# 澳大利亚网络域名爬行与收割研究报告

Paul Koerbin 著
魏来 编译

## 1. 概要

2005年6月—7月，澳大利亚国家图书馆开展了第一次大规模的本国网络域名收割项目。该项目的目的是依据深度广度原则，在有限的爬行期限内尽可能多地收集和保存以.au为结尾的域名，并更好地理解数字化存储的相关问题。IA（Internet Archive）是唯一有过以保存为目的进行如此大规模网络收割经验的代表机构，将代表图书馆负责这一工作。收割爬行连续运行六周，抓取超过18500万个文档，数据量相当于6.69TB。在对IA使用的WayBack机器搜索引擎获取网络文档进行初始分析并对IA提供的索引内容进行更多细节分析的基础上，完成了本报告。WayBack机器搜索引擎界面存在着各种限制，这意味着第一阶段的分析具有局限性。然而，报告显示收割工作非常成功，特别是在爬行的宽度方面。域名爬行为将来国家图书馆制订网络保存策略提供了有价值的标准和基础，同时也促进了澳大利亚文化遗产保存事业的发展。

## 2. 背景

1996年澳大利亚国家图书馆建立了PANDORA（Preserving and Accessing Networked Documentary Resources of Australia）存取计划来有选择地保存澳大利亚网络资源文档。PANDORA保存内容包括网站、网络出版物等网络资源，以PANDORA在线资源选择指南（http://pandora.nla.gov.au/guidelines.html）为基础进行网络文档选取。作为该计划的成果之一，在2005年8月，共存档9315个独立主题。由于一些主题预计采取周期性的存档方式，实际相当于拥有18928个主题文档。排除保存的副本量，在过去9年该项目共获取2600万个文档，相当于925GB的数据量。

PANDORA资源存档的质量标准是存档的全面性和功能性，进而能够反映由于收割器的技术限制而能够达到的存档的最高标准。在收割之前要先从出版商那里得到保存出版物的许可，并且被保存的资源可以通过PANDORA 档案库门户获取。因此，PANDORA档案库是一个高质量、可获取的、澳大利亚网络资源的精选。

澳大利亚国家图书馆认为这是网络存档最为适宜的方法，利用现有的可获得的资源得到最好的结果。但图书馆也承认网络存档方法的选择存在局限性。能够被存档的资源数量仅仅是澳大利亚网络出版总量的一小部分。因此图书馆认为对澳大利亚网络资源尝试大范围收割是非常重要的。

IA的经验为实施并完成澳大利亚首次大规模的网络域名收割计划提供了机遇。IA作为一个非赢利组织，从1996年开始对万维网资源进行爬行收割，具有丰富的经验。为了进行大规模文档爬行和收割保存工作，IA专门设计开发了Heritrix爬行器。而澳大利亚国家图书馆和IA都是国际互联网络保存联盟（International Internet Preservation Consortium，IIPC）的成员，拥有良好的合作关系，这为双方的合作奠定了良好的基础。

## 3. 目标

项目的主要目标包括：

- 在非排他性主要目录和特定参数（主要是.au的域名）的基础上，检验自动爬行大规模网络域名的能力。包括：
  - 爬行宽度的成功性分析

- 爬行深度的成功性分析

- 观察和分析Heritrix爬行器的性能
- 分析自动收割内容获取的限制因素
- 更为清晰的了解澳大利亚网络域名构成
- 通过大范围域名收割获得能够用来获取的真正内容
- 更好地理解对大规模域名收割的过程和成本

## 4. 其他域名爬行项目

近年来，许多组织开展了大范围、自动化的域名爬行工作，其中瑞典和挪威成绩显著。瑞典皇家图书馆于1997年开始进行瑞典域名爬行工作，于2005年1月完成了第12个爬行计划。最近，斯洛文尼亚、希腊和葡萄牙等国家的一些机构也开始进行国家域名爬行工作。通过与其他国家域名爬行工作统计数据的比较，澳大利亚域名爬行规模相对较大。例如：瑞典的最新爬行数量大约为4600万个文件（URLs）即1.6TB数据（约为澳大利亚域名爬行数量的1/4）。从1997年至2005年瑞典的全部域名爬行总量是30600万即10TB。葡萄牙2003年的爬行数量约为380万个文件（URL）即78GB。挪威国家图书馆的Paradigma计划，2003年8月爬行410万。

