Web Archive发展历程与发展趋势研究[1]

李华[1] 吴振新[2] 郭家义[3] 向菁[24]

[1]（沈阳工业大学辽阳校区，辽阳 111003）[2]（中国科学院国家科学图书馆，北京 100190）

[3](北京市信息资源管理中心，北京 100000）[4]（中国科学院研究生院，北京 100049）

[摘要]本文回顾了网络信息资源保存的发展历史，分析了网络信息资源保存在初始实验、应用部署和长远发展这三个阶段中的不同进展和特点，通过总结网络信息资源保存的研究历程和近年来国内外实践，初步预见了未来趋势展望了网络信息资源保存发展趋势，以期对我国网络信息资源保存起到参考作用。

[关键词] 网络信息资源 发展 历程 趋势

[分类号] G250

Web Archive Development Research

Li hua[1] Wu Zhenxin[2] Guo Jiayi[3] Xiang Jing[24]

[1](Liaoyang Campus of Shenyang University of Technology,liaoyang:111003,China)

[2](The Library of Chinese Academy of Science,beijing:100190,China)

[3](Beijing Information Resource Management Center,beijing: 100000,China)

[4](The Library of Chinese Academy of Science beijing:100190,China)

[4](Graduate School of the Chinese Academy of Sciences,beijing:100049,China)

[Abstract] This paper reviews the development history of Web Archive, analyses the progress and characteristics of the three different stages of the initial experiment, deployment and application of web archive. By summing up researches of Web Archive and international practice in recent years, we give an initial view for the future development trend of Web Archive, hope to be a valuable reference for Chinese Web archive researches.

[Keywords] Web information resources development research trend

网络信息资源（以下简称 Web 资源）广义上是指网络信息活动中所有要素的总和，包括与网络相关的信息内容、信息网络、信息人才、信息系统、信息技术等资源；狭义上是指以数字化形式记录本身，以多媒体形式表达，存贮在计算机磁介质、光介质以及各类通讯介质上，并通过计算机网络发布、传递和存储的各种信息资源的集合，也称网络文献。网络信息资源作为全球最大的信息资源库，存在大量对文化遗产、学术研究具有重要价值的信息；随着网络的普及，网络信息已逐渐融入大众生活，网络信息资源的保存（Web Archive，以下简称 WA）也因此变得日益重要。

1、WA 发展历程研究

WA 发展十余年来，从最初的百万数据小规模采集试验发展到大规模数十亿 URL 采集；采集内容也从基于域名的单一采集方式发展到基于事件、主题、复杂域等多种采集方式；项目资金从单个组织机构的投资到组织机构、图书馆、档案馆、商业公司的多方投资，资金数量也是随着采集规模、难度而逐渐递增。按照 WA 发展成熟度的不同，可以将国际 WA 研

[1]本文受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”课题的资助，课题编号为 06BTQ025。

究与实践分为三个阶段。

1.1 初始实验阶段（1996 年-1999 年）

1996 年，Internet Archive（以下简称IA）[1]正式成立标志着WA在全球研究与实践的开始。这一阶段，欧美各国纷纷开始WA实验项目，与此同时，为了保障WA项目的顺利开展，WA组织结构开始建立。

IA 的目标是实现全世界 Web 资源的收集、保存和永久获取，为国际性存档开发技术工具制定标准，鼓励和支持图书馆、档案馆、文化遗产机构实现网络资源的收集和保存，为人类创造一个互联网图书馆。IA 负责开发的“时光倒流机器”（Wayback Machine）按 URL 的主题对网页每两个月进行一次快照，每个爬行器能收集 25 亿张网页信息，每月可累积达到 15GB，目前能实现 110 亿网页信息的全文检索。

1996 年，澳大利亚国家图书馆发起PANDORA项目[2]，保存澳大利亚的在线出版物，包括社会科学、政治、经济、宗教、自然科学、文化等重要文献资产。项目开发了一个“数字信息存档系统（PANDAS）”。从政策上，发展了数字保存政策（Web上提供），提供图书馆保存自有资源的指导以及如何与其它机构合作来获得保存获得的最大收益。从合作上，它建立了与州图书馆及其它学术机构的合作框架，并在国际社会谋求更多的合作；在系统开发上，主要在数字对象存储系统（Digital Object Storage System）、数字对象管理系统（Digital Objects Management System）、数字存档系统（Digital Archive System）等三个领域进行了投入。由澳大利亚国家图书馆领导的另一个重要的项目是关于数字资源保存主题的门户网站－PADI[3]，主要研究目标是提供一种机制，帮助确保数字格式信息能够被有效管理、保存和提供未来的访问。PADI于 1996 年开始正式运行，于 1997 年提供基于Web的版本。

