Web归档生命周期模型的发展研究*

吴硕娜[1] 黄新荣[2]
(1. 中山大学资讯管理学院，广州 511400；2. 西北大学公共管理学院，西安 710127)

摘要：随着互联网的发展，Web信息资源的价值逐渐被人们发现和重视，对Web归档生命周期模型进行发展和完善有利于规范Web信息资源的保存和归档。本文基于信息生命周期管理理论，重点分析Web归档生命周期模型的结构、内容以及显著优势。通过文献阅读和网络调研，发现该模型存在前端控制缺乏、Web档案资源利用不足等问题。针对这些问题，对该模型进行前端和后端扩展，最终得到Web生命周期管理模型，从内容和技术要求上为Web信息资源管理提供了详细指导，有利于更好地发挥Web信息资源的价值，延续Web信息的生命。

关键词：Web归档生命周期模型；Web生命周期管理模型；信息生命周期管理
中图分类号：G270
DOI：10.3772/j.issn.1673-2286.2018.10.006

随着信息技术的发展和应用，海量信息资源随之产生。IDC的报告预测2020年全球数据总量将达到40ZB[1]。为科学管理海量信息资源，充分利用其社会价值，2003年首次出现信息生命周期管理（Information Lifecycle Management）的概念，同年11月在美国网络存储国际大会上被普遍接受和认可[2]。ISO/TC 171技术委员会明确指出信息生命周期管理环节包括信息的生成、获取、标引、存储、检索、分发、呈现、迁移、交换、保护与最后处置或废弃[3]。EMC等数据存储服务商基于信息生命周期管理理念将信息活动划分为搜索、组织、保护/恢复、复制/监控、访问/共享、迁移/归档、删除/销毁[4]。目前关于信息生命周期管理虽然没有统一的环节划分，但各种划分在表述、阶段划分上稍有不同，主体的管理思路仍然是从信息资源的生成、使用到归档，再到归档活动结束后的利用处置。

互联网信息数量急剧增长，截至2017年，中国网站数已达523万家，网民数高达7.72亿人[5]。不仅网站中包含大量Web信息，网民在互联网上的信息行为也产生了大量Web信息，海量信息的出现使Web信息资源的管理和保存变得尤为重要。目前许多国家已经意识到Web信息资源的重要性并且开展了Web归档活动，尝试对网页资源和社交媒体等开展捕获、长期保存等管理活动。网络存档服务组织Archive-It根据众多归档项目的经验，提出Web归档生命周期模型（Web Archiving Life Cycle Model，WALCM）[6]。WALCM是有关Web归档活动的信息管理模型，其对归档展开的管理活动是信息生命周期系统管理流程的一部分，但也根据Web归档实践的特殊性进行了调整。WALCM对Web信息从捕获到长期保存的归档过程进行管理，其管理活动的主要特点：以业务为核心，基于政策，统一路径，异质环境和数据价值关联[4]。本文将探讨WALCM的不足，以期对WALCM进行发展和完善。

1 Web归档生命周期模型概述

目前国外对WALCM的研究较少，国家互联网保存联盟（International Internet Preservation Consortium，IIPC）在归档工具与软件部分对网络归档概念进行系统介绍中提到WALCM[7]，将其描述为一个用于归纳Web归档技术和归档环节程序化的框架。网络调研发现，日本国会图书馆在其网络归档项目（Web Archiving Project，WARP）中应用了WALCM[8]，网络归档生命周期由选择、集合、组织、保存、公开5部分组成，这5部分会跟踪网站信息并随着网站信息的更改发生变化。日本

*本研究得到国家社会科学基金项目“社交媒体文件的归档与管理标准体系研究”（编号：16BTQ093）资助。

国会图书馆在网络存档过程中借鉴了Archive-It提出的WALCM，并且根据实际需要对操作环节做出相应改变。WARP项目中对WALCM的应用也反映了模型的合理性。

1.1 模型的结构

该模型试图规范在开发和管理网络归档程序时所经历的不同操作步骤和阶段。虽然模型分解成单独的环节，但每个操作都不独立，操作步骤与阶段是相关的，它们之间有紧密的重叠。模型由政策带、外圈、元数据描述带、内圈和Web档案集一系列同心圆环组成[9]，在大型项目中，这些阶段和步骤可能会根据项目要求进行循环。政策带位于模型最外侧，这些政策规范基本涉及Web归档的方方面面。模型的外圈代表了机构在建立和管理其网络存档计划时所面临的高层决策，是基于管理层面的操作。Archive-It选择将元数据描述作为一个环，来强调创建、导入和导出元数据是一个持续的过程，与生命周期中的其他活动一起发生。模型的内圈主要描述Web归档业务所涉及的日常任务，与外圈相比，内圈更具体，主要包括评估与选择、归档范围界定、数据捕获、存储和组织、质量保证与分析5个操作层面的步骤。Web档案集是归档活动开展的前提和基础，是Web归档流程的首要环节。当归档单位向现有集合添加新网站、创建全新集合、审阅存档内容、修改网络爬虫软件设置或捕获范围时，归档程序会重新回到Web档案收集环节。

