案，使信息得到整理，使档案得到利用</p> |
| 特藏库网页快照浏览  | 用户想要浏览某一事件的原始社交媒体网页快照档案资源 | 以网页快照的形式为用户提供每一事件相关的社交媒体信息 | <p><b>以两类资源在形式和内容上相互补充</b></p> <p>可浏览的资源包括结构化数据和网页快照，两者在形式和内容上相互补充，网页快照完美的呈现了原始信息形式，而结构化数据展现了基本内容和元数据，为用户的浏览和后续利用提供了便利</p>                                                                                                                  |

表 11 档案浏览功能表

## (1) 事件集结构化数据浏览



图 11 结构化数据浏览指示图



图 12 结构化数据浏览页面

(2) 特藏库网页快照浏览



图 13 特藏库网页快照浏览指示图



## 人民日报



图 14 特藏库网页快照展示图

### 3.2.3.2.2 档案检索

| 具体功能  | 功能需求           | 功能介绍            | 创新性与价值                                                                                     |
|-------|----------------|-----------------|--------------------------------------------------------------------------------------------|
| 多入口检索 | 用户想在不同页面开始检索   | 每个页面都有检索框以供用户检索 | <b>多渠道随时检索</b><br>唯一的档案检索入口不利于用户的检索体验。本平台无需到达指定唯一检索页面开展检索，而是在每个页面都设置检索入口，以使用户随时检索。         |
| 多字段检索 | 用户想要限制不同字段进行检索 | 提供多种字段限制检索内容    | <b>多字段满足检索要求</b><br>档案的内容、成分复杂，而人们进行档案检索的目标各不相同，有人想要提高查全率，而有人更看重查准率。多字段的设置有利于满足不同检索者的不同需求。 |
| 多方式筛选 | 用户想要筛选检索结果     | 提供多种方式二次筛选检索内容  | <b>多方式二次筛选所需</b><br>档案检索结果并不一定能满足检索者的最终需求，而二次筛选可以进一步缩小范围锁定想要寻找的内容。                         |

表 12 档案检索功能表

(1) 多入口检索



图 15 检索页

(2) 多字段检索

用户首先可以限制检索的事件，其次可以限制包括账户名称、账户 ID 号、话题、全文四种字段。



图 16 检索字段展示图

(3) 多方式筛选

可以对检索结果进行时间和用户的二次筛选。时间方面，可以将检索结果以顺序或倒序的方式排列，也可以具体限制某个月份；用户方面，将检索到的相关社交媒体用户以条数的多少按顺序排列，平台使用者可以只查看某一用户的相关档案。



图 17 检索结果筛选页面

3.2.3.3 辅助信息

3.2.3.3.1 关于我们

| 具体功能   | 功能需求                               | 功能介绍                    |
|--------|------------------------------------|-------------------------|
| 查看关于我们 | 用户希望了解本项目的开展背景、项目简介、平台主要内容、平台构建团队等 | 即关于项目本身的背景、目标以及平台的内容阐释。 |

表 13 “关于我们” 功能表

具体查看过程如下（见图 18）。



图 18 查看“关于我们”指示图

3.2.3.3.2 常见问题

| 具体功能 | 功能需求                      | 功能介绍                   |
|------|---------------------------|------------------------|
| 问题解答 | 用户希望获得使用平台的过程中遇到的问题的问题的解答 | 用户可以查看常见的关于平台内容和操作性的问题 |

表 14 常见问题功能表

从用户的角度考虑，平台设计了八个基本问题，包括对平台的基础了解、用户的基本使用和用户权益的保护途径，最后一个问答为用户的其他问题指明路径，获取解答过程如下（见图 19）。



图 19 进入常见问题页面指示图

3.2.3.3.3 免责声明

| 具体功能     | 功能需求                | 功能介绍                             |
|----------|---------------------|----------------------------------|
| 查看平台免责声明 | 用户在使用平台时希望明确平台的权责界限 | 明确平台的权责界限，以构建和谐有序的网络空间环境，避免权益纠纷。 |

表 15 免责声明功能表

具体查看过程如下（见图 20）。



图 20 查看免责声明指示图

### 3.3 科研成果

#### 3.3.1 理论成果

随着平台不断完善，实践过程不断深入，项目计划产出一定的理论成果。主要涵盖重大社会事件社交媒体存档平台开发研究报告、重大社会事件社交媒体存档核心概念的明确与界定、存档的体系化设计以及对实践过程中面临的问题、总结的经验进行理论分析。具体情况如下（见表 16）

| 理论成果                 | 主要内容                                  |
|----------------------|---------------------------------------|
| 重大社会事件社交媒体存档平台开发研究报告 | 对开发研究过程进行全过程回溯和系统性总结                  |
| 核心概念界定               | 明确重大社会事件社交媒体存档的具体内涵、构成要素等核心内容         |
| 存档体系设计               | 分析社交媒体存档的要素，设计存档体系                    |
| 实践问题的理论探讨            | 主要对实践过程中面临的问题，如合规性问题、政策法规缺失的问题，予以探讨   |
| 实践经验的理论总结            | 主要对实践过程中的成功经验，如存档框架、信息归档方案等，在理论层面予以分析 |

表 16 理论成果概览表

在实践过程中，作品涉及了归档、整合、存储与开发利用的档案管理全过程，因而项目团队希望以作品为依托，探讨社交媒体存档的有关问题，明确社交媒体存档的框架、要素、工具等内容，进而推动档案研究与档案工作在信息时代的转型。

### 3.3.2 实践成果

在实践方面，项目也希望建立一套标准化、可参考的重大社会事件社交媒体存档指南或标准。简言之，将从元数据标准、全流程存档框架等内容的制定以及信息归档、信息存储、信息开发利用等实践模式的完善这几个方面入手，对重大社会事件社交媒体存档的全流程、全要素予以明确和规范。

具体而言，已经或将要形成如下实践成果（见表 17）。

| 名称           | 内容                                              | 进展   |
|--------------|-------------------------------------------------|------|
| 社交媒体档案元数据方案  | 明确社交媒体存档应归档的信息元素                                |      |
| 社交媒体存档全过程框架  | 明确社交媒体存档所涉及的要素及各个要素之间的相互作用，形成社交媒体存档参考流程         | 初步形成 |
| 存档范围框架       | 设计存档事件、存档用户的划分标准以及存档平台的优先级，形成可扩充的、合理的、规范的归档范围框架 | 初步形成 |
| 社交媒体信息归档方案   | 对社交媒体信息归档过程中涉及的合规性、技术工具、归档要素等内容予以规定             | 初步形成 |
| 社交媒体信息存储方案   | 设计归档后的社交媒体信息物理存储、结构存储形式、存储时间、存取方式以及安全保密等相关内容    |      |
| 社交媒体信息开发利用方案 | 设计相对系统化、成体系的社交媒体信息开发流程及利用形式                     |      |

表 17 实践成果概览表

## 4 运作模式

### 4.1 STP 模式

#### 4.1.1 目标市场

(1) 首先,公众在网站中可以获取、浏览到重大社会事件的社交媒体信息,并且可以参与归档过程,对构建国家记忆和民族记忆有重要意义。

(2) 其次,专家学者可以从中获取重大社会事件的原始信息和数据,从而为其学术研究提供支撑。

(3) 此外,处于档案共同体的人可以在“档案领域”专题中寻找和构建档案共同体的专属记忆,且网站本身也可以作为档案学者研究社交媒体存档的实践案例,从而促进社交媒体存档的理论探索。

#### 4.1.2 市场定位

市场定位即设计出自己的产品和形象,从而在目标顾客中确定与众不同的有价值的地位。“与众不同”指该产品有独特的个性,从而与竞争产品有明显的差异;“有价值”指产品的这种个性和差异与目标顾客的需求相吻合,二者互相统一,缺一不可,以达到定位的有效性。<sup>[25]</sup>

本产品具有兼具社会价值和专业价值,拥有资源优势、科技含量、成果多样等多种优势(具体分析见 2.项目特色及优势)。而这些独特的优势与亮点正是为符合目标市场而设计的。基于此,本产品的定位如下:

一个开放、参与、互动的资源空间。用户可在该平台浏览、检索、获取重大社会事件社交媒体存档资源,并加以利用。同时,通过编研室、“故事树”等核心功能,用户可以参与社交媒体存档过程,推动构建多元叙事空间。

---

<sup>[25]</sup> 蓝进.试论市场定位、产品定位和竞争定位之间的关系[J].商业研究,2007(10):51-53.

## 4.2 产品策略

### 4.2.1 产品层次

产品层次是营销学上将产品的本质区分成五个层次，以探讨消费者对产品的感觉，以及如何进一步刺激消费者消费的改进方向。具体层次及解析如图所示：

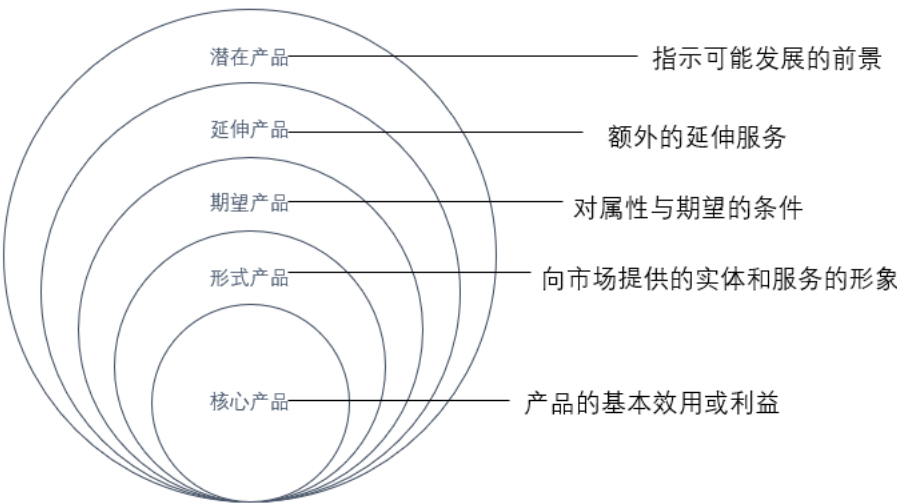


图 21 产品层次解析图

#### （1）核心产品

为访问者提供重大社会事件社交媒体存档信息的资源库，并设计、开放多种形式鼓励用户参与重大事件社交媒体存档过程。

#### （2）形式产品

以网站为主要形式展现组合产品，增强内容的可获得性；设计淡雅简洁的页面，使访问者使用的过程中赏心悦目。

#### （3）期望产品

访问者可以与网页、其他用户互动，上传重大社会事件社交媒体档案或信息开发作品，同时也可以与其他用户共同构建重大社会事件的相关资源。

#### （4）延伸产品

访问者可以获得存档全过程框架等科研或实践成果。

### （5）潜在产品

产品的平台未来可能会进一步深化信息技术，提高自动化水平；拓展信息来源，丰富档案形式；开发多种平台，扩大作品受众；创新利用方式，拓展平台功能。以从多角度多方面形成更加高技术、多资源、多功能的产品。

