基于网络信息资源保存的生命周期管理研究*

□ 李成文 王志庚 李春明 周晨 曲云鹏 / 中国国家图书馆 北京 100081

摘要：网络信息资源生命周期管理研究以网络信息资源保存为基础，研究生命周期理论的规范性方法和最佳实践。文章在分析总结国外研究现状的基础上，从生命周期模型、生命周期管理成本控制以及生命周期过程控制三个方面分析了网络信息资源生命周期管理的相关理论和方法，为后续研究工作作铺垫。该文为2009年第七期“网络信息资源保存”专题文章之一。

关键词：网络信息资源，生命周期管理
DOI：10.3772/j.issn.1673—2286.2009.07.007

1 前言

信息是一种有生命周期的资源，一般由资源的生产、采集、传递、处理、存储、传播与利用等阶段组成[1]。国内外对于传统数字资源的生命周期管理起步较早，研究也深入而具体。网络信息资源作为数字资源的一个重要组成部分，与传统数字资源有一致性的方面，在生命周期管理中有许多可以借鉴的地方。但同时网络信息资源也有着自己的特点，需要在管理中有针对性的策略。

网络信息资源产生于网络，具有广泛性和共享性的特点，信息变化快、稳定性差。在信息的内容上常常有交叉重复的现象，信息质量往往良莠不齐。网络信息资源的长期保存是数字资源长期保存工作的一个重要方面，从90年代中期发展至今已有多个国家图书馆及组织参与其中。网络信息资源的获取、管理、利用是一个复杂的系统工程，研究网络信息资源生命周期管理不仅对于如何规范资源获取、管理以及更好的利用有着重要的意义，同时可以有效的解决网络信息资源长期保存过程中面临的三个主要问题：

- 一是信息资源的脆弱性。网络信息资源从产生开始的整个生命周期内，都容易受到环境的变化（存储介质的改变、科学技术的发展）而带来的影响（如数据迁移、数据格式变化），从而导致资源不可用；
- 二是网络信息资源生命周期的任一阶段内的活动，都会对后续的阶段产生重要影响；
- 三是为了保持网络信息资源可用性这一目的，必须保证资源在整个生命周期内数据的真实性和完整性。

网络信息资源生命周期管理研究就是研究如何对各个阶段的工作进行科学、有效、规范的管理，从而保证资源的持续利用以及不断的增值。

2 信息资源生命周期管理研究的发展与现状

国外关于信息资源生命周期管理的研究自20世纪90年代中后期开始。如1999年Hodge和Garroll完成的“把信息生命周期的理论用于数字存档的最佳实践”[2]，研究了19个典型的数字存档项目，其中涉及数据中心、第三方存储机构、出版者以及法定存储机构（国家图书馆和国家档案馆），总结出最佳实践并给出一个较为详细的生命周

* 本文系国家社会科学基金项目“网络信息资源保存的理论与方法研究”（项目编号：06BTQ025）的研究成果之一。

期阶段划分模型；加州大学伯克利分校的数字图书馆项目“重造学术信息的传播与使用：发展用于改善学术信息生命周期模型的工具与技术”[3]；英国利兹大学的研究项目（1998-2002）“数字馆藏管理”，该项目主要研究数字资源的保存问题，其中涉及了信息生命周期阶段的划分、纸质文献与数字资源生命周期的比较、数字馆藏管理的相关因素等问题[4]；大英图书馆Helen Shenton 在《生命周期馆藏管理》一文中提出信息生命周期包括选择、获取、编目著录、预保存、存贮、检索等过程；Maureen Pennock在“Digital Curation: A Life-Cycle Approach to Managing and Preserving Usable Digital Information”一文中给出一个简单的生命周期模型：产生、使用、评估和筛选、传递、存储和保存、访问和重用；DINI certification program for German institutional repositories研究项目中着重提出了质量控制、数据鉴定在信息资源生命周期中的意义。

