国际主要 Web Archive 项目介绍与评析

向 菁 吴振新 司铁英 曲云鹏 李成文

摘 要 从国家、联盟、项目三个层面对典型网络资源保存项目和网络资源保存联盟进行介绍和分析,为我国网络资源长期保存研究和实践提供参考。表 1。参考文献 19。

关键词 网络资源 长期保存 web archive

1 引言

网络信息资源(简称 Web 资源)作为全球最大的信息资源库,不但是重要的人类文化遗产,对学术研究也具有非常重要的价值。各国图书馆、档案馆、文化遗产机构等纷纷投入网络信息资源保存(Web Archive,简称 WA)研究活动,在全球范围内开展了不同目的、不同形式的众多 WA,对 Web 资源的采集、存储、访问的整个生命周期的各个环节进行研究分析,涉及保存系统的研发、保存风险研究、标准规范、权益等等,并获得了一系列的成果。但由于 Web 资源自身具有海量、复杂多样、质量不一、技术依赖性强等特点,使得 WA 研究与实践也面临着巨大的困难和挑战。

2 Web Archive 项目分类及特点研究

近年来国外很多与文化遗产保存相关的机构都在开展 WA 的研究工作,已完成或正在推进的全球大大小小的项目已达近百项。这些项目从不同的来源获得研究资助,对 WA 活动中的策略、技术、方法、法律、合作等方面进行了不同程度的研究。综观主要的 WA 项目,大致可分为国家层面、组织联盟、项目三大类别。

2.1 国家层面 WA 项目特点及案例研究

从项目的参与主体上来看,国家层面的 WA 项目组织结构模式主要以国家级的图书馆为领导者,广泛吸纳国内其他组织机构参与。从项目的使命来看,国家层面 WA 项目是为了保存国家领域的 Web 资源,多是采用完整性保存策略,进行大规模基于国家域的网络采集活动,通常采用自动爬行的采集方式。该类项目的主要资金通常来自负责的国家图书馆或由国家图书馆从政府、其它基金组织获得资助。该类项目一般还负责国家 WA 基础设施的建设,负责机构通常需要为其它参与者提供技术和系统上的支持。该类项目工作内容比较宽泛,会比较完整地涉及与 WA 相关的各项工作,除了系统的研发,还负责国家保存计划、存档原则及相关政策、策略的制定,协调各方关系,特别是在解决版权问题方面能够发挥重要作用,如争取从法律制度上将网络出版物纳入国家呈缴法,与资源所有者签署保存协议等。国家图书馆通常作为代表负责开展国际交流合作。

2.1.1 澳大利亚国家图书馆 PANDORA

澳大利亚国家图书馆发起和领导了 PANDORA 项目[1],与州图书馆及其他学术机构等 10 个合作伙伴构建全国范围的合作框架,并代表澳大利亚与 IA、IIPC 等国际联盟开展合作,参与和促进世界范围 WA 研究。澳大利亚国家图书馆承担了项目的主要费用,合作机构只负责本机构的运行经费。此外,澳大利亚国家图书馆制订了国家数字保存政策,提供图书馆保存自有资源、网站和联机出版物采集选择、如何与其它机构合作来获得保存最大收益的指导,与出版商达成合作协议,研发的 PANDAS 系统提供给成员使用并提供技术支持。

2.1.2 英国网络信息保存计划 UKWAC

大英图书馆组织,与英国国家档案馆、联合信息系统委员会(JISC)、苏格兰国家图书馆、威尔士国家图书馆、韦尔科姆图书馆等 5 所权威机构组成联盟,于 2004 年 6 月正式启动英国第一个公众网络信息保存计划——UKWAC[2],对英国网站信息进行选择

性保存。该项目联盟成员共同分担 WA 的成本、风险,分享经验和软硬件设施,采用澳大利亚国家图书馆开发的 PANDAS 系统,对已保存站点内容提供免费检索服务。大英图书馆还与 IIPC、IA 等网络资源长期保存机构合作进行 WA 保存工具的研发。

