

网络信息资源保存发展现状及趋势分析 *

向 菁 吴振新

摘 要 20 世纪 90 年代末,网络信息资源保存(WA)开始引起关注,与 WA 相关的研究和实践随之展开。从国际方面看,WA 在参与主体、研究方式及项目状况、保存内容、技术标准、系统工具、法律政策、经济效益、合作机制等方面都有了不同程度的发展和完善,而我国 WA 发展还相当薄弱。WA 的未来发展趋势为:主题和内容丰富化、系统建设标准化和开源化、工作流程规范化、合作广泛化、利用形式多元化。表 1。参考文献 38。

关键词 网络信息资源保存 发展现状 趋势分析

分类号 G250.73

ABSTRACT This paper analyzes the development of domestic and international Web Archiving, summarizes important issues and weaknesses of current researches, points out some aspects of future developments, and provides a general view of Web archiving. 1 tab. 38 refs.

KEY WORDS Web Archiving. Development analysis. Trend prediction.

CLASS NUMBER G250.73

随着信息技术的发展,网络逐渐成为重要的信息媒体、成为科学文化交流的重要平台,越来越多的信息资源以数字化的形式存在,网络信息资源(以下简称 Web 资源)成为人类文化遗产组成中非常重要的一部分,且呈指数级增长。由于 Web 资源具有更新频率快、易逝等特点,这使得 Web 资源寿命短暂,如果不及时时加以保存,大量具有重要价值的学术、文化、科学信息就会丢失。20 世纪 90 年代末,网络信息资源保存(Web Archive,以下简称 WA)开始引起一些组织、机构的关注,随后图书馆、档案馆、商业公司及相关的机构、组织、专家纷纷开始开展与 WA 相关的研究和实践。

1 WA 国际发展现状分析

1996 年 Internet Archive^[1](简称 IA)的成立标志着 WA 研究与实践的开始,自此,WA 领域出现了一批研究项目和应用系统,逐渐形成了比较可靠的保存体系。综观 WA 十余年的发展

历程可以看出:WA 在参与主体、研究方式及项目状况、保存内容、技术标准、系统工具、法律政策、经济效益、合作机制等方面都得到了不同程度的发展和完善。

1.1 越来越多的国家和机构参与到 WA 活动中来

从项目数量来看,越来越多的国家和机构参与到 WA 活动中来,项目数量呈稳定增长趋势。国际互联网保存联盟^[2](以下简称 IIPC)的机构成员已由 2003 年成立初的 12 个发展到目前的 38 个,主要分布在欧洲、北美洲和亚洲。2008 年,IIPC 对其成员进行问卷调查显示^[3]:该组织的成员 50% 是国家图书馆,10% 是高校图书馆,8% 是其他类型图书馆,3% 是国家档案馆,3% 是内容提供商,26% 是研究机构、政府组织等。由此可见,图书馆特别是国家图书馆已经成为 WA 活动参与的主力,并担任了重要的角色。

除图书馆、档案馆、研究机构、内容提供商

* 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”(编号:06BTQ025)的资助。

作为参与 WA 活动的主要力量外,商业机构也以技术支持者的身份积极参与 WA 的研究与实践。Alexa^[4]为 IA 开发 WA 访问工具 Wayback;挪威的搜索引擎公司 FAST^[5]为 NWA (Nordic Web Archive,简称 NWA)项目开发了适于 WA 的搜索引擎;IBM^[6]为荷兰国家图书馆(KB)的 WA 研究提供技术支持和系统服务。

1.2 WA 研究仍以实验和项目形式为主

在 WA 发展初期,项目以小规模 Web 资源采集的探索性实验为主;而在经历一定理论、技术和经验的积累之后,有些项目开始尝试进行可实际运行的应用部署;目前从全球的 WA 发展来看,很多项目开始建立国家、区域战略合作保存体系。IIPC 对 38 家成员馆 2008 年的 WA 研究状态分析显示^[3]:以项目形式开展研究的有 6 家(15%)、以实验形式开展研究的有 7 家(18%)、可运作但仍处于实验阶段的有 11 家(29%)、完全投入应用的有 5 家(15%)、商业应用的有 4 家(10%)、其他形式的有 5 家(13%)。

