arXiv:2305.01071v1 [cs.DL] 1 May 2023

HTML 正确，JSON 错误：重放采用客户端渲染构建的归档网页所面临的挑战

MICHELE C. WEIGLE， Old Dominion University，
美国MICHAEL L. NELSON， Old Dominion University，美国SAWOOD ALAM， 互联网档案馆，
美国MARK GRAHAM， 互联网档案馆，美国

许多网站正在转变其网页的构建方式。传统模式是将内容嵌入服务器端HTML中，并通过 HTTP 响应返回给客户端。越来越多的网站转向另一种模式：初始 HTTP 响应只包含 HTML 骨架以及 JavaScript，后者会向多个服务器发起 API 调用以获取内容（通常为 JSON 格式），然后在客户端构建 DOM，从而更轻松地定期刷新页面内容，并动态修改内容。这种客户端渲染如今已在 Twitter 和 Instagram 等社交媒体平台上占据主导地位，CNN.com 等新闻媒体也开始采用。当传统网页存档技术（如使用 Heritrix 爬取）应用于客户端渲染内容的页面时，JSON 响应可能会与其嵌入的 HTML 页面不同步，导致重放时出现时间违规。由于违规的 JSON 不会直接显示在页面中（也就是说，不像违规的嵌入图像那样可见），因此这种时间违规很难检测。我们介绍了 CNN.com 顶层页面自 2015年4月 起如何使用客户端渲染，以及这对网络档案造成的影响。2015年4月24日 至 2016年7月21日 期间，我们发现近 15,000 个网页存档快照，其中基础 CNN.com HTML 与用于提供主报道内容的 JSON 响应之间存在超过 2 天的时间违规。缓解这一问题的一种方法是使用基于浏览器的爬取，而不是 Heritrix 等传统爬虫，但目前基于浏览器的爬取速度远慢于 Heritrix 等非浏览器工具。

1 引言

人们越来越多地将网页档案的内容用作证据——例如，证明某位公众人物曾表示支持一项不受欢迎的立场 [27, 29, 35], , 证明某位政治候选人歪曲了自己的背景 [36], , 证明某个政府机构阻止公众获取一份报告 [39], , 甚至作为法律案件中的证据 [25, 30], , 还有许多其他例子[1] [18, 19]。要让人们有理由信赖网页档案，就必须能够相信，从网页档案中重放的内容就是当时捕获的内容；或者至少，用户能够轻易辨别显示内容与捕获内容之间的重大差异。尽管我们知道，归档网页的重放，即网络存档快照，可能会随着时间推移而出现一些差异 [8]，原因包括 JavaScript 执行、回放软件升级，甚至档案迁移 [9], , 但我们仍然期望网页存档快照的主要内容能够反映被捕获的实时网页内容。然而，我们发现了一些案例：使用客户端技术并通过 JSON 文件填充页面内容的网页，可能会造成严重的时间违规，并导致网络档案重建

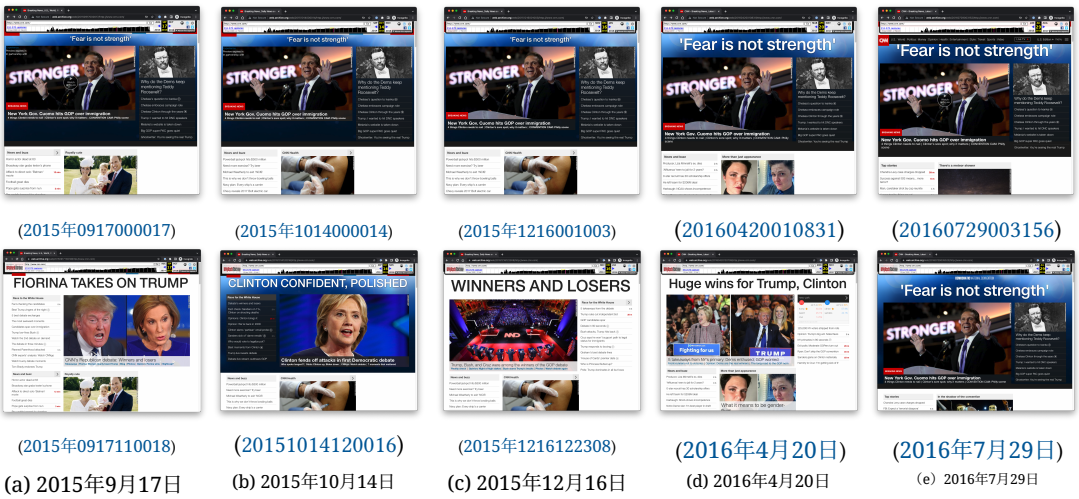
[1] https://web.archive.org/web/20221118192715/https://archive.org/about/news-stories/search?mentions-search=Wayback+Machine

作者地址：Michele C.Weigle，Old Dominion University，美国弗吉尼亚州Norfolk，23529，mweigle@cs.odu.edu；Michael L. Nelson，Old Dominion University，美国弗吉尼亚州Norfolk，23529，mln@cs.odu.edu；Sawood Alam，互联网档案馆，美国加利福尼亚州San Francisco，94118，sawood@archive.org；MarkGraham，互联网档案馆，美国加利福尼亚州San Francisco，94118，mark@archive.org。

允许个人使用本材料。对于其他所有用途，无论是在当前还是未来的任何媒体中，均必须获得IEEE的许可，包括为广告或宣传目的重印/再版本材料、创作新的合集作品、转售或重新分发至服务器或列表，或在其他作品中重复使用本作品的任何受版权保护的组成部分。© 2023 IEEE 经许可在2023年ACM/IEEE数字图书馆联合会议（JCDL）上发表的论文预印本版本

网页可能包含在原始捕获日期时间时并不存在的内容。使用这些技术构建的网页会给网页档案爬虫和回放系统带来严重挑战。

我们之所以关注 CNN.com 的主页面，是因为它是一个热门网页、更新频繁，并且被大量存档。此前已有相关研究，例如 Atkins 等人 [7],分析不同新闻网站在其主网页上报道的内容。在他们开展研究时，由于 CNN.com 主页面的网页存档快照无法在类似 Wayback 的回放系统中重放，Atkins 等人不得不将 CNN.com 排除在分析之外 [12]。尽管如今 CNN.com 主页面的网页存档快照已经可以重放，但我们发现，在某些情况下，该网页的回放仍存在其他问题。其中许多问题并非 CNN.com 所独有，在使用客户端网页构建技术的其他网页上也可能出现。需要指出的是，这一具体示例以美国为中心，因为美国以外的客户端和网络档案通常会被重定向到 CNN.com 的国际版本（https://edition.cnn.com）。CNN 主页面的该版本也存在与我们在此探讨的问题类似的问题，但我们重点研究 https://www.cnn.com 版本，因为可供我们研究的网络存档副本要多得多。然而，我们发现的问题并非美国特有，任何采用类似内容加载技术的网页都可能受到影响。



(2015年0917000017) (2015年1014000014) (2015年1216001003) (20160420010831) (20160729003156)

(2015年0917110018) (20151014120016) (2015年1216122308) (2016年4月20日) (2016年7月29日)

(a) 2015年9月17日 (b) 2015年10月14日 (c) 2015年12月16日 (d) 2016年4月20日 (e) 2016年7月29日

图1. 2015年9月至2016年7月期间存档的 CNN.com 网页。上排：Wayback Machine 中重放的存档 CNN.com 网页，尽管捕获时间相隔数月，但其顶层内容仍然相同。下排：来自相同日期的网页存档快照，但 Hero 故事内容正确。然而，第二层内容在时间上仍然不正确。

图 1 展示了互联网档案馆的 Wayback Machine 中，2015年9月至2016年7月期间 CNN.com 网页存档快照重放后的部分截图。[2] 这些网页存档快照中的顶层内容，包括“头条”（最醒目）新闻和“热门新闻”面板，尽管主页的抓取时间跨度接近10个月，却完全相同。遗憾的是，由于 CNN.com 并未在主页上显著显示日期，因此这些时间违规很难被发现。例如，其中一栏（图1e）包含了当天正确的内容，但如果不进行进一步调查（或不了解民主党全国代表大会的举办时间），就很难判断究竟是哪一天。并非所有2015年9月17日至2016年7月29日的网页存档快照都显示了相同的“恐惧并非力量”标题，但在该时间范围内抽取的600个网页存档快照中，我们发现有45个快照显示了这一标题。

[2]为了清晰起见，图注中的 Wayback Machine 网络存档快照，我们会加入指向网页存档的链接，但只显示日期时间，该日期时间可通用 URL 格式为 https://web.archive.org/web/**日期时间**/https://www.cnn.com/。

这种行为是由 CNN.com 通过客户端调用 API 获取额外资源来构建页面内容所导致的，这些资源返回的可能是 HTML 文件，也可能是包含 HTML 的 JSON 文件。图 1 中所示示例更加复杂，因为在 2015—2016年期间，CNN.com 的内容交付方式并不一致，具体原因下文将予以说明。这些额外资源并非都在与基础 CNN.com HTML页面相近的时间被归档。由于 Wayback式网页档案会使用距离最近的可用日期时间来补全缺失的资源，因此在重放时可能造成时间违规。这类似于此前在某些归档 Twitter 页面中发现的情况 [20], : Twitter 账户页面中各个面板的内容由 JSON 文件填充，而这些文件可能是在不同时间归档的。