从这些项目现有的出版物中可以看到，它们要解决的共同问题包括：

- 用于自动收割的网络域名特征描述和转换问题
- 从文档内容中自动分析和抽取元数据以支持获取
- 存档内容的描述

在处理这些问题的时候要保持与其它域名收割机构的密切联系，相互交换经验、相互学习。IIPC是便于各机构间的联系的有效方式之一。

## 5. 与IA签订的协议内容

IA于2005年5月12日同澳大利亚国家图书馆签订协议，代表澳大利亚国家图书馆从事网络域名爬行工作。

协议内容包括国家图书馆要求IA提供服务的说明：

- 按照技术、时间和资源限制进行为期4周的澳大利亚全部域名的爬行工作
- 提供分析报告
- 提供Wayback机器格式界面以保存内容
- 构建索引库
- 提供馆藏内容的内部管理

## 6. 爬行技术说明

### 6.1 爬行时间

爬行工作从6月13日开始连续进行4周。但实际上，在没有提供额外费用的情况下，IA将爬行工作延长了2周。因此，全部爬行时间为2005年6月13日到2005年7月25连续6周。在爬行工作正式开始之前，IA组织了为期一周（6月6日—6月10日）的爬行测试。爬行测试阶段不包括任何内容的收割。

### 6.2 爬行说明

爬行说明包括：

- 爬行宽度：全部“*.au*”域名
- 利用IA的经验和不完全执行机制自动进行域名服务器查找与爬行页面链接的非*.au*网页，并利用澳大利亚IP地址鉴别非*.au*域名
- 爬行深度：鉴别网站的全部内容，保证对爬行内容的收割
- 规定每篇文档不超过100MB。超过100MB的文档将不被收割，并提供关于这些文档的报告
- 下载超过20分钟的文档将不被收割

- 遵守Robots.txt规则

6.3 主要目录

使用的主要目录表由IA提供，以之前进行的澳大利亚域名Alexa爬行为基础。Alexa于2004年11月进行了爬行工作共获取342296个以*.au*为域名的主机，形成大量的主要目录，并提供广泛的爬行数据。除了Alexa和IA提供的主要目录，国家图书馆还提供了大约530个URIs，主要是政府和高等教育部门的URIs。IA进一步确认了以前Alexa爬行列表的230个网站的50000个网页，约40个网站被排除，主要是为了避免占用过多爬行时间收割大量商业网站，同时重视政府和高等教育网站的爬行收割。

## 7. 报告摘要与爬行结果统计

7.1 爬行范围统计

爬行主机总量：811,523

爬行文件总量：189,824,119

爬行数据字节总量：7,360,187,145,622(6.69 TB)

被压缩的ARC*文件总量： 4,964,818,275,410(4.52 TB)

被压缩的DAT**文件总量：90,888,523,022(84.65 GB)

被压缩数据总量： 5,055,715,039,125(4.60 TB)

7.2 MIME（Multipurpose Internet Mail Extension）类型数量统计

域名爬行识别出 976 个 MIME 类型。如表 1 所示。

表 1MIME 类型数量统计

| MIME 类型 | 文档数量 | 比例 |
|---|---|---|
| text/html | 126,587,753 | 67% |
| image/jpeg | 32,414,376 | 17% |
| image/gif | 20,716,296 | 11% |
| application/pdf | 3,071,252 | 1.6% |
| text/plain | 1,521,619 | 0.8% |
| image/png | 913,104 | 0.48% |
| text/css | 808,571 | 0.42% |
| application/xjavascript | 429,700 | 0.22% |
| application/msword | 392,140 | 0.21% |
| application/xshockwave | 355,840 | 0.18% |

7.3 IA 全部报告列表

爬行任务完成后，IA 提供了系列报告，包括：

- 爬行报告：爬行范围、文档数量等(4 KB)
- MIME 类型报告 (88 KB)
- 代码响应报告(4 KB)
- 单个主机简要报告 (59 MB)
- 文档排除报告 1:下载超过 20 分钟的文档(5.3 MB)
- 文档排除报告 2：超过 100MB 的文档即使发现也不获取(799 KB)
- 文档排除报告 3：按照 robots.txt 规则排除文档 (3.4 GB)

