DOI:CNKI:11-2746/G2.20130114.1631.001

国外 Web Archive 研究与实践进展*

王 芳 史海燕

摘 要 Web Archive 采集并保存 Web 内容,满足当前和未来的访问和使用,其重要性已得到图书馆、档案馆、政府、企业等机构的广泛认可。本文在文献与网络资源调研的基础上,以 Web 采集、Web Archive 的保存、访问和使用为主线,构建了一个系统化的研究框架,并在此框架下梳理国外的相关研究与实践进展。国外的研究与实践值得国内借鉴,如:多主体参与、广泛的交流与合作、注重标准和规范的建设、构建类型多样的 Web Archives、对访问与使用的优化等方面,但该领域仍面临法律与伦理、新 Web 应用内容的归档、存档内容的长期保存、存档内容的多元化应用等问题和挑战。参考文献 61。

关键词 Web Archive Web 采集 网络归档 数字存储 网络档案馆

分类号 G273

Progress of Foreign Research and Practice in Web Archive

Wang Fang & Shi Haiyan

ABSTRACT Web archive is used to harvest and preserve Web contents, provide content access and use for current and future generations. Its significance has been widely recognized by libraries, archives, governments and businesses. On the basis of website investigation and literature review, this paper formulated an analysis framework along the main line of Web harvesting, preservation of Web pages, access to Web archive and use of archived Web contents. Under this framework, the progress in foreign practices and researches of Web Archive was combed. It was found that new advances, such as multi-agency participation, extensive exchanges and cooperation, standard development, building of different types of Web archives, optimization of access and use, and so on, as well as problems and challenges, including legal and ethical issues, archival of new Web application contents, long-term preservation of archived contents and multi-agency application of Web archives, in foreign practices and researches of Web archive, are worth learning by corresponding departments and agencies of China. 61 refs.

KEY WORDS Web Archive. Web harvesting. Web archiving. Digital preservation. Internet archives.

1 引言

随着网络技术应用的日益深入,互联网已经成为最重要的信息汇聚地与发散地。与此同时,网络信息也成为人类社会历史文化风貌的重要记录形式和宝贵的社会历史文化遗产。然而网络信息具有海量、异构、分布式管理、容易消失等特点,一旦消失将难以复原,会给组织或社会信息资源的长久保存和历史传承造成难以挽回的损失。因此,探索网络信息资源的归档与保存策略,满足当代及未来人们访问和使用的需求,成为信息资源管理研究的迫切任务。目前,世界许多国家的政府、档案馆和图书馆都在积极进行网络信息资源归档保存的

* 本文系教育部人文社会科学研究规划项目"基于多利益相关者价值焦点分析的电子政务整体性评价研究"(项目编号:10YJA870021)的阶段性研究成果。

通讯作者:王芳,Email:wangfangnk@nankai.edu.cn

理论研究和实践探索，Web Archive 是主要研究领域之一。

Web Archive(WA)在国内有多种翻译方法，如网络信息资源保存[1]、网页信息存档[2]、网页档案馆[3]、网络信息档案馆[4]等。这些翻译基本可以分为两类：一类侧重于归档保存的活动或行为，一类侧重于归档保存所形成的虚拟实体。相应地，对于 WA 的理解也可以划分为两类，一类认为 WA 是指有关主体有选择地对具有长期保存价值的网络信息进行捕获、归档、存储等档案化管理的过程，其基本目标是通过网络信息资源的存档，更全面真实地反映和再现社会活动的本来面貌，并满足相关主体对网络信息的长期利用需求[5]，这一观点是将 WA 作为归档保存活动或行为的典型代表；另一类则认为 WA 是建立在现代信息技术基础上，利用网络信息采集、整合、保存、发布等技术对网络信息资源进行管理、并通过网络存取的超大规模、分布式数字信息系统[4]。“Archive”的含义包括存档(动词)、档案或档案馆(名词)，而存档这一过程国外通常采用“Archiving”[6]。因此，笔者更倾向于后一类观点，将 WA 视为通过对网络信息的采集、归档、保存所形成的虚拟网络实体，并可以通过网络访问和使用，其实质是一个网络空间中的数字资源系统。需要说明的是，所有具有保存价值的网络信息资源均应纳入归档保存的范畴，但目前 WA 所关注的归档保存对象主要来自 Web(万维网，采用 HTTP 协议)空间，如网站、网页以及从网站或网页中抽取的内容，对于其他网络应用形式(如 FTP、Telnet)所承载的信息较少涉及。

国外 Web Archive 的实践已有十多年的历史，涌现出了各类 WA 项目，如国家层面的 PANDORA(澳大利亚国家图书馆)、联盟形式的互联网档案馆(Internet Archive，IA)、项目形式的“处于风险中的网络”(Web at Risk)等[1]，其研究和实践吸引了众多参与者，包括国家级的图书馆、档案馆、大学图书馆和研究机构、商业机构等，此外还创建了诸如国际互联网保存联盟(International Internet Preservation Consortium，IIPC)组织。可以说，国外 WA 的研究和实践积累了丰富的成果和经验，但是网络环境的飞速发展以及新技术与新应用的不断涌现，也对 WA 提出了新的挑战。目前我国 WA 的实践还处于起步阶段，相关研究涉及网络信息的采集[7-11]、国外项目的介绍评析与分析[1-3,12-13]、Web Archive 的一般性理论与策略[14-16]、特定类型网络信息资源的归档保存[17-18]等，缺少对 WA 过程的系统研究，对于 Web 存档保存过程中存在的各种问题及应对策略缺少全面考察。因此，十分有必要借鉴国外 WA 的实践经验和理论研究成果。本文在文献与网络资源调研的基础上，面向 Web 归档保存过程，以 Web 采集、Web Archive 的保存、Web Archive 的访问和使用为主线，构建一个系统化的 Web Archive 研究框架，并以此框架为基础梳理国外 Web Archive 相关研究与实践进展，分析 Web 内容归档保存过程中所面临的主要问题，发现对这些问题已有的应对方法和可能的应对策略。