除了理论上的研究，还有一些很好的实践项目。如：1996年澳大利亚国家图书馆开发的PANDORA项目[5]，服务于网络信息资源归档。该项目不仅仅关注网络信息资源的采集，同时包含后续一系列的诸如资源管理、获取、保存、访问、发布等；2002年美国国会图书馆开始创建数字资源生命周期框架（Digital Life Cycle Framework），提供了一套架构用于理解并服务于资源获取及其相关的活动、政策、最佳实践和数字资源管理。

随着网络信息资源保存的发展，本研究受到愈来愈多的重视，很多机构和项目开展了相当的研究和实践，主要体现在三个方面：①生命周期模型研究；②生命周期管理的成本控制；③生命周期管理的过程控制。

本文选择了几个典型的项目进行分析研究。

3 信息资源生命周期模型研究

3.1 Marchand和Horton信息生命周期管理模型

生命周期模型是生命周期管理的基础，根据资源形式的不同（音频、视频、网络等）采用适合的模型与指导方针对整个生命周期管理有着至关重要的意义。

美国著名的信息资源管理学家Marchand 和Horton早在1986就提出“信息生命周期管理”的概念，认为信息管理逻辑上存在相关联的若干阶段或步骤，每一步都与上一步的行为紧密相关。如图1所示信息生命周期管理模型的管理阶段包括信息的创建、采集、组织、开发、利用、清理六个部分，其中几乎每个阶段都可能使信息得到增值。

- 创建阶段是信息生命周期的初始阶段，在网络环境下，信息的发布具有很大的自由度和随意性，如果创建人忽略了后续各阶段的重要性，数字信息就可能无法被充分利用而丢失。因此在信息的创建阶段必须注意保持文档格式、规范以及元数据描述的一致性[6]。

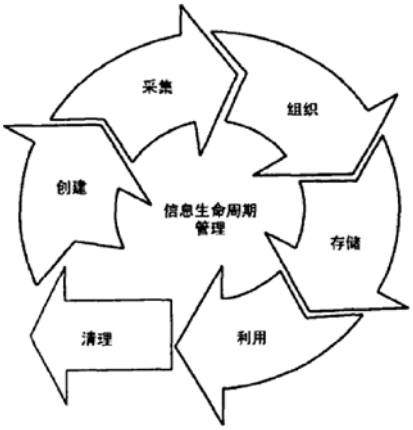

图1 Marchand和Horton信息生命周期管理模型

- 信息采集阶段是将分散的信息采集、积聚起来的过程，这个阶段的重要工作是制定采集政策，对所采内容的范围进行限制和规范。
- 由于网络信息资源往往未经过加工和审核，缺乏可靠性，因此需要在信息组织阶段采取一定的方式将大量分散的、杂乱的信息进行标引和著录使之变得系统、可用。
- 信息存储是实现信息价值的基础，该阶段的主要任务是依托网络存储应用技术，将存贮在网络中的信息从不可得状态变为可得状态，可得状态变为可用状态，低水平的使用状态变为高水平的使用状态的过程[7]。
- 信息利用是信息生命周期管理的最终目的，信息利用是用户对所提供的信息有效的运用的过程。
- 当信息随着时间的迁移逐渐失去利用价值时，进入信息清理阶段。这一阶段的主要工作是制定相应的策略，对失去价值的信息进行迁移或销毁。

以上模型根据信息运动的规律给出了基于生命周期进行管理的模型，科学的分析了信息管理各阶段的主要工作和问题，对信息管理具有重要的参考作用。

3.2 DCC生命周期管理模型

DCC数字资源生命周期模型[8]用图像的方式生动地展示了从最初概念产生之时就需要的成功的数

据管理与保存的各个步骤。此模型可用于机构或联盟内部活动的规划以确保所有必要的活动都能按照正常程序实施。此模型还使得零散的功能得到图像解析、明确作用与责任并建立实施标准与技术的框架。此模型有助于明确特定情况下必要和不必要的行为并确保政策充分落实。

