中国图书馆学报(双月刊)2004 年第 3 期
ZHONGGUO TUSHUGUANXUEBAO

●赵俊玲

国外关于网络信息资源保存的研究

摘 要 很多国家认识到保存网络信息资源的紧迫性,为此进行了一系列研究和实验。今后的研究重点将是关于保存策略、保存机构间的合作模式等。参考文献 13。

关键词 网络信息资源 长期保存 保存策略 保存研究
分类号 G253

ABSTRACT In this paper, the author introduces the researches and practices of the preservation of network information resources in foreign countries, and thinks that we should pay our attention to the following aspects in our future researches: preservation strategy, cooperation among preservation institutions, etc. 13 refs.

KEY WORDS Network information resources. Long-term preservation. Preservation strategy. Preservation research.
CLASS NUMBER G253

网络信息资源日渐丰富,然而它在呈指数增长的同时,消失的速度也很惊人。和其他物理形态载体的信息相比,网络信息相对来说生命更脆弱,更容易破坏,一个内容丰富、规模庞大的网站用一个命令就可以删除。尽管现在对网络信息寿命没有确切的数据,但一些相关机构和学者已作出估计:Internet Archive 的创始人 Brewster Kahle 在 1996 年就曾经估计"网页产生 75 天后就会消失"[1]。再后来美国数字信息基础架构和保存项目(NDIIPP)的报告中指出网络信息的平均寿命为 44 天[2]。相当一部分有价值的学术、文化和科研方面的网络信息资源面临着消失的危险。特别是一些时政性质的网络信息,比如说在澳大利亚,和 2000 年悉尼奥运会相关的很多网络信息资源已经消失。对网络信息进行保存显得十分迫切。

1 网络信息保存面临的挑战

1.1 技术

网络信息资源保存必须解决所有数字文献保存面临的技术过时问题。尽管从上个世纪 90 年代以来,一些基础的网络协议和标准相对稳定,但是在网站管理方面却有很大的变化,比如说越来越多的网络信息是通过动态数据库发送,因此在保存网络信息内容的时候还要保存相应软件和硬件方面的元数据信息。同时还面临网络信息资源自身特点所带来的一些问题,比如网络信息不是孤立存在的,而是相互联系的,这样就导致网络信息保存对象的边界很难限定。网络信息的完整性和真实性如何界定?如何通过网页重要程度和预测网页更新周期来确定抓取周期?如何抓取动态网络信息资源?这些都是需要解决的技术问题。

1.2 责任体系

到底应该由哪个机构或哪些机构来负责网络信息资源的长期保存?学术界对这个问题看法并不一致。有人认为网络信息的生产者和出版者是保存的第一道防线,但是这种选择很不稳定,当出版者无力或不愿意保存时,就会导致数据的永远消失。长期保存需要有固定和长期收入(资助)的机构来承担保存的任务,因此有的学者提出应该由那些能够"维持几百年以上的专门的长期保存机构,比如图书馆、档案馆"进行保存[3]。IFLA/IPA 的联合声明也指出,"出版者应该担负短期保存的责任,长期保存的责任由图书馆承担"[4]。还有人认为网页之间相互联结的特性超越了国家之间的界限,应该建立国际机构来承担网络信息的保存,但目前来看,建立一个有稳定收入来源的国际性组织还不现实。到底哪些机构应该并且能够承担网络信息的保存,合作模式有哪些?是网络信息保存中的一个不可回避的问题。

1.3 法律

新的知识产权法对数字文献的知识产权持肯定态度,网络信息资源同其他任何出版物一样都受到知识产权的保护。因此网络信息归档系统是否有权

利复制和使用网络信息?是否能够及时对原先收藏的网络信息进行各式转换,以便在新的软硬件环境下使原来的信息可以使用? Alenka Kavcic 指出,网络信息保存面临的法律问题主要有 3 个环节:收集网络信息、提供存取以及长久保存[5]。比较一致的看法是建立和完善数字呈缴制度并修改相应的知识产权法,为网络信息保存提供必要的法律支持。但目前仅有少数国家,如挪威和丹麦,将网络信息资源纳入到呈缴之列。

