国家图书馆WEB数据增量采集设计及其实现

季士妍 赵丹阳
（国家图书馆，北京 100081）

摘要：本文详细介绍网络资源保存技术策略现状，并从国家图书馆网络资源采集的实际业务需求出发，制定并设计符合国家图书馆业务需求的增量采集技术策略，简述国家图书馆基于Heritrix3.4的增量采集实现方法和实验效果，以期为业界提供有益的参考和借鉴。

关键词：国家图书馆；增量采集；Heritrix
中图分类号：G255　　DOI：10.3772/j.issn.1673-2286.2021.01.005
引文格式：季士妍，赵丹阳. 国家图书馆WEB数据增量采集设计及其实现[J]. 数字图书馆论坛，2021 (1)：32-37.

国际互联网保存事业开始于20世纪90年代，作为国家公共文化服务体系的重要组成部分，国家图书馆最早于2003年开展网络信息资源采集与保存试验项目（Web Information Collection and Preservation，WICP）[1]，针对中国发生的具有较大影响力的重特大事件进行专题采集；同时，对在中国境内注册的政府网站（.gov.cn）进行定期镜像存档。多年来，国家图书馆在采集策略、采集方法、采集效率等方面不断寻求突破。在开展规模性采集中，一方面采取全面采集和重点采集相结合的方式，尽量实现覆盖范围更广的有效采集；另一方面，充分考虑采集内容的延续性和数据的有效性，不断修正网络资源采集范围。在采集方法上，构建全流程、自动化、分布式的网络资源保存与服务平台，高效开展互联网资源采集保存。与此同时，随着互联网信息爆炸式的增长，WEB网站的内容信息在以天甚至小时为单位快速变更，国家图书馆每年进行一次全站采集的频率，很难保障网络资源保存内容的时新性和数据更新的及时性，这样一定程度上减弱了网络资源保存的意义；如果以现有的全站采集方式缩短采集频率，又会面临新的局限和挑战。

无论是受限于国家图书馆采集业务覆盖范围广、采集数据量大、存储硬件资源有限、网络带宽受限等因素，还是归因于WEB网站资源量大、WEB网页多版本重复数据量大、IT网页技术快速发展等客观因素，仍使用WEB网站全站采集与保存的策略与方式，将会缺失较多版本的WEB网站的变化信息，也不能利用有限的存储空间保存更多数量的WEB网站。因此，国家图书馆根据自身业务特点和需求，选择基于Heritrix最新版本，进行程序定制开发，以增量方式采集网站，提高采集效率，灵活数据更新机制，较好地避免了周期性整站采集所产生的大规模网站采集效率低、大量网站重复数据保存以及存储空间利用浪费等问题，以实现国家图书馆网络资源保存项目中大批量WEB网站的增量采集和有效管理。

1 国内外网络资源保存现状分析

网络信息时代，互联网信息逐渐成为人类文明记忆的载体，及时记录时代记忆、保留网络烙印变得非常重要。但随着互联网信息的爆炸式增长，WEB信息以指数级速度发生着变化，新数据快速增长，旧数据不断更新与迭代，给Web Archives机构带来巨大的压力和挑战。因此如何优化网络资源采集策略和采集技术以广泛和完整地保存互联网资源成为业界研究热点。

1.1 国外网络资源保存现状分析

近年来，国外的Web Archives机构对互联网资源的保存展开了广泛的研究并积极开展多个项目，多个国家均开展了网络资源采集与保存相关工作，并从本国的实际出发，不断扩大保存范围、丰富保存资源类型、完善采集技术，WEB数据“增量采集”[2-4]也是各个国家图书馆探讨的热点和研究实践的方向。美国IA（Internet Archive）[5]是目前世界上规模最大的网络资源存档与研究项目，其收集的网络资源内容丰富，包括网页、音乐、视频资料等内容，主要通过网络爬虫自动收集，使用Wayback Machine实现网页回放。美国国会图书馆2000年开始启动MINERVA项目[6]，通过图书馆、大学、档案馆、出版社、企业等机构的合作，共同制定全国范围的资源采集、保存战略，并研究最新网络资源采集技术，鉴别濒临消失的网络资源。英国开展网络资源保存项目UK Web Archive[7]，对英国网站进行选择性保存，并联合大英图书馆（British Library）、英国国家档案馆（National Archives）等机构共同制定采集策略。澳大利亚网络档案馆开展了Pandora Archive项目[8]，基于IIPC的开源软件开发Pandas系统，以支持图书馆和其他项目参与者采用的选择性网络存档方法所涉及的任务和工作流程。日本国立国会图书馆启动WARP（Web Archiving Project）项目，颁发了修订后的《国立国会图书馆法》[9]，明确了日本国立国会图书馆可以收集、保存国家公共机构的网站信息，积极推进相关立法、灵活采用采集策略、充分挖掘采集资源。