1.2 模型的优点

WALCM具有以下优点。①完整的网络归档指南。WALCM将Web归档实践概括为两条主线，一是政策层面，从管理者角度，帮助即将开展网络归档的机构明确归档目标、归档的最佳管理方法和流程，并制定相应的政策；二是实践层面，对Web信息资源进行评估与选择、归档范围界定、数据捕获、存储和组织、质量保证和分析操作，为归档机构提供合理、具体的可操作步骤，助其快速组织归档工作，完成归档目标。两条主线可帮助机构在短时间内确立较好的管理制度和操作步骤，为未来的Web归档实践提供可借鉴的操作指南，使网络归档的环节规范化，保存社会记忆。②模型简洁明了。WALCM以直观、可视的环状图呈现，便于用户理解和实践操作。一个归档机构进行归档活动时首要考虑的是目前有关Web归档的相关标准和政策，因此，政策带位于模型的最外层，是机构外部的大环境。机构在开展归档活动时，要明确机构的归档目标和确定管理制度等上层建筑，这对归档实践活动的开展非常重要。元数据和描述与WALCM的内圈和外圈都有显著重叠，因此位于两个圆环之间。围绕Web档案集的是Web归档的实践流程，即WALCM的内圈。

2 Web归档生命周期模型在实践应用中的不足

2.1 前端控制缺乏

档案部门在电子文件的管理中应用前端控制，将许多纸质文件的控制手段提前到电子文件管理的最前端。Web信息是在社会实践中直接生成的，这些信息符合文件的真实性、可靠性，具有鲜明的档案属性，是档案数字资源的重要组成部分。前端控制思想是档案领域应对电子文件时代的新思维，是信息技术发展的必然产物。在网络环境下同样要注重对前端控制的思考与发展，有利于避免社会记忆的缺失，保持文化的传承[10]。前端控制有利于建立统一的标准、保证信息内容的真实性[11]及保证后续管理环节的展开。

前端控制的思想在国外Web归档实践中也有体现，如美国国会图书馆在其网页收集项目中明确规定对网页生成和收集的5个要求，即技术特点、格式、交付方式、元数据和技术措施[12]。在技术特点中，首先要求使用站点地图，稳定的URL和开放格式，还应遵循无障碍标准。英国网络归档联盟在制订网络归档计划时也充分体现了前端控制思想[13]，联盟成员在开展归档前需要获得网站所有者的许可，同时合作制定兼容的选择策略，并调查收集和应对网络存档中可能涉及的复杂技术难题，做好统筹安排，使归档活动可以顺利开展。前端控制的思想也影响了东亚部分Web归档项目，在韩国国家图书馆的OASIS项目中，详细规定了网页的爬虫访问设置、网页标题、网址信息更改等设计标准及收集方面的相关政策[14]。考虑到著作权问题日本国会图书馆WARP项目在采集中制定了详细的采集标准，这一点也体现了前端控制思想[15]。

国外Web归档项目中有关前端控制思想的实践表明了前端控制在Web归档实践中的重要性。从实践案例中可以看出WALCM缺少前端控制。在生命周期中增加前端控制不仅有利于网页的获取和保存，更有助于整

个归档工作的开展。WALCM是从Web信息资源收集开始，在Web信息资源生成后，围绕网络资源进行捕获、组织、存储、回放/再利用，只是针对归档这一管理活动的模型。信息生命周期理论是从信息产生开始研究，将信息的生成作为管理的最前端，并且在前端控制思想的指导下，归档管理活动已经提前介入到资源的生成阶段，Web归档生命周期作为信息生命周期的下位概念，也应注重前端控制的应用。因此，从实践经验和理论基础上可以看出其不足体现在模型前端控制的缺乏。

