

对美国各级政府文献信息及其永久性获取的研究

在电子文献的概念提出来之前,政府信息向公众的永久性开放就是一个极具挑战性的话题。1813年,国会立法指定国会参众两院所有的会议记录都要交给联邦寄存图书馆收藏,并向公众永久开放。现如今,政府信息绝大部分以电子形式发布,只需轻点鼠标就可删除,要确保公众的永久获取更是困难重重。

政府机构删除信息出于多方面的原因:其一,防止敏感信息的外泄;其二,网页重建的需要;其三,对某些重要文献的历史意义缺乏认识,如此等等。1998年,国家卫生研究院发现,应纳入联邦寄存图书馆体系(Federal Depository Library Program, FDLP)的政府机构文献有78%没有提交。前政府印刷局员工詹姆斯认为政府机构内部出版物无视FDLP将会使我们的后代对当前政府缺乏足够的认知。

研究综述

上个世纪90年代早期,电子信息载体的诞生促使图书馆员和政府官员对政府文献的传播、获取和保存进行了思考。他们意识到在不久的将来,政府信息只能以电子形式存在。为此,1991年4月25日,印刷联合委员会举办了一次以“政府信息:公共财产”为主题的听证会,图书馆员和相关人士认为一旦政府信息只能通过电子途径获取,联邦寄存制度就必然会成为一纸空文。而且,当时许多政府电子文献并没有对公众开放,有的文献用户需要付费才能利用。图书馆界人士担心,随着电子文献数量的增加,也会有越来越多的政府文

献被清除。

1993年美国国会制定《加强政府印刷局电子信息开放法》(GPO Electronic Information Access Enhancement Act)规定政府印刷局(Government Printing Office, GPO)应保存所有形式的联邦政府出版的信息,并确保公众能从寄存图书馆中自由获取。为此,同年,信息产业协会提交了一份报告,描述了政府文献和地方文献保存的重要性,概述了一系列的政策原则以确保政府文献资源持续永久性开放,并强调无论政府信息出现在何种媒体上,公众都有同样的获取权。

1995年GPO成立专班小组,开始着手研究政府信息自由获取的支撑系统。联邦寄存图书馆指导委员会在同年秋天召开的会议上,承诺了他们对寄存图书馆的责任,并决定与联邦政府机构合作以确保政府电子出版物的可获取性。1997年,据GPO统计,仍有50%的政府出版物尤其是科技报告没有纳入寄存系统,其原因在于:一是出版物的传播渠道单一,机构之间缺乏交流;二是政府机构没有完全依法执行寄存制度;三是由于私营部门的参与而涉及到版权问题。可以看出很多政府部门规避了寄存制度,为此GPO通过与机构的合作协商、寻求国会干预等途径来实现更多出版物的纳入。

1998年O'Mahony提出政府信息由于缺少搜索和保存途径,其丢失的速度值得我们关注,甚至会在不久的将来消失殆尽,这也是由电子信息自身特性所决定的。为了避免这种情形的出现,GPO开始利用统

一资源定位器技术 (Persistent Uniform Resource Locators, PURLs), 无论出版物的链接地址如何改变, 它都可以将用户引导至正确的链接。在随后进行的一项有关失效链接的调查中, 367 个被测试的链接中有 213 个是无效的, PURLs 可以记录下这些问题链接, 再由 GPO 核实。同时, GPO 开始寻求合作伙伴, 例如美国能源部报告数据库, 后者可以为用户提供政府电子信息的多种检索途径。

美国国家图书馆信息科学委员会于 1999 年提交了“政府信息出版物评估报告”, 认为政府信息政策仍然很不健全, 并且各级政府机构在对出版物的管理上缺乏协调、机构内没有专人管理, 从而导致信息分散、混乱。与此同时, 他们对于公众获取政府信息的重要性也认识不足。

人们对流失于传播范围之外的文献数量也颇为关注。2001 年, 美国法律图书馆协会政府文献部成立短期文献委员会, 其任务是“确认和报告短期政府文献, 并促进此类文献的传播。”2002 年, John Shuler 认为寄存图书馆应重新审视他们在电子信息传播中的作用, 图书馆员过于重视短期信息的获取, 而忽略了如何将信息有效提供给公众。图书馆应对政府电子信息的保存和发展起指导性作用, 并按有效方式来获取和组织这些信息。政府印刷局的 George Barnum 概述了该机构通过扩展书目控制技术和新的元数据来实现政府信息永久获取的工作模式, 介绍了对此类信息确认、编目并为公众提供检索入口和利用等一系列工作流程。与此同时, 他也认为为了实现政府电子信息永久获取, 寄存图书馆的角色应该有所变化。