WA 研究经历了十几年的发展,研究的内容和深度逐步拓展,WA 系统和工具不断得到应用和完善。运作较为成功的 WA 项目多采取将项目成果投入实际应用和商业化运作的服务模式。如 Hanzo Archives^[7]公司提供企业级存档服务,WebCite^[8]调动多方用户参与保存 Web 参考文献,IA 于 2005 年推出商业化应用服务——Archive-it^[9]。总体而言,目前 WA 研究仍以实验和项目研究的形式为主,真正投入实际运作和商业化应用的服务还较少。

1.3 WA 采集方式、策略逐渐多样化,采集内容范围不断扩大

Web 资源采集方式已从单一的 Web 资源一次采集发展到 Web 资源二次采集(Lazy Preservation)、数据库采集(深层网采集)和事务型采集等多种方式。采集策略也逐渐从基于主题、事件的选择性采集为主演变为混合策略、复杂域、大规模采集。采集内容从政治、社会文化、健康到艺术、人文,基本涉及人类的各个知识领域。

IIPC 2008 年对成员馆实施采集策略现状的调查显示^[3]:基于主题、事件的选择性采集占 52%,整个国家域采集占 21%,大规模采集占 11%,地区域名采集占 9%,其他形式的占 7%。澳大利亚、丹麦、加拿大采取国家域选择性采集方式;法国采取联合采集方式;瑞典采取全面自动采集方式;荷兰采取与出版者合作采集方式。澳大利亚国家图书馆已与 IA 合作进行过两次大规模域的采集活动,作为 PANDORA^[10]选择性存档计划的补充。法国国家图书馆(BnF)^[11]对经过选择的网站进行持续的自动采集,对于不能自动采集的深层网站进行人工采集,对于某些极易消失的网站进行基于事件的专题采集。基于主题事件域的采集主要针对具有社会、文化、政治意义的重大主题或事件进行专题采集。Minerva 项目^[12]与 IA 合作,对美国 2001 年总统大选、911 事件、2002 年冬季奥运会、107 届国会会议等事件实施专题采集。中国国家图书馆 WICP 项目^[13]已经完成对非典(SARS)、中国载人航天工程、2008 北京奥运会的专题存档。

1.4 逐渐完善 WA 系统技术、标准框架

WA 领域广泛接受并遵循 OAIS 模型。IIPC 提出了基于 OAIS 的 WA 系统通用概念框架,该框架覆盖了 WA 工作链中的所有过程,包括摄取(Ingest)、存储(Storage)、访问(Access)和索引与检索(Index & Search)四大部分,并开发了涵盖 WA 核心功能的完整工具集。IIPC 技术委员会下设的四个子委员会负责对 WA 的摄取、保存、访问、索引与检索进行深入的研究和实践工作。

标准规范方面,WA 同样关注数据层次相关标准规范和系统层次相关标准规范的研究,对 Web 资源采集、存储、访问、索引与检索阶段所涉及的标准规范不断予以研究完善。涉及的标准规范主要包括:获取阶段的存档资源标识(Archival Resource Key,简称 ARK)、统一资源命名(Uniform Resource Names,简称 URN)等数据唯一标识,多任务并发管理协议(HIP)、蜘蛛协议(Robert protocol);存储阶段的存档文件格式(Archive file format,简称 ARC)、Web 存档文件

格式 (Web Archive file format, 简称 WARC); 保存元数据实施策略 (Preservation Metadata: Implementation Strategies, 简称 PREMIS)、元数据编码与传输标准 (Metadata Encoding and Transmission Standard, 简称 METS)、元数据对象描述框架 (Metadata Object Description Schema, 简称 MODS) 等元数据标准; 索引与检索阶段的 CDX、BDB (Berkley Database) 索引结构, 以及开放档案信息系统 (Open Archival Information System, 简称 OAIS)、内容聚合 (Really Simple Syndication, 简称 RSS)、网络服务 (Web Service)、网络服务检索协议 (Search/Retrieve for the Web, 简称 SRW)、URL 服务检索协议 (Search/Retrieve URL Service, 简称 SRU) 和元数据开放搜寻及获取协议 (Open Archives Initiative Protocol for Metadata Harvesting, 简称 OAI-PMH) 等涉及 Web 层次互操作的标准规范。