2.2 后端利用不足

通过对现有Web归档实践项目进行观察，发现大量归档项目对归档资源展开多种形式的利用，充分挖掘归档信息中包含的资源。目前IIPC网站上提供了许多对Web归档资源利用的案例，主要是从链接分析和文本挖掘利用两个角度对数字档案资源展开利用。①链接分析法角度。不同站点和网页的相互联系构成网络链接，对链接进行分析能够反映出隐藏的网络结构。特别是创建者放置的链接或者用户之间分享的链接，可以视为对目标页面的认可，一般都指向有用或相关的资源[16]。目前主要有英国国家图书馆对英国Web档案进行的链接分析[17]，Babel 2012网络语言连接[18]和Zyxt实验室分析常见的抓取网站[19]。②文本挖掘利用角度。文本挖掘指有效地从文档内容及其描述中抽取知识，进行分类、聚类、趋势预测等[20]，可以满足用户的多样化需求，使用户能够短时间、全方面、准确地找到所需信息。目前IIPC展示的文本挖掘实践项目[21]有英国国家图书馆利用短语利用率可视化工具N-Gram，在英国网络存档中找到随时间变化的用户定义的搜索词或短语每月出现的次数；阿姆斯特丹大学利用网络档案搜索工具，重点收集荷兰新闻汇总网站，显示与主要新闻事件相关的词频可视化及随时间变化的词汇进行共生分析。

总体来看，不论是链接分析还是文本挖掘，都是对Web档案的分析利用。国外的机构和组织对Web归档的实践活动不仅停留在Web档案的长期保存和访问，更多地关注归档后Web档案的价值，从而更好地满足公众需求，促进档案事业和社会的发展。Web归档生命周期作为信息生命周期的子概念，也应遵循整体性原则，注重对归档资源的开发利用。因此，WALCM止步于Web档案的长期保存具有一定的局限性，忽略了Web生命周期中信息的形成阶段和在归档结束后对Web档案的分析利用。

3 Web归档生命周期模型的改进

3.1 Web生命周期管理模型的提出

针对目前WALCM模型的局限性，对其进行补充和扩展，可以得到一个系统、完整的Web生命周期管理模型，涵盖Web资源从生成到最后分析利用的管理全过程，即围绕网络资源的形成、获取、保存、组织和利用处置。Web生命周期管理模型扩展了WALCM的前端和后端。前端的扩展要求以WALCM为基础，强调对网络资源收集的控制始于Web归档流程的开端，并且贯穿于Web归档的整个过程。通过提前规定网络资源的格式，再对符合相关标准的网页进行捕获和收集，这样有利于保证捕获质量，使网络资源在生成时就符合Web归档捕获标准和Web档案保存要求；同时也要加强与IIPC、W3C等国际组织合作，共同制定涵盖网络资源从生成到后续各个管理环节的相关标准和规范。后端加强对归档后Web信息资源的分析、利用，通过对资源进行链接分析、文本分析、语义挖掘等，深度挖掘网络资源的价值；还可以对归档网络信息资源展开编研，将编研成果以线上展览等多元化的方式呈现，更好地为用户提供利用。通过对WALCM的前端和后端进行扩展，构建一个完整的Web生命周期管理模型。

Web生命周期是指Web信息资源本身从产生到最后利用处置的过程中，信息数量、效用价值、热度等信息生命指标的变化[22]，展示了网络信息运动的本质和规律。Web生命周期管理是一种管理模型，是对Web信息生命运动的各个阶段进行管理[23]，以保证Web信息资源发挥最大价值。本文中扩展得到的Web生命周期管理模型是从归档保存角度对Web信息资源从产生到利用处置各个生命阶段进行管理，对网络资源进行保存和归档，有助于延续网络信息资源的生命，归档为信息建立了一个镜像副本，即使原链接被删除，也可以通过利用归档的链接对原内容进行分析、利用。

3.2 Web生命周期管理模型的结构

Web生命周期管理模型的坐标轴图（见图1）清晰地呈现了在操作层面对WALCM的扩展，在WALCM的基础上前端增加了Web资源生成时的规范和标准，以及向后延续了对Web档案开展链接分析、文本挖掘等利用。在操作层面，Web生命周期管理模型的结构由Web资源

生成、评价与选择、归档范围界定、数据捕获、存储和组织、质量保证和分析、利用处置构成；在政策层面，在明确机构的归档愿景和目标后，增加对网络资源生成进行前端控制，对机构资源进行审查，制定Web归档计划，并在风险管理后，加入对Web档案的利用、处置。