从图2可以看出数据是生命周期管理的核心，包括数字对象和数据库。以数据为中心向外扩展的几个层次展示了完整的生命周期行为，包括描述与呈现信息，即为长期保存做好元数据描述与控制；保存规划，包括所有生命周期管理行为的管理计划；行业监测与参与，即保持对适当行业行为的监测，并参与共享标准、工具与软件的研发；管理与保存，即在生命周期管理的整个过程中，保持敏感并实施旨在促进管理与保存的管理行为。

图2的最外层描述了信息资源生命周期的各个阶段，包括创建、评估、转移、保存、存储、利用、迁移、销毁。

DCC的生命周期模型详尽的描述数字资源从产生之初到销毁所经历的各个阶段，为数字资源的科学管理提供了理论基础。

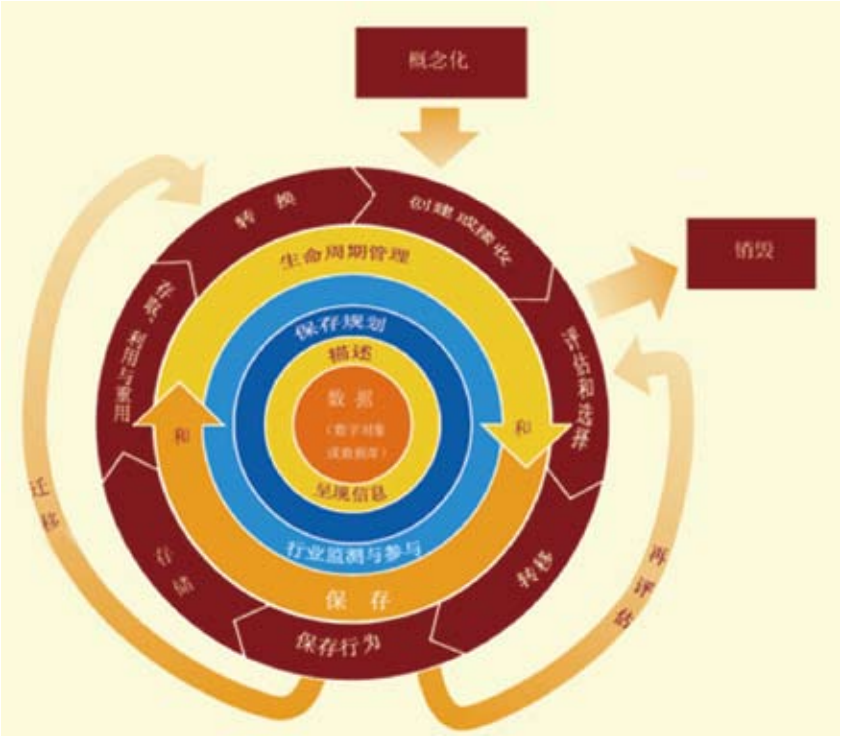

图2 DCC生命周期管理模型

4 生命周期管理的成本控制

基于数字资源生命周期模型的管理成本控制是生命周期管理研究中的一个重要方面。成本控制就是依据所选定的生命周期模型，采用科学的分析方法，分析各个管理环节对整个生命周期管理成本的影响，找出薄弱环节制定解决方案或开发新的管理工具，从而有效的控制成本，在有限的资金和技术的前提下提高管理效率和质量。

电子文献生命周期信息（Life Cycle Information for E-Literature, 简称LIFE）[9]是一个由伦敦大学学院（University College London）和英国图书馆合作的项目。该项目的目标是为数字资源管理的费用计算和管理传递重要信息，这些信息随后可以提供给任何一个对于数字资源保存和存取感兴趣的机构，为它们的活动提供帮助。

项目的研究成果将包括一系列具有实际意义的指导方针以及一个应用框架。利用这个框架可以解决数字资源管理费用计算问题，并回答如下问题：

- 什么是数字资源长期保存的成本；
- 由谁来做成本计算工作；
- 相同的出版物，纸质文献和电子文献长期保存成本的比较；
- 与纸质文献存档相对应的数字存档有哪些相关的风险。