## 4.2.2 产品生命周期应对策略

产品生命周期，是指产品从准备进入市场开始到被淘汰退出市场为止的全部运动过程，主要是由消费者的消费方式、消费水平、消费结构和消费心理的变化所决定的。一般分为导入期、成长期、成熟期、衰退期四个阶段。典型的产品生命周期的四个阶段呈现出不同的市场特征，产品的推广策略也就以各阶段的特征为基点来制定和实施。

### （1）导入期策略

此时市场的需求小，但潜在需求在日益增长。此时应该进行合理的定位、有效地开拓目标市场，传播产品性能、用途信息、刺激需求。

### （2）成长期策略

此时市场需求大，大量访客涌入。应该树立产品形象，使访问者对产品产生喜爱，并不断改进产品功能，提高产品质量和服务质量。

### （3）成熟期策略

此时产品在时间的沉淀中获得了大多数人的认可和关注，应注重宣传产品理念和社会意义以树立形象，并不断提高知名度和提高产品质量，完善服务体系。

### （4）衰退期策略

此时由于需求偏好的转移或社会的发展，部分产品不具有实用性和研究性，此时应调整管理范围或集中管理少数相对有一定优势和不断在使用的产品，集中做好突出部分。<sup>[26]</sup>

---

<sup>[26]</sup> 黄文馨.产品生命周期的研究[J].商业研究,2003(17):13-15.

### 4.3 资金运转来源

资金运转来源即本产品的后期维护和持续性发展所需要的资金支持的来源。

由于本产品为公益性产品，因此获得的经费主要来自于社会支持。政府鼓励、科研经费、研究人员和其他共识之士的主动资助是项目团队得以持续性发展以为个人记忆、专业领域和社会发展做出贡献的主要来源，同时项目也可以通过众筹平台或宣传活动以获得更多的社会支持。

部分资金也来源于后期为个人或机构提供重大事件社交媒体存档与开发利用服务所获得的报酬。

## 5 结语

2020 年的新冠肺炎疫情、国庆 70 周年阅兵以及全国两会等重大社会事件对我国社会生活均产生了深远影响，这种影响超出现实世界，直达网络空间。社交媒体中的各类主体，全方位记录了这些重大社会事件。本产品聚焦与重大社会事件紧密相连的社交媒体信息，通过形成资源库，构建网站和资源开发，记录这一重大事件，讲述数字时代的中国故事。

网络空间中新生成的信息不断覆盖旧有信息，长此以往，社交媒体中关于重大社会事件的信息有多少可以留存便不得而知。在该作品中，项目团队主动归档相关社交媒体信息，同时也鼓励用户积极上传亲身经历或真实事件，从而更加全面且尽可能真实的反映重大社会事件，帮助构建中国记忆，促进民族认同。

### 5.1 现有不足

资源来源和类别以及平台形式单一。在现阶段，该作品所存储的社交媒体信息均源自新浪微博，且只有网页这一平台。社交媒体信息来源单一，不利于从社交媒体信息的角度对重大社会事件进行全面记录。纳入归档范围的社交媒体信息类别和平台形式缺乏多样性，对用户分享社交媒体信息和后续的二次开发与利用有所限制。

已归档资源开发利用形式较为单一。档案资源的二次开发是深入发掘档案价值的重要手段。目前，该作品的二次开发利用部分已初显规模，但面对丰富的社交媒体资源以及日益多元化的用户需求，开发利用部分仍需进一步拓展。

合规性问题被搁置。合规合法的作品能有效确保利益相关者的利益得到维护。目前，该作品在新浪微博对社交媒体信息爬取的过程中并未获得平台授权，由此而导致的信息爬取的合规性问题存在侵权风险。

## 5.2 未来发展方向

### 5.2.1 资源的扩充与联动

（1）本项目将归档更多的重大社会事件，不断扩充事件集。

（2）本项目将充实事前、事中、事后三个阶段的全过程信息资源收录，动态追踪与存档每个事件的不同阶段，以使事件过程更加完整，推动事件资源的持续性生长与扩充。

（3）本项目力图推动同其他形式的档案资源（如政府文件、个人档案等）进行联动，构建多来源多种类的档案集合，促进不同档案资源的联动发展。

### 5.2.2 技术与工具的优化

（1）于数据捕获，本项目力图采用自动化或半自动化资源捕获技术，以提升社交媒体存档的效率。于档案的组织，本项目力图引入自然语言处理等技术助力档案管理实现自动标引、自动分类，从而丰富和整序平台的事件集资源。于档案的开发利用，本项目希望采用 3D 建模等数据可视化工具在形式和内容上创新档案的开发利用方式，同时将该平台延伸至网站以外的形式，如 APP、微信小程序等。

（2）以实践为基础构建社交媒体存档工具方案，推动其常态化使用，将此方案推广至其他社交媒体归档项目的数据捕获、组织和开发利用等方面。

### 5.2.3 存档共同体的构建

（1）于资源，本项目希望与不同机构、行业等达成合作，促进双方资源的相互补充，推动多样形式资源的融合与利用。

（2）于技术，本项目希望同技术人员或相关行业组织达成合作，共享合作双方的技术支撑，以项目社交媒体存档工具方案推动其社交媒体或网页存档，

以对方所掌握的技术能力推动本平台技术与工具的深入。

（3）于专业理论，本项目希望同档案以及相关领域如图情领域的研究人员或相关机构达成合作，共同探索存档规范与指南的制定,推动社交媒体存档的发展。

#### **5.2.4 专业理论与方法的系统化建构**

（1）于理论，目前已经初步形成包括存档框架和学术论文等的科研成果，后期将在不断实践的基础上提炼出重大社会事件社交媒体存档平台开发研究报告、核心概念界定、存档体系设计等理论成果，以为相关领域的研究人员提供借鉴。

（2）于方法，本项目力图为社交媒体存档和网页归档方面的规范与指南的建立、发展与完善贡献力量，从实践与理论的结合中推动方法、规范与指南的可操作化、系统化的发展。

（3）同时，本项目致力于将平台发展为一个有价值、有内涵、可挖掘的范例，为计算档案学或其他相关领域提供研究案例。

## 6 参考文献

- [1] 大英图书馆.UK web archive[EB/OL].<https://www.webarchive.org.uk/en/ukwa/index>,2004-00-00.
- [2] 国际网络保存联盟内容开发工作组.IIPC 奥运会/残奥会收藏[EB/OL].<https://netpreserveblog.wordpress.com/?s=Helena+Byrne>,2010-00-00.
- [3] 杜克大学.杜克大学档案馆网络档案收藏[EB/OL]. <https://library.duke.edu/rubenstein/findingaids/uawebarchive/#collectionoverview>,2010-00-00.
- [4] 浙江省档案局.浙江省档案局关于做好防控新型冠状病毒感染的肺炎疫情期间档案工作的通知[EB/OL].[http://www.zjda.gov.cn/art/2020/1/30/art\\_1378509\\_41963671.html](http://www.zjda.gov.cn/art/2020/1/30/art_1378509_41963671.html),2020-01-30.
- [5] 江苏省档案局.关于面向社会各界征集抗击新冠肺炎疫情防控工作档案资料的公告[EB/OL].[http://www.dajs.gov.cn/art/2020/2/14/art\\_155\\_48850.html](http://www.dajs.gov.cn/art/2020/2/14/art_155_48850.html),2020-02-14.
- [6] 黄新荣,曾萨.网页归档推进策略研究——基于网页归档生态系统视角[J].图书馆学研究,2018(16):63-70+16.
- [7] 闫晓创.日本网络资源存档项目实践研究[J].浙江档案,2017(12):20-23.
- [8] 黄新荣,高晨翔.国内外社交媒体存档研究与实践述评[J].图书情报工作,2019,63(04):122-134.
- [9] 周文泓.全球社交媒体归档行动概览与展望[J].浙江档案,2016(12):16-19.
- [10] 全国档案事业发展“十三五”规划纲要[J].中国档案,2016(05):14-17.
- [11] Archive Social. Social media archiving [EB / OL]. [2019-08-18]. <http://www.archive-social.com>.
- [12] 周文泓.互联网信息社会化保存的冷思考与热展望[J].图书馆论坛,2020,40(01):87-95.
- [13] 周文泓.我国网络空间中档案领域的缺位审视和参与展望——基于社交媒体信息保管行动的解析[J].档案与建设,2019(09):13-17.
- [14] WILEY J.The Onslaught of Big Data and the Information Governance Imperative Information Governance for SocialMedia [M] //ROBERT F.Smallwood. Information Governance: Concepts, Strategies, and Best Practices. New York City: Wiley, 2014: 1-22.
- [15] BUSHEY J.Convergence, Connectivity, Ephemeral and Performed: New Characteristics of Digital Photographs[J].Archives & Manuscripts, 2014 (1): 33-47.
- [16] 周文泓.社交媒体平台作为档案库的内涵及特点解析[J].档案学研究,2019(02):104-110.
- [17] 王新才, 徐欣欣. 国外个人数字存档的实践经验及其启示[J]. 信息资源管理学报, 2016(4): 109-115.
- [18] Helen Hockx-Yu. Preserving Social Media in the Context of Non-print Legal Deposit [EB/OL]. [2016-09-09] <http://library.ifla.org/999/>.
- [19] 高晨翔, 黄新荣. 我国社交媒体文件的保管政策研究[J]. 图书馆研究, 2017(7): 47-55.
- [20] 周文泓.社交媒体信息保管模式与策略探析[J].档案与建设,2018(09):17-21.

- [21] 新浪微博数据中心. 2018 微博用户发展报告[EB/OL].  
<https://data.weibo.com/report/reportDetail?id=433>, 2020-04-15.
- [22] 蒋冠,冯湘君.国家综合档案馆在社交媒体信息存档中的角色定位研究[J].档案学研究,2018(01):10-13.
- [23] 特里·库克,李音.宏观鉴定与职能分析[J].中国档案,2012(01):51-53.
- [24] 冯惠玲. 数字记忆:文化记忆的数字宫殿[J]. 中国图书馆学报,2020,v.46;No.247(03):4-16.
- [25] 蓝进.试论市场定位、产品定位和竞争定位之间的关系[J].商业研究,2007(10):51-53.
- [26] 黄文馨.产品生命周期的研究[J].商业研究,2003(17):13-15.