2 Web 采集

Web 采集(Web harvesting)也称 Web 收集或 Web 收割，是依据某种策略，采用特定方法和工具对 Web 内容进行收集的过程，是构建 Web Archive 的基础，主要涉及采集范围的确定、内容发现和内容获取等问题。

2.1 采集的范围

采集的范围(scope)，即确定“采集什么”的问题，是 Web Archive 采集策略的核心。Web 信息数量庞大且更新迅速，对其中“重要”内容如不及时采集则有可能永远丢失，而 Web Archive 受资金、人力、物力等限制无法完成对所有 Web 内容的采集，因此，确定恰当的采集范围是关键。目前，国外 WA 在确定采集范围时主要采用批量采集策略和选择性采集策略。

(1)批量采集(bulk harvesting)。互联网档案馆(Internet Archive，IA)[19]是 1996 年由 Brewster Kahle 建立的美国非营利组织，是批量采集的典型代表。IA 对全球范围内的 Web 内容进行广泛收集，目的是为研究人员、学者、历史学家和普通大众

提供互联网上数字格式文献的永久访问与免费使用,目前,已拥有文本(电子图书)、音频、活动影像、软件、网页等各类资源。

(2)选择性采集(selective harvesting)。面对海量的 Web 信息,全面的采集并不现实,IA 也仅采集表层网(surface Web),大多数的 Web Archive 采用了选择性采集策略。选择性采集意味着采集"重要"内容而忽略其他部分,但对特定 Web 内容现在及未来的重要性进行鉴定并不容易,信息内容、信息形式、归档主体需求、法律、成本等方面的因素均会影响采集信息的选择[20]。目前已采用的选择标准包括域、主题、资源类型等。

基于域(domain-centric)的采集是国家层面的 Web Archive 常用的方法,出于保存本国历史文化遗产的目的,采集国家域名或特定通用顶级域名(如.com,.edu 等)下的 Web 内容。此外,地理信息、服务器位置、目标受众、语言、域名的所有者或出版者等[21]也是基于域的采集中参考的标准。

基于主题(thematic/topical)的采集通常由特定研究需求驱动。研究者在研究过程中经常会被 Web 内容的暂留性(ephemeral)所困扰,Web 站点生命周期的特点也无法满足科学验证或提供持久参考的需求。因此,一些研究机构和大学图书馆开始创建基于主题的 Web Archive,如德国海德堡大学图书馆的 DACS(Digital Archive for Chinese Studies)[22]和美国哥伦比亚大学图书馆覆盖 15 个主题的 Web Archive[23]。另一类基于主题的采集是由特定事件驱动的,也称为基于事件的(event-centric)的采集,典型的驱动事件如总统选举,法国国家图书馆 BnF(Bibliothèquenationale de France)的选举 Web Archive 即属此类[24]。

基于资源类型(resource-specific)的采集面向特定类型的资源展开。资源类型的界定可从多个角度进行,如来源、媒体类型、应用模式等,相应的 Web Archive 也可依此划分。荷兰格罗宁根大学的 Archipol(Archive of web sites of political parties in the Netherlands)[25]和英国的政府 Web Archive(the UK government Web archive)[26]仅采集政党或政府网站;法国的 Ina 项目[27]从 2009 年开始采集与音视频媒体相关的 Web 站点;2010 年,Twitter 将其所有 tweet 数据捐献给美国国会图书馆存档保存,澳大利亚的 PANDORA 项目采集部分博客站点[28],等等。Web 2.0 的应用日益广泛,所汇聚的资源也日益丰富,但对 Web 2.0 内容的存档保存还很不充分,仅有少量 Web Archive 涉及某些应用。此外,并没有对在很多领域得到广泛应用的 Web 日志进行采集,这也是未来应引起关注的问题之一。

在确定采集范围时还有一些更为细节的问题需要考虑,如是否遵循 robots.txt 协议,是否排除指向范围外域名上资源的转向地址(redirects),是否排除离开"域"所在范围的被链资源(如 PDF 文件、视频文件、html 页面)等。国外 WA 一般都遵循 robots.txt 协议,认为该协议规定的内容是在采集范围之外的,但是图片、动画、音视频文件等内容除外。对于是否排除指向范围外域名上资源的转向地址,目前并没有标准的最佳实践,而离开"域"所在范围的被链资源可能会被纳入采集范围,也可能被排除,但通常会排除所有的 MIME 类型[21]。