1.2 国内网络资源保存现状分析

在我国，图书馆、高校等科研机构开展了一系列网络信息采集研究与部分实践。北京高校图书馆网络“教学科研数字图书馆”研究项目[10]从文献信息采集与整合的总体思路、采集的基本步骤、加工的标准与规范、操作规范等方面进行了数字资源采集与整合。中国科学院文献情报中心开展“国际重要科技机构网络信息采集与存档管理平台”项目[11]，基于IIPC[12]基本工具的选择和使用，搭建技术平台，探索了重要网络科技信息资源的保存和利用。国家图书馆于2003年开展WICP，开始实验性地对中国境内的互联网资源进行采集与保存；在2005年将网络信息采集成果服务网站上线，提供热点专题和政府网站存档资源浏览服务。此后国家图书馆一直在致力于中国的网络信息资源采集与保存相关工作的研究与实践。

可以看到，国际上网络资源采集与保存工作发展迅速，部分国家已经将网络资源的保存上升到国家战略层面，而这些国家的网络资源保存工作能够大规模开展、不断地进行研究和优化并取得巨大成果，正是与国家层面的相应政策以及图书馆、保存机构，甚至企业之间的密切合作和联合行动密不可分的。国内在网络资源保存方面的研究和实践上相对起步较晚，网络信息资源的保存环境缺乏政策扶持，舆论上也缺乏社会和公众对网络资源存档的认可；各个研究机构相对独立，研究和实验的展开多以单一角度或单一资源类型为切入点，只有极个别的机构开展了规模性的网络资源保存项目，研究在技术上借鉴国外同类项目经验的同时，缺乏连贯性、深入性和完善性。放眼全球，网络资源的保存不是一个机构能够做的，需要战略指导以及分地区、跨领域的合作，自上而下地大范围地开展规模性的、适应国情的网络资源保存项目，并在项目中不断完善和优化保存策略和技术。

2 网络资源采集技术分析

2.1 网络资源采集策略分析

网络资源保存涉及很多的策略技术问题（如网络资源重要性评估、网页资源的深层挖掘等），由于各个国家的技术架构不同，其采用的基础技术策略和采集方法也不同，如美国主要采用压缩文档的方法和基于特征提取的方法；澳大利亚Pandora Archive项目主要采用多文件采集；瑞典、挪威、芬兰等主要基于文件格式进行迁移。网络资源保存的技术策略分类没有统一的标准，目前较为认可的包括基于整个WEB的信息采集、面向主题的WEB信息采集[13]、增量式WEB信息采集、分布式WEB信息采集[14]、个性化的WEB信息采集[15]、多技术结合的信息采集[16]和基于元搜索的信息采集[17]等。

2.2 网络资源采集工具分析

在网络资源保存的获取工具方面，有基于Java语言的Heritrix、Apache Nutch等，基于Python语言的Scrapy、cola等，基于C#的ccrawler等，但国际上主流Web Archive机构均采用国际互联网保存联盟IIPC研发的开源工具Heritrix[18]，可实现抓取完整的站点内

容，支持网络资源采集的爬虫定义和网页过滤技术，具有较为高效的可配置功能。同时，Heritrix采用优先算法，技术模块包括核心类和插件模块，用户可以自行选择模块，也可以在此基础上进行不同需求的定制开发。IIPC也开发了网络资源存档相关软件工具，包括存储方面的工具HTTrack2ARC、WARC Tools等，其数据封装有WARC\ZIP、AFF等不同格式，国际上广泛应用的主要为WARC[19]格式，以实现对内容提取、相关主题验证、数据存储、数据封装和格式转换等各种功能的操作。同时，IIPC各成员机构根据自身的业务需求，在IIPC开源工具的框架上独自开发相关软件，最为广泛使用的软件是Wayback Machine，其次还有Time Travel Portal、Nutch WAX等。随着技术的发展，IIPC致力于优化WEB存档的工具和保存标准，不断更新完善Heritrix版本，满足各个机构的实际网络资源保存业务的个性化定制和开发需求，促进国际合作。