9.11 事件对政府信息获取产生了重大影响, 流失文献信息的确认变得更为困难。比如, 一个题为“美国大陆大型公共地面供水系统的区域特征”的 CD-ROM 文献就被政府印刷局确认由于当初出于保密制度的

考虑现在已经流失。白宫指示各政府机构随时删除网站上虽不保密但仍然较为敏感的信息; 教育部经常重建网页, 删除了几百条报告; 能源部和国防科技信息中心删除了网站上数千条涉及安全威胁内容的文献信息。此外, 一些非图书馆文献来源机构对信息传播也慎之又慎, 如美国国会管理和预算办公室、美国科学家联盟, 甚至纽约时报等等。Heintz 建议图书馆员应积极与政府机构建立合作关系, 以促进网络上有用信息的传播, 并为用户提供政府机构网站的检索入口。同时, 他还建议政府印刷局应简化登录电子文献的流程, 从而减轻政府机构负担, 使双方合作更为顺利。

尽管文献信息的流失已引起人们极大的重视, 但其趋势已不可改变。2004 年, Kennedy 发现越来越多的政府信息愈来愈难以被公众所获取。研究人员陆续开发了一些项目来研究搜索、保存这些数量巨大电子信息的工具。政府印刷局联合斯坦福大学及其余几家联邦寄存图书馆开发了 LOC-KSS (多备份资料保存) 系统。该项目主要致力于解决数字化信息的永久保存与利用问题。在 LOCKSS 管理系统中, 政府印刷局设置一台电子文献服务器, 直接从政府机构获取文献资料, 图书馆获得授权后, 可利用 LOCKSS 工具在本地创建一个费用低廉、永久保存的数字化信息缓存站点, 其电子资源馆藏不受政府法令及丢失等问题的影响, 从而可为用户提供持续的、永久性的电子资源存取服务。这样, GPO 的角色就从对印刷文献的处理转移到电子信息的 manages 了。

2006 年, Mart 撰文提出因特网信息的删除对国家安全影响甚小, 而对信息用户影响巨大。她指出, 很多网站由于“网络清洗”被关闭, 文献被莫名其妙地删除, 这些文献主题涉及员工隐私保护、妇女健康、环境污染、公民权等各个方面的内容。同时, 她还建议利用网络技术监测信息状态。

2008年,伊利诺斯州立大学克莱因民主研究中心公布了题为“白宫粉饰历史”的报告,指出白宫在未给出任何通知的情况下改变了多份新闻稿,这无异于改写历史。研究中心是通过对比保存在Internet Archive和Lexis Nexis内的备份后有此发现的。Internet Archive的主管Brewster Kahle指出网页平均生命周期只有100天,暗示数字记录必定要被一次次地重写,而且未必是利用同一种方式。Jacobs将政府电子信息的现状描述为黑暗期,政府与其民众的交流渠道应发生根本性变化,这种变化并非源于技术要求,而是社会、经济、政治等因素的要求。他认为我们必须明确哪些内容的政府信息具有历史意义,将此类信息的维护、保存负责到具体人员,确认哪些信息应该保存和被获取,同时呼吁创建一种保存历史记录的新体系。

文献研究表明,在过去20年中有越来越多的图书馆员和其他领域的研究人员对电子信息的永久保存和使用赋予极大关注。人们对普通政府文献如州立政府和地方政府文献是否也给予了同样的关注?这些文献是否也存在同样的问题?本文接下来会对各级政府信息包括联邦、州立和地方政府文献的稳定性展开进一步研究。

研究方法

为了检测政府信息的稳定性,本文作者遴选一定数量的联邦、州立、地方各级政府电子信息作为文献实例和研究对象,对它们展开了为期一年的跟踪研究,观察一年后保留和消失信息的数量、种类。我们会对这些文献实例定期检测,以确定其有效性,每种文献的相关数据都会记录在电子表格中。

(1) 联邦政府文献的选择。作者利用政府分类名录和文献管理局分类项目来挑选机构。各个政府部门中,我们注重大型机构和小型机构的混合搭配。机构数量取

决于所属部门的规模,也就是说,规模较大的部门选择的机构会多一些,反之则会少一些。此次研究选择了15个白宫行政部门、22个独立联邦机构,而后从中选出100个办事处作为文献来源机构。接下来,研究人员将这100个机构分成两组,每个研究者负责其中50个,并从中选取一些具有代表性的文献作为研究对象。文献格式包括PDF格式、TXT文本格式和html格式,各种文献需要登录的信息包括通讯机构名称、文献名称、发布日期、URL地址、PURL地址、URL或PURL是否改变、改变后的地址为何、检测时是否为有效地址等。