2.2 Web 内容的发现

确定了 Web 采集的范围,在具体实施时首要的问题是如何发现采集范围内的 Web 内容。自动化的内容发现方法一般通过网络爬虫追踪超链接来实现,首先要预设网络爬虫的种子列表,在抓取相关页面后抽取其中的超链接,并从中发现新的资源。在基于域的采集中,网络爬虫的种子列表可以由分配域名的公司提供,或由网络提供商提供,或者通过与已经获取了相当数量域名的组织合作获取[29]。在基于主题或资源类型的采集中,通常是由相关专家或专业人员提供不断更新的种子列表。此外,一些 Web Archive 在其网站上设置站点推荐功能,由用户向其提供 URL。

对于 Web Archive,自动化的内容发现方法是主要途径。网络爬虫最初用于搜索引擎,而最早将网络爬虫技术应用于 Web 内容保存兴起于 1996 年的瑞典,之后的 WA 实践中既有采用现有爬虫工具的项目,也有自行开发工具的项目。爬虫中较为特殊的是主题爬虫(focused crawler 或 topical crawler)

和特定语言爬虫(language specific crawler),可分别用于特定主题的 Web 内容发现和特定语言的 Web 内容发现。理想的主题爬虫应仅下载与特定主题相关的页面并避免下载其他内容,具体的实现方法有多种,如在实际下载页面前通过超链接的锚文本判断相关性[30],或通过预先训练好的分类器对下载完毕的页面内容进行相关性分析[31]。Tamura 等提出了一种用于特定语言爬虫的方法,该方法可以不借助域名来发现特定语言的网页内容,基本的思路是通过预设的语言识别器来判断已下载的网页是否是目标语言[32]。

2.3 Web 内容的获取

获取(acquisition)是采用各种方法和工具从 Web 内容来源站点获得其复本的过程,既包括在线获取,也包括离线传送。具体而言,获取的方法可以分为客户端归档(client-side archiving)、事务性归档(transactional archiving)和服务器端归档(server-side archiving)[33]。

(1)客户端归档。也称为远程采集(remote harvesting),以客户端的形式采用网络爬虫获取 Web 内容,是 WA 实践中广泛采用的方法,常用的工具有 Heritrix、HTTrack、Wget 等。但网络爬虫有以下局限性:某些网站会用 robots. txt 文件限制网络爬虫对特定内容的访问;爬虫陷阱的存在;无法采集深层网(the deep Web);对于特定内容如流媒体无法下载。

(2)事务性归档。事务性归档是指对浏览器和 Web 服务器之间交互的事务(transaction)进行记录并归档,即保存浏览器和服务器间的请求/响应对,可用于对特定网站内容的证据性保存,但它的实现需要在服务器端安装软件,需要服务器的配合。Los Alomos 实验室与 Old Dominion 大学合作的 Memento 项目是采用事务性归档方法的代表[34]。

(3)服务器端归档。服务器端归档指直接从 Web 服务器访问并获取资源而无需采用 HTTP 协议。这一方法相较于事务性归档,更加需要 Web 站点所有者的积极参与,只能在法定缴送(legal deposit)的框架下采用,目前还未见有较成熟的应用。

深网采集是 WA 领域的一个重要问题。Web 中大部分内容隐藏在深层网,其访问与获取涉及与数据库的交互同时受访问权限的限制,一般网络爬虫无法完成,需要特殊的方法和工具。法国 BnF 和澳大利亚国家图书馆分别开发了 DeepArc[35] 和 Xing[36] 两个工具。DeepArc 可以将关系数据库映射为 XML 模式,将关系数据库中的内容导出为 XML 文档,之后 Xing 可以在线传递这些内容。尽管网站原始的布局和行为不能被精确保存,但 Xing 允许基本查询与检索功能的复制。互联网技术的发展也给 WA 带来挑战,新的问题不断出现,多媒体内容是重要问题之一。从 RealPlayer 文件到播客(podcast),WA 所面临的问题不仅来自这些内容本身的复杂性,还有其传送系统的复杂性。数字媒体在互联网上的传送机制可以大致分为两类,一类可以通过 HTTP 协议传送和下载,另一类则是流媒体。流媒体不会在客户端形成任何形式的复本,不可以通过 HTTP 协议下载,对流媒体的捕获和再现涉及诸多技术难题。澳大利亚图书馆[37]和英国国家图书馆[38]进行了有益的尝试,同时新的技术——HTML5 的出现也为解决这一问题带来福音。

3 Web Archive 的保存

保存(preservation)是 Web Archive 的首要任务,是保证对 Web Archive 现在及未来访问和使用的基础,涉及 Web 内容的存储及长期保存。

3.1 Web 内容的存储

Web 内容的存储是将采集获取的 Web 内容保存于 Web Archive 中的过程,需要考虑存储空间、存档格式、元数据、存储系统、复本管理等问题。

Web Archive 对 Web 内容的归档保存是一项持续性的活动,其存档的资源数量将不断增长,因此有必要估计所需的存储空间及其部署的位置[39]。影响 WA 所需存储空间的主要因素是:采集的类型是增量式采集还是非增量式采集,前者对

相同文件仅保存一次,后者则保存每一个文件的每一个复本[29]。Web Archive 中的文件存档格式有多种,如 ARC、WARC、CDX 等,IIPC 推荐使用 WARC。WARC(Web ARChive)存档格式规定了一种将多种数字资源与其相关信息(如元数据)整合为一个存档文件的方法,用以更好支持 Web Archive 的采集、访问和信息的交换[40]。