瑞典在 1996 年设立名为Kulturarw3 的Web信息资源采集项目[4]。该项目以瑞典Web信息资源为对象制定了一揽子收集的方针，通过网络机器人进行数据采集域为.se .com，.org，.nu 的网络资源。2002 年 3 月，开始提供公共服务，为NWA项目的开启提供借鉴经验。

1997 年北欧图书馆启动了NWA项目[5]。项目来源于斯德哥尔摩皇家图书馆 1996 年发起的Kulturarw3 项目，1997 年它以论坛的形式定期讨论交流Web收割及保存经验，重点关注网络资源的获取访问，NWA项目正式开启。1998 年北欧图书馆承担起NWA的主要责任机构的角色，成立专门指导小组负责NWA的协调工作，主要目标是根据保存、访问的要求制定相关的技术规格，协助国家项目的协调发展，联合北欧各国图书馆建立欧洲网络资源长期保存的合作机制。

1997 年美国国会图书馆启动网络信息保存试验项目—Minerva Prototype[6]，为有关Web信息的数字化、元数据、选择和采集、长期保存与获取方面的实际问题提供试验，从而为美国国会图书馆运行一个大规模的Web信息保存项目提供指导和经验。在该项目实施的第一阶段，由项目协调小组、参考咨询专家、数字转换专家、编目专家、网络开发及MARC维护的代表专家构成Minerva项目的核心小组负责收集、保存书目记录，并提供记录的开放获取。第二阶段，国会图书馆通过与Internet Archive和WebArchivist.org共同合作创建 “2000 年总统选举”网页保存项目。Internet archive为此项目提供“时光机”（Wayback machine）采集技术，为收集到的网页建立索引、网页时间依赖的浏览，同时提供临时的数据存储，用户可按日期、网站、类别获取信息。此项目平均每天采集 150-200 个网站信息，于 2001 年 3 月结束。除此之外，国会图书馆还与Internet Archive、WebArchivists.org、Pew Internet & American Life Project（皮尤互联网与美国人生活计划）共同发起“911 网页保存”项目保存国内外个人、团体、媒体有关“911”的网页信息。计划采集对伊拉克战争、2004 年总统选举、联邦政府等主题的网页信息。

1.2 应用部署阶段（2000-2004 年）

在这个阶段 WA 已经获得欧美各主要国家的广泛重视，各国开始从国家层面进行战略计

划和思考，开始建立国家或区域战略合作保存体系，开始数字资源长期保存活动的规模化应用部署。具体分析，该阶段主要有以下 4 个特点:

（1）国家或区域战略合作保存体系迅速发展

2000 年美国国会图书馆提出了“国家数字信息基础设施及保存计划”（National Digital Information Infrastructure Preservation Program ，NDIIPP)[7]，其中包括立即收集和保存可能瞬间即逝的Web数字信息。保存范围主要集中在网页、电子图书、电子期刊、数字电影、数字音频、数字电视等 6 个方面，其中，网页保存被视为该计划的重要组成部分，尤其是那些仅以数字格式出现的资源，如美国大选、卡特里娜飓风、奥运会、伊朗人的博客等。

2003 年 6 月，由来自多个国家的 12 个成员机构组成的国际网络保存联盟(IIPC：International Internet Preservation Consortium）[8]正式成立。该联盟最初成员 12 名，并与成员正式达成共同出资参与项目和工作组的合作协议。其目标是：保存丰富的来自全球的Internet内容，使其能够持续地提供访问；为国际互联网保存联盟制订出一个联合规范；设计和开发出网络资源保存工具；促进发展和使用公共工具、技术和标准来生成国际性存档；鼓励和支持各国进行Internet存档和保存。研究工作按访问、内容管理、深层网络、框架、规则与测试、研究人员需求共 6 个工作组进行，将提供体系结构、标准、访问需求分析、公共工具等，截至 2008 年 5 月，该联盟已有 37 个成员机构。