**文本 2：**  
**重大社会事件社交媒体存档平台建设方案**

## 引言

为确保“重大社会事件社交媒体存档平台”的开发满足档案专业性与合规性，完整的开发方案需要得到建构，从而提供全过程的指导与具体依据。因而，为保证实践过程的科学规范、遵循现有法律法规和行业标准、保证归档事件的全面性和代表性以及技术的可实现性，通过制度调查、案例研究、文本分析等方式，最终形成重大社会事件社交媒体存档平台建设方案。

具体而言，第一，参照了档案以及相关领域的政策法规，为保证重大社会事件社交媒体存档提供制度依据。第二，调研国内外网络信息存档案例，基于从整体到要素层的提取、解析与归纳，初步形成了重大社会事件社交媒体存档的全过程框架。第三，调研与考察重大社会事件社交媒体信息情境，基于其内容、形式、背景等特征设计通用的归档范围。第四，充分分析国家政策以及行业对信息捕获的相关规定，确认重大社会事件社交媒体存档的合规要求。第五，针对存档需求，结合现有工具和方法，形成技术指南。上述行动能充分保证实践过程的专业性、科学性和合法性。

重大社会事件社交媒体存档平台建设方案具体如下：

目 录

1 社交媒体存档专业制度依据 ..... 1

2 存档全过程框架 ..... 2

3 存档范围 ..... 5

    3.1 社交媒体平台的选取 ..... 5

    3.2 重大社会事件的识别 ..... 6

    3.3 主体层级的账户选择 ..... 7

    3.4 信息层级的范围设定 ..... 8

    3.5 归档示例 ..... 9

4 合规性要求 ..... 12

    4.1 社交媒体存档合规性总调查 ..... 12

    4.2 爬虫使用的法律合规性调查 ..... 12

5 技术指南 ..... 22

    5.1 信息捕获 ..... 22

    5.2 信息存储 ..... 23

    5.3 网站建设 ..... 24

6 参考文献 ..... 27

## 1 社交媒体存档专业制度依据

由于缺乏针对性和系统的重大社会事件社交媒体存档依据，项目通过调查现有的政策法规与规范标准，识别和选取可参考的内容，形成用以全程指导存档平台开发的制度依据。例如，本团队调研和参考了国内外相关政策法规与标准规范等相关制度（见表1）。

| 制度参照            |                             | 参照内容                                |
|-----------------|-----------------------------|-------------------------------------|
| 标准号             | 名称                          |                                     |
| ISO 16175       | 《信息与文献——电子办公环境中文件管理原则与功能要求》 | 信息管理的指导原则与功能要求，以保证归档过程符合规范、满足要求     |
| ISO 23081       | 《信息与文件 文件管理流程 文件元数据》        | 信息的管理流程以及元数据的规范标准                   |
| DA/T 80—2019    | 《政府网站网页归档指南》                | 网页归档的收集与整理的一般方法，以及应记录的相关元数据，如标题、作者等 |
| GB/T 18894—2016 | 《电子文件归档与电子档案管理规范》           | 有保存价值的电子信息的收集、整理、归档与管理方法            |
| DA/T78—2019     | 《录音录像档案管理规范》                | 录音录像电子文件的收集、整理、著录及归档要求              |
| DA/T63—2017     | 《录音录像类电子档案元数据方案》            | 录音录像电子档案的元数据实体及其元数据构成               |
| DA/T54—2014     | 《照片类电子档案元数据方案》              | 照片类电子档案的元数据实体及其元数据构成                |
|                 | 《都柏林核心元素集》                  | 参考其中的信息著录格式、及元数据构成                  |
|                 | 《重大活动和突发事件档案管理办法》           | 对重大活动的划分及其对档案管理流程的有关规定              |

表1 存档的制度依据

## 2 存档全过程框架

为得出重大社会事件社交媒体存档的参考框架，以指导重大事件社交媒体存档工作，项目首先对国内外代表性的社交媒体存档案例进行调研。经搜索引擎检索、档案馆与图书馆网站及 IIPC 网站调查可发现，美国、加拿大、英国、德国、丹麦、乌克兰、中国、新加坡、马来西亚、澳大利亚等均有代表性的社交媒体存档项目，建设出大大小小的重大社会事件数字记忆资源库，涵盖政治、文化体育、经济、社会等各领域，覆盖从官方组织、主流机构到社会群体与个人形成的各类社交媒体。

随后，结合档案学客体全宗将信息之间的来源联系明确为同一事件过程中信息的有机关联，以此认识重大社会事件的信息对象的前提依据、社交媒体存档的基本要素如主体、对象、方法、流程（包括捕获、整合、保存、发布等），对项目具体内容进行分类提取，结合各实践项目可供提取的内容，经由归纳和调整，明确了存档主体、社交媒体对象、存档方案、存档结果、存档保障这五大要素及其分要素。

最后，基于对现有重大社会事件社交媒体存档项目的归纳和分析，形成了相对通用的行动框架，具体如下（见图 1）：

从实践来看，相对完备的重大社会事件社交媒体存档框架包括如下要素及其作用：

（1）存档主体，是重大社会事件信息的利益相关者，主要包括发起者与协同者。发起者也可视为存档的主导者或主要行动主体，贯穿倡议、启动、统筹、设计与实施等全流程的活动。作为存档主体，发起者与协同者基于职责、目标以及能力共同发现重大社会事件及其社交媒体信息，深化对信息形成情况、价值以及特征的认识，依此形成社交媒体存档需求与行动的基本认知，为存档形成组织机制，明确存档目标与方案。

(2) 社交媒体信息对象，即重大社会事件在网络平台的相关记录。它是存档信息的来源，也是存档的作用对象，依据事件性质对群体和深度进行记录。它影响着信息形成的基本情况如数量、形式、内容等，体现了不同的价值与特征，这对判定存档主体的存档必要性、程度、所需资源和具体方案都产生影响。其中，更为重要的是通过相对全面的调查帮助梳理和完善重大社会事件的主要脉络和维度，为从背景层理解并认识信息和整体关联提供基础。

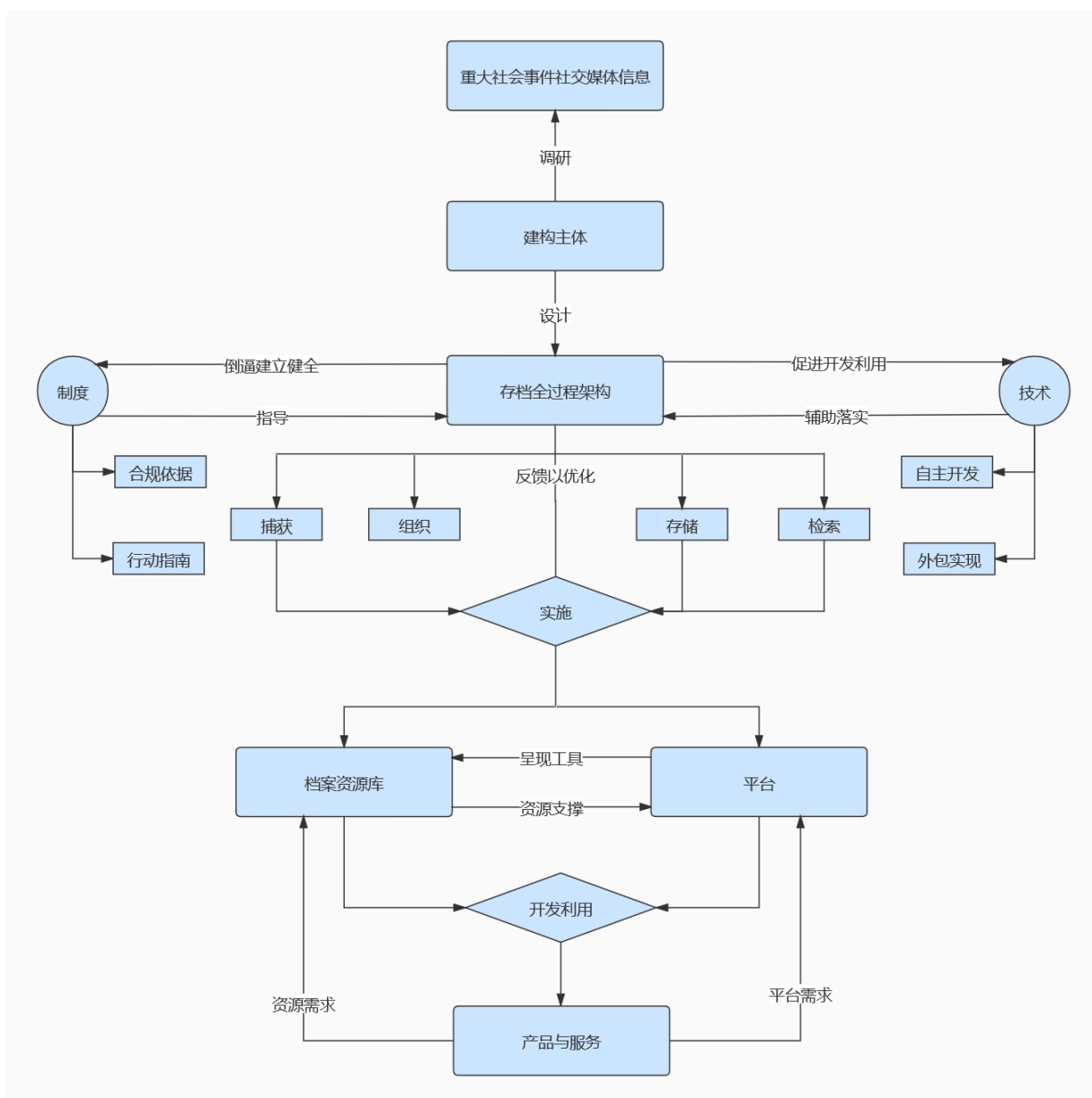


图1 重大社会事件社交媒体存档参考框架

(3) 存档保障，用以全程指导与落实存档方案的实施。制度在广义上包含

用以指导重大社会事件社交媒体存档的规则集合，帮助形成存档方案。重大社会事件社交媒体往往具有跨背景、多信息生产者的特征，因而信息边界与权责难以简单划分，全方位存档指导不可缺失。制度的作用主要体现在两个方面，首先是为存档从必要性和风险规避提供合规依据，如信息所有权、利用权、隐私、知识产权等。然后是为存档提供行动指导。从理论基础、方法、原则、规范等层面落实存档行动的整体模式、流程与活动细节。

另一个存档保障是技术，它是落实存档方案的关键辅助。社交媒体存档从方案到落地实施亟需技术的配套使用，从捕获到发布乃至开发利用均需技术支持。存档主体既可根据需求自主开发，也可向专业领域寻求“外包”支持或改造共享开放的技术。