LIFE项目尝试开发一种基于生命周期的方法来解决上述问题。项目组选取伦敦大学学院的电子杂志、英国图书馆的网络存档资源和英国图书馆VDEP数字资源为研究案例，解决不同文献类型长期保存过程中的成本计算问题。

LIFE的成本计算公式如下：

$$L(T)=Aq+I(T)+M(T)+Ac(T)+S(T)+P(T)$$

其中L是整个生命周期内从时间0到T的成本总和。其它各项含义为：资源获取（Aq）、摄入（I）、元数据（M）、存取（Ac）、存储（S）、保存（P）。

上述公式试图提供一种通用性的数字资源长期保存的成本计算方法，同时公式中的每一项又可以细分为更小的子项，从而适用于一些

特定的生命周期管理模型。

由上述公式可以看出，整个生命周期主要分成6个部分，每个部分可以继续细化，LIFE中定义的子项目并非是最全面的，也并非是全部必须采用的，具体项目中可以按照自己的需求进行增减。

生命周期模型的设计初衷是适用于所有的数字馆藏资源。其中定义的各个生命周期阶段并不一定是必须的，但它提供了一个适应于大多数关于数字资源管理方面的框架。

LIFE提供了一个功能强大、简洁易用的数字资源管理成本计算模型。利用这一模型进行成本计算，输出结果与使用的数据量大小成正比，也就是说，数据集合中的数据越多，这一模型的计算结果就越精确。

5 基于网络信息资源生命周期的过程控制研究

网络信息资源生命周期过程研究以生命的周期模型为基础，对组成生命周期各阶段进行研究。从资源的采集到加工、保存、访问利用，科学的划分子过程，并对流程进行严格的管理和质量控制从而确保信息资源生命周期管理的顺利完成。本文将讨论各个阶段涉及到的主要工作和最佳实践。

5.1 采集

采集与获取是生命周期中的第一个阶段。当数字对象产生以后，成为物理或虚拟上的一个整体，此时，开始进行采集进而归档保存。这一阶段主要需要考虑两个方面的问题：采集策略与采集程序。

在大多数国家，传统出版物与电子出版物在采集策略上的最大差别在于数字资源是否被纳入当前法定存档范围。在这种没有统一标准的情况下，制定合理的指导方针可以用来帮助界定采集的种类及范围。

加拿大国家图书馆的策略强调选取策略需要保证待存档资源的具有持久的文化价值和研究价值[10]。”澳大利亚国家图书馆的PANDORA（Preserving and Access Networked Document Resources of Australia）项目的选择标准是只保存澳大利亚的网络出版物[11]。

知识产权问题在网络资源获取过程中仍然是一个关键性的问题。资源存档组织类型的不同导致知识产权问题解决的方法也有较大的差异。在许多国家里，针对网络资源法律上还没出台相关的法规，各图书馆必须由自己来决定采集策略。PANDORA项目在进行资源存档前会试图寻求版权拥有者的许可。相比之下，瑞典和芬兰国家图书馆项目并不联系所有者而直接进行自动归档。

5.2 唯一标识和编目

编目、生成唯一标识是网络信息资源生命周期中的第二个阶段。编目和唯一标识都是保证信息归档者随着时间的推移仍然能管理资源的重要因素。唯一标识为每个数据对象提供唯一标识符，提供数据检索及链接到其它对象的需要。编目，利用元数据支持信息资源的组织、访问以及数据管理。编目和唯一标识的实行通常与正在归档的内容及能取得的用于管理存档的资源密切相关。

在所研究的项目中，涉及到多种元数据类型。澳大利亚国家图书馆的PANDORA档案完全采用了MARC编目方法。EVA则采用了Dublin Core元数据格式。Dublin Core元数据格式可以很灵活的直接从发布者获取元数据，剔除了图书馆编目的扩展项，简便易用。

关于永久标识问题，大多数档案仍然使用URL作为标识符定位数字对象。但也有一些项目在这方面进行了改变。OCLC档案使用PURLs(http://purl.oclc.org)，即将容易变化的URL映射为永久性标识符。美国化学学会使用数字对象标识符（Digital Object Identifier，http://www.doi.org）作为期刊论文的永久性标识符，同时也保留了原始文件出版发行时所分配的编号。美国国防情报资料中心使用由CNRI开发的Handle系统（http://www.handle.net） 生成永久性标识符进行标识。

