高校网络信息资源自动化处理与长期保存策略研究*

夏立新 杨元 郭致怡
(华中师范大学信息管理学院，武汉 430079)

摘要：本文在全面梳理国内外网络信息资源自动化处理与长期保存现状和高校文献信息资源保障现状的基础上，指出高校网络信息资源自动化处理与长期保存的必要性，并对高校网络信息资源自动化处理与长期保存最佳实践案例密歇根大学本特利历史图书馆网页归档项目进行深入分析，据此以完善高校文献信息资源体系为导向，从资源、技术和管理三个维度，制定我国高校网络信息资源自动化处理与长期保存策略。

关键词：网页归档；网络信息资源；文献信息资源保障；高校图书馆
中图分类号：G250 DOI：10.3772/j.issn.1673-2286.2021.09.001
引文格式：夏立新，杨元，郭致怡. 高校网络信息资源自动化处理与长期保存策略研究[J]. 数字图书馆论坛，2021 (9)：2-10.

文献信息资源作为一种社会智力资源，是人类活动与知识的载体。随着互联网的发展与普及，网页已经逐渐成为人们日常获取、记录信息的重要平台，网页所载文字、图片、音像等成为记录和反映当代社会人类活动与知识的重要信息资源。从文献信息资源保障的视角看，以一定采集策略筛选获得的网络信息资源是当代社会新兴的一类文献信息资源。与传统文献信息资源相比，网络信息资源的内容与形式更加丰富、体量庞大。然而，网络信息资源的易变性、不稳定性、流动性、不可再生性以及对软硬件环境的强依赖性，也为网络信息资源的长期保存与开发利用带来较大挑战[1]。如何确保网络信息资源的可靠使用和永续利用是互联网时代文献信息资源保障工作亟需解决的问题。

我国网页归档实践尚处于起步阶段，对网络信息资源的自动化处理与长期保存的理论研究和实践经验十分有限，当前仅有中国国家图书馆和北京大学较系统地开展了相关实践。本文面向拥有大量活跃互联网用户和丰富内容与形式的高校网络信息资源，通过分析国外高校网络信息资源自动化处理与长期保存的优秀案例，总结其网络信息资源管理策略，为我国网络信息资源建档、归档工作提供方法与路径。

1 相关研究

本文回顾已有研究成果，发现当前高校网络信息资源自动化处理与长期保存策略相关研究主要围绕网络信息资源自动化处理与长期保存现状和高校文献资源保障两方面展开。

1.1 网络信息资源自动化处理与长期保存现状

国际互联网保存联盟（IIPC）将网页归档定义为采集万维网的一部分内容并且以档案形式保存，并支持档案的后续访问和使用[2]。其中，网络信息资源的自动化处理与长期保存的技术与策略是网页档案建设的主要内容，经过自动化处理并进行长期保存的网络信息资源集合称为网页档案资源。在实践方面，美国互联

* 本研究得到国家社会科学基金重大项目“新时代我国文献信息资源保障体系重构研究”（编号：19ZDA345）资助。

网档案馆于1996年率先开展网页归档相关实践。同年，澳大利亚、瑞典、法国也相继于20世纪末展开网页归档实践[3]。目前，欧美国家的网络信息资源归档主题已不断细分，涵盖国家历史文化、社会生活、突发事件、政府信息等。随着项目的推广和深入，国外学者在利用网页档案的过程中产生了多学科、多主题的研究成果，同时不断提出网络信息资源的新需求与相应的保障策略[4-7]。我国的网页归档实践开始于21世纪初，2001年北京大学计算机系网络与分布式系统实验室发起了“中国Web信息博物馆”项目，该项目能够采集我国绝大多数的静态网页并提供网页搜索和数据分享功能[8]。中国国家图书馆于2003年发起了“中国国家图书馆的网络信息资源保存试验项目”（Web Information Collection and Preservation，WICP）对中国境内的网络资源进行采集与保存实验，并于2019年启动“互联网信息战略保存项目”，建设覆盖全国的分级分布式中文互联网信息资源采集与保存体系[9]。然而国家图书馆的互联网资源尚处于建设阶段，还未开展相关服务。在网页归档的理论研究方面，我国网页归档的有限实践导致我国网络信息资源保障工作的相关研究大多集中于理论研究，学者对网络信息资源的采集、保存相关的技术与策略进行研究，而利用网页档案开展的研究成果较少，我国历史网络信息资源尚未得到有效地开发与利用。我国网页归档实践尚处于起步阶段，且现行网页归档项目较少且尚未面向社会进行网页档案资源保障工作。