元数据也是 Web Archive 存储的重要问题。一些 Web Archive 在采集过程中会记录 Web 内容的某些信息,如 URL、校验值、采集时间等,这些数据可以作为元数据使用。数字资源保存领域的元数据标准如 PREMIS(PREservation Metadata: Implementation Strategies)和数字图书馆领域的元数据标准如 METS(Metadata Encoding and Transmission Standard)为 WA 元数据的选取提供了参考[41]。IIPC 在 2005 年提出一个用于 Web Archive 的元数据集,包括与文件相关的数据、与爬虫和服务器相关的数据、与网络爬行过程相关的数据、与选择过程相关的数据等[42]。此外,互联网档案馆的 WAT(Web Archive Transformation)描述了一种从 WARC 文件中抽取结构化数据的方法,WAT 数据可用于大规模数据集上的数据分析[43]。

Web Archive 对 Web 内容的存储意味着重建一个可以提供用户访问的系统,在理想状态下,WA 应同构于其存档的 Web 内容(包括层级结构、文件名、链接机制、文件格式等)。Web Archive 存储更多的挑战来自于对 Web 信息系统的重现,即重现所存档信息的内容、形式、结构等。但 Web 信息系统呈现出复杂的信息结构,所采用的操作系统、服务器配置、应用环境各不相同,为 Web Archive 的重现带来很多困难。目前,解决的策略主要有三种:第一种是在本地建立目标网站的复本,并以与 Web 相同的方式浏览这些复本;第二种是建立 Web 服务器,在这个环境中向用户浏览器提供服务内容;第三种是依据不同的命名、地址和再现逻辑重新组织文件。这三种策略各有优劣,适用于不同的 Web Archive[33]。

从技术角度看,Web 内容的存储还有一些更为具体的问题,如复本的管理。由于 Web 本身的属性,WA 在采集过程中不可避免地产生很多复本,如不同 URL 指向同一文件内容、多次采集的 Web 内容没有更新或仅有少量更新,对于这些复本的保存会浪费大量的存储空间,复本管理对 WA 而言是十分有益的。目前的研究已关注到这一问题,并提出一些解决的方法。对于不同 URL 指向同一文件的问题,可以通过使用统一资源名(Universal Resource Name, URN)解决[44];对于多次采集的 Web 内容没有更新的问题,可以通过历史数据估计 Web 更新周期以避免重复采集[45];对于部分更新的情况,可以采用三角洲存储(delta storage)[46]。但 Daniel Gomes 等提出,以上各种方法均有局限性,去除部分重复的做法并不适合 WA,而基于文件指纹去除完全重复的轻量级方法更为适宜[47]。

3.2 Web 内容的长期保存

Web 内容的长期保存属于数字信息资源长期保存的范畴,是要保证对存档内容在未来的长期访问和使用,是 Web Archive“存档”这一含义的重要体现。Web Archive 中存档的信息不但数量庞大而且不断增长,同时,内容类型多样,对象间存在复杂的链接关系,不同时间段采集的内容同时存在,相较于其他数字资源的长期保存,WA 的长期保存面临着更为严峻的挑战,包括处理不断更新的文件和软件版本,为在不同时间段采集的内容提供同时访问,维护相互链接对象间现时的、结构化的情境和关系,以及维护在这些对象间历时浏览的能力[48]。

IIPC 的保存工作组一直在探讨其他数字资源的长期保存策略对 Web Archive 的适用性,并致力于识别 WA 长期保存所面临的特殊性问题,为保证 WA 的长期可访问性,提出迁移、仿真、存档访问软件和相关文档、风险识别、记录转换和替代的访问路径等可能的长期保存策略[48]。Michael Day 提出 Web 之所以是一类特殊的保存对象,首先是因为 Web 本身是一个具有迷惑性的复杂对象,其次是 Web 的动态属性,而这种复杂性和动态性则反映了一个更深层次的问题,即缺乏对 Web 边界清晰、精确的界定[49]。此外,一些 Web 归档保存的项目也对长期保存的技术模型有所探讨,涉及的问题包括

导入(ingest)的工具、处理格式过时(obsolescence)和重复(replication)的方法、访问的工具等。Joseph 等提出系统化的 Web 内容长期保存的技术模型，需处理的问题包括数字对象的封装、技术演化的有效管理、有效的风险管理和灾难恢复机制、确保内容可用性和完整性的有效机制、信息发现及内容获取与保存的能力、导入率、容积和处理能力的可扩展性、兼容组织变化的能力等[50]。

4 Web Archive 的访问与使用

访问与使用是 Web Archive 价值的具体体现，也有利于其建设者监测 Web 采集是否达到预期目标。

4.1 可访问性

WA 可访问性面临的主要问题是确定提供何种方式的访问或允许何种人访问，不仅涉及具体的技术问题，更重要的是有关法律与伦理问题。Web Archive 提供了一种特殊的数字媒介，对 WA 的访问带来了不同于 Web 访问的问题，需要对现有的访问方法和工具进行调整，特别需要注意的是 WA 的时间维度，因为 WA 中往往保存着一个 Web 文件的多个版本。WA 的潜在用户对 WA 所提供的数据、信息和服务往往拥有非常不同的兴趣和期望。因此，除了以访问公开 Web 的方式来访问 WA 的需求外，WA 还要考虑其他类型的访问需求，如数据挖掘。现有的访问方式(浏览、索引、查询)对于以研究为目的访问是安全的，但更普遍的访问则依赖各国著作权法和法定缴送制度的完善[51]。