同时,WA 领域非常注重协作共享。在系统应用层面,WA 在系统开发的初始阶段就十分关注系统的互操作,通过建立一个开放的模块化系

统框架和进行功能模块化开发,实现系统的开放性、协作性、互操作性。IIPC 提出了 WA 系统体系框架并开发了一系列的开源软件;Netarchive-Suite^[14] 工具包中的存储模块 (Archive Module) 支持分布式资源存储和不同软硬件平台的应用;FAST 公司为 NWA 开发的可扩展、分布式的检索工具很好地解决了大规模分布式资源的共享问题。在数据层面,IIPC 对格式规范、长期保存的元数据框架、永久标识符也进行了系统研究,它在扩展 ARC 格式基础上建立的 WARC (Web ARChive)^[15] 格式能更好地支持 Web 资源的采集、访问,可被 Heritrix^[16]、LibWARC^[17]、Wget^[18]、HTTrack^[19] 等多个资源采集器识别,使得采用不同采集器的系统和机构更容易进行资源共享。

1.5 WA 系统和工具得到发展与应用

伴随 WA 的不断发展,专门用于 Web 资源采集、索引、访问的系统和工具得到很好的发展,并不断得到改进和完善。目前 WA 领域大多数工具都是开源的,可以免费下载使用 (见表 1)。

表 1 常用的 WA 系统和工具 (* 表示为开源工具,按工具使用广泛性排序)

	工具/系统	开发者	工具/系统功能特色	项目来源/资助方
采集系统	PANDAS	澳大利亚图书馆	网络出版物采集的数字化存档系统	PANDORA
	WAS	加利福尼亚数字图书馆	基于 Web 的网络保存服务工具	Web At Risk
	Tumba	里斯本大学 XLDB 小组	提供葡萄牙 WA 搜索引擎服务	无
采集工具	Heritrix *	IA、北欧图书馆	可扩展的采集工具	IIPC 资助
	HTTrack *	法国图书馆	站点复制和离线浏览器工具	BnF Web Archive
	WCT *	Sytec Resources 公司	选择性网络采集的过程管理工具	IIPC 资助
	NetarchiveSuite *	丹麦皇家图书馆	综合性的网络存档软件包	Netarchive. dk
	Nutch *	Apache	开源的网络搜索引擎	Apache 子项目
	Smart Crawler *	英国图书馆、美国国会图书馆、法国图书馆、IA	智能爬虫	IIPC 资助
	DeepArc *	法国图书馆	基于数据库驱动的深层网页迁移工具	BnF Web Archive (IIPC 资助)
	Wget *	GNU	命令行工具、没有交互式界面	GNU 子项目
	NEDLIB crawler	芬兰科学计算中心	缺少用户界面,仅在北欧小范围使用	NEDLIB

续表

	工具/系统	开发者	工具/系统功能特色	项目来源/资助方
索引工具	NutchWAX *	Internet Archive 使用	全文本索引工具	IIPC 资助
	XTF *	加利福尼亚数字图书馆	索引与查询工具	IIPC 资助
	BAT	法国图书馆	ARC、DAT、CDX 管理文件的 API	IIPC 资助
访问工具	Xinq *	澳大利亚图书馆	XML 保存的查询工具	IIPC 资助
	Wayback *	Alexa 公司	访问工具	Internet Archive
	WERA *	IA、挪威国家图书馆	访问工具	Nordic Web Archive
	天网	北京大学网络实验室	Web 搜索引擎	Web Infomall

PANDAS^[20] 是澳大利亚 PANDORA 项目开发的基于采集的数字化存档系统,为国内参与合作的各个州立图书馆构建了合作者分布式使用的功能;WAS(Web Archiving Service)^[21] 是美国 Web At Risk 项目构建的基于 Web 的分布式仓储构建、存储和管理工具,对资源提供者、服务提供者和用户从服务层次上进行具体的规划。

IIPC 资助开发的系列软件 Heritrix、Nutch-WAX^[22]、WERA^[23] 得到了广泛的应用和推广。IA 与 Alexa 公司合作开发的 WA 访问工具 Wayback^[24] 已成为 WA 领域使用率最高的访问工具。NWA 与 IIPC 合作开发的 Web 采集过程管理工具 WCT^[25] 已成为应用率较高的保存工具。NetarchiveSuite^[26] 由丹麦皇家图书馆和丹麦州立大学图书馆联合研发,作为 Netarchive. dk^[27] 项目采集工具,自 2004 年以来已经采集了超过 70TB 的 Web 资源。