---

* ARC 文件是 Heritrix 使用的用于保存存档内容的档案文件格式。

** DAT 文件是用于描述 ARC 文件内容的元数据文件。

- 爬行结束时主要的 URIs 列表报告(960 KB)

## 8. 爬行过程管理

### 8.1 澳大利亚国家图书馆和 IA 之间的联系

爬行计划主管是数字化保存部的 Paul Koerbin，IA 的两个主要联系者是网络保存部主任 Michele Kimpton 和爬行工程师 Igor Ranitovic。

### 8.2 同公众的联系

在爬行进行期间和爬行测试期间与 11 个网站管理员建立了联系，联系内容包括：

- 两个网站管理员认为该项计划具有入侵的性质，一个管理员认为这项计划是滥用权利；另外一个来自商业网站的管理员对爬行计划完全否定
- 两个以上的网络管理员直接要求停止爬行工作
- 一个网络管理员开始要求停止爬行，但是经过解释允许爬行继续进行
- 三个网络管理员报告他们服务器日志出现错误，要求爬行停止
- 三个网络管理员对于他们服务日志产生疑问，要求知道更多爬行网站内容的原因。

关于停止爬行的所有抱怨和要求都表现得非常含蓄，同时所有的问题都得到了很好的响应和满意的解决。

## 9. 分析

### 9.1 方法

目前在本报告中描述的对已经进行的域名收割的内容的分析（包括可视化内容样本）只应作为第一阶段分析结果。很明显，第二阶段的分析应该包括 IT 工作人员对收割的内容采用更多的技术性方法进行的分析。收割内容和索引有望于 2005 年 11 月末交付给图书馆，则第二阶段的分析启动。

#### 9.1.1 问题和局限

分析使用的主要方法是获取和检验收割的内容。Per Host 简要报告提供了每个服务器收割文档数量的统计数据，确切地说是被收割的内容的统计数量，其功能只能视收割的内容来决定。

然而这种方法自身存在局限，原因在于目前仅仅能够通过 IA 提供的 Wayback 机器检索界面获取检索内容，只允许获取指定的 URL。Wayback 机器使用 JavaScript 添加 *WayBackCGI* 字符串重写网页原始链接来传递文档内容，并且添加 BASE 标签来解决传递目标的相关链接。一般来讲，要准确知道要检索的 URL，Wayback 机器才能够最好地发挥其功能。由于这些原因，内容是否被收割的结论不是直接就能解决的任务。

#### 9.1.2 爬行宽度分析方法

在分析内容覆盖宽度过程中，采用下面两种方法鉴别网站（或网站内的某一具体页面）：

1. 通过已知网站列表爬行收割，如政府和教育网站。这些网站入口在收割过程中作为浏览专门网站的起始点被获取。而且 Wayback 机器界面能够从实际检索入口中鉴别专门网站，直接进行专门网站检索。如果收割入口的浏览器出现问题，可以通过保存文档内容来鉴别和检索实际网站的网页。在多数情况下，在检索专门网站时显示未收割的网站能够被查找。
2. 通过使用 Google 搜索引擎查找任意网站（或者专门网页）。通过使用 *google.com.au* 网站限制 Google 搜索澳大利亚资源，然后使用随机检索词（地点、人物、主题的名称）和应用额外的限制条件如 *site:.au*（限制.*au* 域名的内容）和 *site:.asn.au*（限制 *asn.au* 域名网站）实现网站查找。

在很多情况下，特别是检验非.*au*网站和weblogs的IP地址的范围是否属于澳大利亚时（使用[www.dnsstuff.com](http://www.dnsstuff.com)网站的工具），这种方法是非常必要的。这也是决定一个网站能否被收割的一种方法。Robots.txt文档也经常被检测，以确定自动排除规则是漏检内容的一个原因。

9.1.3 爬行深度分析方法

爬行深度分析方法存在更多问题。爬行深度是指内容保存的程度，我们可以认为爬行已经收割了所有公共内容文档和支持网站显示和功能的文档，如 stylesheets 和 JavaScript 文档。

限制爬行深度的主要问题包括：

- 缺少现存网站和保存网站文档数量的比较统计数据
- 这种方法过多依赖视觉检验和可操作性不强的样本分析
- 依靠视觉检验保存页面内在的链接。Wayback 机器用来向浏览器传递内容的复杂系统意味着不可能单纯依靠视觉检验来权威的评价被收割内容
- 使用 Wayback 机器产生对保存内容的无效浏览问题