实践中不同的 Web Archive 采用了不同的访问策略。新西兰实行法定缴送，允许其国家图书馆保存任何已有的新西兰网站并提供对网站存档复本的访问。美国国会图书馆对其所有存档的网站编制书目记录，该书目记录允许公开访问，但只有已获取制作者许可的网站复本才允许公开访问。很多 Web Archive 是黑色存档或只能在特定地点访问，如芬兰、挪威、瑞士和奥地利等国家级的 Web Archive。一些可公开访问的 Web Archive 为避免和网站所有者竞争，对其资源的访问会有特定延时和功能的减少。如哈佛大学图书馆的 WAX 项目，从对某一网站的采集到将其存档复本在 WAX 中进行显示之间，存在至少 3 个月的延时，在 IA 的 Wayback Machine 中，这一延时是 6 至 12 个月[52]。

4.2 功能与服务

Web Archive 所承担的角色日趋多元，除保存外，提供科研服务和认证服务也是未来 Web Archive 的重要角色[53]。为满足用户多元化的访问与使用需求，Web Archive 应提供丰富的功能和服务。IIPC 的访问工作组于 2006 年发布《访问 Internet Archives 的用例》报告，将所有用例分为五类，每一类都需要若干功能和服务的支持[54]。Jinfang Niu 将 Web Archive 的功能划分为查询参数、查询结果、浏览等七类[52]。结合已有研究和国外 Web Archive 的实践，本文将 Web Archive 的功能和服务划分为以下几类。

(1)浏览。浏览一般可按字顺、主题、区域或媒体类型进行，要求 Web Archive 对其资源按相应方式进行组织。对同一 URL，可以借助 WayBack Machine 一类的工具在其不同存档版本间浏览。

(2)查询。信息检索领域已经发展了丰富的技术和方法，WA 可以充分借鉴，但目前 WA 实践中所提供的查询功能与实际 Web 的查询功能还有很大差距。Web Archive 常用的查询途径包括 URL、关键词、域名等，日期和媒体类型可以作为限定查询的方法，关键词查询需要 WA 为其存档资源建立全文索引，这一技术问题令部分 WA 仅提供 URL 查询。大部分 WA 只提供简单检索功能，少量提供高级检索，如加拿大政府的 Web Archive[55]。而信息检索领域的热点技术如多媒体检索、智能检索、自然语言检索等在 WA 查询中则完全没有体现。此外，Web Archive 的查询也有其特殊性，如在结果处理方面，WA 需要揭示来自同一网站页面间的层级关系、为满足认证和引用的需要为每一存档页面(包括不同版本)分配唯一永久标识符、提供页面打印功能等。

(3)数据挖掘。将 Web Archive 中的收藏应用

于学术研究是当前的一个趋势,数据挖掘即是一个主要应用方向。WA 中大量的累积性数据为数据挖掘提供了无限的研究可能,如美国康奈尔大学的 Web Library 基于互联网档案馆的数据所进行的数据挖掘研究和日本基于 Web Archive 的社会感知系统研究[56]。但这些研究仅小范围开展,WA 中大量有价值的收藏还未得到充分利用。从技术角度看,数据挖掘要求 WA 提供编程化、自动化的应用程序接口(API),这是目前制约其研究发展的主要因素之一。

(4)个性化服务。提供诸如"My Archive"之类的个性化服务可以提升用户体验,推动 Web Archive 的应用,但目前还未见这一类服务。

(5)站点重构。Web Archive 可以利用其存档内容帮助丢失的网站进行恢复,如 IA 利用其采集和存储的信息帮助过很多网站进行恢复。

(6)Web Archive 的分析。对 Web Archive 自身的分析,主要包括资源分析和用户使用分析,有利于 WA 更好的发展。澳大利亚的 PANDORA 每月发布一组关于 Archive 数据规模和月增长量的统计数据,此外还有每月一次的用户使用报告、对相邻两月进行比较的新增资源报告。英国的 Web Archive 除提供每月一次的资源统计数据外,还利用数据挖掘技术提供三项可视化服务:为 Web Archive 中的短语或词生成 N-gram、标签云和 3D 墙[57]。

5 对我国的启发及未来的发展方向

Web Archive 是一个实践性很强的领域,同时也包含众多待研究的问题。我国从 2001 年开始 Web 内容归档保存的实践,先后有北京大学网络实验室开发建设的"中国 Web 信息博物馆"、国家图书馆开展的网络信息资源采集与保存试验项目(Web Information Collection and Preservation,WICP)等。"中国 Web 信息博物馆"可通过其网站(http://www.infomall.cn/)访问,目前维护有约 400 亿的网页,提供 URL 查询和事件查询,并可通过 API 访问。但与国外相比,国内研究与实践的深度和广度都存在一定差距,而中文 Web 信息的独特价值和不断增长的数量又使国内 WA 研究和实践的重要性日益凸显,借鉴国外 WA 经验并探索适宜中文 Web 信息归档保存的理论、方法、技术成为迫切需求。国外 WA 实践有诸多成功之处,如多主体参与、广泛的交流与合作、注重标准和规范的建设、构建类型多样的 WA、对访问与使用的优化等。同时,仍有许多问题和挑战,需要国内的研究和实践加以关注:

(1)法律与伦理问题。Web 内容同其他出版物一样都受到知识产权的保护,WA 对 Web 内容的归档保存面临的法律问题主要有三个环节:收集网络信息、提供存取以及长久保存。建立和完善数字呈缴制度并修改相应的知识产权法来解决这一问题是目前较为一致的看法[58],但其具体实施并不容易。而伦理问题则更为复杂。WA 中有大量历史性数据,借助于更为先进的工具,使用者也许会发现内容创建者并不希望他人发现的信息,隐私权和数据保护的问题由此而生[59]。

(2)新 Web 应用内容的归档保存。Web 是异常活跃的领域,新的应用不断出现,如 Web 2.0。这些新应用形式同样是人类历史文化风貌的重要记录,但相较于传统 Web 内容,新的 Web 应用内容更难监测,隐私性更强,更新更为迅速。是否对其进行归档保存、保存哪些内容、如何保存等都是 WA 领域应深入研究的问题。

(3)存档内容的长期保存问题。Web 信息的长期保存是数字信息资源保存领域一个较为特殊的课题,而 WA 的时间维度又增加了其长期保存的复杂性。此外,也有档案学者提出 WA 的采集机制有悖于档案学中的鉴定原则[60],是否 WA 中所有的内容均需要长期保存或如何将鉴定原则应用于 Web 内容的归档保存也是值得思考的问题。

(4)存档内容的多元化应用。虽然 WA 的使用问题受到越来越多的关注,但与功能和服务都日益强大和丰富的 Web 空间相比,WA 的应用还有不小的差距。除伦理与法律问题外,技术的适用性问题是制约 WA 应用的主要因素之一。Web 空间的技术和工具一般不能直接应用于 WA,需要适当调

整，对 WA 的构建带来不小的挑战。然而，WA 的多元化应用特别是应用于学术研究是目前的趋势，如“利用 Web Archive 的数字研究”（Digital Research Using Web Archives）工作组正致力于研究利用 Web Archive 开展科学研究的法律与伦理、可利用性和限制、以“大数据”方式利用 WA 的技术需求等问题[61]。因此，积极向 Web 空间学习并探索各项技术在 WA 中的适用性，对于充分体现 WA 的价值是十分必要的。

参考文献：

[1] 向菁，吴振新，司铁英，等. 国际主要 Web Archive 项目介绍与评析[J]. 国家图书馆学刊，2010(1)：64－38.（Xiang Jing, Wu Zhenxin, Si Tieying, et al. Introduction and reviews of international principal Web archive projects [J]. Journal of the National Library of China, 2010(1)：64－38.）

[2] 耿磊. 起步阶段的网页信息资源长期保存[J]. 上海档案，2012(2)：13－15.（Geng Lei. The long-term preservation of Web pages in its infancy[J]. Shanghai Archives, 2012(2)：13－15.）

[3] 闫晓创. 国外 Web Archive 项目对我国的启示——以澳大利亚 PANDORA 为例[J]. 浙江档案，2011(10)：29－32.（Yan Xiaochuang. The inspiration of foreign Web archive projects to us：Using Australian PANDORA as a case study[J]. Zhejiang Arhives, 2011(10)：29－32.）

[4] 杨道玲. 中文网络信息资源保存问题探讨[J]. 档案学研究，2006(3)：39－42.（Yang Daoling. Exploration of the preservation of Chinese Web resources[J]. Archives Science Study, 2006(3)：39－42.）

[5] 周毅. 网络信息存档：档案部门的责任及其策略[J]. 档案学研究，2010(1)：70－73.（Zhou Yi. Web archive：The liabilities and maneuvers of archival departments[J]. Archives Science Study, 2010(1)：70－73.）

[6] Wikipedia. Web archiving[EB/OL]. [2012－06－10]. http://en.wikipedia.org/wiki/Web_archiving.

[7] 刘兰，吴振新，张智雄，等. Web Archive 的采集策略研究[J]. 现代图书情报技术，2009 (1)：10－15.（Liu Lan, Wu Zhenxin, Zhang Zhixiong, et al. Study on the harvest strategies in Web Archive[J]. New Technology of Library and Information Service, 2009 (1)：10－15.）

[8] 刘兰，吴振新. Web Archive 信息采集流程及关键问题研究[J]. 情报理论与实践，2009，32(8)：113－117.（Liu Lan, Wu Zhenxin. The work flow and its key issues of information gathering in Web Archive[J]. Information Studies：Theory & Application, 2009, 32(8)：113－117.）

[9] 刘兰，吴振新. 网络存储信息采集方式研究[J]. 图书馆杂志，2009 (8)：28－31.（Liu Lan, Wu Zhenxin. Study on Web harvest methods in archive preservation[J]. Library Journal, 2009 (8)：28－31.）

[10] 付光宇. 国外网络信息资源采集研究及其启示[J]. 图书情报论坛，2009 (4)：40－42.（Fu Guangyu. The harvesting of foreign Web information resources and its enlightments[J]. Library & Information Science Tribune, 2009 (4)：40－42.）