2 国外主要研究项目情况分析

从 1994 年开始,就有一些国家的图书馆、档案馆等机构开展各种网络信息保存的实验项目。影响比较大的有:澳大利亚国家图书馆 1996 年开始 PANDORA 项目(保存和存取澳大利亚的网络信息资源,Preserving and accessing networked documentary resources of Australia),根本目的是"在建立一个经过选择的澳大利亚联机出版物档案系统的同时,并且发展一个可以长久保存的国家策略"[6]。到 2003 年 9 月,该项目已经收集大约 4682 个标题,16841954 个文件。1996 年,美国圣弗朗西斯科的一个非营利组织——Internet archive 开始收集所有可以公开检索到的网络信息。从 2001 年起,开始通过一个名为 Wayback Machine 的工具为用户提供检索服务。美国国会图书馆 1997 年开始进行网络信息保存试验项目——Minerva Prototype,该项目的主要目标是"为有关网络信息的选择和收集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网络信息保存项目提供指导和经验"[7]。OCLC 2002 年开始一个 web document digital archive 项目,目的是建立一个可以提供长期存取网络文献的持续性服务系统。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进行一项名为 Nordic web archive 的项目。该项目的目标是"通过北欧几个国家图书馆在技术和方法上的合作,保存北欧的网络资源,以便为今天和以后的用户提供公共获取,为研究服务。"[8]北欧各国还有自己的网络信息保存项目,如瑞典的 kulturarw3 项目、丹麦的 Netarchive、奥地利的联机归档系统(The Austrian On-line, AOLA)、挪威的 the Paradigma Project。1998 年开始的 NEDLIB网络化欧洲存储图书馆(Networked European Deposit Library),是一个合作项目,始于 1998 年 1 月,由荷兰国家图书馆领导,参加方包括其他 8 个欧洲国家的国家图书馆、1 个国家档案馆和 3 家出版商。该项目的主要目标是建立欧洲网络化存储系统的基础构架。英国国家图书馆从 2001 年到 2002 年发起组织了一个网络信息保存的小型项目 Britain on the web。法国、德国、荷兰、新西兰、日本等也都开展了网络信息保存的实验性项目。

2.1 多方参与、分散保存的项目特点

数据归档小组 1996 年曾经指出,"发展建立数字归档系统的最有效、可行的办法是构想一个分散的、而不是集中的框架,来收集信息对象、保护长期完整性从而保证未来的使用。在这种分散的框架下,将保存任务赋予那些对数字对象非常关注,并且能深入理解数字信息的价值的机构。"[9]目前,网络信息长期保存成分散保存的模式,没有一个单独的机构专门负责保存。

国家图书馆是目前网络信息保存的主力。在 IFLA/IPA发表的联合声明中指出,"国家图书馆应该同其他主要图书馆一起领导数字出版物的长期保存"[10]。继澳大利亚和瑞典之后,奥地利、芬兰、法国、新西兰、美国等相继开始了网络保存的实验性项目。

国家档案馆也开始参与网络信息资源长期保存的实践与研究,其重点放在政务信息的保存。2001 年,澳大利亚国家档案馆和公共记录管理局颁布了面向网站管理员的详细的电子记录管理指导。2001 年 1 月,美国的国家档案文件署(NARA)要求所有的联邦机构对他们的公共站点进行快照。英国的公共档案局将唐宁街 10 号网站的快照在 2001 年 6 月大选之前传送给国家档案馆。

大学和协会也在支持或进行小规模的网络信息保存的试验。如芬兰赫尔辛基大学领导的 EVA 项目。澳大利亚墨尔本大学专门成立了网络信息资源保存专门小组。美国的史密斯学会对该站点的信息进行长期保存,并完成了从 html 到 xml 格式的迁异。德国海德尔堡大学汉语研究所开展汉学研究数字归档项目(DACHS, Digital Archive for Chinese Studies),保存和汉学研究相关的网络信息资源并提供存取,特别是政治和社会方面的资源。另外媒体也开始进行网络信息保存,如英国广播公司、美国 CNN 对自身网站的内容进行长期保存。