（4）存档方案，用以落实捕获、整合、保存、发布的过程要素总设置。存档方案的设计是一种过程性活动，例如确认存档范围、捕获频率与方式、信息保存格式、质量标准、发布条件等。需要说明的是，整合是将社交媒体有序化的过程，包括组织、描述、鉴定等具体环节，且社交媒体信息的特征是开放，因而保持合规的“线上可见”特质也是存档重要内容。同时，存档方案将随着实施效果及反馈情况予以动态调整，并倒逼制度的建立健全与技术的优化完善。

（5）存档结果。一是网络档案库与呈现平台，即存档方案实施的直接成果，是留存重大社会事件记忆与证据的信息资源集成与展示的窗口。它的内容与形式直接反映存档方案的实施成效。二是网络档案产品与服务。除了向社会提供重大社会事件的社交媒体信息原材料，存档主体可逐渐通过深层次开发，形成“信息密度”更高的产品与服务。

### 3 存档范围

存档范围的划分和确定将直接对现阶段信息存档的全面性和代表性产生影响，同时其规范性和可拓展性也会对未来补充事件和资源发挥作用。因此，设计一个符合现阶段要求并兼顾未来需求的归档范围极为重要。目前，项目主要从社交媒体平台的选取、重大社会事件的识别、主体层级的账户选择以及信息层级的范围设定四方面对存档范围予以明确。

#### 3.1 社交媒体平台的选取

通过对社交媒体平台的梳理与调研，分析当前各平台的信息密度、分布规律以及具体的内容、形式、载体等特征，结合当前的技术水平和开发能力，将本次存档的平台划分为以下三个层次：

| 存档优先级 | 社交媒体平台                    | 原因                          |
|-------|---------------------------|-----------------------------|
| 第一级   | 微博、微信                     | 信息密度高，且便于获取                 |
| 第二级   | 今日头条、抖音、快手、知乎、豆瓣          | 多为二次传播信息，存档价值较低。            |
| 第三级   | Twitter、Facebook等国外社交媒体平台 | 相关信息少而分散，且存档国外社交媒体平台存在现实困难。 |

表 2 存档平台优先级列表

在第一层级，微信的隐私内容相较于微博更多，其信息的集中程度与呈现方式较之微博都有不足。所以，目前项目的归档平台限定为第一层级的微博。

### 3.2 重大社会事件的识别

为兼顾归档事件的全面性和代表性，项目针对重大社会事件的识别与划分主要从两个方面展开，一是从地域的角度，二是从行业的角度。

首先，从地域的角度而言，项目通过考察重大社会事件的地域归属及其显著程度，将重大社会事件划分为国家级或地区级。国家级的重大社会事件一般指在全国范围内产生重大影响，且国家层面以一定身份参与进来，并在社交媒体上引起全民广泛关注或热烈讨论的事件，如全国两会、国庆 70 周年阅兵等；地区级的重大社会事件一般指对本地区产生重大影响的事件，如 2021 年成都大运会、上海全面实行垃圾分类等。这类事件的利益相关者往往为地区内的政府机关、社会组织以及社交媒体上的个人用户，因而归属地区级重大社会事件。

其次，考虑到部分事件的地域属性及其显著程度有限，但其行业或专业属性显著。因此，项目还依照国民经济行业分类标准 (GB/T 4754-2002) 将这部分事件按照不同的行业进行存档和划分。此外，作为档案共同体的一员，项目将原本属于行业的“档案领域重大事件”抽离出来，作为一个独立的事件集合，从而促进档案共同体的专业认同，塑造集体记忆。

最后，若需要对某一事件进行归档，则先根据其地域属性进行判断，是否可以归入国家层面或地区层面；如若地域属性不明，则考虑将其归入行业层面。

具体而言，目前纳入归档范围的事件及其所属层次如下（见表 3）。

| 层级   | 事件         | 归档情况  |
|------|------------|-------|
| 国家层面 | 新冠肺炎疫情     | 基本完成  |
|      | 国庆 70 周年阅兵 |       |
|      | 民法典草案通过    |       |
| 地区层面 | 2021 成都大运会 | 正在进行中 |
|      | 上海率先实行垃圾分类 |       |
|      | 2020 乌镇戏剧节 |       |

|      |      |              |       |
|------|------|--------------|-------|
| 行业层面 | 教育行业 | 高考改革         | 正在进行中 |
|      | 文体行业 | 中国电竞——IG 夺冠  |       |
|      | 信息行业 | 5G           |       |
| 档案领域 |      | 2020 国际档案日   |       |
|      |      | 《档案法》修订      | 正在进行中 |
|      |      | 互联网信息社会化保存项目 |       |

表 3 归档事件识别表

### 3.3 主体层级的账户选择

在社交媒体上活跃着众多性质、类别各异的用户群体，不同的群体往往会就同一事件发布信息、陈述事实或发表观点，这些社交媒体信息都是对该事件的反映。因此，通过对信息主体进行合理的层次和类别划分，将有效了解信息源头，全面反映并记录事件。

根据社交媒体中用户的特点及相关规范，项目首先将信息主体划分为个人和非个人信息主体。由于非个人信息主体涵盖范围较广，涉及主体较多，宜进一步细化为党组织机构、国家机关、新闻媒体和其他社会组织。在第三层级，则对上述主体做进一步细化，如国家机关进一步细化为全国及地方各级人大、国务院及地方各级人民政府、最高司法机关及地方各级司法机关。具体如下（见表 4）。

|      | 第一层级    | 第二层级   | 第三层级                           |
|------|---------|--------|--------------------------------|
| 信息主体 | 非个人信息主体 | 党组织结构  | 中央组织机构、地方组织机构、基层组织机构           |
|      |         | 国家机关   | 全国人大及地方各级人大、国务院及地方各级人民政府、司法机关、 |
|      |         | 新闻媒体   | 主流央级媒体、地方媒体、自媒体                |
|      |         | 其他社会组织 | 社会团体、企业、事业组织                   |

|  |        |  |                  |
|--|--------|--|------------------|
|  | 个人信息主体 |  | 个人用户 1、个人用户 2、…… |
|--|--------|--|------------------|

表 4 主体账户归档列表

最后，从事件出发，分析可能与该事件产生关系或有明显业务活动的相关主体，形成归档主体列表。

### 3.4 信息层级的范围设定

在确定主体层级的同时，主体的哪些社交媒体信息应进行存档、存档的社交媒体信息应该包含哪些要素这些问题也需要加以考虑。

一方面，就信息内容而言，归档主体的哪些信息应当存档主要从与事件的关联性大小来考虑，此时较为稳妥的办法是通过设计筛选条件对信息进行过滤。但与传统的信息不同，社交媒体信息往往自身会带有一定的标签。此外，针对某一具体的事件，社交媒体平添往往不会按照一定的逻辑对事件进行细致的划分，进而把事件按照这一划分结果加以管理。

所以在确定相关信息筛选条件的时候，首先应该对平台内涉及该事件的社交媒体信息进行一定的调研，初步整理出涉及该事件的一系列标签或话题。随后归纳出标签或话题的共性特征，得出关键词。最后再对关键词进行一定的补充或表述方式的改变，从而形成一组筛选条件。此外，在归档过程中，为最大限度保证查全率，则需将筛选条件以“或”关系进行关联。

另一方面，信息的哪些要素（如背景、形式或内容等）应该归档需要考虑信息开发利用的有关问题，但同时也必须保证信息的真实性、完整性、可用性、安全性。

面对这一问题，首先应该了解平台内的社交媒体信息涵盖哪些要素，既包括信息的显性数据（如信息发布时间、位置信息等），也应包括信息的隐性数据（如平台配置数据）。其次，在了解信息要素的基础上，依照《信息与文献 文件管理

过程 文件元数据 第 1 部分：原则》（GBT 26163.1-2010）等原则和规范的元数据标准，确定必须捕获的元数据要素。从而从信息内容、信息背景、信息形式等多方面记录数据，保证数据的真实性、完整性、可用性、安全性。最后，在适当考虑开发利用需求的基础上，合理补充应捕获的数据要素

### 3.5 归档示例

最后，为进一步理解存档范围如何从事件层逐步细化至数据层，下面以“2021 成都大运会”的存档范围确定为例进行说明。

首先，项目选定“新浪微博”为此次的归档平台，同时由于该事件的地域属性明显，且社交媒体上该事件的主要利益相关者为地方政府，故将其归入地区层面的重大社会事件。

其次，从该事件为“地区层面的重大社会事件”这一定性出发，可以得出应归档的主体大致应该包括：（1）成都市地区和基层党组织；（2）成都市的国家机关（包括人大、政府以及司法机关等）；（3）四川及成都地方的主流媒体、自媒体等；（4）各类文体组织、大运会协会等社会组织；（5）就大运会发布过信息的个人。具体归档用户列表如下（见表 5）

| 用户性质  | 用户名称                                    | 资源类型 |
|-------|-----------------------------------------|------|
| 党组织机构 | 成都共青团、青聚锦<br>官城、文明成都、天府龙泉、武侯发布、<br>郫都青年 |      |

|      |                                                             |                                                                                             |
|------|-------------------------------------------------------------|---------------------------------------------------------------------------------------------|
| 国家机关 | 四川发布、成都市政府门户网站、成都发布、成都高新、双流黄水、运动成都                          | 1、业务活动，如第 36 届国际大体联全体会议、大运会筹备工作会议等。                                                         |
| 新闻媒体 | 大运频道、微成都、四川日报、四川观察、成都新闻广播、一点橙子、听堂 FM、成都交通广文艺广播、HELLOChengdu | 2、赛事筹备信息，如场馆建设情况、志愿者培训情况、配套设施建设情况等赛事筹备进程。<br>3、宣传活动信息，如创意文案征集活动、文创产品销售推广活动以及大运进社区、大运进校园等活动。 |
| 社会组织 | 成都 2021 世界大运会、FISU 国际大体联、成都同城、成都志愿者 volunteer               |                                                                                             |
| 个人用户 | 成都娱乐家、锦城时报、微言四川                                             |                                                                                             |

表 5 归档用户概况表

第三是筛选条件的确定。通过对平台内 2021 成都大运会的相关内容进行初步调研，列出了微博中关于的大运会的几个话题和标签，具体有“成都大运会”、“成都 2021 世界大运会”、“成都世界大运会日期确定”等。随后抽取出其中的共性特征，即“成都大运会”，最后对这一共性特征进行一定的补充或重组形成“2021 成都大运会”、“成都大运会”、“2021 世界大学生运动会”、“2021 年成都大运会”等筛选条件，依照这些关键词对归档用户发布的社交媒体信息进行筛选。