[11] 吕淑萍，朱兵. 网络信息资源采集内容的甄选——国家图书馆“专题存档”的实践[J]. 国家图书馆学刊，2004 (2)：30－33.（Lv Shuping, Zhu Bing. Content selection of gathered Web information resources：The practice of “thematic archiving” by the National Library of China[J]. Journal of the National Library of China, 2004 (2)：30－33.）

[12] 徐健. 英国网络信息保存联盟计划（UKWAC）及其启示[J]. 图书馆论坛，2007(4)：81－84.（Xu Jian. UKWAC project and its significance to China[J]. Library Tribune, 2007(4)：81－84.）

[13] 朱莲花，刘春燕. 韩国的国家知识门户网站与 Web Archive 现状研究[J]. 情报理论与实践，2010，33 (7)：120－123.（Zhu Lianhua, Liu Chunyan. Research on the status quo of Korea's National Knowledge Portal website and Web Archive[J]. Information Studies：Theory & Application, 2010, 33 (7)：120－123.）

[14] 李华，吴振新，郭家义，等. Web Archive 发展历程与发展趋势研究[J]. 现代图书情报技术，2009(1)：2－9.（Li Hua, Wu Zhenxin, Guo Jiayi, et al. Study on the progress and trend of the development of Web Archive[J]. New Technology of Library and Information Service, 2009(1)：2－9.）

[15] 周林兴. Web Archive 保存研究:现状、意义与发展策略[J]. 档案管理, 2009(5): 26 -28. (Zhou Linxing. Study on Web Archive: Current situation, implication and developing strategies[J]. Archives Management, 2009(5): 26 -28.)
[16] 周毅. 论网络信息存档权及其生成[J]. 中国图书馆学报, 2011, 37(1): 102 -108. (Zhou Yi. On the discovery and recognition of Web Archive right[J]. Journal of Library Science in China, 2011, 37(1): 102 -108.)
[17] 谢春枝. 博客长期存取的现状和对策研究[J]. 图书情报知识, 2009(6): 81 -86. (Xie Chunzhi. Research on long-term preservation and access of blogs[J]. Document, Information & Knowledge, 2009 (6): 81 -86.)
[18] 唐琼. 政府网络信息资源长期保存研究[J]. 图书馆理论与实践, 2007(2): 62 -64. (Tang Qiong. Study on the long-term preservation of government Web information resources[J]. Library Theory and Practice, 2007 (2): 62 -64.)
[19] Internet Archive[EB/OL]. [2012 -06 -10]. http://archive.org/.
[20] 仇壮丽,许冬玲. 归档网络信息选择策略的影响因素研究[J]. 档案学研究, 2011 (3): 63 -66. (Qiu Zhuangli, Xu Dongling. Analysis on the impacting factors of selecting strategy in Web archiving[J]. Archives Science Study, 2011 (3): 63 -66.)
[21] Michaela M. International Internet preservation consortium harvesting practices report 2011[EB/OL]. [2012 -06 -10]. http://readyresources-leslierknoblauch.com/2011/08/17/internet-harvesting-practices-report/.
[22] DACS[EB/OL]. [2012 -06 -10]. http://www.sino.uni-heidelberg.de/dachs/.
[23] Web archive[EB/OL]. [2012 -06 -10]. http://www.archive-it.org/organizations/304.
[24] Masanès J. Web archiving methods and approaches:A comparative study[J]. Library Trends, 2005, 54(1): 72 -90.
[25] Archipol[EB/OL]. [2012 -06 -10]. http://www.archipol.nl/english/index.html.
[26] The UK government Web archive[EB/OL]. [2012 -06 -10]. http://www.nationalarchives.gov.uk/webarchive/.
[27] INA Web archive[EB/OL]. [2012 -06 -10]. http://www.ina.fr/.
[28] PANDORA[EB/OL]. [2012 -06 -10]. http://pandora.nla.gov.au.
[29] Hakala J. Archiving the Web: European experiences[J]. Electronic Library and Information Systems, 2004, 38(3): 176 -183.
[30] Wikipedia. Focused crawler[EB/OL]. [2012 -09 -20]. http://en.wikipedia.org/wiki/Focused_crawler.
[31] Chakrabarti S, Berg M, Dom B. Focused crawling: A new approach to topic-specific Web resource discovery[J]. Computer Networks, 1999, 31(11 -16): 1623 -1640.
[32] Tamura T, Somboonviwat K, Kitsuregawa M. A method for language-specific Web crawling and its evaluation[J]. Systems and Computers in Japan, 2007, 38(2): 10 -20.
[33] Masanès J. Web archiving: Issues and methods[J/OL]. Web Archiving, 2006:1 -53[2012 -09 -20]. http://www.researchgate.net/publication/226948825_Web_Archiving_Issues_and_Methods.
[34] Memonto[EB/OL]. [2012 -06 -10]. http://mementoweb.org/depot/native/odusource/.
[35] DeepArc[EB/OL]. [2012 -06 -10]. http://deeparc.sourceforge.net/.
[36] Xing[EB/OL]. [2012 -06 -10]. http://www.nla.gov.au/xinq/.
[37] Crook E. Web archiving in a Web 2.0 world[J]. The Electronic Library, 2009, 27(5): 831 -836.
[38] Hockx-Yu H, Crawford L, Coram R, et al. Capturing and replaying streaming media in a Web archive: A British library case study[EB/OL]. [2012 -09 -20]. http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/hockxyu -44.pdf.
[39] Farrell S, Ashley K, Davis R, et al. A guide to Web preservation[EB/OL]. [2012 -06 -10]. http://jiscpowr.jiscinvolve.org/files/2010/06/Guide -2010 -final.pdf.
[40] WARC[EB/OL]. [2012 -06 -10]. http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml.
[41] Pennock M. Beyond the harvest: Long term preservation of the UK Web archive[EB/OL]. [2012 -06 -10]. http://www.dpconline.org/component/docman/doc_download/397 -0907pennockmissinglinks.
[42] IIPC. IIPC Web archiving metadata set[EB/OL]. [2012 -06 -10]. http:// iwaw.europarchive.org05/masanes2.pdf.
[43] WAT[EB/OL]. [2012 -06 -10]. https://webarchive.jira.com/wiki/display/Iresearch/Web + Archive + Transformation + (WAT) + Specification, + Utilities, + and + Usage + Overview.