1.2　我国高校文献信息资源保障现状

肖希明[10]提出，文献信息资源保障工作的总目标是最大限度地满足用户对文献信息最广泛的需求。刘敏等[11]提出高校图书馆应为高校教学、科研提供“纸质文献信息—电子文献信息—共享文献信息”的全方位服务。蒋岩波等[12]以江西省昌北高校图书馆联盟为例，认为图书馆联盟中各高校应当注重资源采购计划的针对性，凸显本校学科特色；完善重点学科三级文献资源保障体系建设。现有研究多以保障高校内部用户的文献信息需求为目标展开，忽略了高校图书馆在我国文献信息资源保障工作开展中承担的使命与责任，导致高校文献信息资源建设模式相对封闭，主要侧重通过常规采集方式获得的各种文献信息资源，满足高校用户教学、科研的文献信息需求，忽略了记录以及反映高校知识和历史的网络信息资源的采集保存工作，从而严重制约了此类重要资源的开发利用。

综上所述，网络信息资源已成为互联网时代记录和反映人类生产生活的重要文献资源之一，然而我国的网络信息资源的自动化处理和长期保存工作尚处于起步阶段。一方面，我国开展网页归档实践的组织与机构较少，且尚未正式对公众开展网页档案资源服务；另一方面，作为我国文献信息资源保障工作的重要基础性机构的高校图书馆对于网络信息资源的长期保存意识较为淡薄，忽略了网络信息资源的文献价值。基于以上问题，本文借鉴国外优秀高校网络信息资源归档项目——美国密歇根大学本特利历史图书馆（以下简称“本特利历史图书馆”）的网页归档实践，深度剖析该项目的实践情况，在总结其归档资源特征、自动化处理和长期保存具体工作流程的基础上，研究适合我国高校网络信息资源自动化处理与长期保存的策略。

2　文献信息资源保障视角下高校网页归档的必要性分析

互联网记录和传输信息的便捷性，使得人类越来越多地将信息记录和分享在各类网站以及互联网平台上。从文献信息资源保障的角度看，高校网络信息资源已然成为能够广泛、多形式记录和反映高校知识成果和历史发展的重要文献信息资源之一，然而，网络信息资源易丢失、难保存的特点又为其可靠使用和永续利用带来严峻挑战[8]。此外，高校作为我国文献信息资源保障工作开展所依托的重要机构，其网页归档实践对我国网页归档事业发展具有重要参考价值。因此，高校网页归档实践是当前文献信息资源保障工作所面临的重大挑战，又是互联网时代文献信息资源保障的必要工作。

2.1　高校网络信息资源是反映高校发展历史相关信息与知识的文献信息资源

互联网时代，网络信息资源是记录人类活动和知识的重要文献信息资源之一，实现对网络信息资源的长期保存是文献信息资源保障工作的重要内容之一[8]。高校网络信息资源的内容和形式丰富多样，是记录和反映高校发展历程、管理制度、科研成果、学术活动、学生生活、校园文化等多方面历史信息的第一手资料，是互联网时代记录高校管理制度和发展历史的重要原始文

献。因此，探索网页信息的自动化处理与长期保存策略对保留高校历史资料和知识产出具有重要意义[12]。

2.2 网络信息资源的自动化处理与长期保存能够丰富文献信息资源保障类型

文献信息资源保障工作的目标是最大限度地满足用户的文献信息需求。网络信息资源作为互联网时代新兴的文献信息资源类型，既是人类数字记忆的重要组成部分，又是教育学家、历史学家等研究者的重要参考文献[12]。一方面，随着Web2.0时代的到来，网络信息资源相比传统文献信息资源以文档、图片、视频等多种形式，更加全面翔实地记载了人类的知识与活动，弥补了传统文献信息资源的记录空白；另一方面，网络信息资源高度依赖于其所在的软硬件环境，易丢失且难恢复。因此，网络信息资源作为一种新时代我国文献信息资源采访与自建工作的重要对象，只有通过有效的自动化处理与长期保存工作及时识别重要、具有长期保存价值的网络信息资源并对其进行分类、归档和存储，才能保障网络信息资源的可靠使用和永续利用，促进立体化、多样态文献信息资源体系的形成。