本文介绍了 CNN.com 主页的构建方式，并找出了实际包含主要内容的 JSON 文件。我们还为这些变化提供了时间线，帮助研究 CNN.com 网络档案的人了解某些时间段可能出现的问题。对 CNN.com 进行存档时，应确保这些额外资源也能在接近基础页面日期时间的时间点存档，最有可能的做法是使用基于浏览器的爬虫。我们还调查了互联网档案馆和 Archive-It 中 CNN.com 主页的时间违规情况。我们发现，互联网档案馆中 2015年4月24日 至 2016年7月21日 期间的近 15,000 个网络存档快照，在 CNN.com 主页的第二层内容上存在超过 2 天的时间违规。这一时间违规在 2015年10月8日 达到峰值，时间差为 90 天。最后，我们就如何缓解这些时间违规提出了建议。

2 背景与相关工作

网络档案以与搜索引擎大致相同的方式爬取网络，但网页档案不会用爬取资源的最新版本替换先前版本，而是保留所有版本，并根据其存档时间对其进行索引，以供访问（称为*Memento-Datetime* [33]）。传统的网络爬取和存档方式（例如 Heritrix [32] 和 wget [17], 所采用的方式）会下载一个 HTML 页面，提取页面中的 URL（包括链接和嵌入式资源），将这些 URL 添加到爬取队列中，然后爬取队列中的下一个 URL。队列可以按广度优先 或深度优先进行索引；为了提高网站级一致性，大多数爬虫都配置为采用广度优先爬取方式 [10, 11, 16, 37]。例如，先爬取从顶层页面（第 0 层）链接到的所有页面，然后爬取从这些页面（第 1 层）链接到的所有页面，依此类推。

这种广度优先策略可能导致嵌入式资源（例如 JavaScript、CSS、图像和内嵌框架）的抓取时间与嵌入它们的页面大相径庭。这些嵌入式资源（例如徽标、样式表）通常会被网站上的多个页面共享，因此，在尝试缩小整个网站 HTML 页面抓取的时间跨度时，它们的抓取顺序就没那么重要。

近年来，人们开始采用基于浏览器的爬取 [13, 24, 38] 进行网页存档，以执行所有 JavaScript 并确保动态加载的资源都能被保存。虽然这种方法更为完整，因为渲染页面所需的所有资源都会被存档，但它的速度要慢得多 [15]，例如，比互联网档案馆的标准爬虫 Heritrix 慢。archive.today [34] 和 Perma.cc [40],等服务采用基于浏览器的爬取，一次存档一个页面，且每次只发送一个请求；它们不会爬取多个页面，更不用说整个网站了。自 2020 年起，互联网档案馆的订阅服务 Archive-It 向订阅者提供 Heritrix 和基于浏览器的爬虫 Brozzler [13]。互联网档案馆的 Wayback Machine 中的“立即保存页面”功能会对所请求的单个页面使用基于浏览器的爬取 [22],，即使大多数其他爬取仍由速度快得多的 Heritrix 执行。由于许多此类服务仅针对单个页面，因此它们具有内容完整的优势，而且嵌入式资源（例如图像、样式表）与嵌入它们的 HTML 页面之间的时间跨度也较短。

Perma.cc 和 Archive-It 等订阅制网络档案服务允许订阅者更精细地控制其收藏中的网页。例如，如果两位不同的订阅者在同一时间爬取同一页面，就会生成两份不同的副本，而不是共用一份副本。这样，订阅者便可以将某些页面标记为“私有”、删除页面，并总体上确保自己的副本与网络档案中其他位置不同收藏中的副本相互隔离。虽然 Archive-It 爬取的页面也会成为更大的 Wayback Machine 的一部分 [6]（即 URL 不同，但 Memento-Datetime 相同），但在 Archive-It 内部，不同收藏之间的资源并不共享。由于订阅制网络档案的用户受数据额度限制，他们的收藏可能较为稀疏（即副本很少），即便是 Twitter.com 或 CNN.com 这类广受欢迎的页面也不例外。由于收藏之间不共享资源，在 Archive-It 中重放页面时，只能使用该特定收藏中的资源。

集合间缺乏共享，再加上 Wayback Machine 及其同类系统（例如 Open Wayback [23] 和 PyWB [28]）处理时间偏好的标准方式，可能会导致很大的时间跨度[3]。重放的存档网页中的链接和嵌入式资源会被改写为指定与其所在页面相同的 Memento-Datetime。实际上，这些资源很少是在同一 Memento-Datetime 存档的，因此 Wayback Machine 会通过 HTTP 重定向，将请求导向时间上最接近的存档版本，无论该版本来自未来还是过去[1, 2]。这种跨度可能导致时间违规：所在页面与嵌入式资源的组合从未在实时网络上同时存在过 [4]。单凭时间跨度很大，并不一定意味着发生了时间违规（例如，一张十年来未曾更改的图像徽标或样式表，即使时间跨度为 9 年仍然有效）；但实际上，时间跨度越大，发生时间违规的可能性就越高。

近期的一种网页设计趋势，对网页存档以及成功重放归档网页产生了重大影响。如今，通过解析 URL 返回的往往不再是所有信息都已在服务器端组装完毕的完整 HTML页面，而是服务器返回一个由 HTML 和 JavaScript 构成的骨架页面，随后该页面会向其他网络服务器发起一系列 API 调用。这些 API 响应会返回数据（通常为 JSON），JavaScript 随后利用这些数据构建页面的其余部分，这一过程称为客户端渲染（CSR）[31]。使用 Heritrix 这类非基于浏览器的爬虫时，这些 API 调用不会发出，其响应也不会被存档。因此，重放时归档网页会出现内容不完整和错误等问题，例如 Twitter [20]和 Instagram [14, 26]中所述的问题。这种页面客户端渲染方式非常适合社交媒体，因为它们的用户界面具有高度交互性且不断更新。不过，我们发现新闻网站也采用了客户端渲染，例如 CNN.com，而这对重放其归档网页有着重大影响。

3 CNN.com 的客户端构建

互联网档案馆的 Wayback Machine 自 2010 年以来收录了近 200,000 个针对 www.cnn.com 且 HTTP 状态码为 200 的网页存档。利用这一宝贵资源，我们分析了 CNN.com 数年来的网络存档快照，发现主页面的客户端构建始于 2015年4月24日。[3] 我们发现，有多个 zone-manager.html文件通过客户端渲染（CSR）被请求，其中包含用于构建 CNN 主页面各个部分（即各个区域）的 HTML 内容。[4] 这些文件的 URL 格式为

http://www .cnn.com/data/ocs/section/**{baseUri,uri}:id**/views/zones/common/zone-manager.html，带有 h

[3]采用此构建方式之前的最后一个网页存档是 20150424134720，而第一个采用这种新构建方式的网页存档是 20150424150304。 [4]在互联网档案馆中找到的最早的 zone-manager 文件是 https://web.archive.org/web/20150427192038/http://www.cnn.com/data/ocs/section/index.html:homepage4-zone-5/views/zones/common/zone-manager.html。

baseUri、uri和id 定义在 CNN.Zones 变量中（参见图 2）。[5] 加载哪些特定文件，取决于执行 JavaScript 代码时浏览器窗口的最小宽度。我们的分析基于 CNN.com 的桌面版，因此只关注宽度至少为 800 像素的情况。

```
CNN . 区域 = {
  [...]
" 区域 ": {
" baseUri ": " index . html ",
"
mi
ho
inc
inj
{
" h
ho
ho
ho
```

图 2. CNN.com 于 2020年6月18日（20200618234848）的基础 HTML 页面中包含的 JavaScript 代码，用于定义网页区域，包括 baseUri、uri和 id变量

图 3 展示了 2020年6月18日的一份网页存档，并对其进行了标注，以说明这种构造方式。每个版块左侧标出了区域名称，右侧则显示 data-zone-label （见 HTML）。对于没有 data-zone-label 的区域，我们提供了该版块标题的示例。这份网页存档加载了 11 个区域管理器文件，包括 homepage-injection-zone-1 和 homepage-injection-zone-2，它们包含细分隔横幅的 HTML 代码。主要新闻内容位于 homepage1-zone-1、homepage2-zone-1和 homepage3-zone-1 这些区域中，其他区域则加载更多促销类内容。由于我们关注的是影响显著的时间差异，因此本文余下部分只讨论这三个顶部区域。

为开展这项分析，我们首先使用无头浏览器 selenium-wire[6]，从互联网档案馆加载 http://www.cnn.com/ 的网页存档快照，并捕获发出的请求及其状态码。通过这种方法，我们请求了 2013 至 2020 年间的共 1131 个网页存档快照，以研究 CNN.com 主页的请求模式，

[5] 例如，区域 homepage-injection-zone-1 定义于 http://www.cnn.com/data/ocs/section/_homepage-zone-injection/index.html:homepage-injection-zone-1/views/zones/common/zone-manager.html，而区域homepage3-zone-1 定义于 http://www.cnn.com/data/ocs/section/index.html:homepage3-zone-1/views/zones/common/zone-manager.html。 [6] https://pypi.org/project/selenium-wire/