5.3 存储

存储通常被认为是整个生命周期中的一个被动过程，但是存储介质和存储格式的变化可能会导致存档资源的永久丢失。磁盘容量、磁带容量、磁带机的驱动机制以及操作系统都会随着时间而发生变化。大多数资源存档组织认为存储介质更新的周期是3－5年。

对于这一问题最常用的解决方法就是数据迁移，将数据从原存储介质迁移到新的存储系统中。数据迁移需要很高的代价，并且迁移过程中时刻要考虑数据的完整性，防止数据丢失。数据一致性检测算法在数据迁移中起着极端重要的作用。

数据中心在存储介质迁移方面

有很丰富的经验。橡树岭国家实验室（Oak Ridge National Laboratory）的大气辐射监测中心（Atmospheric Radiation Monitoring Center）计划每隔4－5年进行一次存储技术更新。每一次的数据迁移都将数据复制到新工艺下的存储介质上，每次数据迁移工作将持续6－12个月。

5.4 长期保存

长期保存是存档管理的一个重要方面，不但要保存资源的内容同时需要保持数字对象的原貌（look and feel）。当前如何定义长期保存的时间没有完全一致的看法，时间限可以看作是技术和用户群体变革的时间。由于使用的技术的不同以及存档所涉及的学科不同，项目管理者估计随着软件、硬件的变更周期，数据迁移周期在2-10年。

数据库软件、表格处理软件及文档处理软件每隔两到三年就会发布一个新版本，在这期间还会有软件小的升级或打补丁等。虽然软件厂商的产品一般都会有自己相应的迁移策略并具有向下兼容性，但并不能保证两代或两代以上产品都能很好的做到这一点。为了防范软硬件迁移过程中的主要问题，组织一般都会尽量选择主流的商业技术。例如，美国化学学会和美国环境保护局都选择使用Oracle的产品，原因不仅是它的优秀数据管理能力同时也是由于公司长久发展的能力以及在标准规范发展过程中的影响能力。

就目前而言，还没有一个系统可以提供可实际使用方法实现上述技术，尤其是如何使存档对象能在旧的技术平台上使用。最重要的是并没有相应的法律法规要求制造厂商支持仿真信息。在可预见的未来中，长期保存过程中向新的软件硬件迁移仍然是最可行的方法。

5.5 存取

网络信息资源生命周期的前几个阶段都是为保证存档资源能被持续的访问。成功的实践必须要考虑到在未来相当长一段内存取机构的变化、权限管理及数据安全需求。

大多数项目的管理者认为访问与展示机制是数字环境下另一个变化源。现在是通过Web形式进行资源访问，但是将来通过什么样的方式却很难确定。随着未来数字化技术和浏览技术的提高，存档资源的展示质量也会相应提高。美国国家医学图书馆（NLM）的Profiles in Science产品创建了一个基于图片、文本、视频等格式的电子档案馆。这一电子档案馆能适应未来访问机制的变化。建立存档格式的同时也保留了原始的版本。该项目的管理者Alexa McCray认为，技术的革新已经表明，当进行存档格式的数据转换时，无论当初保存了多少的细节信息，未来都会显出不足之处。随着时间的变化，新的软件和硬件会使得资源的获取和展示有更高的质量，但前提是我们一定要保存好资源的原始类型。

权限控制问题是数字存档访问中最难解决的一个问题。为档案赋予什么样的权限，用户群应该赋予什么样的访问权限，存档资源的拥有者应该有什么样的权限，存取机制与档案资源元数据应该如何交互以保证各种权限的管理。权限管理包括正确的分配或限制访问并随着版本和安全级别的变化而能做出相应的改变。