2.1.3　法国国家图书馆 BnF Web Archive

法国国家图书馆(BnF)早在 1999 年就开始进行 Web Archive 试验项目[3],重点关注深层网络资源采集。它采用资源自动采集和人工干预相结合的采集策略,对法国国内网络资源进行保存,并利用具有国际合作研发技术优势积极参与国际合作,与 IIPC、美国国会图书馆、大英图书馆联合开展的智能爬虫(smart crawler)、机构仓储的相关研究,所有成果均以开源软件的形式公布。

除了上述介绍的项目外,美国国会图书馆 Minerva Prototype 网络信息保存试验[4]、瑞典斯德哥尔摩皇家图书馆 Kulturarw3[5]、日本国立图书馆的 WARP 项目[6]、新加坡图书馆理事会资助的 Web Archive Singapore 项目[7]、中国国家图书馆 WICP[8] 等项目都是国家层面网络资源保存项目的典型代表。

2.2　联盟形式的 WA 项目特点及案例研究

联盟形式的 WA 项目组织结构模式主要以 WA 机构组织建立国家或区域战略合作保存体系为依托,进行广泛的研究合作或在一定范围内进行规模化应用部署。项目的使命是鼓励和支持各国进行 Web 资源保存,积极推动全球 WA 的研究发展。该类项目形成以图书馆为主,档案馆、文化遗产保护机构等相关组织机构广泛参与的格局,在一定的国际或区域战略合作的框架和相关机制下,成员机构之间共同分担经费、分享技术观点和经验,在标准制定、资金管理、人员培训、技术研发等方面具有规模化发展的优势。从目前的发展来看,联盟形式的 WA 项目发展迅速,成效卓越,对全球的 WA 活动起到了非常重要的推动作用。

2.2.1　Internet Archive(IA)

IA[9] 是非赢利性的组织,该组织 1996 年的成立标志着 WA 研究的兴起。它采取高度分散合作机制与美国国会图书馆、IIPC 等组织机构保持着密切合作,并为全球很多机构提供技术和数据的支持。它定期采集保存全球网站可抓取信息,提供开放访问服务。IA 与 Alexa 公司合作开发的 WA 访问工具 Wayback Machine 是目前 WA 领域使用最多的访问工具,对全球的 WA 发展起到了积极的促进作用。

2.2.2　Nordic Web Archive (NWA)

1997 年,丹麦、挪威、芬兰、冰岛和瑞典 5 个国家的国家图书馆联合启动 NWA 项目[10],主要目标是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,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。该项目研发了多个开源的功能组件,在联合采集的基础上实现了大规模跨域资源的 web archive 访问,并以此为基础积极参与 IIPC 的相关系统开发,在 WA 系统架构和技术方法的发展上发挥了重要作用。

2.2.3　International Internet Preservation Consortium (IIPC)

2003 年成立的 IIPC[11] 在推动国际 Web Archive 研究方面起着非常重要的作用,它采用责任平等的合作机制,鼓励世界范围内的文化遗产保护机构一起参与网络信息资源保存的工作,目前已汇集了全球多个国家的 38 个国家图书馆、档案馆等保存机构。其指导委员会负责进行成员间协调、沟通、财务管理、技术讨论,成员机构共同分担经费、分享技术和经验。该联盟在 WA 系统构架、标准规范、元数据等方面建立一系列技术规范,并资助其成员开发了网络资源采集到提供访问服务的一系列高质量、易用的开源软件工具,包括选择性网络 WCT、Heritrix、DeepArc、Smart Crawler 等采集工具, NutchWAX、XTF、BAT 等索引工具,WERA、Xinq 等访问工具。