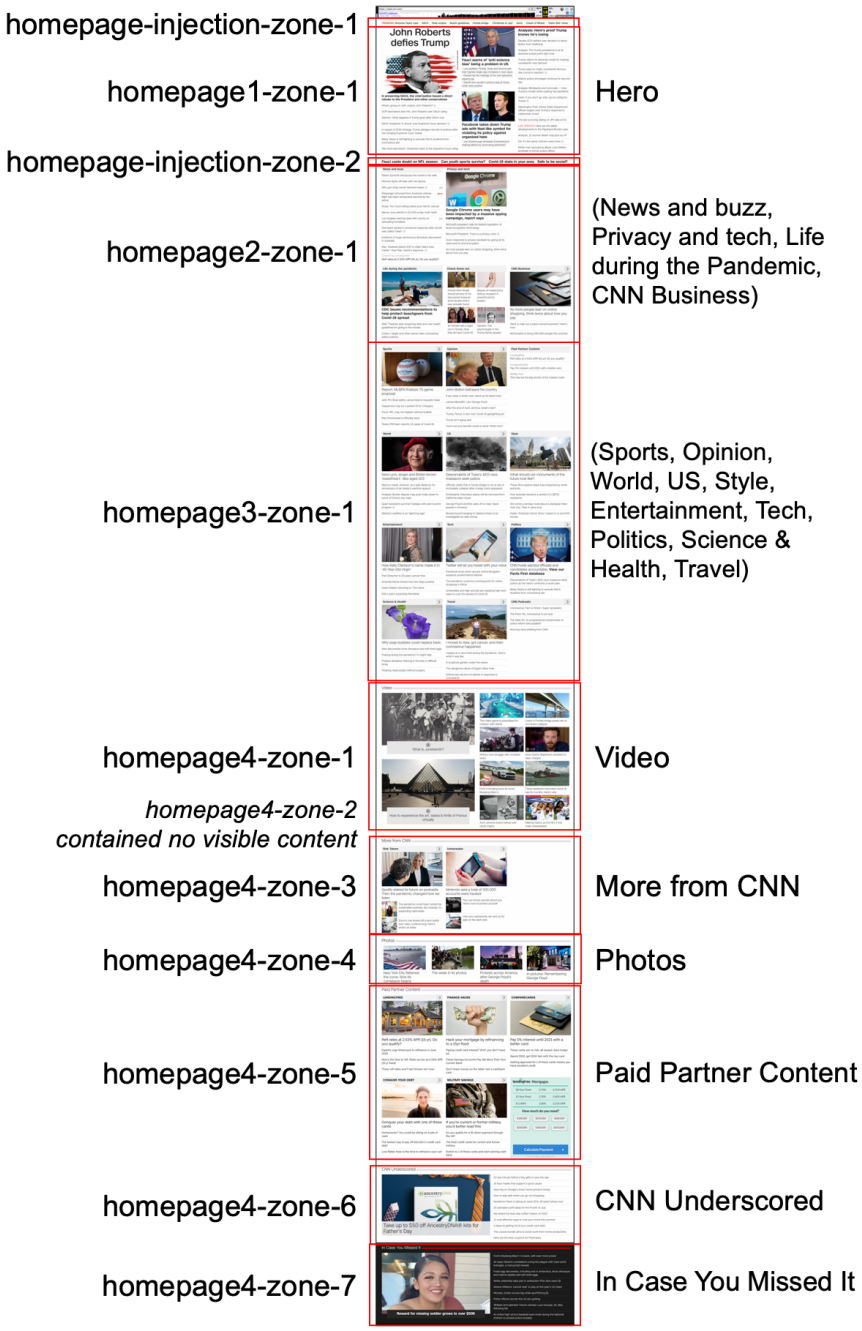


图3。网络档案 2020年6月18日的 CNN 主页面（20200618234848），标注了区域名称和标签。请注意，homepage4-区域-2 包含 无可见内容。

表1. CNN.com 区域管理器时间线

Date	区域内容交付
2月18日, 2015	基础 HTML 中的内容划分为多个区域
2015年4月24日	zone-manager.html文件中通过客户端渲染请求的 Hero 区域以外的内容
2015年9月17日	Hero 区域（homepage1-zone-1）有时 通过客户端渲染请求
2016年10月18日	区域管理器格式更改为 .izl.json
2016年11月1日	通过客户端渲染请求的所有区域
2017年1月31日	zone-manager 扩展名已更改为 .izl

时间。如前所述，我们能够确定区域管理器文件于2015年4月24日引入。此后，除 homepage1-zone-1 外，所有区域均始终通过客户端渲染请求。

然而，直到2016年11月1日，homepage1-zone-1 才始终通过客户端渲染请求。（这也是两个首页注入区域首次出现。）我们原以为，未请求 homepage1-zone-1 的网页存档快照会缺少 Hero 内容，但检查 Wayback Machine 中的重放后，我们发现 Hero 内容在时间上是正确的。为进一步调查，我们使用 Wayback Machine 的 id_选项，下载了2010年至2022年间1316个 http://www.cnn.com/ 的网页存档快照的原始 HTML。我们发现，当 homepage1-zone-1 未通过客户端渲染加载时，HTML 中会出现一个带有 id="homepage1-zone-1" 属性的 <section> 标签，其中包含 Hero 内容。图4展示了其中一段代码。第1行显示了 id="homepage1-zone-1" 属性和 data-zone-label="Hero"。标题“黄石公园惊现重磅发现”出现在第4行，使用 <h2> 和 标签进行格式化。当 homepage1-zone-1 通过客户端渲染请求时，HTML 中不会出现 id="homepage1-zone-1" 属性。通过检查2015年4月24日至2016年11月1日期间1069个网页存档快照的原始 HTML，并将其与同一时期无头浏览器记录的请求进行比较，证实了这一点。

```
1 <section class =" zn zn - homepage1 -zone -1 zn - left - fluid - right - stack zn --idx -0 t - light zn -
lo
pt
" z
hi-
te
```

图 4. 包含 `id="homepage1-zone-1"` 属性和 Hero 内容，来自 2015年4月24日（20150424150304）。

区域管理器文件经历了数次格式和文件扩展名变更。2016年10月，`zone-manager.html` 文件被改为 JSON 文件，其中 HTML代码作为值，并命名为`zone-manager.izl.json`。随后在2017年1月，JSON格式的区域管理器文件被重命名为`zone-manager.izl`。图中展示了一个`homepage1-zone-1 zone-manager.izl`示例，见图5，其中显示了图3 在线版中的“约翰·罗伯茨挑战特朗普”标题。表1 总结了文件名和格式更改的时间线。

1 {" izlData ": {" cards ": [ ]} , 2 " html ": 3 "<section class =\" zn zn - homepage1 - zone -1 zn - left -
fl
cc
ho
>4
ar

图5. zone-manager.izl 在2020年6月18日针对 homepage1-zone-1 的片段，展示了一个采用 JSON 格式的文件中的 HTML 代码，其中包含标题“约翰·罗伯茨挑战特朗普”，如图 3所示。

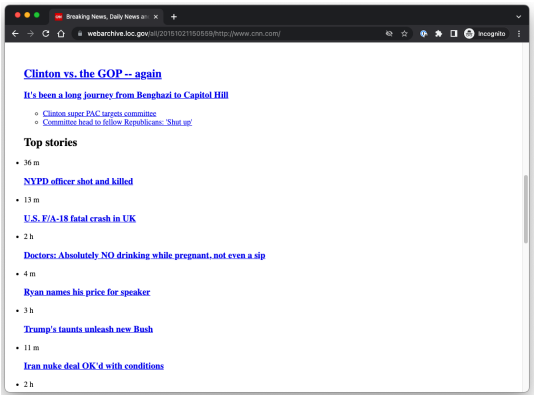
表 2. 2015年4月24日之后公共网页档案中 http://www.cnn.com/ 的网页存档数量（通过 MemGator）

网络档案	网络存档快照
web.archive.org	483,611
wayback.archive-it.org	35,941
archive.today	3276
webarchive.loc.gov	1666
arquivo.pt	1439
swap.stanford.edu	495
perma.cc	139
www.webarchive.org.uk	11
web.archive.org.au	8

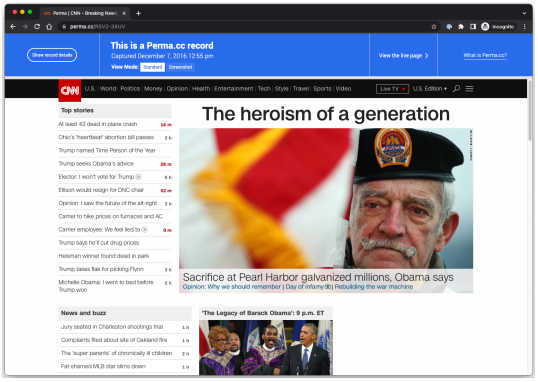
4 对网页档案的影响

CNN 的客户端渲染模型与标准网页存档模型相冲突。由于网页内容是由 HTML 或 JSON 数据块构建而成，如果这些区域管理器文件未与基础 HTML 同时存档，就会产生时间违规。回放时，CNN.com 将使用在不同时间存档的文件中的数据填充，而其日期时间由网络档案横幅指示。此外，CNN.com 不再在主页上显示日期，也不会在其区域中要显示的任何内容里包含日期，因此这些时间违规将难以被发现。由于这些文件中的内容具有时效性，采集间隔越长，违规情况就越严重，尤其是使用未来内容填充时。

我们于2022年11月使用 MemGator [5] 从公共网页档案中收集 http://www.cnn.com/ 的 TimeMap（网络存档快照列表），其中包含2015年4月24日之后的 Memento-Datetimes。表格2 显示，在这一时间段内，互联网档案馆（web.archive.org）和 Archive-It（wayback.archive-it.org）拥有的网络存档快照数量远远最多。本研究的对象 CNN.com 使用地理定位来决定要提供哪个网页，因此美国境外的网页档案（包括 archive.today、arquivo.pt、webarchive.org.uk 和 web.archive.org.au）很可能实际上并未捕获 http://www.cnn.com/，而是捕获了 http://edition.cnn.com/。这取决于网络服务器向网页档案提供的内容，而不是网页档案本身或其捕获技术。