3 国家图书馆增量采集技术路线设计

增量采集方式替代原有整站采集方式，可以有效节省采集服务器的存储空间资源和网络带宽资源，缩短周期性采集整站的时间，提高采集效率。

3.1 采集思路

国家图书馆采用的WEB网页增量采集，是在采集整站WEB网页数据基础上，以采集新出现的网页、变更的网页和已经删除的网页为目标的采集，且以主流静态索引类型网站为主，保障增量采集数据内容不遗漏、不重复。基于需要实现的采集目标和主流静态索引型网站的特点，考虑将网站划分为网站索引页、索引页链接、页面内容和内容变更频率4个部分分别制定识别策略，来分流出采集目标并做采集性能优化处理；同时，考虑到国家图书馆网络资源存档的连续性、现有的硬件环境以及项目后续的优化和深入，综合选择高版本、高性能的采集工具，结合制定的增量采集策略完成程序的定制开发，以实现增量采集目标、提高硬件运行效率和规模性的生产采集。

3.2 采集工具选择

Heritrix是基于Java语言实现的开源网络爬虫软件，当前主要有Heritrix 1.14.4和Heritrix 3.x两个主要版本。国家图书馆2016年开发的网络资源保存与服务平台是基于Heritrix 1.14.4版本实现的个性化采集模块程序开发，考虑到Heritrix 1.14.4具有AdaptiveRevisitingFrontier模块，根据自定义的间隔时间重复访问解析到的URI，进而可以实现网页资源的增量采集。但是AdaptiveRevisitingFrontier模块对URI的重复访问功能，只在同一个采集任务周期内有效，当该采集任务自动结束或管理员手动终止时，Frontier中记录的所有URI的信息会被清除，进而导致Heritrix 1.14.4的增量功能不稳定且无法满足实际的工作需求。因此，国家图书馆在增量采集的设计和实现上，根据采集业务的要求选择了合适的Heritrix版本，舍弃了一直沿用的Heritrix 1.14.4版本，采用Heritrix 3.4版本，以实现增量采集模块的定制开发。

Heritrix 3.4主要由WEB控制台Web Administrative和爬虫功能模块Engine构成。Engine主要由中央控制器CrawlController、边界控制器Frontier、中央处理器组Processor Chain等部分组成。工作流程如图1所示。

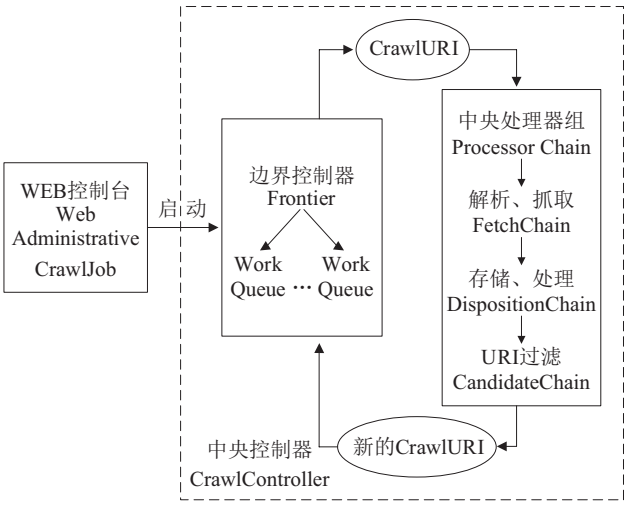

图1 Heritrix工作流程

边界控制器Frontier内部存储了很多的URI队列，并对URI进行调度，根据Heritrix相应的策略，分配不同的URI到WorkQueue中，然后根据不同的策略从WorkQueue中取出URI送给中央处理器组Processor Chain，中央处理器组Processor Chain中的FetchChain、DispositionChain和CandidateChain处理链分别对接收到的URI进行解析、抓取、过滤、存储等操作，将筛选出来的新的URI再送回至边界控制器Frontier中，进行下一个操作循环。

3.3 采集策略设计

国家图书馆基于Heritrix 3.4版本定制开发的增量采集要实现的目标包括以下内容。首先，整体程序对所运行的环境较为友好，可以运行在一个较低成本的硬件环境中；整体程序在运行中较少占用服务器内存，进而可以提升服务器的运行效率。其次，开发实现的增量采集功能，可以面向主流类型网站实现更新网页的识别、采集，增量采集实现内容不重复、不遗漏。最后，开发实现的增量采集功能，可以支持多任务、高并发的采集任务安排，对不同的采集任务可以较为灵活地配置采集参数。