最后，具体至数据层面，新浪微博上每条社交媒体信息及用户信息的数据要

素如下（见表 6）。

| 信息类别 | 信息要素                                                                                                 |
|------|------------------------------------------------------------------------------------------------------|
| 用户信息 | 用户 id、用户昵称、性别、生日、所在地、教育信息、职业信息、阳光值用、微博注册时间、微博数、关注数、粉丝数、用户简介、主页地址、头像 url、高清头像 url、微博等级、会员等级、认证类型、认证信息 |
| 微博信息 | 微博 id、微博正文内容、头条文章 url、原始图片 url、视频 url、微博发布位置、微博发布时间、点赞数、转发数、评论数、微博发布终端、结果文件                          |

表 6 数据要素表

在参考相关标准之后，最后拟捕获的数据主要包括内容信息、责任者（用户）信息以及背景信息。具体情况如下（见表 7）。

| 数据类别 | 归档要素                                                |
|------|-----------------------------------------------------|
| 背景信息 | 微博发布时间、微博发布位置、微博发布终端、结果文件                           |
| 微博信息 | 微博 id、微博正文内容、头条文章 url、原始图片 url、视频 url、微博点赞数、转发数、评论数 |
| 用户信息 | 用户 id、用户昵称、用户性别、所在地、用户简介、职业信息、教育信息、学习经历、认证类型、认证信息   |

表 7 拟存档数据表

## 4 合规性要求

由于社交媒体信息的存档涉及多方主体的多样利益,为明确整个项目开启的合规性以及运行过程中需要注意的法规政策等,本团队开展了对相关文献和政策法规的合规性调查,并将其分为总调查和存档的重点环节——爬虫获取资源两部分。

### 4.1 社交媒体存档合规性总调查

(1)《全国档案事业发展“十三五规划纲要”》中指出,要研究制定重要网页资源的采集和社交媒体文件的归档管理办法,但至今还未出台具体文件。

(2)我国仅有《档案法及实施办法》《电子公文归档管理暂行办法》《文书类电子文件元数据方案》《电子档案移交与接收办法》等与社交媒体记录捕获归档有一定关系,但尚未制定与捕获社交媒体信息直接相关的法规政策。比如:“电子文件应当连同元数据一并收集。”

(3)《档案法》《著作权法》《个人信息保护法》《电子证据法》《政府信息公开法》等相关法律制度,对政务社交媒体文件在捕获、归档和利用过程中可能涉及的知识产权、隐私权、文件的法律证据效力、公民信息权保护等相关法律问题提供必要的法律依据。

(4)我国《档案法》、《公共图书馆法》没有规定合法收割范围、收割行为、开发利用方式,可能会增加与网站沟通的难度,产生版权、个人隐私等纠纷。

总的来说,经对文献和法律法规的调研,未能发现对社交媒体存档过程直接相关的政策规定,现有文件大多是侧面制约。

### 4.2 爬虫使用的法律合规性调查

由于目前尚未有直接规定社交媒体存档各方面的法律法规,本项目对社交媒

体存档中最基本的一个环节——使用爬虫获取资源，展开了合规性调查。

### 4.2.1 使用爬虫的法律风险

#### 4.2.1.1 不遵守 Robot 协议的法律风险

Robots 协议是技术界为解决爬取方和被爬取方之间通过计算机程序完成关于爬取的意愿沟通而产生的一种机制。根据中国互联网协会《互联网搜索引擎服务自律公约》第七条，机器人协议（robots 协议）是指互联网站所有者使用 robots.txt 文件，向网络机器人（Webrobots）给出网站指令的协议。具体而言，robots 协议是网站所有者通过位于网站根目录下的文本文件 robots.txt，提示网络机器人哪些网页不应被抓取，哪些网页可以抓取。

关于 Robots 协议的法律属性，我国当前法律并没有明确规定，但《互联网搜索引擎服务自律公约》第七条规定，“遵循国际通行的行业惯例与商业规则，遵守机器人协议（robots 协议）”。第八条规定“互联网站所有者设置机器人协议应遵循公平、开放和促进信息自由流动的原则，限制搜索引擎抓取应有行业公认合理的正当理由，不利用机器人协议进行不正当竞争行为，积极营造鼓励创新、公平公正的良性竞争环境。”

虽然《互联网搜索引擎服务自律公约》仅适用于中国互联网协会会员单位和自愿加入《中国互联网行业自律公约》的互联网从业单位，但在司法实践中，robots 协议已经被认定构成互联网行业搜索领域内工人的商业道德：北京市第一中级人民法院在百度诉奇虎不正当竞争案件中，将为行业内所公认的 Robots 协议认定为互联网行业搜索领域内公认的商业道德。法院在判决中指出“在被告推出搜索引擎伊始，其网站亦刊载了 Robots 协议的内容和设置方法，说明包括被告在内的整个互联网行业对于 Robots 协议都是认可和遵守的。其应当被认定为行业内的通行规则，应当被认定为搜索引擎行业内公认的、应当被

遵守的商业道德”。将 Robots 协议定性为行业惯例、自律、道德规范与行业规范,也是当前的主流观点。

作为一种柔性规范, Robots 协议不对任何一个主体进行强制。然而,就民法的基本原则而言,首要的平等原则要求按照市场制度的互惠行事,在互联网领域体现为互联网公司自觉遵守协议这样的行业惯例和道德契约,因此爬虫抓取他站的数据必然受到他站的限制。其二,禁止权利滥用原则,是指民事主体在民事活动中应当正确行使民事权利,若行使权利会损害到同样受到保护的他人利益和公共利益时,则构成权利滥用。网站设置协议是对自身权益的合理保护,声称打破垄断者可能恰恰是损害同样受到保护的平等主体和公共利益的滥用权利者。其三,公序良俗原则是指民事主体的行为应当遵守公共秩序,符合善良风俗,不得违反社会公共秩序和一般道德。协议作为延续至今的行业惯例,现有状态下遵守之就是互联网行业的一般道德。协议的设置有助于保护隐私、个人数据,平衡着各方利益,保证着行业的有序竞争和可持续发展,反之,必然损害行业未来和公共利益,违反公平竞争。根据等价有偿原则,民事主体在民事活动中应当等价交换,不得无偿占有、非法侵害他方利益。据此,如果违背网站协议肆意抓取数据,即便是非盈利行为,也不能构成“合理使用”等抗辩事由。

由于 Robots 协议属于格式条款,因此其合同效力应以是否为合理使用为标准进行区分:一般情况下,如果协议限定抓取范围的内容属于合理使用,则其同网络运营商用户协议类似的格式条款为有效条款,搜索引擎爬虫抓取行为一经开始,即为对该站协议的承诺,合意达成,合同成立并生效,若爬虫按照协议的规定抓取所需数据、信息,则为守约,反之违约;若协议限定抓取范围的内容系滥用,由于爬虫抓取动作已经开始,合同已成立并生效,但合同条款全部或部分无效,违反其限定或部分条款进行抓取的,不构成对相应条款的违

反；对于直接对爬虫进行选择，禁止某些爬虫进入数据库的协议内容，若其为合理使用，则被禁爬虫不进入该站的不作为行为构成对该协议的承诺，合同成立，若仍然进入进行数据抓取，则不构成违约责任，但直接产生前文所述侵权责任、竞争法律责任等之可能；若禁止爬虫的协议内容为滥用，与前述同理，违反其内容进入该站数据库的，显然不构成对协议的承诺，无合意，合同未成立，但由于未经行会或主管部门认定，擅自闯入仍可能受到扰乱行业秩序的行政处罚。

根据微博的 Robots 协议，除百度、360、谷歌、搜狗、必应、好搜、易搜外的爬虫不允许抓取的内容为主目录，即微博所有内容。因此本项目使用爬虫爬取微博数据是违背该协议的，虽然我国法律无具体规定，但违反该协议属于违背行业规范的行为，仍具有一定法律风险。同时，微博禁止爬虫的协议内容是否构成滥用也对其效力构成影响。

#### 4.2.1.2 抓取特定类型信息的风险

##### 4.2.1.2.1 抓取信息属于受著作权法保护的作品

被著作权法保护的作品有以下特点：

（1）是具有特定内容的智力创作。作品是直接从事文学、艺术、科学创作活动的智力活动的产物，而不是其他智力劳动，如技术发明的产物。

（2）具有独创性。独创性是指作品由作者独自构思、运用自己的技巧与方法直接创作、反映了自己的个性与特点。抄袭、剽窃产生的“作品”由于不具备独创性，因此不能成为受保护的對象

（3）具有一定的表达形式。文学、艺术、科学创作必须通过文字、符号、语言、声音、造型、色彩等某种形式表达出来，使人们能够感知到它的存在，由此才能认可它是“作品”。

（4）可以以有形的方式复制。具有可复制性即能被固定在某种载体上，只

是作品应该具有的属性，但并不是说所有作品，尤其是口头作品、舞台艺术作品等，只有先被固定在某种载体上，才能被认可。

根据《著作权法》，不具备以上几种特点，不受保护的作品有：抄袭、剽窃产生的作品不受保护；法律、法规，国家机关的决议、决定、命令和其他具有立法、行政、司法性质的文件，及其官方正式译文；时事新闻；历法、通用数表、通用表格和公式；已过著作权法保护期的作品；依法禁止出版、传播的作品。

基于著作权自动取得的原则，微博一经创作完成，作者就享有著作权。但《著作权法》保护的对象必须有独创性。即不是所有的微博都享有著作权，就文字微博来说，如果只是表达博主今天的心情或者简单陈述一件事，并不具有独创性。但是，如果角度独特、内容新颖，自然属于《著作权法》意义上受到保护的作品。

协议下搜索引擎数据挖掘行为版权侵权责任的成立以其抓取的内容属于《著作权法》保护客体为前提。网络环境下，版权法关注的核心由复制权转向传播权，《信息网络传播权保护条例》就体现了由保护复制到保护传播的理念变化。就数据提取和使用行为而言，如果爬虫控制者抓取信息后在自己的网站上公开传播抓取到的信息，则还有可能进一步侵犯信息网络传播权。如果一项复制行为并不以传播为目的，或没有引起传播的后果，那么这种复制对于权利人来说并无损害。在自身商业模式正当性的限度内，搜索引擎数据挖掘行为对所抓取的数据并不产生传播的后果，而仅仅是提供链接、简介等桥梁服务，没有直接向公众发行、展示作品。

我国《著作权法》规定，信息网络传播行为的构成需同时满足“以有线或者无线方式向公众提供作品、使公众可以在其个人选定的时间和地点获得作品”两个要件。其中提供作品，根据 WCT 第 8 条的原文释义，指提供作品的可能性，而没有要求确实获得了作品。根据爬虫的工作原理，快照系复制网页才能获取，因此必然以网页制作者在网络上发布为前提，据此，快照是基于作者在网络上的首