IIPC 对成员机构 WA 项目所使用的采集器、保存工具和访问工具的现状进行相关调查显示^[3]: 使用率高的保存工具包括 WCT (13%)、PANDAS (7%)、NetarchiveSuite (7%)、WAS(7%);访问工具以 Wayback 为主 (47%), 其次是 WEAR(7%)。

WA 领域的工具、系统有趋同的发展趋势。几家具有较强技术实力的机构开始合作开发和完善现有的软件和工具。目前,IIPC 与合作机构正在研发新一代智能爬虫 (Smart Crawl-

er)^[28],以提高爬虫自动采集的效率。

1.6 WA 法律政策制定不断完善

目前已经有 100 多个国家制定了图书等出版物呈缴法律法规,也已经有一些国家将数字资源(主要是数字出版物)纳入呈缴范围之中。澳大利亚、英国、挪威、瑞典、荷兰、阿根廷等国制定了数字资源的呈缴规定或签署了相关协议,从法律上保证数字资源长期保存的可靠性;加拿大、法国、奥地利、瑞典等国将数字出版物纳入呈缴之列,但都没有明确将 Web 信息资源纳入到呈缴法。目前,仅有挪威、丹麦制定了 Web 资源呈缴法,并对呈缴的责任者、内容、方式、格式作出相应规定。为防止出版商因停业、倒闭等可能出现的因素导致 Web 资源无法保存,大多数国家希望在 Web 信息产生 30 日之内实现呈缴。IIPC 2008 年年会的数据显示^[3]: 在 IIPC 的 38 个成员国家中,已经制定或者通过有关数字资源呈缴法律的有 13 个 (36%),即将通过的有 6 个 (17%),没有制定相关法律的有 19 个 (47%)。

目前部分尚未制定 Web 资源呈缴法律的国家也在采取与出版者和版权局协商、发布版权声明、立法试验等方式积极完善 WA 的法律。发布版权声明是目前 WA 项目使用较多的解决法律问题的方法之一,主要有采集前征求出版者许可 (opt-in)、允许出版者提出剔除请求 (opt-out) 和混合型的解决方案。PANDORA 项目采

取“opt-in”方案,规定所有采集对象在采集前都要征得所有者的许可;IA 只采集可公开获取、不受“Robots.txt”文件保护的网站,同时提出法律声明,允许出版者提出剔除请求。

1.7 WA 所需投入较大,成本分析及风险管理研究日益受到重视

由于 WA 对存储设备、技术要求高,花费时间长,人力成本高,人员专业性要求高,因此 WA 项目需要大量的资金投入,且呈逐年上升的趋势。荷兰国家图书馆 2005 年的资金投入接近于 2004 年的 3 倍,2006 - 2009 年继续增加预算。由英国 JISC (Joint Information Systems Committee)资助、伦敦大学学院 (University College London,简称 UCL)和英国国家图书馆联合承担的 LIFE 项目^[29]对英国网络资源保存联盟 (UK Web Archiving Consortium,简称 UKWAC)的长期保存总体费用估算从 2004 年项目之初的 493,169 英镑,增至 5 年后的 915,219 英镑,20 年后预计达到 1,617,541 英镑。

WA 项目的资金主要来源于国家图书馆、政府部门及基金组织的资助。目前澳大利亚、荷兰已经将 WA 纳入国家图书馆的业务开支,意味着 WA 已成为图书馆的正常业务,获得稳定的资金支持。从目前 WA 项目经费使用上看,经费主要用于基础设施建设 (主要是技术方面)、人员费用及日常业务支出,其中基础设施建设的投入所占比例较高。IIPC 2007 年工具开发投入了 17,625 美元,占总投入的 37%;2008 年,投资增长到 178,364 美元,占总预算的 78%^[29]。