(a) 使用传统爬虫归档的 CNN.com



(b) 使用基于浏览器的爬虫归档的 CNN.com

图6。CNN.com 从（a）美国国会图书馆网页档案（webarchive.loc.gov）中重放，存档于2015年10月21日，采用传统爬虫；以及（b）Perma.cc，存档于2016年12月7日，采用基于浏览器的爬虫。

我们还使用 MemGator 在互联网档案馆和 Archive-It 之外的其他档案中查找区域管理器文件的网络存档快照，但只找到少量结果。美国国会图书馆的网页档案（webarchive.loc.gov）有 11 个 zone-manager.izl 文件，全部来自2019年6月22日。唯一其他的区域管理器网页存档是 Arquivo.pt 中一个来自2016年3月15日的 homepage4-zone-1 时光副本。由于这些区域管理器文件包含 CNN.com 主页的实际内容，如果没有这些文件，2015年4月24日之后其他档案中的 CNN.com 网络存档快照要么会受损、显示很少或不显示内容，要么会存在严重的时间违规。例如，图6a 展示了2015年10月12日美国国会图书馆网页档案中 CNN.com 一个网络存档快照的重放。图像向下滚动，以显示 HTML 中包含的 Hero 和头条新闻区，但 CSS 及其他区域内容并未加载。使用基于浏览器的爬取的网页档案（例如 archive.today、Perma.cc（图6b中的2016年12月7日示例）以及互联网档案馆的立即保存页面功能）不会遇到非基于浏览器的爬虫所面临的同类问题，因为浏览器在渲染该页面的 DOM 时，会请求并存档所有区域文件。如前所述，基于浏览器的爬取的缺点是速度远慢于 Heritrix 等传统爬虫，因此不适合对整个网站进行“广泛”爬取。

由于 CNN.com 的网络存档快照数量最多，因此我们在后续分析中重点关注从互联网档案馆以及 Archive-It 中的特定馆藏重放的 CNN.com 网页存档快照。

4.1 互联网档案馆

根据互联网档案馆 CDX API 的数据，在2015年4月24日至2022年9月23日期间，互联网档案馆中针对 http://www.cnn.com/ 的 HTTP 200 状态码网页存档共有172,776个。图7 展示了互联网档案馆（IA）中可用的 homepage1-zone-1、homepage2-zone-1和homepage3-zone-1区域管理器文件的网络存档快照分布情况。Memento-Datetime 沿 x轴排列，各区域文件沿 y轴分开显示。每个圆点代表当天至少存在一个网页存档。相同颜色圆点的堆叠不具有任何实际含义，仅用于避免数据点重叠。由于 www.cnn.com 每天至少有一个网页存档，因此图中未显示 www.cnn.com 的时光副本。图中还标注了不同时间段所使用的文件扩展名。

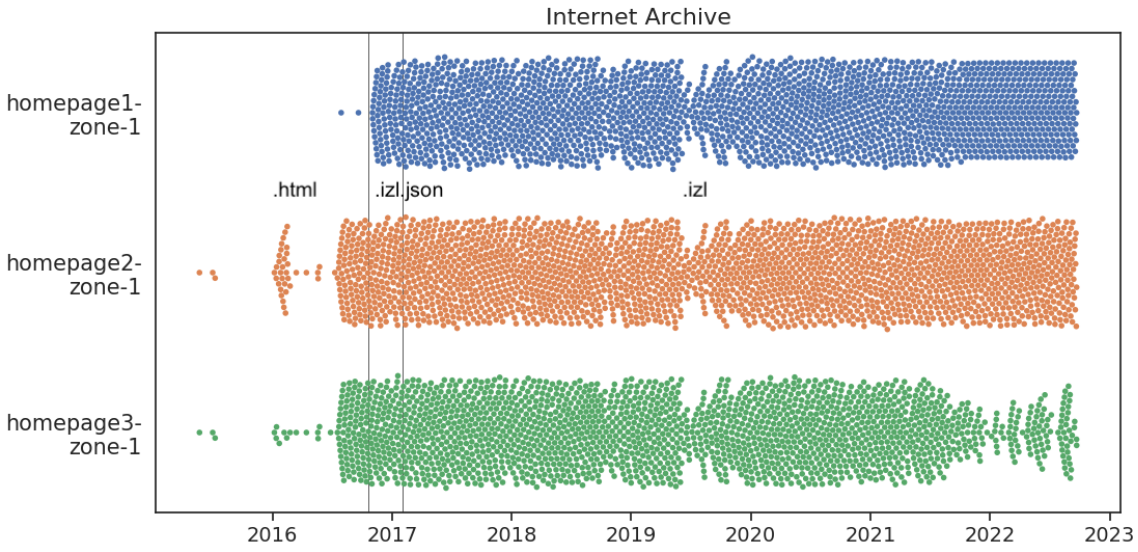
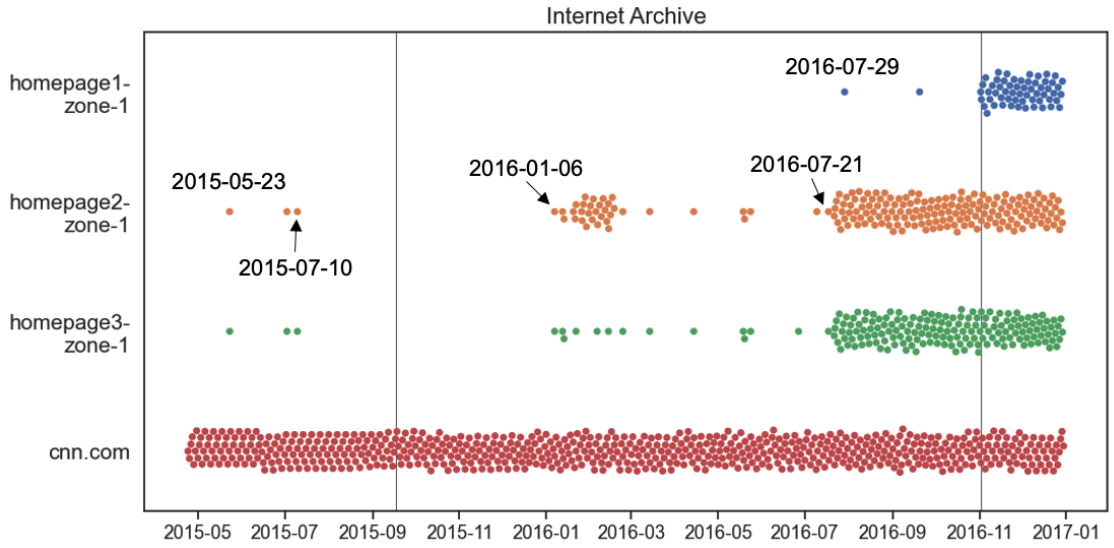


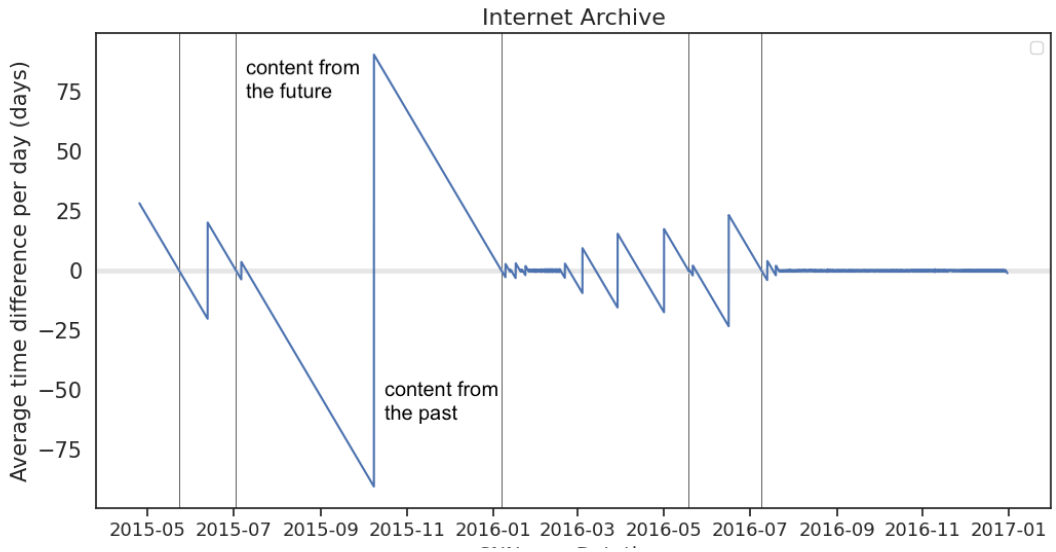
图 7。互联网档案馆中 homepage1-zone-1、homepage2-zone-1、homepage3-zone-1 区域管理器文件的网络存档快照。每个圆点代表当天至少有一个网页存档。

从图 7可以看出，2017年之后，区域管理器网页存档的存在情况相当稳定，因此我们将分析重点放在2015年4月24日至2016年12月31日这一时间段内。图 8 展示了这一时间段的两种不同视图。图8a展示了互联网档案馆中 www.cnn.com 及其排名前三的区域的网页存档快照，并标注了几个日期。第一条垂直线表示我们首次观察到 homepage1-zone-1 被请求的日期（2015年9月17日），第二条垂直线则表示所有区域开始通过客户端渲染定期加载的时间（2016年11月1日）。如第 3节所述，在这一时间段内，homepage1-zone-1 仅偶尔通过客户端渲染请求。暂时，我们将分析重点放在两个主要区域上，这两个区域在2015年4月之后始终通过客户端渲染请求，稍后再回到 homepage1-zone-1 。对于另外两个区域，我们知道，在这一时间段内，这些区域始终通过客户端渲染加载，因此，图 8a 中这些行上的任何空白区域都表示存在时间违规。这意味着，当重放 CNN.com 的网页存档时，将没有该区域时间相近的网页存档，因此内容将来自时间上最近的区域网页存档，而该存档可能早在数天（甚至数月）之前，也可能晚在数天（甚至数月）之后。为简化讨论，在接下来的分析中，我们将重点关注homepage2-zone-1，因为其结果与 homepage3-zone-1的结果类似。