（2）拓展 WA 项目研究内容及深度

在此阶段，WA 项目扩展研究的范围和内容，更为关注保存工具使用、深层网络采集和用户呈现等内容。

2001 年，挪威国家图书馆开始实施WA项目Paradigma（Preservation，Arrangement &Retrieval of Assorted Digital Materials）[9]，以确定收集和保存网络信息资源的技术、方法和组织方式，并且使国家图书馆能够在呈缴制度的框架下提供网络信息资源存取服务。

2001 年，英国国家图书馆启动WA的试验性项目Domain.UK[10]，探索有关网络存档的一系列问题。该项目运用半年时间选择 100 个英国站点网页资源进行保存，图书馆取得网页所有者明确同意后对这些网页进行采集保存，不提供公众获取。

2002 年，法国国家图书馆的WA项目（Bnf Web Archive）[11]主要关注如何重新定义网络爬虫的参数来提高采集资源的效率；测试网络资源保存的每一流程步骤，特别对深层网络资源采集予以重点关注。该项目采集了有关法国大选的 1900 个站点和域名为.fr网页资源。在深层网页资源的保存方面，BnF选择 100 个网站进行深层网页保存的试验，使用DeepArc从数据库中抽取元数据，并以XML格式存放。目前BnF正在进行 2007 年法国大选网页的采集，参与由IIPC、LC、BL联合开展的智能爬虫（smart crawler）的相关研究，计划进行机构仓储的建立等。

Web at risk项目[12]是 2004 年美国国会图书馆资助的八个数字保存项目之一，该项目试图建立一套国家政治文化遗产长期保存的内容识别、选择、获取的方法，建立国家保存文化遗产信息的保存网络。

（3）国际机构、会议对 WA 日益关注

除了以国家图书馆为代表的 WA 项目外，一些国际会议也将 WA 进行专题讨论，WA 日益受到相关国际组织、机构的重视。

2002 年，第 68 届 IFLA 理事会对以往国家层面 WA 采集（如 NWA）经验予以总结，探讨 WA 的法律问题。

自 2001 年ECDL（European Conference on Research and Advanced Technology for Digital Libraries ，ECDL）会议成立第一届WA专题组IWAW（International Web Archiving Workshops）[13]以来，IWAW每年都围绕WA从开源工具、技术、政策、法律、已有项目经验总结与展望、未来发展重点等几个方面对WA予以讨论和关注。随着WA的推动和发展，IWAW会议讨论的

主题从项目经验学习到工具的开发到提高工具效率、解决WA采集、保存、访问的技术难题、网页呈现，主题的范围和深度也在不断的扩大和加深。

2001 年主要是介绍德国、丹麦、瑞典、法国、美国 Minerva 项目经验及未来的发展计划；2002 年第二届会议重点关注 WA 涉及的技术、采集策略、合作平台的构建、关注智能爬虫工具、描述网页结构的 METS 等。法国、芬兰、丹麦、美国国会图书馆合法采集策略，以 Xyleme 工具为例探讨了其采集网页上 XML 数据、管理 HTML、保存带宽、计算网页排名和变化频次等功能，使用自动分类技术聚合不同应用域，实现语义整合 XML DTDs 的目的；2003 年会议更加关注 WA 开源工具情况，更广范围探讨了 WA 项目的现状，更深入探讨了 WA 采集的相关问题：如基于主题采集、政治主题采集的经验和面临的挑战。2004 年起关注提高采集器爬行的准确性和效率，改善采集器的强健性、灵活性和可维护性，特别是对避免重复的下载而对于同一资源进行重复性的保存——爬行器的增量爬行技术问题予以探讨。

（4）WA 系统的应用与发展日益得到重视

构建 WA 系统，对授权保存的 Web 信息资源进行长期保存，是国际图书馆界正在采取的一项重要行动。欧美等发达国家的图书馆和联盟已经建成了诸如 e-Depot、Portico、LOCKSS 等投入实际服务的长期保存系统，以多种方式对数字学术资源（网络信息资源的一种）进行长期保存，以确保当前的某些数字信息资源在将来能够被特定的用户持续利用。