2.2 基本框架

除了少数项目,如 Internet Archive ,目前大多数系统都是遵循 OAIS 模型。该模型由美国空间数据

系统咨询委员会于1999年提出,全称为开放档案信息系统参考模型(Reference Model for an Open Archive Information System)。该模型主要包括摄取(Ingest)模块、存储(Archival storage)模块、数据管理模块、检索存取(Access)模块、系统管理(AD)等模块。其中,SIP(submission information package)是提交信息包,DIP(dissemination information packae)为发布信息包,AIP(archival information package)是存储信息包。其中SIP和DIP比较简单,而AIP需要经过一系列的检验描述、重新封装等加工处理,即较复杂。

由于该模型只是一个参考模型,一些项目在设计自己的归档系统时作了一些调整,如NEDLIB就在OAIS模块的基础上增加了一个长期保存(preservation)模块,保证所有存储的数据流在原有应用系统废弃的情况下仍然可以使用,并对数据迁移和仿真技术作了描述。

2.3 收集策略

从目前项目实施情况看,主要有两种技术方案:一种是以澳大利亚PANDORA为代表的选择性保存,另一种是以Internet Archive为代表的全面自动保存。

PANDORA项目还没有真正实施之前,澳大利亚国家图书馆就着手制定联机出版物的选择标准。Law曾经解释这么做的原因是:"收集和保存数字信息非常复杂、费时和昂贵,因此现阶段国家图书馆职能将精力集中在那些现在和未来具有研究价值的资源上"[11]。该项目的特点就是高度选择。基本的选择标准包括:关于澳大利亚内容方面的、是否为澳大利亚作者所著、研究价值、出版物是否有纸本、在出版环节是否有任何质量控制、公众对该主题的关注程度、出版物是否已经被权威标引机构标引等。全面自动保存主要是利用机器人、爬虫等网络搜索工具对所有相关的网络信息资源进行抓取。Internet Archive采用Alexa Crawler定期对网络信息资源进行收集。瑞典的kulturarw3使用Combine Crawler,其他欧洲国家使用芬兰CSC的NEDLIB收集器。

选择性保存的优点是,不用将有限的精力浪费在保存很多垃圾信息上。但甄别筛选是非常费力的事情,并且后人可能对今天选择的标准进行非议。Lloyd Sokvitne曾经指出:"我们不知道未来的人需要什么样的信息,我们的判断标准不一定科学。"[12]尽管人们很尽力,但仍旧会失去很多有价值的信息。全面保存可能会保存很多没有价值的网络信息,但是会节省人力。美国国会图书馆Minera项目的负责人威廉姆·亚姆斯在项目报告中指出,该项目使用的选择策略的成本大约为Internet Archive使用的全面策略的100倍。但是目前的自动收集工具只能收集那些静态信息资源(有的称浅层网络资源,surface web),还不能处理那些所谓的深层网络资源(deep web),如动态信息、数据库驱动的站点(database-driven sites),这些信息目前只能通过人工方式进行收集。因此,越来越多的项目倾向于综合这两种方法,如法国国家图书馆尽可能地用自动方式,必要时人工介入。

还有的学者将网络信息资源的收集模型分为推送模型(push model)和拉取模型(pull model)。所谓推送模型就是指建立网站的机构负责将网站上资源传到负责保存的机构,这种方式以数字呈缴法为依托,丹麦就是基于这种模型;而拉取模型则是指负责保存的机构使用相应软件自动收集网站资源。

2.4 存取

由于网络信息保存的法律框架还没有建立起来,很多网络信息保存项目还不能提供存取,或者提供限制性存取,比如只能在图书馆进行检索,还有需要提供检索口令。澳大利亚PANDORA的经验值得借鉴,所有商业出版物都是在征求出版者的同意后进行收集,因此可以提供检索。日本的WARP计划等纷纷仿效,和出版商合作进行网络信息资源保存,从而保证网络信息资源的收集和存取。

3 研究展望

目前国外对网络信息资源保存的研究已经起步,有了一些初始项目,还召开了几次国际会议,如欧洲数字图书馆会议网络信息资源保存研讨会,从2001年开始每年举行1次,连续举办了3次。2002年1月30日,在东京召开了"网络信息保存国际论坛"。2002年3月25日,数字保存联合会在伦敦举行"管理和保存联机文献和记录:网络信息保存论坛"。

目前的研究重点是网络信息资源保存的技术问题,特别是对于网络信息收集技术以及保存元数据方面的问题进行分析。对于同样是技术性问题的长期保存策略不够重视。有的学者已经认识到这个问题,指出:"当前的项目更加侧重于资源的收集,相对来说,对于保存技术的关注则要少得多。短时间内,