2.3 高校网页信息归档实践能够为我国网络信息资源保障事业发展提供参考

高校网络信息资源归档实践较易开展且能够获得较丰富的实践经验。一方面，高校是一个管理体制较完善且拥有本校网络信息资源知识产权的组织，其网络信息资源归档实践面临的外界阻碍较小；另一方面，高校网络信息资源内容与形式丰富，拥有其不同部门及附属单位，甚至由教职工和学生创建、管理，服务于学校各项业务或高校成员业余生活的各类网站，其网络信息资源归档实践能够为不同主题、不同信息资源类型、不同更新频率、不同运营机构特征的网络信息资源的自动化处理与长期保存工作提供经验和参考。因此，研究和开展高校网络信息资源归档工作能够推进我国网络信息资源保障事业发展。

3 国外高校网页归档最佳实践——以密歇根大学本特利历史图书馆为例

美国国家数字化管理联盟（National Digital Stewardship Alliance，NDSA）2016年和2017年的网页归档项目调查报告显示，近年来美国高校开展网页归档项目的机构数量明显增长，美国高校图书档案机构成为网页归档的重要实践单位[13]。本文以高校网页归档最佳实践案例本特利历史图书馆的网页归档项目为例，进行深入分析其网页归档实践中自动化处理与长期保存策略。

3.1 网页归档实践概况

密歇根大学所属本特利历史图书馆成立于1935年，其主要职能是收集并管理密歇根大学相关历史的第一手证据和数据并促进对它们的历史研究，以确保运营的连续性和有效的管理，履行法律、监管和财政责任，并优化其对空间和时间的利用。自2010年以来，本特利历史图书馆一直在通过网页归档实践来识别、评估和选择能够反映大学运营管理和具有档案收藏价值的网站并定期进行对这些网站进行自动化处理与保存。截至2021年6月1日，本特利历史图书馆已建立了9个网页档案，共归档2 803个网站[12]。这些网页档案向公众开放，用户可通过Archive-It官网、U-M Library（密歇根大学图书馆的在线公共访问目录库）或BHL Finding Aid 3个网站对其网页档案信息进行访问。同时，本特利历史图书馆还积极地与其他档案机构合作分享它的检索工具，以便公众和远程研究人员能够了解本特利的馆藏并加以利用[14]。

3.2 网页归档资源及资源特点

本特利历史图书馆于2019年修订的《档案政策与程序手册》中明确了其筛选归档网站所必须满足的5项条件：①网站由大学所有且用于开展大学相关业务；②网站反映与大学相关的基本功能或活动；③网站是对现有档案和手稿收藏的补充；④网站填补了收藏中的空白；⑤网站包含定期更新的独特且有意义的内容[12]。

目前，本特利历史图书馆已建立档案的内容包括密歇根州的历史收藏以及密歇根大学管理、校友和粉丝、体育、卫生系统、新闻与活动、附属单位（学校、学院、研究、中心和研究所、学生组织）、mBLog（移动博客）。从数量上看，密歇根大学附属单位的网络档案所包含的网站数量最多（1 283个），而密歇根大学校友和粉丝网络档案所包含网站数目最少（19个）。从内容上看，本特利历史图书馆进行归档的网页包含八大主题，

分别为大学与图书馆、社会与文化、艺术与人文、博客和社交媒体、科学与健康、自发事件、计算机与技术和政府－美国各州。此外，本特利历史图书馆所收藏网站既包含密歇根大学附属的学院、研究机构和学生组织，也包含学校的教职工和学生合作或独立创建的网站[15]。通过对归档网站的筛选原则和密歇根大学已归档网络信息资源的调研分析，本文发现密歇根大学的网页归档资源具有以下特点。

（1）网站为密歇根大学所有。本特利历史图书馆所归档网站均由密歇根大学附属单位、教职员工或学生创建、拥有或使用。此类网站所记录的信息资源，不仅在内容上与密歇根大学密切相关，具有一定保存价值，而且其知识产权归密歇根大学所有，合理规避了潜在知识产权纠纷。

（2）网站服务于该校的各项工作与活动且能够反映其开展情况。本特利历史图书馆要求网站用于大学且能够反映有关的业务、功能或活动。此类网站能够从不同视角广泛、形象地记录和反映密歇根大学的管理事务、校园活动和发展历程，是密歇根大学的重要历史遗产，可以帮助解释事件发生的方式或原因，为历史学家、教育学家、新闻工作者等提供优质信息源。