[44] Daigle L, Gulik van DW, Iannella R, Faltstrom P. Uniform Resource Names (URN) name space definition mechanisms [EB/OL]. [2012-09-20]. http://www.ietf.org/rfc/rfc3406.txt.

[45] Cho J, Garcia-Molina H. Estimating frequency of change[J/OL]. ACM Transactions on Internet Technology, 2003(3) [2012-09-20]. http://oak.cs.ucla.edu/~cho/papers/cho-freq.pdf.

[46] MacDonald J. Versioned file archiving, compression, and distribution[EB/OL]. [2012-09-20]. http://reference.kfupm.edu.sa/content/v/e/versioned_file_archiving__compression__a_17205.pdf.

[47] Gomes D, Santos A L, Silva M J. Managing duplicates in a Web archive[C]// Proceedings of the 2006 ACM Symposium on Applied Computing, 2006:23-27.

[48] IIPC Preservation Working Group. Preserving access—Making more informed guesses about what works[EB/OL]. [2012-06-10]. http://netpreserve.org/publications/preservingaccess.pdf.

[49] Day M. The long-term preservation of Web content[M/OL]// Web Archiving. Berlin: Springe, 2006: 177-199[2012-06-10]. http://www.ukoln.ac.uk/preservation/publications/2006/web-archiving/md-final-draft.pdf.

[50] JaJa J, Smorul M, Song S. Tools and services for long-term preservation of digital archives[EB/OL]. Trends in Digital Preservation, 2009. [2012-09-10]. https://wiki.umiacs.umd.edu/adapt/images/2/2a/Indo-US-Workshop-jaja.pdf.

[51] Hallgrimsson T. Web archiving: Access and finding aids[M/OL]// Web Archiving. Berlin: Springe, 2006: 131-151. [2012-06-10]. http://www.springerlink.com/content/jx881q7544545757/.

[52] Niu J. An overview of Web archiving[J/OL]. D-Lib Magazine, 2012, 18(3/4)[2012-06-10]. http://www.dlib.org/dlib/march12/niu/03niu1.html.

[53] Leetaru K L. A vision of the role and the future of Web archive[EB/OL]. [2012-06-10]. http://netpreserve.org/publications/Kalev2012.pdf.

[54] International Internet Preservation Consortium Access Working Group. Use cases for access to Internet archives[EB/OL]. [2012-06-10]. http://netpreserve.org/publications/iipc-r-003.pdf.

[55] Government of Canada Web Archive[EB/OL]. [2012-06-10]. http://www.collectionscanada.gc.ca/webarchives/index-e.html;jsessionid=85D78F4590C129BD2EFB9C4593A6379E.

[56] 吴振新，张智雄，孙志茹. 基于数据挖掘的 Web Archive 资源应用分析[EB/OL]. [2012-06-10]. http://ir.las.ac.cn/bitstream/12502/629/1/基于数据挖掘的 WebArchive 资源应用分析.pdf. (Wu Zhenxin, Zhang Zhixiong, Sun Zhiru. An analysis of the application of Web Archive resources based on data mining[EB/OL]. [2012-06-10]. http://ir.las.ac.cn/bitstream/12502/629/1/an analysis of the application of Web Archive resources based on data mining.pdf.)

[57] UK Web archive[EB/OL]. [2012-06-10]. http://www.webarchive.org.uk/ukwa/.

[58] 赵俊玲. 国外关于网络信息资源保存的研究[J]. 中国图书馆学报, 2004,(3): 80-83. (Zhao Junling. A study of the preservation of network information resources in foreign countries[J]. Journal of Library Science in China, 2004,(3): 80-83.)

[59] Rauber A, Kaiser M, Wachter B. Ethical issues in Web archive creation and usage—Towards a research agenda[EB/OL]. [2012-06-10]. http:// iwaw.europarchive.org/08/IWAW2008-Rauber-pres.pdf.

[60] Theimer K. NARA and the Web harvest: A discussion of the issues[EB/OL]. [2012-06-10]. http://www.archivesnext.com/?p=137.

[61] Digital research using web archives[EB/OL]. [2012-09-20]. http://digital-research.oerc.ox.ac.uk/programme/web-archive.

王 芳 南开大学商学院信息资源管理系教授，博士生导师。
通讯地址：天津市南开区卫津路 94 号。邮编：300071。

史海燕 南开大学商学院信息资源管理系博士研究生，河北大学管理学院讲师。通讯地址同上。

(收稿日期：2012-10-12)