由美国Stanford 大学图书馆发起并组织实施的LOCKSS (Lots of Copies Keep Stuff Safe) 项目[14]，受美国国家自然基金、Sun Microsystems Inc 以及Andrew W. Mellon 基金支持，致力于解决电子期刊的长期保存和利用。该项目通过分布式地保存多个数据存档来实现永久稳固的保存。它通过建立出版商与图书馆之间的协作关系，允许图书馆利用LOCKSS提供的开源保存系统在本地创建一个低费用、永久保存的数字化的信息缓存站点，实现电子期刊信息的本地化收集、存储、管理和服务。Stanford大学于 2001 年发布了一个开源的长期保存管理工具LOCKSS，并建立了LOCKSS联盟，该联盟通过获取国家及各种基金资助和收取会员费来保障其正常发展和有序运行，在全球范围吸纳了越来越多的成员加入，目前包括 180 多家图书馆和 50 多个出版商，系统中的数据内容每周都在增加中。

荷兰国家图书馆（KB）数字存档系统e-Depot[15]是由荷兰国家图书馆与IBM公司合作开发的一套完全自动化的数字资源保存系统，长期存储着国际上主要出版商的e-journals，主要用于保存研究领域已出版的艺术、人文和社会科学、科学、技术和医学以及数字化的文化遗产。e-Depot为出版商提供存档格式的耐久性检查，并指导出版商如何创造最持久的电子出版物。e-Depot不接受单一的出版物，愿意使用该服务的出版商必须与其签订一项存档协议，大批量的提交内容和指定的元数据。KB计划与全球最大的 20 家出版商签订存档协议，目前Elsevier，Springer，Blackwell，Biomed，Oxford，Taylor & Francis，Sage，Brill and IOS已经与KB签订了存档协议。截至 2007 年 11 月，e-Depot已摄取了 1 千多万个数字对象。

Portico项目[16]最初由JSTOR于 2002 年发起，后于 2005 年获取了JSTOR、 Ithaka、美国国会图书馆和Andrew W. Mellon 基金的联合资助。目的是构建一个可信赖的第三方存档系统以保存电子格式的正式出版的学术资源。Portico主要通过存档的出版商和机构的资助来保障其正常运行，同时它也积极寻求其它基金和政府的资助。Portico从出版商直接获取电子期刊文献，通过特定的转换或规范化处理把多样化的提交格式文档转换成更适于长期保存的统一存档格式。到 2008 年 6 月底共有 468 家图书馆、56 家出版商提交了 7，447，642 篇文章。

1.3 长远发展阶段（2005—）

在 WA 相关项目的实验研究、应用部署发展的基础上，相关的组织机构、项目对如何更好的将 Web 资源呈现给用户，提供检索服务，并进行相应的数据挖掘以用于学术研究、追踪动态等 WA 长远发展问题更为关注。

2005 年后，IWAW会议讨论的议题更为广泛[17]。2005 年IWAW会议扩展到数字资源保存的问题，关注IIPC开发的一系列开源工具集、ARC格式的扩展和升级版——WARC、视/声频采集问题、时间维中无法获取的URL、死链接、链接深度的问题。2006 年IWAW会议继续关注WA优化和注释、流媒体的保存、工具；2007 年会议对WA目前的开源工具进行详尽分析，关注WA编目注释方法探讨。

当前，IIPC将其成员范围扩大到致力于Web资源保存的其他图书馆、档案馆、博物馆和文化遗产机构。截至 2008 年 5 月，该联盟已有 37 个成员机构，目前亚洲的日本国会图书馆已作为新成员加入IIPC[18]。该联盟在功能构建、标准API、保存格式（WARC：Web ARChive file format）、元数据、永久标识等方面建立技术标准规范，拥有从数字资源获取到检索全过程的技术工具，这些工具均是开源、扩展性强，适用于不同环境、不同系统的保存，并建立组织内外有关Web信息保存的信息分享平台。目前，IIPC已对网络信息分类问题达成一致意见，对Web信息系统界面规范、文件格式和元数据标准提出相应意见，完善新检索工具的识别规范，开发网络信息保存的控制、规划的保存工具，开发完善WARC (Web ARChive)工具，定义保存网络爬虫抓取的文件模块，支持相关内容模块，并开发一系列开源软件。

2、WA 发展趋势分析

综观 WA 的发程历程，和近年来国际 WA 实践，可以预见未来 WA 的发展将呈现以下发展趋势：

2.1 主题和内容丰富化趋势

随着 WA 的发展，采集的主题内容、形式也日益丰富化。采集主题内容从政治、社会文化、健康到艺术、人文，基本涉及人类的各个知识领域。国际 WA 领域的重要会议之一的 Iwaw（International Web Archiving Workshop）2005 年对 WA 的视频、声频保存专题进行讨论，2006 年的会议又将流媒体保存予以专题讨论，可见多媒体动态内容的采集也成为 WA 采集中重要的形式。