这不会有什么错误,但是从长远考虑,必须要分析研究各种不同的保存策略(迁移、移植、更新)对于网络信息资源的适用性以及建立一个基于 OAIS 模型的可信赖的存储系统。"[13]另外,技术不能解决网络信息保存的问题,还需要相应的经济、法律、责任体系等机构建设。随着网络信息保存项目的日益发展,会有越来越多的学者关注长久保存的问题以及社会支撑体系的建立。

合作模式的研究也将成为一个热点。即使所有的技术问题都解决了,由于网络的全球性特点以及确定网络资源国家界限的困难,有时候很难确定某一网络信息资源应该由哪个国家的哪个机构负责。因此探讨合作方式至关重要,不仅是国家之间的合作,还有各类型机构之间的合作。Andreas Rauber 曾经分析在欧盟框架内国家图书馆、研究图书馆和其他机构合作进行网络信息保存的可行性。在第 2 届 ECDL 会议上,Michele Kimpton 就曾经提出 Internet Archive 联盟的构想,设想了 Internet Archive 和国家图书馆合作的方式:国家图书馆负责选择标准的制定、收集和提供检索,Internet Archive 负责技术上的支持以及研发新的工具。

对 Internet Archive 的资源进行选择性评价的结果表明,未来的网络信息保存应该以特别用户团体为中心来进行。尽管我们不能确切预测后世用户的需要,但是对于当前用户的需要是能够掌握和了解的。因此今后的保存服务应该是在保存人类文化遗产的基础上,以用户需求为中心开展服务。

参考文献

1 Brewster Kahle. Archiving the Internet. http://www.nettime.org/Lists-Archives/nettime-1-9710/msg00014.html
2 National Digital Information Infrastructure and Preservation Program. http://www.digitalpreservation.gov/index.php? nav=3&subnav=2
3 Johan Mannerheim. The WWW and our digital heritage — the new preservation tasks of the library community. http://www.ifla.net
4,10 IFLA/IPA Preserving the memory of the world in perpetuity: A joint statement on archiving and preserving digital information. http://www.ifla.net
5 Alenka Kavcic. Archiving the world wide web — some legal aspects. 68th IFLA Council and General Conference. 2002.
6 Warwick Cathro, Colin Web and Julie Whiting. Archiving the Web: the PANDORA archive at the national library of Australia. http://www.nla.gov.au
7 William Y Arms. Web preservation project: final report. http://www.loc.gov
8 Svein Arne Brygfjeld. Access to web archives: the Nordic Web Archive Access Project. http://www.ifla.org
9 Preserving digital information: Task force on archiving of digital information. http://www.rlg.org/ArchTF/tskforce.html
11 Law, C. PANDORA: the Australian electronic heritage in a box. International Preservation News, 26, December, 13-17.
12 Lloyd Sokvitne. Our Digital Island: web preservation issues and solutions at the state libray of Tasmania.
13 Michael Day. Collecting and preserving the world wide web. http://www.jisc.ac.uk/uploaded-documents/archiving-feasibility.pdf

赵俊玲 河北大学管理学院讲师,北京大学信息管理系在职博士生。通讯地址:河北大学管理学院。邮编 071002。

(来稿时间:2003-10-29)

(上接第 79 页)选择离开,转投向其他商业信息机构寻求帮助。

参考文献

1 刘荣.图书馆信息服务与管理.北京:北京图书馆出版社,2002
2 张晓林.走向知识服务:21 世纪中国学术信息服务的挑战与发展.成都:四川大学出版社,2001
3 李朝民,日本文部省学术情报中心简介.图书馆理论与实践,2003(1)
4 张宇萌,张树华.信息服务与知识导航.中国图书馆学报,2003(1)
5 李爱国,汪社教.学术信息资源整合工具——SFX 及其启示.现代图书情报技术,2003(3)
6 邵敏,李旭.合作虚拟参考咨询服务.现代图书情报技术,2003(3)
7 刘喜申等.因特网资源及其应用.北京:北京图书馆出版社,2002

黄连庆 佛山科技学院图书馆馆员。通讯地址:广东佛山。邮编 528000。

(来稿时间:2003-10-28)