3.3.1 索引页更新识别

当前互联网上大多数网站提供的均是索引型网页，即网页展示内容不是由页面直接提供，而是通过程序生成链接进而将网页内容呈现出来。国家图书馆在实现网站采集时，首先抓取网站的索引页作为该网站的种子链接，将网站首页索引页以及网站各版面索引页一同形成该网站的原始种子列表。后续在实现增量采集时，根据种子列表进行判断，将所有发生变化的链接均加入任务队列，最终通过任务队列进行全部采集和保存。

3.3.2 索引页链接更新识别

制定程序建立了数据库管理和记录比对所有的索引页中的链接数据，判断并产生更新种子链接列表。数据库中的每一个链接数据以时间戳记录、ID顺序记录、无规律的记录3种规则来识别该链接的唯一性。时间戳记录和ID顺序记录规则可以以上一次采集周期结束为标记判断是否有更新的链接；无规律的记录规则以压缩后的URI字节数/组作为Key值，比对Key值取得发生更新的URI。

3.3.3 页面内容相似性识别

制定程序通过建立阈值识别方法，来准确识别网页内容的更新。针对网页内容，通过过滤网页的干扰并提取该网页内容的主题词，设定该页面的更新阈值，后续可以对不同时刻的页面内容进行比对，一旦更新值超过阈值，就表示该页面内容出现了更新，需要进行更新采集。

3.3.4 页面变化频率识别

所有网站的页面更新，都是有其自身规律的。国家图书馆在目前阶段将继续沿用原有的采集策略，综合考虑采集网站的数据量大小、更新时间等因素，人为设定网络更新采集频率。与此同时，还将建立一个网站更新频率数据库，详细记录网站数据量、更新频率、更新量等，进而为未来利用机器学习的方式自动检测网站更新频率、自动触发增量采集打下基础。

4 国家图书馆增量采集技术实现

国际上广泛应用的网络资源保存格式主要为WARC格式，为方便国际交流，国家图书馆十多年来一直采用WARC格式保存，增量采集选用WARCWriter Processorwen文件处理器。遵照增量采集设计策略，网站增量采集实现流程如图2所示。

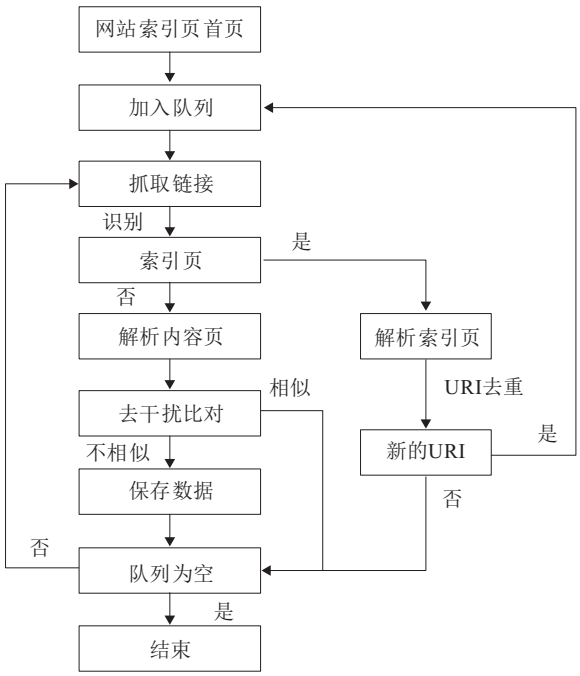

图2 网站增量采集流程

网站增量的完整过程是一个多次判断、多次循环的过程。在网站采集任务开始时，首先将待保存网站所有页面的索引页的首页作为种子链接，形成网站的种子列表。当采集任务启动后，系统按照采集队列中的种子列表，依次采集每一个链接，并根据链接地址字符识别采集页面是索引页还是内容页。如果是索引页，则根据定制的增量采集逻辑解析页面中的目录链接，过滤出

更新链接加入采集队列，进行下一次循环，直到采集队列为空，该采集任务结束。如果是内容页，则需按照采集内容阈值设定进行对比判断，如为不相似内容则触发更新采集，如为相似内容则不更新采集，依次循环，直到采集队列为空，任务结束。