（1）WA 保存的内容和主题日益丰富

Web At Risk项目[19]涉及的子项目内容涉及政治、经济、社会文化等多方面，包括加利福尼亚州政治博客及相关政治团体的网站、伊斯兰教、中东国家政治团体及反政府团体的信息、人口统计数据、贫困、HIV/AIDS相关信息、保存美国有关劳动、就业问题的信息、国际经济发展等。IA[20]保存网页信息；电影、动画片、新闻、游戏、演讲、电视采访等视频；音乐、录音、广播节目等音频，特别对现场音乐会、总统演说录音、有关新闻、公共事务的广播节目等予以保存。目前已保存 19，000 场音乐会及 1，000 场演讲；公共领域的书籍和文档；游戏引擎制作的引擎电影（Machinima）、速度测试、经典软件等。目前已保存 10，000 多个软件；中美各大学免费的课程视频录像、演讲、课堂补充学习资料等教育资源

（2）WA 采集的形式从传统的静态网页的采集向多媒体动态内容的采集方向转变，目前也有将 Web2.0 软件形式纳入采集的意向

Web2.0 是 Web 发展的重要组成部分，在学术研究、科学产出、e-learning 领域都有应用，越来越多的学术团体将 Web2.0 资源作为重要的参考源，进行 Web2.0 资源的保存是十分必要的。但 Web2.0 资源的保存面临保存责任者难以界定、隐蔽网采集难度、存储难度等多方面的挑战，各种 Web2.0 的应用模式又各自具备自身的特点及保存中需要考虑的问题。如博客保存要考虑如 Web 上存在对博客内容的讨论，存在链接对象不存在、无法识别的问题；博客的保存范围问题（是保存完整的博客？是否需要保存其他人的评论和回溯链接？）；很多博客在服务器系统的主机上，实际的内容则是保存在数据库中，是隐蔽网的一部分；博客软件提供自身内部的保存系统，允许一些爬虫进行完整性的采集，存在持久保存、用户授权采集的问题。分享网站保存（如 YOUTUBE 视频分享网站）：属于隐蔽网，存在用户许可的

问题；本身提供存储；用户可创建自身的社交网络。数据 mash-up 要考虑保存责任者难以界定问题，不同数据源采集难度问题等。

（3）WA 的内容管理，特别是保存资源评价、质量控制方面日益重视

加拿大政府网络保存项目（Government of Canada Web Archive）[21]开发了索引和质量控制的工具（IQ App），允许工作人员对保存的站点进行索引和质量控制；开发的爬虫管理工具CMD通过Heritrix JMX界面管理爬虫的所有活动。对于采集的非政府网站信息、网站首页内容多次呈现等问题，该项目通过使用Heritrix 的profile设置， Nutchwax 修剪工具从索引中移除不相关的索引内容；通过在Nutchwax 检索顶层增加一层内容来进行检索结果过滤，通过Wayback排除参数设置对某些保存内容不予显示，从而更好的进行WA内容的管理和质量控制。

2.2 系统建设标准化和开源化趋势

（1）WA 项目在标准规范方面不断予以改进

IIPC致力于WARC标准的推广，及ARC向WARC转换工作，完善转化框架和工具开发。IIPC制定出管理WARC文件内容的综合、可扩展、高速的框架——Libwarc[22]，并计划于 2008 年 10 月推出libwarc 1.0 版本和应用的 1.0 版本。WA服务机构Hanzo[23]开发WARC操作工具，可在第三方上运行。