2.3　项目形式的 WA 项目特点及案例研究

项目形式的 WA 参与主体极其广泛,包括图书馆、档案馆及相关的文化遗产保存组织机构。该类项目多为基金资助项目,由一个机构独立进行,在一定时期内,对 WA 领域的某个问题或者从某个角度进行探索和实践,项目通常持续时间不长,项目规模有限,采取选择性采集方式为主策略,主要是进行实验性的探索工作。该类项目研究内容涉及广泛,关注 Web 资源的存储、呈现、利用、WA 系统开发等方面的研究。这些项目的研究和实践常常为后续的研究项目奠定基础,进而形成和开展更大范围、更深层

次的 WA 研究。通常是由一个机构独立进行,独自发展政策、系统开发和研究等。该类项目通常都会取得良好的进展,但难以进行规模化的发展,因此如果有合适的潜在合作伙伴时,项目都会积极寻找合作的机会,从而寻求更大的发展。

2.3.1 Web at risk

Web at risk 项目[12]是美国 NDIIPP(National Digital Information Infrastnuture and Presenation Program,美国教学信息保存计划)资助、运作较为成功的数字保存项目之一,开发了一套能根据管理者所在的环境创建保存计划、提供向导、保存网页[13]各个版本的 Web Archiving Service(WAS)系统,建立一套完善、规范的工作流程,实行了跨部门、跨组织的合作机制。

2.3.2 Kulturarw3

瑞典 1996 年开始了名为 Kulturarw3[5]的网络信息资源收集试验项目。该项目以瑞典网络信息资源为对象制定了完整性收集的策略,通过网络机器人进行数据收集试验并开始提供公共服务。Kulturarw3 项目对采集策略和采集工具进行了全面的研究,并成为随后开展的北欧图书馆联盟 NWA 项目的基础。

2.3.3 Domain. UK

2001 年英国国家图书馆启动 WA 试验性项目 Domain. UK[2],探索围绕在网络存档周围的一系列问题。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。该项目还保存了 2001 年大选、口蹄疫等网页信息。Domain. UK 为图书馆未来网络信息资源长期保存发展奠定基础。

其他运作较为成功的 WA 项目还有美国圣地亚哥超型计算机中心(SDSC)负责、美国自然科学基金资助的 Chronopolis 项目[15],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;葡萄牙里斯本大学开发的 Tumba 搜索引擎[16]关注大规模网络资源不同时间点不同版本的"原貌"呈现;德国马普学会计算机研究院开发的 YAGO 搜索引擎[17]实现 web 环境下大规模网络资源基于本体的语义搜索;东京大学知识循环社会信息教育研究中心研发,MEXT 资助研发的基于 WA 数据抽取可视化显示网页历史内容的搜索引擎 Page History Explorer[18],可挖掘网页的历史信息,以时间云图的可视化方式呈现历史网页信息的演变过程和内容。

3 主要项目比较和评析

随着数字资源长期保存相关技术的发展,WA 已成为欧美发达国家的研究热点之一。从以上介绍的项目看,国外已投入大量的人力和物力进行 WA 项目的相关研究。下面对国际主要网络信息长期保存项目从负责机构、项目启动时间、合作机构、开发工具/系统等方面进行分析。

表 1 主要网络信息长期保存项目比较(按字母顺序排列)

项目名称	性质	牵头负责机构	启动时间	合作机构	开发工具/系统	合法呈缴
BnF Web Archive	国家层面	法国国家图书馆	2001	IIPC、LC、英国图书馆	HTTrack,IIPC 工具包	有
Chronopolis	项目形式	SDSC 圣地亚哥超型计算机中心	1998	美国自然科学基金 NSF 资助	SRB 及开源版本 iROD	无
UKWAC	国家层面	英国国家图书馆	2001	JISC 等 5 所权威机构、IIPC、IA	PANDAS,IIPC 工具包	无
IIPC	联盟形式	国际联盟组织	2003	37 个成员机构	IIPC 工具包	
Infomall	项目形式	北京大学	2002	北京大学计算机网络与分布式系统实验室	中国网页历史信息存贮与展示系统	无

