《专题：数字资源长期保存》

澳大利亚国家图书馆的网页存档

澳大利亚国家图书馆 著　余敏 编译

澳大利亚国家图书馆的网页存档计划可以追溯到 1996 年 PANDORA 档案计划的实施。由于 PANDORA 档案计划囊括了 10 年之前至今的各种内容，国家图书馆所采用的网页存档方法从一开始就是有选择性、高效而又实用的。

国家图书馆始终坚持在 PANDORA 档案内容的建设中采取一种广泛协作的方法，并积极地促进澳大利亚国立图书馆、各地图书馆以及其他主要的资源搜集组织的参与。目前，已经有超过 10 个图书馆组织参与到 PANDORA 档案内容的选择和存档计划中去，这些图书馆包括：澳大利亚大陆各州立图书馆、北方图书馆（the Northern Territory Library）、国家声像档案馆（the National Film and Sound Archive）、澳大利亚战争纪念馆（the Australian War Memorial）以及澳大利亚原住民及托雷斯海峡居民研究协会（the Australian Institute for Aboriginal and Torres Strait Islander Studies，AIATSIS）。

由于国家图书馆开发出的一套被称之为PANDAS（PANDORA数字化存档系统）网页存档流程管理系统，使得开展协作化的网页内容搜集以存档变成可能。这个基于网络的应用系统允许各参与馆的负责人通过由国家图书馆负责维护的一些设施开展网页资源的存档工作。PANDAS最初是作为一个研究成果于 2001 年 6 月份投入应用，2002 年，PANDAS发布了第二代增强版本。预计在 2007 年，PANDAS将会推出一个经过重新设计和功能加强的第三代系统。

对于国家图书馆而言，对已存档网页内容的存取是一个很重要的问题。通过采用一些特定的网页存档的方法，目前已经可以在一定范围内从那些从事网页存档的出版商处获得必要的许可，并且，还可以实现从这些已通过选取的网页资源中生成 MARC 记录。对存档内容的获取是通过 PANDORA 网站入口来实现的，该网站能够提供 Lucene 全文搜索引擎来以及存档资源的主题和标题列表。此外，PANDAS 存档管理系统还提供了限制模块功能，该模块使得各参与馆的负责人可以方便地在需要的时候对一些存档内容施加获取限制。

已存档文章:	13,719
已存档实例:	27,933
文件数量:	34,541,963
数据量（单位：千兆）:	1.3

表 1 至 2006 年底 PANDORA 档案统计数据

1 澳大利亚域名网页存档

PANDORA 档案计划的有选择性的存档方法建立起了一整套的资源存档。这些存档内容的搜集必须十分关注选取标准、存档过程的质量评估、以及获得出版商的允许并提供对于内容的获取。尽管这本身就是选择性存档的优点，但同时网页存档的可能范围也将被限制。基于此，国家图书馆于 2005 年与著名的因特网档案计划（Internet Archive）展开协作进行了大规模的域收割活动以作为 PANDORA 选择性存档计划的补充。迄今为止，已经完成了两次大规模的资源“爬行”，第一次是在 2005 年的 6－7 月，第二次则是在 2006 年的 8－9 月份。

这两次大规模的资源“爬行”过程中采用了自动的 geoIP 查询识别机制，其目标就在于在澳大利亚境内主机中广泛深入地抓取尽可能多的采用.au 顶级域名以及那些非.au 域名的网页资源。

澳大利亚域名“爬行”内容的获取提供是国家图书馆的主要问题之一。尽管目前并不提供该内容的公众获取，但是我们期望新近提出的相关立法提案可以更好地支持图书馆资源的搜集和保存。2006 年著作权法案修正案（Amendments to the Copyright Act）对于图书馆的数字搜集和保存活动需求给予了一定的重视和支持。与此同时，我们也寄希望于在不久的将来，澳大利亚联邦的合法保存法案会将范围扩展到数字资源，这样，我们就能够在网页资源的有效搜集以及相关存档内容的提供方面得到更好的支持。

收割域名	2005	2006
搜集的独立文件	185,549,662	596,280,285
爬行域名数:	811,523	1,046,038
数据量（单位：千兆）	6.69	19.04
爬行持续时间:	4 weeks	5 weeks

表 2 澳大利亚域名网页的收割统计数据如下

2 策略及协作联盟

通过 10 多年的对于网页的有选择性存档以及两年的域名收割，国家图书馆积累了大量的经验，足以对未来澳大利亚网页资源的存档策略做一个前瞻性的预见。这些策略应该着重于以下几个方面：保证对不断增长的网络出版物的高效搜集，重视存档过程中实效性和质量的价值，对于有用及实用描述的需求，对资源发现路径的需求，以及随着时间的流逝，仍能实现对存档内容持续获取的保存方法的重要性。

2006 年 12 月，为了更加紧密地将国家图书馆网页存档计划和数字化保存活动结合起来，在图书馆馆藏管理部（the Library's Collections Management Division）内部成立了一个新的分部。这个由 Colin Webb 领导的新的网页存档和数字化保存部门，其战略目标就在于更好地结合网页资源描述和搜集功能，并在存档数据中发展和应用数字化保存管理。

国家图书馆将继续积极参与到各协作联盟的工作中，通过促进相应标准，政策以及工具的发展来推进网页存档和保存工作的深入开展。在国际上，国家图书馆早在 2003 年就已经是国际因特网保存联盟（International Internet Preservation Consortium ， IIPC）的成员，并将继续担任该组织的执行委员会成员。在国内，国家图书馆目前已经参与到澳大利亚可持续知识仓储伙伴计划（Australian Partnership for Sustainable Repositories ， APSR)中，在这个联盟内，国家图书馆将继续致力于建立一套基于 PREMIS 和 METS 的功能框架，以满足对于保存元数据的需求以及建立起一套自动申报过时系统（Automated Obsolescence Notification System ， AONS）。

编译自： Web Archiving at the National Library of Australia PANDORA: Australia's Web Archive. http://www.ndl.go.jp/en/publication/cdnlao/058/581.html. [2007-4-1]