（3）网站中网页信息具有永久且持续记录价值的内容。本特利历史图书馆要求所归档网页信息能够填补已有收藏的空白并会定期进行更新。考虑到长期保存的成本问题，本特利历史图书馆馆员在进行网页归档前会对网站的信息资源内容进行评估，删除明显重复和历史价值不足的网站。

3.3 网页归档的工作流程

网页归档涉及网络信息资源的采集、归档、编目、存储4个关键步骤。本特利历史图书馆负责确认采集对象、规范网站建设、提供访问接口和管理知识产权等问题，同时通过Internet Archive推出的Archive-It程序，进行网络信息资源的收集、归档和保存工作。

本特利历史图书馆网页归档的具体工作流程如图1所示，依托Archive-It程序并制定辅助Archive-It顺利开展网站识别、网络信息资源爬取和编目工作的相关制度与规范，实现网络信息资源的自动化处理与长期保存。

本特利历史图书馆网页归档过程中的工作分为三个阶段：一是在采集网络信息资源前，制定便于Archive-It程序进行自动化处理的网站建设规范，并提供便于网站预归档名单、网站信息资源采集方案，如对各网站信息资源采集的时间和频率；二是在采集网络信息资源过程中，选用Archive-It程序对网络信息资源定期进行自动化

图1　本特利历史图书馆网页归档工作流程

识别和采集，与Archive-It相关负责人员及时沟通，根据实际情况调整网页归档计划；三是在完成网络信息资源采集工作之后，继续利用Archive-It程序对网络信息资源进行编目、归档和存储。

Archive-It在密歇根大学网页归档实践中主要负责在特定的时间点获取所需归档网站的快照并创建网站的存档副本，将副本文件规范化存储于Internet Archive的互联网数据库和密歇根大学的机构资料库Deep Blue中，实现多服务器网络信息资源存储与服务。具体而言，Archive-It的工作分为网络信息资源采集和网络信息资源编目与存储两个阶段：在网络信息资源采集过程中，Archive-It在不干扰网站访问的情况下通过爬虫软件进行网络信息资源的爬取。大多数网络信息资源的爬取工作每年仅运行几次并持续几天，在网络信息资源采集工作完成后，爬虫软件将停止与服务器进行交互。此外，Archive-It会对密歇根大学预归档名单中的网站进行持续性跟踪和监测，提供网站最佳采集时间的建议，当网站发生重大变更时，提醒档案管理员增加临时采集计划。在完成网络信息资源采集工作后，Archive-It采用柏林核心元数据和网络资源存档国际标准WARC格式对网络信息资源进行规范化著录和保存。

3.4 网页归档过程中的自动化处理与长期保存策略

通过对密歇根大学网络归档工作流程的梳理，可以总结出本特利历史图书馆网页归档工作各环节采取的关键策略如下。

（1）选择性网络信息资源采集。本特利历史图书馆根据网页归档资源的筛选原则，预先确定需要采集的网站，忽略保存价值低的网络信息资源。对于高校网页归档工作而言，受到成本、法律等多方面外界因素限制，采取选择性采集策略有以下优势：一是通过人工的预先筛选能够保证所采集网络信息资源的内容质量；二是大大缩小网络信息资源采集的范围，能够降低网络信息资源采集的技术、设备成本，也有利于网络信息资源的知识产权合规管理，有效规避高校网页档案资源开放的法律风险。

（2）规范化的网站建设。为提升本校网站的可访问性，方便网络信息资源的自动化识别和归档，本特利历史图书馆发布了《网站可访问性指南》对密歇根大学的各网站创建出规范化要求，具体包括：①所有大学网站和印刷材料上均应包含版权行；②要确保网页有效且符合HTML规范；③在网站站点的robots.txt文件开头添加规定代码，明确允许Archive-It对站点进行归档；④要求网站在HTML标头中使用描述性元数据元素来提供有关网站的文档。

（3）定期捕获和及时捕获相结合的网络信息资源采集。本特利历史图书馆在利用Archive-It程序进行网页归档的自动化处理过程中，本特利历史图书馆根据需要归档网站内容的一般变化情况，确定各网站的采集时间和频率，将网站地址及其捕获频率提供给Archive-It。Archive-It按照本特利历史图书馆设定好的捕获频率，定期对本特利选定的网站进行数据爬取，创建网页存档副本并进行存储。此外，当密歇根大学对其网站进行临时性重大更改时，本特利历史图书馆可以在Archive-It人工添加新的捕获计划。