在网络资源采集任务管理上，定制程序对于增量采集与全站采集采用了不同的管理策略，对增量采集任务实施任务标记，进而明确区分增量采集与全站采集两种不同的采集任务。Engine是Heritrix的主要爬虫功能模块，通过Heritrix的CrawlJob创建一个常规的爬虫任务后，Heritrix会在系统目录下创建任务文件。本定制程序实现的增量采集以创建文件夹的形式，以上一次采集开始为时间准则，在任务清单中添加爬虫任务。增量采集任务的标记放在任务启动开始前，当采集人员通过任务界面启动增量采集任务时，系统会发送请求，根据jobName区分爬虫任务，然后将对应的爬虫任务启动，同时调用Engine创建增量采集任务的接口，在增量采集过程中noteFrontierState会调用接口中jobState状态。在增加采集过程中，定制程序实现了索引页识别、链接去重、页面内容过滤等功能。

（1）索引页识别。索引页的识别和采集主要通过Heritrix的页面处理接口实现，并可以根据链接字符区分索引页和内容页进行处理。定制程序实现了html格式的网站索引页的处理，利用Java自带包解析html，导入Java包javax.swing.text.*和javax.swing.text.html.*，利用Heritrix的org.archive.modules.extractor包进行定制。

（2）链接去重。定制程序通过Heritrix中的org.archive.crawler.datamodel.CrawlURI类实现网页链接去重，将采集任务处理的变量以及URI相关的数据全部封装在CrawlURI中。CrawlURI将页面中解析出的新链接放到outLinks中，在CandidatesProcessor执行DecideRuleSequence对Link进行过滤，将符合要求的链接持续加入Frontier，进而执行采集循环。

（3）页面内容过滤。增量采集执行的程序以原始网页源码中<p>或者
或者标签中间的内容为增量网页采集的目标内容数据，程序通过MD算法对目标数据内容进行信息摘要，生成一个128bit的定长字符串保存到数据库中。当实施增量采集时，程序只需要比对待采集的网页与数据库中记录的MD编码值是否一致即可，如果MD编码值不一致，表示产生了页面内容的改变，则触发增量采集任务的启动。

5 国家图书馆增量采集示范

国家图书馆用增量采集的方式实验性地采集了包括文化和旅游部、中国政府网、外交部、教育部等政府门户类网站，人民网、新华网等新闻类网站，北京北图文化发展中心、国家图书馆等文化服务类网站。以中国政府网（https://www.gov.cn/）以及文化和旅游部网站（https://www.mct.gov.cn/）为例，由于网站自有策略、国家图书馆网络带宽流量控制策略和馆网用网高峰流量控制等因素限制，因此，单位时间的采集数据量不同，不同网站单位时间采集数据量不能同向比较。

中国政府网采集深度参数设置为3层，为保证数据的时新性，以季度为采集频率周期进行全站采集，第一次全站采集数据量达6 144MB，采集时长47小时，采集HTML数量达7 436页；第二次全站采集数据量为7 356MB，采集时长60小时，采集HTML数量达8 082页；但是，采用增量采集策略后（增量采集在第一次全站采集基础上进行），数据量为455MB，采集时长为8小时，采集HTML数量1 820页，采集时长缩短83%，存储节约93%。

文化和旅游部网站采集深度参数设置为2层，第一次全站采集数据量达514MB，采集时长14小时，采集HTML数量达1 971页；第二次全站采集数据量达638MB，采集时长25小时，采集HTML数量达2 488页；但第二次采用增量采集策略后，数据量为53.5MB，采集时长4小时，采集HTML数量194页，采集时长缩短71%，存储节约90%。

从上面的对比数据可以很明显地看出，使用同一采集策略和方式，不同的采集深度，网站的采集数据量和时长均不同。虽然采集网站深度为2层所采集的数据量相较于3层深度数据量骤降，但是2层深度不能很好地保证存档网站的原貌，因此在实际工作中，国家图书馆的网站保存多采用3~5层深度。在采用WEB网站全站采集方式时，由于全站采集获得的数据大部分为重复性数据，进而导致前后两次全站采集的数据量相近。因此，如果长期多次采用WEB网站全站采集的方式，会给存储带来相当大的压力，并且采集工作会长期占用网络带宽资源。与之形成较为明显的对比，则为WEB网站增量采集方式，这种采集模式会大大减少网络资源存档量、较为显著地缩短采集时间，其采集效率比整站采集明显提升，不仅保证了采集保存的网站数据的时新性，而且采集服务器存储占用骤降，有效地解决了全