在图8b中，我们展示了请求的 CNN.com 日期时间与已加载的 homepage2-zone-1 区域管理器文件的日期时间之间，按天计算的平均差值（以天为单位）。差值为正表示嵌入内容（来自 homepage2-zone-1 区域管理器文件）来自未来；差值为负则表示嵌入内容来自过去。黑色竖线表示部分稀疏的区域管理器网页存档的日期时间（图8a中的橙色圆点）。如果嵌入资源在与主页相同的日期时间不存在，Wayback Machine 会尝试查找时间最接近的可用网页存档。在图中，差值突然增大，说明下一个网页存档比前一个更接近，此时嵌入内容将来自未来，而非过去。我们看到，在2015年10月8日，



(a) CNN.com 和排名前三的区域在互联网档案馆中的网页存档。

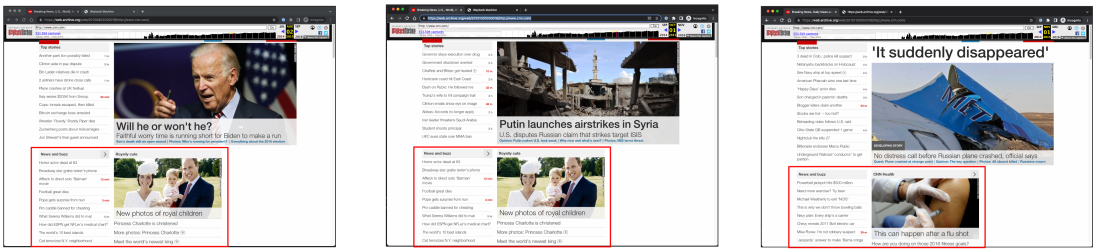


(b) CNN.com 日期时间与互联网档案馆中 homepage2-zone-1 区域管理器文件最近的网页存档之间的平均天数差。

图 8. 2015年4月24日至2016年12月31日期间互联网档案馆网页存档的分析。(a) 每个圆点代表有网页存档的一天；(b) www.cnn.com 与最近的 homepage2-zone-1 网页存档之间的差异。

差值飙升至90天左右。那一天前后，重放的 homepage2-zone-1 内容会先来自2015年7月10日，随后来自2016年1月6日，这分别对应时间最近的区域管理器网页存档。

图 9 展示了一些涉及 homepage2-zone-1 内容的时间违规示例。图 9a展示了日期时间为2015年8月2日的 CNN.com 网页存档快照，图 9b 则展示了日期时间为2015年10月1日的快照。Hero 内容包含在基础 HTML 页面中，因此不存在时间违规。然而，其余内容（区域 homepage2-zone-1至 homepage4-zone-5）来自2015年7月。同样的 homepage2-zone-1 内容也可以在图 1a中 Hero 内容下方看到。图 9c 展示了一个未来内容被填充的示例，其中 homepage2-zone-1 内容来自2016年1月6日。此示例对应于图 8b中的尖峰，该尖峰出现在2015年10月8日。在此之前，网页存档快照重放时使用2015年7月10日的 homepage2-zone-1内容；此后，网页存档快照重放时使用2016年1月6日的内容，直到2016年1月7日出现新的 homepage2-zone-1 内容。



(a) 2015年8月2日（20150802000019） (b) 2015年10月1日（20151001000018） (c) 2015年11月1日（20151101000016）

图9. 2015年由于区域管理器文件存档稀疏，Wayback Machine 重放中出现 homepage2-zone-1 时间违规的示例。(a) 和 (b) 中的内容来自过去，即2015年7月10日（20150710001845）；(c) 中的内容来自未来，即2016年1月6日（20160106233405）。

我们计算了区域 homepage2-zone-1 中具有不同时间违规级别的网页存档数量（根据图8a，来自 homepage3-zone-1 的内容也会得到类似结果）。在表3中，我们考察了采用时间范围阈值对三个时间段内加载 homepage2-zone-1 内容的影响：2015年4月24日至5月23日（无网页存档）、2015年5月23日至2016年7月21日（网页存档稀疏）以及2016年7月21日至2022年9月23日（网页存档密集）。换言之，我们考察了有多少 CNN.com 时光副本（以及涉及多少天）会在第二级（及第三级）区域中没有内容，前提是不允许加载时间差异超过这些阈值的区域管理器文件。我们可以看到，由于2016年7月21日之后区域管理器网页存档数量增加，时间违规的影响有所减轻。然而，在此前的450天里，影响会相当显著。

www.cnn.com 的重放中出现的时间违规，在发生于 homepage1-zone-1Hero 区域时更加明显，如图 1所示。由于 CNN.com 主页不包含日期时间，因此这些违规很难被检测到，因为 HTML 没有提供可与网络档案中的日期时间值进行对比的日期时间值。我们对 homepage1-zone-1 内容进行了一些分析，但由于这些文件并不总是会被请求（如第 3 节所述），我们不得不下载 www.cnn.com 的 HTML，以确定是否存在 `id="homepage1-zone-1"` 属性，然后才能确定该区域何时通过客户端渲染加载，并判断其是否可能发生时间违规。

我们注意到，homepage1-zone-1 区域管理器最早的网页存档快照日期为2016年7月29日（20160729003156）。在此日期之前，任何请求通过客户端渲染（CSR）加载 homepage1-zone-1 的 CNN.com 主页网页存档快照，都会导致头条故事出现时间违规。图 10 展示了 homepage1-zone-1 网络存档快照的时间线，以及通过客户端渲染（CSR）请求该区域的部分 CNN.com 主页样本。与其他同类型图表一样，一个圆点表示至少存在一个网页存档快照

表3. 从互联网档案馆加载 homepage2-zone-1 内容时，受各种时间阈值影响的网络存档快照

日期范围	总计 网络存档快照	Days	阈值	受影响 网络存档快照	Days
2015年4月24日–2015年5月23日	788	29	1 hr	786	29
			2小时	785	29
			6小时	780	28
（无）			24小时	754	27
			48小时	724	26
2015年5月23日–2016年7月21日	17,017	425	1 hr	16,863	425
			2小时	16,738	425
			6小时	16,244	410
（稀疏）			24小时	15,142	370
			48小时	14,178	342
2016年7月21日–2022年9月23日	154,917	2256	1 hr	95,009	1826
			2小时	67,992	1425
			6小时	13,893	190
（密集）			24小时	1812	37
			48小时	526	8