（2）WA 项目中所使用的采集、索引、访问工具基本都是开源的，而且在提高开源软件的效率、性能、规范化方面还在进行不断努力和探索

IA 与 Nordic 国家图书馆联合开发开源、可扩展的网络爬虫 Heritrix、索引工具 NutchWAX；新西兰国家图书馆和大英图书馆共同开发 WCT 工具；IA 与挪威图书馆开发的资源访问工具 WERA 等。IIPC 在访问、内容管理、隐蔽网、框架构建、试验等方面都作出一些有力的尝试和实践，与 IA、IIIPC 的合作图书馆合作开发了开源的 WayBack、NutchWAX 、WERA 访问工具，WCT 内容管理工具，XML 隐蔽网软件，构建 WARC 记录的标准流程及 ARC 一系列工具，并对采集、索引进行相关试验。IIPC 目前正与 LC、BNF、BL 资助开展智能爬虫的相关研究，包括复本还原（Duplicate reduction）、增强优先级（Enhanced prioritization）等，计划在 2008 年继续进行适应性再爬行（adaptive recrawling）的探索实践工作，在国家分类、网络陷阱识别、字符编码（简体/繁体）、网络地址发生变动、视频实时浏览等方面进行实验研究，所有成果均以开源软件的形式公布。

2.3 工作流程规范化趋势

随着 WA 十余年的发展，项目在工作流程规范化，提高工作效率方面尝试和探索越来越多，对 WA 的采集、法律问题、编目、保存、访问等环节制定了详细、规范的流程。

法国图书馆在深层网络采集方面有规范化的工作流程[24]。法国图书馆组建包括图书馆各部门负责法律呈缴的专业人员的专门小组，处理数字环境下网络资源保存问题，小组主要利用访问工具和观察法来评估爬虫的结果和采集的质量。采集器进行目标采集时根据小组建议来确定URL采集的变量。运用自动和手工相结合的方式对深层网络信息进行采集。自动化的技术发现深层网的信息源，然后和出版者进行协商，由出版者以email、FTP、CD等方式呈交或发送给存档机构，由存档机构本地上传。手工采集深层网络信息方面：法国国家图书馆同时开发了一种用来采集深层网页的方法策略——呈缴跟踪（deposit track），由图书馆员提供评价一系列的站点，然后和站点所有者进行协商，并希望他们将站点内容呈缴予以保存。如果所使用的网络信息采集工具不能很好的保证采集效果，会要求网站管理员将网络信息资源以email、FTP、CD等方式发送给保存机构，由保存机构进行本地上传。为了实现对数据库网络存档，法国国家图书馆开发了DeepArc工具，把数据库内容转化为XML文件，在进行网络采集和存档，该工具成为IIPC的主要工具之一。

Minerva Prototype项目[25]工作流程完整，过程包括计划、法律、选择、权限管理、采集、

编目、保存、界面开发等过程，整个模型建立于Pandora模型的基础之上。Minerva按照MARC21 和AACR2 建立文件，使用OCLC的CORC系统进行编目。如果在CORC中发现某一个资源的编目记录，则按照国会图书馆复制编目的原则将该记录作为一个MARC记录合并到图书馆集成管理系统中；如果在CORC中没有发现记录，则通过CORC的元数据发现功能自动生成准备编目数据。

2.4 更为广泛的规范化合作趋势

国际 WA 领域开始构建更大范围的长期保存网络合作模式，共享 WA 系统和资源，利用分布式的系统和资源构成网格和协作网络，构建异地分布的 WA 合作框架，以促进实践中的资源共享、职责与费用分摊以及交流等。其中比较典型的合作项目有 NWA 基于工具合作机制、SDSC 基于大规模存储网格合作机制、PANDORA 基于采集合作机制等。

（1）NWA 基于工具合作机制

NWA项目[26]开发了一系列网络信息资源保存的工具包，包括有文本检索工具、导出工具、访问模块、附加的组件有搜索引擎、搜索引擎抽象层等，旨在构建欧洲领域WA框架。在索引和检索模块上，对于挪威、瑞典、丹麦、冰岛、芬兰的异地分布的存储结点，主要是通过与挪威的搜索引擎公司FAST开发的索引工具实现索引和访问。该搜索引擎具有良好的扩展性，采用一个适用于任何国家的超级分布式结点，在五国设置分布式结点，每个分布式结点下再继续细分为n个结点。当用户进入访问界面提出检索请求，模块向前台分布式结点向北欧所有的索引发出请求，得到结果后再将结果返回给访问模块，模块再将此结果传递到浏览器供用户浏览。