站采集带来的存储不够和带宽限制的问题，明显地提高了工作效率。

6 结语与展望

国家图书馆基于Heritrix 3.4实现了网页增量采集和数据管理，可以针对静态索引类型的网站实现规模性的网站增量采集，保证资源内容时新性的同时，极大地缓解了存储空间的压力和网络带宽的负担。国家图书馆还将继续完善增量采集的方法，如利用机器学习方法进行网页更新频率管理，增加针对多种类型相匹配的增量采集程序，将增量采集程序包装成独立的软件包进而可以分享给其他机构进行增量采集操作；同时，根据时代网络资源的特点，针对更丰富的资源类型进行采集方法的探究，如移动端资源的采集研究等，以期为业界提供参考和借鉴。

参考文献

[1] 张炜，张文静. 中国网络信息采集工作研究现状分析 [J]. 图书馆建设，2008（7）：43-51.

[2] CHO J，GARCIA-MOLINA H. The evolution of the web and implications for an incremental crawl [C] //In Proceedings of the 26th International Conference on Very Large Database，Cairo：2000.

[3] 杨颂，欧阳柳波. 基于Heritrix的面向电子商务网站的增量爬虫研究 [J]. 软件导刊，2010，9（7）：38-39.

[4] Managing duplicates in a web archive [EB/OL]. [2020-11-16]. https://www.docin.com/p-1406798808.html.

[5] Internet Archive's Terms of Use，Privacy Policy，and Copyright Policy [EB/OL]. [2020-08-01]. https://archive.org/about/terms.php.

[6] For Researchers [EB/OL]. [2020-11-16]. https://www.loc.gov/programs/web-archiving/about-this-program/.

[7] 徐健. 英国网络信息保存联盟计划（UKWAC）及其启示 [J]. 图书馆论坛，2007（2）：81-84

[8] About Pandora [EB/OL]. [2020-11-16]. http://pandora.nla.gov.au/about.html.

[9] 张妍，程瑾，刘雪涛，等. 日本国立国会图书馆网络资源收集保存事业（WARP）及其启示 [J]. 中华医学图书情报杂志，2017（12）：41-45.

[10] 林屹，李军英，熊丽，等. 北京高校图书馆网络"教学科研数字图书馆"研究项目信息采集与整合的解决方案 [J]. 现代图书情报技术，2003（5）：86-88.

[11] 吴振新，谢靖，张智雄，等. 国际重要科研机构网络信息采集与存档管理平台构建 [J]. 图书情报工作，2014（11）：100-104.

[12] IIPC [EB/OL]. [2020-11-05]. http://netpreserve.org/.

[13] 李春旺. Web信息主题采集技术研究 [J]. 图书情报工作，2005（4）：77-80.

[14] 金岳富，范剑英，冯扬. 分布式Web信息采集系统的设计与实现 [J]. 哈尔滨理工大学学报，2010（2）：116-123.

[15] 吴丽辉，王斌，张刚. 一个个性化的Web信息采集模型 [J]. 计算机工程，2005（11）：86-88.

[16] 王征清，成全. 基于Multi-Agent的分布式主题信息采集结构模式研究 [J]. 情报理论与实践，2007（3）：419-422.

[17] 王忠，程磊. 基于元搜索引擎的个性化Web信息采集 [J]. 计算机工程与设计，2009（13）：3117-3119.

[18] Heritrix [EB/OL]. [2020-11-16]. https://Webarchive.jira.com/wiki/display/Heritrix/Heritrix.

[19] ISO 28500:2009（en）Information and documentation—WARC file format [EB/OL]. [2020-11-16]. https://www.iso.org/obp/ui/#iso:std:iso:28500:ed-1:v1:en.

作者简介

季士妍，女，1978年生，硕士，副研究馆员，研究方向：数字资源保存和服务、数字图书馆资源整合，E-mail：jisy@nlc.cn。
赵丹阳，女，1988年生，硕士，馆员，研究方向：数字资源保存和服务、数字图书馆资源整合，E-mail：zhaody@nlc.cn。

Design and Implementation on the Web Data Deduplicated Crawlers of the National Library of China

JI ShiYan　ZHAO DanYang
(National Library of China, Beijing 100081, China)

Abstract: This paper introduces the current situation of web archiving technology strategy in detail, and designs the deduplicated crawlers technology strategy based on the actual practices of web archiving in the National Library of China. It describes the realization method of duplicated crawlers based on heritrix 3.4, so as to provide useful reference for the industry.

Keywords: National Library of China; Duplicated Crawlers; Heritrix

（收稿日期：2020-12-12）