先发表才使公众有获取该作品的可能，因此快照的商业模式不会侵犯信息网络传播权。但通常意义上的快照业务，整个实现过程是自动完成、没有人工干预的，因而不成立侵权责任。但利用入口资源截取快照、抄袭其他搜索引擎搜索结果的行为，其分析、建立索引等数据挖掘过程中，体现了根据自己需求对内容的选择，如存储时间的长短、对象等的情况，不能以技术、自动、机器等避风港原则中的要素免责，由于整个过程并非自动完成及没有人为干预，构成复制行为，侵犯了版权人权利束中的复制权以及网页作品的复制权或非独创性的信息财产权。可见，提供快照的商业模式下，也应做到合理使用。

就本项目而言，收集的信息内容涉及新闻、政府公布的实时数据和政策法规等不属于《著作权法》保护范畴内，而对于满足《著作权法》要求的作品，使用网页快照的形式不会侵犯其信息网络传播权，但涉及人为干预的情况下，需合理使用信息，例如仅提供缩略图或限制信息长度，否则可能构成直接或间接侵权。

#### 4.2.1.2.2 抓取信息属于用户个人信息

根据《中华人民共和国网络安全法》第七十六条，“个人信息，是指以电子或者其他方式记录的能够单独或者与其他信息结合识别自然人个人身份的各种信息，包括但不限于自然人的姓名、出生日期、身份证件号码、个人生物识别信息、住址、电话号码等。”

根据《民法总则》第 111 条，“任何组织和个人需要获取他人个人信息的，应当依法取得并确保信息安全。不得非法收集、使用、加工、传输他人个人信息”；《网络安全法》第四十四条，“任何个人和组织不得窃取或者以其他非法方式获取个人信息。”因此，如果爬虫在未经用户同意的情况下大量抓取用户的个人信息，则有可能构成非法收集个人信息的违法行为。

如果涉及到个人信息，不仅抓取行为本身具有很大法律风险，其后续行为也可能构成违法犯罪。我国《刑法》修正案（九）中将刑法第二百五十三条进

行了修订，明确规定在未经用户许可的情况下，非法获取用户的个人信息，情节严重的也将构成“侵犯公民个人信息罪”。根据《最高人民法院最高人民检察院关于办理侵犯公民个人信息刑事案件适用法律若干问题的解释》第五条规定，对“情节严重”的解释，（1）非法获取、出售或者提供行踪轨迹信息、通信内容、征信信息、财产信息五十条以上的；（2）非法获取、出售或者提供住宿信息、通信记录、健康生理信息、交易信息等其他可能影响人身、财产安全的公民个人信息五百条以上的；（3）非法获取、出售或者提供第三项、第四项规定以外的公民个人信息五千条以上的便构成“侵犯公民个人信息罪”所要求的“情节严重”。此外，未经被收集者同意，即使是将合法收集的公民个人信息向他人提供的，也属于刑法第二百五十三条之一规定的“提供公民个人信息”，可能构成犯罪。

与用户个人信息密切相关的是隐私权。在世界范围内，隐私没有统一、成熟的定义，各种学说对隐私的精神内核均无争议，但隐私的外延是模糊的。因为隐私的范畴与社会文化相关，随着社会经济文化发展而变化。陈笑认为，依据与人格利益是否相关这一标准划分隐私的外延是周延的，并且与今天我国互联网发展态势相适应。在互联网领域，用户要求网络运营商负担切断用户网络行为与身份对应性的义务，其隐私需求主要在于其网络行为与身份的不对应，这种情况下，对网站个人用户而言，姓名、肖像也是隐私，而只能根据个案具体问题具体分析，因此，在网络运营商和用户的关系中，“凡是不想让别人知道的就是隐私”的主观主义标准仍然具有合理性。这一层面，用户的网络行为等个人数据，只要没有与用户身份关联而使他人能将这些数据与该用户本身对应，就不属于隐私。[2]

在微博、微信、博客等社交平台，用户可以上传自身的信息后设置访问限制，或者使用加密功能只有自己可以访问，因此，该等信息仍有可能属于用户的个人隐私。因此，如果爬虫控制者绕开上述限制，而抓取用户的该等隐私，又或者在

抓取后公开传播该等信息，造成对用户的损害后果的，则有可能侵犯了相关用户的隐私权。

#### 4.2.1.3 强行突破网站设置的技术措施的法律风险

由于爬虫的批量访问会给网站带来巨大的压力和负担，许多网站经营者会采取技术手段，以阻止爬虫批量获取自己网站信息。常见的技术措施包括：

（1）通过 UA 识别爬虫、（2）设置 IP 访问频率，如果超过一定频率，弹出验证码（3）通过并发识别爬虫；（4）请求的时间窗口过滤统计；（5）限制单个 ip/apitoken 的访问量；（6）识别出合法爬虫（7）蜜罐资源等。

虽存在多种不同的技术，但该等技术本质都是网站运营者采取的阻止爬虫批量访问和抓取信息的技术手段，因此从法律上并没有实质性区别。而针对该等技术手段，爬虫开发者可以通过优化自己的代码、使用 IP 池等多种方式规避上述技术措施，实现对网站信息的批量抓取和复制。

由于网络爬虫会根据特定策略尽可能多的访问页面，因而爬虫的使用将占用被访问网站的网络带宽并增加网络服务器的处理开销，甚至无法正常提供服务。另一方面，虽然网站上的信息是公开的，但在没有爬虫的情况下，一般经营者较难批量复制网站上的信息，例如抓取竞争对手的价格以进行比价，或者实时调整自己的价格。而在使用爬虫技术的情况，则能够实现上述模式。

《反不正当竞争法》第十二条第二款规定“经营者不得利用技术手段，通过影响用户选择或者其他方式，实施列妨碍、破坏其他经营者合法提供的网络产品或者服务正常运行的行为：…其他妨碍、破坏其他经营者合法提供的网络产品或者服务正常运行的行为。”虽此仅是一个《反不正当竞争法》网络条款中的兜底条款，但其体现了法律禁止通过技术手段进行对他人产品的非法干扰的总体原则。因此，如果网站运营者已经采取了一定的反爬虫措施，而爬虫控制者基于经营目的、强行突破网站运营者采取的反爬虫技术措施，并客观导致

被抓取网站的正常运行，则大可能构成上述规定所规制的不正当竞争行为。当然，此种情形下，鉴别实施者身份仍然是追责的巨大障碍。

此外，强行突破某些特定被爬方的技术措施，还可能构成刑事犯罪行为。

《刑法》第二百八十六条规定，违反国家规定，对计算机信息系统功能进行删除、修改、增加、干扰，造成计算机信息系统不能正常运行，后果严重的，构成犯罪，处五年以下有期徒刑或者拘役；后果特别严重的，处五年以上有期徒刑。而违反国家规定，对计算机信息系统中存储、处理或者传输的数据和应用程序进行删除、修改、增加的操作，后果严重的，也构成犯罪，依照前款的规定处罚。

#### 4.2.1.4 违反用户协议的法律风险

根据最新版《微博服务使用协议》1.3款，“用户使用微博服务过程中，对自身使用微博服务的行为以及通过微博服务发布、公开的任何信息享有合法权利，承担相应法律责任。同时，用户同意：微博运营方对微博内容（微博内容即指用户在微博上已发布的信息，例如文字、图片、视频、音频等）享有使用权；未经微博运营方事先书面许可，用户不得自行或授权、协助任何第三方非法抓取微博内容，“非法抓取”是指采用程序或者非正常浏览等技术手段获取内容数据的行为。”以及1.4款“用户在使用微博服务过程中应当严格遵守微博运营方所发布的Robots协议。未经微博运营方事先书面同意，任何用户不得以任何方式自行或委托任何第三方以违反上述规定的方式访问微博平台或收集任何微博内容。”4.5款“为维护微博平台的稳定运营，确保用户体验质量，未经微博运营方事先书面许可，任何人不得擅自在微博平台上实施自动化行为或发布垃圾信息。自动化行为，是指未经微博运营方同意，用户自行或授权、协助第三方采用自动化手段或明显异于常人的、远高于正常用户频率地发布微博、评论、私信、头条文章或作出关注、点赞、抓取数据等的行为。”

#### 4.2.2 关于爬网的意见与建议

为规避使用爬虫技术带来的风险，爬虫控制者在使用爬虫时应：（1）严格遵守网站设置的 robots 协议；（2）在规避反爬虫措施的同时，需要优化自己的代码，避免干扰被访问网站的正常运行；（3）在设置抓取策略时，应注意编码抓取视频、音乐等可能构成作品的数据，或者针对某些特定网站批量抓取其中的用户生成内容；（4）在使用、传播抓取到的信息时，应审查所抓取的内容，如发现属于用户的个人信息、隐私或者他人的商业秘密的，应及时停止并删除。

基于微博用户服务协议的要求，建议先向微博运营方提出申请，在获得其书面许可后再使用爬虫爬取数据。

## 5 技术指南

在著录规则、元数据标准以及归档范围和事件等档案领域的专业要素确定后，需要对技术方案予以明确，从而对信息捕获、信息存储以及信息呈现方式加以规定。由此，需要对技术方案进行整体介绍。

### 5.1 信息捕获

Python 具有众代码简洁的特点，且可直接使用的模块应用程序数量多，可有效降低开发的时间成本。其中的 **Selenium** 模块应用程序为 web 浏览器的自动化提供了各种工具和依赖包，能够很好的自动化测试各种各样的浏览器，能够调用浏览器接口模拟浏览器的一系列行为。因此，项目使用选择 **Selenium** 模块应用程序捕获相关事件及用户的社交媒体信息。具体流程如下：

（1）抓取环节是第一步。在对新浪微博进行爬取的过程中，通过读取并分析整理基于宏观鉴定和职能分析而确定的归档用户列表，访问上述用户的主页面，抓取网页上的文字、链接、图片等信息。

（2）过滤环节是指页面信息被爬虫抓取之后，会被暂时存放在临时数据库中。这些暂时存放在临时数据库中的网页信息将被根据页面的质量接受筛选过滤处理，从而决定该页面是否被收录与否。

（3）通过筛选环节页面则会进入到收录、排序环节。在该环节，依据来源原则，把页面快照按不同来源进行归档、排序，结构化信息直接存储在数据库中。

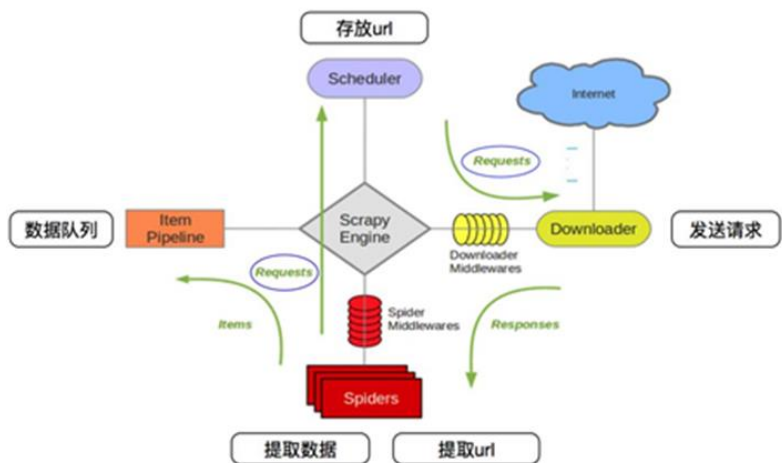


图 2 爬虫原理图

## 5.2 信息存储

目前，平台共存储了两种类型的数据，一是数据库存储的结构化信息，二是网页快照。

面对结构化数据，项目在实践中选择利用 MySQL 数据库进行存储。该数据库具有开源、多线程编程的特点，提供的接口支持多种语言连接操作，拥有快速且稳定的基于线程的内存分配系统，运行过程十分稳定。