（4）国际标准化的网络信息资源编目与存储。本特利历史图书馆网页归档资源的编目和存储也在Archive-It程序的辅助下进行，在网络信息资源的存储上，采用WARC格式（网络资源存档国际标准ISO 28500：2009）进行网页数据的存储。在对所采集网络信息资源的描述上，采用国际上广泛使用的柏林核心元数据集对网络信息资源的文件类型、标题、内容、URL、主题及发布者等进行描述与著录。此外，Internet Archive还开发了一种能够从WARC文件中抽取结构化数据的方法WAT（Web Archive Transformation），便于对大规模数据集进行数据分析。采用国际通用的数据描述标准和存储范式，有助于所收集网络信息资源的整合共享、高效使用、二次开发和永续保存。

（5）多副本多服务器的网络信息资源存储。本特利历史图书馆的网页档案不仅保存在Internet Archive的互联网数据库中，还备份存储在密歇根大学的机构资料库Deep Blue中，支持通过Archive-It官网或密歇根大学的数字图书馆扩展服务访问其网页档案。这种保存策略不仅能够增强网络信息资源存储的安全性，而且能够支持该校网页档案的多途径访问和利用，更好地满足校内和社会用户的相关文献信息需求。

3.5 网页归档所提供服务与应用现状

本特利历史图书馆提供多渠道网页档案资源检索服务、网页档案资源索引与指南服务。一方面，用户利用Archive-It的时光机项目（Wayback Machine）、本

特利历史图书馆检索工具BHL Finding Aid和密歇根大学图书馆的数字图书馆扩展服务进行网页档案的检索和访问，便利本校用户和公众对本特利历史图书馆网页档案的访问和利用。另一方面，本特利历史图书馆的档案管理员整理并提供了网页档案的描述性指南与索引，列出了网页存档信息资源和网页档案的名称、主题、摘要、创建者、采集日期等内容，方便用户确认自己所需的网络信息资源或按某一分类标准进行获取具有特定特征的网页存档信息资源和网页档案。

本特利历史图书馆的网页归档实践已初具规模，然而，其自动化处理和长期保存工作中依旧面临网站存档版本不完整的问题。具体而言，本特利历史图书馆的网页归档策略是针对html格式的静态网页，在对其他类型的网络信息资源进行采集和存储时，难以保留其完整形式、功能和内容，主要包括：①存储在不同域或子域上的链接内容；②动态脚本或应用程序，如JavaScript或Adobe Flash；③具有视频或音频内容的流媒体播放器；④受密码保护的材料；⑤需要与网站进行交互的表单或数据库驱动的内容[12]。

4　本特利历史图书馆网页归档实践对我国文献信息资源保障工作的启示

网络信息资源是互联网时代记录和反映人类生产生活的重要文献信息资源，是新时代我国文献信息资源保障体系建设中关键的组成部分。然而，我国网页归档工作尚处于起步阶段，尚未形成系统性的网络信息资源自动化处理和长期保存机制，学界、业界对历史网络信息资源潜在价值的二次开发与利用十分有限。由此，以习近平总书记“融合发展思想、开放发展理念”为指导，将网络信息资源纳入我国文献信息资源保障体系，建设便于社会各界获取与利用的网页档案尤为必要。鉴于我国在网页归档实践中存在的问题，本文提出优先进行高校网页归档实践，及时保留高校网站所记录和反映的高校知识与历史的网络信息资源，为我国网页归档事业提供经验和参考方案。

4.1　探索我国高校网络信息资源归档与保障模式

在借鉴国外最佳实践的基础上，要实现我国高校网络信息资源的自动化处理与长期保存，完善高校文献信息资源体系，确保高校重要网络信息资源的可靠使用和永续利用，就要建立适合我国高校发展特点的网络信息资源归档与保障模式。本文从文献信息资源保障工作的核心内容文献信息资源的建设和服务两方面出发，结合本特利历史图书馆的最佳实践经验，构建我国高校网络信息资源归档与保障模式，如图2所示。

（1）网络资源层。该层是服务于高校成员办公和日常活动的各类知识产权归本校所有的网站中所承载各类数据及资源的集合。高校官方网站、高校网页论坛和高校成员用于高校各类活动自建的网站所承载的高