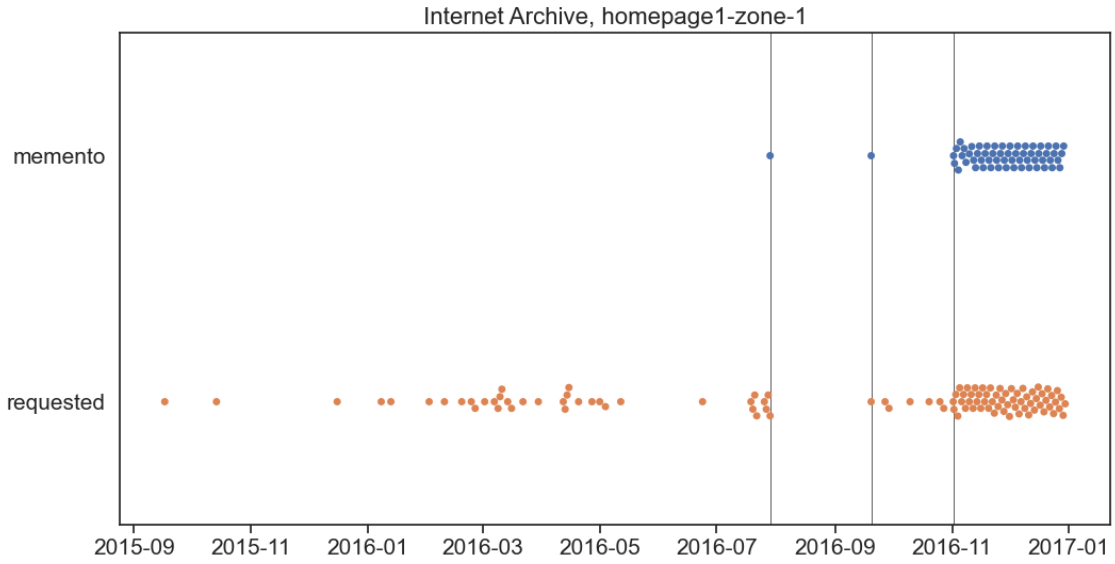


图 10. 2015年9月至2016年12月期间，互联网档案馆中 homepage1-zone-1 区域管理器文件的网页存档快照和请求时间线。

当天存在（或发出的请求）。垂直线与该区域可用的前三个网页存档相对应。第一条垂直线之前的所有“请求”圆点均代表存在时间违规的网页存档，并且有多个这样的实例。

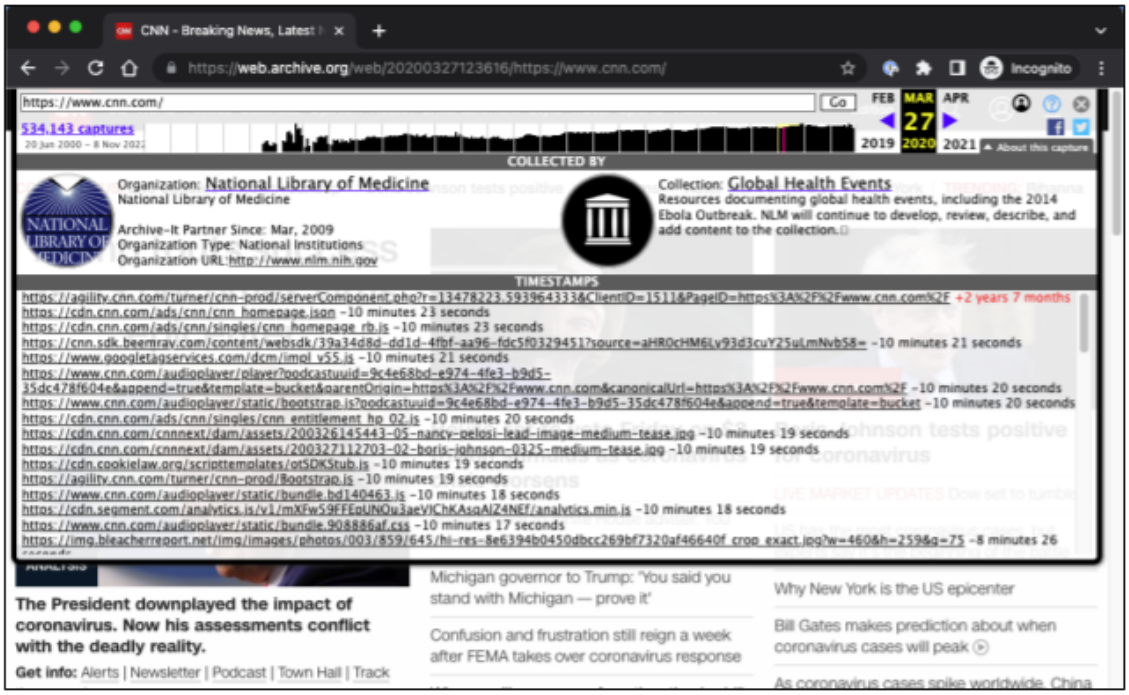


图 11. CNN.com 的“关于此捕获”面板（20200327123616），显示了 200 多个嵌入式资源中的 15 个 .

在第二行和第三行之间。需要注意的是，“requested”区域并未显示 homepage1-zone-1 通过客户端渲染发起请求的所有日期，而只显示我们采样的日期。请注意，从 2016 年 11 月开始，一旦 homepage1-zone-1 定期通过客户端渲染发起请求，homepage1-zone-1 网页存档的数量便会增加。

2017 年，互联网档案馆在存档横幅中添加了“关于此捕获”服务 [21]。点击横幅中的链接会列出用于重放该页面的所有资源（例如图像、样式表），并显示它们相应的日期时间，以及相对于根 HTML 页面而言，它们是在未来还是过去被抓取的（图11）。然而，这种显示方式可能会给用户带来不便：正如第 2 节所述，并非所有时间跨度都属于时间违规[4], ，而且对于包含数百个嵌入式资源的页面，图形用户界面也难以良好呈现。尽管存在这些问题，JSON 文件目前仍未纳入“关于此捕获”显示，因此 CNN.com 上的时间违规会悄无声息地发生。然而，即使将 JSON 文件纳入其中，尤其是在 CNN.com 的情况下，普通用户也不会知道名为 `zone-manager.izl`的文件有何重要性。

4.2 Archive-It

Archive-It 中的公共资源可在互联网档案馆的 Wayback Machine 中访问，但各个 Archive-It 馆藏无法访问其馆藏范围之外的资源。由于 Archive-It 馆藏网页存档仅限于该特定馆藏中的资源，因此基础页面日期时间与内容 JSON 日期时间之间的差距可能会更大。

美国国家医学图书馆的全球卫生事件合集[7]（合集 4887）位于 Archive-It，是一个规模庞大（包含 22,000 个种子 URL）且持续时间较长（自 2014 年 10 月起）的合集。由于其专门收录与 COVID-19 大流行相关的网页，因此具有特别重要的价值。图 12展示了该合集的网络存档快照及差异图表。该合集在 2015 年 5 月至 2022 年 12 月期间包含 600 个状态码为 200 OK 的网页存档快照。两张图中的垂直线标记了首次通过客户端渲染（CSR）请求 homepage1-zone-1的时间（2016 年），以及间隔一段时间后首次捕获区域管理器网页存档的日期。图 12a 展示了该合集中的可用网页存档快照分布图，而图 12展示了 CNN.com 主页面的日期时间与 homepage1-zone-1 和homepage2-zone-1 区域管理器文件之间相应的时间差。由于区域管理器文件的网页存档快照是在同一时间捕获的，因此这些线条彼此重叠。请注意，此处的时间差以天而非小时为单位。我们注意到，尽管该合集从 2015 年开始就对 CNN.com 进行了存档，但直到 2020 年 6 月才首次捕获区域管理器文件的网页存档快照。

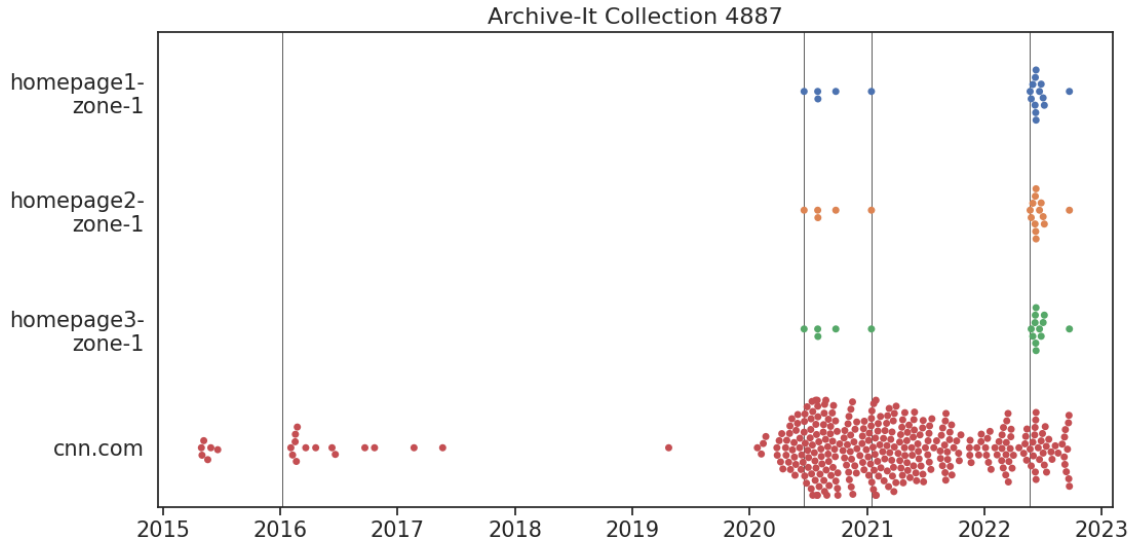
不过，其他 Archive-It 馆藏（由不同机构拥有）确实保存了大量 CNN.com 时光副本及相关资源。例如，馆藏 7678[8] 由 Wayback Machine 的主管 MARK GRAHAM 创建，其种子 URL 只有一个：`http://www.cnn.com`。该馆藏收录了 2016年7月 至 2019年9月期间 CNN.com 主页的 2619 个网络存档快照，以及近 13,000 个各种区域管理器文件的网络存档快照。图 13a 展示了 Archive-It 馆藏 7678 中 CNN.com 及排名靠前的区域管理器文件（zone-1）可用网络存档快照的分布。相应的差值线（以天为单位）见图 13b。垂直线表示该馆藏中最后一个 homepage2-zone-1 区域管理器网页存档（2019年6月4日）；不过，CNN.com 主页仍持续被捕获到 2019年9月。y轴上 0 附近的阴影区域表示 1 天的阈值。可以看到，该馆藏提供了一组高质量的 CNN.com 捕获记录，大多数差值都在 1 天范围内。如果允许其他 Archive-It 馆藏借用该馆藏的区域管理器文件，回放效果将会大大改善。