(2) 州立政府文献的选择。作者列出50个州,分为两组,每组25个,文献选取流程同上,文献登录信息包含文献所属州名、文献名称、发布日期等其他和以上相同的内容。

(3) 在地方政府文献选择过程中,有25个州被选出,并分为两组,每位研究人员负责一组,每个州选取一个城市,从该市政府网站上挑选一篇文献作为研究实例,这些城市规模不一,既有大型城市,也有中、小型城市。文献登录信息包括文献所属城市和州的名称、文献名称、发布日期等与以上相同的内容。

文献实例确定后,2008年1月中旬研究人员将文献信息登录到电子表格,并从1月31日开始检测这些文献的链接,每月最后一天检测,一直持续到同年12月。检测方法是將文献的URL地址复制到浏览器地址栏检验它所在网页是否仍然有效,或文献是否仍可被下载。联邦政府文献如同时有URL和PURL地址,两者都会检测。如是有效地址,表格中应被登录为“yes”;如URL地址无效或是文献已从网站中删除,研究人员要继续检索上一级链接看其链接地址是否已改变,并将新地址登录到电子表格中;如文献从上级链接和因特网中都无法获取,表格中会登录为“no”,但在下

一周期仍会继续检测。对于带有 PURL 地址的联邦政府文献,研究人员会利用政府印刷局的 PURL 搜索表单工具,每月检索其有效性或是否变更。在这些文献实例中,研究人员发现,联邦政府文献有 85% 为 PDF 格式,州立政府文献中 PDF 格式占 92%,地方政府文献中 PDF 格式占 84%。

讨论

(1) 2008 年底,研究表明 100 篇联邦政府的文献实例中,有 3 篇文献被彻底删除,占其 3%,其中两篇是国防资料,另一篇是农业信贷署资料,也就是说,13 篇国防文献实例,一年后流失了其中的 15%,而农业信贷署的这篇文献则是被新的版本替代。改变 URL 地址的文献有 5 篇,其中一篇在变更链接前中断了一段时间。在这 100 篇文献实例中,有 22 篇提供了 PURL 地址,一年后仍有 20 篇的地址有效,而国家教育统计中心和联邦商务委员会分别提供的 PURL 地址都已关闭,但其文献仍可在这些机构网站其它地方获取。没有一篇文献实例的 PURL 地址被删除,这说明政府印刷局对于维护在线信息稳定性方面的工作已取得很大进展。但与此同时,要逐一保存数量巨大的电子政府信息仍然任务艰巨。

(2) 50 篇州立政府文献实例删除率则很高。从亚利桑那州、堪萨斯州、蒙大拿州、佛蒙特州、马萨诸塞州、怀俄明州选取的文献实例被完全清除,占总数的 12%。其中,亚利桑那州和佛蒙特州的文献被新版本替代,堪萨斯州文献链接被指向联邦政府中同一主题文献。加利福尼亚、纽约、德克萨斯州的文献实例变更了链接地址;夏威夷、科罗拉多文献实例中断一个月。

(3) 参与调研的 25 篇地方政府文献只有一篇特罗伊市的文献实例被删除,占总数的 4%,另外三篇变更了链接地址,占总数的 12%。

对于联邦政府文献,我们还需要展开进一步调查。目前的研究表明州立和地方政府文献保存存在较为严重的问题,而程度如何、是否在继续扩大以及对公众有何影响尚需更为深入和纵向的调研。

结论

毫无疑问,人们对于政府电子信息给予了极大关注和努力,但研究表明这无力改变此类信息不断流失的现状。我们发现研究对象中 3% 的文献一年后被清除,其原因在于政府机构行为。值得警惕的是,三篇清除的联邦政府文献中有两篇是国防资料。州立和地方政府文献清除比率较高,对于这些政府文献的保存,我们也应该赋予关注。

我们还发现,新版文献公布后,旧版均未予以保留,而这种清除则与其内容主题无关。此外,由于网站重建而造成的 URL 地址变更也为用户的利用带来了不便。

研究表明,电子环境下各级政府文献的稳定性存在问题,州立和地方政府文献尤其如此,而对联邦政府电子文献的评估还应进一步深化。图书馆员、研究机构和信息用户应携手努力,使政府和公众都认识到信息保存的重要性和现状的严峻。未来,我们还有许多问题尚需思考,比如电子信息的流失既然不可避免,那么将其控制在何种程度才是合理的?要采取何种措施才能尽量降低其流失率?

编译自: Claudene Sproles, Angel Clemons.
Permanent Electronic Access to Government
Information: A Study of Federal, State, and
Local Documents Electronic Journal of Aca-
demic and Special Librarianship, v.10 no.2

编译: 袁俊华 武汉大学图书馆

校对: 胡春娟 武汉大学图书馆