（2）SDSC 基于大规模存储网格合作机制

SDSC的Chronopolis项目[27]是运用最新的存储技术和网格设备搭建长期保存的环境，建立基于网格的概念性长期保存框架，在满足潜在用户团体需求的基础上，建立长期保存的风险成本模型、服务模型，并与内容提供商、合作者、用户签署谅解备忘录、服务水平协议来建立互信的合作关系。数据网格由地理上分散的 3 个站点组成，将每份数据资源存在 3 个独立管理的复本，同时增加的说明层为数据管理和数据完整检验提供了额外的安全保证，系统、数据等之间的异构性由SRB（Storage Resource Broker ：存储资源代理）及iROD（存储资源代理的开源版本）进行处理。

（3）PANDORA 基于采集合作机制

PANDORA项目[28]的合作伙伴分布在澳大利亚的各个州，为加强各州合作，PANDORA开发了一个综合的、基于网络的、可以远程工作的数字存档系统PANDAS。每一个合作伙伴在各自不同在采集策略基础上均采用HTTrack爬行器来采集网页的内容，也可以直接上传文件或者出版物到PANDAS系统之中。英国网络保存机构UKWAC[29]也是采用PANDAS系统作为其网络信息资源保存系统，采取基于采集的合作方式。

2.5 利用形式多元化趋势

WA 利用由以前的提供检索访问发展到网站恢复重现、保存 Web 文献参考链接信息、搜索引擎结果的改进、分析 Web 技术演进等利用形式。

（1）网站恢复重现

网站重现将 WA 存储器中存储的网站内容以其原有的样貌展现给用户，让用户感觉就像是在访问原始网站一样。网站重现需要恢复原有的超链接情景和同一网站不同时间采集的版本。目前主要有超链接重写、HTML 重写、Proxy URL 技术方法来实现网站的重现。

（2）保存 Web 文献参考链接信息

Web上参考链接文献具有重要的保存价值，目前为非期刊Web参考文献设计的Archive系统——WebCite[30]，文章的作者利用此系统可将引用的来源网站（页）进行存档，并提供一个网址来连接这个存档的网页，呈现所存档的网址及存档日期，使用户能随时浏览真实的

网页。

（3）搜索引擎结果的改进[31]

WA 用于搜索引擎结果改进是针对搜索引擎的爬行范围有限、网站内容的动态变化等原因，使用链接结构分析来评价网页的质量和排名的方法导致一些拥有高质量信息的页面排名靠后，用户难以获取。通过分析排名靠前的页面以往历史版本内容，通过对比一些排名靠后的页面以往版本与排名靠前网页的核心内容有着很高的相似性方法，提升排名靠后网站的排名，为用户获取。

（4）分析Web技术演进[32]

WA 用于分析 Web 技术演进是通过查看 WA 中特定时间、特定版本的信息，从而可以追踪信息使用的技术、系统生命周期，跟踪文件格式、交互标准和创建某些信息的服务器技术，根据某项技术的使用程度，还可以决定是延长使用或早些淘汰此类技术。

3、结语

国外 WA 相关研究与实践为我国 WA 的开展提供了丰富的经验和案例支持，对于我国 WA 的发展具有重要的借鉴意义。我国应当充分借鉴国外发展的相关经验，把握 WA 的发展趋势，重视 WA 的研究与实践工作，重点开展以下几个方面的工作：

（1）制定 WA 保存国家战略。在国家层面制定相关政策，确保 WA 工作在国家规划范畴内进行。

（2）重视国家图书馆的保存作用。强化国家图书馆 Web 信息资源的保存主体作用，保障国家图书馆在 WA 的主导地位。

（3）加强合作。需要加强三方面的合作，一是国内保存机构间的合作；二是保存机构与出版商、技术支持厂商的合作；三是要加强国际合作。

（4）建立健全有利于 WA 的法律环境，特别是要完善 Web 信息资源的呈缴制度，使得 WA 工作制度化，长久化。

（5）加强标准化和规范化研究。在与国际标准接轨的同时，加强中文 Web 信息资源的特点，尽快出台一系列中文 Web 信息资源保存的标准。

（6）加强技术策略研究与保存系统开发，支撑 WA 工作的开展。

作者简介：

李华，女，馆员，沈阳工业大学辽阳校区图书馆，联系方式：辽宁省辽阳市沈阳工业大学辽阳校区图书馆，111003

吴振新，女，副研究员，中国科学院国家科学图书馆，联系方式：北京市北四环西路 33 号中国科学院国家科学图书馆，100190

郭家义，男，副研究馆员，北京市信息资源管理中心，联系方式：北京市朝阳区北辰西路 12 号数字北京大厦 7 层北京市信息资源管理中心，电话：01084371831，电子邮件：guojy@beijingit.gov.cn