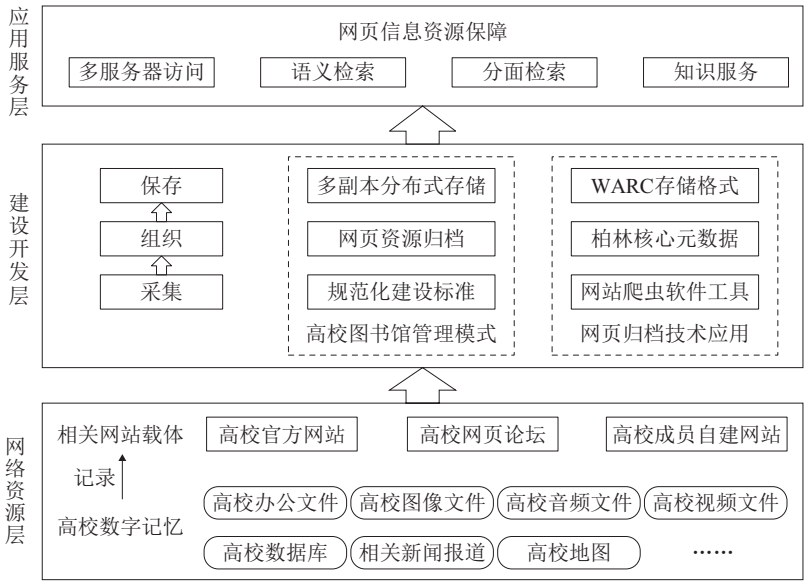

图2　高校网络信息资源归档与保障模式

校办公、图像、音频、视频的文件及高校数据库、相关新闻报道、高校地图等数字资源共同构成高校的数字记忆[16]。网络资源层的涵盖范围广泛，资源内容丰富，资源类型复杂且资源质量不一，需要高校图书馆员对网络资源层的海量资源进行筛选，保留具有长期保存价值的网络信息资源。

（2）建设开发层。该层主要包含高校网络信息资源的采集、组织、保存三方面的工作。本文从高校图书馆在建设开发中所需要采取的管理模式和网页归档所需的相关技术两方面出发，梳理高校网络信息资源建设开发过程各阶段的组织和技术保障需求。在网络信息资源的采集方面，需要图书馆为本校网站建设制定统一标准，便于网络信息资源的爬取和著录。同时，在此阶段，需要高校图书馆根据所需归档的网络信息资源特点，选取成本合适、能够可靠爬取网络信息资源的爬虫软件。网络信息资源的组织过程中，需要根据网络信息资源的内容特征进行归档，自动化处理软件应该按照统一标准进行编目。在网页档案资源保存阶段，图书馆需要确保网页档案资源保存的安全性，采取多副本分布式存储策略，网页档案资源的存储格式应与WARC相一致。

（3）应用服务层。该层主要包含实现高校网页档案资源有效可靠保障的各类服务。通过多服务器存储拓宽高校网页档案资源的服务对象范围，便于高校用户和社会用户对高校网页档案资源的开发利用。一方面，提供全文搜索和浏览列表等多种检索方式，将托管的网页档案集合直接链接到机构本地的搜索页面。同时，推出便于网页档案资源开发利用的相关知识服务，提供数据驱动研究方法，如网络分析、文本与数据挖掘、纵向内容分析等扩展用户访问和分析归档网页资源的方式[13]，确保高校网页档案资源的保障效果。

4.2 构建我国高校网络信息资源自动化处理与长期保存策略框架

我国高校网络信息资源的自动化处理与长期保存工作，不仅要探索适应我国高校当前发展水平和特点的网络信息资源归档与保障模式，更要构建能够长期有效指导的我国高校网络信息资源自动化处理与长期保存的策略框架，从而保证高校网络信息资源建设与服务能够适应时代发展，不断提升高校文献信息资源的保障水平。本文从文献信息资源保障的视角出发，通过分析本特利历史图书馆在高校网络信息资源的自动化处理与长期保存中运用的技术和管理策略，发现网页归档中的策略制定主要包括三方面内容，分别是网页归档的对象网络信息资源，网页归档所用的自动化处理与长期保存相关技术，以及网页归档所需要的组织管理。因此，本文从资源、技术和管理三个维度构建我国高校网页档案资源的自动化处理与长期保存策略框架（见图3）。