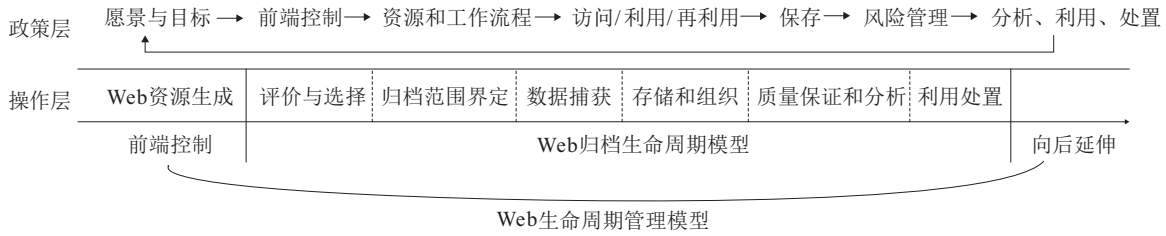

图1 Web生命周期管理模型坐标轴图

Web生命周期管理模型的坐标轴图仅横向展示Web生命周期管理的阶段和环节，没有对Web生命周期管理进行整体描述。Web生命周期管理模型环状图（见图2）弥补了这一不足，可从整体角度向用户展现各阶段的相互关系，便于从全局角度理解Web生命周期管理模型，是对坐标轴图的扩展。

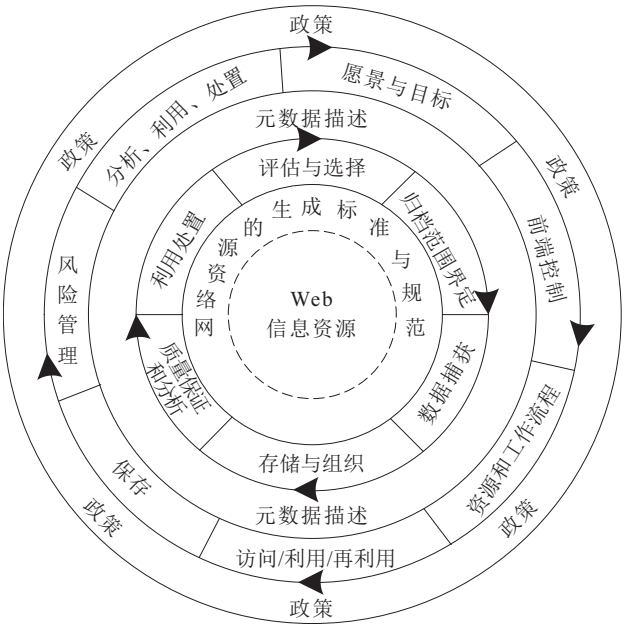

图2 Web生命周期管理模型环状图

在环状图中，Web生命周期管理模型的中央由Web信息资源的标准与规范构成，用虚线进行分隔，表明生成时标准与规范是对Web信息资源生成时的约束，Web信息资源是模型的核心。Web生命周期管理模型不只是对WALCM的前端和后端进行扩展，其最外侧的政策带含义也发生了改变。WALCM的政策是和Web归档有关的政策决定，这些外部的政策环境有可能会影响归档机构内部的政策变化和收集活动。Web生命周期管理的政策是针对Web信息从产生到利用、处置各个环节的相关政策决定、标准制定；互联网档案馆（IA）、IIPC联盟等国际组织的相继出现和影响力日益提升，表明了Web信息管理项目呈现从独立走向合作的趋势，在政策层面表现为：要加强和W3C等互联网组织合作，发展Web规范，解决Web应用中不同平台、技术、开发者导致的不兼容问题，确保Web信息的交换和迁移，督促Web应用开发者和内容提供者遵循这些标准[24]，使Web信息资源利于保存和分析利用。

4 结语

WALCM总结了网络归档时所经历的不同步骤和阶段，对Web归档实践具有指导意义。但这一模型也有较多的局限性，通过修正得到了Web生命周期管理模型，Web生命周期模型的提出使Web生命周期的各个管理环节更加具体，具有可操作性，同时也厘清了生命周期模型与生命周期管理模型的区别。下一步应密切关注Web生命周期管理模型与管理主体的关系，Web信息资源在不同的运行阶段，其管理的主体是不一样的，在资源生成阶段，主体是发布者或平台提供者；在归档阶段，主体变成两类：一类是信息资源的所有者，另一类可以是档案馆、图书馆或第三方机构。由于网页的共享性，归档主体不受限制，但是对于其他网络资源（如社交媒体），归档后资源的分析利用会产生权属问题，主要体现在知识产权方面。因此，有关Web信息资源所有权问题，即Web归档的风险管理是下一步研究应重点考虑的问题和关注的方向。

参考文献

[1] 数据存储单位从B/KB/MB/GB到NB/DB，如何转化？全球数据量有多少［EB/OL］.［2016-06-30］. http://www.360doc.com/

content/16/0630/08/982782_571834184.shtml.

[2] 唐竟. 基于信息生命周期管理的数据迁移技术研究 [D]. 长沙：湖南大学，2009.