面对网页快照，项目在实践中采取 MHTML 作为数据档案的存储格式。MHTML 也称网页归档，可将一个多附件网页(如包含大量图片、Flash 动画、Java 小程序)存储为单一档案，即网页封存档案。选择 MHTML 作为数据格式，是因为 MHTML 是一种 WEB 电子邮件档案，其保存的网页只具有一个文件，这样利于档案保存和分类。

两类数据的存储结构如下（见图 3—图 4）。

```
mysql> desc user;
+-----+-----+-----+-----+-----+-----+
| Field | Type | Null | Key | Default | Extra |
+-----+-----+-----+-----+-----+-----+
| id    | varchar(20) | NO | PRI | NULL |  |
| nickname | varchar(30) | YES |  | NULL |  |
| gender | varchar(10) | YES |  | NULL |  |
| location | varchar(200) | YES |  | NULL |  |
| birthday | varchar(40) | YES |  | NULL |  |
| description | varchar(140) | YES |  | NULL |  |
| verified_reason | varchar(140) | YES |  | NULL |  |
| talent | varchar(200) | YES |  | NULL |  |
| education | varchar(200) | YES |  | NULL |  |
| work | varchar(200) | YES |  | NULL |  |
| weibo_num | int(11) | YES |  | NULL |  |
| following | int(11) | YES |  | NULL |  |
| followers | int(11) | YES |  | NULL |  |
+-----+-----+-----+-----+-----+-----+
13 rows in set (0.00 sec)

mysql>
```

图 3 结构化数据存储结构

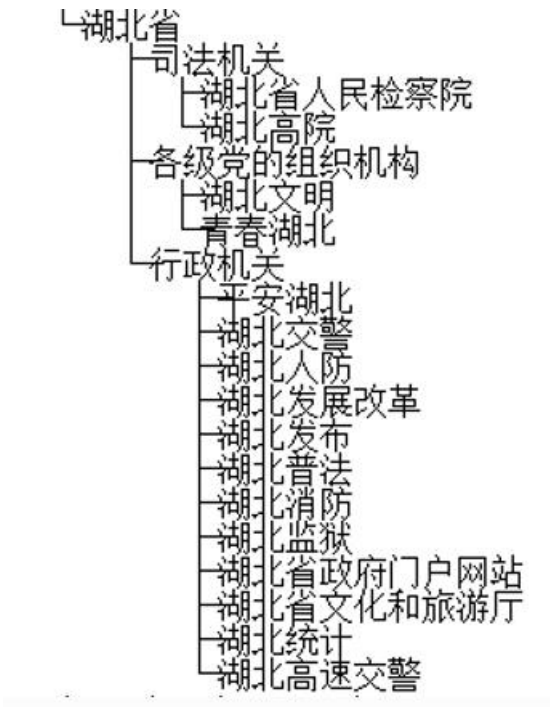


图 4 网页快照存储结构

5.3 网站建设

超文本标记语言（HTML）是用来描述网页的一种语言,是整个网页的基础骨架，定义网页内容。它包括一系列标签，通过这些标签可以将网络上的文档格式统一，使分散的 Internet 资源连接为一个逻辑整体。HTML 文本由 HTML 命令组成，可以说明文字，图形、动画、声音、表格、链接等，用于搭建网页的结构和

骨架。

层叠样式表(CSS)是 CSS 是一种定义样式结构如字体、颜色、位置等的语言，定义如何现实 HTML 元素，规定网页布局。开发人员只需修改 CSS 文件，就可以改变站点的所有页面，大大提高了网页制作的效率，有助于实现模块化开发。

JavaScript 是一种脚本编写语言，它是连接前台（HTML）和后台服务器的桥梁,具有读写 HTML 元素，对浏览器事件做出响应，控制 cookies 等功能，进而定义网页的行为。

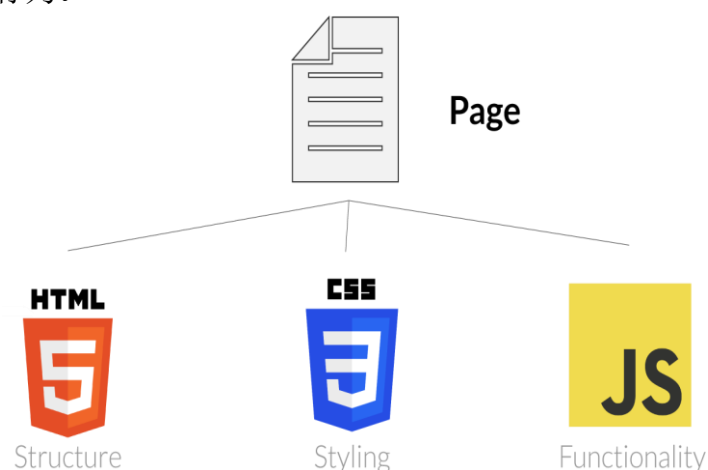


图 5 HTMLCSS&JS 功能图

该网站架构可分为“前端”、“后端”、“数据层”、“运行环境”四层。前端由后台管理系统和网页客户端组成，前者由项目团队负责，用来管理整个系统；而后者则面向用户，负责成果展示。后端以 Flask 框架为基础，负责诸如“流量监控”、“业务处理”、“请求响应”、“日志记录”等功能。数据层则以 MongoDB 为基础框架，利用网络爬虫技术获取所需数据后，将其按照一定组织结构存放于数据库中，以供上层调用。运行环境层方面，拟租借云服务器，采用 Nginx+gunicorn +CentOs7.2 的部署模式进行整个底层环境的构建。具体内容如下（见图 6）。