最后一点需要进一步的解释。通过 Wayback 机器界面获取一个页面并且从这个页面出发开始浏览一个网站是可能的，然而在多数情况下通过这种浏览可能不会发现一个页面，相反通过直接获取文档可能会发现该页面。

这些限制表明实施爬行过程中，获取内容数量即爬行深度受到一定程度的限制。然而，另一方面，内容爬行执行分析包括了对爬行深度定性方法的分析，成功爬行了一些主题，并在大型网站内做了抽样测试。这项测试包括一些由于存在特定文档格式或者传输机制（如使用 Flash 技术）而产生问题的网站。这个过程也证明了在爬行实施过程中得出确定的结论是有疑问的，包括上面已经提到的原因，也因为通过爬行器是不可能决定内容是怎样被定位的。这些方面的数据分析还需要大量工作才能得出有效的结论。

9.2 关于专门域名和格式的评价

9.2.1 政府和教育域名网站

政府和高等教育网站在 IA 提供的主要目录中占有绝对的优势，范围广泛，覆盖了所有级别的政府网站。然而许多政府网站和高等教育网站规模很大，通常包括动态内容传递及应用复杂的传递模式和传递功能程序，因此不能保证收割的深度和完整性，浏览这样的网站不可能达到很深的水平。

9.2.2 联合域名网站（asn.au）

为了研究.au 域名的一个特定附属域名获取范围，检测 asn.au 网站的一个随机样本。使用 Google 搜索引擎，限制参数为 *site:.asn.au* 和检索词“Australia”随机选取 40 个网站。Google 检索结果列表显示的 40 个随机选取的网站（或者网页）中 38 个网站（或者网页）的域名能够被收割，样本数据显示收割成功率为 95%。

9.2.3 非.au(如：.com)网站

采用相似的方法对非.*au* 网站进行样本检测分析。对非.*au* 域名网站覆盖范围抽样分析的目的在于评估域名自动爬行功能。在被爬行网页相关链接识别过程中非.*au* 域名的网站被登记并通过 GeoIP 数据库判断该网站是否属于澳大利亚域名。IA 强调这个功能仅仅是在试验阶段，处于部分执行状态，并且不能保证这种机制的成功性。再次选取一个 40 个网站的样本，使用 Google 澳大利亚检索限制来自澳大利亚的网页，检索条件：-*site:.au* (以过滤.*au* 域名网站)和检索词“web”，来获取检索结果。同 *asn.au* 网站样本分析一样，来自 Google 检索结果列表的 40 个随机选取的网站只有两个网站域名不能被收割，样本成功率为 95%。

9.2.4 博客

博客是当前非常有影响的一种网络出版形式，却给域名收割工作带来了一定的困难。在许多方面博客和其他网站一样，在PANDORA存档方面一般不会产生主要的技术问题。多数情况下，博客存在的主要困难在于鉴别其是否属于澳大利亚域名范围，不仅仅需要鉴别服务器位置，还需要鉴别内容和作者。当然一些注册.*au* 域名的博客能够被收割。其他非.*au* 网站上注册澳大利亚IP地址的博客，可以通过自动DNS查找功能被收割。然而许多澳大利亚人的博客位于专门的博客主机上，如Blogger (http://www.blogspot.com/)。这种主要的博客网站，

Google将其识别为澳大利亚以外的网站，因此这样网站上的“澳大利亚”人的博客不能被收割。然而奇怪的是使用博客名录——Blogwise来鉴别一个澳大利亚博客的样本，大量使用非澳大利亚博客主机的网站都能够被收割，但是仅仅能收割第一页。

#### 9.2.5 存有疑问的文档格式

一些分析探索了收割器怎样处理已知有问题文档格式，分析结果是非结论性的，还需要进一步的工作。然而，现有的分析显示，Heritrix 处理有问题文档格式的能力比 HTTrack 强甚至更好。

在 PANDORA 项目中，一些文档由于存在一定问题在域名爬行中不能完整保存。例如 HTTrack，不能成功收割.*ehtml* 扩展名网站。而且，一些用 JavaScript 文档控制选择菜单的网站，HTTrack 同样不能成功收割。