[3] 李铭. 看国际动态 找国内差距 促技术发展 [J]. 数字与缩微影像，2002（2）：25-29.

[4] 索传军. 基于信息生命周期的数字馆藏管理研究 [J]. 大学图书馆学报，2005，23（1）：26-29.

[5] 第41次《中国互联网络发展状况统计报告》[EB/OL]. [2018-03-05]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/t20180305_70249.htm.

[6] Announcing the Web Archiving Life Cycle Model [EB/OL]. [2018-03-11]. https://archive-it.org/blog/post/announcing-the-web-archiving-life-cycle-model/.

[7] Tool & Software [EB/OL]. [2018-08-15]. http://netpreserve.org/web-archiving/tools-and-software/.

[8] ウェブアーカイブのライフサイクル [EB/OL]. [2018-03-27]. http://warp.da.ndl.go.jp/contents/reccommend/mechanism/mechanism02.html.

[9] The Web Archiving Life Cycle Model [EB/OL]. [2018-09-03]. http://ait.blog.archive.org/learn-more/publications/.

[10] 杨艳. 档案学前端控制思想的学术渊源和实践需求 [J]. 北京档案，2012（9）：16-18.

[11] 金鑫. 电子文件呼唤前端控制 [J]. 档案学研究，2004（6）：48-51.

[12] Section 508 Standards for Electronic and Information Technology [EB/OL]. [2018-03-30]. https://www.access-board.gov/guidelines-and-standards/communications-and-it/about-the-section-508-standards/section-508-standards.

[13] Frequently asked questions：Is permission asked first before a website is archived? [EB/OL]. [2018-08-15]. https://www.webarchive.org.uk/ukwa/info/faq/.

[14] OASIS 온라인 보관 검색 인터넷 소스 [EB/OL]. [2018-03-27]. http://www.oasis.go.kr/about/guide.do.

[15] インターネット資料収集保存事業の概要 [EB/OL]. [2018-08-15]. http://www.ndl.go.jp/jp/collect/internet/index.html#anchor02.

[16] 汪涛. 基于链接分析的网络信息行为研究 [D]. 合肥：安徽大学，2013.

[17] Link Analysis [EB/OL]. [2018-04-01]. https://www.webarchive.org.uk/ukwa/visualisation/ukwa.ds.2/linkage.

[18] Babel 2012 Web Language Connections [EB/OL]. [2018-04-01]. https://github.com/norvigaward/2012-naward25/wiki/Babel-2012---Web-Language-Connections.

[19] Study of ~1.3 Billion URLs：~22% of Web Pages Reference Facebook [EB/OL]. [2018-04-01]. http://zyxt.com/post/26851542949/ study-of-13-billion-urls-22-of-web-pages.

[20] 李锐. 网页文本分类挖掘的几种算法研究 [J]. 福建电脑，2008（10）：36，59.

[21] Text Mining [EB/OL]. [2018-04-01]. http://netpreserve.org/web-archiving/case-studies/.

[22] 朱建军，周强. 互联网信息生命周期研究 [J]. 铁路计算机应用，2015，24（3）：45-49.

[23] 索传军. 试论信息生命周期的概念及研究内容 [J]. 图书情报工作，2010，54（13）：5-9.

[24] 周洪喜. 基于人工标注技术的网页内容抽取系统开发 [D]. 上海：复旦大学，2010.

作者简介

吴硕娜，女，硕士研究生，研究方向：网络信息资源管理、信息资源整合与档案数字化，E-mail：945946601@qq.com。
黄新荣，男，博士，副教授，研究方向：电子文件管理、网络信息资源管理。

Research on the Development of Web Archive Life Cycle Model

WU ShuoNa[1] HUANG XinRong[2]
(1. School of Information Management, Sun Yat-sen University, Guangzhou 511400, China;
2. School of Public Administration, Northwest University, Xi'an 710127, China)

Abstract: With the development of internet, the value of Web information resources has been gradually discovered and valued by people. Developing and perfecting the life cycle of Web archives model is conducive to standardizing the preservation of Web information resources. Based on the theory of information lifecycle management, this article focuses on the structure, content, and significant advantages of the model of Web archive life cycle. Through literature reading and network research, it is found that there are problems such as lack of front-end control and insufficient utilization of Web archive resources in this model. To solve these problems, the front-end and back-end of the model are extended to finally obtain the Web lifecycle management model. The Web lifecycle management model provides detailed guidance for Web information resource management from the content and technical requirements, which is conducive to the better use of the value of Web information resources and the continuation of the life of Web information.

Keywords: Web Archiving Life Cycle Model; Web Lifecycle Management Model; Information Life Cycle Management

（收稿日期：2018-09-03）