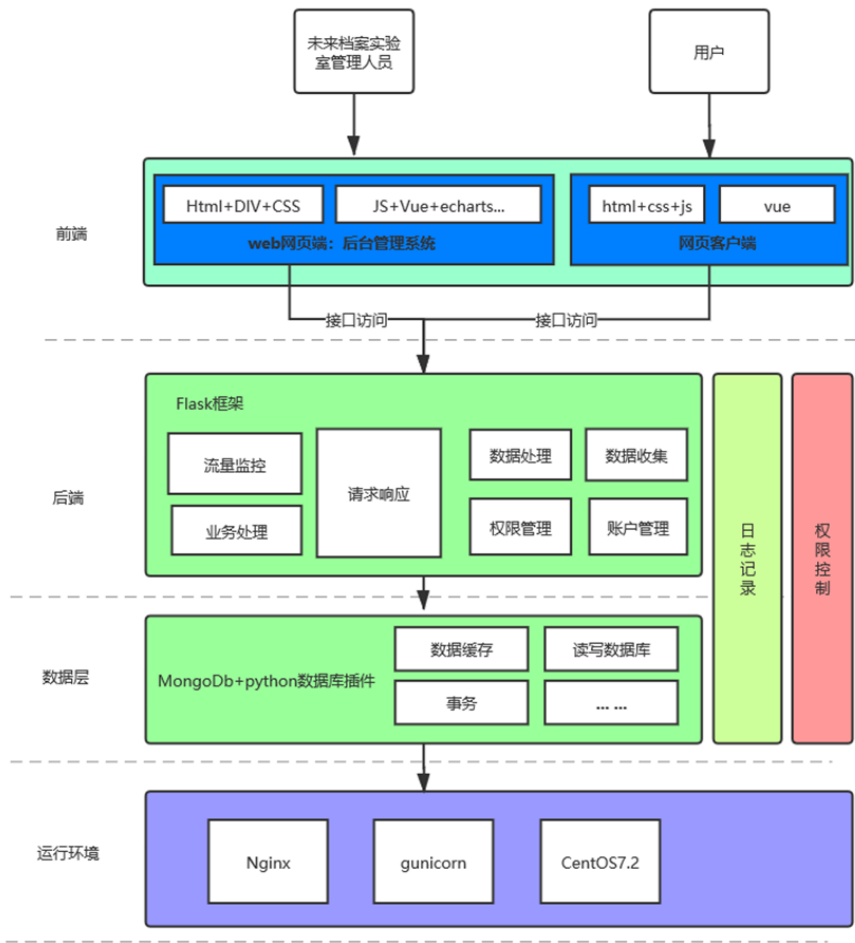


图 6 网站架构图

## 6 参考文献

- [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 web archive[EB/OL]. [2020-02-20]. <https://www.webarchive.org.uk/en/ukwa/index>.
- [4] TAMIMEN LIBRARY OF NEW YORK UNIVERSITY. Special collections[EB/OL]. [2020-02-20]. [https://specialcollections.library.nyu.edu/search/?f%5Brepository\\_sim%5D%5B%5D=tamwag](https://specialcollections.library.nyu.edu/search/?f%5Brepository_sim%5D%5B%5D=tamwag).
- [5] UNIVERSITY OF TEXAS AT AUSTIN LIBRARY. Web archive[EB/OL]. [2020-02-20]. <http://legacy.lib.utexas.edu/hrdi/>.
- [6] DUKE UNIVERSITY ARCHIVES. Uawebarchive[EB/OL]. [2020-02-20]. <https://library.duke.edu/rubenstein/findingaids/uawebarchive/#collectionoverview>.
- [7] NATIONAL LIBRARY OF THE CZECH REPUBLIC. Webarchiv[EB/OL]. [2020-02-20]. <https://www.webarchiv.cz/en/>.
- [8] NATIONAL ELECTRONIC ARCHIVES OF UKRAINE. Collections[EB/OL]. [2020-02-20]. <https://tsdea.archives.gov.ua/en/>.
- [9] DEUTSCHE NATIONALBIBLIOTHEK. Web archive[EB/OL]. [2020-02-20]. [https://www.dnb.de/EN/Professionell/Sammeln/Sammlung\\_Websites/sammlung\\_websites\\_node.html#doc272684bodyText1](https://www.dnb.de/EN/Professionell/Sammeln/Sammlung_Websites/sammlung_websites_node.html#doc272684bodyText1).
- [10] NATIONAL LIBRARY OF NEW ZEALAND. New Zealand web archive[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>.
- [11] INTERNATIONAL INTERNET PRESERVATION CONSORTIUM. CDG collections[EB/OL]. [2020-02-20]. <https://netpreserveblog.wordpress.com/?s=Helena+Byrne>.
- [12] JIAYILIUJIAYI. 2020nCov individual archives[EB/OL]. [2020-02-20]. [https://github.com/jiayiliujiayi/2020nCov\\_individual\\_archives](https://github.com/jiayiliujiayi/2020nCov_individual_archives).
- [13] NATIONAL ARCHIVES. Webarchive[EB/OL]. [2020-02-20]. <http://www.nationalarchives.gov.uk/webarchive/>.
- [14] BANQ. Services[EB/OL]. [2020-02-20]. [https://www.banq.qc.ca/services/bibliotheque\\_nationale/depot\\_legal/collecte\\_sites/](https://www.banq.qc.ca/services/bibliotheque_nationale/depot_legal/collecte_sites/).
- [15] NATIONAL LIBRARY OF AUSTRALIA. Pandora overview[EB/OL]. [2020-02-20]. <http://pandora.nla.gov.au/overview.html>.
- [16] LIBRARY OF CONGRESS. Web-archiving[EB/OL]. [2020-02-20]. <https://www.loc.gov/programs/web-archiving/about-this-program/>.
- [17] 台湾大学. NTU[EB/OL]. [2020-02-20]. <http://webarchive.lib.ntu.edu.tw/eng/default.asp>.
- [18] 陈永生, 王沐晖, 苏焕宁, 杨茜茜. 基于互联网政务服务平台的文件归档与管理: 治理观[J]. 档案学研究, 2019(06): 4-11.
- [19] 华芳芳, 叶楚健. 我国网络信息存档研究的现状述评[J]. 档案与建设, 2018(10): 42-45+37.
- [20] 黄新荣, 曾萨. 网页归档推进策略研究——基于网页归档生态系统视角[J]. 图书馆学研

究, 2018(16):63-70+16.

[21]陈永生, 苏焕宁, 杨茜茜, 王沐晖. 基于互联网政务服务平台的文件归档与管理:事由观[J]. 档案学研究, 2018(02):4-13.

[22]谭彩敏. 社交媒体文件归档管理研究:鉴定标准与管理模式[J]. 浙江档案, 2018(03):10-12.

[23]冉朝霞. 档案学视角下网络舆情大数据的采集、研判和归档研究[J]. 档案管理, 2018(01):57-58.

[24]周文泓, 耿越, 黄榆涵. 个人社交网络信息归档管理研究[J]. 浙江档案, 2017(08):33-35.

[25]孙红蕾, 郑建明. 互联网信息资源长期协作保存机制研究[J]. 图书馆学研究, 2017(10):20-25+10.

[26]高晨翔, 黄新荣. 我国社交媒体文件的归档政策研究[J]. 图书馆学研究, 2017(07):47-55.

[27]罗刚《网络爬虫全解析:技术、原理与实践》, 电子工业出版社, 第65-66页

[28]陈笑. 论 Robots 协议下搜索引擎数据挖掘行为的法律责任[D]. 北京邮电大学, 2014.

**文本 3：**  
**重大社会事件社交媒体存档平台相关附件**

目 录

附件 1：常见问题文本 ..... 1

附件 2：项目简介文本 ..... 3

附件 3：免责声明文本 ..... 5

## 附件 1：常见问题文本

### 1. 网站的构建者是谁？

本网站的指导老师为周文泓（总指导）及乔健教授（总顾问），主要构建者包括四川大学未来档案员实验室团队的四名成员：张晓宇、吴琼、贺谭涛、黄小宇；以及四川大学计算机学院的两名本科学生：王淼、白立忱。

### 2. 网站包括哪些重要内容？

本网站共包含三个重要内容，一是“馆藏档案”，包括支持多样字段检索的事件集档案和支持主体检索的快照特藏库；二是“用户贡献”，即与社交媒体平台用户的互动板块，用户可上传相符的档案以共同铸就更加丰富的重大社会事件记忆；三是“开发利用”，包括专题展览、线上编研室、“故事树”和“回忆沙龙”板块。

### 3. 网站的目的是什么？

本网站旨在构建以“新浪微博”为主要社交媒体平台的“重大社会事件记忆”，以期为想要了解和研究重大社会事件的人们提供有组织、有价值的丰富的档案资源。

### 4. 我如何发现记忆？

(1)通过主页面的一框式检索或事件集中各事件的一框式检索获取记忆；(2)通过特藏库中的主体检索可以获取您想要寻找的相关主体的记忆；(3)通过浏览专题展或编研室成果可以获取相关记忆。

### 5. 我如何分享记忆？

可通过导航栏的“用户贡献”进入贡献档案界面，按要求描述相关内容即可。

### 6. 我可以分享什么样的内容？

在社交媒体上形成的与某一重大社会事件有关内容都可以进行分享，可以图片、视频、文本等形式；可以是聊天记录、推送、动态等内容；可以是事件期间

的感想、做过的事情等的记录。

**7. 我上传的记忆将作何使用？**

本网站将您上传的记忆呈现于网站内，用以丰富本网站的重大社会事件记忆，以共享给有兴趣的人们和做研究的学者。同时，本网站的相关条款保护您的记忆不作商业用途。

**8. 如果发现与我相关的记忆但不想再公开在本网站，我要如何处理？**

您可以立即与我们联系，我们会在第一时间作相关处理。

**9. 我还有其他问题怎么办？**

您可以立即与我们联系，我们会在第一时间回复您的问题。

## 附件 2：项目简介文本

### 1. 项目背景

重大社会事件指在一个地区、国家或民族产生深远影响的事件，对塑造国家和民族的记忆与认同具有重要意义。从 2020 年初的新冠肺炎疫情，到即将在成都举办的 2021 年世界大学生运动会，再到档案领域备受关注的互联网信息社会化保存项目，不难发现社交媒体这个数字空间成为了展现和反映这些重大社会事件的重要组成部分。在此种现实情况下，本项目组聚焦于国家、地区、行业以及档案领域的重大社会事件，以社交媒体为阵地，展开了对重大社会事件社交媒体存档的实践探索。

### 2. 项目目标

本网站为四川大学重大社会事件社交媒体信息存档项目组的成果之一，旨在综合运用信息技术、档案管理、资源管理、人文、艺术等多学科领域的方法、技术和工具，构建以“新浪微博”为主的社交媒体平台的重大社会事件记忆，并基于此进行档案的二次开发利用，鼓励用户参与重大社会事件社交媒体存档，构建多元主体共建的叙事空间。

### 3. 项目内容

本网站共包含三个重要内容，一是“馆藏档案”，包括支持多样字段检索的事件集档案和支持主体检索的快照特藏库；二是“用户贡献”，即与社交媒体平台用户的互动板块，用户可上传相符的档案以共同铸就更加丰富的重大社会事件记忆；三是“档案服务开发”，包括线上编研室、专题展览、“故事树”和“回忆沙龙”板块。

### 4. 团队介绍

本作品由四川大学公共管理学院秘书档案系与四川大学未来档案实验室指导建构完成，为国家社会科学基金青年项目“基于多元数字技术的网络空间参与

式归档研究”（项目号 18CTQ037）所支持，是“重大社会事件社交媒体存档”项目实践成果，具体团队成员为：

项目顾问：乔健教授

总负责人：周文弘副研究员

学生成员：四川大学公共管理学院 2017 级档案学专业张晓宇；四川大学公共管理学院 2018 级档案学专业吴琼、贺谭涛、黄小宇；计算机学院 2018 级计算机科学与技术专业王淼、物联网工程专业白立忱

### 附件 3：免责声明文本

访问者在接受本网站服务之前，请务必仔细阅读本条款并同意本声明。访问者访问本网站的行为以及通过各类方式利用本网站的行为，都将被视作是对本声明全部内容的无异议的认可；如有异议，请立即与本网站协商，并取得本网站的书面同意意见。

第一条 访问者在从事与本网站相关的所有行为（包括但不限于访问浏览、利用、转载、宣传介绍）时，必须以善意且谨慎的态度行事；访问者不得故意或者过失的损害或者弱化本网站的各类合法权利与利益，不得利用本网站以任何方式直接或者间接的从事违反中国法律、国际公约以及社会公序良俗的行为，且访问者应当恪守下述承诺：

- 1、传输和利用信息符合中国法律、国际公约的规定、符合公序良俗；
- 2、不将本网站以及与之相关的网络服务用作非法途径以及非正当用途；
- 3、不干扰和扰乱本网站以及与之相关的网络服务；
- 4、遵守与本网站以及与之相关的网络服务的协议、规定、程序和惯例等。

第二条 本网站郑重提醒访问者：请在转载、上载或者下载有关内容时务必尊重该内容的版权、著作权；如果您发现有您未署名的内容，请立即和我们联系，我们会在第一时间加上您的署名或作相关处理。

第三条 除我们另有明确说明或者中国法律有强制性的规定外，本网站用户原创的内容，本网站及作者共同享有版权，其他网站及传统媒体如需使用，须取得本网站的书面授权，未经授权严禁转载或用于其他商业用途。

第四条 本网站内容仅代表作者本人观点，不代表本网站的观点和看法，与本网站立场无关，相关责任作者自负。

第五条 本网站有权将在本网站内发表的作品用于其他用途，包括网站、电子杂志等，作品有附带版权声明者除外。

第六条 未经本网站和作者共同同意，其他任何机构不得以任何形式侵犯其作品著作权，包括但不限于：擅自复制、链接、非法使用或转载，或以任何方式建立作品镜像。

第七条 本网站所刊载的各类形式（包括但不限于文字、图片、图表）的作品仅供参考使用，并不代表本网站同意其说法或者描述，仅为提供更多信息。对于访问者根据本网站提供的信息所做出的一切行为，除非另有明确的书面承诺文件，否则网站不承担任何形式的责任。

第八条 访问者在本网站注册时提供的一些个人资料，本网站除您本人同意以及第九条规定外不会将用户的任何资料以任何方式泄露给任何一方。

第九条 当政府部门、司法机关等依照法定程序要求本网站披露个人资料时，本网站将根据执法单位之要求或为公共安全之目的提供个人资料。在此情况下之任何披露，本网站均得免责。

第十条 由于用户将个人密码告知他人或与他人共享注册账户，由此导致的任何个人资料泄露，本网站不负任何责任。

第十一条 本网站如无意中侵犯了某媒体、公司、企业或个人等的知识产权，请来电或致函告知，本网站将在规定时间内给予删除等相关处理，若有涉及版权费等问题，请及时提供相关证明材料并与我们联系，通过友好协商公平公正原则处理纠纷。

第十二条 以上声明内容的最终解释权归重大社会事件社交媒体存档平台网站所有。