向菁，女，中国科学院国家科学图书馆 2007 级硕士研究生，联系方式：北京市北四环西路 33 号中国科学院国家科学图书馆，100190

参考文献

[1] Internet Archive [EB/OL]. [2008-06-15] http://www.archive.org/index.php.
[2] PANDORA[EB/OL]. [2008-06-15]http://pandora.nla.gov.au/.
[3] PADI [EB/OL]. [2008-06-15]http://www.nla.gov.au/padi/.
[4] Kulturarw3 Project[J/OL]. .[2008-06-15].http://www.ifla.org/IV/ifla66/papers/154-157e.htm.
[5] Nordic Web Archive (NWA). [2008-07-03].http://nwa.nb.no/.
[6] Minerva [EB/OL]. [2008-06-15].http://www.loc.gov/minerva/.
[7] National Digital Information Infrastructure and Preservation Program[EB/OL]. [2008-06-20]. http://www. digitalpreservation．Gov/index. php?rlav=3&subnav=2.
[8] IIPC [EB/OL]. [2008-06-15].http://www.netpreserve.org/about/index.php.
[9] Paradigma [EB/OL]. [2008-06-15].http://www.digitalpreservation.gov/.
[10] The British Library: Web Archiving Challenges and Initiatives [J/OL]. [2008-06-15].http://www.diglib.org/forums/fall2005/presentations/tuck-2005-11.pdf.
[11] Web Archiving at BnF[EB/OL]. [2008-06-15]. http://www.bnf.fr/pages/version_anglaise/depotleg/pdf/BnFnews200609.pdf.
[12] Web At Risk[EB/OL]. [2008-06-15].http://web3.unt.edu/webatrisk.
[13] IWAW [EB/OL]. [2008-06-15]. http://www.iwaw.net/.
[14] LOCKSS [EB/OL]. [2008-06-15].http://www.lockss.org/.
[15] e-Depot[EB/OL]. [2008-06-15].http://www.kb.nl/dnp/e-depot/e-depot-en.html.
[16] Portico[EB/OL]. [2008-06-15].http://www.portico.org/
[17] IWAW [EB/OL]. [2008-06-15]. http://www.iwaw.net/.
[18] IIPC [EB/OL]. [2008-06-15].http://www.netpreserve.org/about/index.php.
[19] Web At Risk[EB/OL]. [2008-06-15].http://web3.unt.edu/webatrisk.
[20] Internet Archive [EB/OL]. [2008-06-15].http://www.archive.org/index.php.
[21] Gillian Cantello, JOHN STEGENGA . Government Web content in Canada A national library web archive perspective [J/OL]. [2008-08-10]. http://www.ifla.org/IV/ifla74/papers/130-Cantello_Stegenga-en.pdf.
[22] Libwac[EB/OL]. [2008-08-10].http://code.google.com/p/warc-tools/.
[23] Hanzo[EB/OL]. [2008-08-10].http://www.hanzoweb.com/.
[24] Web Archiving at BnF[EB/OL]. [2008-06-15]. http://www.ifla.org/VI/4/news/ipnn40.pdf.
[25] Minerva [EB/OL]. [2008-06-15].http://www.loc.gov/minerva/.
[26] Nordic Web Archive (NWA). [2008-07-03].http://nwa.nb.no/.
[27] Chronopolis[EB/OL]. [2008-07-14].http://chronopolis.sdsc.edu/.
[28] PANDORA[EB/OL]. [2008-06-15].http://pandora.nla.gov.au/.
[29] The British Library: Web Archiving Challenges and Initiatives [J/OL]. [2008-06-15].http://www.diglib.org/forums/fall2005/presentations/tuck-2005-11.pdf.
[30] WebCite[EB/OL]. [2008-07-14]. http://www.webcitation.org/.
[31] Jatowt A., et al. Using Web Archive for Improving Search Engine Results[J]. Proceedings of the Eighth Asia Pacific Web Conference (AP-Web2006), 2006:893–898
[32] Rauber A., et al. Uncovering Information Hidden in Web Archives[J/OL]. D-Lib Magazine, 2002,8(12):1082-9873. http://www.dlib.org/dlib/december02/rauber/12rauber.html.