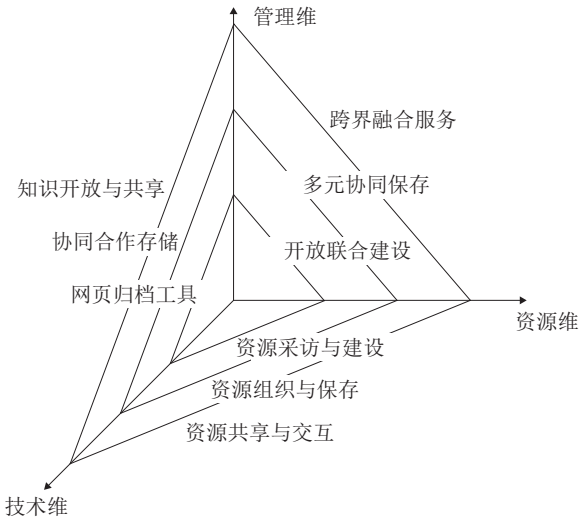

图3 高校网络档案信息自动化处理与长期保存策略框架

（1）基于开放联合的高校网络信息资源规划与采集。当前高校文献信息资源建设的边界仍需拓展，不仅要建设传统文献信息资源和数字出版物，还要关注网络信息资源的文献价值，将高校有关网站及各互联网平台上与高校相关的文献信息资源纳入高校文献信息资源建设的整体规划，进行网络信息资源采集与建设的探索和实践。具体而言，高校应主动承担起本校重要网网络信息资源的归档工作，对外主动联合先进的网络信息资源管理机构，吸收先进的网页归档自动化处理技术与方案；对内规范网站建设，筛选出具有长期保存价值的网站，主导各类网络信息资源的采集、网页档案自建工作，丰富高校文献信息资源保障类型，形成立体化、多样态高校文献信息资源体系。

（2）基于多元协同的高校网页档案资源组织与长期保存。在开放环境下，高校应积极开展高校间及社会各文献资源保障机构间的合作交流，促进各主体网页档案资源的共建共享，探索多主体协同合作的文献信息资源长期保存模式。具体而言，高校在进行网页归档实践时应该充分吸收社会各界的技术、经验，吸收先进的网络信息资源自动化处理和长期保存技术，确定统一

的网页档案资源编目标准和存储格式，一方面方便网络信息资源的整合、更新和共享；另一方面，便于网页档案资源的多副本分布式存储，增强网页档案资源的容灾性和安全性[17]，保证网页档案资源的可靠使用和永续利用。

（3）基于跨界融合的高校网页档案资源开放与共享。高校作为知识、发现和教育的中心以及公共资助的机构，应该积极承担文献信息资源建设、服务与创新性实践的使命与职责。在我国网页归档事业的发展进程中，率先开展实践，探索高校网络信息资源的自动化处理与长期保存方案，对内完善网站建设规范，采集、组织和保存有历史价值的网络信息资源，形成更加立体化、多样态的高校文献信息资源保障体系；对外提供开放接口，向社会研究人员与机构提供优质的信息源，促进高校网络信息资源的社会化开发。

5 结语

网络信息资源是互联网时代我国文献信息资源体系的重要组成部分，高校网页归档实践不仅完善了高校文献信息资源体系，为挖掘高校历史和开展相关研究提供了优质文献信息源；同时，还能够为我国网页归档事业提供参考。本文通过本特利历史图书馆的网页归档实践，总结其对网络信息资源自动化处理和长期保存的策略，从网络信息资源的规划采集、组织保存、开放共享三方面构建了我国高校网络信息资源自动化处理与长期保存策略的系统框架。

参考文献

[1] 魏大威，季士妍. 国家图书馆网络信息资源采集与保存平台关键技术实现 [J]. 图书馆，2021（3）：45-50.

[2] IIPC. WEB ARCHIVING [EB/OL]. [2021-08-24]. https://netpreserve.org/web-archiving/.

[3] 李子林，龙家庆. 欧洲网络存档项目实践进展与经验启示 [J]. 图书馆学研究，2020（15）：56-64.

[4] WEIKUM G，NTARMOS N，SPANIOL M，et al. Longitudinal analytics on web archive data：It's about time! [C] //CIDR. 2011：199-202.

[5] COSTA M H，GOMES D，SILVA M J. The evolution of web archiving [J]. International Journal on Digital Libraries，2017，18（3）：191-205.

[6] GOMES D，MIRANDA J，COSTA M. A survey on web archiving initiatives [C] // Research and Advanced Technology for Digital Libraries-International Conference on Theory and Practice of Digital Libraries，TPDL 2011，Berlin，Germany，September 26-28，2011.

[7] TOYODA M，KITSUREGAWA M. The History of Web Archiving [J]. Proceedings of the IEEE，2012，100：1441-1443.

[8] 王静. 中美网页归档项目的对比研究 [J]. 档案与建设，2015（7）：14，19-23.