5 讨论

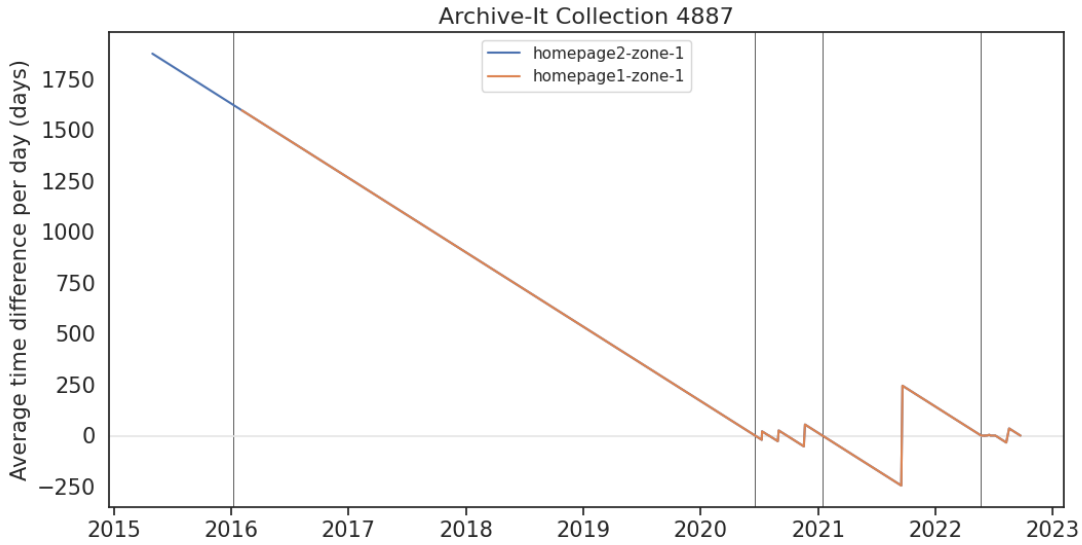
这项分析提出了几个值得进一步讨论的问题。首先，从短期来看，互联网档案馆等网页档案应如何处理这些存在时间违规的网页存档？我们已经看到，对于 CNN.com 这样可能包含数百个嵌入式资源的网页，仅仅显示各种嵌入式资源之间的时间差异并不能有效解决问题。网页档案可以考虑在请求不同日期时间的网页存档快照之前，先检查重定向资源的时间差异。这可能需要专用代码以及对受影响网页进行分析（类似于我们在本文中所做的工作），因为正如我们所提到的，对于许多嵌入式资源（例如图像）而言，时间差异是几分钟、几天还是几个月可能并不重要。对于 CNN.com，网络档案可以确保扩展名为 .html、.izl.json 或 .izl 的资源不会以较大的时间差异加载。这可以解决过去出现的时间违规问题，但无法阻止 CNN.com 在未来某个时候更改文件扩展名。此外，不包含这些资源会导致网页存档快照在重放时出现大量缺失内容。鉴于互联网档案馆拥有大量存档材料，它还有另一种处理 Hero 区域中存在时间违规的网页存档快照的方案——静默删除这些网页存档快照。对于我们在 homepage1-zone-1 中识别出的每个存在时间违规的网页存档快照，都存在另一个包含该内容的网页存档快照，其中该内容已包含在 HTML 中，因此不存在时间违规。这一情况仅适用于 2016年11月1日之前捕获的网页存档快照。对于 Archive-It，是否有可能让所有在 2016年7月—2019年6月期间存档过 CNN.com 网页的馆藏都.....

[7] https://archive-it.org/collections/4887

[8] https://archive-it.org/collections/7678



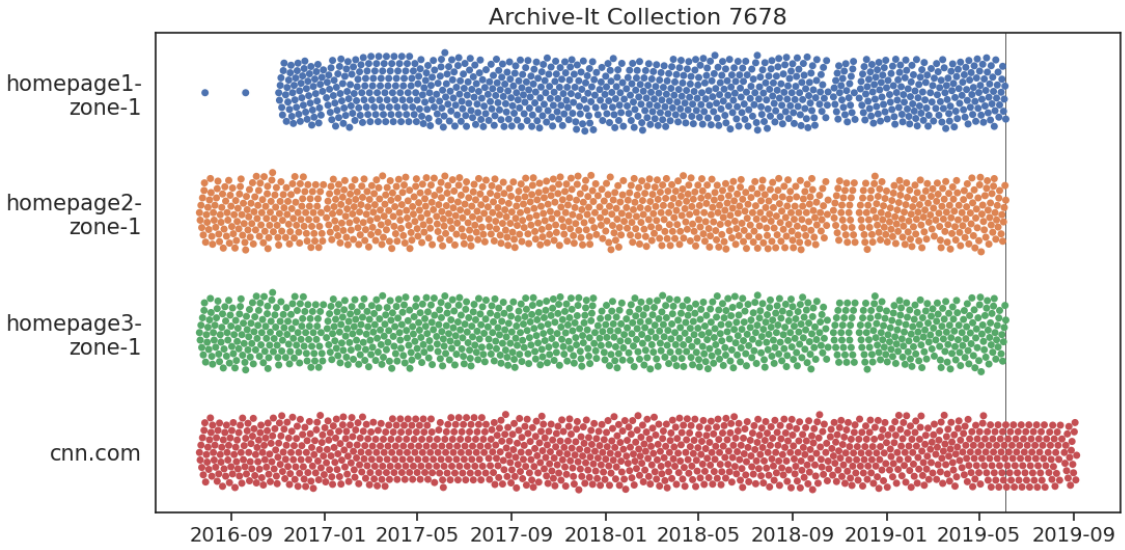
(a) 前三个区域和 CNN.com 的网络存档快照时间线。



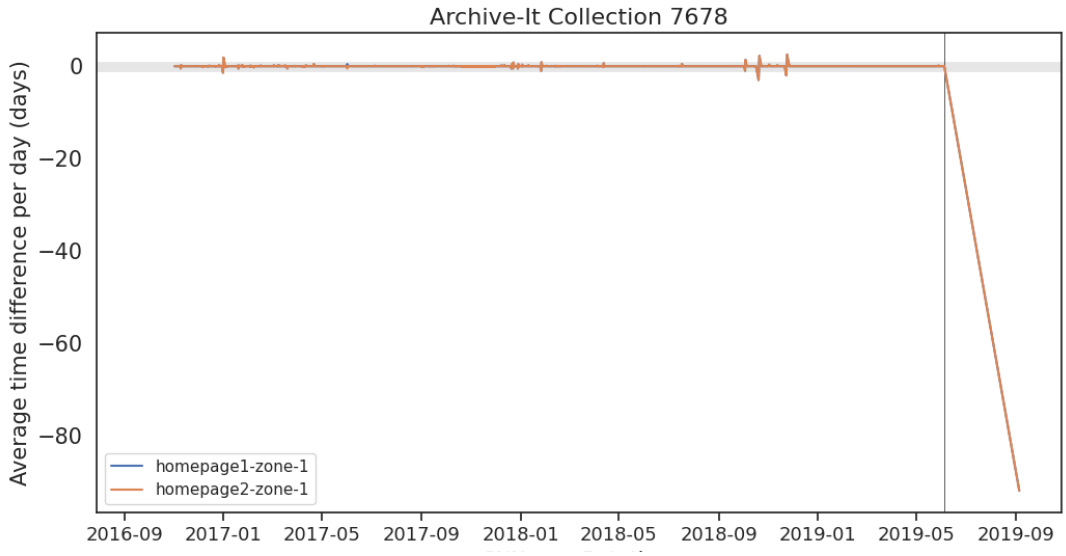
(b) CNN.com 日期时间与 homepage1-zone-1 和 homepage2- 最近网页存档之间的平均天数差 zone-1 区域管理器文件

图 12. Archive-It 馆藏 4887 中 CNN.com 网络存档快照的分析

借用馆藏 7678 中的区域管理器资源（或从互联网档案馆的一般馆藏中借用）？ 这很可能会改善许多 Archive-It 馆藏中 CNN.com 时光副本的重放效果。



(a) 排名前三的区域和 CNN.com 的网络存档快照时间线。



(b) CNN.com 日期时间与 homepage1-zone-1 和 homepage2- zone-1 区域管理器文件中最近的网页存档之间的平均天数差

图 13. Archive-It 收藏集 7678 中 CNN.com 网络存档快照的分析

此外，网页档案应采取哪些措施来避免今后出现这类问题？这个问题并不局限于社交媒体网站和 CNN.com。我们观察到（但尚未全面测量），《亚特兰大宪政报》主页也存在类似情况。

《亚特兰大宪政报》的 AJC.com。鉴于 CNN.com 于2015年改用客户端渲染，这并非一种新兴的网页设计技术。肯定还有许多网页受到类似问题的影响，只是它们不够引人注目，因而未被发现。应使用基于浏览器的爬虫对这些网页进行存档，类似于互联网档案馆如何对某些网页，以及通过“立即保存页面”功能请求抓取的网页，采用基于浏览器的爬取。

除了 CNN.com 特有的启发式方法——通过查找 HTML 中是否缺少 id="homepage1-zone-1"，来确定 Hero 内容何时通过客户端渲染加载——我们还确定了另一种启发式方法，用于检测主要仅提供客户端渲染模板的 HTML。我们发现，当 CNN.com 通过客户端渲染请求所有区域时，基础 HTML 中可能属于内容的文本量（以小写字符开头的单词）为 15 个或更少。这是在移除所有 HTML 标签、脚本和样式内容之后得出的。该启发式方法适用于 CNN.com，但未来可以开展研究，以了解这一方法或类似方法的适用范围有多广。随后，对于检测到大量采用客户端渲染构建的页面，应标记为使用基于浏览器的爬虫进行抓取。

6 结论

由于 HTML页面采用客户端渲染，其中 API 调用产生的 JSON 被用于构建 DOM，因此使用非基于浏览器的网页存档工具来存档和重放此类页面可能十分困难。Heritrix 就是这样一种工具，它针对速度和站点级抓取进行了优化；相比之下，基于浏览器的工具更加完整，但运行速度要慢得多。众所周知，Heritrix 等传统工具通常无法充分存档社交媒体网站，但我们已经证明，这同样会对 CNN.com 等新闻网站产生影响。因此，在某个日期时间存档的页面，在重放时可能会显示从违规 JSON 响应中加载的新闻标题，而这些标题相对于主 HTML页面的日期时间可能处于遥远的未来或过去。由于违规 JSON 并不像图像那样能够以相同方式直接在页面中观察到，因此检测时间违规可能十分困难。这会带来问题，因为网络档案的内容曾被用于举证，而捕获内容与重放内容之间的巨大差异应当能够被用户轻易察觉。