续表

项目名称	性质	牵头负责机构	启动时间	合作机构	开发工具/系统	合法呈缴
Internet Archive	联盟形式	美国非盈利性组织	1996	Alexa Internet、美国国会图书馆	Wayback Machine	无
Minerva	国家层面	美国国会图书馆	1997	IA、Pew Internet & American Life	HTTrack, CORC 软件, Wayback machine	无
Nordic Web Archive	联盟形式	挪威、瑞典、丹麦、冰岛、芬兰国家图书馆	1997	FAST、IIPC	NWA 工具包	有
Pandora	国家层面	澳大利亚国家图书馆	1996	国内 9 所图书馆、文化机构	PANDAS	无
WARP	项目形式	日本国会图书馆	2002		Heritrix(IIPC)	无
Web Archive S ingapore	项目形式	新加坡国家图书馆	2005	新加坡图书馆理事会	IIPC 工具包及自有系统	无
Web at risk	项目形式	美国国会图书馆	2004	加利福尼亚数字图书馆、北得克萨斯大学、纽约大学	Web Archiving Service (WAS)	无

(1)从项目牵头负责的机构来看,WA 项目研究者主要由国家级图书馆、联盟组织(IA、IIPC)、专业研究机构(如 SDSC)组成,其中国家级图书馆是 WA 项目负责实施的主体,在网络资源保存中发挥着主导作用。国家图书馆作为国家层面 WA 领导者,比较容易争取政府在法律、政策方面的支持和政府、基金组织的资助,在代表国家寻求多方合作、构建项目国际合作的框架和体系具有优势。

(2)从项目的开展上看,WA 项目研究始于 20 世纪 90 年代末,欧美的图书馆、文化、研究机构及相关组织陆续开始此方面的研究,迄今为止已经积累了十几年的技术经验和相关的人才储备,很多项目逐步进入成熟化的运作阶段。相比而言,亚洲在 WA 研究起步较晚,日本国立图书馆开展的 WARP、新加坡图书馆理事会资助的 Web Archive Singapore 以及中国国家图书馆和北京大学都在进行相应的研究,努力推进亚洲 WA 研究进展,但在标准化、知识产权以及系统的研发,特别是国际国内合作等方面还是有许多待完善的地方。

(3)从项目合作来看,由于 WA 所涉及的技术、资金、管理、人力问题的复杂性决定其不能单靠某一个机构力量完成,建立国际性、区域性合作框架,实行跨国家、跨区域、跨组织、跨部门的合作是 WA 项目主流趋势,通过合作来有效利用资源、分担责任、降低风险、最大获益。欧美国家在此方面进行了有益的尝试和探索,如 Pandora 建立基于采集的合作模式;IIPC 构建了基于工具开发的国际合作框架;SDSC 建立基于网格存储的合作框架。

(4)从 WA 的技术角度来看,有两个主要特点:需要专有工具和系统;系统工具大部分是开源系统。目前的 WA 已经形成了较好的模块化体系架构,采集、存储、索引、访问等环节都提供了开源的专门模块和工具,随着 WA 研究不断深入,这些工具组件的开发也在不断的予以完善提高。同时出现了一些遵循 OAIS 模型的投入实际服务的长期保存系统,以多种方式对网络信息资源进行长期保存。如 PANDAS 系统[19]、LOCKSS[20] 系统、e-Depot 系统[21]。

(5)从合法呈缴方面来看,目前全球范围只有个别国家已经有了相关法律规定,实际上仅有丹麦、新西兰明确提出了可以采集网络资源,其他国家呈缴范围仅限于数字出版物。法律法规的不完善为各国网络信息资源保存活动的发展造成了一定的阻碍,建立网络信息资源的合法呈缴制度是各国保存者积极努力的方向。

4 结语

通过分析国际主要 WA 项目,可以为我国从不同层面开展网络资源长期保存活动提供借鉴意义。

(1)建立权责明确的国家层面的 WA 责任体系,国家级的图书馆作为 WA 项目的参与主体,应承担主要责任,鼓励和带动相关机构进行 WA 研究,积极从政府和各种基金寻求更多的资金资助和政策支持。