在一些情况下，数据显示 Heritrix 可能执行的更好。例如 Flash 网站使用 Flash 技术传递和链接内容，全部域名收割有时显示执行情况和 HTTrack 相似。即一些内容能够被收割但是不能从保存文档成功实现再次交付（尽管 PANDORA Flash 文档可以使用 HEX 编辑器进行编辑以实现其功能）。使用视觉检测分析 Heritrix 功能是存有疑问和非决定性的，然而数据显示 Heritrix 比 HTTrack 执行效果更好。例如 Bell Shakespeare Company 网站——一个 Flash 网站，在域名收割中收割了 465 个文档，使用 HTTrack 进行的收割测试仅仅收割了 6 个文档，数据说明 Heritrix 性能更好。然而通过视觉检测不可能说明 Heritrix 是否具有更好的解析内在链接的能力。全部域名爬行收割范围和 PANDORA 收割范围是不能严格进行比较的。使用 HTTrack 进行的 PANDORA 收割是以专门参数（聚合文档）为基础，使用一个专门的 URL 列表实现的。这方面的分析还有待进一步探索。

#### 9.2.6 网页限制

IA 提供的 HTTP 响应代码报告显示 444,214 个 URL 报告 401 个响应代码。这是“非授权”的响应代码，包括用户 ID 和密码的服务器响应。这可能是一个不明确指标，却显示仅有全部 URLs 的 0.234％被爬行。

## 10. 结论

### 10.1 优势

全部域名收割的文档数量超过 PANDORA 保存项目 9 年来收割文档总数的 7 倍，也远远超过 2004 年 11 月进行的 Alexa 爬行。爬行结果显示在.*au* 域名覆盖范围的广度上是成功的。

许多网站在分析过程中被访问但是没有被计算，因此不能提供一个全面的精确的数据。然而对于一些专门类型的网站包括非.*au* 网站和 *asn.au* 网站，随机选取样本被收割的成功率达到 95％。相对于全部内容样本比例较小，但是样本是没有任何偏好随机选取的。尽管样本视觉分析过程限制得出明确的结论，数据显示对.*au* 域名的覆盖范围可以持有乐观评价。

另外一个优势在于一定数量的使用澳大利亚 IP 地址的非.*au* 域名能够被自动识别和收割。

实现深度内容收割，一些大型网站显示出相当可观的内容深度。例如，ASIC 网站拥有 11857 个文档。相反 ABS 网站仅仅拥有 5641 个文档，ABS 网站有自动排除目录，可能会限制收割内容。大约 13612 个来自 *nla.gov.au* 域名的文档被收割。

值得指出的是通过 IA 的服务和技术尤其是 Heritrix 和 ARC 文档格式进行的这次收割项目意味着保存文档标准格式正逐渐形成，这与 IIPC 的目标是一致的。

### 10.2 劣势

收割的一个弱点在于不能应用专门的主题聚合过滤器。然而这种大范围收割不需要过滤器来限制爬行，过滤器通常用来促进收割，聚合内容，辅助文档以及在不能成功鉴别和收割

情况下规范文档框架，即过多的依赖爬行器解析代码和跟踪链接的能力。对 robots.txt 排除规则的依赖也会削弱收割的成果，robots.txt 规则在鉴别相关收割内容的多数情况下支持爬行器运行。然而，robots.txt 不是专门直接面向收割器的文档，而是面向一般意义上的爬行器，包括索引器。由于爬行器不能分析规则，因此不能判断哪些文档遵循规则，哪些文档没有，通过遵循 robots.txt 排除规则，一些内容被排除而不能被收割。

通过使用 DNS 查找和解析链接能够获得一定数量的非.*au* 域名的网站，往往需要以 IP 地址为基础进行鉴别。这种收割方法的缺点在于位于澳大利亚之外却涉及澳大利亚内容的网站不能被自动鉴别和包括在收割范围之内。

10.3 计量限制

域名收割分析的一个限制因素在于没有一个现存的标准或计量方法来评估它。即没有直接适用的计量方法来确定目前全部澳大利亚域名爬行的比例关系，也没有可获得的数据来评估独立网站爬行的全面性。因此，爬行的成功与否也不能得出明确的、以统计数据为基础的结论。

大部分可以获得的澳大利亚网络域名是以域名登记或者 IP 地址分配数量为基础的。到 2005 年 9 月 *AusRegistry* 报告有 585756 个注册的.*au* 域名。

(see: http://www.ausregistry.com.au/pdf/PUBLIC200509print.pdf).