对于 CNN.com 的主页，我们测量了 Hero 内容以及第二级和第三级内容中的时间违规。在区域管理器文件未归档或仅零星归档的 453 天期间，第二级和第三级内容存在显著的时间违规，日期差异最高达 90 天。我们还发现 Hero 内容中存在显著时间违规的情况，但可以通过删除这些特定的网页存档来缓解，因为在那些日期还存在其他网页存档，其中的 HTML 已包含 Hero 内容。这一影响并不限于 CNN.com，也可能波及任何使用客户端渲染且经常更新的网页。要维持页面一致性，也就是确保 JSON 响应与初始 HTML 响应同时归档，就需要采用基于浏览器的爬取。遗憾的是，基于浏览器的爬虫目前尚未优化到足以大规模快速爬取整个网站的程度，因此我们可能需要一些方法来识别哪些页面需要基于浏览器的爬取，哪些页面适合使用 Heritrix 等速度更快的工具。

参考文献

[1] Scott G. Ainsworth 和 Michael L. Nelson。2013 年。通过测量穿越网络档案的随机游走中的时间漂移，评估滑动目标策略和粘性目标策略。载于*ACM/IEEE*数字图书馆联合会议论文集（*JCDL*）。39–48。[2] Scott G. Ainsworth 和 MichaelL.Nelson。2015 年。通过测量穿越网络档案的无环游走中的时间漂移，评估滑动目标策略和粘性目标策略。《国际数字图书馆期刊》16，2（2015），129–144。[3] Scott G. Ainsworth、Michael L. Nelson 和 Herbert Van de Sompel。2014 年。复合*Memento*时间一致性评估框架。技术报告 arXiv:1402.0928。

[4] Scott G. Ainsworth、Michael L. Nelson 和 Herbert Van de Sompel。2015 年。五个归档网页中只有一个曾按所呈现的样子存在。载于 ACM 超文本会议论文集。257–266。[5] Sawood Alam 和 Michael L. Nelson。2016 年。MemGator——便携式并发 Memento 聚合器。载于 ACM/IEEE-CS 数字图书馆联合会议第 16 届会议论文集（美国新泽西州纽瓦克）（JCDL '16）。243–244。https://doi.org/10.1145/2910896.2925452[6] Archive-It。2023 年。常见问题。https://archive-it.org/blog/products-and-services/archive-it-faqs/。[7] Grant Atkins、Alexander Nwala、Michele C. Weigle 和 Michael L. Nelson。2018 年。衡量十个美国新闻网站之间的新闻相似度。载于国际数字保存会议（iPres）论文集。马萨诸塞州波士顿。[8] Mohamed Aturban。2020 年。归档网络资源固定性验证框架。博士学位论文。Old Dominion University。[9] Mohamed Aturban、Michael L. Nelson 和 Michele C. Weigle。2021 年。网络档案去哪儿了？载于数字图书馆理论与实践会议（TPDL）论文集。[10] Myriam Ben Saad 和 Stéphane Gançarski。2011 年。使用页面变化模式进行网页存档：案例研究。载于 ACM/IEEE数字图书馆联合会议（JCDL）论文集。113–122。[11] Myriam Ben Saad、Zeynep Pehlivan 和 Stéphane Gançarski。2011 年。使用模式对网络档案进行面向一致性的爬取和导航。载于数字图书馆理论与实践会议（TPDL）论文集。Springer，421–433。https://doi.org/10.1007/978-3-642-24469-8_42[12] John Berlin。2017 年。自 2016年11月1日以来，CNN.com 一直无法归档。https://ws-dl.blogspot.com/2017/01/2017-01-20-cnncom-has-been- unarchivable.html。[13] Karl-Rainer Blumenthal。2020 年。堆栈：使用 Brozzler 大规模高保真网页采集。Archive-It 博客，https://archive-it.org/blog/post/the- stack-brozzler。[14] Haley Bragg、Himarsha R. Jayanetti、Michael L. Nelson 和 Michele C. Weigle。2023 年。虚假信息十二人的 Instagram 账户页面中，可重放的不到 4%。载于 ACM/IEEE数字图书馆联合会议（JCDL）论文集。[15] Justin F. Brunelle、Michele C. Weigle 和 Michael L. Nelson。2015 年。使用双层爬取方法归档延迟表示。载于 iPRES 2015：第 12 届国际数字对象保存会议论文集。（另见 arXiv:1508.02315）。[16] Dimitar Denev、Arturas Mazeika、Marc Spaniol 和 Gerhard Weikum。2009 年。SHARC：面向质量感知网络归档的框架。《VLDB 基金会会刊》2，1（2009），586–597。[17] 自由软件基金会。2005 年。GNU wget。http://www.gnu.org/software/wget/wget.html。[18] Lesley Frew。2022 年。记者如何使用网络档案数据集。https://github.com/lesleyodu/webarchives-user-tasks/blob/main/webarchives_记者_用户_任务_clean.csv。[19] Lesley Frew。2022 年。大众媒体中的网络归档 II：记者用户任务。https://ws-dl.blogspot.com/2022/08/2022-08-04-web-archiving-in- popular.html。[20] Kritika Garg、Himarsha R. Jayanetti、Sawood Alam、Michele C. Weigle 和 Michael L. Nelson。2021 年。重放归档 Twitter：当你的鸟坏了时，它会把你拖垮吗？载于 ACM/IEEE数字图书馆联合会议（JCDL）论文集。160–169。https://doi.org/10.1109/JCDL52503.2021.00028[21] Mark Graham。2017 年。Wayback Machine 回放……现在带有时间戳！http://blog.archive.org/2017/10/05/wayback-machine-playback-now- with-timestamps/。[22] Mark Graham。2019 年。Wayback Machine 的立即保存页面功能焕然一新。http://blog.archive.org/2019/10/23/the-wayback-machines-save-page-now-is-new-and-improved。[23] 国际互联网保存联盟（IIPC）。2019 年。OpenWayback。https://github.com/iipc/openwayback/。[24] 国际互联网保存联盟（IIPC）。2022 年。面向所有人的基于浏览器的爬取系统。https://netpreserve.org/projects/browser- based-crawling/。[25] 互联网档案馆。2023 年。法律常见问题。https://archive.org/legal/faq.php[26] Himarsha Jayanetti。2020 年。Instagram 的归档情况如何？https://ws-dl.blogspot.com/2020/11/2020-11-04-how-well-is-instagram.html。[27] Andrew Kaczynski 和 Em Steck。2022 年。共和党参议员候选人 Blake Masters 删除了竞选网站上称 2020 年选举被特朗普窃取的相关表述。CNN（2022年8月）。https://www.cnn.com/2022/08/29/politics/blake-masters-campaign-website-changes/index.html[28] Ilya Kreymer。2013 年。PyWB——面向所有人的网络归档工具。https://github.com/ikreymer/pywb。[29] Alexi McCammond 和 Andrew Solender。2022 年。大清洗。Axios（2022年8月）。https://www.axios.com/2022/08/31/republicans-midterms[30] Kieren McCarthy。2018 年。美国上诉法院法官裁定，Archive.org 的 Wayback Machine 是合法的法律证据。《The Register》（2018 年 9 月）。https: // www.theregister.com/2018/09/04/wayback_machine_legit[31] Jason Miller 和 Addy Osmani。2022 年。网页渲染。https://web.dev/rendering-on-the-web/。[32] Gordon Mohr、Michele Kimpton、Michael Stack 和 Igor Ranitovic。2004 年。Heritrix 简介：高质量网络爬虫。载于第 4 届国际网络归档研讨会（IWAW）论文集（英国巴斯）。[33] Michael L. Nelson。2011 年。Memento-Datetime 不是 Last-Modified。http://ws-dl.blogspot.com/2010/11/2010-11-05-memento-datetime-is-not- last.html。[34] Michael L. Nelson。2013 年。Archive.is 支持 Memento。https://ws-dl.blogspot.com/2013/07/2013-07-09-archiveis-supports-memento.html。[35] Laurie Roberts。2022 年。Blake Masters 的堕胎立场改造仍在继续：他悄悄删除了网站内容。《亚利桑那共和国报》（2022年8月）。https: // w w w. a z c e ntral.com/story/opinion/op-ed/laurieroberts/2022/08/25/blake-masters-abortion-makeover-continues-website-scrub/7898088001

[36] Liz Skalka。2022。陷入争议的美国共和党候选人 J.R. Majewski 还夸大了自己的学历。 *HuffPost* （2022年10月）。 https://web.archive.org/web/20221117145333/https://www.huffpost.com/entry/jr-majewski-ohio_n_634832a0e4b03e8038d16cdf[37] Marc Spaniol、Arturas Mazeika、Dimitar Denev 和 Gerhard Weikum。2009。“抓得到我就算你厉害”：网页存档中连贯性缺陷的可视化分析。载于 第九届国际网络存档研讨会（*IWAW*）论文集。[38] Webrecorder Software。2022。Browsertrix Cloud。 https://browsertrix.cloud/。[39] Davis Winkie。2022。美国陆军在报道指出自杀防范规定逾期未落实后，封锁了公开报告。 《陆军时报》 （2022年8月）。 https://www.armytimes.com/news/your-army/2022/05/27/army-locked-public-report-after-story-on-overdue-suicide-regulation[40] Jonathan Zittrain、Kendra Albert 和 Lawrence Lessig。2014。Perma：界定并解决法律引文中的链接失效与参考文献失效问题。《法律信息管理》 14，02 （ 2014 ），88–99。