(2)构建良好的合作机制是推动我国 WA 项目发展的有力保障。构建国际或地域范围内的协作网络,在采集、工具开发、存储等方面进行合作,共同分担保存中国网络资源的责任和风险,积极加入国际联盟组织(IA、IIPC),共享全球 WA 的成果。

(3)完善我国有关 WA 的相关法律法规建设,尽快将 Web 资源纳入合法呈缴的范围,采集和保存具有历史、文化、研究价值的 Web 资源为用户利用。

(4)在技术方面,应遵循国际 WA 的统一标准,建议采用国际 WA 项目常使用的开源工具、系统,并根据自身项目情况进行改进和完善。在经费和能力允许的情况下,应积极参与国际联盟的工具和系统的研发活动。

从以上对国外 WA 项目的介绍和评析可以看到,随着长期保存技术的逐渐成熟,WA 呈现主题日益丰富化、遵循统一的架构和标准规范、工作流程规范化、采用开源系统为主要工具、建立规范化合作框架的发展趋势。

参考文献

1 PANDORA[EB/OL]. [2008 - 06 - 06]. http://pandora. nla. gov. au/.

2 UKWAC[EB/OL]. [2008 - 06 - 23]. http://www. webarchive. org. uk/.

3 Archiving the web:experiments at BnF[J/OL]. [2008 - 07 - 23]. http://www. dpconline. org/graphics/events/presentations/pdf/Masanes. pdf.

4 Minerva Prototype[EB/OL]. [2008 - 06 - 13]. http://www. loc. gov/minerva/.

5 Kulturarw3 Project[J/OL]. [2008 - 07 - 13]. http://www. ifla. org/IV/ifla66/papers/154 - 157e. htm.

6 Web Archiving Project[EB/OL]. [2008 - 07 - 28]. http://warp. ndl. go. jp.

7 Web Archive Singapore[EB/OL]. [2008 - 08 - 02]. http://was. nl. sg/.

8 WICP[EB/OL]. [2008 - 07 - 16]. http://webarchive. nlc. gov. cn.

9 Internet Archive[EB/OL]. [2008 - 08 - 09]. http://www. archive. org/index. php.

10 Nordic Web Archive (NWA)[EB/OL]. [2008 - 07 - 03]. http://nwa. nb. no/.

11 International Internet Preservation Consortium[EB/OL]. [2008 - 08 - 10]. http://www. netpreserve. org/about/index. php.

12 Web at risk[EB/OL]. [2008 - 08 - 13] http://www. cdlib. org/inside/projects/preservation/webatrisk/.

13 Chronopolis[EB/OL]. [2008 - 08 - 14]. http://chronopolis. sdsc. edu/.

14 Tumba[EB/OL]. [2008 - 08 - 15]. http://www. tumba. pt.

15 YAGO[EB/OL]. [2008 - 08 - 17]. http://www. mpi-inf. mpg. de/ ~ suchanek/downloads/yago/.

16 Adam Jatowt, Yukiko Kawai. Visualizing Historical Content of Web Pages[J/OL]. [2008 - 08 - 18]. http://www2008. org/papers/pdf/p1221 - Jatowt. pdf.

17 PANDAS[EB/OL]. [2007 - 11 - 3]. http://pandora. nla. gov. au/pandas. html.

18 LOCKSS[EB/OL]. [2008 - 08 - 18]. http://www. lockss. org/.

19 e-Deport[EB/OL]. [2008 - 08 - 19]. http://www. kb. nl/dnp/e-depot/e-depot-en. html. .

(向 菁 中国科学院国家科学图书馆 中国科学院研究生院 2007 级硕士研究生,吴振新 中国科学院国家科学图书馆,司铁英 曲云鹏 李成文 国家图书馆)

收稿日期:2009 - 05 - 15