这个数据小于实际爬行域名数量，表明相当数量的非.*au* 域名包括在爬行范围之内。

10.4. 获取限制

爬行方法和再次交付机制限制了内容的获取。内容获取将促进纯文本内容索引优先传递给国家图书馆，因为这将为 Wayback 机器 准确获取 URL 提供关键词。关于内容的浏览仍然存在困难，这种困难也许会随着内容获取工具的发展而降低。

## 11. 经验总结

准确测定澳大利亚域名被鉴别和收割的数量是不可能的，但是通过大量的主要目录和使用 Heritrix 进行爬行的这六周时间，可以推断收割数量非常大，特别是在收割广度上。

IA 提供了大量的可获得的主要目录作为爬行内容的补充，对于我们专门需要的内容直接进行爬行收割，这也扩大了爬行的范围。关注目标主要目录的发展是关注未来收割的有效途径。这次爬行的范围广泛，特别关注.*au* 域名的鉴别和自动鉴别位于澳大利亚网络服务器的其他内容。因此，在这个项目初始过程中，澳大利亚域名是从地理上来限定的。很明显，虽然这种方法在内容上是广泛的、真实的和相关的，但是缺少对网站内容的实际考虑，因为大量与澳大利亚相关或者由澳大利亚人发表的内容，其所在的网站服务器位于澳大利亚之外。最明显的例子是 weblogs，而且大量国际性网站使用.*com* and .*net* 域名。服从 robots.txt 规则可能对收割公共性的可获得的内容产生不可预见的限制。网络管理员使用 robots.txt 排除文档可能导致网站的内容不能被收割。

依靠内容的视觉分析比预期的更为困难和局限，主要在于通过 Wayback 机器进行的内容获取和内容交付方式存在局限性。

IA 进行的爬行和收割工作非常有效和成功，按照国家图书馆的要求提供了工作成果。

## 12. 未来选择

这次域名爬行和收割工作为将来大范围爬行提供了一个标准和计量方法，形成了关于澳大利亚网络域名存档快照的馆藏资源。这种资源为将来存档快照的建立提供了良好的基础，将来存档快照应当包括：

- 周期性的爬行规范相似的域名，提供快照年表以及关于域名增长和组成的计量方法
- 关注特别的附属域名的爬行（如.gov, .edu, .org），缩短运行期限，实现完全覆盖
- 拓宽域名爬行识别范围，澳大利亚域名的定义不仅局限于主机服务器的地理位置

这种馆藏资源为提高澳大利亚网络存档效率提供了基础，表现在：

- 这些重要的澳大利亚文化遗产数字化资源，支持对版权法案的修正并允许公众获取这些内容。即这样的资源不再是理论上的，而是实际的能够被研究者和澳大利亚公众获得
- 它为国家图书馆未来网络保存战略的发展提供了真实的和有价值的馆藏资源

为了探索未来选择和明确图书馆的网络保存的未来战略，需要下列行动：

- 扩大电子出版物合法保存的复兴行动
- 对已经编入索引的收割内容进行更为详细的数据分析，更好地理解关于收集和提供获取之间的问题
- 提供有许可的公众获取收割内容的执行工具
- 考虑澳大利亚在线资源域名收割保存未来发展的政策和过程

## 13. 建议

- 纯文本索引完成并将内容传递给国家图书馆后，对收割内容进行更为彻底的进一步分析
- 增强对 ARC 文档内容获取工具的研究和实施
- 利用域名收割文档促进版权法案关于合法保存规定的变化
- 考虑法律的、技术的、组织的问题以及提供公众获取域名收割内容的费用。包括：
  - 同 PANDORA 合作研究是否一些收割内容现在能够合法的提供获取
  - 研究国家图书馆是否可以通过图书馆阅览室提供有一定限制的网络获取
- 以这次收割工作为基础定义能够作为基准的计量方法，为将来大规模域名收割提供计量方法
- 研究以内容和其他非地理因素为基础的澳大利亚域名鉴别和界定的有效的和实用的方法
- 安排 2006 年下半年的第二次域名收割工作

编译自：Paul Koerbin. Report on the Crawl and Harvest of the Whole Australian Web Domain Undertaken during June and July 2005. http://pandora.nla.gov.au/documents/domain_harvest_report_public.pdf . [2006-01-09].

（李麟 校译）