为了持续进行 WA 研究,对费用成本及风险管理的研究也日益受到关注。目前主要有 LIFE 项目以及 NASA 用于研究太空、地球数据保存的费用估计工具 CET^[30]等,但尚未建立专门针对 WA 的成本费用模型。另外,著名的数字资源长期保存专家尼尔·比格利 (Neil Beagrie)、朱莉娅·克鲁兹 (Julia Chruszcz)、布莱恩·拉沃伊 (Brian Lavoie)等人对英国 4 所大学的考古典藏中心开展数据长期保存研究的花费情况进行了调研,为制定一个详细的费用框架打下

了基础^[31]。

1.8 WA 合作范围扩大,合作机制加强

WA 项目需要政策、经济、法律、技术、管理等多方面的支持来协调一系列社会利益和利用足够的社会资源,单靠某一个组织的力量很难完成如此庞大的工程。目前国际上 WA 项目大多数是由多个机构合作完成,并在合作过程中呈现出多种合作形态,合作已成为 WA 活动发展的趋势。

从 WA 合作范围来看,从国家内部、组织间的合作发展到地区合作,再到国际合作。瑞典国家图书馆 Kulturarw3 项目^[32]、Web At Risk 项目^[33]是国家内部合作;欧洲的 NEDLIB 项目^[34]、北欧的 NWA 项目^[35]是地区间的合作;IA、IIPC 是 WA 组织机构国际合作的典型;PANDORA 同时采取多种合作形式。从 WA 合作机制来看,主要分为高度集中机制、责任平等机制、高度分散机制和独立工作机制。PANDORA 项目是高度集中合作机制的典型代表;IIPC 是责任平等合作机制;IA 是高度分散的合作机制;Kulturarw3 项目是独立工作机制。从合作层面上来看,有基于存储、技术、战略、工具的合作模式。SDSC 的 Chronopolis^[36]框架是基于网格实现大规模存储;PANDORA 项目与各个州立图书馆采取基于采集的协作方式;NWA 项目通过工具的合作研发,建立起欧洲网络信息资源保存的合作机制。

权责明确的责任体系、明确的合作目标、实现技术共享以及保持良好的沟通是影响 WA 合作的关键因素。构建 WA 的合作框架时需要根据项目的实际情况选择合适的合作机制;组织机构也要根据该项目的合作机制是否适合自身情况来选择要参与的合作项目。

2 国内 WA 发展现状分析

我国 WA 的研究始于 20 世纪末。随着数字资源使用的日益普及,长期使用的问题日益突出。由此,WA 的必要性和重要性日益受到重视,并开展了一系列相关的理论研究和具体

实践。

我国对 WA 比较系统的理论研究主要集中在中国科学院国家科学图书馆和少数其他单位。国家科学图书馆一直以来关注数字资源长期保存的宏观支撑机制和问题框架的研究,目前正在进行的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”,对 WA 的理论、技术予以探索研究。国家图书馆和高校的一些研究人员对元数据方案、服务模式、WA 面临的问题进行了探讨。

在积极开展国际跟踪和理论研究的同时,一些文献情报机构从 2002 年起开展了 WA 的试验项目(WICP)采集保存了自 2003 年以来 .cn 域名下的网站和所有中文(编码)网站,积累 Web 数据达 150G;对政府网站、电子报刊、国学的 Web 资源进行镜像存档;选择了 2008 北京奥运、中国载人航天工程等专题进行专题存档,并对专题存档的数据进行质量控制、数据挖掘。国家图书馆已经成为 IIPC 的成员,并积极推动和促进 Web 资源呈缴法的起草,以解决 WA 长远发展过程中的法律障碍。中国 Web 信息博物馆(Web Infomall)^[37]是在国家“973”和“985”项目支持下,由北京大学“计算机网络与分布式系统实验室”主持开发的中国网页历史信息存贮与展示系统,采集和保存了自 2002 年以来 30 亿页的中文 Web 资源。该项目在 WA 应用方面进行了探索,提供历史网页检索、真实呈现等服务,通过数据挖掘追踪重大历史事件发展进程的全貌,有力地支撑了社会科学的研究。但由于我国还没有制定 Web 信息合法呈缴的相关法律,这些项目均采用版权声明方式暂时解决 WA 的法律问题。

与此同时,我国研究人员还积极参与国际交流,国家科技图书文献中心(NSTL)与国家科学图书馆于 2004 年、2007 年两次承办“数字资源长期保存国际会议”(iPRESS),为国内保存领域的研究人员参与国际长期保存合作,促进长期保存的可持续高水平发展提供了良好的机会。