[9] 国家图书馆将启动互联网信息战略保存项目 [EB/OL]. [2021-08-22]. http://www.xinhuanet.com/book/2019-04/12/c_1210106680.htm.

[10] 肖希明. 我国文献资源保障体系论纲 [J]. 图书馆，1996（6）：8-12.

[11] 刘敏，吕先竞，张建. 构建三级文献保障体系 全面服务高校教学科研——地方多科性大学文献保障服务理论与实践 [J]. 情报资料工作，2008（3）：39-42.

[12] 蒋岩波，生修雯. 地方高校重点学科区域性文献资源保障体系建设问题研究——以江西省昌北高校图书馆联盟为例 [J]. 图书情报工作，2015，59（11）：89-93.

[13] 张莉，颜祥林. 美国网页归档项目发展的新动向——基于NDSA2016年和2017年调查报告的分析 [J]. 档案与建设，2019（10）：33，39-42.

[14] 吴晓茹，陈丹.密歇根大学网页资源归档实践研究及启示 [J]. 档案管理，2020（6）：112-114.

[15] Arichive-it. University of Michigan Bentley Historical Library [EB/OL]. [2021-08-01]. https://archive-it.org/organizations/934.

[16] 陈慧，乐茜，罗慧玉，等. 社会记忆视角下网络信息资源归档路径探析——以PANDORA项目为例 [J]. 数字图书馆论坛，2020（6）：15-21.

[17] 王芳，史海燕. 国外Web Archive研究与实践进展 [J]. 中国图书馆学报，2013，39（2）：36-45.

作者简介

夏立新，男，1968年生，教授，博士生导师，研究方向：信息组织与检索。
杨元，女，1998年生，硕士研究生，通信作者，研究方向：信息组织与检索，E-mail：yangyuan2016@mails.ccnu.edu.cn。
郭致怡，女，1999年生，硕士研究生，研究方向：信息组织与检索。

Research on the Automatic Processing and Long-term Preservation Strategy of University Network Information Resources

XIA LiXin YANG Yuan GUO ZhiYi
(School of Information Management, Central China Normal University, Wuhan 430079, China)

Abstract: Based on a comprehensive review of the domestic and foreign research status of automatic processing and long-term preservation of network information resources and the literature information resource guarantee in colleges and universities, this article points out the necessity of universities' automatic processing and long-term preservation of network information resources. An in-depth analysis of the best practice cases of automatic processing and long-term preservation of university network information resources, based on the Bentley History Library web archiving project, based on this, to improve the university literature information resource system as the guidance, from the three dimensions of resources, technology and management develop strategies for automated processing and long-term preservation of network information resources in Chinese universities.

Keywords: Web Archive; Network Information Resources; Literature Information Resources Guarantee; University Library

（收稿日期：2021-08-02）

书 讯

《汉语主题词表》

《汉语主题词表》自1980年问世以后，经1991年进行自然科学版修订，在我国图书情报界发挥了应有作用，曾经获得国家科学技术进步二等奖。为适应网络环境下知识组织与数据处理的需要，由中国科学技术信息研究所主持，并联合全国图书情报界相关机构，自2009年开始进行重新编制工作，拟分为工程技术卷、自然科学卷、生命科学卷、社会科学卷四大部分逐步完成。目前工程技术卷和自然科学卷已出版。

《汉语主题词表（工程技术卷）》共收录优选词19.6万条，非优选词16.4万条，等同率0.84，在体系结构、词汇术语、词间关系等方面进行了改进创新。《汉语主题词表（自然科学卷）》共收录专业术语12.4万条，包含数学、物理学、化学、天文学、测绘学、地球物理学、大气科学、地质学、海洋学、自然地理学等学科领域，收词系统、完整，语义关系丰富、严谨，每条词汇都有相应的学科分类号表现其专业属性，并与同义英文术语对应。同时，建立《汉语主题词表》网络服务系统，提供术语查询、文本主题分析、知识树辅助构建等服务。《汉语主题词表》可用于汉语文本分词、主题标引、语义关联、学科分类、知识导航和数据挖掘，是文本信息处理及检索系统开发人员不可或缺的工具。

《汉语主题词表（工程技术卷）》已于2014年由科学技术文献出版社出版，分为13个分册，总定价3 880元。《汉语主题词表（自然科学卷）》已于2018年5月由科学技术文献出版社出版，分为5个分册，总定价1 247元。两卷均可分册购买。