6 结语

网络信息资源生命周期管理基于网络信息资源的保存研究与实践。随着网络信息资源保存工作的进一步推进，目前具备了生命周期管理研究的理论基础，在生命周期模型、成本控制、过程控制方面取得了一定成果，但是实践中仍然面临一定的问题。网络信息资源的特点是类型丰富、变化快、复杂度高，网络环境也在不断的发展变化，这些因素都给生命周期管理的研究带来了一定的影响，未来还需要在如下方面进行更深入的研究和实践：

（1）制定一个统一的生命周期模型。生命周期模型是整个管理工作的基础，目前对生命周期各阶段的划分理论大致相同，但也存在一定的差别，DCC的生命周期模型也缺少一定的实践活动进行验证和修正，因此无论在理论上还是实践中，寻找最佳的网络信息生命周期模型是关键工作。

（2）进行针对性的研究和实践活动。目前的研究成果和最佳实践大多基于信息资源，缺少针对网络信息资源的实践活动。网络信息资源是信息资源的一种，但由于来源于网络，其生产者的不同对信息质量带来重大影响，因此从资源类型出发，在吸收信息资源生命周期管理研究的基础上总结更多的针对网络信息资源的研究理论与实践十分必要。

（3）标准规范工作有待加强。网络信息资源生命周期过程控制是保证资源保存和永久访问的重要环节，而控制工作有赖于各个环节标准的工作流程和采用的标准协

议、规范方法等。目前各界的实践活动中，对标准有一定的认识，但尚未形成统一思想。

（4）生命周期管理成本控制研究需要进一步完善。LIFE项目提供了完整的模型进行生命周期管理的成本计算，选取了一些资源保存项目进行实践验证。但由于网络信息资源保存的复杂性，成本模型和更为精确的计算理论仍然需要进一步完善。

希望本文的研究成果能为国内网络信息资源的长期保存提供有益借鉴。

参考文献

[1] MARCHAND D A, HORTON F W, Jr. INFOTRENDS: Profiting from You Information Resources. 1986.20-25.
[2] HODGE G M. Best Practices for Digital Archiving:An Information Life Cycle Approach[J]. D-Lib Magazine, 2000(1).
[3] UC Berkeley Digital Library Project: Re-inventing Scholarly Information Dissemination and Use[EB/Ol]. [2009-05-01]. http://elib.cs.berkeley.edu.
[4] The Cedars Guide to: Digital collection Management[EB/Ol]. [2009-05-01]. http://www.leeds.ac.uk/cedars/guideto/collmanagement/ arch 2002.
[5] [EB/Ol]. [2009-05-01]. http://www.nla.gov.au/policy/plan/pandora.html.
[6] SHAW K A, HICKOK G J. Life cycle information management:A case study[J]. Information Management Joural,2000,(4) :24-36.
[7] 罗贤春.网络信息生命周期[J] .图书馆学研究,2004,(2):51-53.
[8] [EB/Ol]. [2009-05-01]. http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf.
[9] [EB/Ol]. [2009-05-01]. http://www.life.ac.uk/.
[10] National Library of Canada, Electronic Collections Coordinating Group. Networked Electronic Publications Policy And Guidelines, October 1998.
[11] National Library of Australia. Selection of Online Australian Publications Intended for Preservation by the National Library of Australia[EB/Ol]. [2009-05-01]. http://www.nla.gov.au/scoap/guidelines.html.

作者简介

李成文（1981-），重点研究网络信息采集在图书馆的应用。通讯地址：中关村南大街33号国家图书馆北412室 100081。E-mail: lichwen@nlc.gov.cn

Research on Resource Lifecycle Management Based On Web Archive

Li Chengwen, Wang Zhigeng, Li Chunming, Zhou Chen / National Library of China, Beijing, 100081

Abstract: The research on resource lifecycle management based on the web archive, aiming to find out the criterial method and the best practice of lifecycle theory. After summing up the situation of current research of other countries, this article analyzes the proper theory and method from the aspect of lifecycle model, control of the cost of lifecycle management and the control of lifecycle process, which will benefit to the following research.

Keywords: Web archiving, Lifecycle management

（收稿日期：2009-05-15；责任编辑：贾延霞）