3 WA 发展趋势分析

3.1 WA 研究重点及薄弱环节

目前,WA 项目和系统从各个角度及不同层面研究 WA 所涉及的技术、政策、法律问题,获得了很多的经验和教训,取得了良好的进展。但笔者也注意到,当前的 WA 研究中还存在一些薄弱环节。

采集方面,采集的策略、流程、工具一直是 WA 研究的重点。目前许多 WA 项目已具备较为完整的采集流程、采集策略,Heritrix、HT-Track、Web Curator 等采集工具在各个项目中得以广泛应用,但在深层网采集、增量采集、流媒体采集、采集工具效率等方面的研究还有待更深入的探索。

存储和长期保存方面,WARC 的出现标志着 WA 存储格式趋于统一和规范,部分 WA 项目通过采用 XML、HTML 格式规范来实现 Web 资源的长期保存,但 Web 资源长期保存的其他方面还没有更进一步的探索实践。另外,海量数据的有效存储及易扩展的存储体系结构都是 WA 领域极具吸引力的课题。

在开展一系列 WA 的采集、长期保存实践后,提高 Web 资源的检索效率、更好地实现用户呈现,特别是实现大规模数据索引、访问、检索、质量控制、数据挖掘、智能检索等问题的研究也显得尤为重要。在刚刚结束的 2008 年 IAWW(International Web Archiving Workshop)会议上,访问已成为议题的焦点。同时,利用数据挖掘、语义技术对学术研究动态进行追踪等 WA 长远发展问题也逐渐获得更多关注。

技术方面,WA 的技术框架、标准规范得到了很好的发展和共享,开发了大量(开源)工具和系统,并对一些技术难点、关键点的性能和效率不断进行完善,旨在改善工具的强健性、灵活性和可维护性,其投入也逐年增加。目前 IIPC 正在开发的智能爬行器旨在加强对深层网络采集的效率;WCT 规范了采集流程管理和采集质量审核机制,旨在提高采集效果。

由于目前大部分国家还没有建立 Web 资源

呈缴法,在一定程度上影响 WA 采集的效率和成本,使得 WA 在采集、提供访问上受到一定阻碍。此外,目前还没有专门针对 Web 的成本和风险管理的模型,如何估算和控制 WA 各个环节的成本、资金投入以保证 WA 长期可持续发展,还需要深入的研究和实践。

3.2 未来 WA 发展趋势分析

(1) 主题和内容丰富化

WA 保存的内容和主题日益丰富,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。目前已有机构开始 Web2.0 网站采集的研究实践。WA 的内容管理,特别是保存资源评价、质量控制方面日益受到重视。

(2) 系统建设标准化和开源化

WA 项目在标准规范方面不断改进, 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。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、规范化方面还在不断努力和探索。

(3) 工作流程规范化

随着 WA 十余年的发展,项目在工作流程规范化、提高工作效率方面的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。

(4) 合作广泛化

WA 领域的合作范围不断扩大,合作内容不断深化。利用分布式的系统和资源构建网格和协作网络,共享 WA 系统和资源,促进资源共享,实行职责与费用分摊。

(5) 利用形式多元化

除了网站恢复、保存 Web 文献参考链接信息等 WA 利用形式外,WA 领域还运用多种技术集成的手段来追踪 Web 发展演变,用于科学研究的支持决策。美国弗吉尼亚多米尼恩大学计算机学院、康奈尔大学、日本东京大学、北京大学 Web InfoMall、欧盟资助的 LiWA (Living Web Archives) 项目^[38]运用数据挖掘、知识抽取、本体语义等技术进行了 WA 领域未来多元化利用形式的前沿探索与研究。

参考文献:

- [1] Internet Archive [EB/OL]. [2008-11-09]. <http://www.archive.org/index.php>.
- [2] IIPC [EB/OL]. [2008-11-12]. <http://www.net-preserve.org/about/index.php>.
- [3] Abbie Grotke. IIPC Membership Report [R]. Canberra: IIPC General Assembly Meeting, 2008.
- [4] Alexa [EB/OL]. [2008-11-12]. <http://www.alexa.com/>.
- [5] FAST [EB/OL]. [2008-11-12]. <http://www.fastsearch.com/>.
- [6] IBM [EB/OL]. [2008-11-12]. <http://www.ibm.com/>.
- [7] Hanzo Archives [EB/OL]. [2008-11-12]. <http://www.hanzoarchives.com/>.
- [8] WebCite [EB/OL]. [2008-11-12]. <http://www.webcitation.org/>.
- [9] Archive-it [EB/OL]. [2008-11-12]. <http://www.archive-it.org/>.
- [10] PANDORA [EB/OL]. [2008-11-12]. <http://pandora.nla.gov.au/>.
- [11] Bibliothèque nationale de France [EB/OL]. [2008-11-12]. <http://www.bnf.fr/>.
- [12] Minerva [EB/OL]. [2008-11-12]. <http://www.loc.gov/minerva/>.
- [13] WICP [EB/OL]. [2008-11-12]. <http://webarchive.nlc.gov.cn>.
- [14] NetarchiveSuite [EB/OL]. [2008-11-12]. <http://netarchive.dk/suite>.
- [15] WARC [EB/OL]. [2008-11-12]. <http://www.digitalpreservation.gov/formats/fdd/fdd000235.shtml>.
- [16] Heritrix [EB/OL]. [2008-12-24]. <http://crawler.archive.org/>.
- [17] LibWARC [EB/OL]. [2008-12-24]. <http://code.google.com/p/warc-tools/issues/list>.
- [18] Wget [EB/OL]. [2008-12-24]. <http://en.wikipedia.org/wiki/Wget>.
- [19] HTTrack [EB/OL]. [2008-12-24]. <http://www.httrack.com/>.
- [20] PANDAS [J/OL]. [2008-12-24]. <http://www.nla.gov.au/nla/staffpaper/2003/mphillips1>.

- ml.
- [21] Web Archiving Service [EB/OL]. [2008-12-24]. http://www.cdlib.org/inside/projects/preservation/webatrisk/web_archiving.html#information
- [22] NutchWAX[EB/OL]. [2008-12-24]. <http://archive-access.sourceforge.net/projects/nutch/>.
- [23] WERA [EB/OL]. [2008-12-24]. <http://archive-access.sourceforge.net/projects/wera/>.
- [24] Wayback [EB/OL]. [2008-11-12]. <http://www.archive.org/web/web.php>.
- [25] WCT[EB/OL]. [2008-11-12]. <http://webcurator.sourceforge.net/>.
- [26] NetarchiveSuite [EB/OL]. [2008-11-12]. <http://netarchive.dk/suite>.
- [27] Netarchive.dk [EB/OL]. [2008-11-12]. <http://netarchive.dk/index-en.php>.
- [28] smart crawler[EB/OL]. [2008-11-12]. <http://crawler.archive.org/>.
- [29] LIFE Project final report [EB/OL]. [2008-11-12]. <http://eprints.ucl.ac.uk/1854/1/LifeProjectMaster.pdf>.
- [30] Cost Estimation Toolkit. [EB/OL]. [2008-11-12]. <http://opensource.gsfc.nasa.gov/>.
- [31] Keep research data safe: A cost model and guidance for UK university [J/OL]. [2008-11-12]. <http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf>.
- [32] Kulturarw3 Project [J/OL] [2008-11-12]. <http://www.ifla.org/IV/ifla66/papers/154-157e.htm>.
- [33] Web At Risk Project[EB/OL]. [2008-11-12]. <http://web3.unt.edu/webatrisk/>.
- [34] NEDLIB [EB/OL]. [2008-11-12]. <http://nedlib.kb.nl/>.
- [35] NWA [EB/OL]. [2008-11-12]. <http://nwa.nb.no/>.
- [36] SDSC Chronopolis[EB/OL]. [2008-11-12]. <http://chronopolis.sdsc.edu/premis.html>.
- [37] Web Infomall[EB/OL]. [2008-11-12]. <http://www.infomall.cn/>.
- [38] LiWA[EB/OL]. [2008-11-12]. <http://www.li-wa-project.eu/>.

向 菁 中国科学院国家科学图书馆2007级硕士研究生。通讯地址:北京市北四环西路33号。邮编100190。

吴振新 中国科学院国家科学图书馆副研究馆员。通讯地址同上。

(收稿日期:2008-12-08;修回日期:2008-12-25)